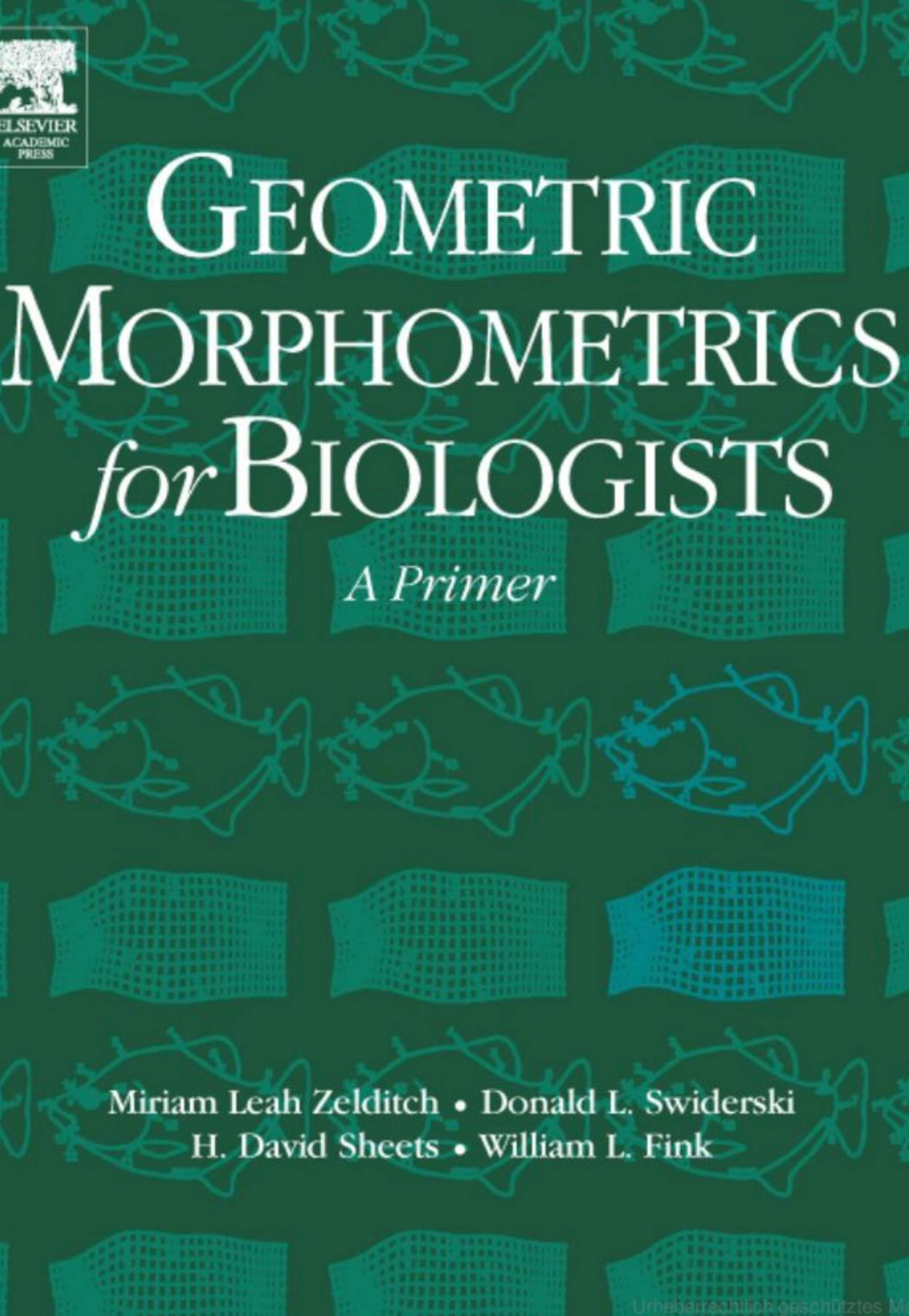
ELSEVIER
ACADEMIC
PRESS



Geometric Morphometrics for Biologists

A Primer

Miriam Leah Zelditch • Donald L. Swiderski
H. David Sheets • William L. Fink

Urheberrechtlich geschütztes M

GEOMETRIC MORPHOMETRICS FOR BIOLOGISTS: A PRIMER

GEOMETRIC MORPHOMETRICS FOR BIOLOGISTS: A PRIMER

Miriam Leah Zelditch, Donald L. Swiderski,
H. David Sheets and William L. Fink

ELSEVIER
ACADEMIC
PRESS
Amsterdam • Boston • Heidelberg • London • New York • Oxford • Paris
San Diego • San Francisco • Singapore • Sydney • Tokyo

This book is printed on acid-free paper

Copyright © 2004, Elsevier (USA). All rights reserved

No part of this publication may be reproduced, stored in a retrieval system,
or transmitted in any form or by any means electronic, mechanical, photocopying,
recording or otherwise, without the prior written permission of the publisher.

Permissions may be sought directly from Elsevier's Science & Technology
Rights Department in Oxford, UK: phone (+44) 1865 843830,
fax: (+44) 1865 853333, e-mail: permissions@elsevier.co.uk.
You may also complete your request on-line via the Elsevier homepage
(http://www.elsevier.com), by selecting 'Customer Support' and then
'Obtaining Permissions'

Elsevier Academic Press
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
http://www.elsevier.com

Elsevier Academic Press
84 Theobald's Road, London WC1X 8RR, UK
http://www.elsevier.com

Library of Congress Catalog Number: 2004 107927

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0–12–77846–08

Printed and bound by Maple-Vail, Kirkwood, New York, USA

04 05 06 07 08 09 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries

www.elsevier.com | www.bookaid.org | www.sabre.org

ELSEVIER BOOK AID International Sabre Foundation

Contents

List of contributors vii

Preface ix

Abbreviations xi

1 Introduction 1

Part I Basics of Shape Data 21

2 Landmarks 23

3 Simple size and shape variables: Bookstein shape coordinates 51

4 Theory of shape 73

5 Superimposition methods 105

6 The thin-plate spline: visualizing shape change as a deformation 129

Part II Analyzing Shape Variables 153

7 Ordination methods 155

8 Computer-based statistical methods 189

9 Multivariate analysis of variance 209

10 Regression 229

11 Partial least squares analysis 261

Part III Applications of Morphometric Methods to Complex Hypotheses 291

12 Disparity and variation 293

13 The relationship between ontogeny and phylogeny 321

14 Morphometrics and systematics 363

Part IV Last Things 383

15 Beyond two-dimensional configurations of landmarks 385

Glossary 409

Bibliography 429

Index 437

Contributors

Miriam Leah Zelditch is an Associate Research Scientist at the Museum of Paleontology, University of Michigan, MI. She studies the evolution of ontogeny, focusing on the developmental basis of morphological disparity and variance.

Donald L. Swiderski is the Adjunct Research Investigator, Mammal Division, at the Museum of Zoology, University of Michigan, MI. He is an evolutionary morphologist, interested in the relationships between the morphological and ecological diversity of mammals.

H. David Sheets is Chair of the Department of Physics at Canisius College, Buffalo NY. A physicist by training, his interest in dynamical processes led to work on the processes and patterns of evolutionary change. The need for robust, powerful methods to characterize shape spurred an interest in geometric morphometrics, leading to the development of analytic methods and the Integrated Morphometrics Programs software series.

William L. Fink is the Professor of Biology at the University of Michigan, and the Curator of Fishes at the Museum of Zoology, University of Michigan, MI. His research interests include the biology and systematics of fishes (particularly neotropical species), higher classification, and the theory of systematics and biogeography.

Preface

This is a textbook on shape analysis for biologists, covering both its basic theory and its practice. We think that a textbook is needed because the field has changed tremendously in recent years, primarily due to rapid developments in the mathematical theory of shape. Most of the work done during the past decade has focused on the mathematical foundations of methods, or on extensions of theory to ever more mathematically complex situations, and so most treatments of the subject have been written by and for mathematicians. Biologists reading those treatments may find the subject abstract, even obscure and mysterious, when the reality is that conducting a biological shape analysis requires no more background in mathematics and statistics than most biologists acquire in their undergraduate training. The discouragement caused by reading the often highly technical works in the field is unfortunate, because the tools of shape analysis have great utility in biology. Furthermore, a major achievement of geometric morphometrics is the ability to draw pictures of morphological transformations; we can literally *see* one morphology transforming into another, which should make geometric morphometrics intuitively accessible to morphologists. Accordingly, we have written this primer emphasizing applications to biological questions and illustration of results, and we have written it presuming that the reader's background consists only of a basic course in statistics and some familiarity with elementary geometry and algebra.

We wrote this for largely selfish reasons: we teach morphometrics to advanced undergraduate and graduate students, and we needed a textbook. Like many biologists, our students ask sophisticated biological questions and require methods that can answer those questions, but have little (if any) experience with matrix algebra, non-Euclidean geometry or multivariate statistics. Also, like many others who are learning new methods, our students want to apply them (often immediately on learning them), not just to learn their rationales. Accordingly, we have emphasized the biological questions answered by various methods and provided examples of applications to both simple and complex biological questions. We also make software available electronically to conduct all the analyses explained in the book, and incorporate the manuals within each chapter. This combination of explanation of methods, examples, illustration of results, software and manuals will allow students to begin applying the methods as they are learning them.

We strongly encourage all students (and faculty) to begin collecting data as soon as possible – ideally while reading Chapter 2. Although some datasets are contained within the software packages so that you can practice analyzing them, it is best to practice on your own data. They are more familiar and far more interesting to you than any we can provide. Consequently, you will learn the material most quickly and thoroughly by applying it to your own organisms, in context of your own biological questions.

We are very grateful to our students, both those in our regular courses at the University of Michigan and the State University of New York, Buffalo and Canisius College, and those who participated in our workshops at the University of California, Berkeley and the University of Massachusetts, Amherst. They all improved the text considerably. We are also grateful to others who read earlier versions of this text and pointed out errors and ambiguities, and also asked probing questions that sometimes required us to rethink as well as rewrite. We especially thank Barbara L. Lundrigan, Ian Dworkin, and Rebecca German and her laboratory group. We also thank Jason Mezey and Markus Bastir, who provided prepublication versions of the methods for comparing subspaces (Chapter 12) and for evaluating competing hypothesis of morphological integration (Chapter 11), respectively. Finally, we are especially grateful to Dr Charles R. Crumly of Academic Press, who worked with us patiently, but not too patiently.

M.L.Z., D.L.S., H.D.S., W.L.F.
April 2004

Abbreviations

The following abbreviations or symbols are used regularly throughout this book, or are used commonly in statistics or morphometrics. Definitions are in the Glossary.

BC	Bookstein coordinates
BTR	Bookstein two-point registration, or Bookstein coordinates
CS	Centroid size
CVA	Canonical variates analysis
D_F	Full Procrustes distance
df, d.f., or dF	Degrees of freedom
df1, df2	Degrees of freedom for the within- and between-group factors in an F-test
D_p	Partial Procrustes distance
GLS	Generalized least squares
K	Number of landmarks used in a study
LCS	Logarithm (to base(e) or base(10)) of centroid size
MD	Morphological disparity
m	(1) The number of dimensions of a landmark (either two or three), or (2) slope of regression line, as in $Y = mX + b$
P	Procrustes distance
PCA	Principal components analysis
RFTRA	Resistant-fit theta-rho analysis
SBR	Sliding baseline registration
SVD	Singular value decomposition
V–C, V/C matrix	Variance–covariance matrix
Var–Covar	Variance–covariance, as in variance–covariance matrix
X	A component along the Cartesian X-axis
Y	A component along the Cartesian Y-axis
Z	Typically represents a complex number, but can also be a component along the Cartesian Z-axis in analyses of three-dimensional data
δ	Delta, used to indicate a small change
Σ	The variance–covariance matrix
$\sum_{i=1}^{n}$	The summation from 1 to n
Λ	Lambda, usually, Wilk's Lambda

1

Introduction

Shape analysis plays an important role in many kinds of biological studies. A variety of biological processes produce differences in shape between individuals or their parts, such as disease or injury, ontogenetic development, adaptation to local geographic factors, or long-term evolutionary diversification. Differences in shape may signal different functional roles played by the same parts, different responses to the same selective pressures (or differences in the selective pressures themselves), as well as differences in processes of growth and morphogenesis. Shape analysis is one approach to understanding those diverse causes of variation and morphological transformation.

Frequently, differences in shape are adequately summarized by comparing the observed shapes to more familiar objects such as circles, kidneys or letters of the alphabet (or even, in the case of the Lower Peninsula of Michigan, a right-handed mitten). Organisms, or their parts, are then characterized as being more or less circular, reniform or C-shaped (or mitten-like). Such comparisons can be extremely valuable because they help us to visualize unfamiliar organisms, or focus attention on biologically meaningful components of shape. However, they can also be vague, inaccurate or even misleading, especially when the shapes are complex and do not closely resemble familiar icons. Even under the best of circumstances, we still cannot say precisely how much more circular, reniform, or C-shaped (or mitten-like) one shape is than another. When we need that precision, we turn to measurement.

Morphometrics is simply a quantitative way of addressing the shape comparisons that have always interested biologists. This may not seem to be the case because conventional morphological approaches typical of the qualitative literature and traditional morphometric studies appear to produce quite different kinds of results. The qualitative studies produce pictures or detailed descriptions (in which analogies figure prominently), and the morphometric studies usually produce tables with disembodied lists of numbers. Those numbers seem so highly abstract that we cannot readily visualize them as descriptors of shape differences, and the language of morphometrics is also highly abstract and mathematical. As a result, morphometrics has seemed closer to statistics or algebra than to morphology. In one sense that perception is entirely accurate: morphometrics *is* a branch of mathematical shape analysis. The ways we extract information from morphometric

Geometric Morphometrics for Biologists
ISBN 0–12–77846–08

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

data involve mathematical operations rather than concepts rooted in biological intuition or classical morphology. Indeed, the pioneering work in modern geometric morphometrics (the focus of this book) had nothing at all to do with organismal morphology; the goal was to answer a question about the alignment of megalithic "standing stones" like Stonehenge (Kendall, 1977; Kendall and Kendall, 1980). Nevertheless, morphometrics can be a branch of morphology as much as it is a branch of statistics.

This is the case when the tools of shape analysis are turned to organismal shapes, and when those tools allow us to illustrate and explain shape differences that have been mathematically analyzed. The tools of geometric shape analysis have a tremendous advantage when it comes to these purposes: not only does this method offer precise and accurate description, but also it serves the equally important purposes of visualization, interpretation and communication of results. Geometric morphometrics allows us to visualize differences among complex shapes with nearly the same facility as we can visualize differences among circles, kidneys and letters of the alphabet (and mittens).

In emphasizing the biological component of morphometrics, we do not discount the significance of its mathematical component. Mathematics provides the models used to analyze data, including the general linear models exploited in statistical analyses, and the models underlying exploratory methods (such as principal components analysis). Additionally, mathematics provides a theory of measurement that we use to obtain data in the first place. It may not be obvious that a theory governs measurement, because very little (if any) theory underlay traditional measurement approaches. Asked the question "What are you measuring?", we could give many answers based on our biological motivation for measurement – such as (1) "Functionally important characters;" (2) "Systematically important characters;" (3) "Developmentally important characters;" or, more generally, (4) "Size and shape." However, if asked "What do you mean by 'character' and how is that related, mathematically or conceptually, to what you are measuring?", or even if just asked "What do you mean by 'size and shape'?", we could not provide theoretically coherent answers. A great deal of experience and tacit knowledge went into devising measurement schemes, but they had very little to do with a general theory of measurement. It was almost as if each study devised its own approach to measurement according to the particular biological questions at hand. There was no general theory of shape, nor were there specialized analytic methods adapted to the characteristics of shape data.

The remarkable progress in morphometrics over the past decade resulted largely from precisely defining "shape," then pursuing the mathematical implications of that definition. The most fundamental change has been in measurement theory. Below we offer a critical overview of the recent history of measurement theory, presenting it first in terms of exemplary data sets and then in more theoretical terms, emphasizing the core of the theory underlying geometric morphometrics – the definition of shape. We conclude the conceptual part of this Introduction with a brief discussion of methods of data analysis. The rest of the Introduction is concerned with the organization of this book, and available software and other resources for carrying out morphometric analyses.

A critical overview of measurement theory

Traditionally, morphometric data have been measurements of length, depth and width, such as those shown in Figure 1.1, which is based on a scheme presented in a classic

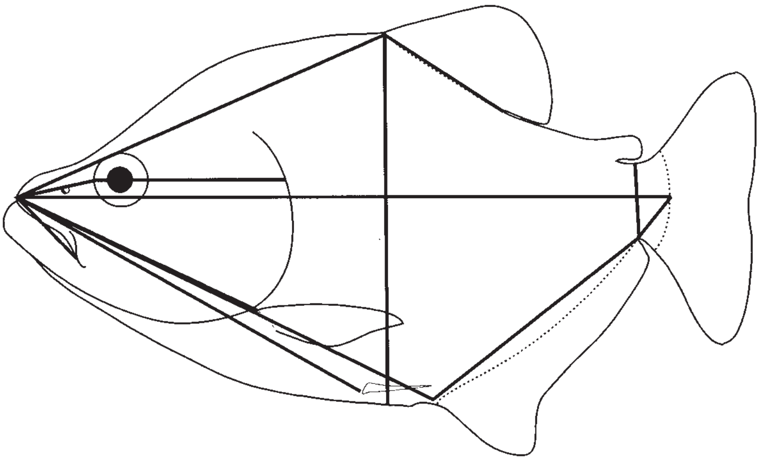


Figure 1.1 Traditional morphometric measurements of external body form of a teleost, adapted from the scheme in Lagler et al., 1962.

ichthyology text (Lagler et al., 1962). Such a data set contains relatively little information about shape, and some of that information is fairly ambiguous. These kinds of data sets contain less information than they appear to hold because many of the measurements overlap or run in similar directions. Several of the measurements radiate from a single point, so their values cannot be completely independent (which also means that any error in locating that point affects all of these measurements). Such a data set also contains less information than could have been collected with the same effort, because some directions are measured redundantly, and many of these measurements overlap. For example, there are multiple measurements of length along the anteroposterior body axis and most of them cross some part of the head, whereas there are only two measurements along the dorsoventral axis, and only two others that are measurements of post-cranial dimensions. In addition, the overlap of the measurements complicates the problem of describing localized shape differences like changes in the position of the dorsal fin relative to the back of the head. Also missing from this type of measurement scheme is information about the spatial relationships among measurements. That information might be given in the descriptions of the measured line segment, but it is not captured in the list of observed values of those lengths, which are the data that are actually available for analysis. Finally, the measurements in this scheme may not sample homologous features of the organism. Body depth can be measured by a line extending between two well-defined points (e.g. the anterior base of the dorsal fin to the anterior base of the anal fin), but it can also be measured wherever the body is deepest, yielding a measurement of "greatest body depth" wherever that occurs. This measurement of depth might not be comparable anatomically from species to species, or even from specimen to specimen, so it provides almost no useful information. When all of the limitations of the traditional measurement scheme are considered, it is apparent that the number of measurements greatly overestimates the amount of shape information that is collected.

The classical measurement scheme can be greatly improved, without altering its basic mathematical framework, by the box truss (Figure 1.2) – a scheme developed by Bookstein and colleagues (Strauss and Bookstein, 1982; Bookstein et al., 1985). This set of

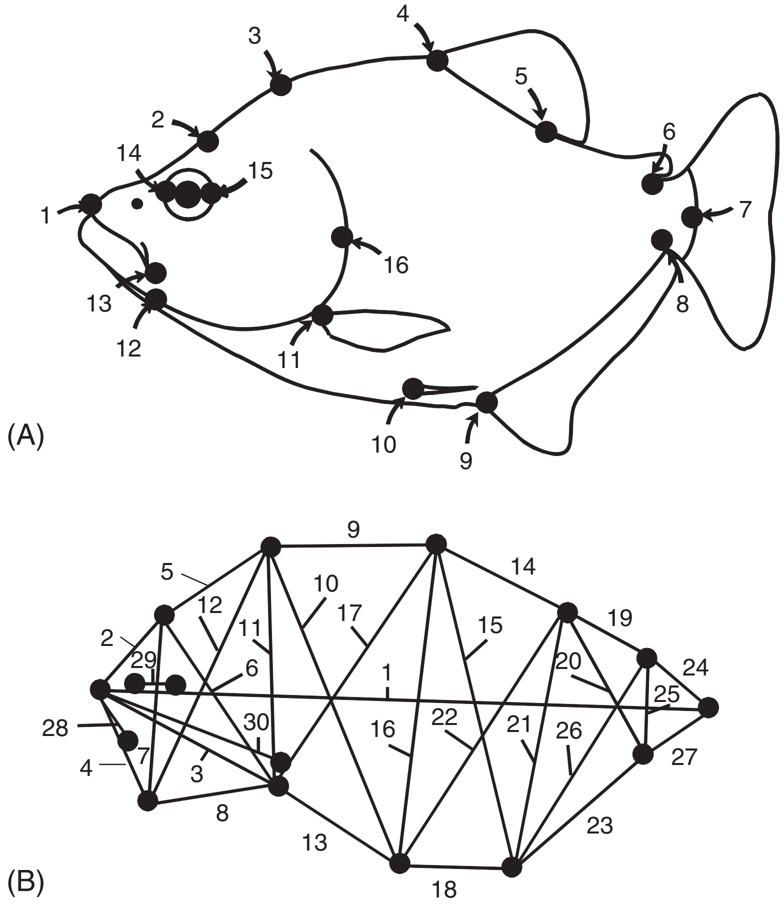


Figure 1.2 Truss measurement scheme of external body form of a teleost: (A) well-defined endpoints of measurements; (B) a selection of 30 lengths, arranged in a truss.

measurements samples more directions of the organism and the measurements are more evenly spaced; the set also contains many short measurements. Additionally, the endpoints of all of the measurements are biologically homologous anatomical loci – landmarks. Although these features make the truss a clear improvement over the classical measurement scheme, this approach still produces a list of numbers (values of segment lengths), with all the attendant problems of visualization and communication.

One problem shared by the two measurement schemes is that neither collects all of the information that could be collected. The truss scheme shown in Figure 1.2 contains 30 measurements, but this is only a fraction of the 120 that could be taken among the same 16 landmarks (Figure 1.3). Of course, many of the 120 are redundant, and several of them span large regions of the organism. We would also need extraordinarily large samples before we could perform the necessarily mathematical manipulations or perform valid tests of hypothesis. In addition, the results would be incredibly difficult to interpret because there would be 120 pieces of information (e.g. regression coefficients, principal component loadings) for each specimen, for each trend or difference. We might be tempted

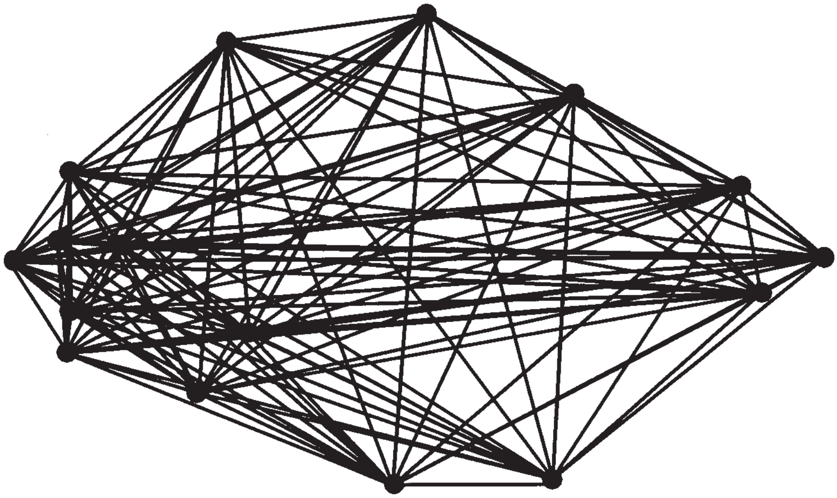


Figure 1.3 All 120 measurements between endpoints defined by the 16 landmarks of Figure 1.2.

to cull the 120 measurements to those that seem most likely to be informative, but until we have done the analysis we cannot know which to cull without altering the results. Clearly, we need another way to get the same shape information as the 120 measurements, but without the excessive redundancy.

Another problem that the truss shares with more traditional schemes is that it measures size rather than shape – each length is the magnitude of a dimension, a measure of size. This does not mean that the data include no information about shape – they do – but that information is contained in the ratios among the lengths, and it can be surprisingly difficult to separate information about shape from size. Some studies have analyzed ratios directly, but ratios pose serious statistical problems (debated by Atchley et al., 1976; Corruccini, 1977; Albrecht, 1978; Atchley and Anderson, 1978; Hills, 1978; Dodson, 1978). The more usual approach is to construct shape variables from linear combinations of length measurements, such as Principal Component (PC) loadings. Here, one component, usually the first (PC1), is interpreted as a measure of size, and all the others are interpreted as measures of shape. However, PC1 includes information about both shape and size, as do all the other PCs. The raw measurements include information about both shape and size, and so do their linear combinations.

Not only are the methods of separating size from shape problematic; the *idea* of size and shape has been one of the most controversial subjects in traditional morphometrics. One reason for this controversy is the multiplicity of definitions of size (and also of shape), several of which are articulated by Bookstein (1989). Virtually any approach to effecting this separation can be disputed on the grounds that the notion of "size" that is separated from "shape" is not really "size." Another reason for the controversy is that some workers argue that no such separation is biologically reasonable (see, for example, the discussion of studies of heterochrony based on growth models in Klingenberg, 1998). However, even if we accept the argument that size and shape are intimately linked by biological processes, we still want to know more about their relationship than the mere fact of its existence.

Extracting the relationship between size and shape from a set of measurements can be especially difficult when the organisms span a broad size range. When some organisms are 20 mm long and others are 250 mm, *all* measurements will differ in length. Even if shape is not much influenced by a ten-fold change in size, all measurements will still be correlated with size; quantifying this fact is merely restating the obvious. In fact, we should expect size to be the dominant explanation for the variance in traditional morphometrics because these measurements *are* measurements of size. Instead, we should be concerned about the possibility that the variance in shape is not fully explained by the variance in size, but is simply overwhelmed by it. For instance, in analyses of ontogenetic series of two species of piranha (one being the running example throughout this chapter), we find that 99.4% of the variance is explained by the PC1 in both species. This suggests that there is nothing else to explain in either species, because it is hard to imagine that the remaining 0.6% is anything but noise. And yet, we do not actually know what proportion of shape variation is explained by size; nor do we know whether different proportions or patterns of shape change are explained by size in these species.

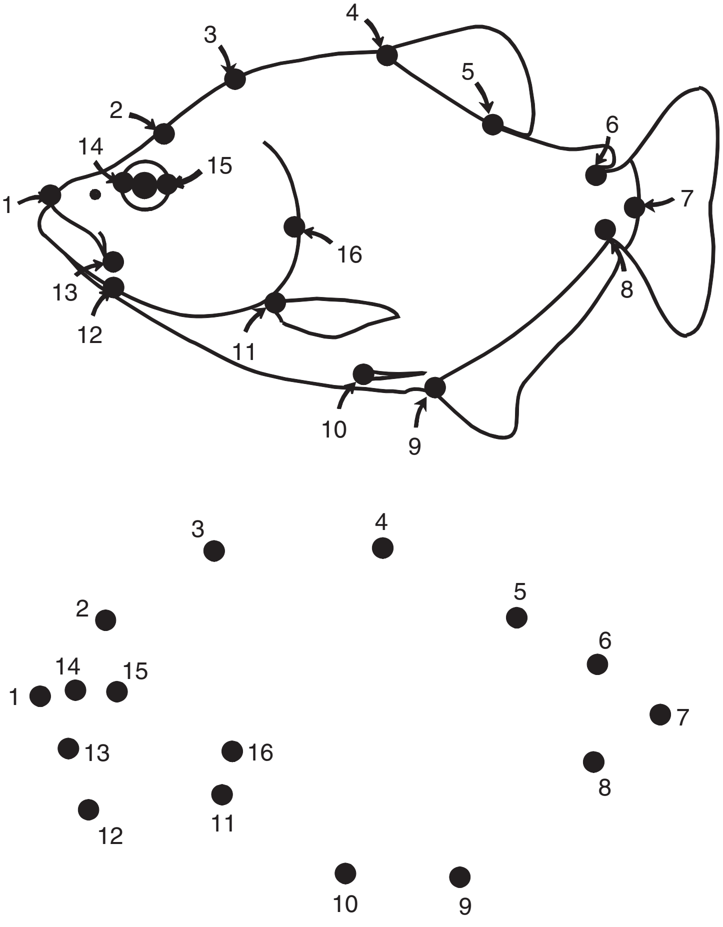


Figure 1.4 The 16 landmarks, stripped of the line segments connecting them.

One other serious limitation of traditional morphometrics is that the measurements convey no information about their geometric structure. If we strip off the line segments connecting the landmarks in Figure 1.3 and just look at the position of the landmarks on the page (Figure 1.4), we can see that some are close to each other (e.g. 12 and 13) and others are far apart (e.g. 1 and 7); some are dorsal (3 and 5), others are more posterior (6–8). That information about relative positions, which is so important to morphologists, is contained in the coordinates of the landmarks but not in the list of distances among them – not even in the comprehensive list of 120 measurements. In fact, the list of 32 coordinates contains all of this positional information in addition to all of the information contained in the 120 distances (the distances can be reconstructed from the coordinates if the units of the coordinate system are known). More importantly, simple algebraic manipulations allow us to partition the information captured by the coordinates into components of size and shape (and to strip off irrelevant information like the position and orientation of the specimen). Afterward, we have slightly fewer than 32 shape variables (because information about size, position and orientation has been separated from information about shape), but we still have the information about the geometric structure of our landmarks that was captured when we digitized the specimens, and we have the information that is present in the full list of 120 measurements without the redundancy. Consequently, we do not need to cull the data in advance of the analysis, and so we do not lose any information we might have had prior to that culling. In addition, partitioning the morphological variation into components of size and shape means that variance in size does not overwhelm variance in shape even when the variance in size is relatively large. In the two species mentioned above (in which PC1 accounts for 99.4% of the variance), size explains 71% of the variance in shape in one species, but only 21.7% in the other.

An important advantage of analyzing landmark coordinates is that it is relatively easy to draw informative pictures to illustrate results. In Figure 1.5, the shape changes that occur during the ontogeny of one species of piranha are shown as vectors of relative landmark displacement and as a deformed grid interpolating among those vectors. In both representations, it is quite clear that the middle of the body becomes relatively deeper while the postanal region becomes relatively short, especially the caudal peduncle (between landmarks 6 and 7). Both pictures also show that the posterodorsal region of the head (above and behind the eye) becomes relatively longer and deeper while other regions of the head become relatively shorter. (We emphasize that these are *relative* changes, because the piranha becomes *absolutely* larger in every dimension and region mentioned.)

It is possible to present traditional morphometric results in graphic form by placing the numbers on the organisms, as in Figure 1.6. This, like Figure 1.5, shows that the middle of the body grows faster and becomes deeper than the rest of the animal. The limitation of this representation (and of the analysis) is exemplified by the difficulty of interpreting the large coefficient (1.23) of the posterior, dorsal head length – it is not clear whether the head is just elongating rapidly, or if it is mainly deepening, or if it is both elongating and deepening. We also cannot tell if the pre- and postorbital head size increases at the same rate, because the measurement scheme does not include distances from the eye to other landmarks. None of these ambiguities arose from the geometric analysis of the landmark coordinates; the figure illustrating that result showed the information needed to understand the ontogenetic changes in these specific regions. This ability to extract and communicate information about the spatial localization of morphological variation

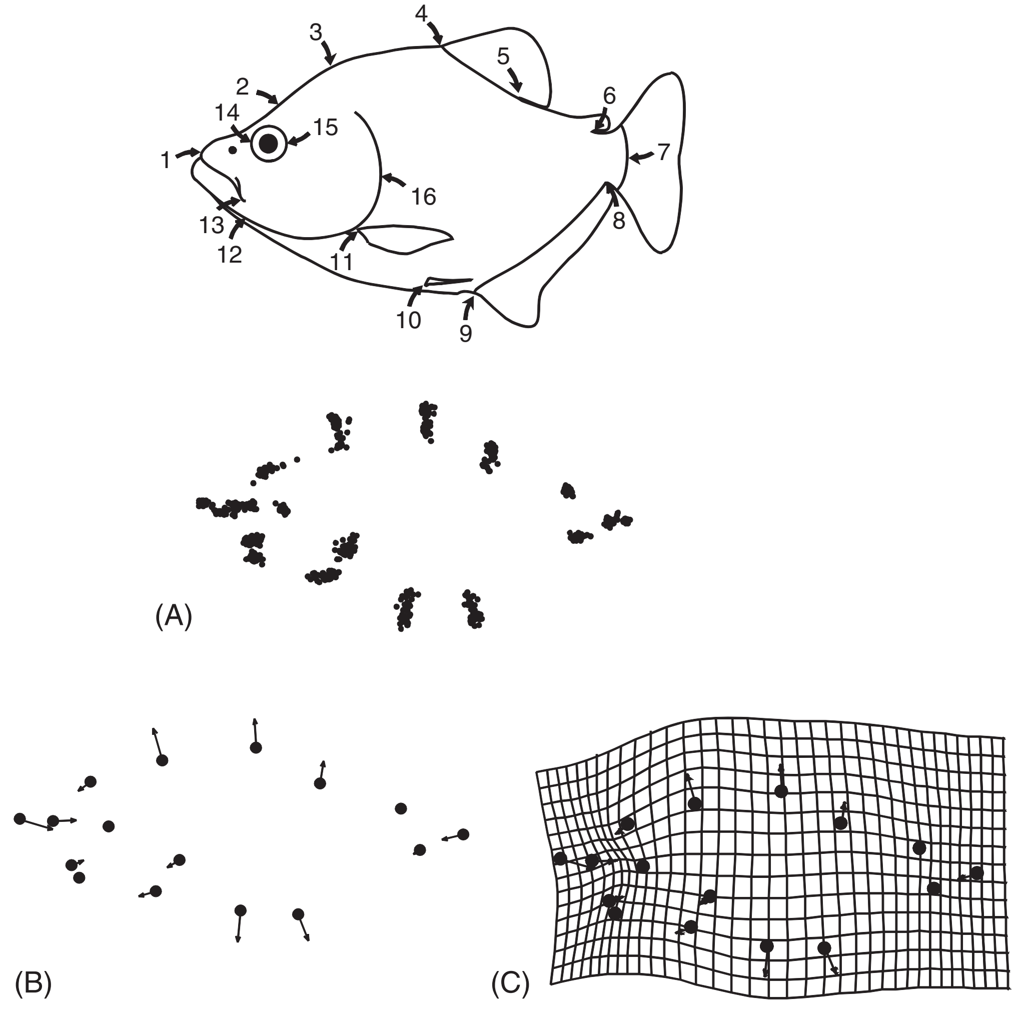


Figure 1.5 Ontogenetic shape change depicted in two visual styles. (A) Landmarks of all specimens; (B) vectors of relative landmark displacement; (C) deformed grid.

(its magnitude, position and spatial extent on the organism) is among the more important benefits of geometric morphometrics.

Geometric morphometrics does not solve all of the problems confronting traditional methods, and one remaining problem becomes evident when we try to examine the changes in head profile over the piranha's ontogeny (Figure 1.7). We can see that the average slope on either side of landmark 2 must get steeper, but we cannot tell whether the profile becomes more S-shaped, C-shaped or any other shape. This uncertainty arises because the three landmarks provide no better a sample of the curve's shape than do the line segments connecting them. Clearly, any solution of this problem will require analysis of points on the curve that are not landmarks (Figure 1.8). Methods for analyzing curves are being developed and used (we discuss them in Chapter 15), so this limitation of geometric morphometrics will likely prove transitory.

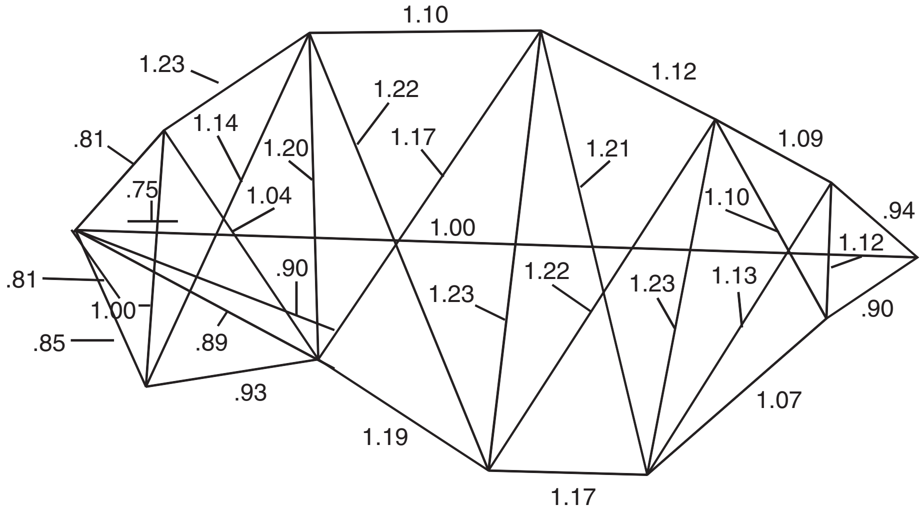


Figure 1.6 Allometric coefficients of traditional morphometric measurements, plotted on the organism.

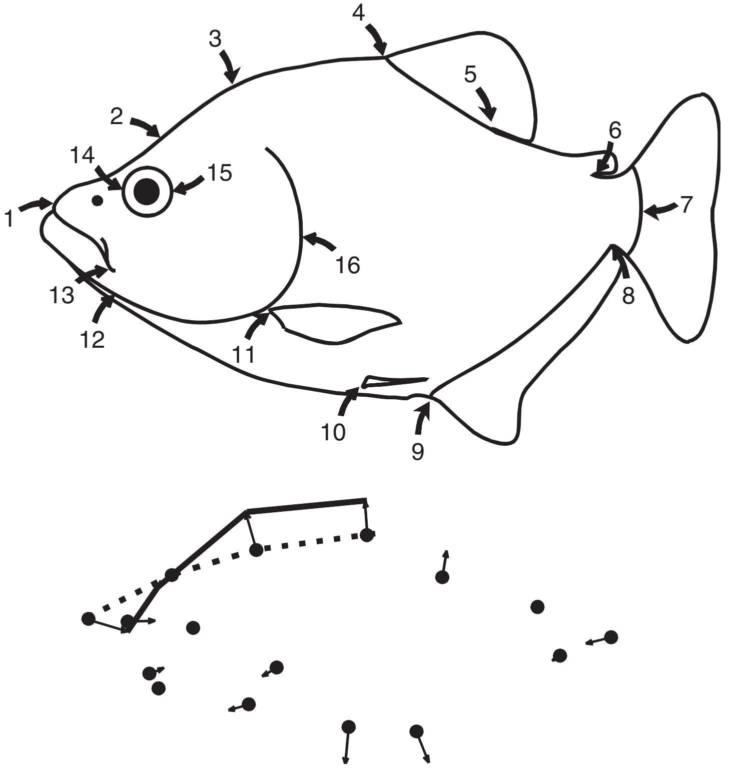


Figure 1.7 Ontogenetic change in head profile as implied by changes in the orientation of straight lines drawn between landmarks of the head.

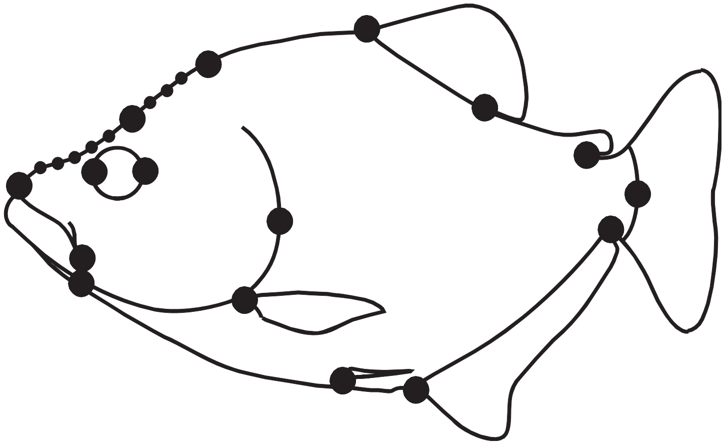


Figure 1.8 Additional points on the head profile, which are not landmarks.

Geometric morphometrics may also appear to have a limitation that does not confront traditional methods: the restriction to two-dimensional data. The reality is that mathematical theory poses no obstacles to analysis of three-dimensional shapes. Instead, the obstacles lie in other constraints restricting biologists to two-dimensional data, notably (1) the cost of the equipment for obtaining three-dimensional coordinates (which is also time-consuming to use) and (2) the difficulty of depicting the results on static, two-dimensional media like the pages of a journal. Traditional morphometric studies need not face these obstacles because, if the equipment required for three-dimensional digitizing is exorbitant (in time or money), specimens can always be measured with calipers. However, in using calipers we do not collect three-dimensional coordinates, so this approach sidesteps rather than solves the problem. The difficulty of depicting results on a two-dimensional page does not arise when results are tables of numbers, which is another case of sidestepping rather than solving the problem.

Geometric analyses of landmark coordinates do solve many of the problems confronting traditional methods of measurement. Those that remain involve analyses of curves with few or no landmarks, and the illustration of three-dimensional results. Without denying that these are real issues, we can still obtain a great deal of information about shape and size from geometric studies.

Shape and size

The rapid progress in geometric morphometrics has resulted largely from having a coherent mathematical theory of shape, which requires articulating a precise definition of the concept. Like the definition of any word, that of "shape" is entirely a matter of semantics. However, semantics is not trivial. We cannot have a coherent mathematical theory of an undefined concept; the definition of shape is the foundation for a mathematical theory of shape. Whether that theory applies to our biological questions depends on whether it captures what we mean by shape. Thus it is important to understand the concept of shape underlying geometric morphometrics, and also, because the concept of size is so closely related to that of shape, we cannot fully understand one without understanding the other and also how they are related to each other.

Shape

In geometric morphometrics, shape is defined as "all the geometric information that remains when location, scale and rotational effects are filtered out from an object" (Kendall, 1977). The earliest work that depends on this definition of shape began the analysis with the coordinates of points; consequently, the "objects" are sets of those coordinates – i.e. configurations of landmarks, such as that shown in Figure 1.4. An important implication of Kendall's definition is that removing the differences between configurations that are attributable to differences in location, scale and orientation leaves only differences in shape. These operations and their consequences are illustrated in Figure 1.9. In Figure 1.9A there are two configurations, side by side. This difference in location has no bearing on their shape difference, so in Figure 1.9B both have been translated to the same location. The two configurations still differ in scale, which also has no bearing on their shape difference, so in Figure 1.9C they are converted to the same scale. The two configurations still differ in orientation (their long axes are about 45° apart), which also has no bearing on their shape differences, so in Figure 1.9D they are rotated to an alignment that leaves only the shape differences. After removing all the differences that are not shape differences, and provided that this is done in a way that does not alter shape, we are left with only the shape differences. We can now use the coordinates of the final configurations (Figure 1.9D) to analyze these shape differences.

Representing an organism solely by a configuration of landmarks leaves out some aspects of what we might normally mean by shape, such as curvature. Curvature *is* a feature of an object that remains after filtering out location, scale and rotational effects, but it is not necessarily captured effectively by the coordinates of a set of landmarks. Because curvature fits the broad definition of shape, we can anticipate eventually having a theory of shape analysis that applies to the shapes of curves and is consistent with the theory that applies to configurations of landmarks.

Size

Kendall's definition of shape mentions scale as one of the effects to be removed to extract differences in shape between two configurations. The implication of this statement is that scale provides a definition of size that is independent of the definition of shape. The concept underlying geometric scale is quite simple, and may be intuitively obvious by visual inspection – in Figure 1.9A the landmarks are generally further apart in one configuration than in the other, which is what we would expect when a configuration is larger. Before computing geometric scale, we need to determine the location of the center of the form (its "centroid") and calculate the distance between each landmark and the centroid. Figure 1.10 shows the location of the centroid and the segments connecting the landmarks to the centroid for one of the piranhas we have been discussing. Now we compute geometric scale by calculating the square of each of those distances, summing all the squared distances, and then taking the square root of that sum. This quantity is called "centroid size."

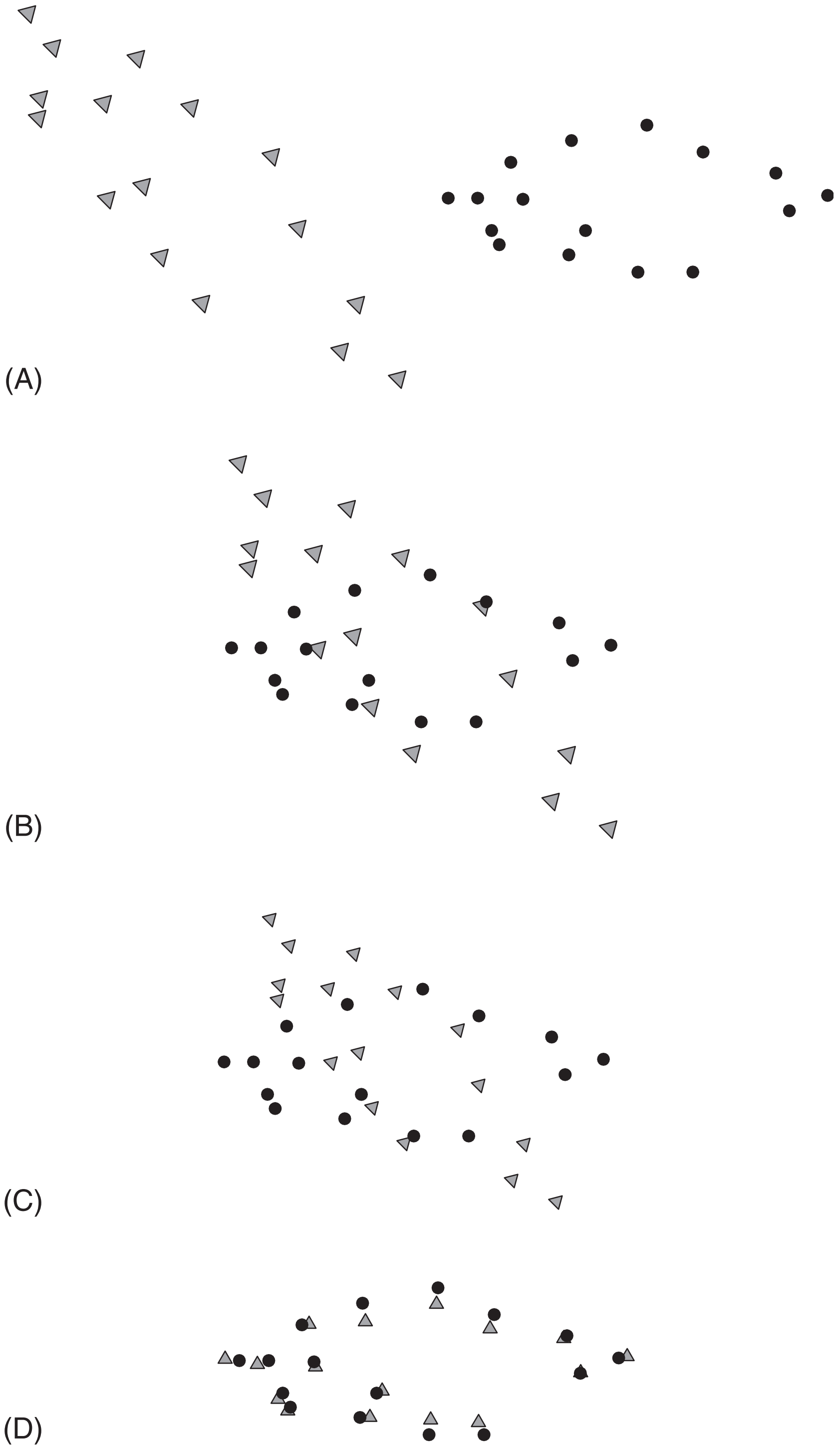


Figure 1.9 Removing variation due to differences in position, scale and orientation. (A) Two original configurations; (B) after removing differences in location; (C) after removing differences in scale; (D) after removing differences in orientation, leaving only differences in shape.

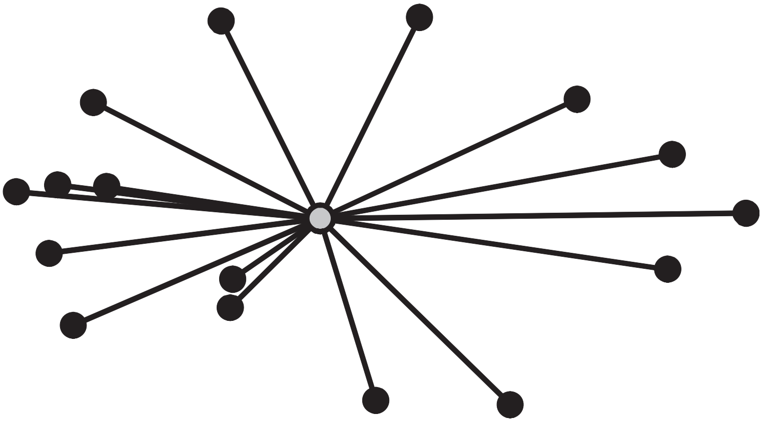


Figure 1.10 A visual representation of centroid size as computed for 16 landmarks on a piranha. The open circle is the centroid; the segments connecting the centroid to the landmarks represent the distances used to compute centroid size.

Centroid size is the one measure of size that is *mathematically* independent of shape. Empirically, centroid size may often be correlated with shape because larger organisms are usually shaped differently than smaller ones. The fact that we have measured shape and size separately does not mean that we lose any information about the relationship between them, any more than measuring shape and age separately bars us from analyzing their relationship. We can easily evaluate the empirical relationship between shape and size using those conventional statistical methods that can be applied to both size and shape data.

Methods of data analysis

Replacing the distances of traditional morphometrics with landmark coordinates does not force us to sacrifice conventional statistical analyses of shape. We can ask all the questions we have ever asked. Such questions often comprise two parts, the first of which Bookstein (1991) termed the "existential question": *is* there an effect on shape? We answer that by determining the probability that the association between variables is no greater than could have arisen by chance. The second question, "*what* is the effect?", calls for a description. In the ontogenetic series of piranhas discussed earlier, we can analyze the relationship between shape and size by computing the centroid size of each configuration of landmarks, and then computing the configurations of landmarks from which differences in position, scale and rotational effects have all been removed. These new configurations, shown in Figure 1.11A, represent the shapes of all the specimens. To answer the first question about the existence of an effect, we regress shape on centroid size using multivariate regression in which "shape" is the dependent variable and "centroid size" (or its logarithm) is the independent variable. For this example, we can conclusively reject the null hypothesis of no effect at $p < 1 \times 10^{-5}$ (we can also determine that 71% of the shape variation is explained by size). To answer the second question about the description of the effect, we present the pictures showing relative landmark displacement (Figure 1.11B) or the deformed grid computed by interpolation (Figure 1.11C).

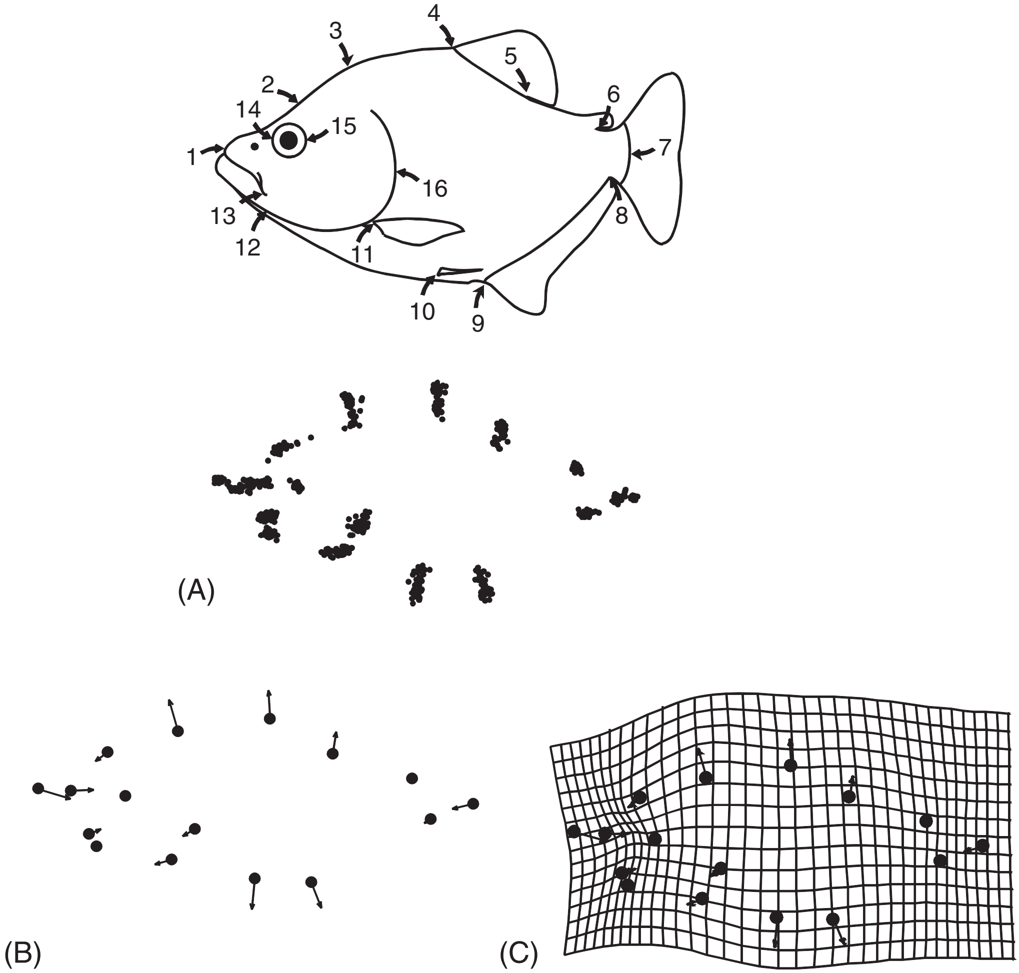


Figure 1.11 Analyzing the impact of size on shape by multivariate regression. (A) Configurations of landmarks from which differences in position, scale and orientation have been removed; (B) the covariance between size and shape depicted by vectors of relative landmark displacements; (C) the covariance between size and shape depicted by a deformed grid.

Replacing distances with coordinates also does not require us to abandon familiar ordination methods, such as principal components analysis and canonical variates analysis. These are methods that are used frequently to explore patterns in the data; their results include scatter plots of specimens that describe patterns of variation among individuals or differentiation among groups. These patterns often provide hints about the causes of variation or differentiation; hints that are reinforced by the accompanying graphics of the dimensions along which specimens most vary (Figure 1.12) or groups most differ (Figure 1.13).

The one important distinction between analyses of geometric shape data and those of conventional morphometric data is that all analyses of landmark configurations are necessarily multivariate. By definition, shape is a feature of the whole configuration of landmarks. Even the simplest shape, a triangle, cannot be analyzed univariately; more

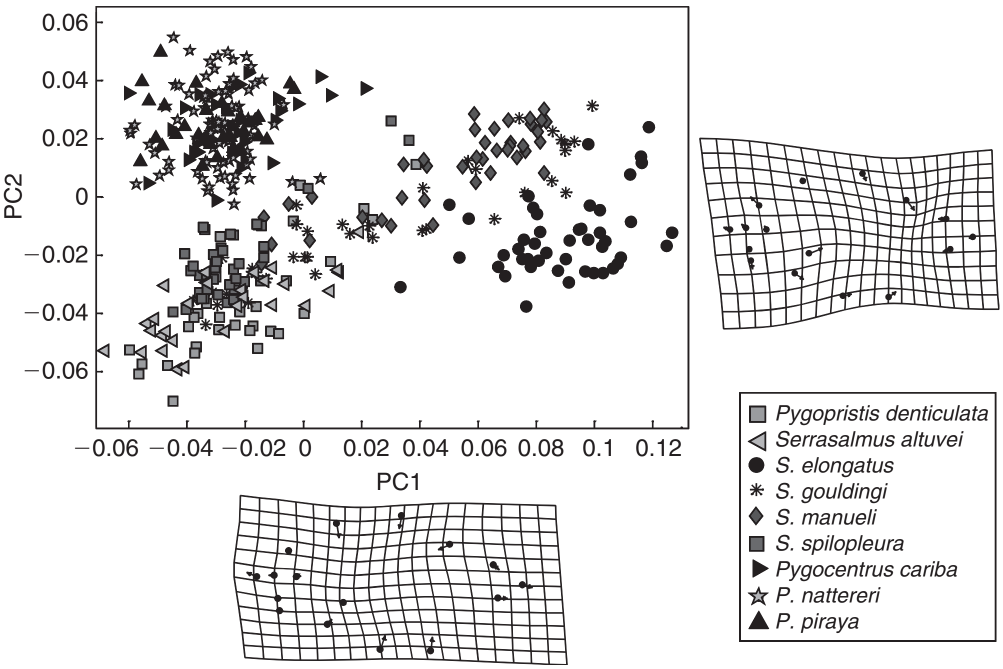


Figure 1.12 Principal components analysis of piranha body shape.

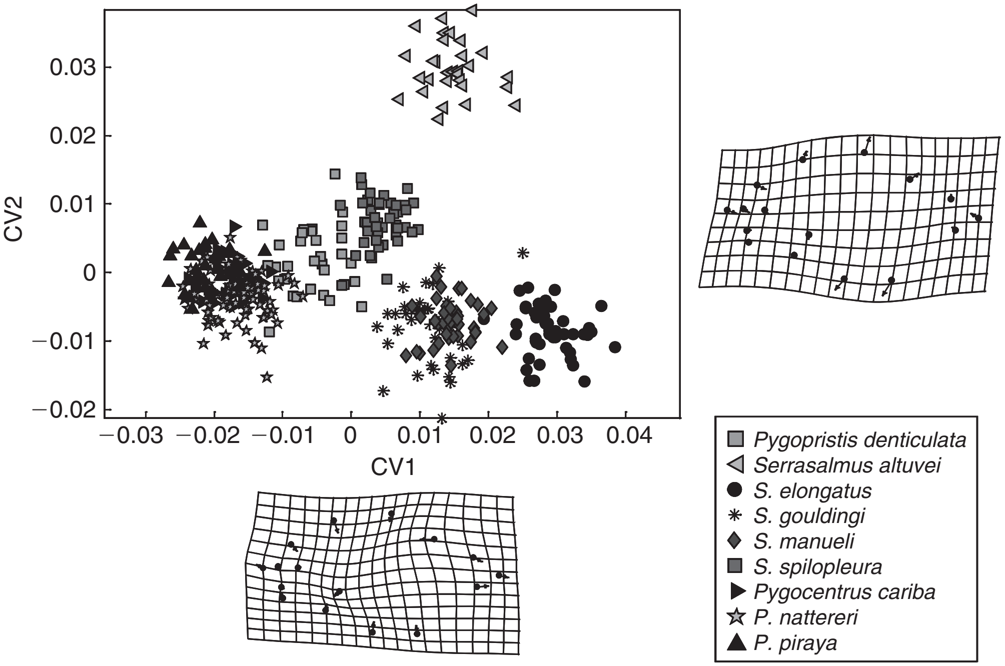


Figure 1.13 Canonical variates analysis of piranha body shape.

than one variable is needed to describe differences among triangles completely. We cannot simplify analyses (or interpretations) by partitioning the configurations of landmarks into subsets; subsets of landmarks are different shapes, not traits dissected from the whole. We cannot regress the coordinates of only one of the 16 piranha landmarks on size and consider the resulting coefficients to be a valid result about a part of the configuration of 16 landmarks. We cannot even regress the coordinates of 12 of the 16 landmarks on size, and consider the resulting 24 coefficients, taken together, to be a valid result about a part of the configuration. Because we have defined shape in terms of the whole configuration of landmarks, our analyses must be of that whole. However, this does not prevent us from subdividing an *organism* to analyze relationships between parts. For example, we could divide the piranha into the cranial and postcranial regions, and analyze the landmarks from each region as a separate configuration; we could then ask how the shapes of these two regions covary by analyzing the relationship between configurations. The requirement that *configurations* be analyzed multivariately and therefore as wholes does not force us to treat *organisms* as unitary wholes (although we may find out that they are).

Biological and statistical hypotheses

Few of the hypotheses of interest to biologists are as simple as the allometric hypothesis examined earlier. Only rarely can the more complex hypotheses be wrestled into the form of a statistical null hypothesis and its alternatives. The first difficulty is that the statistical null merely states that the factor of interest has no effect; this is the hypothesis we hope to reject in favor of the alternative hypothesis that the factor does have an effect. In this situation we have two hypotheses that are diametrically opposed, mutually exclusive. In contrast, many biological hypotheses are more complex, stating multiple alternative theories of causation, and these alternatives may not be mutually exclusive. Thus the real goal of many studies is to discriminate between expected effects, not to reject a hypothesis of no effect. Perhaps we are interested in the evolution of claw shape in crabs. We probably already know that claw shape has evolved; the more interesting (and difficult) question is whether the derived claw shape arose to enhance the ability to burrow into a muddy substrate or was intrinsically constrained by development (or both).

Another difficulty posed by realistic biological studies is that there may be other alternative hypotheses beyond the few we have chosen to test. For example, other explanations for the derived claw shape of the crabs might be an enhanced ability to block a burrow entrance or even to attract mates. We also might have several alternative theories about how development could constrain the evolution of claw shape.

Yet another obstacle to translating a biological hypothesis into a statistical one is that the complexity of the biological hypotheses rarely allows for adequate testing by any single method. To test whether the evolution of crab claw shape was intrinsically constrained by development, we must first determine whether development demonstrates any signs of constraint and then show that constraint could explain the evolution of claw shape. We should also show that the various adaptive hypotheses predict different evolutionary transformations than those specified by the developmental constraint hypothesis, so that we can rule out these biological alternative hypotheses.

In emphasizing the complexity of biological hypotheses we do not mean to say that they cannot be tested rigorously – they can be. However, doing so requires far more effort and creativity than testing the simple hypothesis that size affects shape. It also requires understanding what various analytic methods do, what their limits are, and how they are mathematically related. Far too often biologists use a limited array of techniques to analyze multivariate data, regardless of their questions. Throughout this book we emphasize the biological questions prompting a morphometric analysis, and underscore the applications of each method as we discuss them in turn. However, only after a variety of methods have been introduced (and mastered) can we begin to address questions of realistic biological complexity.

Organization of the book

We begin this book with a series of chapters covering the basics of shape data – what landmarks are and how to select them (Chapter 2), and how their coordinates are transformed into the shape variables that will be used in subsequent analyses (Chapters 3–6). The next section covers analytic methods: exploratory tools (Chapters 7 and 8) and more formal methods of hypothesis testing (Chapters 9–11). We then demonstrate the application of these methods to complex biological questions, which may require using multiple methods, both exploratory and hypothesis-testing (Chapters 12–13). The final two chapters cover issues that require continued development: Chapter 14 discusses the use of morphometric analysis in phylogenetic studies, and Chapter 15 covers some methodological topics on which there is still not complete consensus regarding either technical or graphical issues, but which are likely to yield promising new methods in the near future.

In presenting the basics of shape data, we follow the discussion of landmarks (Chapter 2) with a simple method of producing shape variables (Chapter 3) – namely the two-point registration that yields Bookstein's shape coordinates (Bookstein, 1986, 1991). These variables are easily understood, easily calculated by hand, and do not require an understanding of the general theory of shape. Presenting them first allows us to discuss a number of general issues (including the interpretation of results) before presenting the more abstract theory of shape analysis in Chapter 4. That theory provides the framework for generating (as well as analyzing) shape variables. After reviewing the basic theory, we return to the subject of shape variables in Chapter 5. Chapter 6 discusses the thin-plate spline, an interpolation function useful for depicting results by means of a deformed grid (as in Figures 1.11–1.13), and also for obtaining a set of shape variables that can be used in conventional multivariate analyses.

The second section of the book concerns methods for analyzing shape variables. In a sense, all these methods are used to produce the biologically interesting variables – the ones that covary with the biological factors of interest. Unlike the variables produced by the methods of the previous section, the variables produced by these analytic methods have a biological meaning. They answer such fundamental questions as "What impact does size have on shape?", or "By how much, and in what way, do these species differ in their ontogenies?", or "Do these populations vary along a single latitudinal gradient?", or even "What shape has the highest fitness in this population?" Each of these questions is answered in terms of a shape variable – the shape covariates of size or age, of latitude or fitness, or of any other factor of interest. When we do not have any such factors in mind in advance of a study, we can explore the data algebraically, using the methods of matrix algebra to determine if any interesting patterns emerge (principal components analysis, PCA, is an example of this kind of algebraic exploration).

Because many biologists begin a study by exploring patterns in the data, the section on analytic methods begins with an overview of ordination methods (Chapter 7). These are useful for extracting simple patterns from complex multidimensional data because they provide a space of relatively low dimensionality, capturing most of the variation among specimens (PCA), or most of the differences among groups (canonical variates analysis, CVA). We explain the algebra underlying these methods, compare them, and discuss when each is appropriate in light of particular biological questions.

The next three chapters cover methods of statistical analysis. We begin with an overview of computer-based statistical methods, i.e. computer-intensive methods for constructing confidence intervals and/or hypothesis testing, such as bootstrapping and Monte Carlo simulations (Chapter 8). The next two chapters discuss the two broad classes of hypotheses that are conventionally tested statistically. Chapter 9 addresses hypotheses about the effects of an independent categorical variable – Hotelling's T^2-test, analysis of variance (ANOVA), and multivariate analysis of variance (MANOVA); Chapter 10 addresses hypotheses about the effect of a continuous variable on shape (regression). The final chapter in this section, Chapter 11, covers a method new to morphometric studies, one that analyzes the covariance between two blocks of variables, partial least squares analysis.

The third section covers applications of morphometric methods to realistically complex biological hypotheses, addressing more than just existential questions and requiring more of the answers than just descriptions. We begin with hypotheses that are often stated only in words, discuss framing them in the terms of more precise formal models, and then reframe these models into terms suited to statistical analysis. Once a hypothesis has been framed in the last set of terms, data analysis can proceed in a quite straightforward fashion, combining an array of techniques. As examples of complex biological questions we include those posed by studies of disparity and variance (Chapter 12), the analysis of relationships between ontogeny and phylogeny (Chapter 13), and also systematics (Chapter 14). The latter chapter represents a bridge between complex but tractable questions and subjects in need of additional tools.

The final chapter of this book (Chapter 15) briefly discusses two important areas in which a full set of tools have not been developed yet: (1) methods for analyzing three-dimensional coordinate data, and (2) methods for analyzing shapes of curves where no discrete anatomical loci can be found (by locating and analyzing points called "semi-landmarks"). Neither of these subjects is properly part of a primer that focuses on well-developed, uncontroversial methods, but both are important for biologists, and both are subjects of intensive ongoing work. In presenting these subjects we concentrate on the major points of departure (both conceptual and practical) from the primary subject of this book, the analysis of two-dimensional configurations of landmarks.

The terminology of statistical shape analysis can be daunting – there are many unfamiliar words and many terms differ by only a single letter or subscript. Thus we conclude this book with a glossary of terms, including general statistical terms (e.g. population, sample) and more specialized terms of shape analysis (e.g. Procrustes distance, partial warps).

Software and other resources

Geometric morphometrics studies require fairly specialized software, not so much to analyze the data as to depict the results graphically. Fortunately, the necessary software is readily and freely available. As Mac users will soon realize, virtually all the compiled software runs under Microsoft Windows.

At present, one major source of software is located at the SUNY Stony Brook website: http://life.bio.sunysb.edu/morph. Follow the link to **Software** (the rest of the links go to other valuable resources, including information about meetings, courses, and a directory of many people interested in morphometrics, with links to their webpages). We recommend that anyone planning a morphometric study downloads the videodigitizing program, **TPSDig**. Not only is this a well-designed and extremely useful program, but also many writers of morphometric software assume that is the one used for data collection, so the format in which it outputs the data (TPS format) has become the standard input format for several programs. There are other useful programs in the TPS series, but we generally do not provide detailed instructions for using them because we can neither anticipate nor control any changes in them.

Another major source of morphometric software is located at the website: http://www.canisius.edu/~sheets/morphsoft.html. This software, called the Integrated Morphometrics Programs (IMP), is written by one of us (HDS) and every method of analysis discussed in this book can be implemented by software in this series. There are three categories of software: (1) General Release; (2) Undocumented Software (which lacks manuals but the programs run and have been extensively used in research), and (3) Beta-Software (which has not been used in any serious research project, so may need considerable reworking before it is fully useful). There are some additional programs available that have been used in published research and so are made available; these can be found at the end of the “Update Information.” At the end of most chapters of this book, we provide instructions for using the relevant software. These instructions are based on versions of the programs that have been frozen, so that you can run all the programs using these instructions. We do, however, anticipate upgrading the software; these upgrades will be available on the website and will (eventually) be documented. Major changes will be detailed in the “Update Information” on the bottom of the morphsoft webpage.

Running the IMP programs, which are written in Matlab (Mathworks, 2000) and compiled to run under Microsoft Windows, requires first installing a large package of software, **mglinstaller** (detailed instructions for installing it, and for installing other programs in the IMP series, are given below). Different versions of Matlab are often incompatible with each other (both upwardly and downwardly), so programs written in the future, using a newer version of Matlab, will require installation of a new version of **mglinstaller** (in a different directory).

Another important resource is the listserver Morphmet. It is useful to subscribe to this list, if only to be informed of new software and notified of any mathematical mistakes or bugs in the programs. Additionally the list is sometimes quite active, discussing topics of general interest, including conceptual issues like the meanings of size and shape, and practical issues like dealing with preservational artifacts. Some recent posts have also provided extensive bibliographies of morphometric studies of mollusks and fishes.

To subscribe to this list, send an email to majordomo@wfubmc.edu and include the following single line in the body of the message: subscribe morphmet.

Downloading and installing mglinstaller

Before you use any program in the IMP series, you need to download and run the self-expanding **mglinstaller** (megalo-installer). This will create the directories (folders) where the other IMP programs must be installed. To download **mglinstaller**, go to the IMP website (http://www2.canisius.edu/∼sheets/morphsoft.html or www.biocollections.org), find **mglinstaller** and click on it. This is a very large file, so it may take a while to download. After the download is complete, you need to create the directory (folder) where you want **mglinstaller** to be expanded. We recommend you call this folder **Matlab6** so you can keep track of the version of Matlab used to write the software. Now expand **mglinstaller** in that directory. It will create a folder in **Matlab6** called **bin**, and a folder in **bin** called **win32**; it will also unpack a series of files needed to run the other IMP programs. The other programs are also packaged as self-expanding files. After you download them, they *must* be expanded into the folder **win32**. If they are not installed in **win32**, they will not run.

References

Albrecht, G. (1978). Some comments on the use of ratios. *Systematic Zoology*, **27**, 67–71.

Atchley, W. R. and Anderson, D. (1978). Ratios and the statistical analysis of biological data. *Systematic Zoology*, **27**, 71–78.

Atchley, W. R., Gaskins, C. T. and Anderson, D. (1976). Statistical properties of ratios. I. Empirical results. *Systematic Zoology*, **25**, 137–148.

Bookstein, F. L. (1986). Size and shape spaces for landmark data in two dimensions. *Statistical Science*, **1**, 181–242.

Bookstein, F. L. (1989). “Size and shape”: A comment on semantics. *Systematic Zoology*, **38**, 173–190.

Bookstein, F. L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.

Bookstein, F. L., Chernoff, B., Elder, R. et al. (1985). *Morphometrics in Evolutionary Biology*. The Academy of Natural Sciences of Philadelphia.

Corruccini, R. S. (1977). Correlation properties of morphometric ratios. *Systematic Zoology*, **26**, 211–214.

Dodson, P. (1978). On the use of ratios in growth studies. *Systematic Zoology*, **27**, 62–67.

Hills, M. (1978). On ratios – a response to Atchley, Gaskins and Anderson. *Systematic Zoology*, **27**, 61–62.

Kendall, D. (1977). The diffusion of shape. *Advances in Applied Probability*, **9**, 428–430.

Kendall, D. G. and Kendall, W. S. (1980). Alignments in two-dimensional random sets of points. *Advances in Applied Probability*, **12**, 380–424.

Klingenberg, C. P. (1998). Heterochrony and allometry: the analysis of evolutionary change in ontogeny. *Biological Reviews*, **73**, 79–123.

Lagler, K. F., Bardach, J. E. and Miller, R. R. (1962). *Ichthyology*. John Wiley & Sons.

Strauss, R. E. and Bookstein, F. L. (1982). The truss – body form reconstructions in morphometrics. *Systematic Zoology*, **31**, 113–135.

PART I

Basics of Shape Data

2

Landmarks

Landmarks are discrete anatomical loci that can be recognized as the same loci in all specimens in the study. Because landmarks play a fundamental role in geometric morphometrics, it is important to understand their function in a shape analysis. It is equally important to understand which functions they do *not* serve, as that understanding also influences the selection of landmarks. Criteria for selecting landmarks differ from those applied to choosing traditional morphometric variables, so some rethinking may be required. This chapter begins with a summary of some of the basic differences between conventional and landmark-based studies that bear on landmark selection and on how we may need to change the way we think about selecting variables. Next is a review of the criteria for choosing landmarks in light of both biological and mathematical considerations, focusing on general criteria and principles. This is followed by three concrete examples, each explaining why particular landmarks were chosen. The chapter concludes with a practical guide to collecting landmark data.

Changing the way we think about selecting variables

One major difference between conventional and landmark-based techniques is most important for thinking about the selection of variables: conventional morphometric variables are selected *a priori*, meaning that we choose variables before we conduct the analysis; in landmark-based studies, that is not the case. In studies of traditional measurements only the variables chosen in advance of the analysis are available for analysis (unless we go back and remeasure the specimens), and for that reason much emphasis has been placed on choosing "meaningful" variables. If we do not select the meaningful ones we will not have them to aid our interpretations, and if we include many that have no particular biological significance, they could complicate our interpretations of those that do. There are many considerations that might enter into the decision regarding which variables are meaningful, including relevance to understanding (1) biomechanics, (2) developmental processes, (3) systematics, and (4) evolutionary processes. For example, in a study designed to analyze the biomechanics of chewing we would conscientiously select variables for their relevance to chewing, which might not be the same variables as those that capture the information

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

relevant for understanding jaw development or relationships among taxa or evolutionary processes. Consequently we might need as many as four different measurement schemes, each one designed in light of the substantive biological questions addressed by the study.

In landmark-based studies, variables are not selected *a priori* (although landmarks are); the meaningful variables are discovered *by* the analysis. All the variables that we could have measured between any pair of landmarks are included in the analysis, so we do not need to decide among them before we begin. As mentioned in Chapter 1, given 16 landmarks we would have to measure 120 variables (i.e. all the lengths, depths and widths that could be measured by distances between pairs of landmarks) to capture as much information as we have in the coordinates of the 16 landmarks. Having all that information available to us does not complicate the interpretation of the results, because the variables do not enter into the interpretation unless they are *found* to be relevant. For example, if we are interested in feeding performance and measure 16 landmarks on jaw bones and teeth, we will discover which are relevant to feeding performance by analyzing the covariance between the landmarks and measures of performance. We do not need to know which variables covary with performance when we begin the analysis – the objective is to discover that at the end.

Landmarks should provide a sufficiently comprehensive sampling of morphology that the features of biological significance can be discovered. This emphasis on discovery does not mean that you should avoid thinking about variables that might be important for your biological questions. The landmarks you select do determine what you may discover. If you are interested in the biomechanics of lever arms, you should locate landmarks on those lever arms or else you will not have the data required to analyze them. However, you will not lose or dilute information about biomechanics of lever arms by including other landmarks of unknown relevance – if they are not relevant, they will not covary with measures of performance. If your *only* question is "What is the mechanical advantage of this jaw compared to that one?", then there is no reason to do a shape analysis – the question you are asking is about mechanical advantage, not shape. As Bookstein (1996) pointed out, geometric methods might be "overkill" in such purely biomechanical studies. However, if you want to place those lever arms in a broader morphological context, geometric morphometrics helps to provide one.

As a general rule, landmarks should be chosen so you can quantify any differences that you can see. A quantitative study should capture at least as much information as does an informal, qualitative inspection of specimens. However, the morphologies should also be sampled more broadly so that you can discover more than is evident by visual inspection, and, if you want to pin down where the changes occur, you will need even and fairly dense coverage of the form. Below we discuss the general criteria for selecting landmarks in more depth, and then we provide examples of three data sets, explaining what the landmarks are and why they were chosen.

Criteria for choosing landmarks

Ideally, landmarks are (1) homologous anatomical loci that (2) do not alter their topological positions relative to other landmarks, (3) provide adequate coverage of the morphology, (4) can be found repeatedly and reliably, and (5) lie within the same plane.

Homology

The concept of homology plays a crucial role in landmark-based morphometrics. Although many traditional morphometric studies have been concerned with homology, homology was not a fundamental concern when selecting measurements. If it had been, some standard variables (such as "greatest skull breadth" or "least interorbital width") would not have been measured. Such variables are not necessarily homologous because they may be measured at very different points on the skull in different organisms, depending on where the skull is widest or the interorbital region is narrowest. Consequently, we cannot say how broadening of the skull or narrowing of interorbital width has been affected by alterations in skull shape. We cannot trace the changes in skull breadth to the changes in shape of the skull because we are not measuring the same things on all skulls. In contrast, homology has been stressed above all criteria for selecting landmarks in geometric morphometrics, and it is undeniably the most important one. For both mathematical and biological reasons, homology is the paramount consideration when it comes to the selection of landmarks.

Understanding the role that homology plays both mathematically and biologically requires an intuitive feel for the mathematics as well as for the biological issues. Sometimes there are reasons for including landmarks in a study, even though their homology is somewhat dubious (in the last chapter of this book we discuss "semi-landmarks," points that aid in studies of regions that lack homologous landmarks). However, it is important to recognize the compromises resulting from using landmarks of doubtful homology. Most systematists will presume that homology is a central concern regardless of any mathematical arguments, but functional morphologists and developmental biologists may be less convinced that homology is actually necessary. However, there are mathematical arguments that reinforce the biological ones, as discussed in depth in Chapter 4. Because you will select your landmarks before you read that chapter, and you need an intuitive feel for the primary mathematical issue before choosing them, the basic mathematical ideas bear mentioning here. The primary mathematical issue is the interpretation of biological change as a deformation: a (smooth) mapping of one set of points to *corresponding* points in another form. The mapping only makes sense if the points are truly "corresponding," and that correspondence requires more than that landmarks have the same name. It requires a careful consideration of what "correspondence" means.

Correspondence need not imply biological homology – we might think of correspondence in functional (or developmental) terms. For example, we might view points as corresponding to each other because they are located at the end of an input lever arm in two different organisms, even if those lever arms are in different locations. In a purely functional sense those points might indeed correspond, but if our aim is to describe the transformation from one form to another, and we are using a mathematical model of a deformation, we need a more restrictive view of "correspondence." The landmark is not just serving a corresponding function; it must also be *the same anatomical locus*.

The importance of biological homology to morphometric analysis has been obscured somewhat by the definition of homology that sometimes appears in the morphometric literature. The semantic discrepancy between the notion of homology used by some morphometricians and the one favored by biologists is partly historical (workers in the two fields have traditionally used the term differently), but there is an important conceptual distinction as well because of the conceptual gap between the subject matter of homology assessments in biological and mathematical contexts.

Biologists usually think about homology in terms of organismal parts or characters, whereas mathematicians think about homology in terms of the individual loci (i.e. points) on those parts. As biologists, our objective in choosing landmarks is to permit making inferences about the regions between them – we are not interested in the landmarks *per se*, but in the shapes of the morphological structures on which those landmarks lie. The role of the landmarks is to pin down those structures at discrete points that we can recognize as the same on all organisms. However, this means that our data *are* the landmarks, the individual loci, and so we also need to think about the homology of those points. Fortunately, this is not a wild conceptual leap. We recognize structures are homologous as structures because they are discrete (distinct from other structures) and recognizable in all specimens. We can apply the same criteria to intersections of structures (as at sutures), or to their centers, or to their tips (ends). If discrete and recognizable structures are homologous as structures, then discrete and recognizable locations on them are arguably homologous as points.

The mathematical framework for thinking about homology is the idea of a deformation, which extends the correspondence of sampled points to unsampled points lying between. Using a model of a deformation, such as the thin-plate spline (Chapter 6), we can draw a picture of a change in shape that extends that change over the whole form, even though we only sampled it at selected points. In that sense, the deformation *imputes* homology to intervening points. For that reason, the mathematical models for deformations have sometimes been termed "homology functions" (see, for example, Bookstein et al., 1985). To understand this idea more fully, consider a sample of landmarks on a skull (Figure 2.1); when looking at the results, we can see changes in the relative positions of landmarks that imply changes in the proportions of structures sampled by them. We can visualize the impact of those changes for the shape of the skull using the deformed grid that stretches where regions are relatively enlarged and contracts where regions are relatively reduced. A highly literal interpretation of that picture could make us uncomfortable because we do not know where every single point on one skull is located on the other – we cannot read the intersections of the grid, for example, as if they are at homologous anatomical loci. However, we are not trying to impute homology to all those points; rather, we are inferring changes in shape that are implied by the homologous landmarks. If we are willing to consider that the structures, such as premaxilla and maxilla (and the sutures between them), are homologous, and that the presphenoid, sphenoid and basisphenoid (and the sutures between them) are also homologous, and that foramina are also homologous, we are specifying correspondences among points. The mathematical analysis uses that information to infer the changes in shape between the landmarks. If our sample of landmarks is sparse, we have good reason to worry about inferring changes between them – interpolating from sparse data is always a cause for concern. However, homology of points between the landmarks we sample is not imputed or determined by the deformation; homology is established in advance, by biological arguments. For a deformation to make mathematical sense, the points in one form must correspond to the points in another.

Sometimes it may seem that points cannot be homologous, as in ontogenetic studies, because bony tissue is added during growth. Thus the cells located at the suture between bones at one developmental stage are not the ones found at that suture at another developmental stage. *Histologically*, the points are not homologous, but it is nonetheless important

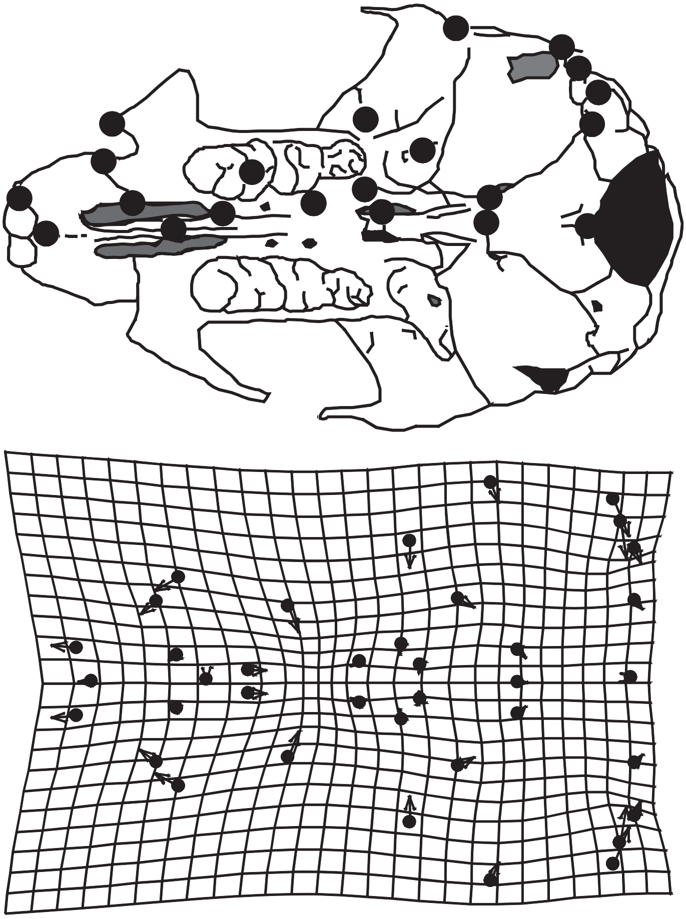


Figure 2.1 Landmarks sampled on the skull to show the interpolation of changes between landmarks based on the analysis of displacements of landmarks (relative to other landmarks).

that the anatomical parts be so, especially if we hope to compare growth from age to age or from species to species. We want to compare rates of growth of comparable parts. That the landmarks are located in different cell populations does not matter, but the comparability of the parts does.

Consistency of relative position

Morphometric methods cannot be applied properly when shapes are too different. For example, if bones are so radically altered in their topology that points on one have moved past other points (e.g. a foramen that was anterior to a tooth is now posterior to that tooth), the shapes may be too different to analyze geometrically. Also, in some cases landmarks may disappear altogether – as when a foramen is present in one taxon (or age) but not in another. Such changes, while undeniably interesting, are not suitable for a morphometric analysis. They are not matters of changes in shape so much as of changes in topology. Shape analysis is about shape, which makes the techniques fairly limited in their application, and the constraint can be serious. To some extent the constraint is implied by the idea of "shape analysis" (which is obviously an analysis of shape). It is also partly due to the mathematics of geometric methods, which rely on linear approximations. When shapes differ by too much, linear approximations are problematic. We obviously need methods to decide whether the changes are "too large" (an issue we discuss in more detail later, Chapter 4). However, if the changes are in topology, rather than shape, then the landmarks recording those topological changes are not suitable for a shape analysis.

Adequate coverage of the form

A third important criterion is adequate coverage of the form, or, as Roth (1993) put it, comprehensive coverage. That we need comprehensive coverage should be self-evident because we cannot detect changes without data, and the landmarks are the data. Additionally, we cannot find changes within particular regions unless we have landmarks within them. One way to decide if you have met this criterion is to draw a picture of the landmarks without tracing the rest of the organism. Given only that sample of landmarks, can you see the form of your organism? For example, Figure 2.2 shows two sets of landmarks for the same morphology (a squirrel scapula, one of the examples discussed later in this chapter). In Figure 2.2A the form of the scapula is present, even if the outline of the structure is erased; in Figure 2.2B it is virtually impossible even to tell that the structure is a scapula. Given the landmarks shown in Figure 2.2B, we cannot tell what is happening between the peripheral points (meaning those on the outline). Therefore, if there are any interesting and localized changes in scapula shape, we will not find them.

Sometimes we simply cannot find any landmarks in a particular region, and there is no choice but to accept sparse coverage; at other times sparse coverage is not acceptable, and so we may need to compromise and relax the criterion of homology. However, this relaxation must be done with great caution. For example, in studies of piranhas we need information on the changes in position and size of the eye within the head, even though there are no discrete points that can serve as landmarks just anterior and posterior to the eye. If we strictly enforce the criterion of homology, we could place a landmark in the middle of the eye; we could then detect changes in the location of the eye within the head. However, we would not have any information about the diameter of the eye, although changes in proportions of the eye are one of the most visually obvious ontogenetic changes in shape. We do not want to sacrifice that information. Thus, we place points that mark the anterior and posterior boundaries of the structure. Their homology is an untenable hypothesis even though the eye is a homologous feature of piranhas, but to provide the needed information we relaxed the criterion of homology and put landmarks at the same geometric location in every specimen.

It is dangerous to relax the criterion of homology too far. Some landmarks would be rejected by the criterion of homology, and cannot be justified by the criterion of comprehensive coverage. For example, traditional morphometric studies of mammals often include the measurement "least interorbital breadth." That measurement is taken as a transect across the frontal bone; where it is chosen is a function of where the distance between orbits is smallest – and where that distance is found might be arbitrary with respect to homology. Unlike the landmarks at the anterior and posterior of the eye in piranhas,

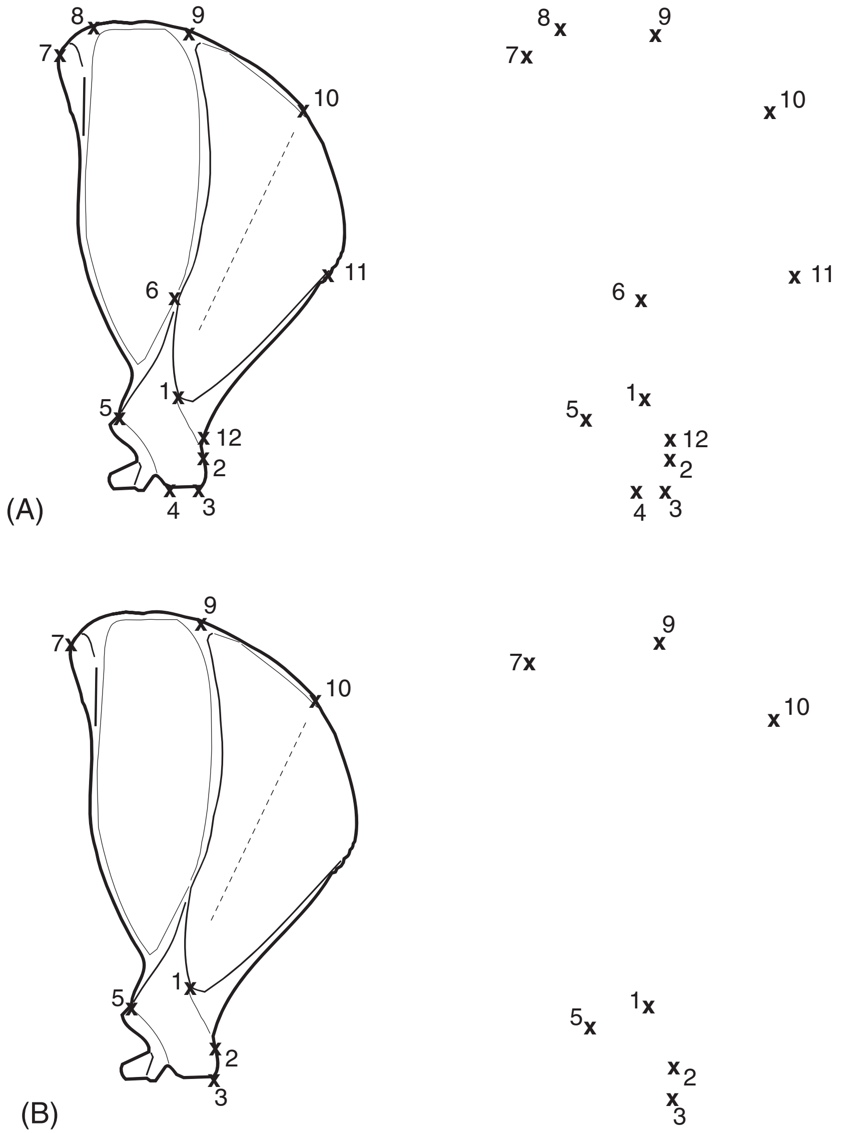


Figure 2.2 Landmarks on a squirrel scapula to show varying degrees of coverage: (A) comprehensive coverage; (B) limited coverage.

which are approximately constant in location, the endpoints of least interorbital breadth are not. They may be on entirely different parts of the frontal bone in different specimens. That is not a debatable case of homology – from the definition of the measurement it is obvious that it has no connection to homology.

Sometimes the homology of landmarks is debatable. For example, the anterior point of the dorsal fin base may be located on different structural elements in different species, but the point could be considered homologous as the anterior of the dorsal fin base. When debatable points are chosen, they need considerable justification.

Repeatability

The fourth criterion for selecting landmarks is that they can be located and relocated without error. Sometimes the amount of error is surprising. Some points that seem as though they ought to be difficult to find repeatedly, such as the anterior and posterior points on the eye (which are not discrete, clearly demarcated points), actually might be less prone to error than other points that appear more discrete and well-defined. Also, points that seem very fuzzy (such as blurs on x-rays) can sometimes be more reliable than you might imagine. It is probably best to avoid prejudging the landmarks (unless you simply cannot find them on several specimens) and instead check their repeatability empirically (methods for doing so are discussed in the next chapter).

Some landmarks are prone to error in only one dimension – for example, it might be easy to find its position along the anteroposterior axis but harder to determine its location along the dorsoventral axis. This can be a real problem for points that might otherwise be well defined, such as points on a suture. Sutures that generally follow a body axis sometimes wander, taking a complex path. It may be easy to pin down the anteroposterior location of a point along the suture, but more difficult to decide its mediolateral position. When a landmark is difficult to find in only one direction, the error will be concentrated in that direction; it will be biased, not random. Biased error is a more serious problem than a large random error because biased error will look like something that merits an explanation. However, the difficulty that you perceive in the course of digitizing may not be reflected in the actual variability of the point. At the outset of the analysis, before deciding that a point is unrepeatable in one or both directions, digitize it and then check its error. You can always delete it if you find that the error is biased.

Coplanarity of landmarks

The fifth criterion for selecting landmarks is related to the problem of analyzing three-dimensional organisms in two dimensions. It is not strictly necessary to reduce three-dimensional organisms into two dimensions because you can always examine more than one plane, and there are techniques for obtaining and analyzing three-dimensional landmarks (these are still relatively undeveloped and beyond the scope of this book, but we discuss them briefly and provide sources of further information about them in Chapter 15). Still, many readers will use two-dimensional approaches if only because the technology for three-dimensional data collection is expensive, so the possibility of distortion due to projecting a three-dimensional organism into a two-dimensional plane must be considered. To avoid this distortion, specimens must be consistently oriented under the camera, and one particular plane must be chosen for that orientation. Points not in that plane may be inconsistently oriented or difficult to interpret. The two-dimensional analysis will suggest that the points have moved within the plane of photography, but it is possible that they actually have moved toward or away from that plane. What you will see is the projection of a change in that third dimension onto the plane of the photograph. This can be a serious problem, and it turned out to be an issue in the analysis of cotton rat (*Sigmodon fulviventer*) skull ontogeny (Zelditch et al., 1992). One of the characteristic features of mammalian skull ontogeny is the change in orientation of the skull base. As a result, points initially on the posterior end of the ventral surface move dorsally, out of the picture plane with increasing age. Thus some points could not be included in the data set because they could not be seen at all ages in consistently oriented photographs. Even more problematically, other points that appear to be on the lateral boundary of the skull in a two-dimensional view are on the lateral *surface* of the skull. It was not possible to tell if they moved in the anteroposterior and mediolateral directions (the plane of photography) or if they instead moved dorsoventrally. In hindsight, those lateral points should have been excluded as too ambiguous.

Bookstein's typology of landmarks

Bookstein (1991) classified landmarks into three categories: Type 1, Type 2 and Type 3 (see Roth, 1993, for another discussion of these types). Type 1 landmarks are optimal, Type 2 are more problematic and Type 3 might not even be considered landmarks at all. The classification is based on two interrelated considerations: one is that landmarks ought to be locally defined, and the degree to which they are locally defined determines their classification; the other is the type of "epigenetic" explanation in which they can enter. The first consideration is relatively easy to summarize while the second is more difficult.

Landmarks are locally defined when they are located by particular structures close to the point. For example, the intersection between three bony sutures is locally defined. Bookstein refers to these as points at discrete juxtapositions of tissues, although they need not be juxtapositions of different tissue types – by his usage, the juxtaposition of three bones is a juxtaposition of tissue types. For these Type 1 landmarks you do not need to mention any structures far away from that point. In contrast, "the point furthest away from the tip of the snout along the dorsoventral axis" is not defined by any structures surrounding or near that point; instead, it is defined solely by being at an extreme distance from another point. This kind of landmark represents the other extreme, the Type 3 landmarks. The intermediate class, the Type 2 landmarks, includes such points as the tip of a tooth or end of a bony process; these are located at local minima and maxima of curvature, such as a bulge or tip of a structure. Like Type 1 landmarks they are defined in terms of specific local features, but like Type 3 landmarks they are defined as extremes of curvature or points furthest along (or away from) some structure.

Bookstein distinguishes these as different in kind partly because they enter into different kinds of explanations. Type 1 landmarks allow you to identify directions of forces that impinge on a structure, or to recognize the effects of processes moving the landmarks (e.g. bone deposition). This is because Type 1 landmarks are surrounded (in all directions). In contrast, Type 2 landmarks lack information from surrounding tissues in at least one direction such that you cannot distinguish between several possible directions in which forces might be applied. For example, one possibility is that forces are applied laterally to a structure, along its boundary, but another possibility is that some combination of forces is applied perpendicular to the boundary, some outward and some inward. From Type 2 landmarks you cannot decide between these alternatives. The lateral landmarks on the *Sigmodon* skull exemplify this case. Bookstein's Type 3 category might seem to include such "almost locally defined extrema", like the endpoints of eye, but these are probably Type 2, being extreme with respect to a very small, local structure. Points that are truly Type 3 are like those at the endpoints of our measurement of least interorbital breadth.

Examples: applying ideals to actual cases

Having discussed some general principles and theory, we now turn to specific and concrete examples of data. These examples are all taken from our own work, both because we can explain our own reasoning and because these cases will be used to demonstrate methods of analysis throughout this book.

Landmarks on the lateral surface of the squirrel scapula

Figure 2.3 shows the major anatomical features of a tree squirrel scapula. Also shown are 12 landmarks that were digitized in a study of changes in scapula shape associated with the evolution of burrowing in chipmunks and ground squirrels (Swiderski, 1993). Studies of scapulae of other mammals have found important changes in the blade, acromion and metacromion associated with functional shifts (Oxnard, 1968; Taylor, 1974; Stein, 1981). These same studies found little or no change in the coracoid process and the bell-shaped structure that articulates with the humerus (hidden behind the metacromion in Figure 2.3). A preliminary survey of squirrel scapulae indicated that they may have a similar anatomical distribution of changes. This pattern dictated that the squirrel scapulae should be digitized from the lateral view, because this is the only view in which the blade, acromion and metacromion could be seen in all taxa. Fortunately, the one feature of the bell that was considered potentially relevant to a functional analysis was also visible in the lateral view. That feature, the "neck" between the blade and the bell, is expected to change in thickness

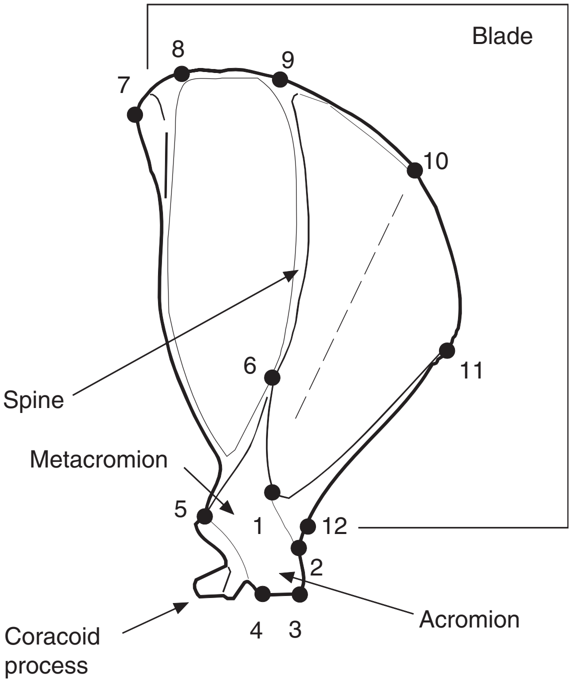


Figure 2.3 Landmarks on a squirrel scapula.

to reflect the magnitude of the forces transmitted to the scapula from the humerus. Thus, before any decisions were made about inclusion of specific landmarks, functional considerations were used to decide which general aspects of scapula shape would be analyzed.

The anticipated importance of changes in the acromion and metacromion meant that concerns about the distortion of three-dimensional aspects of shape could not be ignored, and also that landmarks could not be deleted if the distortion was expected to be large. Instead, concerns about distortion were addressed by standardizing the protocol used to capture the images that were digitized. As is usual for morphometric analyses based on photographs or video images, the scapula was placed in a standard orientation so that differences in orientation would not be interpreted as differences in shape. In addition, the distance of the camera lens from the scapula was adjusted for each specimen so that the blade always occupied the same proportion of the field. Then, if the height of the spine and sizes of the acromion and metacromion were proportional to the size of the blade, the acromion and metacromion would also occupy a constant proportion of the field. More importantly, the pattern of landmark displacement that would occur if these proportions changed could be predicted and tests for these patterns could be performed. No evidence of such patterns was found in the data.

After deciding which view to digitize, a major concern was coverage: finding enough landmarks to represent adequately the shape of the scapula. Structurally the scapula is rather simple, which means there are few points that can be uniquely defined. This is especially true of the main portion of the scapula, the semi-circular or triangular "blade"; the blade is nearly flat and has only two ridges crossing it – the large scapular spine on the lateral surface, and the smaller subscapular ridge on the medial surface. The margin of the blade is also rather featureless, having few corners and no spines, only more ridges or thickenings.

Despite the shortage of potential landmarks, it was still considered important to define them so that they could reasonably be considered homologous. For example, the ends of ridges may seem to be good landmarks, but quite often these are gently tapered, making it difficult to define precisely where they end. Usually, when a ridge ends abruptly, it ends at an intersection with some other structure. On the scapula blade, landmarks 8, 9 and 10 are points where two ridges intersect. Landmark 6, on the metacromion, is another intersection, marking the attachment of the metacromion to the spine. Landmarks 7 and 11 are points on the margin of the blade where the end of a marginal ridge is associated with a corner. Landmark 5, on the metacromion, is another corner associated with the end of a marginal ridge. Landmark 1 is one of the few places on the blade where a ridge (the scapular spine) ends abruptly without intersecting another structure.

Concern for homology extended to the corners as well as the ends of the ridges. Landmarks 2, 3 and 4 are at the only corners that are not associated with the ends of ridges. Other anatomical information was used to infer their homology. Landmarks 2 and 3 are corners where the acromion terminates in a flat surface that articulates with the clavicle. The corner labeled as landmark 4 appears to mark the boundary between the acromion and metacromion. This interpretation is reinforced by the point's proximity to the line of the scapular spine, which separates anterior and posterior components of both the scapula and the attached muscles.

The grounds for inferring homology are weakest for landmark 12. This is the only point on the articulating structure, the "bell," that could be seen in lateral view in all taxa. If more points on this structure were visible, landmark 12 might not have been used. This point is identified only as the cranial edge of the neck, which is the narrowest region between the blade and the bell of the articulating structure. This criterion for recognizing a landmark is harder to apply than the criteria for recognizing the other 11 landmarks because the boundary between bell and blade is not marked by a corner or other distinctive feature. In this regard the neck of the scapula may seem similar to the least interorbital width of the skull, as being poorly defined and of doubtful homology. However, unlike least interorbital width, the neck of the scapula marks the boundary of two functionally distinct components of the scapula. In addition, analysis of digitizing error indicated that this point was not substantially harder to locate than other landmarks. Therefore, doubts about the homology of this point were set aside in favor of having at least one landmark on this structure.

Landmarks on the external body of piranhas

Figure 2.4 illustrates the landmarks used in several studies of shape change in piranhas. These points were originally intended for analyses of shape by trusses (see Strauss and Bookstein, 1982), so they were chosen to allow for constructing a series of boxes and diagonals over the form. In addition, because the truss analysis was to be compared to more traditional measurement schemes in ichthyology, some landmarks were chosen to allow duplication of those measures. Traditional measurements between these landmarks were used in a systematic study of *Pygocentrus* (Fink, 1993), and in several geometric morphometric studies of the evolution of piranha ontogeny and the diversification of their body forms (e.g. Zelditch et al., 2000, 2003a).

Selecting landmarks on the lateral body of piranhas is relatively straightforward because specimens are essentially two-dimensional. Most of the shape variation can be seen in that view, and little distortion is caused by viewing the animal in a plane. Specimen bending can occur at fixation or during preservation, and such bent specimens were not included in analyses unless they could be manually straightened with no resulting distortion in the

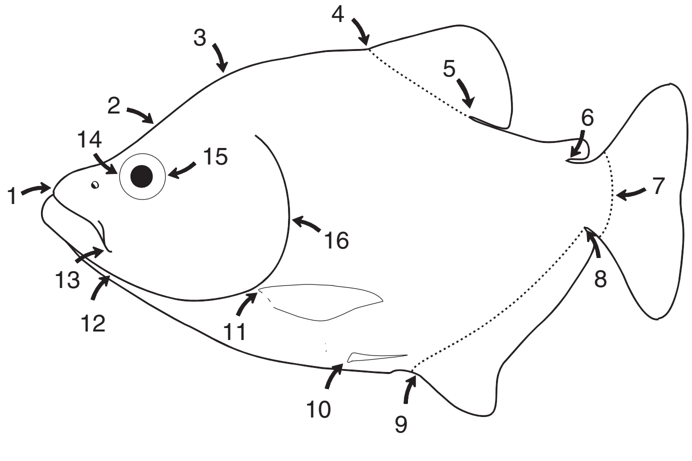


Figure 2.4 Landmarks on the external body of a piranha.

lateral body shape. Data acquisition consisted of placing the specimen in a standard view, using a specially designed container that kept the animal's midline in the plane defined by the top edges of the container. A piece of metric graph paper was placed on the container's edge in the same plane for calculating size. In some cases, insect pins of various sizes were used to make landmarks more visible. The camera was placed so that each specimen occupied approximately the same area in the viewing field, in order to minimize distortion.

There are few landmarks on the postcranial lateral body of piranhas, and almost all landmarks chosen are from around the perimeter of the body. Had the data been taken from radiographs, some internal osteological landmarks could have been used. However, it was decided that data would be taken from entire specimens, partly to facilitate application to identification keys. Most of the landmarks chosen are at boundaries or extremes of structures, or are skeletal features accessible without x-rays.

Landmark 1 represents the anterior point of the head, and is taken where the two premaxillary bones articulate at the midline. Because this point is directly on a vertical from the plane of the specimen, no special marking is required. The landmark involves soft tissues, and thus could be affected by desiccation of the specimen.

Landmarks 2, 3, 7 and 12–16 all represent skeletal features, representing extremal points, intersections of structures, or borders of bones. Landmark 2 is the anterior border of the epiphyseal bar – a small extension of bone that spans a large fossa in the dorsal neurocranium – and was chosen to provide information on the shape of the head. The landmark is found by inserting a pin through the skin of the midline dorsal to the orbital region, where the pin just penetrates past the bar into the brain cavity. Although this landmark is constantly available in piranhas, some related fishes show ontogenetic change in the width of the bar, such that the bone grows anteriorly as the fish grows, independent of head shape changes. Landmark 3 lies at the posterior tip of the supraoccipital bone of the neurocranium. It lies just under the skin at the dorsal midline, and is found by moving a fingernail along the midline until the junction between bone and muscle is found. A pin is inserted at that point for purposes of digitizing. Landmark 7 represents the posterior termination of the hypural bones of the caudal skeleton, traditionally a point used in the calculation of standard length (tip of snout to base of caudal fin). In piranhas there is a concavity in the hypural bones at the lateral midline such that the bone lies anterior to the rest of the posterior border of the caudal skeleton, so the actual point measured is where the bone would be in other teleosts. This is less problematic than it might seem, since the actual measurement is done at the area where the caudal fin base can be bent laterally. Until some experience in finding this landmark is gained, it may be difficult to be consistent in reproducing this point. An inexperienced person usually has error in the anteroposterior axis. This is a landmark for which some argument regarding homology must be made. This is because the internal skeleton may not be consistent with the point used externally. However, consistently measured as the posterior termination of the body at the lateral midline, the point may be considered homologous.

Landmark 12 represents the ventral side of the articulation between the quadrate bone and the mandible. It thus lies lateral to the midline, although it usually lies on a vertical from the ventral midline. This point is located by placing a fingernail in the joint between the two bones, and then a pin is inserted in the joint. Landmark 13 lies at the intersection of the maxillary bone and the infraorbital bone that defines the "cheek" area of the face. The point lies well lateral to the midline, but marks an important area of the skull, approximating the length of the upper jaw. This point is marked by slipping the pin under the infraorbital bone adjacent to the posterior border of the maxillary. This landmark is composed of an extreme point (the posterior maxillary border) as well as an intersection of two structures. The homology of this landmark may be questioned. Landmarks 14 and 15 capture the width of the bony orbit. Each point lies at the extreme of the orbit along the anteroposterior body axis. Both of these landmarks are of questionable homology, but are taken because the eye has been shown to be highly allometric and has been used in traditional measurement schemes. With practice, these landmarks can be taken with little error.

Landmark 16 is perhaps the most difficult to justify in this analysis. It occupies the most posterior point of the bony opercle, the bone that forms the bulk of the gill cover, and its original purpose was to duplicate the landmark used in traditional ichthyology measurements of head length. This landmark was expected to be an articulation point between the opercle and subopercle bones. However, comparisons among several species showed that the position of the articulation varied excessively, and inaccurately represented the posterior of the head. The landmark now taken is simply the extreme along the bone border as measured from the tip of the snout. No reasonable homology argument can be made for this landmark; it may be that it is partially redundant with landmark 11. However, our analyses have shown that this landmark can be consistently digitized and is often informative about alterations in head shape.

Landmarks 4–6 and 8–11 represent points where the fins insert on the body, at the anterior or posterior of the fin base. In most cases these points are measured where the bony fin ray intersects the body. Together these landmarks provide a great deal of information on postcranial body shape. Landmarks 4 and 5 lie at the anterior and posterior of the dorsal fin base, respectively. Ontogenetic variation in anterior fin ray morphology can reduce the repeatability of landmark 4, as discussed in Fink (1993).

Landmarks 8 and 9 represent the posterior and anterior of the anal fin base. Often the fin is collapsed, and a pin must the inserted to make landmark 8 visible. In some piranhas there are accessory spines at the anterior of the fin base, and they are not included. Landmarks 10 and 11 represent the insertion onto the underlying skeletal girdles of the pelvic and pectoral fins, respectively. Both lie dorsolateral to the ventral midline. Landmark 10 is easily visible in larger specimens, but in some smaller specimens the transparency of the fin makes it difficult to find; in this case it can be located by raising the fin laterally and placing a pin at the anterior fin-ray's base.

Landmark 6 lies at the posterior base of the fleshy adipose fin, where the fin meets the skin of the dorsal midline. This point may be difficult to locate unambiguously because it may be obscured by the fin overlapping the skin of the peduncle, so a pin is inserted to mark its location for digitizing. In some of our studies we have attempted to use the anterior insertion of the adipose fin, but its broadly curving profile in many species renders it too difficult to repeat.

Note that landmarks 9–12 represent the ventral area of the body form, but they do not capture the actual convex belly shape of these fishes. A great deal of effort was spent in attempting to find appropriate landmark locations along the ventral profile, but no repeatable and consistent landmarks were found that could be located on all piranha species.

Landmarks on the skull of *Sigmodon fulviventer* and *Mus musculus domesticus*

The landmarks on the skull of cotton rats, *Sigmodon fulviventer* (Figure 2.5), were selected to cover the skull as evenly as possible for the purpose of determining whether ontogenetic changes in skull form are spatially integrated or localized (Zelditch et al., 1992) and to study developmental constraints on variability in that species (Zelditch et al., 1993). Because the studies were designed to analyze the ontogeny of skull shape and its variation, the only landmarks that could be included in the analysis are those that are visible in (approximately) the same plane at all ontogenetic stages. Because mammalian skulls are highly three-dimensional structures, and the cranial base rotates during ontogeny, landmarks that are parallel to the camera at one stage may rotate out of that plane later. This produces what appears to be a change in shape (within the plane). However, omitting all landmarks that might be affected by such a rotation would mean losing vital information about cranial length and width, because the landmarks most strongly affected by extension of the cranial base are the ones marking the juncture between the anterior and posterior cranial base, and those located on the posterolateral braincase. Consequently, landmarks were placed on those locations even though that complicates distinguishing between changes in shape caused by differential growth and apparent changes in shape due to the rotation of bones in the third dimension.

A subsequent study was undertaken to compare skull shape ontogeny of *S. fulviventer* to that of another rodent, the house mouse *Mus musculus domesticus*. A major objective of that study was to examine the relationship between life-history strategy and timing of skull morphogenesis (Zelditch et al., 2003b). Ideally we would have sampled both skulls densely, selecting homologous landmarks that provide a richly detailed description of the ontogeny of both species. However, some landmarks could be seen in only one species or another. For example, in *S. fulviventer* we can locate a landmark on the posterior of the glenoid fossa, but the curve of the glenoid is so smooth in *M. m. domesticus* that we cannot find a distinct point anywhere comparable to the glenoid landmark of *S. fulviventer*. To capture information about skull width in the region of the zygomatic arch of *M. m. domesticus*, a different point had to be chosen, complicating the comparative analysis. Several other points that are readily visible in *S. fulviventer* also cannot be found in *M. m. domesticus*. However, the problem posed by the inability to find landmarks in *M. m. domesticus* that are homologous with those already measured in *S. fulviventer* is partly mitigated because there are landmarks in *S. fulviventer* that had not been previously sampled, but which can be recognized in both species. Thus, in the comparative study, additional landmarks were sampled on *S. fulviventer*. Even so, the set of landmarks common to both species comprises a rather sparse sample of each skull. Therefore, analyses were done separately for each species, using the landmarks providing the densest coverage possible for each species, and the comparative analyses exploited the subset of landmarks common to both.

The landmarks selected for the original analysis of *S. fulviventer* (Zelditch et al., 1992, 1993; Figure 2.5A) include:

1. the lateral margin of the incisive alveolus where it intersects the outline of the skull in the photographic plane (IN)
2. the anteriormost point on the zygomatic spine (ZS)

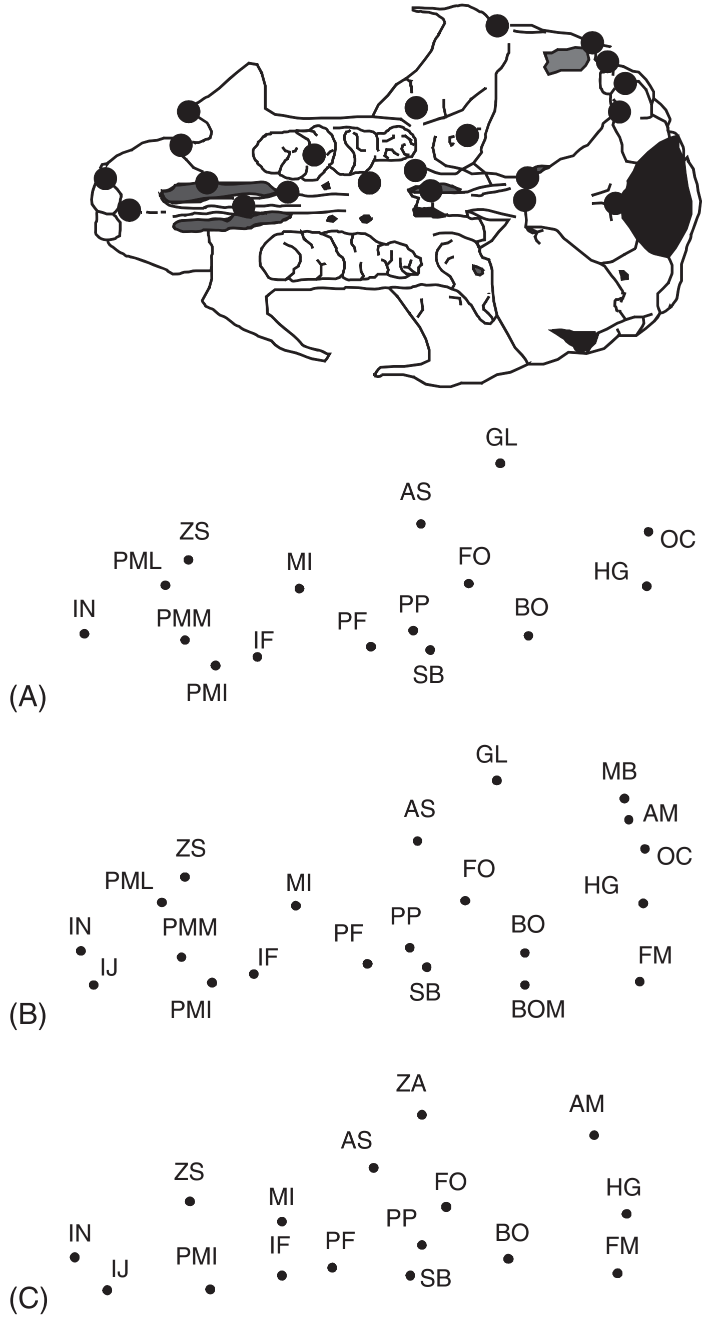


Figure 2.5 Sixteen landmarks on the ventral view of the skull of the cotton rat (*Sigmodon fulviventer*). (A) Landmarks selected for the analysis of ontogenetic change in this species; (B) the landmarks shown in Figure 2.5A supplemented by those that would allow for comparisons to *Mus musculus domesticus*; (C) landmarks sampled on skulls of *M. m. domesticus*. Comparisons between the two species used the landmarks shown in Figure 2.5C except for ZA.

3. the premaxilla–maxilla suture where it intersects the outline of the skull in the photographic plane (PML)
4. the premaxilla–maxilla suture lateral to the incisive foramen (PML)
5. the posteriormost point of the incisive foramen (IF)
6. the median mure of the first molar (M1)
7. the posterolateral palatine pit (PP)
8. the junction between squamosal, alisphenoid and frontal on the squamosal–alisphenoid side of the suture (AS)
9. the midpoint along the posterior margin of the glenoid fossa (GL)
10. the anteriormost point of the foramen ovale (FO)
11. the most lateral point on the presphenoid–basisphenoid suture where it intersects the sphenopalatine vacuity in the photographic plane (SB)
12. the most lateral point on the basisphenoid–basioccipital suture (BO)
13. the hypoglossal foramen (HG)
14. the juncture between the paraoccipital process and mastoid portion of the temporal bone (OC).

Several landmarks were added to these in the later study, designed to compare *S. fulviventer* to *M. m. domesticus*. These additional landmarks (Figure 2.5B) are:

1. the juncture between the incisors on the premaxillary bone (IJ)
2. the midpoint of the basisphenoid–basioccipital suture along the sagittal axis (BOM)
3. the midpoint of foramen magnum (FM)
4. the juncture of mastoid, squamosal and bullae (MB)
5. the juncture between the mastoid and the medial end of the auditory tube (AM).

The landmarks sampled on the skulls of *M. m. domesticus* include a subset of the original *Sigmodon* landmarks, plus the newly added ones, and a point at the interior corner formed by the intersection of the zygomatic arch with the braincase (ZA) (Figure 2.5C).

Image acquisition and manipulation

In this section, we discuss the basics of creating the image files you will be digitizing. Included below are the rudiments of taking a picture (what makes the image and how can you manipulate the camera and the lighting to get a better image), important differences between a photograph and a digital image file, image file formats, and a few things you can do to make the captured image even better. Because of the wide range of hardware and software available for acquiring and editing images and the rapid advancement of the technology behind those tools, most of the following discussion is quite general. You will need to familiarize yourself with the characteristics of your particular system (but then, you will need to do that anyway to get the best possible results). Our goal is that this section provides you with an orientation that gets you through the first stages of that familiarization with a minimum of unnecessary pain.

Inside the camera

The aperture

Arguably the most important part of a camera is the aperture – the small hole that lets light into the box to form the image. The aperture is critical, because light reflecting off the

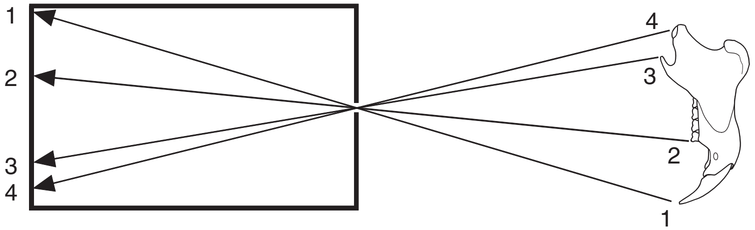


Figure 2.6 Image formation in an idealized pinhole camera. Light rays travel in a straight line from a point on the object (the squirrel jaw) through the aperture to a point on the back wall of the box. The geometric arrangement of the starting locations of light rays is reproduced by their arrival locations.

object is leaving it in many different directions and the aperture functions as a filter that selects light rays based on their direction of travel. The only rays admitted into the camera are those traveling on a path that takes them through the aperture. In theory, if the aperture is small enough (and nothing else intervenes), it insures that the geometric arrangement of the rays' starting locations is exactly reproduced by the geometric arrangement of the rays' arrival locations (Figure 2.6). This is why a child's pinhole shoebox camera works. It is also why the image is inverted.

Because the image is formed from a cone of light leaving a three-dimensional object and arriving on a two-dimensional surface, there are certain artifacts or distortions introduced in the image (Figure 2.7). As discussed below, a good lens system can reduce these effects, but it cannot eliminate them completely. One distortion in photographic images of three-dimensional objects is that an object that is closer to the camera will appear to be magnified relative to an object that is farther from the camera (Figure 2.7A). This occurs because the light rays traveling toward the lens from opposite corners of the object form a larger angle when the object is closer to the aperture, which means they will form a larger image inside the camera. For this same reason, the closer feature will hide more distant features. Another distortion is that a smaller object in the field may appear to be behind a taller object when the smaller object is actually next to the taller one (Figure 2.7B). The reality is that the smaller object *is* farther from the aperture, but not in the expected direction. In a related phenomenon, surfaces of an object that face toward the center of the field of view will be visible in the image, and surfaces that face away from the center will not be visible (Figure 2.7C). This is the reason buildings appear to lean away from the camera in aerial photographs. In the case of a sphere, this means that the visible edge (horizon) will not be the equator, but will be closer to the camera than the equator. If the sphere is not centered in the image, the horizon will also be tilted toward the center of the image (this effect can be a serious obstacle to digitizing landmarks on the sagittal plane of a skull). These phenomena are more pronounced near the edges of the image, so one way to reduce their influence on your results is to be consistent in positioning your specimens in the center of the field of view. Towards the end of the next section we discuss other steps you can take to minimize these distortions and the effects they would have on your morphometric data.

What the lens does

The lens does two things: it magnifies the image, and it makes it possible to use a larger aperture than a pinhole. Both are important advantages, but they come at a cost.

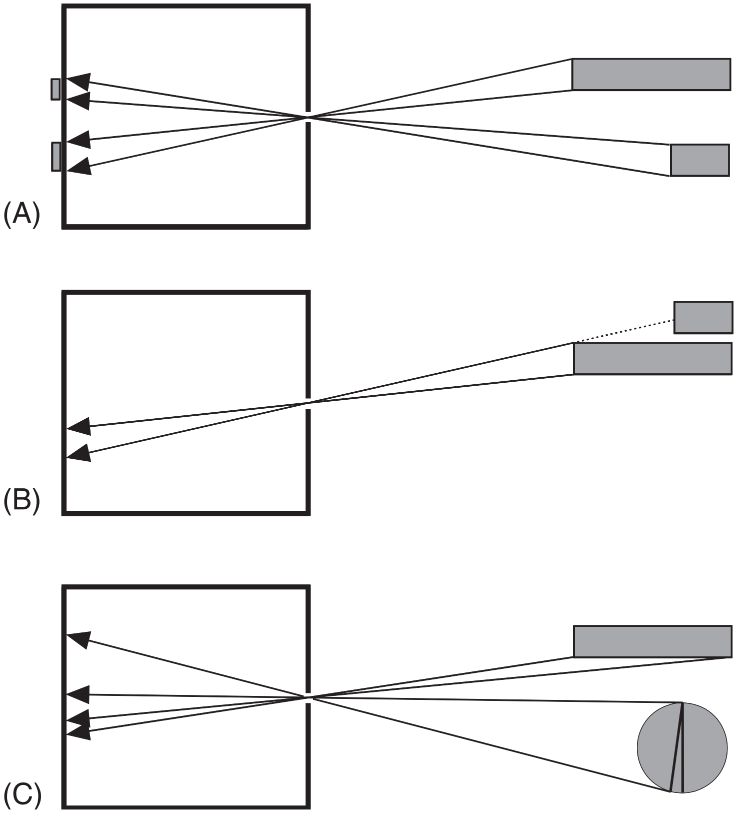


Figure 2.7 Distortions resulting from light leaving a three-dimensional surface and arriving on a two-dimensional plane: (A) the two rectangles have the same width, but the upper ("taller") rectangle produces a larger image (appears to be magnified) because its end is closer to the camera; (B) an object that is farther from the center appears to be behind an object that is closer to the center of the image, especially if the more central object is "taller"; (C) the sides of an object that face the center of the field are visible, and the surfaces that face away from the center are hidden. In the special case of a spherical object, less than half of the surface will be visible; if the object is not centered in the field, the apparent horizon will be tilted away from the expected reference plane (the equator) toward the aperture.

The size of the image in a pinhole camera is a function of the ratio of two distances: (1) the distance from the object to the aperture, and (2) the distance from the aperture to the back of the box. If the object is far from the pinhole, light rays converging on the aperture from different ends of the object will form a small angle. The light rays will leave the pinhole at the same small angle, so the image will be smaller than the object unless the box is very large. One way to enlarge the image is to enlarge the box; another is to shorten the distance between the camera and the object, so that light rays converging on the aperture from different ends of the object will form a very large angle covering the back surface (Figure 2.8A). A lens magnifies the image by changing the paths of the light rays arriving at the lens, so that the angle between them when they depart the lens is greater than the angle between them when they arrived. Consequently the image is larger than it would be without a lens, making the object appear to be closer to the lens than it is (Figure 2.8B).

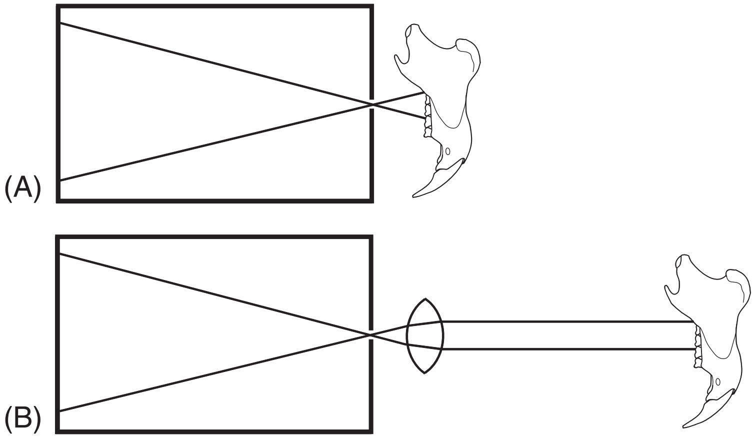


Figure 2.8 Two methods of image magnification: (A) moving the camera and object closer together; (B) using a lens to change the paths that the light travels from the object to the image, thereby changing the *apparent* distance of the object from the aperture.

The amount of magnification produced by a lens depends on several factors. Light striking the surface of the lens at 90° does not change direction, but as the angle of incidence becomes more acute the change in direction increases. Exactly how much the path of the light is bent depends on the properties of the material of which the lens is made and on the wavelength of the light. The light changes direction again when it exits the lens. As well as the advantage of having a larger image, which enhances resolution, there is the additional advantage that the distortions that occur in images of three-dimensional objects are reduced. Because the object is farther away than it would be for a pinhole image of the same magnification, the same small aperture is now selecting a narrower cone of rays leaving the object. This is particularly true for features near the center of the image. Features near the edges of the image are subject to other distortions (see below).

The image in a pinhole camera is faint because the pinhole must be small to be an effective filter of the light rays' directions of travel. A larger aperture would admit more of the light leaving the object, but it would produce a fuzzier image because a larger cone of light leaving each point on the object would reach the back of the box (Figure 2.9A). Consequently, features in the image would have wide diffuse edges, and the edges of adjacent features would overlap, making it impossible to discriminate between those features. The lens corrects this problem by bending the light so that the cone converges again at some point on the other side of the lens (Figure 2.9B). This allows you to increase the size of the aperture, allowing more light from the object to reach the back of the box. The image that results is generally brighter and has more contrast between light and dark areas.

The principal cost of using a lens is that it imposes a particular relationship on the distances from the lens to the object and the image. This relationship is expressed by the following equation:

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i} \tag{2.1}$$

where d_o and d_i are the distances to the object and image, respectively. The value of f is determined by the shape and material properties of the lens, and is called the focal length.

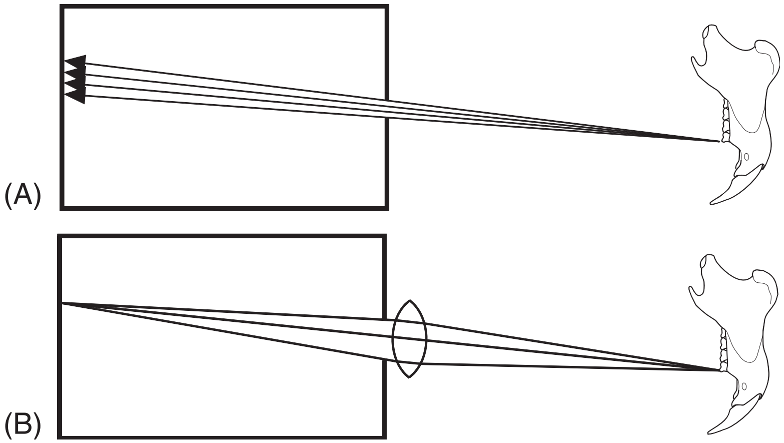


Figure 2.9 The role of the lens in enhancing image resolution: (A) a large aperture admits many rays leaving the object in divergent directions, which produces a fuzzy image because each point on the object produces a relatively large circle of light at the back of the box; (B) the lens bends the light so the diverging rays from a point on the object converge on a point at the back of the box.

For the cone of light from a particular point on the object to converge again at the back of the camera, that point on the object must be a specific distance from the lens. At this distance, that point is "in focus." If a part of the object is not at this optimal distance, light leaving that part does not converge at the right distance behind the lens, and that part of the image is blurred. The thickness of the zone in which this effect is negligible is the depth of field. Greater depth of field means that a thicker section of the specimen will be perceived as in focus. Depth of field decreases with magnification – at higher magnification the light is bent more as it passes through the lens, so the difference in focal points is magnified as much as the areas of the features. Consequently, a thinner section of the specimen is in focus. To further complicate matters, in simple (single-lens) optical systems the slice that is in focus is curved, not flat. Similarly, the surface on which the image is in focus is also curved. The complex lens systems of higher quality optical equipment flatten these surfaces considerably, but you may still find that only the center of a flat object (or a ring around the center) is in focus. The best solution for this problem is to use a lower magnification, increasing the depth of field. There are things you can do to edit the "captured" image, but, as discussed in a later section, these are limited by the initial quality of the image.

Near the edge of the image, additional distortions produced by the lens may become apparent. These distortions arise because the amount that the path of light is bent is not just a function of the properties of the lens; the deflection is also a function of the angle of incidence and the wavelength of the light. Two rays arriving at the center of the lens from locations near the center of the field of view are bent by relatively small amounts because they strike the surface at nearly 90°, but two rays arriving at the lens from locations near the edge of the field of view are not only bent by larger amounts; the difference in how much they are bent is also greater. Consequently, a straight line passing through the field will be curved in the image unless it passes through the center of the field. Closer to the edge of the field, differences in how much different wavelengths of light are bent by the lens may also become apparent as rainbow fringes on the edges of features in the image. This effect is most evident under high magnification or very bright light.

Checking your system

The complex lens systems of higher quality optical equipment greatly reduce all of the distortions discussed above, but none of these distortions can be eliminated completely. Fortunately, there are a few simple things you can do to insure that the effects on your data are negligible. The first is to put a piece of graph paper in the field and note where the rainbow effect, if any, becomes apparent. Next, digitize several points at regular intervals along a line through the center and compute the distances between the points. As you approach the edge of the image, the interval will gradually change. Take note of where this effect begins to be appreciable; you will want to keep the image of your specimen inside of this region. In other words, if the object is large, place the camera at a greater distance so that the image does not extend into the distorted region of the field. Next, get a box or other object with a flat bottom and vertical sides and mark one side of the box at a height corresponding to the thickness of your specimens. Put the box in the field of view, with the marked side at the center of the field. Slowly slide the box toward one edge of the field until you can see the inner surface of the side between the mark and the bottom. Again, you will want to keep your specimens inside this region. Outside of this region, features at the height of the mark will appear to be displaced away from the center of the image. Finally, check your depth of field by putting a sloped object marked with the thickness of your specimen (or one of your larger specimens) in the field. If all the critical features are not in focus at the same time, you should use a lower magnification to avoid guessing where in the fuzz is the feature you want to digitize.

What happens in the back of a camera

Now that you have a minimally distorted image at the back of your camera, you need to "capture" the image with a light-sensitive device (either the detector array in a digital camera, or the film in a conventional camera). These devices contain a large number of light detectors used to record the image (silver crystals in film; pixels in digital cameras). A black-and-white (gray-scale) detector is a single light-sensitive device that reports total light intensity (number of photons per time unit), whereas a color detector is a bundle of three light-sensitive devices recording intensities in three narrow color ranges. The higher the number of detectors, the higher the resolution of the recorded image will be. Detector number can be increased either by reducing the size of the detectors, so more can be packed into the same area, or by increasing the size of the image to span more detectors (as in large format films). Typically, detector number is higher for black-and-white equipment than for color equipment. Because the three-color bundle occupies more area, color devices tend to produce less resolved images than black-and-white devices (although a new color device may be an improvement over an old black-and-white device). Most software for processing color images can convert them to gray scale.

The image captured by your camera is not the image you digitize. If you are using a digital camera and projecting an image to the screen, you may never see the captured image; what you see on the screen is a second image produced from the information collected by the detectors. If the camera image is mapped to the screen one-to-one, pixel for pixel, the two images will have the same resolution. You can enlarge or reduce the size of this image, but you cannot increase the resolution. When the screen image is reduced, information from multiple camera pixels is averaged for display by a single screen pixel. This may produce an image that looks sharper, but that is the result of losing the small-scale, almost imperceptible details that created the original "fuzz." The reduced picture may be easier to interpret, but there is the risk that the features you want to digitize have been merged. When the screen image is enlarged, information from a single camera pixel is displayed by multiple screen pixels, which produces the blocky, stair-step effect. The result seems less resolved because the edges are not smooth, but features that were separated before still are separated. The drawback is that excessive enlargement may make the image difficult to interpret and increase the mental strain of digitizing. If you are using a conventional camera and making a print from the negative, the same principles apply. In addition, the quality of the print will depend on the quality of the lenses in the enlarger and the size of the grains in the paper.

Saving image files

Once you have an image "captured", you must choose the format in which to save it (we assume you want to save the image before editing it). Most image file formats are raster formats (also called bitmap formats), where the image is represented as a set of values assigned to a grid. This format reflects the structure of your screen and the detector array in your digital camera. BMP, TIFF and JPEG are all raster formats; so is the format used in the Windows clipboard. The principal alternative to the raster format is the vector format, in which the image is represented by a series of mathematical formulae that specify a set of geometric shapes. This format has some advantages over the raster format, but it does not work well with photographic images of biological specimens because their complexity requires a large number of geometric shapes to be mapped onto the specimen. Meta formats, such as that used in Windows metafiles (*.WMF), allow data in multiple formats in the same file, permitting the user to build up complex compositions (e.g. a picture, plus a graph, plus text).

The quality of an image reproduced from a raster file depends on the number of bits used to save the information at each cell (pixel) in the grid. The number of bits determines the number of colors or gray tones in the image – a 16-bit image can contain up to 64 000 colors, an 8-bit image can contain only 256 colors. Each pixel displays only a single color, so the advantage of the 16-bit image is that it can have much smaller changes in color from one pixel to the next and thus can more accurately reflect graded changes in color across the object. In practice, the 8-bit image may not be noticeably poorer unless the image size is changed, and has the advantage of requiring much less disk space and less time to load. If space is important and color is not, saving the image in gray scale will considerably reduce disk space and loading time without reducing resolution. If color is important, the most economical format is JPEG (Joint Photographic Experts Group, *.JPG). This is a compressed format analogous to the *.ZIP format. An image that requires 900 K of disk space as a 16-bit BMP or TIFF file might require less than 100 K as a minimally compressed JPEG. More importantly, the 100 K JPEG will look just as good on the screen because there is very little information lost in the compression. In contrast, a 4-bit BMP file requiring about the same disk space will have lost much more information and look considerably worse.

If you are really pressed for disk space but need to preserve as much color information as you can, explore the options in your software. Normal color reduction replaces each pixel with the nearest color in the reduced color set (e.g. emerald, jade and lime will all be replaced with green). This creates large blocks of uniform color that obliterate many details. Various optimizations and diffusion algorithms produce "speckled" images that blend into more natural colors when viewed at a distance, or when reduced. These also do a better job of preserving edges.

Improving the image

Before you capture it

What you see in the image depends on how much light you shine on the object, and how much of the light reflected from the object you allow to reach the detector. There are several options for manipulating light; the trick is to find the right balance so that you can see the features you want to digitize. It is important to understand that the best image for digitizing may not be the most esthetically pleasing image.

When you shine a light from a single source on a three-dimensional object, some parts are likely to be in shadow. Shadows can be advantageous in that they allow you to see the relief, but you want to avoid a shadow so dark that you cannot see anything *in* the shadow. Backlighting allows you to see features in the shadow without obliterating the shadows. This is achieved by using a weaker light, or some kind of reflector (e.g. a piece of white paper) to illuminate the "back" of the object.

The size of the aperture and the amount of time it is open determine the amount of light that strikes the detector. A larger aperture admits more light, but, as discussed above, the image is less sharply resolved. However, minimizing the aperture does not necessarily produce the most useful picture. A small aperture allows very little light to reach the detector from any area, and the resulting image is generally dark. You can compensate for this by decreasing the shutter speed (or its digital analog); this allows light through the aperture to register on the detector for a longer period of time. As the shutter stays open longer, the brighter parts of the specimens become brighter in the image and the dark parts of the image stay dark. In other words, the contrast is increased. Unfortunately, minimizing shutter speed does not necessarily produce the best picture either. The longer the time that light is collected, the longer the fuzzy fringes register on the detector. If you leave the aperture open too long, eventually thin dark features will be obliterated completely and small bright areas will appear larger than they really are. You can also compensate for small aperture size by using brighter lights to illuminate the object. This has much the same effect as increasing the time the aperture is open. More light registers because there is more light from the specimen per unit time.

In summary, getting a decent picture may require a delicate and sometimes annoying balancing act. We strongly recommend that you try many different settings to see what works best, and keep a log of the conditions in which each picture was taken. In your log, you should also take note of the brightness and shininess of the specimen. A dull gray specimen may require a different set-up than a shiny white specimen. You should also take note of what other room lights are on, and if the room where you are working has windows the time of day can be an important factor as well. Have patience. Although there is a lot you can do to edit an image, you can only highlight information that is already there; you can't recover information that was lost by the original.

After you capture it

A quick tour through almost any photo-editing software will reveal a bewildering array of functions you could use to modify your image. Here, we discuss a few tools that are widely available and likely to be useful to a large number of biologists. All of these manipulations reduce the accuracy of the image as a reproduction of the original image. Again, it is important to realize that an esthetically pleasing or artistically interesting image may not be the best one to digitize for a morphometric analysis.

Probably the two most generally useful tools are the ones that adjust brightness and contrast. These functions can be most easily understood if your image editor displays a histogram of pixel luminance (the intensity of light emitted). Increasing brightness makes the whole image lighter, adding the same increment of luminance to every cell, up to the maximum value. Detail is lost at the bright end because pixels near that end converge on the maximum value, and details at the dark end may emerge as they are brought into a range where the differences between adjacent cells become perceptible. Except for the pixels near the bright end, the actual difference in brightness between adjacent cells does not change (the peaks in the histogram move toward the bright end, but they do not change shape). Decreasing brightness has the opposite effect. Increasing contrast makes the dark areas darker and the bright areas brighter, shifting the peaks away from the middle and towards the ends. Decreasing contrast makes everything a homogeneous gray, shifting the peaks toward the middle. The peaks also change shape as they move, becoming narrower and taller with decreasing contrast, and wider and flatter with increasing contrast. Again, differences between adjacent cells are lost as their values converge on the ends or the middle. Adjustments of either brightness or contrast can be used to make features near the middle of the brightness range easier to distinguish; the difference is whether the features that are made harder to distinguish are at one end (brightness) or both ends (contrast) of the range.

As noted above, computer images have jagged edges due to their raster formats. When the image is scaled up, it is also apparent that sharp edges in the original are represented as a transition zone with large steps in brightness and/or color. This creates the problem of deciding exactly where in the zone is the edge you want to digitize. Adjusting brightness is unlikely to solve this problem, because the number of steps and the difference between them stays the same. Increasing contrast can help more, because it makes the steps bigger, but this comes at the expense of making the jaggedness more apparent. Even so, narrowing the zone of transition may be worth the increased jaggedness. Some alternatives to increasing contrast may include sharpening and edge enhancement. These tools use more complex operations that both shift and change the shapes of the luminance peaks, but they also can produce images with thinner edges. In general terms, the effect is similar to increasing contrast, but the computations are performed on a more local scale. Which tool works best to highlight the features you want to digitize will depend on the composition of your picture. Consequently, what works for one image may not work for another.

Digitizing

This discussion of digitizing explains how to use one particular program, **tpsDig**. We recommend using this because all programs for shape analysis can read data in the format output by this program.

1. Getting ready to digitize your first image

Start the program. Go to the **File** menu. Select **Input source**, then **File**. Find the folder containing your pictures and select the correct type of files. Select the file you want to open, or type the file name and extension, and click the **Open** button. The image should appear in the main window. In the toolbar above the image are two buttons (+ and −) that you can use to zoom in or out, and a number that shows the magnification of the image.

Go to the **Modes** menu, and select **Digitize landmarks** (if not already checked).

Go to the **Options** menu, and select **Label landmarks** (this will put numbers next to the digitized points). Return to the **Options** menu, and select **Specimen info...** In the window that opens, type a unique specimen identifier in the **ID:** box. Ignore the other boxes for now. Click OK. Again, go to the **Options** menu, and this time select **Image tools**. In the window that opens, select the **Cursors** tab, and choose the digitizing cursor that you prefer. (You can try out the cursor by moving it over the image. Do not click on the image; if you do, a landmark will be recorded and the image tools window will close.) Now select the **Colors** tab. Here, you can choose the color and size of the circle that will be used to indicate a landmark's position, whether the circle will be closed or open, and the color and size of the number used to label the landmark. You can change these options at any time. Close this window, and you are now ready to digitize.

Notice that one of the buttons on the toolbar looks like a digitizing cursor. This button should be depressed, indicating you are in digitizing mode. Also, the cursor will look like the digitizing cross-hairs, and not your standard mouse pointer. Position the cursor over the landmark and click the left mouse button. A circle and number should appear. The # (pound) button on the toolbar is a toggle to hide or reveal the landmark numbers.

2. Editing digitized landmarks

Click the arrow button on the toolbar to enter edit mode, or click the right mouse button. The cursor will change from the cross-hairs to the arrow. To *move* a landmark, place the tip of the arrow on the circle, press and hold the left mouse button, and "drag" the landmark to the new location. (Note that its number does not change.) To *delete* a landmark, place the tip of the arrow on the circle and click the right mouse button. In the pop-up menu, click on **Delete landmark** (or select one of the other options if you change your mind). Note that every landmark with a higher number will be renumbered.

If you click **Insert landmark** instead of **Delete landmark**, a point will be inserted to the left of the selected landmark. The new point will have the number originally assigned to the selected landmark – the selected landmark and all landmarks with higher numbers will be renumbered accordingly. Thus, you can use **Insert landmark** to add skipped landmarks without redigitizing the entire set. Select the landmark that should be after the skipped landmark, right click, choose **Insert landmark**, and drag the new point to the correct location.

3. Saving data

After digitizing the specimen, go to the **File** menu and select **Save data**. A pop-up window will open with a space for you to enter a file name. Notice that the default is to save the file in the same folder as the pictures. Do not change this. Enter the file name with no extension. It will be saved with a .TPS extension. There is no option here, despite appearances.

You are now ready to digitize a second specimen, but before continuing you should take a look at the file you have just created. Minimize the **tpsDig** window, go to Windows Explorer, find the folder containing your pictures, and open the *.TPS file you just created (it will probably open in Notepad or Word).

Notice the format of your file (this is the TPS format). The first line is "LM=" followed by the number of landmarks. The next several lines are *X*, *Y* coordinates, in numbers of pixels. After the coordinates, the next line is "IMAGE=" followed by the name of the image file. The last line is "ID=" followed by the ID you entered (or a default number that is a counter). The IMAGE= line does not include the path to the file, only the name. To see why this is important, close the file and return to **tpsDig**. Select **File**, then **Input source**, then **File**. In the pop-up window, select the file you just created and open it. (If you do not see a list of TPS files, go to the file type box and select TPS files.) **tpsDig** finds and opens the image file and displays the landmarks on the image. If you had saved the data file to a different directory, **tpsDig** would not be able to find the image file and would refuse to open the data file.

4. Adding more specimens to the file

To digitize a second image and add the data to the existing data file, open the file containing the second image and digitize the landmarks, as above. When you are ready to save, go to the **File** menu and select **Save data**. This time, when the pop-up window appears, select the existing file to which you want to add the new data, and then select **Save**. In the new window, select **Append**. The new data will be added to the end of the selected file.

5. Editing data in a TPS file

Open the TPS file in **tpsDig**. Use the arrow buttons on the toolbar to scroll through the images, and edit the digitized landmarks as described above. To save the changes, go to the **File** menu and select **Save data**. This time, when the pop-up window appears, select the existing file to which you want to add the new data, and then select **Save**. In the new window, select **Overwrite**. This will replace the original data with the modified data, and landmarks and specimens that were not edited will not be changed. (If you select **Append**, you will create a file that has both the original data and the edited data.)

References

Bookstein, F. L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.

Bookstein, F. L. (1996). Combining the tools of geometric morphometrics. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 131–152. Plenum Press.

Bookstein, F. L., Chernoff, B., Elder, R. et al. (1985). *Morphometrics in Evolutionary Biology*. The Academy of Natural Sciences of Philadelphia.

Fink, W. L. (1993). Revision of the piranha genus *Pygocentrus* (Teleostei, Characiformes). *Copeia*, **1993**, 665–687.

Oxnard, C. E. (1968). The architecture of the shoulder in some mammals. *Journal of Morphology*, **126**, 249–290.

Roth, V. L. (1993). On three-dimensional morphometrics, and on the identification of landmark points. In *Contributions to Morphometrics* (L. F. Marcus, E. Bello and A. García-Valdecasas, eds), pp. 41–62. Museo Nacional de Ciencias Naturales, Madrid.

Stein, B. R. (1981). Comparative limb myology of two opossums, *Didelphis* and *Chironectes*. *Journal of Morphology*, **169**, 113–140.

Strauss, R. E. and Bookstein, F. L. (1982). The truss – body form reconstructions in morphometrics. *Systematic Zoology*, **31**, 113–135.

Swiderski, D. L. (1993). Morphological evolution of the scapula in tree squirrels, chipmunks, and ground squirrels (Sciuridae): an analysis using thin-plate splines. *Evolution*, **47**, 1854–1873.

Taylor, M. E. (1974). The functional anatomy of the forelimb of some African Viverridae (Carnivora). *Journal of Morphology*, **143**, 307–336.

Zelditch, M. L., Bookstein, F. L. and Lundrigan, B. L. (1992). Ontogeny of integrated skull growth in the cotton rat *Sigmodon fulviventer*. *Evolution*, **46,** 1164–1180.

Zelditch, M. L., Bookstein, F. L. and Lundrigan, B. L. (1993). The ontogenetic complexity of developmental constraints. *Journal of Evolutionary Biology*, **6,** 121–141.

Zelditch, M. L., Sheets, H. D. and Fink, W. L. (2000). Spatiotemporal reorganization of growth rates in the evolution of ontogeny. *Evolution*, **54**, 1363–1371.

Zelditch, M. L., Sheets, H. D. and Fink, W. L. (2003a). The ontogenetic dynamics of shape disparity. *Paleobiology*, **29**, 139–156.

Zelditch, M. L., Lundrigan, B. L., Sheets, H. D. and Garland, J. T. (2003b). Do precocial mammals have a fast developmental rate? A comparison between *Sigmodon fulviventer* and *Mus musculus domesticus*. *Journal of Evolutionary Biology*, **16**, 708–720.

3

Simple size and shape variables: Bookstein shape coordinates

This chapter presents a method for obtaining shape variables that is both simple and visually informative. Called "the two-point registration," this method produces a set of shape coordinates, sometimes called "Bookstein shape coordinates," that can be used both for graphical displays and formal statistical tests. Bookstein shape coordinates (BC) provide a useful introduction to shape analysis because they are intuitively accessible, their formula is relatively straightforward, and understanding them does not require a general understanding of morphometric theory. We present that theory in Chapter 4, after which we can introduce alternatives to BC (Chapter 5).

To introduce the two-point registration, we first review the meaning of shape (the first of several reviews) because this meaning is crucial to the formula. We then focus on the simplest possible application of the method, the analysis of shapes with only three landmarks (triangles). We also discuss how information about size can be restored (because it is removed in the course of the two-point registration). Once we have shape coordinates and a measure of size, we can then test the hypothesis that two samples of shapes differ statistically or that shape change is correlated with size change. These statistical tests are done directly on the coordinates of landmarks – should a statistically significant difference (or covariance) be found, we can then depict it and describe the variable that differs or changes. In this chapter we also discuss the description of shape variables and the biological interpretation of them, because, to a large extent, it is the descriptive power of geometric morphometrics that makes these methods so useful.

Shape and size revisited

In Chapter 1 we discussed the meanings of shape and size, as they are defined in geometric morphometrics. We defined shape in terms of operations that do *not* alter shape – specifically, translation, rotation and rescaling. These operations can be applied to a simple form, a triangle, allowing us to obtain a coordinate system. For the triangle shown in Figure 3.1, we can translate it so that one landmark is at the origin (0, 0) (Figure 3.1A). We can then

Geometric Morphometrics for Biologists
ISBN 0–12–77846–08

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

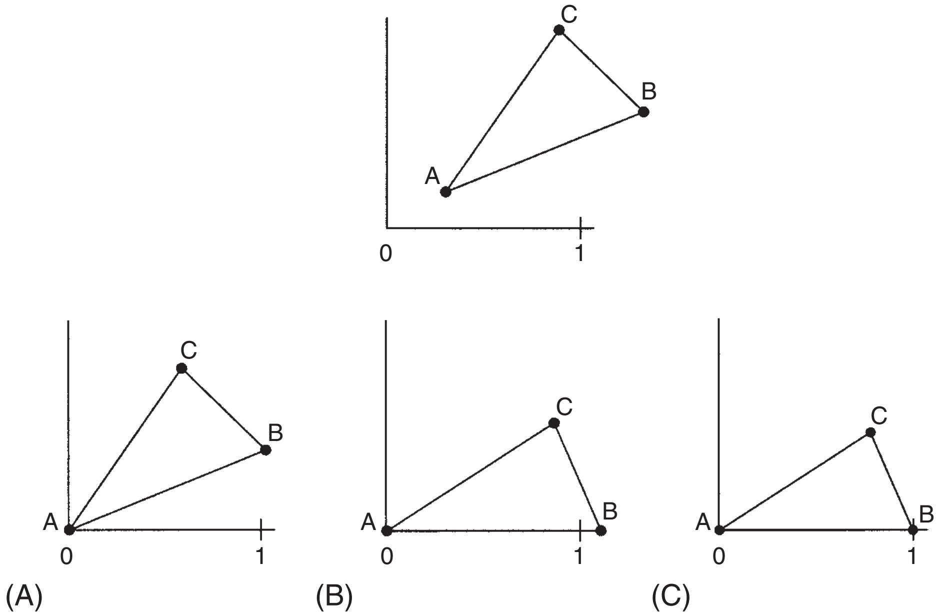


Figure 3.1 Three operations that do not alter shape, applied to a triangle: (A) translation; (B) rotation; (C) rescaling.

rotate it so that the side AB is along the *X*-axis (Figure 3.1B), and finally we can scale it so that the coordinate of landmark B is at point (1, 0) (Figure 3.1C). We can then calculate the coordinate of the third landmark, C, in the coordinate system we have just defined. All of these operations can be applied without worrying about the consequences for shape, because we have defined shape such that none of the operations alter it.

Only those three operations are involved in calculating the coordinates of point C, which is done according to the following formula, in which A_x, A_y, B_x, B_y, C_x, and C_y are the original digitized coordinates, and SC_x, and SC_y are the coordinates of landmark C in the new coordinate system:

$$\begin{aligned} SC_x &= \frac{(B_x - A_x)(C_x - A_x) + (B_y - A_y)(C_y - A_y)}{(B_x - A_x)^2 + (B_y - A_y)^2} \\ SC_y &= \frac{(B_x - A_x)(C_y - A_y) - (B_y - A_y)(C_x - A_x)}{(B_x - A_x)^2 + (B_y - A_y)^2} \end{aligned} \tag{3.1}$$

(The numerators for the two equations really do differ in sign, as well as subscripts; that is not a misprint.)

SC_x and SC_y are the “shape coordinates” of landmark C (which from now on we will simply call C_{xy}). This relatively simple set of operations will be important when we compare the shapes of two triangles.

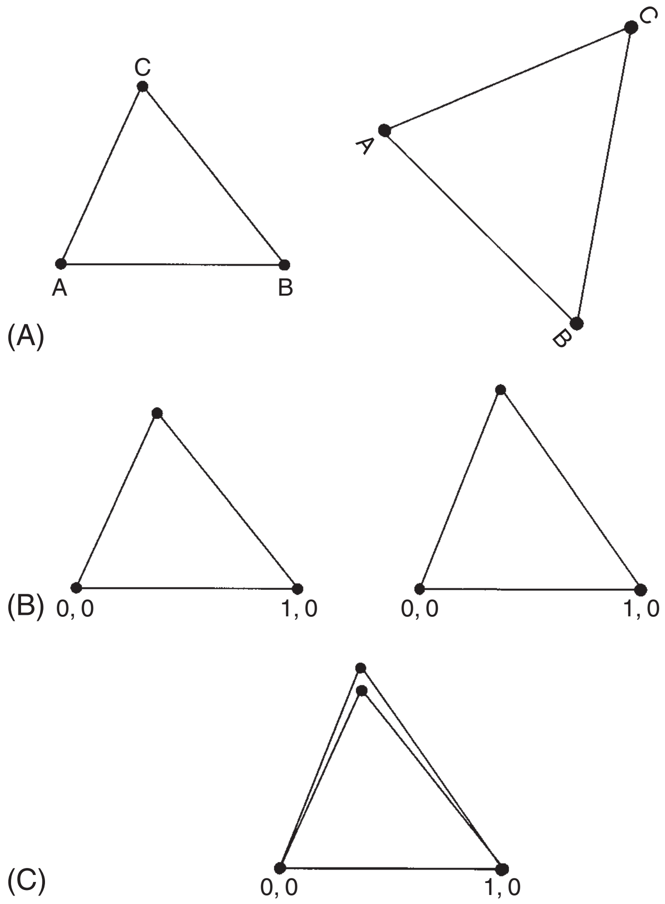


Figure 3.2 Two triangles whose shape difference is the subject of investigation: (A) the two triangles as initially recorded; (B) the same two triangles after being translated, rotated and rescaled by the two-point registration; (C) the same two triangles, superimposed.

Comparing the shapes of two triangles

Our objective in this section is to answer the question: "do the two triangles of Figure 3.2A differ in shape?" To do this, we apply the operations outlined above to both triangles and calculate the shape coordinates of landmark C. That is, we assign the coordinates (0, 0) to landmark A in both triangles, and we assign the coordinates (1, 0) to landmark B in both triangles (Figure 3.2B). As a result, the difference between the two triangles is entirely represented by the difference in the location of the third vertex, landmark C. We can now draw both triangles on the same coordinate system (Figure 3.2C).

While there are programs to do these calculations, they are easily done in any spreadsheet or statistical program that manipulates formulae. As an exercise, take the following three pairs of coordinates for points of a triangle (in the format produced by a common digitizing program), compute the shape coordinates, and draw the triangle. For the moment, pick any two points as the endpoints of the baseline (A and B); we will discuss how to choose them later.

1.	54.00000	306.00000
2.	223.00000	447.00000
3.	632.00000	300.00000

Now take the next three coordinate pairs, and draw that triangle:

1.	11	342
2.	251	520
3.	769	318

Now draw both triangles using the same baseline (with point A and B superimposed), and draw the vector extending between the one free (C_{xy}) landmark on both of the triangles. That vector is the shape variable describing the difference between the triangles.

Comparing many triangles

Of course, we rarely (if ever) compare only two specimens (or triangles). We now consider how to compare many individual triangles (below we discuss comparing forms more complex than triangles). The same procedure (and formulae) still apply, no matter how many triangles or individuals are examined. For example, given a collection of triangles (Figure 3.3A), we assign points A and B the coordinates (0, 0) and (1, 0), and then compare all these triangles (Figure 3.3B) either as whole triangles, or as scatter plots of the one free point (Figure 3.3C).

The scatter-plot is useful for checking the repeatability of your landmarks, as well as for studying the variability of shape or differences in shape. For all these purposes, it is important that the axes of the scatter-plot be sized so that a square shape is shown as a square – that is, the length of the interval from 0 to 1 on the *X*-axis should be the same as the length of the interval from 0 to 1 on the *Y*-axis. Many programs do not do this scaling of axes automatically, so you may have to scale the axes yourself. Often this can be done by first calculating the maximum and minimum values for the *X*- and *Y*-coordinates; the difference between those values, i.e. the range of values should be set equal for both coordinates. For example, if the *X*-coordinate ranges from 0.030 to 0.060 and the *Y*-coordinate ranges from 0.020 to 0.060, both axes should be 0.040 units long (the *Y*-coordinate has the slightly larger range). In this case, the minimum on the *X*-axis could be set to 0.025 and the maximum on the *X*-axis to 0.065. This distributes the extra length equally above and below the observed values, and should enforce a 1 : 1 aspect ratio for the graph.

When the axes are on the same scale, an approximately circular scatter of points indicates that there is an equal amount of variation in all directions. Random digitizing error should be circular; systematic errors, in contrast, will look elliptical. If you have already digitized landmarks, now would be a good time to compute shape coordinates, scale the axes appropriately, and check that your digitizing error is circular. Should you find points that depart substantially from circularity, you should either delete that landmark from your analysis, or take its biased error into account in subsequent analyses.

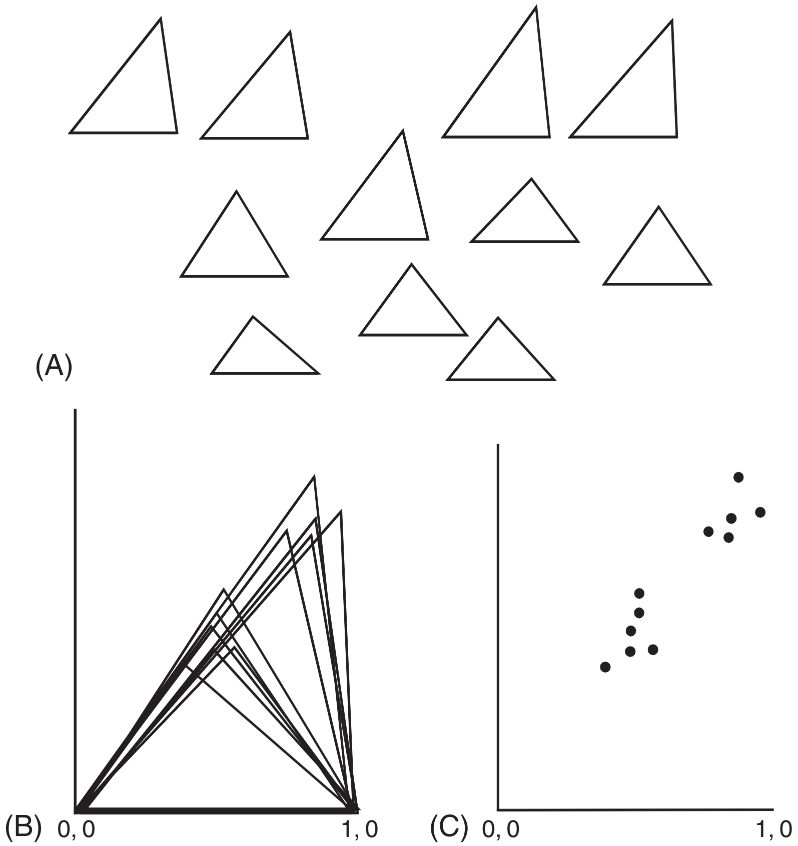


Figure 3.3 Comparing shapes of triangles: (A) the collection of triangles whose shape differences are the subject of investigation; (B) the same collection of triangles, put in a common coordinate system by the two-point registration; (C) scatter-plot depicting the location of the free landmark.

Size

We lost no information about shape when we represented all the triangles by the shape coordinates of point C, but we did remove information about size. Specifically, we removed it by rescaling the baseline to a length of one. We can restore the information about size by using a measure that captures the notion of scale. By scale, we mean the property that changes when an image is enlarged or reduced. There are several possible meanings of size, including a simple measure of the length of an organism along one body axis (e.g. snout–vent length), area, volume, weight or even a linear combination of all measured quantities that captures the positive correlations among them all (as in the case of the first principal component). In geometric morphometrics we use a specific concept of size, one related to geometric scale. One feature of this notion of size is that it is independent of shape. This is not the case for the other possible size variables mentioned above, which may be independent of shape but are not necessarily so in all cases. To see this idea of independence, consider what happens in the case of isometric growth: every dimension

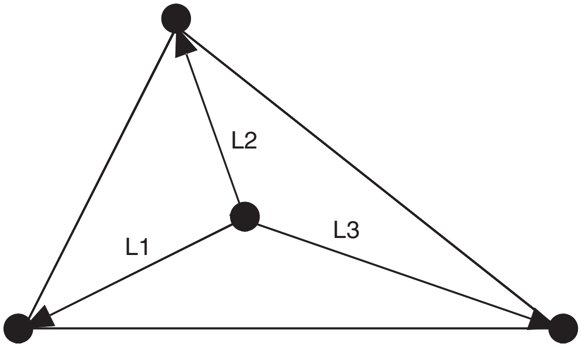


Figure 3.4 A geometric depiction of the calculation of centroid size, which equals the square root of the summed squared lengths of line segments L1, L2, L3.

is enlarged by the same proportion; each coordinate is moved away from the center in proportion to its original distance from the center. The size variable that captures this radial notion of scale is centroid size, graphically illustrated in Figure 3.4.

To compute centroid size, first compute the centroid (center) of the form, which is located at the mean position of all coordinates. This mean is found by simply averaging the X-coordinates and the Y-coordinates. For example, the three landmarks of the triangle might be at $(0, 0)$, $(1, 0)$ and $(0.3, 0.8)$. The average X-coordinate is thus the arithmetic mean of the three X-coordinates $(0, 0.3$, and $1)$, which is 0.433, and the average Y-coordinate is 0.267. Then the squared distance of each of the three landmarks from the center is calculated using the standard formula for a squared distance between two points $(X_2 - X_1)^2 + (Y_2 - Y_1)^2$. This sum gives a measure of size related to area; taking the square root of the sum gives a linearized measure of size. The square root of the summed squared distances of each landmark from the center of the form is centroid size.

Size is thus measured separately from shape, and it is statistically uncorrelated with shape so long as shape changes isometrically (which, by definition, means that shape does not change with size). This is a useful attribute of a size measure, because we do not want size to be intrinsically correlated with shape simply by virtue of its formula. Rather, we want a measure of size that is correlated with shape only when size and shape change together. Of course real data will often show this correlation between shape and scale, because allometry is a common phenomenon. However, allometry is an empirical finding, not an effect of the formula for size. Centroid size is the only size variable that is uncorrelated with shape *in the absence of allometry* (others that are also uncorrelated with shape are variants on centroid size). This independence from shape is one of the main reasons why centroid size is used as a size variable. The other reason is that centroid size has a crucial role in defining the metric for a distance between two shapes (Chapter 4).

Choosing the baseline

When we calculated shape coordinates, we chose one side of the triangle to serve as a baseline. An obvious question is whether our results might depend on that choice. As Bookstein (1991) has proven, the scatters for different sets of shape coordinates of the same triangle to different baselines differ mainly by translation, rotation and rescaling. In effect, all the statistical results are (approximately) the same regardless of the choice of baseline. However, this does not mean the baseline should be chosen arbitrarily. First, some landmarks are difficult to digitize and may be especially difficult to locate – these should not serve as an endpoint of the baseline. This is because the method, in effect, transfers all the variance in the baseline points to all the other landmarks, so if the endpoints of the baseline are highly variable then all the points will be noisy. More problematically, the variance is not evenly distributed across all landmarks; the transfer of variance might therefore introduce a bias into the data. Another consideration that enters into choosing a baseline is its orientation. If the baseline rotates relative to a body axis it does not compromise the statistical analyses, but it can make interpretations based on graphics difficult – it might seem that all the landmarks are moving away from the baseline in the posterodorsal direction, for example, when the baseline rotates in the anteroventral direction. Also, in choosing the endpoints of the baseline, we do not want points that are too close to each other because any highly localized variation in shape may be common to both those points. Just as the noise of the baseline landmarks is transferred to all the others, the variance local to the baseline landmarks is transferred to all the other landmarks. Ideally, therefore, we want endpoints of the baseline to be along the longest diameter of the form that passes through the centroid of the form, so long as those points are not especially unreliable and the longest diameter does not rotate.

It is easiest to interpret results when the baseline lies along an organismal body axis. Even though results can be interpreted in a baseline-invariant way, the interpretations still refer to sides of the triangle. It is most convenient when at least one side is a conventional and familiar reference. Bookstein has put a great deal of emphasis on baseline-invariant interpretations out of a concern for reports free from arbitrary, abiological decisions. However, organismal body axes are neither arbitrary nor abiological – indeed, we often want to make explicit references to organismal body axes in our interpretations. Thus, even though we *can* interpret shape changes without reference to organismal body axes, we might still wish to orient our findings with respect to them. This motivates choosing a baseline along one of those axes.

Statistics of shape coordinates

Once we have shape coordinates, we can answer the basic "existential" questions as defined in Chapter 1, such as "do these samples differ in shape?" These questions have "yes" and "no" answers supplied by statistical tests. All conventional statistical methods and tests can be applied to shape coordinates and centroid size. For example, an average value for the shape coordinate at point C is computed by averaging the X-coordinates for that point across all individuals within a sample, then dividing that sum by the total number of individuals in that sample; the same procedure is then applied to the Y-coordinates. Variances and standard deviations are also calculated by standard formulae. Because the two endpoints of the baseline are fixed, they have no variance and should not be included in statistical analyses. If you use conventional statistical packages to analyze these coordinates, remember to exclude them from the analysis because many programs will not run if the variables do not vary.

Because every landmark has two dimensions (its *X*-, and *Y*-coordinates), statistical analyses are necessarily multivariate. Even if we are asking whether two samples of triangles differ in average shape, we must use a multivariate test. In particular, we would use the multivariate form of the familiar Student's *t*-test, Hotelling's T^2 test (see, for example, Morrison, 1990). When comparing two samples of triangles, the test is applied to the two coordinates of landmark C. When we are comparing more than two samples, we can use Wilks' Λ (Rao, 1973) or one of the related statistics obtained by a multivariate analysis of variance (MANOVA). In studies of allometry, we use multivariate regression.

To apply any of these statistical tests to the data, it is first necessary to decide the appropriate null hypothesis. In many cases, the null hypothesis is that the differences in shape between two or more samples are due solely to chance (the vagaries of sampling). To test this hypothesis, the shape coordinates (for free landmarks only, not for baseline points) are compared by Hotelling's T^2 test (in the two-group case) or by MANOVA (in the multigroup case). This can be done in any statistical package. If, for example, the two samples being compared are two sexes, "sex" is the categorical variable, the factor whose effect is being tested. If the difference is statistically significant, that is evidence of sexual dimorphism. Dimorphism in size can also be tested, which involves a univariate test because size is a one-dimensional variable. To test the hypothesis that males and females differ in shape because they differ in size, and solely for that reason, MANCOVA (multivariate analysis of covariance) is used.

Studies of allometry are equally straightforward. The null hypothesis is that there is no covariance between size and shape beyond that due to sampling effects, and rejection of the null hypothesis means there is a correlation between size and shape – allometry. Again, this test can be done using any conventional statistical package; the shape coordinates of the free landmark(s) comprise the dependent variable (we will refer to it in the singular, as the dependent "variable", even though it has multiple components). Size is the independent variable, and the effect to be tested is that of size on shape.

Describing shape differences

Having documented that shapes *do* differ, or *do* covary with a measured factor, the next step is to describe that difference or covariance. A description comes before any interpretation, because interpretations offer an explanation and we need to know what the effect is before we can explain it. For example, if we want to interpret the impact of size on shape, we first need to know how size affects shape. We can then interpret that effect in light of growth processes or biomechanics. If we detect allometry statistically, and describe the shape variable that covaries with size effectively, we can then seek explanations in terms of growth and biomechanics.

Given a comparison between two triangles, we first find the vector linking landmarks C to C′; that vector has two components, its *X*- and *Y*-coordinates. In Figure 3.5A, the triangle is drawn between the tip of the snout (landmark A), the posterior end of the hypural bones (landmark B) and the free landmark (C) at the anterior dorsal fin base. The difference between the two shapes is entirely along the *Y*-direction of landmark C – the little vector extending between C and C′ points directly upwards (Figure 3.5B). We can describe it as a vertical (dorsad) displacement of the anterior dorsal fin base relative to the baseline.

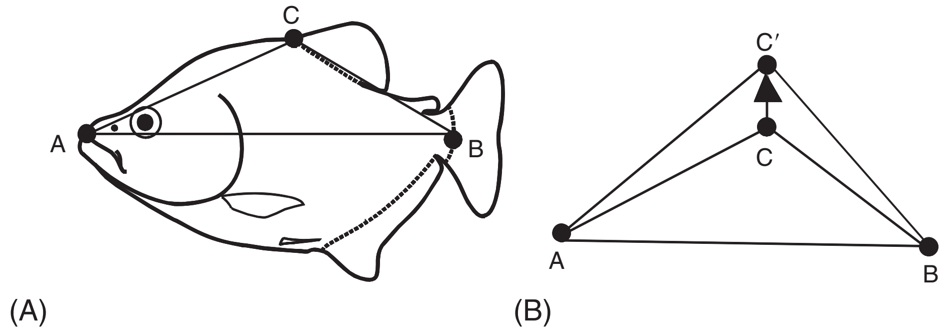


Figure 3.5 (A) A triangle with the baseline along the anteroposterior body axis and the free point at the anterior dorsal fin base; (B) the difference between two shapes depicted by a vector extending between C and C′.

However, framed in those terms, we have not described a change in shape of a triangle – we have not described a change in proportions or angles. Even though a ratio is implicit in our description of a dorsad displacement of the landmark, we need to go further and actually translate the displacement of a landmark (relative to the baseline) into a shape variable.

Figure 3.6 shows three shape changes that we will use to describe such a translation. In Figure 3.6A, the change in location of point C is depicted by a vector in the vertical direction; in Figure 3.6B, the change in location of point C is depicted by a vector in the horizontal direction; and in Figure 3.6C the change in location of point C is depicted by a vector that lies along one of the sides, AC. Our objective is to relate these changes to changes in familiar ratios or angles of a triangle. For Figure 3.6A, the change is in the height of the triangle relative to the length of its baseline, and thus, we can name the shape variable as change in the ratio height : base. We can term that a change in the "aspect ratio of the triangle." We can describe the change implied by the vector in Figure 3.6B in terms of a ratio of two segments of the baseline: the original position of C is projected onto the baseline at point X, and the shape change is an increase in length of the line from A to X ($\overline{AX}$) relative to the length from A to B ($\overline{AB}$). In Figure 3.6C, the shape variable implied by the vector is a modification in the ratio $\overline{AC} : \overline{AB}$.

For each of these changes we can also describe what does *not* change – we can describe the invariant as well as the covariant variables. In the first two cases, the changes are along the axes of the shape coordinates; the change is entirely in one direction, implying no change in the other. That is, a change oriented entirely along the vertical direction (height or aspect ratio) implies no change in the horizontal direction (i.e. no change in $\overline{AX}$: $\overline{AB}$). When the change is entirely directed along side AC, the unchanged feature is the angle at A. It is important to recognize that every change implies an invariant; it is therefore inappropriate to invoke constraints merely because something does not change. *Every* change has an invariant aspect because we can *always* draw a vector at right angles to one depicting the direction of change.

The shape variables we have just named all depend on the baseline and on the particular points we digitized. Had we chosen a different baseline, we would have obtained a different vector and a different descriptor. While the scatters and statistics are unaffected by that

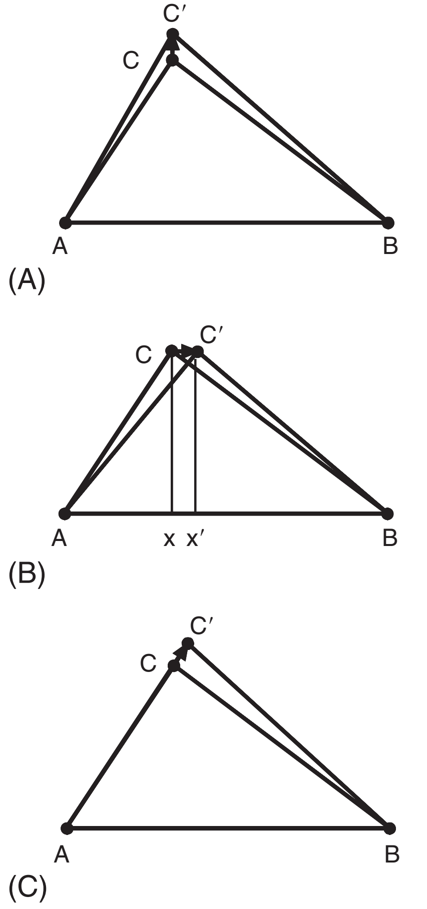


Figure 3.6 Three changes in the shape of a triangle: (A) increasing the ratio of the height to the base; (B) increasing the length of Ax relative to AB; (C) increasing the length of AC relative to AB.

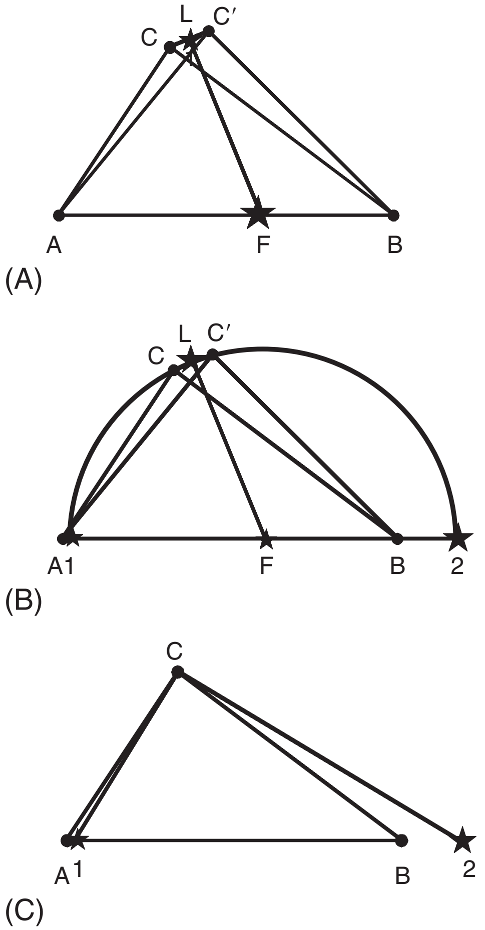


Figure 3.7 The circle construction. (A) The perpendicular bisector of the line segment extending from C to C′ is determined and extended to the baseline (the point of intersection between them is F). (B) A circle is drawn through points C and C′, with its center on F; the circle intersects the baseline (or extensions from it) at two points 1 and 2. (C) Lines C1 and C2 can now be drawn on the triangle ABC. The ratio of 1C/1C′ gives the principal strain in that direction, while the ratio 2C/2C′ gives the strain in the other.

choice, the vectors and verbal interpretations depend on it. We now present a method that yields descriptors that are invariant under changes in baseline, which is especially useful when the baseline is biologically arbitrary. Even when the baseline is not biologically arbitrary, we might still want a description of change in general terms – ones that do not presuppose a fixed side. To that end, we introduce the construction of principal axes, an algorithm for finding a pair of directions (at right angles to each other) – one is the direction of greatest change and the other is the direction of least change.

Principal axes

Principal axes describe change by symmetric tensors, in more specific terms, by relative metric or strain tensors. Vectors like the ones we drew between points C and C′ above, have directions that rotate with the coordinate system. In contrast, the principal axes are invariant under changes in the coordinate system; these axes are perpendicular both before and after transformations (they have also been called biorthogonal directions). In addition to these axes, we will also compute the principal strains, measures of the change in length of each principal axis. Then we can describe the difference between forms by the ratio of the two strains, a metric called anisotropy. We will show how to construct the principal axes by hand, and give the formulae for calculating the strains and anisotropy. Finally, we will discuss naming the shape variables implied by the principal axes.

The circle construction for the principal axes

By assumption, each little piece of the triangle (small portions within the regions between the landmarks) corresponds from triangle to triangle. Also by assumption, the change from one triangle to another is entirely uniform. We can then find principal axes by the following algorithm, called the circle construction. The construction involves four pairs of shape coordinates, those of points A and B (which are the same for both forms) and those of points C and C′ (corresponding to the two locations of C). We first determine the perpendicular bisector of the line segment extending from C to C′, and extend that line to the baseline (Figure 3.7A). To do this, draw the line segment between C and C′ (L), find its midpoint, and draw a line perpendicular to L that extends from that midpoint to the baseline. That point of intersection is called F. Next, draw a circle through points C and C′, with its center on F (Figure 3.7B). The circle intersects the baseline (or extensions from it) at two additional points labeled 1 and 2. These points, like A and B, are unmoved by the shape transformation because we are operating under the assumption that the change is entirely uniform. The angles 1C2 and 1C′2 are both right angles, and hence we have identified the biorthogonal directions – those that are perpendicular both before and after the transformation.

The lines C1 and C2 can now be drawn on the triangle ABC (Figure 3.7C). In some cases, the line segments C1 and C2 lie outside the triangle. These can be placed within it by drawing lines parallel to C1 and C2 that pass through a vertex of the triangle. In one case, the circle cannot be drawn at all: when C′ is displaced purely in the vertical direction (in that case the perpendicular bisector of the line segment CC′ is parallel to the baseline, thus it does not intersect it). For that case, one principal axis is in the direction of the vector connecting C to C′, and the other axis parallels the baseline. When the change in shape is slight – so slight that it is difficult to find a midpoint along the line from C to C′ – the midpoint can be approximated by point C and the perpendicular bisector of line CC′ can be approximated by the perpendicular to the line CC′ at point C (Figure 3.8). The intersection of this line with the baseline estimates point F, the center of the circle. After F has been located, the construction can be completed as described above.

Not only are principal axes directions that are perpendicular before and after the shape change, they are also directions that undergo the most extreme changes in length during that shape change. Although it is common to think of one direction as elongating and the

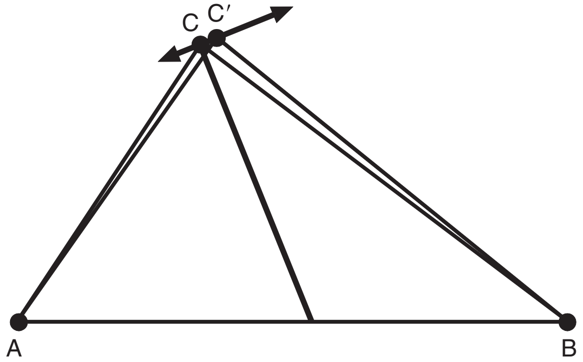


Figure 3.8 Changing shape by such a slight degree that it is difficult to find a midpoint along the line from C to C′.

other as shortening, it actually may be the case that that one is the direction of fastest growth and the other is the direction of slowest growth (or even no growth). In physics, changes in length are called strains, so the changes in length along the principal axes are the principal strains. To calculate the principal strains we need the lengths of the line segments C1, C2, C′1 and C′2. We also need the difference in absolute length of the side AB for the two triangles, i.e. the length of that side before rescaling. The ratio of 1C : 1C′ gives the principal strain in that direction (its relative elongation), while the ratio 2C : 2C′ gives the strain in the other. Because the two baselines might differ in absolute length, those ratios must be adjusted to the proper scale by multiplying each ratio (1C : 1C′) and (2C : 2C′) by the ratio of the unscaled lengths of the two baselines.

The ratio between the two principal strains is called the anisotropy of the shape change. This is a measure of the degree to which the transformation is unequal along the two axes. For the case of a slight change in shape – the case for which we approximated the principal axes (above) – we can approximate the anisotropy as $1 + (d/h)$, where d is the length of the vector from C to C′ and h is the height of C above the baseline (the height of the point above the baseline is the Y-shape coordinate of the landmark). In addition to being directions that undergo the most extreme change relative to their original lengths, principal axes also are the directions that undergo the greatest change relative to each other. In other directions, the strain is intermediate between the principal strains. At 45° to the principal axes (bisecting the angle between them) are two directions that undergo identical strains; these are directions of isotropic change – that is, no relative elongation or shortening.

Naming the variables implied by the principal axes

Earlier we discussed naming shape variables based on vectors that represent a change in the location of shape coordinates. We now turn to naming shape variables based on the orientation of the principal axes. When principal axes are aligned with a side of the triangle, or with one of its angles, the shape change can be described as simple changes of a ratio or angle. At the same time, the feature of the triangle that is invariant can be described just as easily. Also, when the bisectors of the principal axes (the directions of no relative change) are aligned with a side of the triangle or with the bisector of one of the angles, the

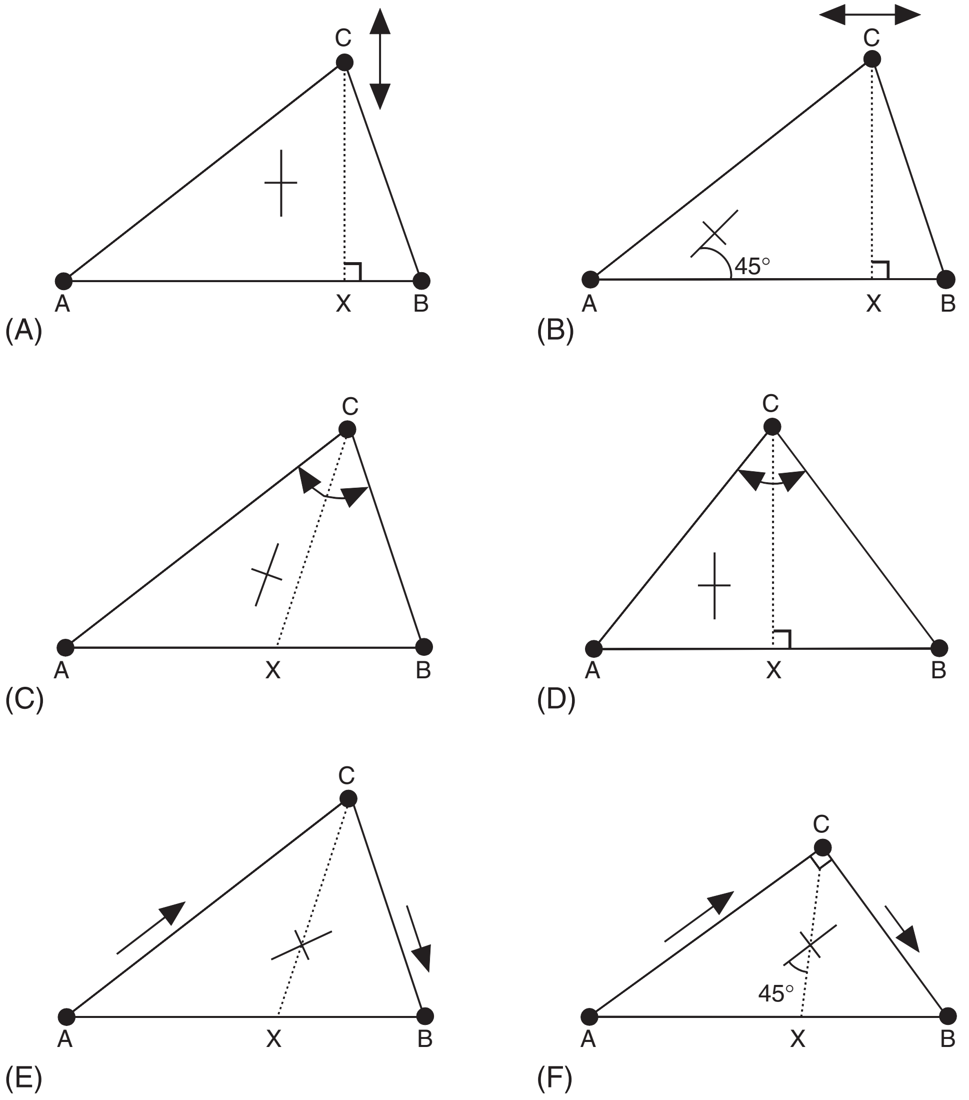


Figure 3.9 Naming shape variables for principal axes relative to sides and angle bisectors of a triangle.

shape changes and the invariant features can be described just as easily. Because there are three sides and three angles in a triangle, and two ways to align the axes to a side or angle, there are twelve possible shape changes that can be expressed in familiar terms. In some special cases (e.g. a right triangle) there are fewer possibilities because some alignments are identical (e.g. 1 and 4 when the angle is the right angle and the side is adjacent to that angle).

When a principal axis is aligned with a side of the triangle, the variable expressing what changes most is the ratio of the length of that side relative to the distance of the third point from that side. For example, in Figure 3.9A we show a case in which one axis is aligned with the baseline AB and the other axis parallels XC, the line through C perpendicular to

AB. (For each triangle there is one pair of these axes, which can be positioned anywhere in the triangle; that pair of axes describes the change of the entire triangle, so we would find the same pair of principal axes anywhere we look in that triangle.) The shape variable implied by this pair of axes can be described as a change in the height of the triangle (at C) relative to the length of side AB, which is equivalent to saying that C is moved toward or away from line AB. The line having arrowheads at both ends indicates the displacement of point C in the vertical direction. The invariant feature is the position of XC relative to AB – in other words, the relative lengths of segments AX and XB. There are several ways we might describe this change, but the choice of description should make biological sense. Sometimes it may be more appropriate to speak of a structure displaced along a line, at other times it may make more sense to speak of a change in the distance between a point and a line relative to the length of a line, and sometimes it might make most sense to speak of a change in aspect ratio of a triangle. It is important here that in talking about principal axes we are not really talking about a change *at* point C, rather, we are concerned with the location of C relative to AB.

When a principal axis is at 45° to a side (i.e. when a bisector of the principal axes is aligned with a side), we can say that the shape change is displacement of the third point parallel to that side. In Figure 3.9B, point C is displaced horizontally, parallel to side AB (rather than vertically as in Figure 3.9A). We again show the direction of displacement by the line with the arrowheads at both ends. Equivalently, we could describe this as a change in the ratio of AX to XB, or as a shearing of the triangle. As may be obvious from the contrast between this shape variable and the one described in the previous paragraph, the feature that changes most in this one is the feature that did not change in the previous one (the ratio of lengths AX and XB).

When one of the principal axes is aligned with the bisector of an angle of the triangle, the feature that changes most is that angle. The invariant feature of the triangle is the ratio of the lengths of the sides adjacent to that angle. In Figure 3.9C, the angle being changed is at point C (shown by the arc with the double arrowhead), which is either opening or closing; the invariant feature is the ratio of lengths of AC and BC. We can describe this shape change in terms of the altered angle at C, or as the displacement of point C along the axis of greatest strain (from X to C) if that axis corresponds to an anatomically or functionally meaningful direction. In the special case when the lengths of AC and BC are equal (Figure 3.9D), the bisector of the angle makes a right angle to side AB so the direction of greatest change is in the direction of the height XC and the direction of least strain is parallel to side AB (as in Figure 3.9A).

Finally, the principal axes can be oriented at 45° to the bisector of an angle of the triangle (Figure 3.9E). If so, the feature which most changes is the ratio of AC to BC, and the invariant feature is the angle C. We can speak of this shape change in terms of the contrasting displacements of landmarks A and B relative to C (i.e. A moves towards C while B moves away, or A moves away from C while B moves towards it). We can also speak of line AB rotating relative to AC and BC. In the special case of a 90° angle at C (Figure 3.9F), the bisector of that angle is at 45° to sides AC and BC, so orienting the principle axes at 45° to the bisector orients the axes parallel to sides AC and BC. This can be interpreted as the displacement of B perpendicular to AC (similar to the first case in Figure 3.9A, but with AC as the baseline and BC as the height of the triangle at B). The invariant feature is the position of B parallel to AC.

The descriptions above can usually be applied when the orientation of the principal axes is close to one of the specified alignments; an exact match is not required. However, sometimes the alignment is not particularly close to one of these convenient special cases and it is difficult to determine which exemplar most closely matches the empirical results. It may then be difficult to select familiar words that convey the results most accurately. Fortunately, the graphics convey the information. Words are useful to summarize the information and to communicate with readers unfamiliar with the graphics. Words are particularly useful when several competing hypotheses predict different directions of shape change and the hypotheses are phrased solely in words. Then, the verbal description of shape change provides a bridge between the hypotheses and the graphical displays of expected results (and the appropriate statistical tests).

Multiple triangles

So far we have concentrated on the simplest possible case: comparisons of a triangle. This is because most of the principles introduced by that simple case extend directly to more complex cases (although some do not). Before introducing the complexities introduced by analyses of more than three landmarks, we first discuss the general principles that do extend unproblematically to the more complex case. We then detail those that do not, setting the stage for the analytic methods that will be introduced later (particularly those in Chapter 6).

Multiple landmarks can all be transformed into shape coordinates using the formulae introduced for computing the shape coordinates of a single moveable point, C. We just apply that same formula to all the additional points. It is not necessary to use the same baseline for all points, but it does ease the task of reporting the changes. Not only is the same formula applicable to the more complex case, but the same basic statistical machinery also applies, with one caveat: the statistical test of a shape difference or covariance cannot be applied to all landmarks simultaneously unless the sample size is minimally twice the number of free landmarks. When sample sizes are smaller than this, the number of variables exceeds the number of observations. This is obviously not enough observations, and in fact we may need four times as many observations as landmarks for an adequate analysis. When sample sizes are (too) small, it may still be possible to test the null hypothesis statistically by applying Hotelling's T^2 or MANOVA to individual landmarks (i.e. to both X- and Y-coordinates of each landmark), then adjusting the p-values according to the number of tests (using whatever approach is preferred for *post hoc* tests). The hypothesis we are testing is that the configuration of landmarks differs; the null hypothesis that we wish to reject at some level of significance, e.g. 0.05, should be tested at that level. If we test landmarks separately, we risk rejecting a true null hypothesis 5% of the time and each test counts as one time – so with multiple landmarks the risk of rejecting a true null hypothesis is actually far more than 5%. One approach to ensuring a table-wide error rate of 0.05 is the Bonferroni approach to multiple comparisons; using it, we divide 0.05 (the α level) by the number of tests, e.g. 10, and reject the null hypothesis at a table-wide level of $\alpha = 0.005$, which is 0.05/10. So long as one variable allows us to reject the null hypothesis at the table-wide level of 0.05, we can reject the null hypothesis for shape.

Another procedure extends unproblematically from one to many triangles – the depiction of shape differences by vectors at the free landmarks (Figure 3.10). As in the case of

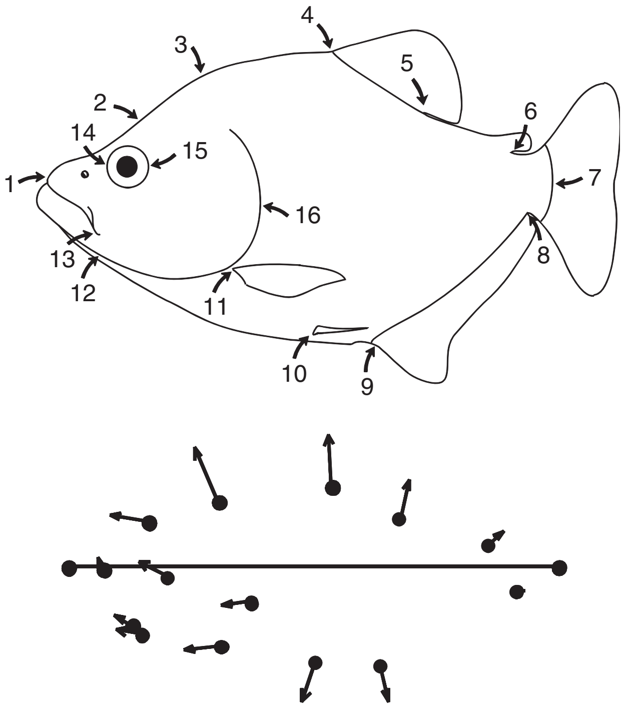


Figure 3.10 Ontogenetic changes in the shape of a piranha, *Serrasalmus gouldingi,* represented by vectors depicting the change in location of Bookstein shape coordinates from their position in a young juvenile.

a single triangle, that depiction depends on the baseline. If this baseline dependence is not seen as a serious problem, the description can proceed in terms of the displacements of landmarks relative to each other, relative to the baseline. For example, in describing the ontogenetic change in shape depicted in Figure 3.10, we would need to take the relative lengths of all the vectors into account. The most anterior free point on the dorsal margin (landmark 2, at the epiphyseal bar) is displaced anteriorly, indicating that the region between it and the baseline point at the tip of the snout is shortened relative to the length of the baseline. The point immediately posterior to landmark 2 (landmark 3, at the tip of the supraoccipital process) is also displaced anteriorly, although most of its displacement is along the dorsoventral body axis. Because the anteroposterior component of this vector is short relative to that of the more anterior point, the region between the epiphyseal bar and supraoccipital process is relatively elongated (relative both to the length of baseline and to the more anterior region just described). Such descriptions can be useful, even if they depend on the baseline. We can also describe and depict two ontogenies relative to the same baseline, either by a comparison between vectors at each point (Figure 3.11A) or by highlighting the implied changes in body profile (Figure 3.11B).

Although all these procedures extend to multiple triangles, we need to consider the special case in which those triangles describe two sides of a bilaterally symmetric organism. If we are interested specifically in their asymmetry, both sides contain relevant information (or, more exactly, the information lies in the difference between the sides). Otherwise, the

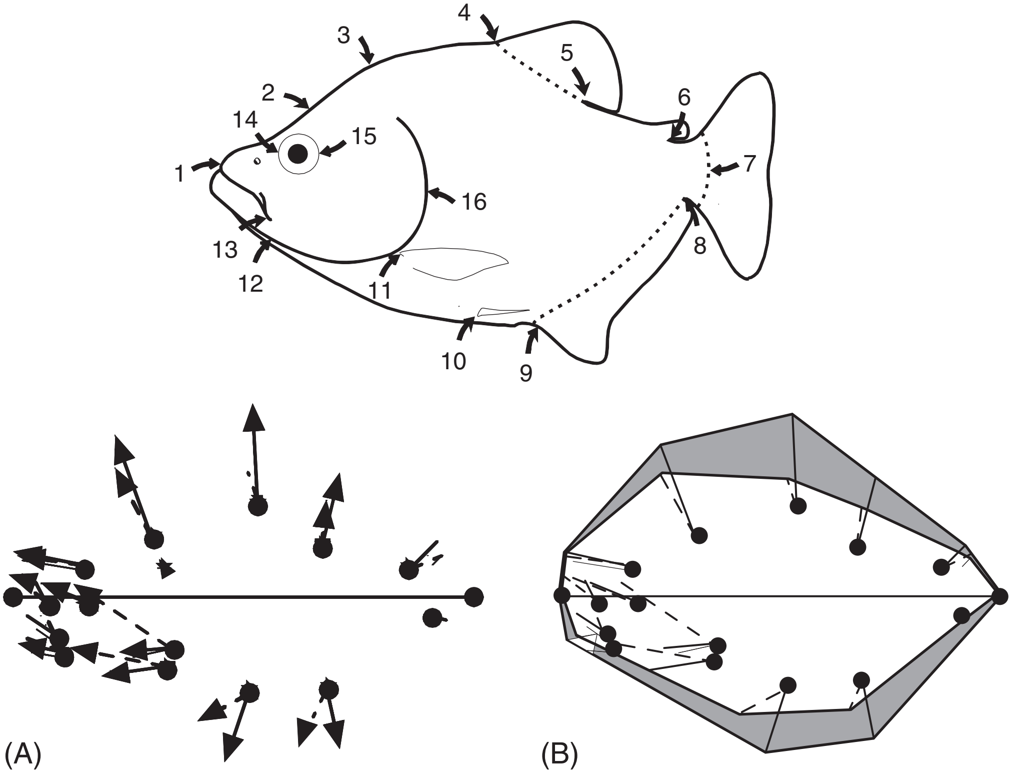


Figure 3.11 A comparison between the ontogeny of *S. gouldingi* and *Serrasalmus elongatus*: (A) depicted by vectors at each landmark, those representing the ontogenetic change in the relative locations of landmarks of *S. gouldingi* are shown as solid lines; those of *S. elongatus* are shown as dashed lines; (B) the implied changes in body profile; the dark shaded regions represent the areas that increase in *S. gouldingi* relative to those same regions in *S. elongatus*.

two sides are redundant; we would not wish to treat them as independent of each other. In effect, unless asymmetry is the topic of interest, we have measured the same shape twice. Doing so creates serious problems for statistical analyses because our degrees of freedom will be inflated (and we will also need far larger sample sizes to analyze the data, as well as many more intact specimens). So, the standard approach to bilaterally symmetric forms is to reflect one side across the midline, averaging the coordinates of the two sides. That approach provides the correct degrees of freedom for statistical analysis, reduces the number of specimens required for testing statistical hypotheses, and allows us to use partially fragmentary specimens with landmarks present on only one side or the other.

Having obtained useful descriptions, both verbal and graphical, we are still left with one serious and unsolved problem; that of describing changes in regions *between* landmarks. This is the topic of later chapters (particularly Chapter 6), but we emphasize it here because it has major implications for descriptions based on shape coordinates. For example, in Figure 3.10 we can see that the point at the anterior dorsal fin base (landmark 4) is displaced vertically, indicating a deepening of the body relative to its length. The same deepening

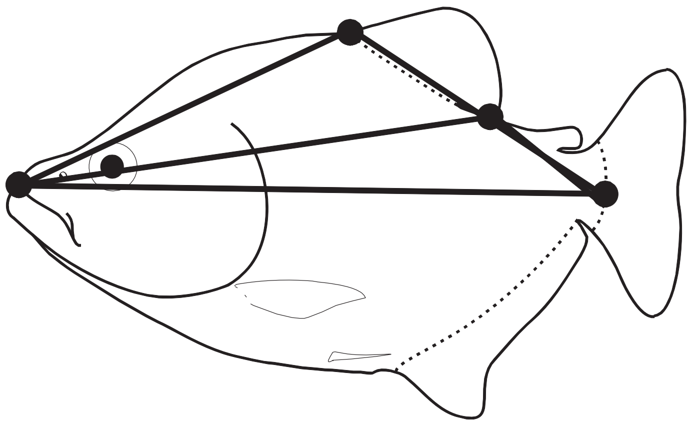


Figure 3.12 Ontogenetic changes in two overlapping triangles. For both triangles the endpoints of the baseline are the tip of the snout and the posterior termination of the hypural bones; the moveable landmarks are the anterior and posterior dorsal fin bases. Owing to their overlap, their changes are unlikely to be independent.

is suggested by the vertical displacement at the posterior dorsal fin base. In this case the "body" being deepened is contained within two overlapping triangles (Figure 3.12), so it would not make sense to describe the changes in these two triangles separately, such as by calculating the principal axes of each. Nor can we fully integrate these two triangles into one single descriptor until we can determine whether the same descriptor applies to both triangles. In effect, we can easily talk about changes at individual landmarks, but we cannot easily talk about regions between the points.

We accomplished the description of changes within regions for individual triangles using principal axes. However, the principal axes are particular to specific triangles and may vary from one triangle to another, even if the triangles are adjacent or partially overlapping. We may get different principal axes if we divide up the shape into different arrangements of triangles, which means that our results can depend on an arbitrary choice – which triangles we draw for a given set of landmarks. The fundamental problem here is that we cannot treat each triangle separately; instead we need to think about all the displacements of all the landmarks relative to all other landmarks, and the analysis must take into account the spatial distributions of these points (i.e. which ones are adjacent, which are far apart, which are anterior to or ventral to each other, etc.). In sum, we need a method that analyzes relative displacements of landmarks in context of their geometry, when we have no *a priori* reason to divide an organism into biologically independent triangles. One such method (and the only one known to us) is introduced in Chapter 6.

Before we go on to talk about baseline-invariant descriptions of change between landmarks, we need to consider other shape coordinates. Bookstein's shape coordinates are especially simple and transparent, and are well suited to conventional statistical analyses. However, they have some serious problems in context of the general statistical theory of shape. Before those reasons can be understood fully, we need to present the general theory (Chapter 4), and then we can discuss alternative methods for obtaining landmark coordinates (Chapter 5).

Software

The program for calculating shape coordinates is **CoordGen** (for “Coordinate Generator”). It produces a variety of shape coordinates, including Bookstein’s shape coordinates (BC), and others discussed in Chapter 5. The program also displays the data graphically (although others produce higher quality graphics with a greater variety of displays because they are designed to depict results of morphometric studies, not just the input data files). In addition to calculating shape coordinates for individual specimens, **CoordGen** also calculates the mean for each sample, as well as means for a specified number of the largest or smallest specimens, which can be saved to a file. A second program is available for bilaterally symmetric specimens – this program, **BigFix**, reflects one side over the midline and averages the coordinates found on both sides. This program accommodates missing data, so long as the data are missing for only one side. A third program, **TwoGroup**, conducts a Hotelling’s T^2-test using Bookstein’s shape coordinates.

Running CoordGen

To use **CoordGen** it is necessary to understand the various file formats that the program can read. Before loading the file you must first specify that format. One format read by **CoordGen** is that produced by the digitizer program **tpsDig**. A second is the format produced by the NIH digitizer (and probably many others). A third is the format produced by **CoordGen** itself, so you can read in files produced by **CoordGen** to transform one type of coordinates into another (e.g. the coordinates obtained by the two-point registration can be transformed into those obtained by a Procrustes superimposition – introduced in Chapter 5). As well as reading multiple formats, **CoordGen** can read files that include a ruler, a scaling factor, or neither.

Data files in TPS format are produced by **tpsDig**, and are used by Rohlf’s software (as well as by other software written to be consistent with his). TPS format was described in the last chapter; **CoordGen** can read TPS files that include a ruler and those that do not include either a ruler or scaling factor. If the file lacks both a ruler and scaling factor, the images must be scaled properly while digitizing if centroid size is to be estimated correctly (this scaling must be done by the digitizer program itself). When selecting one of the two TPS options, the critical factor is that the first line begins with “LM=K” (where K is the number of landmarks, including the two for the ruler if the ruler is in the file). Any information on this line after the number of landmarks is ignored. If you use a digitizing program that outputs two columns of coordinates for each specimen, but does not put “LM=K” in the first line, you will need to enter that keyword and number of landmarks manually.

If you use the NIH digitizer, or another program that produces data for each specimen in a single row, select one of the next two options: X1Y1, with or without a ruler. The final option is the IMP format, X1Y1...CS. The difference between the IMP format and other X1Y1 formats is that the final column of the IMP format is centroid size (CS), whereas it is the *Y*-coordinate of the last digitized point in the other formats.

After selecting the appropriate format, enter the file name that you wish to analyze. If you selected a format that includes a ruler, the next thing you need to do is give **CoordGen** the endpoints of the ruler and the length of the ruler. The default endpoints are the last two points in your file; if those are not your ruler points, type in the correct numbers – for example, if your ruler endpoints are the first two landmarks in your file, type in the "1" and "2". **CoordGen** also assumes that the ruler length is 10 (which could be 10 mm, 10 cm, 10 inches, etc.); again, you can change the default by typing in the correct number in the box. If, for instance, your ruler is 20 mm long, type "20" in the box for ruler length. When the ruler endpoints and length are correctly specified, you can **Carry Out Rescaling** (by clicking on that box). If you do not have a ruler or scaling factor in your file, the data are assumed to be properly scaled already.

When you click **Carry Out Rescaling**, or when you load a file that has no ruler, **CoordGen** will calculate Bookstein coordinates using the first two landmarks in the data file. (If the ruler is the first two points in the file, **CoordGen** will use the next two points as the baseline.) To calculate Bookstein coordinates to a different baseline, enter the numbers of those landmarks in the boxes under **Baseline**. The first landmark (left box) will be assigned coordinates (0, 0); the second will be assigned coordinates (1, 0). The display is not automatically updated to show the new baseline, so now go to the **Display** buttons and click on **Show BC**.

The image that appears in the box can be saved, with or without the axes, by clicking on **Copy Image to Clipboard** or **Save Image to an EPS File** (encapsulated postscript file). You can also print the image to the default Windows printer, which gives a quick method for obtaining a hardcopy of the image. The default is to include the axes as well as the coordinates in the image, so if you want to remove the axes, click on **Clear Axis**. The **Numbers on Landmarks** option displays the number of each landmark on the image near the mean location of that landmark in the data set (the red triangle). The **Figure Options** pull-down menu has several other options for controlling the image, including changing the size of the symbols and filling them. We will not explain the option **Spiffy Fish** – try it.

The program saves the coordinates in two file formats, listed under **Output File Format** (above the blue **Save Coordinates**). The default is X1Y1...CS, but you can change this to TPS format by clicking on that button. The TPS format is required by the software in the TPS series, so it is a good idea to save data in both formats as this will allow you to run programs in both the IMP and TPS series. After choosing the format, go to the blue **Save Coordinates** box and select the type of coordinates you wish to save (which, for the moment, are Bookstein coordinates, BC).

In addition to calculating the shape coordinates for all the specimens in the file, you can also calculate a reference form for analyses based on the thin-plate spline (introduced in Chapter 6). You do not actually need to calculate one because all programs that require it calculate it from the input data; however, you might want to control the choice of reference, or just to display the mean, and you can calculate it here. The default is the mean of the specimens in the file, but you can also calculate the mean of the *N* smallest (if you want to save or display the average of the juveniles, for example) or the mean of the *N* largest (if you want to save or display the average of the largest adults). Set the *N* value by using the *N* = window. The default *N* is 5. You do not need to make your choice now – you can simply reload the file at some later point and calculate the reference when you know what you want.

To load a new data file, use **Clear and Reset** to remove a data set from the program, then go back up to **Load Data** to load a new one.

Running BigFix

BigFix takes bilaterally symmetric raw data from the digitizer program **tpsDig** (the program cannot read files from any other digitizer), reflects landmarks of bilateral points, replaces missing landmarks with the coordinate of the bilateral homologue, and produces files of Bookstein shape coordinates. To use this program, the baseline points must be along the midline (they determine the axis of reflection).

Unlike all other programs in the IMP series, it can accept missing data so long as those landmarks are coded as missing (by "999"). If you cannot enter a 999 where a landmark is missing, keep notes and replace the digitized coordinate by 999 manually. As you might guess, **BigFix** can only handle missing data for *paired* landmarks (it uses the coordinates for the side that is present). If the landmark is available on the side being reflected, the sign of the Y-coordinate will be reversed (so the coordinate produced by **BigFix** will be on the same side as the others). If it turns out that you are missing a midline point, **BigFix** will give you an error message, stating which specimen lacks which landmark.

To proceed, **BigFix** needs to be told which landmarks are paired and which lie along the midline; it also needs all the information required by **CoordGen** (if you haven't read the manual for **CoordGen**, do so before using **BigFix**). The information about the pairing of landmarks must be in a separate file called the **Pair Configuration File**. This file consists of two columns of numbers (separated by a space or a tab). For bilaterally paired landmarks, the line contains the numbers of those landmarks; for midline landmarks, the line contains the number of that point and a 0 (meaning it has no pair). For example, given this **Pair Configuration File**

1	4
2	3
5	0
6	0

BigFix would interpret landmarks 1 and 4 as bilaterally homologous, 2 and 3 as bilaterally homologous, and 5 and 6 as midline points. It would then calculate coordinates of four landmarks: the average of 1 and the reflection of 4, the average of 2 and the reflection of 3, and the coordinates of 5 and 6 (which would have to be the two baseline points because they are the only two points on the midline).

BigFix has an option to show the new numbering of the points – a convenience if you have many paired landmarks that were digitized out of sequence (making it difficult to figure out which is which after they are reflected).

Running **BigFix** is otherwise like running **CoordGen**, so see the instructions for that program.

Running TwoGroup

TwoGroup tests the hypothesis that the two groups differ by more than expected by chance, using a Hotelling's T^2-test applied to Bookstein shape coordinates. It also conducts other tests, but the only one suited to Bookstein shape coordinates is Hotelling's T^2-test. To run the program, load the two groups, first clicking on **Load Data Set 1** and entering the file name, then clicking on **Load Data Set 2** and entering that file name. You will then need to enter the **Baseline Endpoints** for your files (the defaults are landmarks 1 and 7 – the endpoints of the baseline used in the piranha analyses). Then go to **Analytic Tests**, and click on **Hotelling's T^2 (BC)** – the parenthetical BC means Bookstein coordinates. The statistical result will appear in the results window beneath.

Your two data sets will appear in the visualization window. You can also display the mean for each data set by clicking on **Show Means** (to return to the display of both samples, click on **Show Data**). The program offers several options aside from Bookstein shape coordinates, but for now use the option **BC**. You can alter the color of the symbols at the landmarks, their size, and whether they are hollow or filled, by selecting options on the **Symbol Controls** menu on the toolbar at the top of the program. These pictures can also be saved by clicking on **Copy Image to Clipboard** or **Copy Image to EPS File**. Additionally, you can display the difference between the means as vectors of relative landmark displacements by clicking on **Plot Difference in Means**. There is a long list of options for these displays (controlled by the pull-down **Difference Plot Options** menu), including line weight, line color, symbol type, arrowheads, filled or hollow symbols, and symbol size. There are other options as well, but they are for methods that will not be introduced until later in this book. Like the other images, these can be copied to the clipboard or to an EPS file.

References

Bookstein, F. L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.

Morrison, D. F. (1990). *Multivariate Statistical Methods*. McGraw-Hill.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. John Wiley & Sons.

4

Theory of shape

This chapter covers the basic theory of shape, beginning with the definition of shape and proceeding through the characterization of several theoretical spaces. Some of the mathematics may look a bit difficult, but it is important to grasp the basic ideas, which we present verbally as well as mathematically. These ideas will reappear in the next two chapters, because they form the core of geometric morphometrics. Interestingly, many of the techniques used in geometric morphometrics were developed independently of this theory even though they are justified by it. As the theory matured, it became possible to synthesize a large body of techniques that had been developed independently of each other and to explicate their interrelationships. Perhaps most importantly, this theory also allows us to judge whether or not methods are valid. The theory provides the underlying justifications for all our techniques, thereby allowing us to make inferences about shape without worrying that those inferences are somehow based on arbitrary or mathematically faulty choices that we happened to make in the course of our analyses. Freed of such concerns, we can concentrate on the biological meaning of the results.

It would be possible to learn techniques without understanding any of this theory – but don't. Without the theory it is impossible to say why some methods are right and others are not. In effect, you would have to memorize a list of "dos" and "don'ts" by rote without understanding why the "dos" are "dos" and the "don'ts" are "don'ts." Learned in that way, it might seem that there are lots of picky rules and dogma, but these rules are not picky and they are not a matter of dogma. Rather, they all logically (and mathematically) follow from the mathematical theory of shape. In fact, they follow from the definition of shape. Because this definition is central to geometric morphometric theory, we begin there, developing it further than in previous chapters.

The definition of shape

David Kendall's (1977) definition of shape is the basis of all that will follow in this chapter, and indeed of any consideration of the meaning of shape:

> Shape is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object.

ISBN 0–12–77846–08
Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

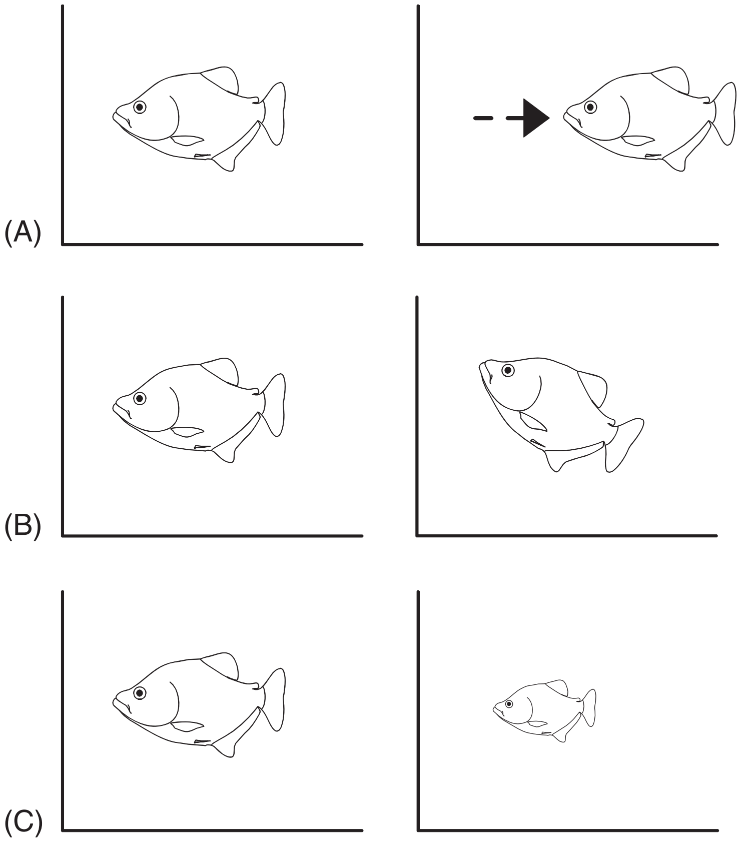


Figure 4.1 The operations that do not alter shape: (A) translation; (B) rotation; (C) scaling.

This statement is both intuitively reasonable and mathematically useful. It suits our intuitions because we can all agree that moving an object from one place to another does not change its shape; that operation, called translation, obviously does not alter shape. For example, Figure 4.1A shows the translation of a shape along an axis, and this motion has no consequences for shape. Likewise, rotating the object does not change shape (Figure 4.1B), and neither does enlarging or reducing an image (a manipulation called rescaling; Figure 4.1C). Although it may be obvious that translation, rotation and rescaling do not alter shape, it may not be obvious that this fact provides a mathematically useful definition of shape.

To a non-mathematician this definition may seem a bit odd, because it defines shape by what *does not* alter it rather than in terms of what shape *is* or by the operations that *do* alter it. However, the definition is useful because it means that any operation *not* on that list *does* affect shape. Also, the list of operations that do not alter shape is useful because we know that we are free to use those operations when we compare shapes mathematically.

The entire theory of geometric morphometrics follows from the definition of shape, so we need to develop it further. First, we need a more precise definition of a landmark. When we discussed the criteria for choosing them in Chapter 2, we emphasized that the criterion of homology has mathematical as well as biological implications. The mathematical implication follows from the formal definition of a landmark (Dryden and Mardia, 1998):

> A landmark is a point of correspondence on each object that matches between and within populations.

The concept of matching encoded in that passage is not necessarily one of biological homology, but the idea of correspondence is essential to the mathematical theory of shape. If the landmarks do not correspond, we cannot compare shapes.

Another crucial idea is that of a *configuration of landmarks*; the full set of landmarks recorded for each specimen. All comparisons of shapes are between matching configurations of landmarks, not between individual landmarks (analyzed separately). An individual landmark is not an object of comparison because it does not satisfy the definition of shape. The objects of comparison are entire configurations comprised of K landmarks (where K refers to the number of landmarks), each of which has M coordinates (i.e. $M = 2$ for planar shapes). For example, in the case of the piranhas introduced in the second chapter, $K = 16$ and $M = 2$. Whatever the number of landmarks and coordinates, our analyses and conclusions are based on the *entire set*. Thus if we have 16 landmarks with two coordinates apiece, we have one shape – not 32 variables. No one landmark (and no one coordinate) is a shape variable in its own right. Instead, we view each shape as the entire configuration and we analyze samples of entire configurations.

This is a very different view of measurement (and variables) from that commonly encountered in traditional morphometrics, where a single measurement might be viewed as a variable, meriting analysis in its own right. It is common to analyze measurements separately and to draw biological conclusions from them individually. Sometimes the conclusions based on one measurement conflict with conclusions based on another, and the inference often drawn in such situations is that the processes are trait-specific. In geometric morphometrics, individual measurements are not traits or even variables. Rather, a shape variable is the entire vector of coefficients representing the complete difference in landmark configurations between samples, or, alternatively, the entire vector of coefficients measuring the covariance between the landmark configurations and some other variable (e.g. size).

This view of shape as a configuration of landmarks is central to the theory of geometric morphometrics. Recognizing that, and conforming to the requirements it imposes on analytic methods, is crucial. It may seem biologically unreasonable to treat an entire shape as a single entity, but the pay-off for doing so is the guarantee that our results do not depend on arbitrary choices we happened to make in the course of an analysis. The reward for following what might seem like a rigid set of rules is the rigor and power of these methods, as well as the visual appeal of the graphics.

Morphometric spaces

Given the definition of shape, we can now develop the mathematical idea of morphometric spaces. We begin by defining some additional terms.

The configuration matrix

A configuration matrix represents an entire configuration of landmarks. It is a $K \times M$ matrix of Cartesian coordinates that describes a particular set of K landmarks in M dimensions (Dryden and Mardia, 1998). When we talk about a $K \times M$ matrix, we mean that the matrix has K rows and M columns; each of the K rows represents a specific landmark on a specimen, with M Cartesian coordinates. For example, the simplest shape we might want to study is a triangle with landmarks located at the three vertices of the triangle. Calling the coordinates of the first vertex X_1 and Y_1, and those of the second vertex X_2 and Y_2, and those of the third vertex X_3 and Y_3, the configuration matrix of triangle **X** is:

$$\mathbf{X} = \begin{bmatrix} X_1 & Y_1 \\ X_2 & Y_2 \\ X_3 & Y_3 \end{bmatrix} \tag{4.1}$$

It is often useful to represent this same landmark configuration as a *row vector*, in which the landmark coordinates are listed along a single row in $K \times M$ columns:

$$\mathbf{X} = [X_1 \quad Y_1 \quad X_2 \quad Y_2 \quad X_3 \quad Y_3] \tag{4.2}$$

This contains exactly the same information, represented slightly differently. Given a set of landmark coordinates in row vector form, you can easily convert it to a configuration matrix (the representation you might prefer at any given time depends on the particular task or software at hand).

For example, the configuration matrix of the triangle shown in Figure 4.2 is:

$$\mathbf{X} = \begin{bmatrix} -1 & -1 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \tag{4.3}$$

The row vector representing the same triangle would be:

$$\mathbf{X} = \begin{bmatrix} -1 & -1 & 1 & -1 & 0 & 1 \end{bmatrix} \tag{4.4}$$

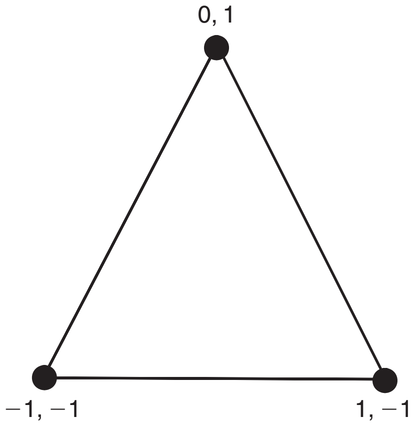


Figure 4.2 Example of a triangle.

Configuration space

The configuration space is a set of all possible $K \times M$ matrices describing all possible sets of landmark configurations for that given K and M. For example, a 16×2 dimensional configuration space is the space of all configurations having 16 two-dimensional landmarks. That space encompasses *all* possible configurations for those 16 landmarks with two coordinates. Should we record the locations of 16 landmarks on a two-dimensional image of a piranha, and 16 landmarks on a two-dimensional image of a rat skull, both configuration matrices are in the same configuration space. Clearly, any group of biologically similar organisms (with matched landmarks) will occupy a relatively small part of configuration space because the locations of their corresponding landmarks will be fairly similar. For example, in the 16×2 configuration space, piranhas will occupy a very small part of a space – that space also contains the 16×2 two-dimensional coordinates of rat skulls.

The configuration space of K landmarks with M coordinates per landmark has $K \times M$ dimensions. To specify the location of any shape in that space, we must specify $K \times M$ components of a vector (or elements in a matrix).

Position or location of a configuration matrix

The position of a configuration matrix is the location of the centroid of that matrix. This centroid is the M-dimensional vector (two in the case of the two-dimensional landmarks of piranhas) whose components are the averages of the X and Y coordinates of the landmarks (in the two-dimensional case), so the centroid position is given by:

$$X_C = \frac{1}{K}\sum_{j=1}^{K} X_j$$
$$Y_C = \frac{1}{K}\sum_{j=1}^{K} Y_j \tag{4.5}$$

For example, Figure 4.3 shows the centroid position of the triangle seen earlier, which is located at $(0, -0.333)$.

A configuration matrix is said to be *centered* if the average of all the coordinates is zero. Centering is useful because it often simplifies the mathematics; it is done by translating the configuration along the X- and Y-axes. That translation is done by adding a constant (positive or negative) to the X- and Y-coordinates. To do this we first calculate the X and Y centroid coordinates of the configuration matrix $\mathbf{X}$ as in Equation 4.5, then subtract the centroid positions from each coordinate to form the centered configuration matrix $\mathbf{XC}$:

$$\mathbf{XC} = \begin{bmatrix} (X_1 - X_C) & (Y_1 - Y_C) \\ (X_2 - X_C) & (Y_2 - Y_C) \\ \vdots & \vdots \\ (X_K - X_C) & (Y_K - Y_C) \end{bmatrix} \tag{4.6}$$

Two configuration matrices that differ only in the position of the centroid are not different shapes (they differ only by translation, one of the operations that does not alter shape).

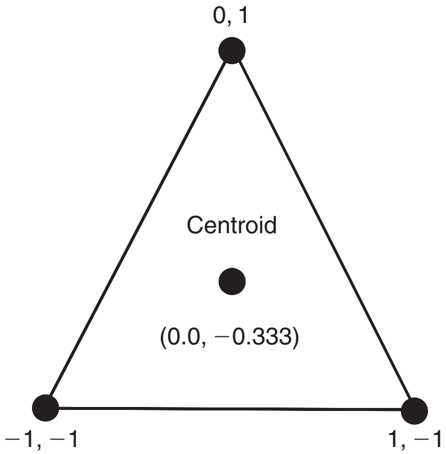


Figure 4.3 The centroid of the triangle in Figure 4.2. The coordinates of the centroid are the averaged coordinates of the three vertices.

Size of a configuration matrix

Before we can coherently talk about *scale*, we need to define what we mean (mathematically) by the term *size*. For configuration matrices, a number of different, non-equivalent size measures have been used. It is not possible to say that one size measure is "correct" or "preferable," but it is important to explain the consequences of making a particular choice. The most commonly used size measure in geometric morphometrics is called *centroid size*, which is favored because it does not induce a correlation between size and shape, hence we restrict our discussion of size to that particular measure.

The centroid size (CS) of a configuration ($\mathbf{X}$) is:

$$CS(\mathbf{X}) = \sqrt{\sum_{i=1}^{K}\sum_{j=1}^{M}(\mathbf{X}_{ij} - C_j)^2} \tag{4.7}$$

where the sum is over the rows i and columns j of the matrix $\mathbf{X}$. $\mathbf{X}_{ij}$ is a standard notation from linear algebra specifying the value located on the ith row and jth column of the matrix $\mathbf{X}$, and in this case C_j stands for the location of the jth component of the centroid. C_1 is the X-coordinate of the centroid and C_2 is its Y-coordinate.

Centroid size is thus the square root of the sum of the squared distances of the landmarks from the centroid. The distances from the centroid to each landmark of the triangle are shown in Figure 4.4; the centroid size of this triangle is simply the square root of the sum of the squared lengths of these lines. Centroid size is not altered by changing the position of the configuration, because this leads to all landmarks (and the centroid) changing by a common amount. Similarly, multiplying the configuration matrix $\mathbf{X}$ by a constant factor increases centroid size by the same factor. Two configurations of landmarks that differ only in centroid size do not differ in shape (they differ only in scale).

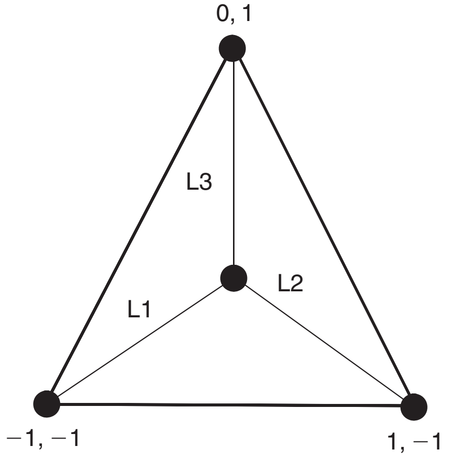


Figure 4.4 Centroid size of the triangle in Figure 4.2, calculated as the sum: $(\mathrm{L1}^2 + \mathrm{L2}^2 + \mathrm{L3}^2)^{1/2} = 2.16$.

Pre-shape space

As we stated above, every configuration of K landmarks having M coordinates can be thought of as a point in a space with $K \times M$ dimensions. (To avoid confusion, we should make it clear that by "point" in this context we mean an individual shape, an entire configuration of landmarks, not one landmark.) Some of the configurations in this space differ only in centroid size; others differ only in location (coordinates of the centroid). We can define a subset of configurations that do not differ in location or size by placing two restrictions on each configuration matrix: (1) that it be centered, and (2) that centroid size be one. These restrictions define a space called pre-shape space (Dryden and Mardia, 1998). In practice, we translate and scale each of the original configurations in our data so that the new configurations meet the restrictions of pre-shape space. In doing this, we are using two of the three operations that do not alter shape. Each of the new configurations is a centered pre-shape.

The shape of pre-shape space

The two requirements imposed on this space mean that the summed squared landmark positions add up to one. The consequences of that property can be understood by considering the set of points satisfying the restriction in an ordinary two-dimensional space: the set of points is centered on the origin $(0, 0)$, and each point in the set has coordinates satisfying the equation $X^2 + Y^2 = 1$. The set of points is a circle of radius one, centered on the origin. This circle is a one-dimensional subspace (a curve) inhabiting a two-dimensional space (a plane). Knowing that all points are equidistant from the center means that we need specify only the direction of a point from the center to define it uniquely; thus, the location of any point on the circle can be described sufficiently by a single dimension (direction). Extending this to a three-dimensional space, we now have the set of all points (X, Y, Z) centered on the origin $(0, 0, 0)$ such that $X^2 + Y^2 + Z^2 = 1$. This is the surface of a sphere of radius one, centered on the origin, and it is a two-dimensional subspace within a three-dimensional space. Again, the constraint that all points are on the surface allows us to describe the location of a point by giving a direction from the center; the only difference from the circle is that we now need two components to describe that direction (e.g. latitude and longitude). So in talking about a pre-shape space we are talking about the surface of a hypersphere centered on the origin, which is the generalization of an ordinary sphere in $K \times M$ dimensions. In that general case, we have:

$$\sum_{i=1}^{K}\sum_{j=1}^{M}(\mathbf{X}_{ij})^2 = 1 \tag{4.8}$$

which states that the sum of all squared landmark coordinates is one. That hypersphere is simply the equivalent of a sphere in more than three dimensions.

We can determine the number of dimensions in pre-shape space by considering the number of dimensions that were lost in the transition from configuration space. One dimension is lost in fixing centroid size to one, eliminating the size dimension of the configuration space. Another, M dimensions are lost in centering the configurations; eliminating the M dimensions needed to describe location (the coordinates of the centroid). Thus in moving from configuration space to pre-shape space, we moved to a space that has $M+1$ fewer dimensions, which is:

$$KM-(M+1)=KM-M-1 \tag{4.9}$$

For two-dimensional configurations of landmarks, pre-shape spaces have $2K-3$ dimensions; so the pre-shape space for triangles has three dimensions. For three-dimensional configurations of landmarks, pre-shape spaces have $3K-4$ dimensions.

Returning to the three-dimensional sphere (because most of us have trouble imagining spaces having more than three dimensions), you should be imagining pre-shape space to be a hollow ball of radius one, centered at the origin $(0, 0, 0)$. Arrayed on the two-dimensional surface of this ball are points representing individual configurations of landmarks. The two restrictions we have imposed on our configuration matrices mean that the configurations in this set do not differ in scale or location; we have used the operations of translation and scaling to remove the effects of (differences in) location and scale. We have not yet rotated the shapes to remove the effects of rotation (that comes later, as we move from pre-shape space to shape space). Thus, configurations of landmarks that differ only by a rotation are located at different points in pre-shape space, as are configurations that differ only in shape. This underscores an important point (which some may find counterintuitive): as we said earlier, configurations that differ only by a rotation (such as those shown in Figure 4.1B) do not differ in shape. Because we have not yet removed all three effects mentioned in Kendall's definition of shape (location, scale and rotation) we have not yet reached shapes. At present we are concerned with pre-shapes, i.e. configurations that may differ by a rotation, by a shape change or by some combination of the two. In pre-shape space, configurations that differ only by rotation are different points, as are configurations that differ only in shape.

Fibers in pre-shape space

To visualize the locations in pre-shape space of configurations that differ only in rotation, we introduce the term *fiber*. A fiber (in the context of our particular discussion of pre-shape space) consists of the set of all the points in pre-shape space that can be obtained by rotating a particular centered pre-shape. The fiber is a circular arc that comprises the set of all points in pre-shape space that can be "reached" by rotating the pre-shape matrix. Figure 4.5 depicts the concept of fibers as an arc on the surface of a sphere (ignoring the higher dimensionality of a pre-shape hypersphere). Two fibers are shown: arcs 1 and 2. Arc 1 is the set of all possible rotations of the pre-shape $\mathbf{Z}_1$, and arc 2 is the set of all possible rotations of the pre-shape $\mathbf{Z}_2$. For a less abstract visualization of the concept of fibers, we have drawn a cartoon (Figure 4.6) representing four fibers (in columns); the triangles within a column differ solely by a rotation, whereas those in different columns also differ in shape. (This visualization is somewhat limited, because a row does not accurately represent the number of dimensions needed to describe shapes of triangles, as explained in the next section.)

With the concept of fiber in hand, it is now possible to talk about the separation of shapes and the distance between them. Figure 4.7 shows the same two fibers on the curved surface of the pre-shape space hypersphere as in Figure 4.5. In addition, Figure 4.7 shows an arc (ρ) crossing the surface from one fiber to the other, and the chord (D_p) that passes through the interior of the hypersphere between the same two surface points. We can draw many such arcs connecting a rotation of the pre-shape $\mathbf{Z}_1$ with a rotation of the pre-shape $\mathbf{Z}_2$. The arc we want is the shortest one – that is, the one connecting fibers at their "point of closest approach." Finding the shortest possible distance between points is a common tactic for defining distances between objects in spaces. When we find that distance, we will find the rotation that is optimal in the sense of being the minimum distance between

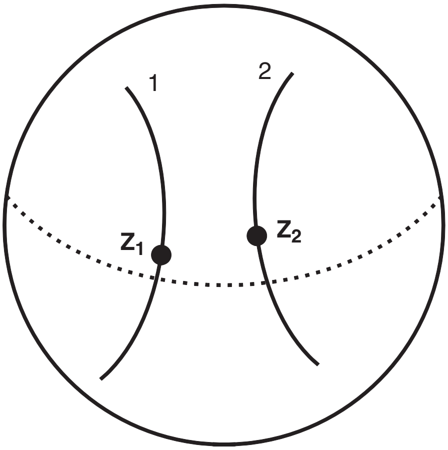


Figure 4.5 Fibers in pre-shape space. The points $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are the locations of pre-shapes on the hypersphere (centered and scaled matrices computed from two original matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, which are not shown). Curve 1 passing through $\mathbf{Z}_1$ is a fiber, the set of all centered and scaled pre-shapes differing from $\mathbf{Z}_1$ only by rotation. Curve 2 is a fiber of pre-shapes differing from $\mathbf{Z}_2$ only by rotation. (The dotted curve is the "equator" of the hypersphere, and does not represent a fiber.)

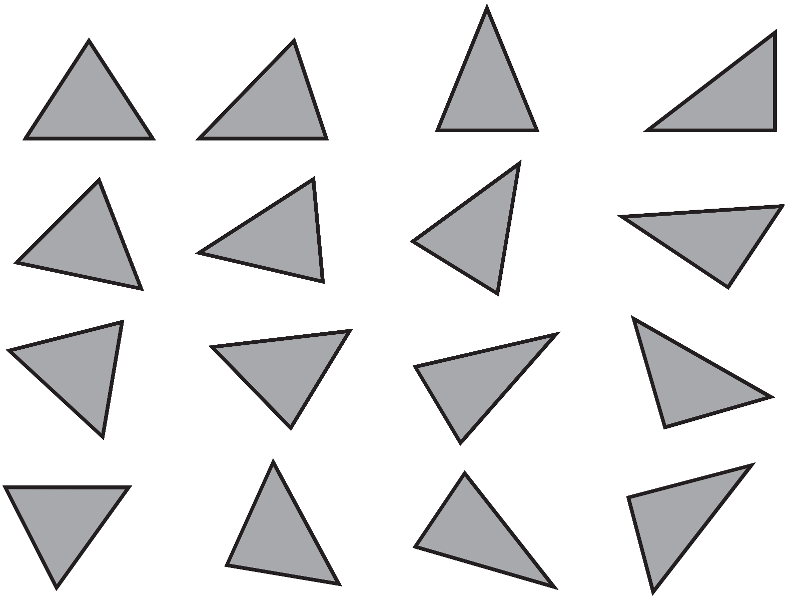


Figure 4.6 An alternative visualization of the concept of a fiber. Each column shows rotations of a single shape; triangles in different columns differ in shape. Each column represents a single fiber.

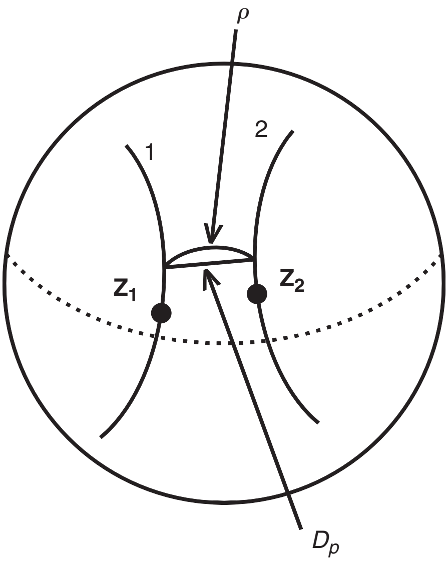


Figure 4.7 Determining the distance between the fibers of pre-shapes. The arc ρ is the shortest distance across the surface of the hypersphere from fiber 1 to fiber 2. The length of the arc is the Procrustes distance. The length of the chord (D_p) is the partial Procrustes distance.

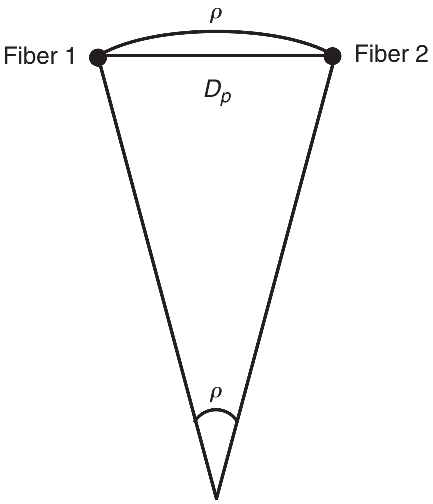


Figure 4.8 A slice through pre-shape space showing the Procrustes distance (ρ) and the partial Procrustes distance (D_p).

shapes. The length of this arc is known as the *Procrustes distance*, and it is quantified by determining the angle between the radii that connect the center of the hypersphere to the point at which the fibers most closely approach each other. Figure 4.8 shows the cross-section through the pre-shape space in the plane defined by those two radii. The angle subtended by the arc is ρ; the chord length is D_p. The length of the arc is equal to ρ (in radians) times the length of the radius. Because we have constrained the radius to a length of one, the length of the arc is the value of the angle. This value ranges from zero to π; at π, the two shapes are on opposite sides of pre-shape space.

Shape spaces

In the previous section, we used the points of closest approach on the pre-shape fibers to define the distance between two shapes. Now, we use the same criterion to construct a shape space. This shape space contains one configuration from each fiber, one rotation of a centered pre-shape. Conventionally, we select a convenient orientation of one pre-shape to serve as the *reference* configuration; every other *target* (or subject) configuration is selected as the rotation corresponding to the point of closest approach of its pre-shape fiber to the reference. That is, the orientation is chosen to minimize the Procrustes distance between the target and reference. The points on those fibers that are farther from the reference differ from it in both shape and rotational effects. By selecting the point of closest approach, we reduce each fiber of pre-shapes to a single point (a shape); consequently, configurations in this set differ only in shape.

The shape space we just described has fewer dimensions than the pre-shape space from which it was derived. The number of dimensions lost in the transition are given by:

$$\frac{M(M-1)}{2} \tag{4.10}$$

where M is the number of landmark coordinates. For two-dimensional landmarks, Equation 4.10 simplifies to one, which reflects the fact that a planar shape can only be rotated about its centroid on one axis (the axis perpendicular to the plane of the shape) and still stay in the same plane. Consequently, shape spaces of two-dimensional configurations of K landmarks have $2K-4$ dimensions. The four lost dimensions are those describing differences in size (-1), translation (-2) and rotation (-1). For three-dimensional landmarks, Equation 4.10 simplifies to three, which reflects the fact that a three-dimensional shape can be rotated about its centroid on three distinct orthogonal axes in the three-dimensional coordinate space. Subtracting three from the $3K-4$ dimensions of the pre-shape space (from Equation 4.9) yields $3K-7$ dimensions for shape spaces of three-dimensional shapes, which simplifies to five dimensions for the shape space of tetrahedra. The seven lost dimensions are those describing differences in size (-1), translation (-3) and rotation (-3).

In the special case of triangles, the shape spaces defined above are the familiar two-dimensional surfaces of three-dimensional spheres. Because this is a reasonably simple geometry to visualize and illustrate, we will focus on triangles before returning to the general case. In Figure 4.9 we show half of a space determined by using the equilateral triangle as the reference. Because we retain the constraints that each triangle is centered and scaled to centroid size of one, the sphere has a radius of one. For convenience, the space is oriented so that the point representing the equilateral triangle configuration is located at the pole. At the equator are triangles with zero height; in other words, various arrangements of three collinear landmarks. The other half of the space would be an exact

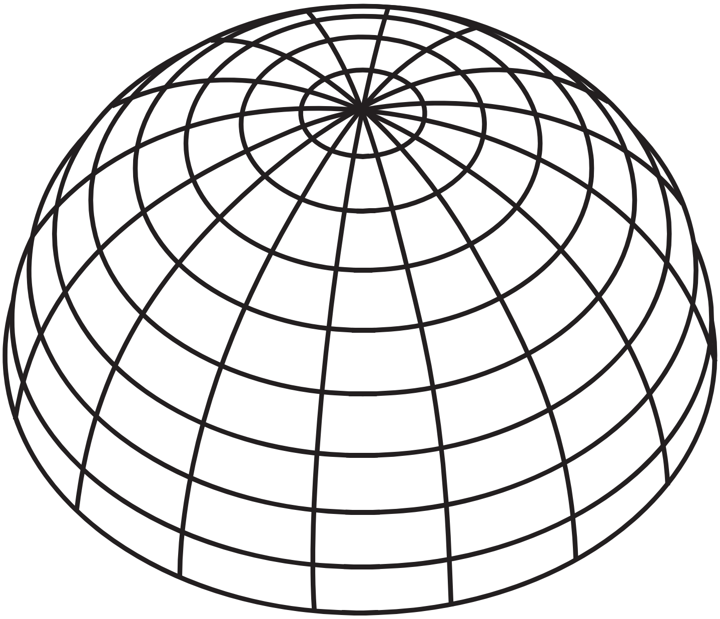


Figure 4.9 Half of the space of triangles that have been centered, scaled to unit centroid size and aligned with a centered, scaled equilateral triangle. The equilateral triangle is at the pole. Lines of "latitude" represent shapes equidistant from the equilateral triangle. The "equator" corresponds to the set of triangles with zero height (three collinear landmarks).

reflection of the one shown. Each shape would be a simple reflection of the shape that is at a corresponding location in the other hemisphere. In the case of triangles, a reflection is equivalent to a rotation of 180^o (albeit on a different axis from the one considered earlier), so we can discard the bottom half because it contains the same shapes as the top.

Although the shape space just described is a useful construction, it does not satisfy the mathematician's urge to find the smallest distances between configurations with those shapes. To illustrate this point, we consider a slice through the polar axis of the hemisphere of triangles just described (Figure 4.10). As in pre-shape space, the distance of a shape (A) from the reference is ρ. The angle and the arc length are unchanged because the dimension eliminated in the transition from pre-shape space to this shape space did not contribute to the measurement of the shape difference. It should be apparent in Figure 4.10 that the arc across the surface is not the shortest possible distance between the two shapes. The chord passing through the interior of the hemisphere would be shorter, but it is still not the shortest possible distance between configurations with those shapes. We obtain that shortest possible distance, and the relevant configurations, by changing the constraint on the centroid sizes of the two configurations. Conventionally, we keep the centroid size of the reference at one, and allow the centroid size of the target to adopt the value that minimizes its distance from the reference. This is equivalent to allowing the target to travel along its radius while the reference stays on the surface. The point along the radius where the second shape is closest to the target is some distance below the surface of the shape space, reflecting a reduction of the centroid size of the target. This point (B) is defined by the line that is perpendicular to the target's radius and passes through the reference's position on the surface. The corresponding centroid size of the target is $\cos(\rho)$; the distance between configurations is $\sin(\rho)$ and is called the *full Procrustes distance* (D_F).

Because $\cos(\rho)$ decreases as ρ increases, scaling each configuration in the shape space to $\cos(\rho)$ (where ρ is its distance from the reference) produces a new shape space sphere with a

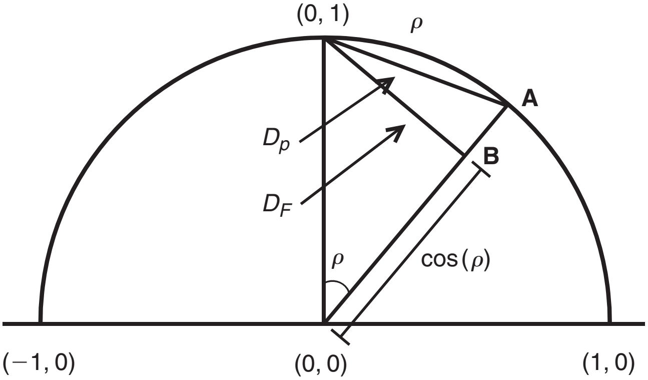


Figure 4.10 A slice through part of the space of aligned triangles at unit centroid size, showing the relationships among the distances between the reference shape (at 0, 1) and **A**. The semicircle is a cross-section of the space, which is a hemisphere of radius one. The length of the arc is the Procrustes distance (ρ), the length of the chord is the partial Procrustes distance (D_p), and the shortest possible distance (obtained by relaxing the constraint on centroid size, producing the configuration **B**) is the full Procrustes distance (D_F).

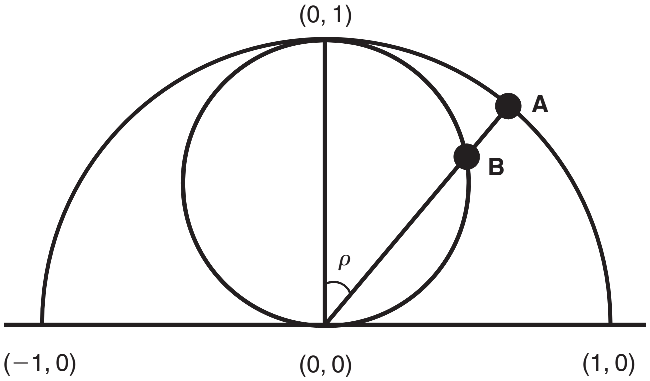


Figure 4.11 The relationship of Kendall's shape space to the space of aligned triangles scaled to unit centroid size. The outer semicircle is the cross-section of the space of aligned triangles scaled to unit centroid size, as in Figure 4.10. The inner circle is a cross-section through Kendall's shape space, which is the sphere of aligned triangles scaled to cos(ρ). Kendall's shape space has a radius of one-half. Points **A** and **B** represent the same shape at $CS = 1$ and $CS = \cos(\rho)$, respectively.

radius of 1/2, tangent to the previous shape space at the reference shape (Figure 4.11). This new space is *Kendall's shape space* for triangles; it is the set of centered shapes in which each is at the size and orientation that minimizes its distance from the reference. It may appear that Kendall's shape space is dramatically different from the previous shape space, but certain key properties remain the same. One of these properties is the distance of the target shape from the reference shape across the surfaces of the shape spaces. In the first shape space, the distance of the target from the reference was ρ, the angle subtended by the arc. In Kendall's shape space, the angle subtended by the arc is now 2ρ, but the radius is 1/2, so the arc length is $2\rho/2$. Although distances between the reference and the targets are not altered, distances between targets are (Slice, 2001). Another key property that remains the same is the number of dimensions. In the transition between shape spaces, the constraint on centroid size was changed; in Kendall's shape space the constraint is cos(ρ) instead of one. This still specifies a single value for each shape; configurations that differ only in size are represented by a single point in Kendall's shape space. Thus, Kendall's shape space for triangles is also the two-dimensional surface of a three-dimensional sphere.

For configurations of landmarks that are more complex than triangles, we can apply the same set of operations to move from pre-shape space to the two shape spaces. Regardless of the number of landmarks and the number of coordinates of those landmarks, the transitions involve: (1) selecting the rotations that are at the minimum distance from the reference in pre-shape space, and (2) finding the centroid sizes that fully minimize the distance from the reference. Describing the geometric relationship of these spaces at higher dimensions is rather demanding (Small, 1996), but near their poles (i.e. near the reference configurations) these spaces are expected to have similar properties to the spaces for triangles (Slice, 2001).

Kendall's shape space and all of the spaces described above are curved, non-Euclidean spaces. This is important because the conventional tools of statistical inference assume a linear, Euclidean space. Consequently, we cannot use those tools to analyze shapes in Kendall's shape space. Much of Kendall's own work concerns statistical inference within the curved space that bears his name, but most biologists do not need to work in that space. As discussed in a later section of this chapter, it is possible to map locations in Kendall's shape space to locations in a Euclidean space tangent to Kendall's shape space. Like planar maps of the Earth, the Euclidean "maps" of shape space distort the relative positions of shapes far from the tangent point. This becomes important when comparing extremely dissimilar shapes. In most biological studies the range of shapes will be small relative to the curvature of the space, so the distortion will be mathematically trivial for any well-considered choice of the tangent point (we discuss criteria for selecting the tangent point in a later section). If you are comparing such highly dissimilar shapes that you need to work in Kendall's shape space, you will need a more detailed understanding of this space than presented here. The excellent texts by Dryden and Mardia (1998) and Small (1996) discuss the variables and procedures for carrying out inference in Kendall's shape space.

Finding the angle of rotation that minimizes the Euclidean distance between two shapes

To determine the angle of rotation required to place one pre-shape at a minimum Procrustes distance from a second, it is sufficient to rotate the first shape (the target) to minimize the summed squared distance between it and the reference. This distance we are minimizing is the partial Procrustes distance. Because the Procrustes distance is a monotonic function of the partial Procrustes distance, this minimization of the partial Procrustes distance also minimizes the Procrustes distance.

An arbitrary rotation of the target form (of two-dimensional landmarks, $M = 2$) by an angle θ maps the paired landmarks (X_{Tj}, Y_{Tj}) of the target to the coordinates $((X_{Tj}\cos\theta - Y_{Tj}\sin\theta), (X_{Tj}\sin\theta + Y_{Tj}\sin\theta))$. The sum of the squared Euclidean distances between the K landmarks of this rotated target and the reference is:

$$D^2 = \sum_{j=1}^{K}\left[(X_{Rj} - (X_{Tj}\cos\theta - Y_{Tj}\sin\theta))^2 + (Y_{Rj} - (X_{Tj}\sin\theta + Y_{Tj}\cos\theta))^2\right] \tag{4.11}$$

where (X_{Rj}, Y_{Rj}) are the coordinates of the landmark in the reference. To minimize this squared distance as a function of θ, we take the derivative with respect to θ and set it equal to zero:

$$-\sum_{j=1}^{K}\left[\begin{array}{l} 2(X_{Rj} - (X_{Tj}\cos\theta - Y_{Tj}\sin\theta))(-X_{Tj}\sin\theta - Y_{Tj}\cos\theta) \\ \quad + 2(Y_{Rj} - (X_{Tj}\sin\theta + Y_{Tj}\cos\theta))(X_{Tj}\cos\theta - Y_{Tj}\sin\theta) \end{array}\right] = 0 \tag{4.12}$$

and solve for θ:

$$\theta = \text{arctangent}\left(\frac{\sum_{j=1}^{K} Y_{Rj}X_{Tj} - X_{Rj}Y_{Tj}}{\sum_{j=1}^{K} X_{Rj}X_{Tj} + Y_{Rj}Y_{Tj}}\right) \tag{4.13}$$

which gives us the angle by which to rotate the target to minimize its distance from the reference.

A numerical example for the simplest case

To make the preceding discussion of theory more concrete and accessible, we apply the ideas to the simplest useful case, the space of triangles (this space has been discussed extensively in Small, 1996; Dryden and Mardia, 1998; Rohlf, 2000; Slice, 2001). We have used this example throughout this chapter, but we now pull all the information together. There are other approaches to constructing the matrices representing shapes in Kendall's shape space, but the sequence of steps we follow here is easily illustrated and requires relatively simple computations.

We begin with two triangles, **X** and **W**, drawn on a flat surface (Figure 4.12). **X** is the triangle from Figure 4.2, with coordinates $(-1,-1)$, $(1,-1)$ and $(0,1)$; triangle **W** has coordinates $(1.07,-1.64)$, $(3.10,-0.72)$ and $(1.55,0.82)$. Each triangle has $K=3$ landmarks with $M=2$ coordinates; thus the configuration matrix for each has six entries:

$$\mathbf{X}=\begin{bmatrix}-1 & -1\\ 1 & -1\\ 0 & 1\end{bmatrix}\quad \mathbf{W}=\begin{bmatrix}1.07 & -1.64\\ 3.10 & -0.72\\ 1.55 & 0.82\end{bmatrix}\qquad (\mathbf{4.14})$$

The six landmark coordinates of each triangle contain six pieces of information needed to determine all the properties of that triangle: size, shape, location, and rotation. Not only do we need all six coordinates to determine these properties; we cannot infer the value of any one coordinate from the other five. Because we need all six coordinates to determine the triangle, we can say there are six *degrees of freedom*. This also helps to explain why the configuration space of triangles has six dimensions.

We can infer from the coordinates that the two triangles have different locations, as suggested in the figure. We confirm this by calculating the coordinates of the centroid

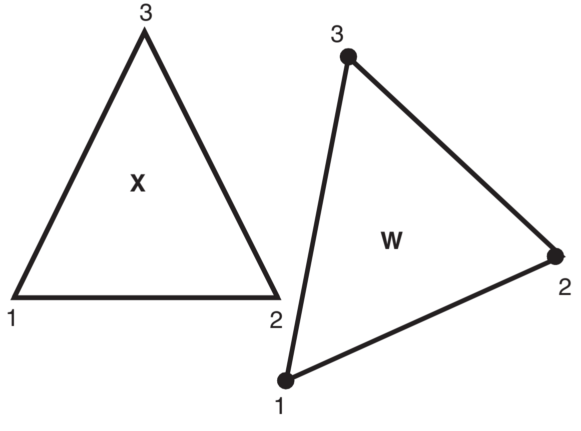


Figure 4.12 Two triangles, **X** (from Figure 4.2) and **W**. The vertices are numbered to indicate their homologies.

using Equation 4.5, reproduced here:

$$X_C = \frac{1}{K}\sum_{j=1}^{K} X_j$$
$$Y_C = \frac{1}{K}\sum_{j=1}^{K} Y_j \qquad (4.15)$$

For triangle **X**, the coordinates of the centroid are $X_C = (1/3)(-1+1+0) = 0$, and $Y_C = (1/3)(-1+-1+1) = -0.333$. For triangle **W**, the coordinates of the centroid are $X_C = (1/3)(1.07+3.10+1.55) = 1.907$ and $Y_C = (1/3)(-1.64+-0.72+0.82) = -0.513$.

We use the coordinates of the centroid to form the centered configuration matrix **XC** by subtracting the centroid coordinate from the corresponding coordinate of each landmark:

$$\mathbf{XC} = \begin{bmatrix} (X_1 - X_C) & (Y_1 - Y_C) \\ (X_2 - X_C) & (Y_2 - Y_C) \\ \vdots & \vdots \\ (X_K - X_C) & (Y_K - Y_C) \end{bmatrix} \qquad (4.16)$$

This produces the centered configuration matrices:

$$\mathbf{X}_{\text{centered}} = \begin{bmatrix} (-1-0) & (-1-(-0.333)) \\ (1-0) & (-1-(-0.333)) \\ (0-0) & (1-(-0.333)) \end{bmatrix} = \begin{bmatrix} -1 & -0.667 \\ 1 & -0.667 \\ 0 & 1.333 \end{bmatrix} \qquad (4.17)$$

and

$$\mathbf{W}_{\text{centered}} = \begin{bmatrix} (1.07-1.907) & (-1.64-(-0.513)) \\ (3.10-1.907) & (-0.72-(-0.513)) \\ (1.55-1.907) & (0.82-(-0.513)) \end{bmatrix} = \begin{bmatrix} -0.837 & -1.127 \\ 1.193 & -0.207 \\ -0.357 & 1.333 \end{bmatrix} \qquad (4.18)$$

The centered triangles are shown in Figure 4.13. One consequence of centering is that the two triangles are now *superimposed*; another is the loss of two degrees of freedom. Knowing that the centroid has coordinates $(0, 0)$, which are the means of the landmark coordinates, we can use the coordinates of any two landmarks to determine the coordinates of the third landmark. Accordingly, the space of centered triangles (which we have not discussed previously) is a four-dimensional space. Another way to think of this is that the two coordinates of the centroid, specifying the location of the triangle, account for two of the six dimensions of the configuration space. Also, now that all individuals have the same value for their centroid coordinates, the variation due to position disappears, collapsing that dimension of variation to a point at the origin.

The centered triangles are not in pre-shape space. To put them there, we need to rescale each so that its centroid size is one. The formula for centroid size is:

$$CS(\mathbf{X}) = \sqrt{\sum_{i=1}^{K}\sum_{j=1}^{M} (\mathbf{X}_{ij} - C_j)^2} \qquad (4.19)$$

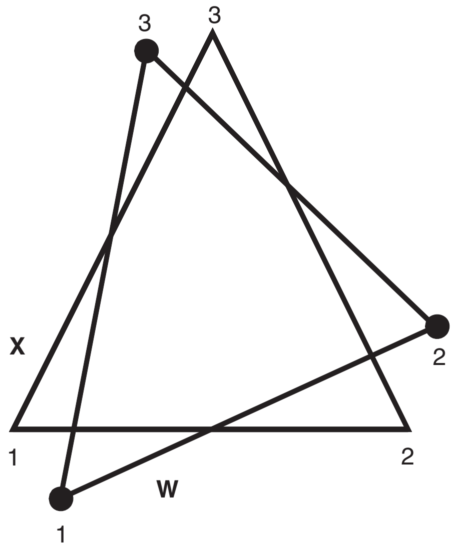


Figure 4.13 Centered triangles computed from **X** and **W**. Computation of the centroids of **X** and **W** is given by Equation 4.15; computation of the landmark coordinates after centering is given by Equations 4.16–4.18. Vertices are numbered to indicate their homology.

which is the square root of the sum of the squared distances of the landmarks from the centroid. Given that the centroids of $\mathbf{X}_{\text{centered}}$ and $\mathbf{W}_{\text{centered}}$ are both at (0, 0), we can simply sum the squared coordinates:

$$\begin{aligned} CS(\mathbf{X}_{\text{centered}}) &= \sqrt{(-1.0)^2 + (-0.667)^2 + (1.0)^2 + (-0.667)^2 + (0)^2 + (1.333)^2} \\ &= 2.160 \end{aligned} \tag{4.20}$$

$$\begin{aligned} CS(\mathbf{W}_{\text{centered}}) &= \sqrt{(-0.837)^2 + (1.127)^2 + (1.193)^2 + (-0.207)^2 + (-0.357)^2 + (1.333)^2} \\ &= 2.311 \end{aligned} \tag{4.21}$$

Dividing each coordinate of the centered triangle by its centroid size produces the pre-shape matrices:

$$\mathbf{X}_{\text{pre-shape}} = \frac{1}{2.160}\begin{bmatrix} -1 & -0.667 \\ 1 & -0.667 \\ 0 & 1.333 \end{bmatrix} = \begin{bmatrix} -0.463 & -0.309 \\ 0.463 & -0.309 \\ 0.000 & 0.617 \end{bmatrix} \tag{4.22}$$

$$\mathbf{W}_{\text{pre-shape}} = \frac{1}{2.311}\begin{bmatrix} -0.837 & -1.127 \\ 1.193 & -0.207 \\ -0.357 & 1.333 \end{bmatrix} = \begin{bmatrix} -0.362 & -0.488 \\ 0.516 & -0.089 \\ -0.154 & 0.577 \end{bmatrix} \tag{4.23}$$

These centered and scaled triangles are shown in Figure 4.14.

Because size differences do not contribute to the differences between $\mathbf{X}_{\text{pre-shape}}$ and $\mathbf{W}_{\text{pre-shape}}$, another degree of freedom has been lost (this is the third degree of freedom

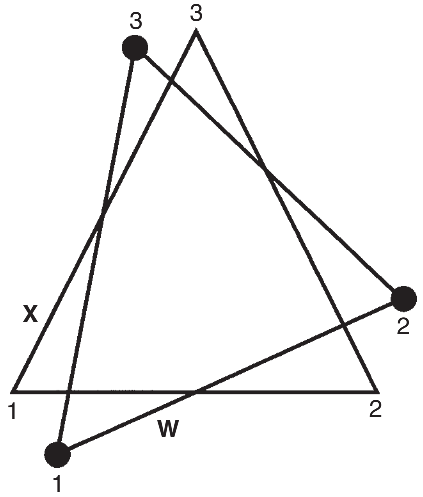


Figure 4.14 Centered triangles from Figure 4.13, scaled to unit centroid size. Computation of centroid size is given in Equations 4.19–4.21. Computation of landmark coordinates after scaling is given by Equations 4.22 and 4.23.

lost). In other words, size is no longer a dimension of possible variation; configurations that differ only in size are considered equivalent. After subtracting the three degrees of freedom representing differences in location and centroid size, we are left with three degrees of freedom to describe differences among triangle pre-shapes – triangles that are centered and scaled to unit centroid size. Accordingly, the pre-shape space of triangles is a three-dimensional space. As explained above, it is the three-dimensional surface of a four-dimensional hypersphere, so it is not an easy space to visualize or illustrate.

To make the transition from pre-shape space to shape space, we begin by choosing one shape and placing it in a convenient orientation; this configuration will be the reference. For this demonstration it is convenient to use **X** in the orientation shown in the last few figures. Choosing **X** as the reference means that **W** will be the target, so the next step is to rotate **W**, in the plane of the page around its centroid through some angle (θ). The rotation places it in the orientation that minimizes the difference between the two sets of landmark coordinates (Figure 4.15). After the rotation, the X- and Y-coordinates of each landmark will be mapped to the new coordinates ($X\cos\theta - Y\sin\theta$), ($X\sin\theta + Y\cos\theta$). Thus, the rotated form of $\mathbf{W}_{\text{pre-shape}}$ will be:

$$\mathbf{W}_{\text{pre-shape, rotated}} = \begin{bmatrix} (-0.362\cos\theta) - (-0.488\sin\theta) & (-0.362\sin\theta) + (-0.488\cos\theta) \\ (0.516\cos\theta) - (-0.089\sin\theta) & (0.516\sin\theta) + (-0.089\cos\theta) \\ (-0.154\cos\theta) - (0.577\sin\theta) & (-0.154\sin\theta) + (0.577\cos\theta) \end{bmatrix} \tag{4.24}$$

Before we can pick the value of θ that will minimize the difference between the reference ($\mathbf{X}_{\text{pre-shape}}$) and the rotated target ($\mathbf{W}_{\text{pre-shape, rotated}}$), we need a criterion to define what is being minimized. The criterion that leads to the shape space discussed earlier is minimization of the square root of the sum of the squared distances between the corresponding

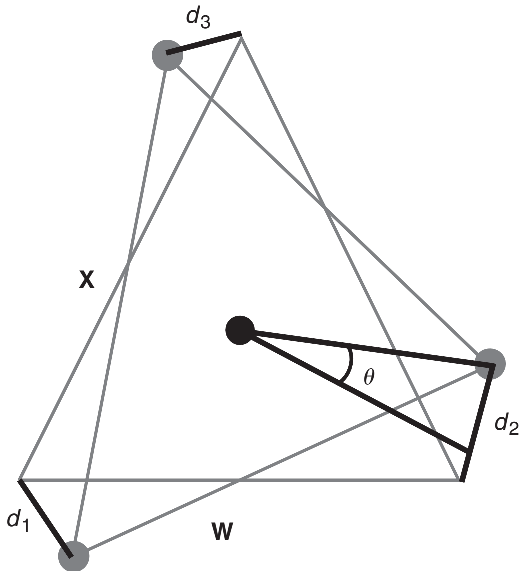


Figure 4.15 Optimal alignment of **W** to **X** will be achieved by rotating **W** around its centroid through an unknown angle θ to minimize the square root of the sum of the squares of distances d_1, d_2, and d_3.

landmarks (the distances d_1, d_2, and d_3 shown in Figure 4.15). This quantity can be computed directly from the squared differences between the corresponding coordinates of the landmarks:

$$D = \sqrt{(X_{11} - X_{21})^2 + (Y_{11} - Y_{21})^2 + \cdots + (X_{13} - X_{23})^2 + (Y_{13} - Y_{23})^2} \qquad (\mathbf{4.25})$$

(There are other criteria that lead to other superimpositions of the two triangles; one is discussed below, others in Chapter 5.)

With this criterion in hand, we can solve for the unique value of θ at which D is minimized. In our example, that value is $\theta = -19.2°$. When we insert this value into the matrix for $\mathbf{W}_{\text{pre-shape, rotated}}$ (Equation 4.22), we get:

$$\mathbf{W}_{\text{pre-shape, rotated}} = \begin{bmatrix} -0.502 & -0.341 \\ 0.458 & -0.254 \\ 0.044 & 0.596 \end{bmatrix} \qquad (\mathbf{4.26})$$

Under the conditions set out above, this is the optimal alignment to the reference form:

$$\mathbf{X}_{\text{pre-shape}} = \begin{bmatrix} -0.463 & -0.309 \\ 0.463 & -0.309 \\ 0.000 & 0.617 \end{bmatrix} \qquad (\mathbf{4.27})$$

Figure 4.16 shows the two triangles under these conditions.

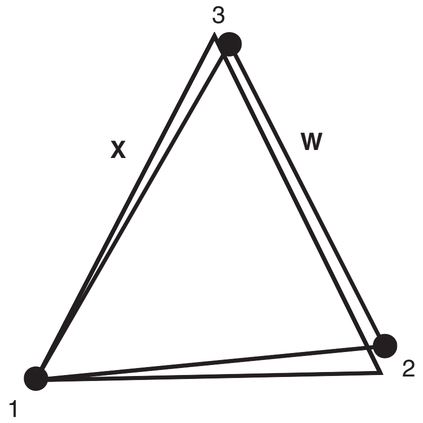


Figure 4.16 Triangles **X** and **W** after rotation of **W** to minimize the Procrustes distance. Computation of the landmark coordinates of **W** after rotation is given in Equation 4.24; the result is given in Equation 4.26. Vertices are numbered to indicate their homology.

The distance minimized above is the partial Procrustes distance, so we will label it D_p from this point forward. The value of D_p in this particular case is:

$$\begin{aligned} D_p &= [(-0.502-(-0.463))^2+(-0.341-(-0.309))^2 \\ &+ (0.458-0.463)^2+(-0.254-(-0.309))^2 \\ &+ (0.044-0)^2+(0.596-0.617)^2]^{½} \\ &= 0.089 \end{aligned} \qquad (\mathbf{4.28})$$

This is the minimum length of the chord connecting the pre-shape fibers of **X** and **W** in the pre-shape space of triangles. Because **W** is superimposed to meet the criterion of minimizing the partial Procrustes distance, $\mathbf{W}_{\text{pre-shape, rotated}}$ is said to be in *partial Procrustes superimposition* on the reference form $\mathbf{X}_{\text{pre-shape}}$. We can solve for the Procrustes distance, the arc length across the surface between $\mathbf{X}_{\text{pre-shape}}$ and $\mathbf{W}_{\text{pre-shape, rotated}}$, because the radius of the hypersphere is constrained to be one. The perpendicular from the chord to the center of the hypersphere bisects the angle ρ (Figure 4.17), which has the same value (in radians) as the arc length. Thus, there is a very simple relationship between D_p and ρ; specifically, $\rho = 2\arcsin(D_p/2)$. In our example, D_p and ρ are so small they cannot be distinguished with fewer than 4 decimal places (0.08941 and 0.08943, respectively), which is not surprising given that ρ represents a very small angle of just 5.1°.

Because rotational effects do not contribute to the differences between $\mathbf{X}_{\text{pre-shape}}$ and $\mathbf{W}_{\text{pre-shape, rotated}}$, another degree of freedom has been lost (the fourth). Rotation, or orientation, is no longer a dimension of possible variation; configurations that differ only by rotation are considered equivalent. After subtracting the four degrees of freedom representing differences in location and centroid size and rotation, we are left with two degrees of freedom to describe differences among triangles. Accordingly, the shape space of triangles is a two-dimensional space. As explained above, it is the two-dimensional surface of a three-dimensional sphere, and is a relatively easy space to visualize or illustrate.

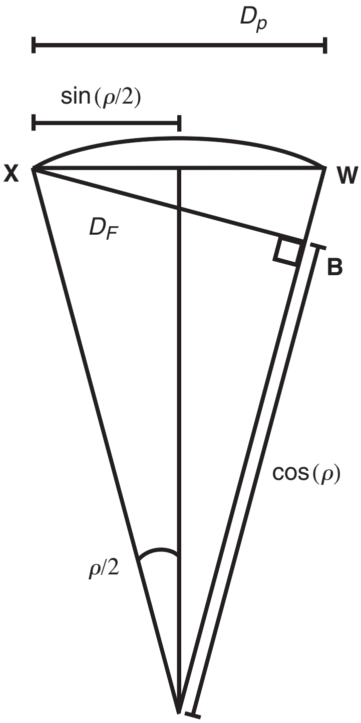


Figure 4.17 The relationships among the Procrustes distance, ρ, full Procrustes distance $D_F = \sin(\rho)$, and partial Procrustes distance $D_p = 2\sin(\rho/2)$. The configuration at point B represents a triangle in Kendall's shape space.

$\mathbf{X}_{\text{pre-shape}}$ and $\mathbf{W}_{\text{pre-shape, rotated}}$ are configurations in a shape space, but they are not yet in Kendall's shape space. To make this final transition, we need to solve for the centroid size that would further reduce the distance between the shapes **X** and **W**; we are taking **W** to **B** (Figure 4.17). As indicated in Figure 4.17, that distance (D_F, the full Procrustes distance) is measured along a line segment orthogonal to the radius of $\mathbf{W}_{\text{pre-shape, rotated}}$, passing through $\mathbf{X}_{\text{pre-shape}}$. In our example, ρ is small (0.0894 radians); its cosine is near one (0.996) so we need make only a very slight adjustment to convert the coordinates of $\mathbf{W}_{\text{pre-shape, rotated}}$ to $\mathbf{W}_{\text{shape}}$:

$$\mathbf{W}_{\text{shape}} = \cos(0.089)\begin{bmatrix} -0.5021 & -0.3414 \\ 0.4583 & -0.2542 \\ 0.0439 & 0.5956 \end{bmatrix} = \begin{bmatrix} -0.5001 & -0.3401 \\ 0.4564 & -0.2532 \\ 0.0437 & 0.5932 \end{bmatrix} \qquad (\mathbf{4.29})$$

This is the triangle with the same shape as **W**, but it is now in Kendall's shape space with the reference at triangle $\mathbf{X}_{\text{pre-shape}}$. Because the full Procrustes distance was used to determine the coordinates of the landmarks in $\mathbf{W}_{\text{shape}}$, we can say that $\mathbf{W}_{\text{shape}}$ is in full Procrustes superimposition on the reference form $\mathbf{X}_{\text{pre-shape}}$.

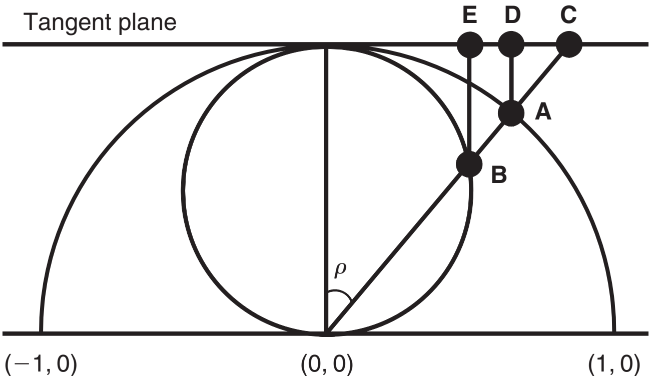


Figure 4.18 Tangent space to shape spaces of triangles and projections onto the tangent space. As in Figure 4.11, the outer hemisphere is a section through the space of centered and aligned shapes scaled to unit centroid size, and the inner circle is a section through Kendall's shape space of centered and aligned shapes scaled to $\cos(\rho)$. The plane is tangent to the sphere and the hemisphere at the point of the reference shape. The configuration at point **B** represents a triangle in Kendall's shape space; **A** is the same shape scaled to unit centroid size. **C** is a stereographic projection of **B** onto the tangent plane. **D** is the orthogonal projection of **A** onto the tangent plane, and **E** is the orthogonal projection of **B** onto the tangent plane.

Tangent spaces

As mentioned earlier in this chapter, the mathematics of statistical inference in Kendall's shape space has been developed by Kendall and others. However, the simple fact remains that the curvature of shape space makes statistical inference more difficult in this space than it is in Euclidean spaces. In addition, most of the familiar methods of multivariate statistical analysis assume a Euclidean space. Therefore, in this section we discuss the replacement of Kendall's shape space with a Euclidean approximation.

The problem of replacing a curved space with a Euclidean approximation is illustrated for the special case of triangles in Figure 4.18. As before (see Figure 4.11), the outer hemisphere is the space constructed by aligning pre-shapes (with centroid size fixed at one) to minimize the partial Procrustes distance (the square root of the summed squared distances between corresponding landmarks). The inner sphere is Kendall's shape space, constructed by scaling the aligned target shapes to centroid size $= \cos(\rho)$. These two spaces share a common point, the reference shape, because the distance of the reference from itself is zero, so $\cos(\rho)$ is one. Tangent to both of these spaces, at the reference shape, is a Euclidean plane. We also need to decide how we will construct the projection of shapes onto the tangent plane, which includes deciding (1) which space will be the source of the configurations projected onto the tangent plane, and (2) what rule we will use to determine the direction of the projection. (We also need to decide how to choose an appropriate reference configuration to serve as the tangent point, which is discussed in the next section.)

Figure 4.18 illustrates two common approaches to projecting from one space onto another. One approach is to project to the new space from the centroid of some reference space. In this case, the reference space is the hemisphere of aligned pre-shapes, so the projections are along the radii of this hemisphere to the tangent space. In this stereographic projection, the shape represented by points **B** and **A** (at centroid sizes cos(ρ) and one, respectively) map to the same location (**C**) in the tangent space. The distance in the plane from the reference to **C** is *greater* than the arc length from the reference to **B** (the Procrustes distance); and the discrepancy between these distances increases as ρ increases and the distance in the tangent plane approaches infinity. The other approach to projecting from one space onto another is to project along lines that are orthogonal to the new space. Point **E** represents the orthogonal projection of **B** onto the tangent plane, and this projection produces distances from the reference in the tangent plane that are *less* than the Procrustes distance. As in the stereographic projection, the magnitude of the discrepancy between the distances increases as ρ increases, but in the orthogonal projections, distances in the tangent plane asymptotically approach the maximum equal to the radius of the shape space.

In the stereographic projection it does not matter whether the projection to the tangent plane is from the hemisphere of triangles in partial Procrustes superimposition, or from the sphere of triangles in full Procrustes superimposition. Both target configurations project to the same point in the tangent space. In the orthogonal projection, it does matter whether the projection from the tangent plane is from the outer or inner hemisphere. The projection from the hemisphere produces distances in the tangent plane that depart less from the Procrustes distance (the arc length) and are closer to the partial Procrustes distance (the chord length). Projection from the sphere produces distances that depart more from the Procrustes distance and are closer to the full Procrustes distance. Furthermore, the projections from the hemisphere of triangles in partial Procrustes superimposition have a higher maximum distance from the reference (one instead of one-half), and approach it more slowly.

In the simple example given in the previous section, we demonstrated that the differences between the Procrustes, partial Procrustes and full Procrustes distances from the reference become negligible as ρ approaches zero. Similarly, the differences among the stereographic and orthogonal projects also become negligible as ρ approaches zero.

Selecting the reference configuration

Many of the steps involved in placing target configurations in shape space, or in the Euclidean space tangent to it, are functions of the reference shape. For example, in the construction of a shape space, each target configuration is rotated to the orientation that minimizes its distance from the reference. Also, in the construction of Kendall's shape space, the scaling of each target configuration is a function of its distance from the reference. Moreover, the tangent space is tangent to shape space at the reference. Perhaps most important, the discrepancies between distances in the tangent space and those in shape space increase as a function of distance between target and reference. Thus the choice of reference can have important consequences.

Most interesting biological questions will be concerned with differences among more than two specimens. The inferences based on analyses of multiple specimens will be based on all of the distances among specimens, not just their distances from the reference. Accordingly, the choice of a reference must consider the effects of that choice on approximating distances among target specimens, not just distances of target specimens from the reference.

Not only will distances from the reference be distorted, so too will the distances among target specimens, and this distortion will also be a function of their distances from the reference. If these distortions are large, inferences based on distances in the Euclidean tangent space will be unreliable.

One possible reference is the average shape of the entire sample (computed using methods discussed in Chapter 5). This approach has the advantage that it minimizes the average distance from the reference, which minimizes the average distortions of interspecimen distances projected to the tangent plane (Bookstein, 1996; Rohlf, 1998). An alternative choice of reference is a shape inferred to represent the starting point of some biological process (e.g. a neonate in a study of ontogenetic transformation – *cf.* Zelditch et al., 1992). This approach has the advantage that the difference between target and reference can be interpreted as a biological transformation as well as a mathematical transformation (Fink and Zelditch, 1995; Zelditch et al., 1998). However, as Rohlf (1998) points out, this approach can have the limitation that the reference is at one extreme of the observed distribution of shapes, thereby increasing the risk of substantial distortions of distances when changes in shape are large. Conceivably, erroneous inferences could be drawn from the analysis. However, Marcus et al. (2000) analyzed differences in skull shape among representatives of several mammalian orders and found that most Procrustes distances are closely approximated by the Euclidean distance in the tangent space. The principal exceptions were the distances from terrestrial taxa (especially the muskrat) to a dolphin (which is not surprising, given the extraordinary reorganization of the cetacean head). This result suggests that most biologists are unlikely to encounter any cases in which the differences among specimens are large enough to worry about the adequacy of the linear approximations. It is unlikely that distances in the tangent space (based on *any* reference) will poorly approximate distances in shape space. Even so, using the average shape of all specimens in the data minimizes the risk that such a problem will occur. The use of any other reference carries with it the responsibility to ensure that Euclidean distances in the tangent space are accurate approximations of the distances in shape space.

Dimensions and degrees of freedom

The issue of degrees of freedom (or the number of independent measurements in a system) is important for statistical analyses, but it can be confusing, especially when talking about shape. To clarify it, we can consider a simple example. Suppose we wish to describe the location of a notebook in a room. We could give its location in terms of three distances from a reference point (such as the corner of the door of the room), and this is equivalent to defining its position by three Cartesian coordinates relative to that reference point. In this example, there are three degrees of freedom for the location of the notebook because three variables are required to describe it. Knowing those variables and the reference suffices to find the notebook. However, if the notebook is on a chair, and all chairs are known to be the same height, specifying the height conveys no more information than saying that the notebook is on a chair. Knowing what we do about the chairs, we only need two additional pieces of information, the X- and Y-coordinates, to specify the location of the notebook in the room. Thus by specifying the constraint that the notebook is on a chair of fixed height, we have removed one of the three degrees of freedom.

We can take this example a step further by specifying that all the chairs are located along walls of the room, with every chair touching the wall. Now, the X- and Y-coordinates can be replaced by the distance (L) around the perimeter of the room from the door to the notebook, and the direction of the measurement (clockwise or counter-clockwise). If we agree that distances around a perimeter are always measured in the same direction, then the value of L is sufficient to describe the location of the notebook. The additional constraints (chairs against the wall, perimeter measured in clockwise direction) have reduced the degrees of freedom from two (X and Y) to one (L). We have not actually eliminated either X or Y; rather, we have merely replaced that pair by L. Nor have we lost any information; given L, and the direction in which L is measured, as well as the height of the chairs, we can reconstruct the original three Cartesian coordinates (X, Y, and Z) of the notebook.

In the case of two-dimensional shapes, we start out with K landmarks in two dimensions, so we have $2K$ coordinates, which constitute $2K$ independent measurements (because each coordinate is independent of the others, in principle). In the course of superimposing the shapes on the reference form, we perform three operations: (1) we center the matrix on the centroid, thereby losing two degrees of freedom; (2) we set centroid size to one, thereby losing another; and (3) we compute the angle through which to rotate the specimen, thereby losing one more. By the end, we have lost four degrees of freedom as a consequence of applying these constraints to the data. However, unlike the notebook example, we still have $2K$ variable coordinates in our data matrix; none of them have been removed or constrained. We have not lost degrees of freedom by removing coordinates, because the loss of degrees of freedom is shared by *all* coordinates – each coordinate has lost some fraction of a degree of freedom because each is partially constrained by the operations of centering, scaling and rotation. Consequently, we have too many variable coordinates for the degrees of freedom. The primary advantage of the thin-plate spline methods (discussed in Chapter 6) is that we can work with $2K-4$ variables, so that the number of variables and the number of degrees of freedom are the same.

Summary

Because there are several different morphometric spaces and distances, some with only slightly different names, we summarize them below.

The *configuration space* is the set of all matrices representing landmark configurations that have the same number of landmarks and coordinates. This space has $K \times M$ dimensions, where K is the number of landmarks and M is the number of coordinates.

The *pre-shape space* is the set of all $K \times M$ configurations with a centroid size of one, centered at the origin. This space is the surface of a hypersphere of radius one. Because of the centering, configurations that differ only in position are represented as the same point in pre-shape space. Similarly, because of the scaling, configurations that differ only in centroid size are represented by the same point in pre-shape space. Consequently, this space has $KM-(M+1)$ dimensions; M dimensions are lost due to centering, and one dimension is lost due to scaling. In pre-shape space, the set of all configurations that may be converted into one another by rotation lies along a circular arc called a *fiber*, which lies on the surface of the pre-shape hypersphere. The distance between shapes in pre-shape space is the length of the shortest arc across the surface connecting the fibers representing those shapes, and is called the *Procrustes distance*. Because the radius of the pre-shape hypersphere is one, the length of the arc is also the value (in radians) of the angle subtended (ρ).

To construct a *shape space*, we select one point on each fiber, removing differences in rotation. The number of axes on which a configuration can be rotated is a function of the number of landmark coordinates: $M(M-1)/2$. This also specifies the number of dimensions that are lost in the transition from pre-shape space to shape space (1 if $M=2$, 3 if $M=3$). The construction of a shape space begins with the selection of one shape in a convenient orientation to serve as the *reference* configuration. Every other shape (called a *target* configuration) is placed in the orientation that corresponds to the location on its fiber that is closest to the reference. This orientation is the position that minimizes the square root of the sum of the squared differences between the coordinates of corresponding landmarks. When minimized simply by rotation, this quantity is called the *partial Procrustes distance*. Configurations that satisfy this condition are said to be in *partial Procrustes superimposition* on the reference. The partial Procrustes distance is the length of the chord of the arc between the fibers in pre-shape space.

After rotation to partial Procrustes superimposition, the square root of the sum of the squared differences between the coordinates of corresponding landmarks can be further reduced by rescaling the target to centroid size of $\cos(\rho)$. Configurations that satisfy this condition are said to be in *full Procrustes superimposition* on the reference; and the resulting distance between shapes (square root of the sum of the squared differences between the coordinates of corresponding landmarks) is the *full Procrustes distance*. The set of shapes in full Procrustes superimposition comprises a hypersphere of radius one-half, inside the hypersphere of shapes in partial Procrustes superimposition, and tangent to the larger hypersphere at the reference. This smaller, inner hypersphere is *Kendall's shape space*.

Problems and exercises on the theory of shape

To get a feel for what the software will be doing for you, do these problems and exercises using pencil, paper and a scientific calculator. The numbers of landmarks are small, to keep the level of tedium to a minimum!

1. Suppose that the configuration matrix for a given shape is:

$$\mathbf{A} = \begin{bmatrix} 0.0 & -1.0 \\ 0.0 & 0.5 \\ 0.7 & -0.2 \end{bmatrix}$$

   a. How many landmarks are there in this configuration? How many dimensions does it have?
   b. Sketch the shape representing this configuration (you may want to use graph paper, if it helps). Number the landmarks.
   c. Write out the row vector form of this landmark configuration.
   d. Find the centroid position of this landmark configuration. How many coordinates are in the centroid position? Sketch the location of the centroid on your picture from (b) above.

e. Write out the centered form of this configuration matrix, by subtracting the value of the X-coordinate of the centroid from each of the values in the first column, and subtracting the value of the Y-coordinate of the centroid from each of the values in the second column.
f. Find the centroid size of this landmark configuration.
g. Now form the pre-shape configuration for **A**. Do this by dividing the centered form of this matrix (solution to (e) above) by the centroid size (solution to (f) above). Remember that when you divide a matrix by a scalar (an ordinary number, like centroid size), you must divide each value in the matrix by the scalar divisor (which is centroid size in this case).

2. Suppose a configuration of dimensional landmarks is given by:

$$B = \{0.3, -1.0, 0.25, -0.4, 0.0, 0.75, -0.2, 0.35\}$$

a. How many landmarks are in this configuration?
b. Write out the configuration matrix for this configuration.
c. Find the centroid for this configuration.
d. Find the centroid size for this configuration.

3. Given the landmark configuration:

$$C = \{0.1, 0.1, 0.1, 0.3, -1.0, 1.1, -0.6, -0.3, 0.2, 0.3, -0.1, 0.15\}$$

can you determine what K and M are?

4. Suppose we have the configuration matrix:

$$\mathbf{X} = \begin{bmatrix} 0.5 & 0.5 \\ -0.2 & 0.3 \\ 0.1 & 0.3 \end{bmatrix}$$

a. Compute a configuration matrix that would represent **X** in pre-shape space.
b. Now, for the truly stout of heart, suppose we have a second configuration matrix in pre-shape space:

$$\mathbf{Y} = \begin{bmatrix} 0.6864 & 0.1961 \\ -0.6864 & -0.0981 \\ 0.0 & -0.0981 \end{bmatrix}$$

Determine the angle that you would have to rotate the pre-shape space matrix form of **X** (from (a) above) to produce a partial Procrustes superposition of **X** on the reference form **Y**.

5. Given two matrices:

$$\mathbf{X} = \begin{bmatrix} 0.7146 & 0.2150 \\ -0.6438 & -0.0913 \\ -0.0709 & -0.1237 \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} 0.6864 & 0.1961 \\ -0.6864 & -0.0981 \\ 0.0 & -0.0981 \end{bmatrix}$$

where **X** is in partial Procrustes superposition with **Y**:

a. Find the partial Procrustes distance between the two.

b. Use the partial Procrustes distance to find the Procrustes distance between the two.

c. Use the Procrustes distance to calculate the full Procrustes distance between the two.

Answers to problems and exercises

(A full solution is given if the calculation has not been seen before.)

1. Looking at the configuration matrix:

a. There are three landmarks ($K = 3$ rows) and each is in two dimensions ($M = 2$ columns).

b. See Figure 4.19.

c. In row form, $\mathbf{A} = \{0.0, -1.0, 0.0, 0.5, 0.7, -0.2\}$.

d. The centroid is located at (0.2333, −0.2333), or $X = 0.2333$, $Y = -0.233$. The centroid position is calculated:

$$X_C = \frac{(0 + 0 + 0.7)}{3} = 0.2333$$

$$Y_C = \frac{(-1 + 0.5 - 0.2)}{3} = -0.2333$$

Figure 4.20 shows the location of the centroid.

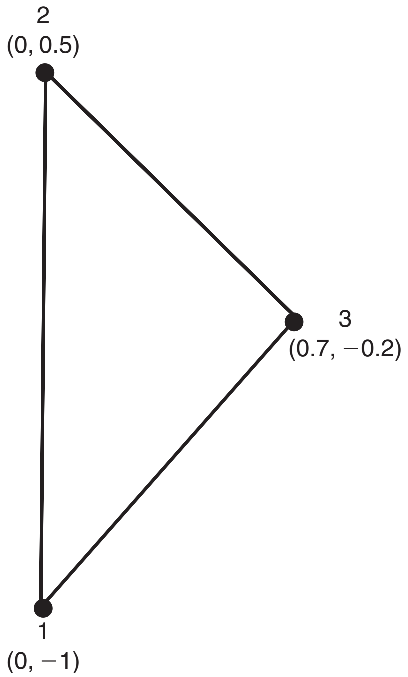


Figure 4.19 Answer to Exercise 4.1b.

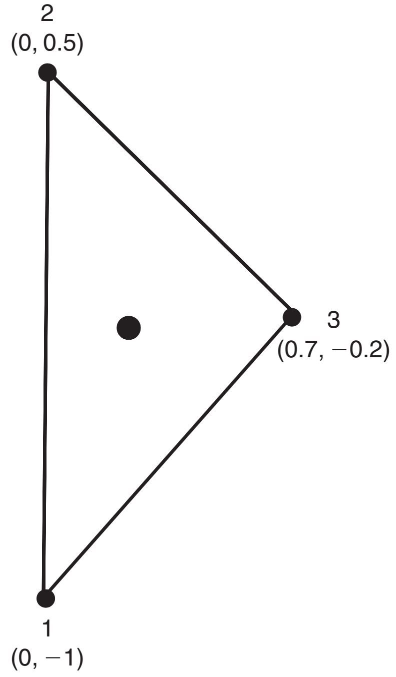


Figure 4.20 Answer to Exercise 4.1d.

e. We simply subtract the X-coordinate of the centroid (0.2333) from the first column of **A**, and subtract the Y-coordinate of the centroid (−0.2333) from the second column. This leaves us with:

$$\begin{bmatrix} -0.2333 & -0.7667 \\ -0.2333 & 0.7333 \\ 0.4667 & 0.0333 \end{bmatrix}$$

Note that if you add up the values in the first column you get zero, which is also true for the second column. Thus the centroid position of the centered matrix is (0, 0).

f. The centroid size is 1.2055. The centroid size is the square root of the summed squared distances of the landmarks from the centroid, which is:

$$\begin{aligned} CS = \{&(0-0.2333)^2 + (-1-(-0.2333))^2 + (0-0.2333)^2 \\ &+ (0.5-(-0.2333))^2 + (0.7-0.2333)^2 \\ &+ (-0.2-(-0.2333))^2\}^{½} \\ = {}& 1.2055 \end{aligned}$$

An easier approach is to use the centered form of the configuration matrix (with the centroid set to zero). With this form, we can take the square root of the summed squared coordinates of the landmarks:

$$\begin{aligned}CS = &\ \{(-0.2333)^2 + (-0.7667)^2 + (-0.2333)^2 \\ &+ (0.7333)^2 + (0.4667)^2 + (0.0333)^2\}^{½} \\ = &\ 1.2055\end{aligned}$$

g. The resulting pre-shape space configuration is

$$\mathrm{A}_{\text{pre-shape}} = \begin{bmatrix} -0.1936 & -0.0630 \\ -0.1936 & 0.6083 \\ 0.3871 & 0.0277 \end{bmatrix}$$

Note that the entries are identical to the centered matrix values (see (e) above) divided by 1.2055.

2. Looking at the configuration:
   a. There are four landmarks.
   b. The configuration matrix is:

$$\begin{bmatrix} 0.3 & -1.0 \\ 0.25 & -0.4 \\ 0.0 & 0.75 \\ -0.2 & 0.35 \end{bmatrix}$$

   c. The centroid is located at $X_C = 0.0875$, $Y_C = -0.075$.
   d. The centroid size $CS = 1.4087$.
3. No! This might be $K = 6$ and $M = 2$ (a two-dimensional system), or $K = 4$, $M = 3$ (a three-dimensional system). If the data are in a row format, you cannot tell the value of K or M from looking at it.
4. Looking at the configuration matrix:
   a. The pre-shape space form of **X** is

$$\begin{bmatrix} 0.7013 & 0.2550 \\ -0.6376 & -0.1275 \\ -0.0638 & -0.1275 \end{bmatrix}$$

   b. The triangles can be iteratively rotated, or Equation 4.13 can be used:

$$\theta = \text{arctangent}\left(\frac{\sum_{j=1}^{K} Y_{Rj}X_{Tj} - X_{Rj}Y_{Tj}}{\sum_{j=1}^{K} X_{Rj}X_{Tj} + Y_{Rj}Y_{Tj}}\right) \qquad \textbf{(4.13)}$$

Substituting the appropriate values of X and Y for the reference (R) and target (T) yields:

$$\begin{aligned}\theta = \tan^{-1}\{&(0.1961 \times 0.7013 + (-0.0981) \times (-0.6376) + (-0.0981)\\ &\times (-0.0638) + (-0.6864) \times 0.2550 + (-0.6864) \times (-0.1275)\\ &+ 0 \times (-0.1275))/(0.6864 \times 0.7013 + (-0.6864) \times (-0.6376)\\ &+ 0 \times (0.0638) + 0.1961 \times 0.2550 + (-0.0981) \times (-0.1275)\\ &+ (-0.0981) \times (-0.1275))\}\end{aligned}$$

$$\theta = -0.0562 \text{ radians} = -3.2175^\circ$$

5. Looking at the matrices:
   a. To find the partial Procrustes distance between the two, we take the square root of the summed squared differences in the landmark coordinates:

$$\begin{aligned}D_p = \{&(0.7146 - 0.6864)^2 + (0.2150 - 0.1962)^2\\ &+ (-0.6438 - (-0.6864))^2 + (-0.0913 - (-0.0981))^2\\ &+ (-0.0709 - 0)^2 + (-0.1237 - (-0.0981))^2\}^{1/2}\\ = {}&0.0933\end{aligned}$$

   b. Because $\rho = 2\ \arcsin(D_p)$, $\rho = 2\ \arcsin(0.0933/2) = 0.0933$ radians; the two are equal through three decimal places.
   c. $D_F =\ \sin(\rho)$, so $D_F = \sin(0.0933) = 0.0932$.

References

Bookstein, F. L. (1996). Combining the tools of geometric morphometrics. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 131–151. Plenum Press.

Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*. John Wiley & Sons.

Fink, W. L. and Zelditch, M. L. (1995). Phylogenetic analysis of ontogenetic shape transformations: a reassessment of the piranha genus *Pygocentrus* (Teleostei). *Systematic Biology*, **44**, 343–360.

Kendall, D. (1977). The diffusion of shape. *Advances in Applied Probability*, **9**, 428–430.

Marcus, L. F., Hingst-Zaher, E. and Zaher, H. (2000). Applications of landmark morphometrics to skulls representing the orders of living mammals. *Hystrix* (n.s.), **11**, 24–48.

Rohlf, F. J. (1998). On applications of geometric morphometrics to studies of ontogeny and phylogeny. *Systematic Biology*, **47**, 147–158.

Rohlf, F. J. (2000). On the use of shape spaces to compare morphometric methods. *Hystrix* (n.s.), **11**, 8–24.

Slice, D. E. (2001). Landmark coordinates aligned by Procrustes analysis do not lie in Kendall's shape space. *Systematic Biology*, **50**, 141–149.

Small, C. G. (1996). *The Statistical Theory of Shape*. Springer.

Zelditch, M. L., Bookstein, F. L. and Lundrigan, B. L. (1992). Ontogeny of integrated skull growth in the cotton rat *Sigmodon fulviventer*. *Evolution*, **46**, 1164–1180.

Zelditch, M. L., Fink, W. L., Swiderski, D. L. and Lundrigan, B. L. (1998). On applications of geometric morphometrics to studies of ontogeny and phylogeny: a reply to Rohlf. *Systematic Biology*, **47**, 159–167.

5

Superimposition methods

In Chapter 3 we presented a simple method for obtaining shape variables: the two-point registration that produces Bookstein's shape coordinates. We began with this method because it is especially easy to understand, even without knowing any of the theory developed in Chapter 4. Having covered that theory, we now can develop a more general approach to the problem of matching up shapes prior to comparison. This matching is termed "superimposition" because the landmark configurations are placed on top of each other (by the mathematical operations that do not alter shape, i.e. translation, scaling, and rotation). Several superimposition methods are available; they differ in how these operations are applied. The objective of this chapter is to explain some of these superimposition methods, compare them, and discuss their relative advantages and disadvantages. We also discuss in some detail the issue of interpreting the pictures of superimposed landmarks. At first sight different methods may appear to suggest different interpretations of the shape differences, but to a large extent the differences are illusory – the pictures might look different, but they have the same meaning.

Before discussing the alternative methods of superimposition, we first explain why we would even want an alternative to Bookstein's shape coordinates (BC). We then describe a superimposition method that is based on a very similar approach. This method, called sliding baseline registration (SBR), also involves a two-point registration, but the two points are not entirely fixed (they are allowed to "slide" along one axis). Next, we present the most widely used method, Procrustes generalized least squares (GLS), followed by an alternative that is similar to it in some respects (variously called Procrustes resistant fit, or resistant fit theta-rho analysis, RFTRA). After presenting all of these methods, we summarize their similarities and differences, discuss the interpretation of their graphical results, and conclude with recommendations regarding their uses.

Why we want an alternative to Bookstein shape coordinates

Recall that Bookstein's shape coordinates (BC) are obtained by the two-point registration method (Chapter 3). This procedure fixes the coordinates of two points at (0, 0) and (1, 0);

Geometric Morphometrics for Biologists
ISBN 0–12–77846–08

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

thus the segment bounded by these landmarks (the baseline) has a standard orientation and length in all specimens, and the coordinates of all the other landmarks indicate their positions relative to the baseline. Fixing the baseline coordinates is the source of the principal advantages of this superimposition method, and also of the main disadvantages. The severity of the disadvantages may persuade you to use one of the other methods, even if the purposes of those methods seem a bit obscure at first.

The most notable advantages of BC arise from standardizing the coordinates of the baseline. Because four coordinates are fixed (X- and Y-coordinates of the two baseline points), the number of variable coordinates is $2K - 4$ (where K is the number of landmarks). Consequently, the number of variable coordinates is exactly the same as the number of dimensions of the shape space they occupy, which is also the number of statistical degrees

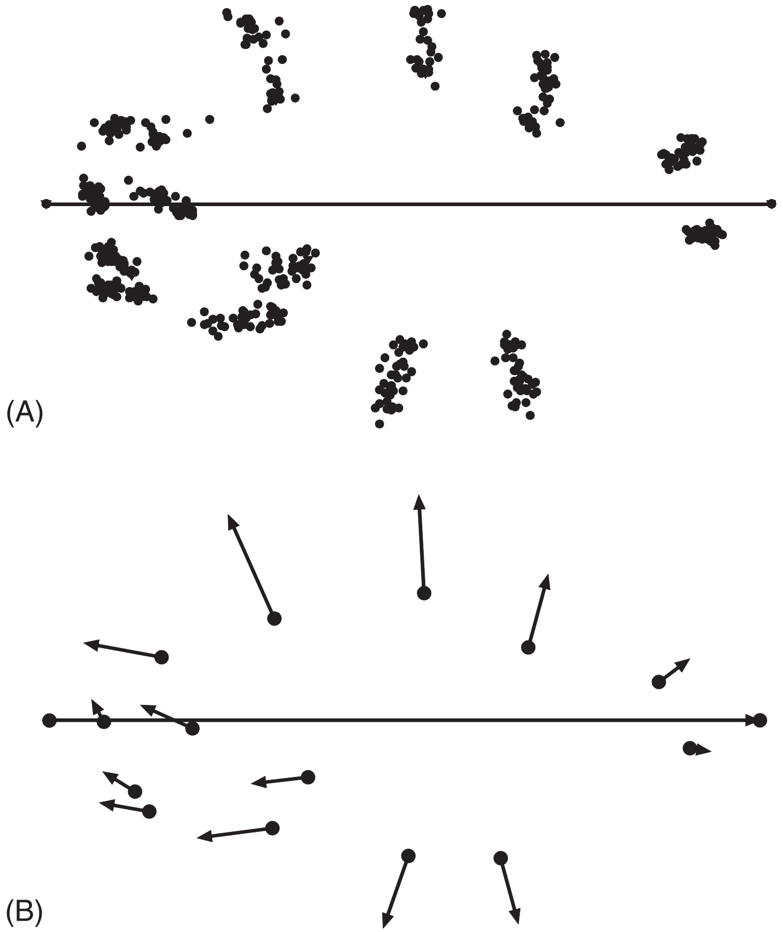


Figure 5.1 Variation in landmark positions relative to a fixed baseline for an ontogenetic series of the piranha *Serrasalmus gouldingi*: (A) coordinates of landmarks relative to a fixed baseline that extends between landmarks 1 and 7 (the tip of the snout and posterior termination of the hypural bones); (B) vectors indicating displacements of landmarks relative to the fixed baseline. (The summed squared lengths of these vectors does not equal the Procrustes distance between juvenile and adult shapes.)

of freedom. The advantage of the correspondence of these numbers is that we can perform analytic tests like Hotelling's T^2 without discarding variables. Because the four coordinates of the baseline are fixed, the orientation of the baseline is standardized. The advantage of this is that the pictures of the superimposed configurations are often quite easy to interpret, especially if the baseline is an important morphological feature like an anatomical axis.

The most notable disadvantages of BC also lie in standardizing the coordinates of the baseline. One of these disadvantages arises because there are no truly invariant landmarks; every landmark varies in location relative to all of the others. Fixing the locations of the two landmarks that serve as endpoints of the baseline means that the variance of those landmarks must be put *somewhere*. In an ontogenetic series of the piranha *Serrasalmus gouldingi*, superimposed at two landmarks near the dorsoventral midline (Figure 5.1A), we can see that the three most dorsal landmarks vary primarily along the dorsoventral axis. When we regress these coordinates on log centroid size (using methods discussed in Chapter 10), we see that these landmarks move away from the midline during growth (Figure 5.1B).

If two of the dorsal landmarks are used as the baseline, the pictures look strikingly different (Figure 5.2). The dorsoventral component has been removed from all three dorsal

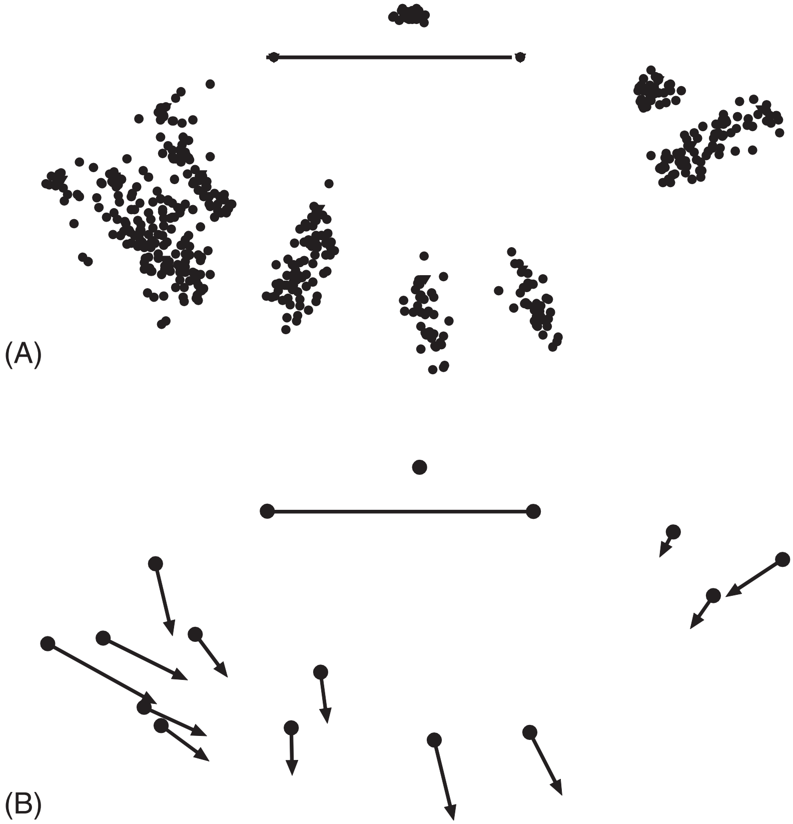


Figure 5.2 Ontogenetic variation in *S. gouldingi* visualized relative to a baseline on the dorsal body (landmarks 3 and 5): (A) coordinates of landmarks relative to the baseline; (B) vectors indicating displacements of landmarks relative to the fixed baseline. (Note the greater apparent magnitude of the variance landmarks compared to Figure 5.1.)

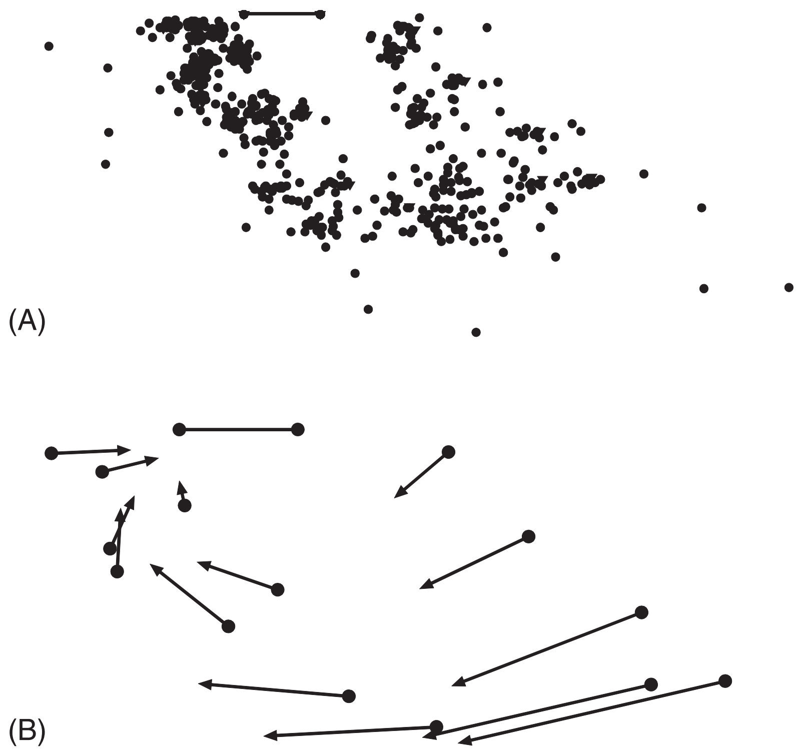


Figure 5.3 Ontogenetic variation in *S. gouldingi* visualized relative to a baseline on the dorsal head (landmarks 2 and 3): (A) coordinates of landmarks relative to the baseline; (B) vectors indicating displacements of landmarks relative to the fixed baseline. (Note the apparent rotation of landmarks around this baseline.)

landmarks (because they covary to a high degree), and is expressed as displacement of every other landmark away from the dorsal edge. In addition, the elongation of the middle of the body relative to the rest of the piranha is now expressed as a relative contraction of the ends towards the middle.

If we use a baseline that rotates relative to most of the other landmarks, the resulting superimposition seems to indicate that the piranha's body rotates around the baseline as it grows (Figure 5.3).

Clearly the baselines used in Figures 5.2 and 5.3 are spectacularly bad choices; however, they simply exaggerate the general problem that the variance of baseline points is transferred to the other landmarks. It should be intuitively obvious, even if not visibly so, that the actual anatomical landmarks are really no more variable than in Figure 5.1; changing the baseline does nothing to the data set but rotate and rescale it. The consequences for our perception of the shape differences can be dramatic, particularly when it makes the data seem inordinately noisy, but they can be understood as the consequences of a change in perspective. What makes this transfer of variance really worrisome is that it is not necessarily unbiased – it is related to the distance of the free landmarks to the baseline (Dryden and Mardia, 1998). Consequently, the transfer of variance can induce correlations among landmarks (compare the relative displacement of the most posterior landmarks across the three baseline registrations).

A second disadvantage arising from standardizing the coordinates of the baseline points is that this superimposition does not minimize the distance between configurations (the summed squared distance between corresponding landmarks). Scaling configurations of landmarks to unit baseline length need not produce configurations of the same centroid size, much less configurations of unit centroid size. Similarly, the rotation to align the baseline is unlikely to be the exact rotation needed to remove rotational effects, as prescribed by Kendall. Consequently, the summed squared distances between corresponding landmarks is not the minimized partial Procrustes distance between shapes. In other words, the configurations produced by the two-point registration do not differ solely in shape (as defined by Kendall), so the graphics based on this superimposition cannot be said truly to embody the differences between shapes. This is not merely a graphical problem; it is also a serious metrical problem. There is a profound conceptual and mathematical inconsistency between the distances measured as summed squared differences of shape coordinates, and the Procrustes distances between shapes in shape space. This discrepancy is likely to be especially large when the baseline points are close together and more variable than most other landmarks (as in Figures 5.2 and 5.3), but no choice of baseline can completely eliminate the problem.

Given these rather substantial disadvantages, it is useful to have alternative superimposition methods. The first alternative we discuss is conceptually similar to the two-point registration, but addresses the problems of scaling and variance transfer. The second alternative addresses the discrepancy between distance metrics. The third alternative is conceptually related to the second, but addresses problems arising from highly localized shape change.

Sliding baseline registration

The sliding baseline registration (SBR) was developed by David Sheets in collaboration with Mark Webster (Webster et al., 2001) and Keonho Kim (Kim et al., 2002) to reduce the disadvantages of aligning landmark configurations along one edge as in the two-point registration. To that end, configurations are scaled to unit centroid size, which directly addresses the conceptual and mathematical inconsistency between the scaling used in the superimposition and the scaling used in the definition of shape. As we will demonstrate below, this also reduces the problem of variance transfer. Because the configurations are scaled to unit centroid size, their baselines will usually differ in length and consequently, the two end points cannot be superimposed simultaneously. Instead, their *Y*-coordinates are fixed at zero and their *X*-coordinates are allowed to vary as necessary to align the *X*-coordinates of the *centroids* at the zero, in effect sliding the baseline along the *X*-axis. (The *Y*-coordinate of the centroid is the average perpendicular distance of the landmarks from the baseline after scaling to unit centroid size.)

At first glance, the SBR superimposition appears to indicate that the *S. gouldingi* ontogeny involves a smaller increase in body depth than was seen using BC (Figure 5.4). Closer examination reveals that the baseline is getting shorter as depth increases, so the increase in *relative* depth is the same (scaling to the same size makes adults, with

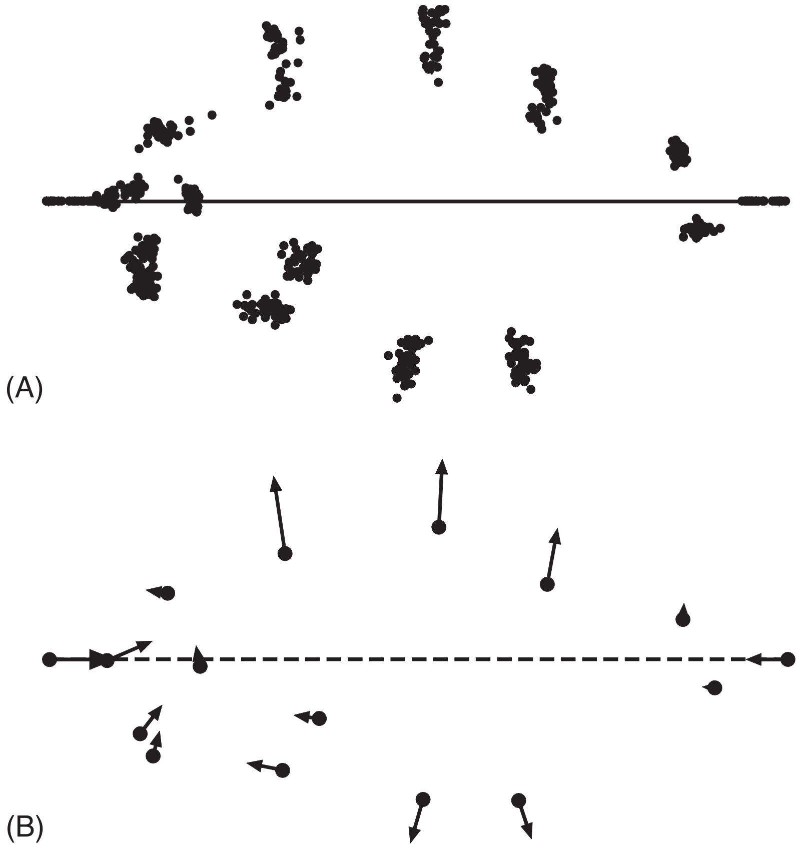


Figure 5.4 Ontogenetic variation in *S. gouldingi* visualized by sliding baseline registration. As in Figure 5.1, the baseline is formed by the two most distant landmarks, 1 and 7. (A) Coordinates of landmarks relative to the baseline; (B) vectors indicating displacements of landmarks relative to the fixed baseline. (Note the lesser apparent magnitude of the variance landmarks compared to Figure 5.1.)

relatively deeper bodies, look both shorter and deeper). This difference arises because part of the variation in the relative positions of the baseline endpoints is expressed in the *X*-coordinates of those points under the SBR superimposition. This also means that SBR transfers correspondingly less variance to the other landmarks.

Although SBR can reduce the variance that is transferred to the free landmarks, it cannot completely eliminate the problem. The tremendous deepening of the midbody of *S. gouldingi* still induces a large covariance among the other landmarks if the dorsal landmarks are used for the baseline (Figure 5.5).

When the baseline also rotates relative to the other landmarks, that rotation can be a more prominent component of the induced covariance under SBR than in BC (Figure 5.6). Therefore, SBR cannot compensate for a poor choice of baseline.

Another problem that SBR shares with BC is that the implied displacements of the landmarks still do not equal the partial Procrustes distance between the shapes. Again, the configurations are not centered on the centroid (although they are closer after SBR),

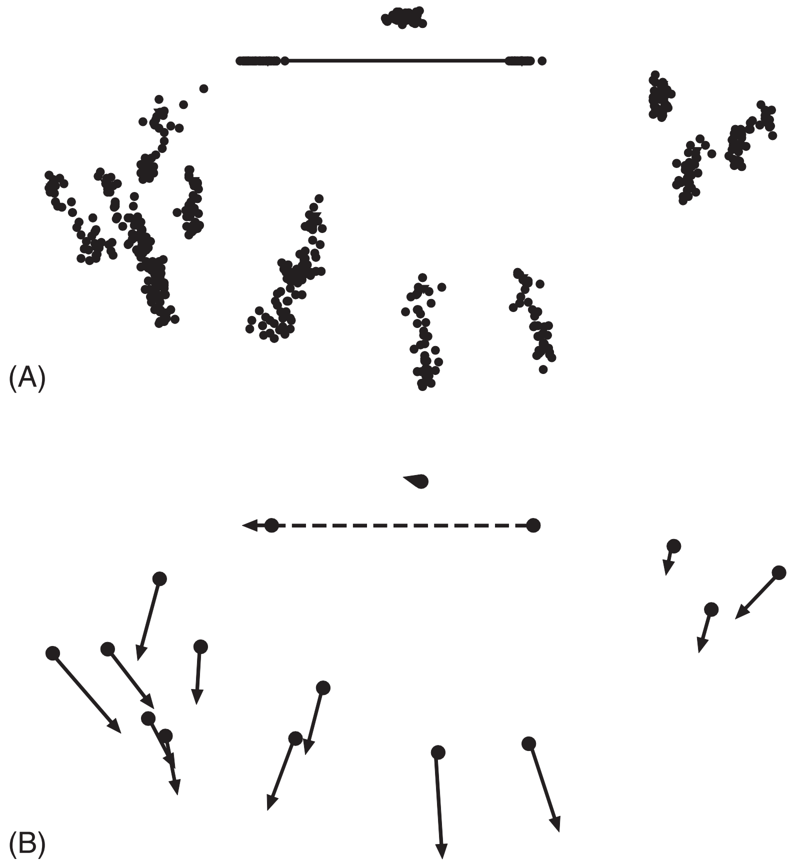


Figure 5.5 Ontogenetic variation in *S. gouldingi* visualized by sliding baseline registration, using a baseline near the dorsal edge of the body (landmarks 3 and 5): (A) coordinates of landmarks relative to the baseline; (B) vectors indicating displacements of landmarks relative to the baseline. Less variance is transferred than was the case when this baseline was fixed, but there are still substantial induced correlations because the variance of the baseline endpoints is mostly perpendicular to the baseline.

nor are they rotated to the orientations that minimize the summed squared distances between the corresponding landmarks. This means that the configurations produced by SBR, like those produced by BC, do not differ solely in shape as defined by Kendall; they are not the same set of configurations as would appear in Kendall's shape space. Even so, the configurations produced by SBR might be preferable if the orientation of the baseline has some biological significance. In that case, we might judge that rotation *does* change shape, so the only rotation permitted during superimposition is that which corrects an earlier misalignment when specimens were digitized. By choosing to rotate each configuration to a specific orientation, we have effectively chosen to include information about orientation (i.e. "rotational effects") in what we mean by "information about shape." This is not a trivial choice; it takes us away from the mathematical theory discussed in Chapter 4.

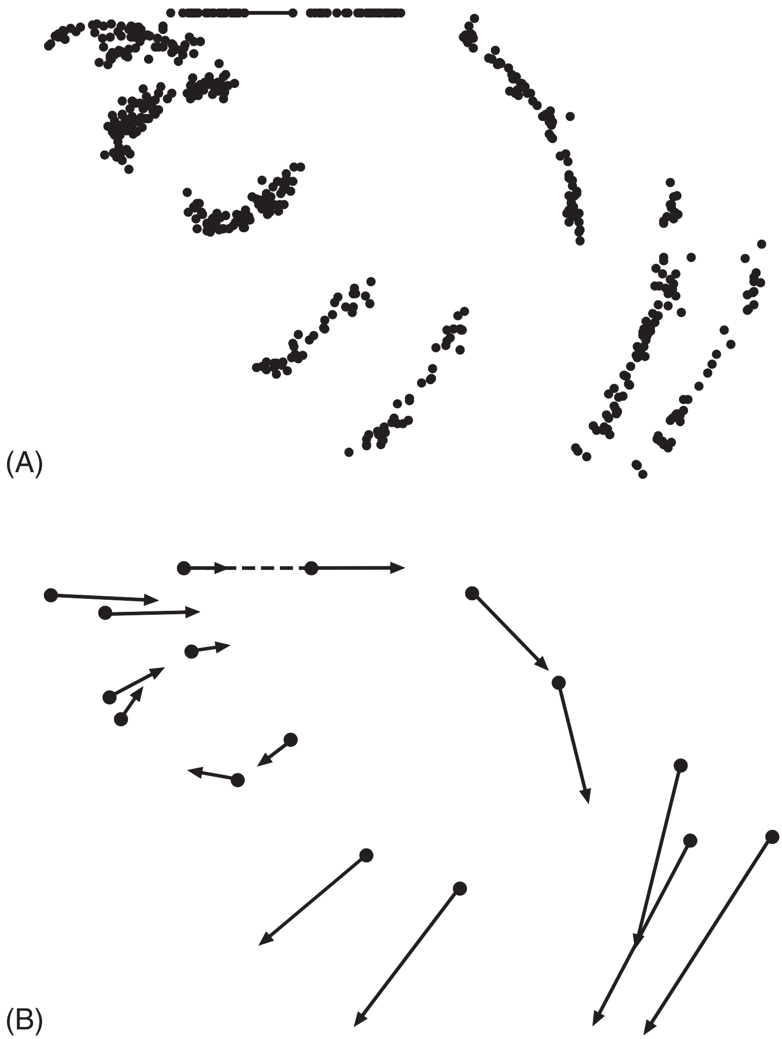


Figure 5.6 Ontogenetic variation in *S. gouldingi* visualized by sliding baseline registration, using a baseline (landmarks 2 and 3) that rotates relative to most other landmarks: (A) coordinates of landmarks relative to the baseline; (B) vectors indicating displacements of landmarks relative to the baseline. Again, there is improvement relative to fixing the length of this baseline, but substantial induced correlations remain.

In addition to the lingering problems shared with BC (variance transfer and constrained alignment), SBR has a new problem that may also limit its utility for data analysis. By allowing the X-coordinates of the two baseline points to vary, two new variables are added to the analysis. Consequently, the number of variables exceeds the number of degrees of freedom by two. Thus, if these coordinates are to be analyzed using conventional statistical methods, such as Hotelling's T^2, we must exclude two coordinates from the analysis. Unfortunately there is no general rule for deciding which to exclude, and there is an obvious risk that the ones selected are chosen because they produce the desired results.

One possible solution is to try all possible pairwise combinations of dropped coefficients, asking if the statistical results are robust to the choice. A better solution is to replace the conventional statistical methods with resampling-based methods, which do not require estimates of the degrees of freedom (see Chapter 8).

Even if you decide that SBR (or BC) is not appropriate for analysis, you can still choose to use it for purely graphical purposes, reserving the statistical analysis for the coordinates introduced in the next section. This will not lead to any inconsistencies between statistical and biological inferences; the statistical inferences should not depend on the choice of variables, and nor should any biological inferences. Obviously the statistical tests must be based on some set of variables, and so must the pictures, but if the analyses and the pictures are done correctly, all complete sets of variables will yield the same statistical results and all illustrations of those results will support the same interpretations.

Generalized least squares Procrustes superimposition

Of all the topics covered in this book, this section on the generalized least squares Procrustes superimposition (GLS) may be the most important. For reasons discussed below, this is the generally favored superimposition method. This is the method we (and many others) use to take the raw data from the digitizer and turn it into the data we analyze. If you understand the coordinates obtained by GLS, you understand enough about the data to use them in studies based on ordination methods, such as principal components analysis (Chapter 7), or in studies using multivariate statistics (Chapters 8, 9 and 10).

The name *Procrustes* comes from Greek mythology; Procrustes fit his visitors (victims) to a bed by stretching or truncating them. In doing so, Procrustes minimized the difference between his visitors and the bed. In this sense the comparison is apt; Procrustes superimpositions minimize differences between landmark configurations. In another sense, the name does not fit; what Procrustes did altered the shape of his visitors, whereas the mathematical superimposition methods only use those operations that do not alter shape. Presumably Procrustes' guests would have preferred that he had done likewise!

As prescribed in Kendall's definition of shape (discussed in Chapter 4), Procrustes superimpositions rely on translation, scaling, and rotation to remove all information unrelated to shape. However, these operations are used in several other superimposition methods for the specific reason that they do not change shape. The crucial distinction of the GLS method is the criterion used to minimize differences between configurations: the Procrustes distance (the summed squared distances between corresponding landmarks – see Chapter 4). The particular combination of translation, scaling and rotation that minimizes the Procrustes distance is considered the optimal Procrustes superimposition. (An interesting historical note on this method is that it was developed before the importance of the Procrustes distance was fully realized.)

A step-wise description of the GLS method was presented by Rohlf (1990). We summarize each step below; they should be familiar from the previous chapter, in which we discussed putting shapes into shape space (we also presented the mathematical details of each operation in that chapter):

1. Center each configuration of landmarks at the origin by subtracting the coordinates of its centroid from the corresponding (X or Y) coordinates of each landmark. This

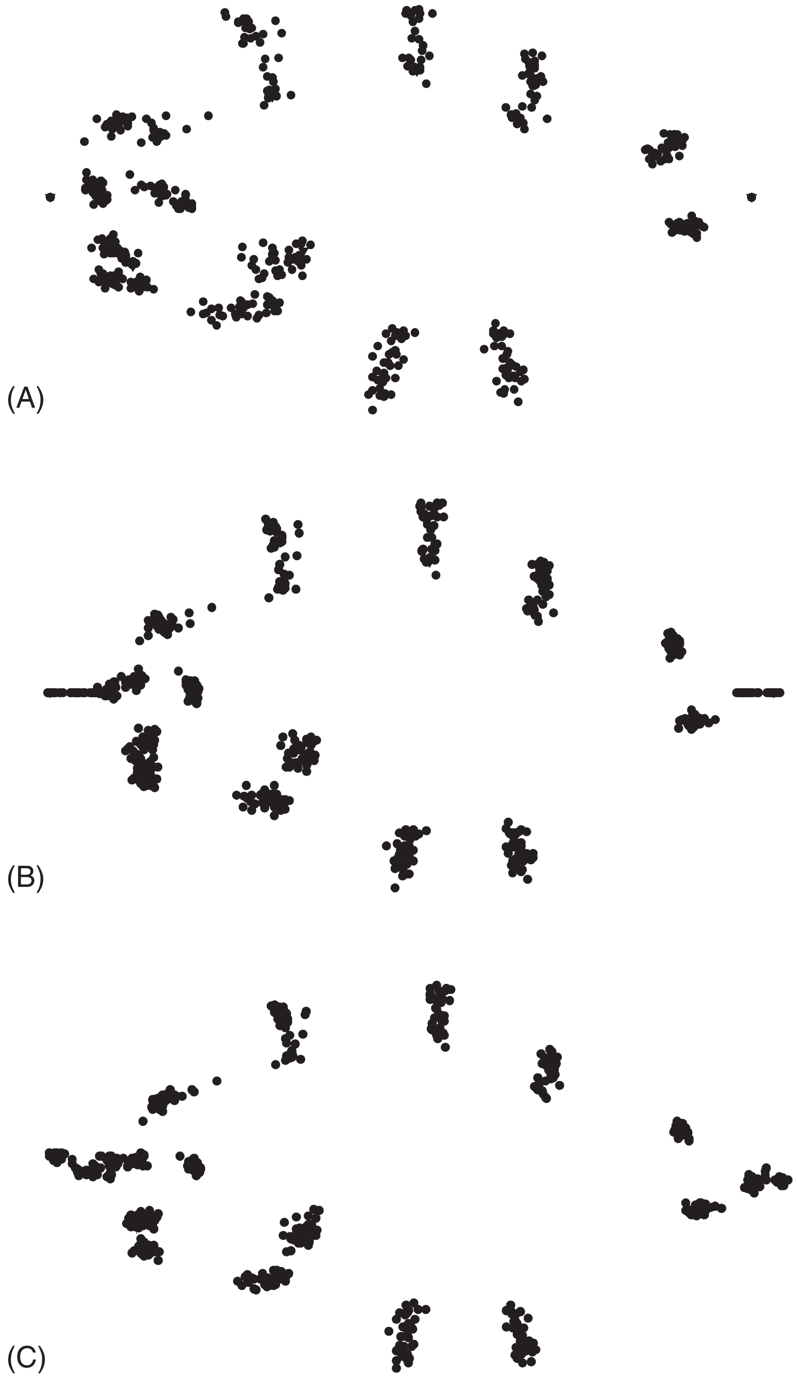


Figure 5.7 Ontogenetic variation in *S. gouldingi* visualized by three different superimpositions: (A) Bookstein's shape coordinates from the 1–7 baseline: (B) sliding baseline registration to the same baseline; (C) Procrustes superimposition.

translates each centroid to the origin (and the coordinates of the landmarks now reflect their deviation from the centroid).
2. Scale the landmark configurations to unit centroid size by dividing each coordinate of each landmark by the centroid size of that configuration.
3. Choose one configuration to be the reference, then rotate the second configuration to minimize the summed squared distances between homologous landmarks (over all landmarks) between the forms. In other words, rotate the second configuration to minimize the partial Procrustes distance.

When there are more than two forms, all are rotated to optimal alignment on the first; the average shape is then calculated and all are rotated to optimal alignment on the average shape, which is the new reference. At this point, the average shape is recalculated. If it differs from the previous reference, the rotations are recalculated using this newest reference. When the newest reference is the same as the previous, the iterations stop (usually only a few iterations are necessary). The final reference is the one that minimizes the average distances of shapes from the reference. Note that this result does not depend on the shape of the first specimen used in the alignment; instead, it depends on the distribution of shapes in the sample.

The protocol outlined above has been called partial Procrustes superimposition (Dryden and Mardia, 1998). The centroid size is set to one for all specimens, so the minimum distance of a specimen from the reference is the partial Procrustes distance (the length of the chord connecting fibers in pre-shape space at their points of closest approach – see Chapter 4). Two steps can be added to the end of this protocol to produce a full Procrustes superimposition, a method that minimizes the full Procrustes distance: the first is to compute the full Procrustes distance and corresponding centroid size from the partial Procrustes distance (this relationship was discussed in Chapter 4); the second is to rescale each configuration to the new centroid size. The partial Procrustes superimposition is more commonly used in biological applications, partly because size is held constant, simplifying the interpretation of shape differences.

The results of the partial Procrustes superimposition (GLS) computed for the piranha ontogenetic series are compared to coordinates generated by BC and SBR in Figure 5.7. The coordinates generated by this superimposition are much like those generated from the two baseline methods when the baseline connected landmarks 1 and 7. These landmarks are the most distant from the center, so they will tend to have a larger influence than other landmarks on the rotation to minimize the partial Procrustes distance (a small angular displacement produces a large linear displacement, which is squared). In addition, the GLS result appears to be more similar to the SBR result than it is to the BC result. This is because GLS and SBR both scale each specimen to unit centroid size, so both show relative deepening as a combination of deepening and shortening.

One advantage of GLS is that the transfer of variance seen in BC is reduced even more than in SBR. This reduction is partly due to allowing both coordinates of all landmarks to vary freely (subject to the other constraints of translation, scaling and rotation). The consequences of this can be seen readily in the pictures of the inferred ontogenetic transformations of *S. gouldingi* (Figure 5.8). In BC (Figure 5.8A) the dorsal and ventral displacements of landmarks are largest, because this is the only way to express deepening relative to the fixed baseline. In SBR (Figure 5.8B) the dorsal and ventral displacements

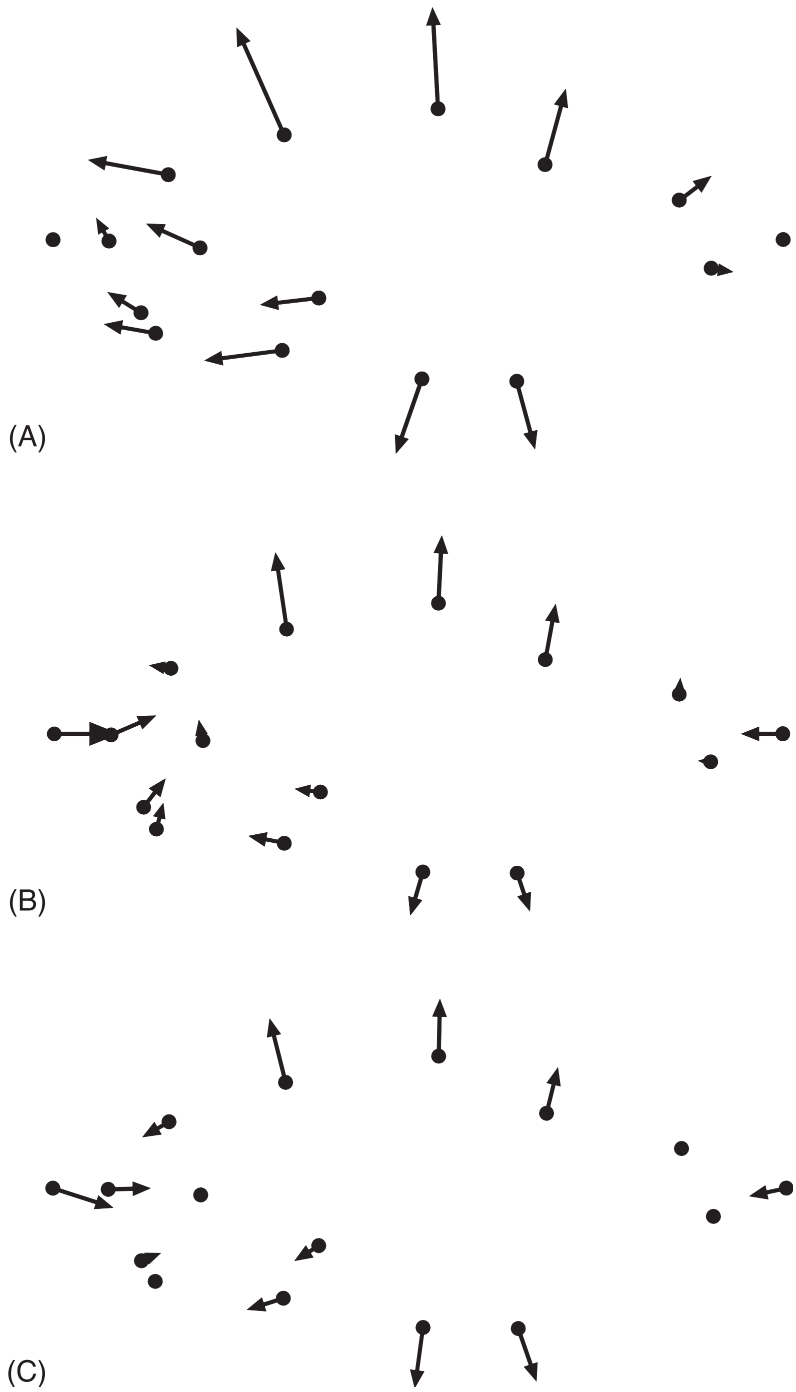


Figure 5.8 Ontogenetic change in *S. gouldingi* depicted by vectors of relative landmark displacement computed from three superimpositions: (A) Bookstein's shape coordinates from the 1–7 baseline; (B) sliding baseline registration to the same baseline; (C) Procrustes superimposition. Note that the points that had been anchored to the baseline are displaced ventrally as well as posteriorly.

are reduced, because relative deepening can also be expressed as relative shortening; consequently, landmarks 1 and 7 are shown moving posteriorly and anteriorly respectively. Both BC and SBR also indicate that the dorsal body deepens faster than the ventral body, and that several landmarks in the head are displaced dorsally. In GLS (Figure 5.8C) we see landmarks 1 and 7 moving toward each other as in SBR, but only in GLS do we see these two landmarks move ventrally relative to all other landmarks. The ventral displacements of these landmarks are an equally valid representation of the greater deepening of the dorsal body, but one that minimizes the covariance of head and dorsal body landmarks. This effect is achieved by minimizing the implied displacements of all landmarks simultaneously.

The more important advantage of GLS is that it is grounded in the mathematical theory of shape. Configurations of landmarks are manipulated using the three operations that do not alter shape as defined by Kendall. These operations are used in a manner that removes all differences that are not shape differences. The configurations produced by this procedure are those that map to points in the shape spaces implied by Kendall's definition of shape. The computed distances between these configurations (the various Procrustes distances) are the distances between points in those spaces, or in certain linear approximations of those spaces. The characteristics of these metrics are well known, providing a secure and stable foundation for biological shape analysis.

One of the main disadvantages of GLS is that it yields the full complement of $2K$ variable coordinates, which is four more than the number of dimensions of the shape space. Fortunately, this is a relatively minor problem that can be circumvented rather easily. One option is to convert the coordinates to the variables discussed in Chapter 6 – the partial warps scores (the two sets of results will be consistent because both use the same distance metric). Another option is to use the resampling-based statistical methods discussed in Chapter 8, which do not require estimates of degrees of freedom. Yet another option is to use statistical tests specifically adapted to the GLS coordinates (e.g. Goodall's F-test, discussed in Chapter 9). Thus the excess number of variable coordinates does not pose an obstacle to valid statistical analysis.

Another disadvantage is that GLS can yield visually unsettling results, such as rotated axes of symmetry. For example, analyses of rodent skulls (Zelditch et al., 2003) use "symmetrized" landmarks on one half of the skull to avoid inflating degrees of freedom (coordinates for one side are reflected onto the other and the coordinates of the two sides are averaged for each specimen; see Chapter 3). In the GLS result (Figure 5.9A) the midline of the skull appears to rotate, but that cannot happen; the midline is the midline regardless of variation in shape. Not only is this apparent rotation of the midline visually troubling, it also complicates the interpretation of the results. These are the problems that SBR was designed to overcome (Figure 5.9B). Because SBR prevents rotation of the baseline, it yields a more realistic representation of the data – in this case, of the ontogenetic change in skull shape. Actually, a very similar picture can be obtained by duplicating the landmark coordinates (except the landmarks on the midline) and reflecting the second set across the midline, then performing GLS superimposition on the reconstructed whole skulls (Figure 5.9C). In general, reconstructing the whole skull makes a more interpretable picture (one that looks more like the organism), so it might be useful to present results in these terms even if the statistical analyses used the GLS coordinates computed for the symmetrized half skull.

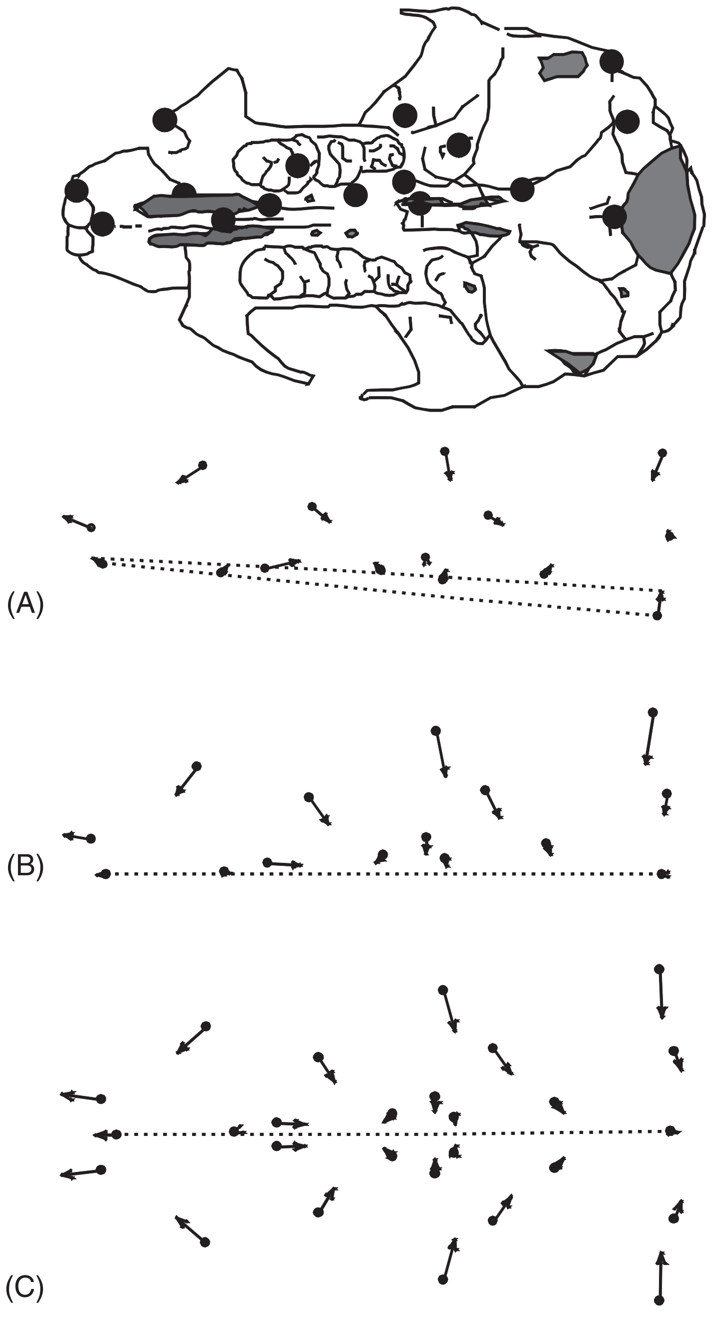


Figure 5.9 Superimpositions of forms with an axis of symmetry – ontogenetic changes in a rodent skull. Dotted lines connect landmarks on the sagittal plane. (A) Landmark displacements inferred from GLS, which appears to indicate translation and rotation of the sagittal plane; (B) landmark displacements inferred from SBR, which does not appear to suggest translation and rotation of the sagittal plane; (C) landmark displacements inferred from GLS on symmetrized and back-reflected configurations, which also does not appear to suggest translation and rotation of the sagittal plane.

(We should note that all the analytic software we provide uses the coordinates obtained by GLS, but other types of superimposition are available for depicting the results – although GLS is usually fine for that purpose as well.)

Resistant-fit superimposition

Like GLS, resistant-fit superimposition methods minimize differences between configurations by minimizing differences at corresponding landmarks over all landmarks. In recognition of this general similarity, the resistant-fit methods have also been characterized as "Procrustes methods" (see Chapman, 1990). However, the crucial difference between resistant-fit methods and GLS is that the former do not use Procrustes distance as the criterion for optimal superimposition. The resistant-fit methods also differ from each other in the optimization criteria they do use. Below we examine the rationale for rejecting the Procrustes distance metric and the general objective of the alternative optimization criteria, then we focus on the oldest and most well-known of these methods, resistant-fit theta-rho analysis (RFTRA; Siegel and Benson, 1982) and examine its optimization criterion in somewhat greater detail.

The general objection is that least squares optimizations like GLS are very sensitive to large displacements at few landmarks. In statistical procedures like regression, a few cases with unusually large deviations from the general pattern ("outliers" or "influential observations") can have a large effect on the results because the procedure minimizes the sum of the squared deviations. In shape analysis, a large change limited to one or a few landmarks is sometimes called the Pinocchio effect (however, the influential landmarks need not be at the tip of a long process). Figure 5.10A shows a hypothetical example in which the only shape change in a tree squirrel scapula is the ventral displacement of the three most ventral landmarks. When GLS is used to superimpose landmark configurations (Figure 5.10B), the Pinocchio effect can have a large effect on the superimposition. The least squares criterion distributes the displacement of the few landmarks across all the other landmarks. In graphical displays, the Pinocchio effect appears to be "smeared out" over all landmarks, which can be unsettling for some workers. The more severe consequence is that the least squares criterion causes the variances of the influential landmarks to be allocated to other points, inducing covariances (Walker, 2000). Ironically, minimizing induced covariances was one of the reasons for using GLS rather than a baseline method.

Resistant-fit methods reduce the influence of the Pinocchio effect by taking a "robust" approach to superimposition. In statistics, "robust" means that the method is relatively insensitive to outliers in the data. Similarly, a robust superimposition method is relatively insensitive to a few landmarks with large relative displacements. A wide variety of error functions have been used as criteria for robust fitting procedures, the interested reader is referred to Press et al. (1988) for a discussion of several alternatives. None of them allow analytic solutions for the rotation and scaling parameters needed to carry out a superimposition; instead they use numerical methods (simplex searches) to find the rotation and scaling necessary to minimize the error function.

The robust approach implemented by RFTRA uses the method of "repeated medians" to determine the scaling and rotation necessary to superimpose one shape on another

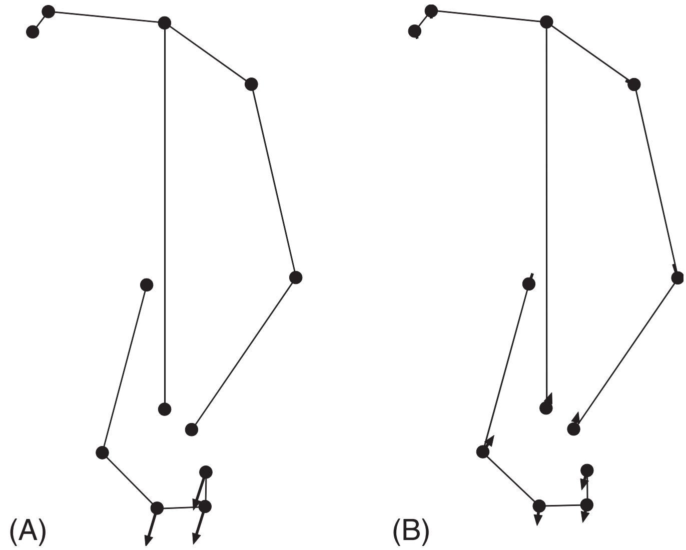


Figure 5.10 Hypothetical example of the Pinocchio effect as exemplified by ventral displacements of the three most ventral landmarks of a tree squirrel scapula: (A) configurations superimposed by a resistant-fit method (RFTRA); (B) configurations superimposed by GLS. The outline of the scapula is approximated by lines connecting the landmarks.

(Chapman, 1990). We describe the steps used to find the scaling factor in some depth; then more briefly describe the steps to find the rotation. For the scaling factor:

1. Compute the pairwise interlandmark distances in both shapes and then compute the ratio of each pair of corresponding distances.
2. For each landmark, find the median of ratios for all segments radiating from that landmark. This will yield one ratio for each landmark.
3. Find the median of the medians generated by step 2. This median of medians is the scaling factor used in the superimposition – in other words, all coordinates of the second shape are scaled by this factor.

After scaling the second form, the rotation angle used by RFTRA can be determined in a similar fashion from the same set of line segments. The first step is to compute the angles between the corresponding segments; the remaining steps find the median angle associated with each landmark and then the median of the medians. (As in GLS, an iterative procedure is used to compute a reference shape and superimpose all the specimens on it.)

RFTRA is robust because medians are relatively insensitive to outliers, and, consequently, large changes at one or a few landmarks will not appreciably alter the median scaling factor or the median rotation angle. This makes RFTRA resistant to the Pinocchio effect, which helps to highlight the region where the effect occurs, as in Figure 5.10. However, in the absence of the Pinocchio effect, superimpositions produced by resistant-fit methods usually do not differ greatly from those produced by GLS. Figure 5.11 shows GLS and RFTRA superimpositions of the real squirrel scapulae that were the basis of the hypothetical example. The real scapulae differ in the relative length of the ventral

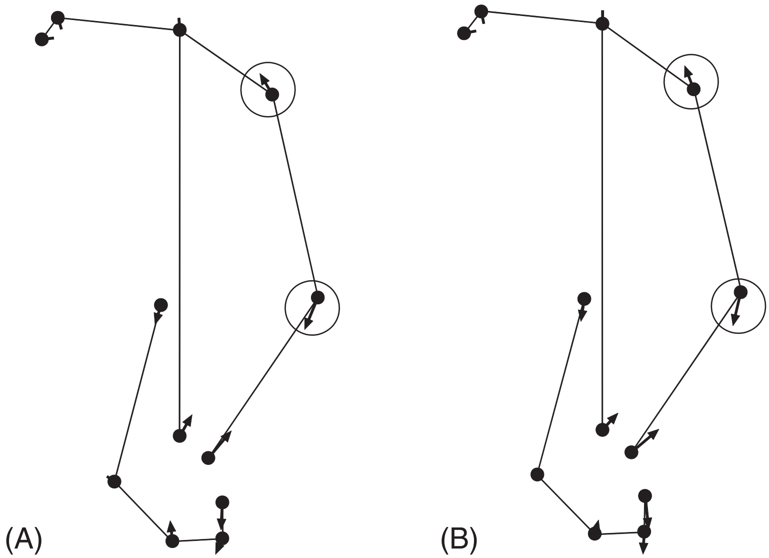


Figure 5.11 Comparison of GLS (A) and RFTRA (B) superimpositions for data that lack a Pinocchio effect: real differences in scapula shape between two squirrels. The circles denote the two landmarks that undergo large relative displacements, in addition to the three most ventral landmarks. The outline of the scapula is approximated by lines connecting the landmarks.

process, as in the hypothetical case, but they also differ in the shape of the anterior edge of the scapula (producing large relative displacements of the two circled landmarks). The difference between the superimpositions is subtle; it is most noticeable at the ventral end, where RFTRA attributes somewhat greater anterior displacements to the more ventral landmarks.

The principal advantage of RFTRA and other resistant-fit methods lies in their ability to address the Pinocchio effect. Their principal disadvantage lies in their departure from the Procrustes distance metric. Comparisons performed under these superimpositions are not covered by the theory developed in Chapter 4.

If you are considering a resistant-fit method, we recommend that you compare several alternative methods. Not all rely on the median of medians, as RFTRA does; some employ an explicit distance metric (or error function), not just the Procrustes distance. Methods that use different metrics will produce different superimpositions, which may lead to different biological inferences as a consequence of the difference between metrics. It is usually difficult to justify using any particular metric, so it is important to compare results based on a variety of methods. That will increase the likelihood that your conclusions reflect the structure of your data and do not depend on the type of software that happened to be available. The program **SuperPoser** (Liebner and Sheets, 2001) carries out robust resistant-fit methods using a variety of error functions. A different robust approach, one based on identifying and excluding highly variable landmarks and then carrying out the superimposition on the remainder (using either generalized least squares or a resistant-fit method), has been developed by Dryden and Walker (1999), but the software for this approach does not seem to be readily available at present.

Summarizing the advantages and disadvantages of different methods

Before summarizing the advantages and disadvantages of the methods, we must admit to our own preference. In general, we favor GLS because of its most notable advantage over all other methods – it is the one consistent with the general theory of shape. The Procrustes distance is the function minimized by this method, so it is the one most consistent with theory (and therefore with our agreed-upon definition of shape). Even the pictures embody that measure. When you look at the pictures, you can see the distance between shapes – it is the square root of the summed squared lengths of the vectors displayed in the graphics. At this point you still may have difficulty appreciating how important this is, but it is the primary advantage of the method and the reason these coordinates are the ones preferred for statistical analyses. Accordingly, we regard it as the "default" method – the one to use unless there is a good reason to choose an alternative. We thus use it as the basis for comparing all other methods.

Another important advantage of GLS over the baseline methods (BC and SBR) is shared with the resistant-fit methods. The advantage is that GLS and the resistant-fit methods do not have landmarks constrained to lie on the baseline, so there is no transfer of variance from these points to all other landmarks. In general, this reduces the induced covariances among landmarks and produces smaller variance ellipses around most landmarks. However, GLS enjoys less of an advantage when there is a strong Pinocchio effect. In this case the resistant-fit methods will have a greater advantage, particularly if the user is not troubled by their departure from the Procrustes distance metric and the associated shape spaces.

One disadvantage of GLS is that the number of variable coordinates exceeds the number of shape variables by four. If the data are analyzed by conventional statistical tests, four randomly selected coordinates must be dropped – which could have a substantial impact on results. Fortunately, there are three generally accepted ways of getting around this problem. One is to replace the landmark coordinates with a set of shape variables that convey the same information but use the correct number of variables (the partial warps scores discussed in Chapter 6). The second is to use resampling based methods that do not require estimates of the number of degrees of freedom (these methods are discussed in Chapter 8). The third option is to use statistical tests specifically adapted to the coordinates produced by GLS (such as Goodall's *F*-test, discussed in Chapter 9). Thus the discrepancy between the number of variable coordinates produced by the GLS superimposition and the number of shape variables is an obstacle that is easily removed.

A potentially more serious problem is that the Procrustes method freely rotates forms to maximize their similarity. This is consistent with the definition of shape that forms the basis of geometric morphometrics, in which rotations do not alter shape. From a biological perspective, however, this attitude toward rotations might seem unreasonable. As Bookstein (1996) observes, the method "happily" rotates shapes around axes of bilateral symmetry. The effect is illustrated in the analysis of rodent skulls (Figure 5.9), in which the Procrustes superimposition appears to rotate and translate the midline, even though that is biologically unrealistic. The effect is not due to asymmetry in the skulls because the data for each individual were "symmetrized" (as described earlier); instead, the rotation reflects the fact that posterior landmarks generally undergo larger medio-lateral displacements than anterior landmarks (this might be most visible in the depiction by the sliding baseline, Figure 5.9B). The change in the relative width of the two ends forces the GLS superimposition to rotate the skull, and the midline with it. The GLS is free to perform that rotation because the orientation of the skull is not information relevant to an analysis of shape. In contrast, our interpretation that the midline has been rotated, and especially our unease with that interpretation, reveals a perspective in which orientation is information relevant to an analysis of skulls.

The conflict between the two views outlined above raises an important conceptual issue. If we regard rotation as inappropriate under some conditions, we need to clarify what we actually mean by equivalent shapes. In particular, if we would not consider two forms to be equivalent when they differ by rotation (or translation) of an axis of symmetry, then we should not place them in the same class or claim that there is no distance between them. This view of rotation is not consistent with the Procrustes metric, or with the idea that rotations do not alter shape. Either we must adopt a different definition of shape (and a new theory of shape analysis to go with it), or we must recognize that sometimes the difference of interest is not purely a difference in shape. The latter approach seems more productive; not only does it retain a well-established theory of shape analysis, but it also recognizes that there is more to morphology than shape. In the case of the rodent skulls, the rotations that were used to reveal shape differences removed an important component of information about skull morphology – namely skull orientation. Usually orientation is viewed as a "nuisance" parameter because it only refers to the orientation of a specimen on a digitizer, but when dealing with axes of symmetry, orientation has a biological significance. The loss of information about orientation is what makes the pictures difficult to interpret.

Fortunately, the conflict between biological and geometric perspectives can be addressed by judicious choice of graphical styles; we do not need to give up one perspective for the other. The statistical analyses can be performed on the symmetrized data (the half skull) to evaluate shape differences, regardless of the graphical representation. One option is to use SBR to depict the results, although this method will not convey the actual Procrustes distances among shapes (the coordinates obtained by SBR can be analyzed statistically, using a resampling method, to check that the results are consistent with those based on the coordinates obtained by GLS). Another option is to duplicate the coordinates of the symmetrized landmarks, reflect them back across the midline to create whole symmetrical shapes, and then perform a GLS superimposition on the reconstructed whole skulls (Figure 5.9C; see also Zelditch et al., 2003). This approach allows us to use coordinates that are consistent with the Procrustes distance metric (hence directly depict the results of any statistical analyses that are done) while avoiding the problem of interpreting inappropriate translations or rotations of the baseline.

Although none of the limitations of GLS are particularly burdensome, there are still times when a baseline superimposition method might be preferred. Generally, these are cases when it is useful to have all shapes aligned to a standardized or conventional orientation. Alignment of the skulls along the midline (as above) is just one such example. Frequently, fixing the baseline simplifies the interpretation of complex shape changes by making it possible to begin with an analysis of each landmark's displacement relative to the baseline (although it remains difficult to talk about changes in all the free landmarks relative to each other). To the extent that this is an advantage, it is a bigger advantage for BC than SBR because BC fixes all four coordinates of the baseline endpoints and SBR fixes only two. In some cases this advantage might be cancelled out by the fact that

SBR coordinates are scaled by centroid size rather than baseline length. Not only does SBR have the advantage of using a concept of scale that is consistent with the theory of geometric morphometrics; the implied shape distances are also closer to the Procrustes distances.

Interpreting graphical results: resolving apparent inconsistencies

The graphical representation of results is one of the main reasons why geometric methods are so useful. Because the graphics can influence the interpretation of results, it is important to understand exactly how the superimposition methods influence the graphics. To illustrate these effects, we use the four superimposition methods that have been the focus of this chapter (BC, SBR, GLS and RFTRA) to depict the results of a single analysis of the ontogeny of body shape in the piranha *S. gouldingi* (Figure 5.12).

Perhaps the most obvious difference among the four panels is the degree to which postcranial landmarks are vertically displaced. It might appear that Bookstein shape coordinates either exaggerate the degree to which the postcranial body is deepened, or else that the other superimpositions understate it. However, this is not the case; all the other superimpositions show a relative shortening of the body, which is equivalent to a relative deepening. Both mean exactly the same thing. Relative body depth is a ratio between depth and length, so it is just as reasonable to think of it as a decrease in length relative to depth as to think of it as an increase in depth relative to length. Increasing body depth increases the ratio by increasing the numerator; decreasing body length also increases the ratio, but by decreasing the denominator. Because we come to the pictures informed by our knowledge

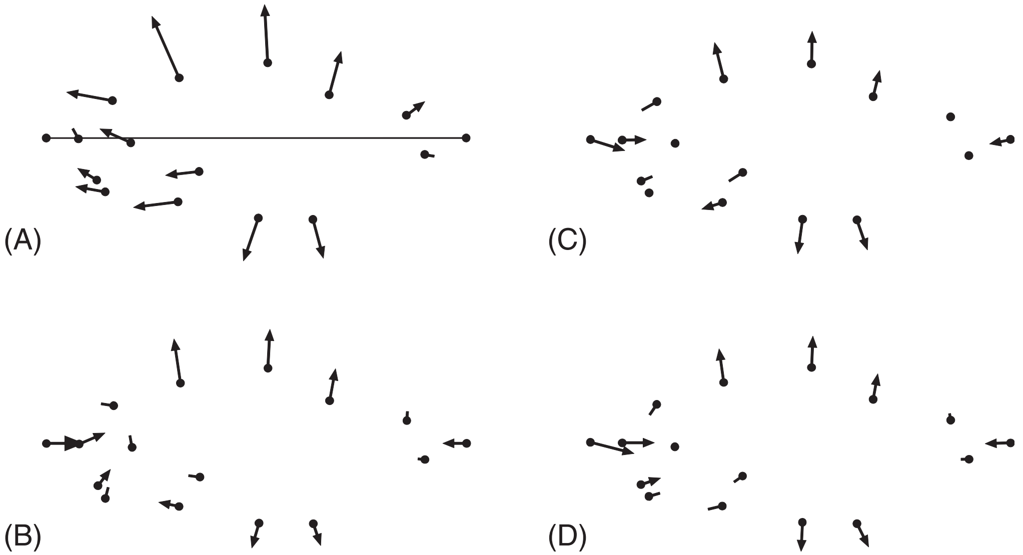


Figure 5.12 Ontogenetic change in body shape of *S. gouldingi* depicted by vectors of relative landmark displacement computed from four superimpositions: (A) BC coordinates from the 1–7 baseline; (B) SBR to the same baseline; (C) GLS; (D) RFTRA.

that body length increases over ontogeny, it may be difficult to grasp that it decreases relative to depth. We would probably avoid saying that body length decreases relative to depth, simply because that phrasing is disconcerting to biological intuition; instead, we would say that depth increases relative to length. When pictures show a *relative* decrease in a feature that is increasing in *absolute* length, readers may need some explanation of the unexpected contrast. In particular, it is important to explain that the decrease is in relative (not absolute) length.

Other apparent inconsistencies between pictures can also be reconciled, usually by concentrating on the changes in *relative position*s of landmarks rather than on the vectors at individual landmarks. It may take a lot of practice before this is easy. For example, look at the circled landmark in Figure 5.13. If you look only at this landmark, the results from the different superimpositions appear to be inconsistent. That landmark appears to "move" quite far anterodorsally in the BC superimposition (Figure 5.13A), but much less and in three different directions in the other superimpositions: anteriorly in SBR (Figure 5.13B); anteroventrally in GLS (Figure 5.13C), and almost entirely ventrally in RFTRA (Figure 5.13D). However, none of these statements actually reflect what the pictures show. None of the pictures show the independent movement of any one point in isolation; rather, what they show is the relative displacements of all points.

We get a better indication of the displacement of the circled landmark relative to neighboring landmarks by "connecting the dots" – drawing line segments between landmark locations to approximate the profile of *S. gouldingi*'s head. In Figure 5.14 we show the same superimpositions and ontogenetic displacements of landmarks as in Figure 5.13, and we add lines to show the relative positions of the landmarks early in ontogeny (dotted lines connecting the bases of the arrows) and late in ontogeny (solid lines connecting the tips of the arrows). Now we can see that the profile of the head is initially fairly shallow

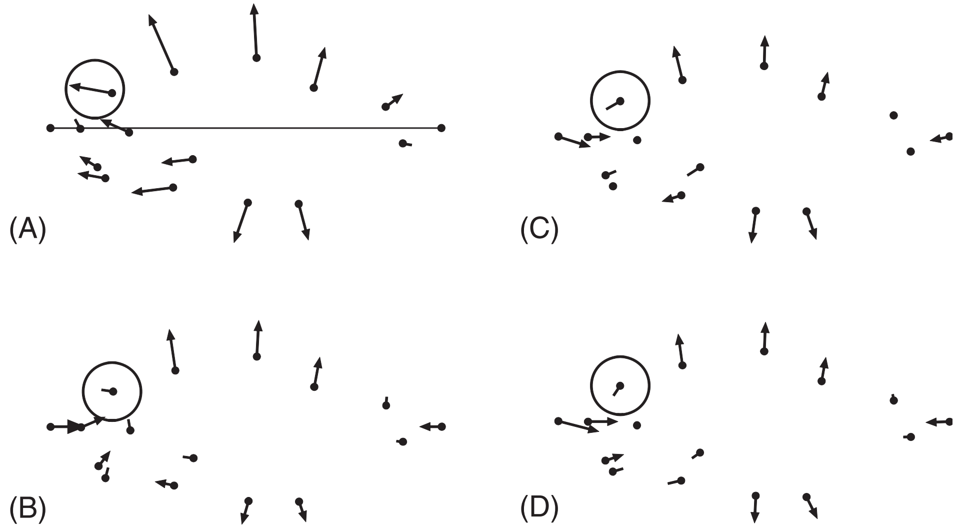


Figure 5.13 Ontogenetic change in body shape of *S. gouldingi*, highlighting the landmark at the epiphyseal bar. Displacements are shown in four superimpositions: (A) BC coordinates from the 1–7 baseline; (B) SBR to the same baseline; (C) GLS; (D) RFTRA.

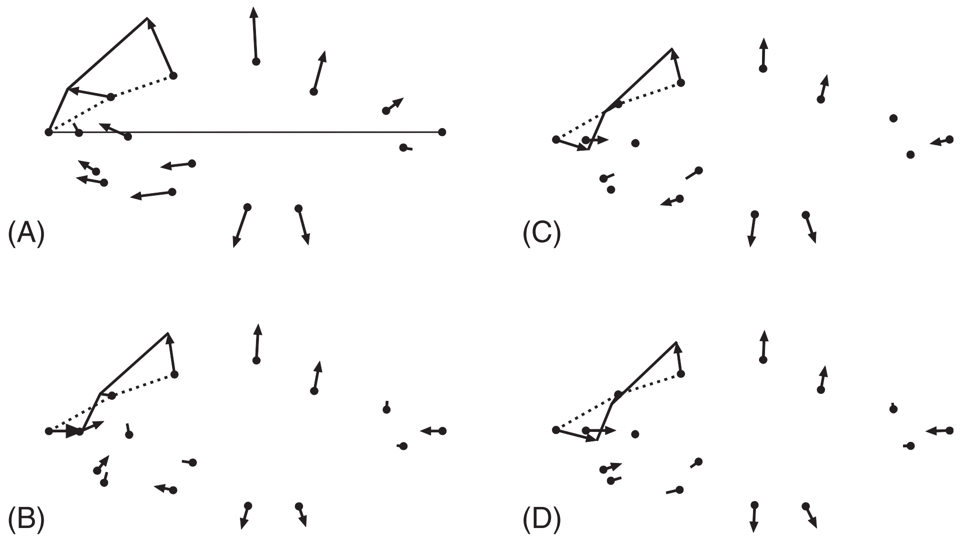


Figure 5.14 Ontogenetic changes in dorsal head profile are highlighted by drawing in line segments between the locations of landmarks at two different ages: early in ontogeny (dotted line) and late in ontogeny (solid line). Displacements are shown in four superimpositions: (A) BC coordinates from the 1–7 baseline; (B) SBR to the same baseline; (C) GLS; (D) RFTRA.

(nearly a straight line across all three points), and becomes much steeper (particularly between the tip of the snout and the second landmark – the one that was circled). All four superimpositions show this same change in profile. Despite apparent discrepancies in the displacements of individual landmarks, the relationships among the landmarks are consistently represented. Before we can interpret the results in terms of these vectors of relative landmark displacement, we must become accustomed to what these vectors represent. The individual vectors do not show changes *at* landmarks; rather, the differences between vectors show changes between the landmarks.

Because the different superimpositions show the same shape change, the primary criterion for choosing which to use to display results is simply the ease of interpretation. In some cases, it is easier to understand the results depicted by Bookstein shape coordinates because the fixed baseline provides a simple, straightforward line of reference. That line makes it easier to interpret and to verbalize the information contained in the pictures; aiding both your ability to understand your results and your ability to communicate your results to others.

Recommendations regarding superimposition methods

1. Unless there is a good reason for choosing an alternative, use GLS.
2. When an axis of symmetry is present, and it is rotated by GLS, decide whether you think that shape is unaltered by rotation. If you are willing to redefine shape, so that it actually *is* altered by rotation, use SBR, and use resampling-based statistical tests. Alternatively, use GLS for statistical analyses and depict the results by back-reflecting

the data. However, if choosing this second approach, realize that ordination methods (like principal components and canonical variates analysis, discussed in Chapter 7) do not necessarily give the same results for the symmetrized and back-reflected data. If using ordination methods, do the analysis on the symmetrized data, save the PC or CV scores, and regress the back-reflected data on those scores to obtain the pictures.

3. When conventional statistical tests are the only option, either use Bookstein's shape coordinates or use the variables explained in Chapter 6 (partial warps). These are the only methods that produce $2K - 4$ variables, equal to the dimensionality of shape space.
4. When the major objective is to depict the differences between forms, use several methods to see how each represents that difference. Only the GLS method displays the difference in terms that are exactly commensurate with the Procrustes distance. If your primary concern is having a picture that unambiguously represents the measured distance between shapes, use the GLS to depict the results. If ease of interpretation or communication is of greater importance, use whatever method produces the most easily interpreted graphics.
5. When there is a great deal of noise that does not respond to alternative choices of baselines, or at least not to alternatives that permit straightforward interpretations, SBR or GLS will be more useful than Bookstein shape coordinates (regardless of interpretability).
6. If you insist on a robust fitting method, look beyond RFTRA. Determine which of the many possible error functions is most appropriate for your data, and use that in weighting the variances.

Software

The program **CoordGen**, introduced in Chapter 3, can also be used to obtain coordinates by SBR, GLS, and RFTRA, in addition to BC. A variety of resistant-fit coordinates can be calculated using **SuperPoser**. Virtually all the software in the IMP series can read coordinate data produced by any kind of superimposition and convert it to GLS for analysis, so you can input the BC coordinates you saved earlier and use them in all subsequent analyses. Each program gives you the option to display your results (based on the GLS) using any of the other available superimposition methods.

When running **CoordGen**, you will follow the same general procedure to obtain superimposed coordinates regardless of the method of superimposition, so most of the directions for using **CoordGen** are given in Chapter 3 (which introduced BC). There is one major difference that applies to GLS and RFTRA. Unlike BC and SBR, GLS and RFTRA do not depend on baseline, so you do not need to enter the endpoints of the baseline. However, during the input of a new dataset, **CoordGen** automatically aligns the specimens to the default baseline. This will be the approximate orientation of your specimens after GLS, unless you change it. To select a more convenient orientation, enter a baseline that is closer to your preferred orientation, click **Show BC**, *then* click **Show Procrustes**. Another option for orienting your GLS superimposition is to click the **Show Procr (PA)** button. This will align the reference configuration so that its principal axis (its long axis, approximately) is aligned with the X-axis. You can also click on the **Vertical Axis (Procrustes PA)** before clicking on **Show Procr (PA)**. This will rotate the reference shape so that its principal axis is aligned with the *Y*-axis. If none of these provide a reasonable orientation, don't panic; other IMP programs have interactive tools that allow you to rotate pictures of results (shape differences) through any desired angle.

References

Bookstein, F. L. (1996). Combining the tools of geometric morphometrics. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 131–151. Plenum Press.

Chapman, R. E. (1990). Conventional Procrustes methods. In *Proceedings of the Michigan Morphometrics Workshop* (F. J. Rohlf and F. L. Bookstein, eds) pp. 251–267. University of Michigan Museum of Zoology.

Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*. John Wiley & Sons.

Dryden, I. L. and Walker, G. (1999). Highly resistant regression and object matching. *Biometrics*, **55**, 820–825.

Kim, K., Sheets, H. D., Haney, R. A. and Mitchell, C. E. (2002). Morphometric analysis of ontogeny and allometry of the Middle Ordovician trilobite, *Triarthrus becki*. *Paleobiology*, **28**, 364–377.

Liebner, D. L. and Sheets, H. D. (2001) Superposer, available on the IMP website www.canisius, edu/~sheets/morphsoft.html

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1988). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press.

Rohlf, F. J. (1990). Rotational fit (Procrustes) methods. In *Proceedings of the Michigan Morphometrics Workshop* (F. J. Rohlf and F. L. Bookstein, eds) pp. 227–236. University of Michigan Museum of Zoology.

Siegel, A. F. and Benson, R. H. (1982). A robust comparison of biological shapes. *Biometrics*, **38**, 341–350.

Walker, J. A. (2000). Ability of geometric morphometric methods to estimate a known covariance matrix. *Systematic Biology*, **49**, 686–696.

Webster, M., Sheets, H. D. and Hughes, N. C. (2001). Allometric patterning in trilobite ontogeny: testing for heterochrony in *Nephrolenellus*. In *Beyond Heterochrony: The Evolution of Development* (M. L. Zelditch, ed.) pp. 105–144. Wiley-Liss.

Zelditch, M. L., Lundrigan, B. L., Sheets, H. D. and Garland, T. Jr (2003). Do precocial mammals develop at a faster rate? A comparison of rates of skull development in *Sigmodon fulviventer* and *Mus musculus domesticus*. *Journal of Evolutionary Biology*, **16**, 708–720.

6

The thin-plate spline: visualizing shape change as a deformation

Shape coordinates of all kinds are fundamentally limited when it comes to depicting transformations in shape – they cannot tell us what happens *between* landmarks. Sometimes it is obvious what happens between landmarks, as in Figure 6.1, where we can see that the snout elongates relative to the eye. That is obvious because the posterior eye landmark is displaced towards the anterior eye landmark, and that anterior eye landmark is *not* displaced towards the snout – so the snout must be lengthening relative to the eye. However, it is not so obvious whether the postorbital region is elongating (relative either to the head or body). Similarly, it is difficult to judge whether the head (as a whole) elongates relative to the postcranial body. The problem is not that we lack landmarks in the relevant regions; rather, it is that so many landmarks are displaced relative to the others that it is mentally exhausting to track what happens between them all. That tracking requires looking at the lengths of *all* the vectors to determine whether several landmarks are displaced to a similar degree in concert, or if some are displaced relatively more than others (thereby either increasing or decreasing the distance between them). Even the landmarks that are not displaced relative to others must be considered. What we need is a method for visualizing changes between landmarks over the entire form.

That visualization is the primary purpose of the thin-plate spline. Using it, we can interpolate between landmarks, taking all displacements of all landmarks relative to all others into account (Figure 6.2). The other major purpose of the spline has been mentioned previously in this text: we need a set of shape variables to use in conventional statistical tests. Specifically, we need a set of variables that spans the entire space of our data but numbering only $2K-4$ for two-dimensional data (more generally, numbering $(KM-1-M-(M(M-1)/2))$ where K is the number of landmarks in M dimensions). The spline provides such a set. Unlike the coordinates obtained by the Procrustes-based superimposition methods, the thin-plate spline coefficients (called *partial warp scores*) can be used in conventional statistical tests without adjusting the degrees of freedom. Also unlike the coordinates produced by the two-point registration, which also have the appropriate number for statistical tests, the partial warp scores employ the correct tangent space

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

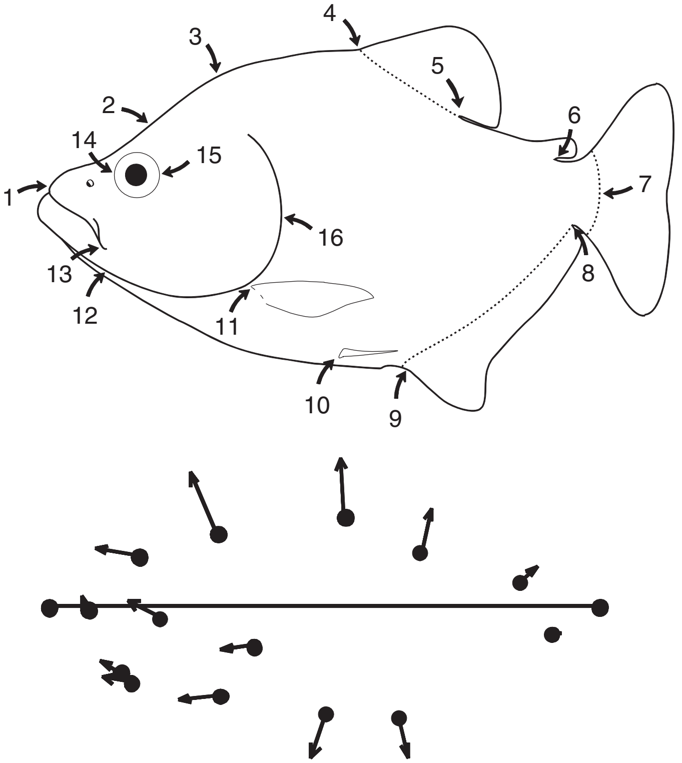


Figure 6.1 Ontogenetic change in body shape of *Serrasalmus gouldingi*, depicted by relative displacements of Bookstein shape coordinates.

measure of distance – the Procrustes distance. Using partial warps you will get precisely the same results as you get using the coordinates obtained by the Procrustes (GLS) superimposition if you correctly adjust the degrees of freedom, or use tests that calculate them properly (like Goodall's *F*).

In summary, the thin-plate spline provides a visually interpretable description of a deformation, with the same number of variables as there are statistical degrees of freedom, and it employs the Procrustes distance as a metric. Even if we were not concerned with the advantages of the spline for graphical analysis, we might still want to use it for purposes of statistical inference. Conversely, even if we were not concerned with the advantages of the spline for statistical analysis, we might still wish to use it for its graphical capabilities. You can use the spline to depict your results, and you can use partial warps in your statistical analyses without worrying that the mathematical details (and complexities) will have any impact on your results. The spline is a convenient tool for visual display and for obtaining variables with the correct degrees of freedom – it is nothing more (or less) than that.

In this chapter, we begin with a basic overview of the mathematical idea of a deformation. We then discuss the mathematical metaphor underlying one particular model of a deformation, the thin-plate spline, and how we can decompose it to yield variables. In general, we present a largely intuitive overview before delving more deeply into the mathematics.

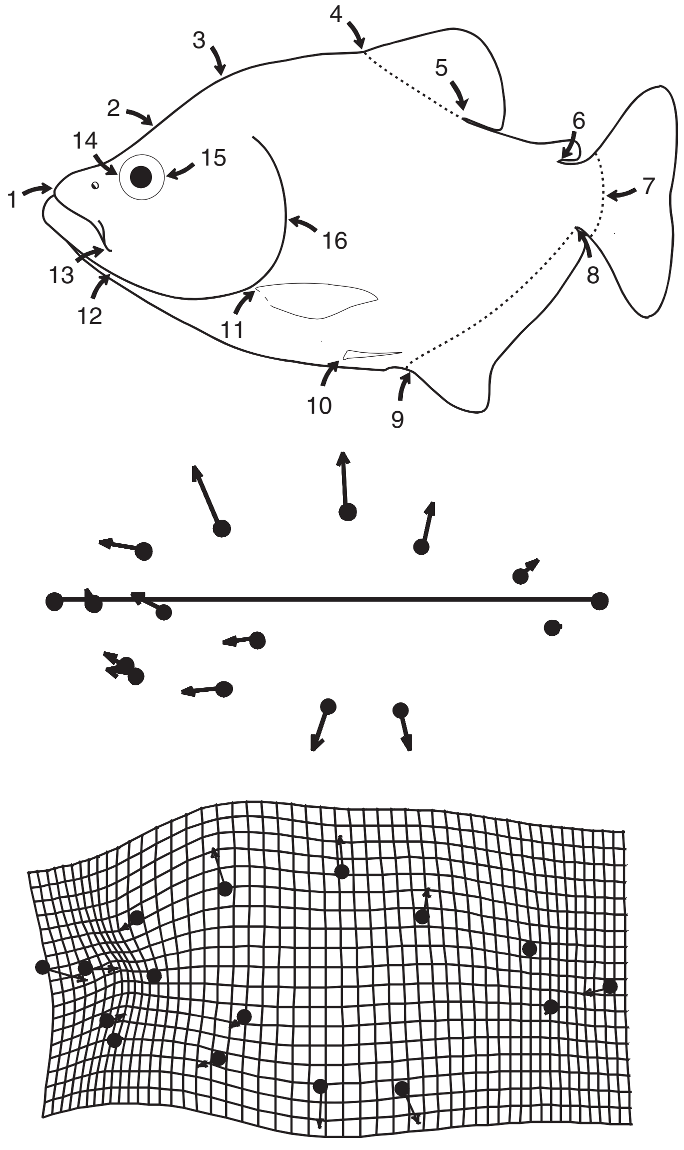


Figure 6.2 Ontogenetic change in body shape of *S. gouldingi*, depicted both by relative displacements of Bookstein shape coordinates and by the thin-plate spline.

Modeling shape change as a deformation

A deformation is a smooth function that maps points in one form to corresponding points in another form. Intuitively, smoothness means that the function goes on without interruptions or abrupt changes. More precisely, it means that the function is continuously differentiable (it can be differentiated, its first derivative can be differentiated, and so can its second, and so forth). To be differentiable, a function must be continuous. For example, the function $Y = X^3$ is continuous, but the absolute value function is not because it has a sharp corner at $X=0$ and so is not differentiable at that point. The Dirichelet function $Y=\{1$ when X is rational; 0 when X is irrational$\}$ is also not continuous – it is not differentiable anywhere. To be continuous, it is not enough to have a first derivative, that first derivative must also be a differentiable function. That deformations are continuously differentiable is important, because it means that the function must extend between landmarks – it cannot be defined only at certain discrete points and disappear in the regions between them.

If a function blows up (becomes infinite or non-differentiable) between points, we cannot use it to interpolate values between them. This is important because we are using the thin-plate spline as an interpolation function, inferring what happens between landmarks from data at given anatomical points. If it is unreasonable to interpolate, it is unreasonable to use the thin-plate spline for that purpose. It is also unreasonable to interpolate between far distant landmarks, just as it is unreasonable to extrapolate a linear regression far beyond the range of the observed data. If our landmarks are far apart, we have too few data to draw conclusions about what happens between them. For example, in Figure 6.2 we are assuming that the changes in regions between postcranial landmarks can be inferred from landmarks on the dorsal and ventral periphery. That assumption can be questioned, because if we actually had more landmarks in that region we might find abrupt changes – small regions where the grid dramatically compresses or expands. We are simply assuming that no such localized changes occur.

Another case in which it would be inappropriate to think of shape change as a deformation is when there is change concentrated at a single landmark. That is equivalent to a function with an abrupt change, which violates the assumption of continuity. Such discontinuities can be detected as displacement of one shape coordinate against a background of invariant points. That pattern may be rare, but one close to it has actually been found in data (Myers et al., 1996). In that study, mice (*Peromyscus maniculatus bairdii*) fed different diets were found to have skulls that differ only in the location of the tips of the incisors relative to the other skull landmarks. This is an extreme case of a Pinocchio effect (as discussed in Chapter 5). Such highly local changes should be ruled out before any deformation-based method is applied; if such highly localized change is found, it is better to rely on shape coordinates.

There is one other case in which a deformation-based approach might be unwise; when the interpolation spans a large amount of extra-organismal space – that is, when it is interpolating the changes over regions of "tissue" outside the organism. This can happen when landmarks are located at tips of long structures, or on structures that extend far laterally. Normally this is not a serious problem because we can simply avoid interpreting the changes in regions between those landmarks, except to say (perhaps) that the long bony structures are relatively elongated or reoriented more laterally. However, this can be a problem when multiple landmarks are located at tips of long structures and no other landmarks serve to pin down what is happening to the regions between them. It is possible to analyze the changes in relative position and length of those tips using shape coordinates, but it may not be wise to draw a grid interpolating changes at those tips to regions between them – there is no organismal tissue there.

If we do not have one of the special cases described above – that is, if we do not have evidence that some landmarks are largely independent of the others – then we can apply an interpolation function to understand changes between landmarks. Because the interpolation function is continuously differentiable, relative displacements of landmarks can be used to calculate the displacement of any location on the organism. These inferred displacements between landmarks can be illustrated using a variety of graphical styles; Figure 6.2 demonstrates the one most often used, a deformed grid in the style of D'Arcy Thompson (1942).

The physical metaphor

The mathematical basis for drawing the picture of the deformed grid is a metaphor – the bending of an idealized steel plate (Bookstein, 1989). According to this metaphor, displacements of landmarks in the X, Y plane (the plane in which we have drawn them in Figure 6.1) are visualized as if they were transferred to the Z-coordinate of an infinite, uniform and infinitely thin steel plate. That is, instead of depicting a landmark as displaced in some direction within the plane of this page, it is visualized as if it were displaced in the third dimension (out of this page).

The metal plate is constrained by little stalks that weld the landmarks in one shape to the landmarks in the other. This is difficult to draw because the imagery is inherently three-dimensional, so imagine two plates and place a configuration of landmarks on each. Now, put one plate above the other, and construct little stalks that attach a landmark on one plate to its homologue on the other plate. If a landmark in one shape is displaced a long distance relative to the other landmarks, construct a long stalk. Thus when the landmark is displaced a long distance in one direction (such as far anteriorly) the stalk is long; conversely, when displaced only a short distance the stalk is short. Therefore the stalks are of uneven lengths, and that unevenness means that one plate cannot be flat. The conformation that plate takes is determined by the relative heights of the stalks, and by the distances between them on the plate.

In some cases the plate simply tilts or rotates (it does not actually bend); in other cases the plate must actually bend, such as when a point in the middle is elevated higher than four surrounding points. That bending may be gentle or quite sharp. For real steel plates, the conformation of the plate tends to minimize the magnitude of bending over the whole plate (as well as the physical energy required to produce that bending). Here we use the expression *tends* to minimize the magnitude and energy of bending, because real steel plates may have flaws, and the situation is not a pure case of work against elasticity. In the ideal case, the bending energy depends solely on the distance between the points and the relative heights of the stalks, and the total amplitude of bending. If we consider two different deformed plates, both describing the same total overall amount of change (the same set of stalk heights) but one with the stalks proportionately closer together, the one that is bent between the more closely spaced points requires more energy than the one that is bent between more widely spaced points.

The bending energy depends on the spacing of the stalks because it is a function of the rate of change in the slope of the plate – i.e. whether the slope of the surface increases rapidly or slowly. In these terms, more energy is required when the slope of the surface changes at a higher rate (for the same net amplitude of change). Imagine a tall stalk surrounded by short ones, which induces a steep slope in the curvature of the plate. The steepness of that slope is proportional to the function being minimized – the rate of change in slope of the surface – and thus the function being minimized is a function of the second derivative (the slope of the surface is the first derivative) integrated over the whole surface of the plate. It can also be termed the integral of the quadratic variation over the plate.

To return from ideal plates to the analysis of a deformation, we now project the changes that were visualized as if in the Z-direction back into the X, Y plane (the plane of our landmark data). The idea of bending that had a physical meaning when we were talking about changes in the Z-direction is now reinterpreted as "spatially local information." This interpretation may not be intuitively obvious, but consider what a relatively rapid increase in slope means – that there are contrasting displacements of closely spaced points. When closely spaced points change in opposite directions it requires more energy to bend the plate between them; so there is an inverse relationship between the spatial scale of the change and its metaphorical bending energy. Minimization of bending energy is equivalent to minimization of spatially localized information.

It is always possible to envision changes as highly local by assuming that the plate flattens out immediately after rising, then rises again just at the next stalk, then flattens again, then rises again, etc. The argument against doing so is that this would be the most unparsimonious interpretation possible. By minimizing bending energy, we obtain a more parsimonious description of the change. We do not assume highly localized change unless the data demand doing so.

Uniform and non-uniform components of a deformation

Some transformations require no bending energy at all; these are equivalent to tilting or rotating the plate. These are often called *affine* or *uniform* transformations, meaning that they leave parallel lines parallel. The terms "affine" and "uniform" are both used to describe the same component of a deformation; "affine" is favored by mathematicians, but "uniform" appears more often in the geometric morphometric literature. Consequently, we will use "uniform" for this component and "non-uniform" for its complement. In our example (Figure 6.2), if the entire fish simply elongates relative to its depth, without any disproportionate lengthening of one region relative to another, that is a uniform elongation. Uniform elongation is equivalent to uniform narrowing, as should be recalled from our discussion of shape variables in previous chapters. Because it is uniform, meaning that the same change occurs everywhere, we need only one descriptor for the change of the whole organism. In contrast, the *non-uniform* or *non-affine* deformations (which involve the metaphorical bending) have regionally differentiated effects.

A deformation can be broken down into uniform and non-uniform components, as in Figure 6.3. Most real biological transformations will have both uniform and non-uniform components. These components are computed separately, so we describe them separately (first the uniform, then the non-uniform), but it is important to bear in mind that a complete description, and an accurate illustration, requires specifying *all* the components.

Uniform (affine) components

There are six distinct types of uniform deformations for landmarks in two dimensions, and they are independent of each other (meaning that they are mutually orthogonal). Figure 6.4 shows these six operations carried out on a square configuration of landmarks. The

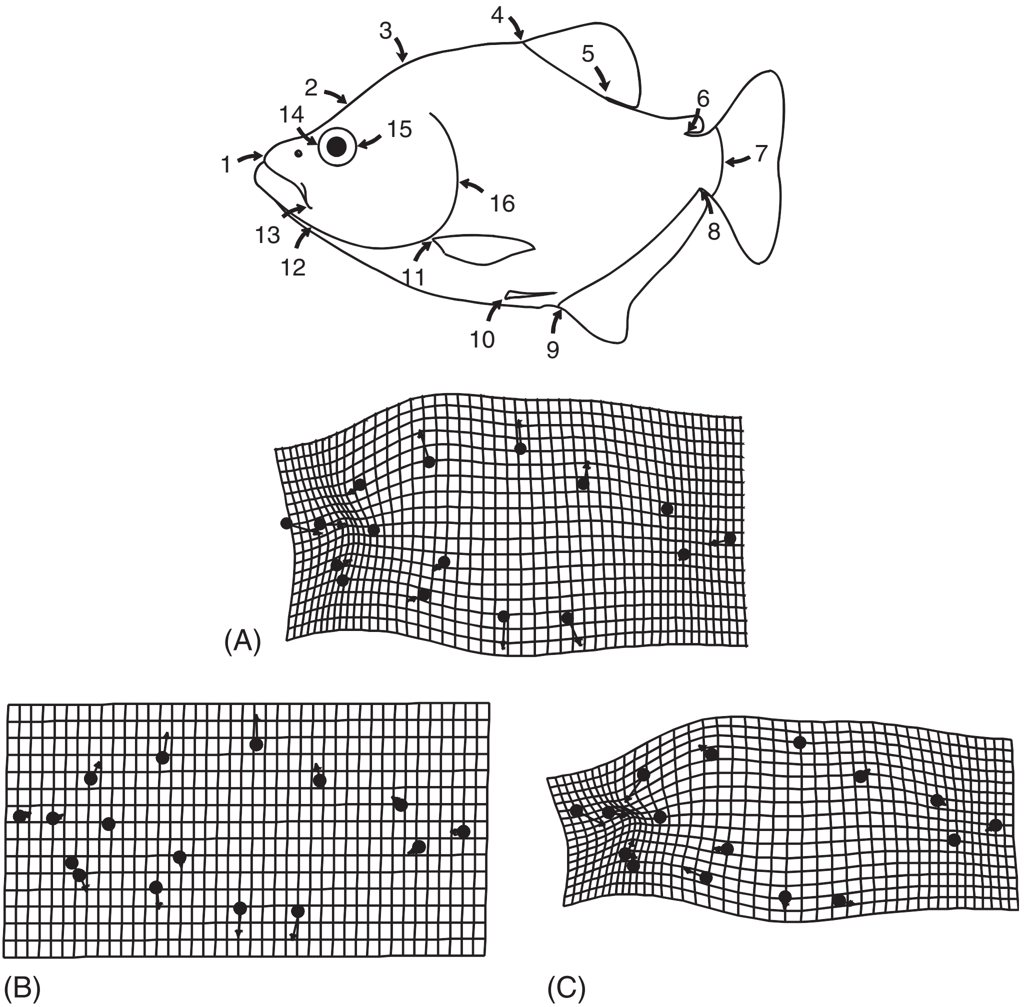


Figure 6.3 Ontogenetic change in body shape of *S. gouldingi*, depicting: (A) the total deformation and its two components; (B) uniform component; (C) non-uniform component.

first four are the familiar ones that do not alter shape: translation along two perpendicular axes (Figures 6.4A, 6.4B), scaling (Figure 6.4C) and rotation (Figure 6.4D). These are all used in superimposing shapes. The other two uniform deformations do alter shape: compression/dilation (Figure 6.4E) and shear (Figure 6.4F). Compression/dilation refers to the case in which one direction has expanded (the vertical or *Y*-direction in Figure 6.4E) while the other has contracted (the horizontal or *X*-direction). Shearing refers to translating landmarks along one axis by a distance proportional to their location along the other axis.

Because compression/dilation and shear alter shape whereas translation, rotation and scaling do not, it is common to talk about the two that alter shape without mentioning the ones that do not. All of them need to be tracked, so we will refer to compression/dilation

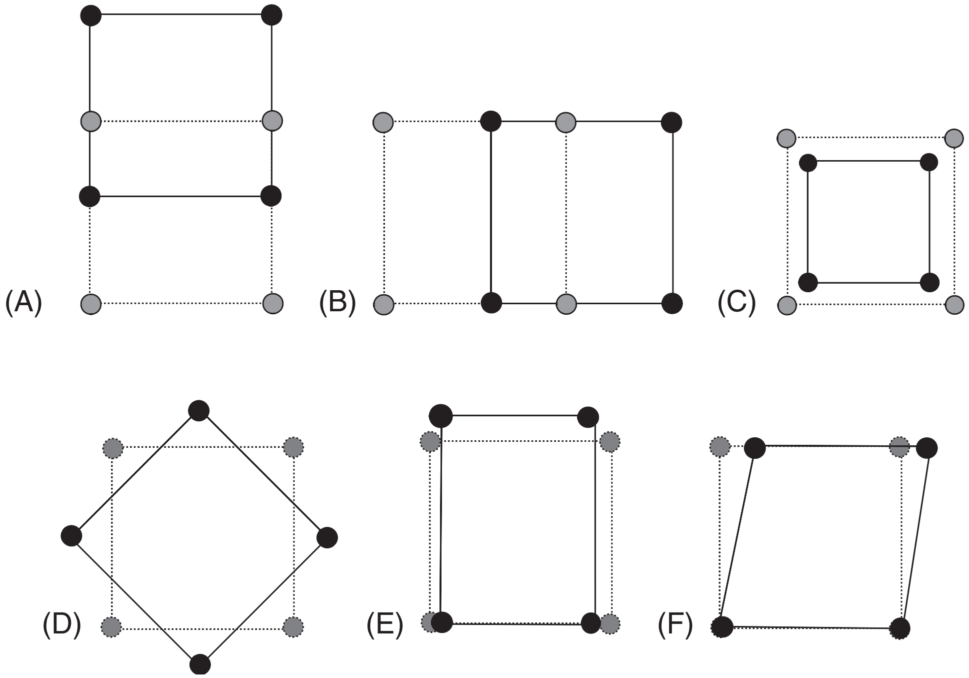


Figure 6.4 The six uniform (affine) transformations: (A) translation along the vertical axis; (B) translation along the horizontal axis; (C) scaling; (D) rotation; (E) compression/dilation; (F) shearing. The original (or reference) square is shown with dotted lines, while the deformed shape is shown with solid lines.

and shear as the *explicit uniform deformations* or *explicit uniform terms* because they are the ones explicitly tracked. We will refer to the others as the *implicit uniform deformations* or *implicit uniform terms*. They are implicit because they can be mathematically determined from the superimposition method used, the explicit uniform components, and the non-uniform components of a deformation – they are the translation, rotation and scaling that must have been carried out. Both explicit and implicit uniform terms are needed, in addition to the non-uniform terms, to draw the deformation correctly.

Each deformation has an inverse. Applying the inverse of a deformation is equivalent to traveling backwards along the path that was taken until we arrive back at the starting point. We can think of the deformation in terms of a $2K$-dimensional vector (i.e. two dimensions per landmark). There would be a vector at each landmark indicating the direction in which that particular landmark will be mapped under the deformation (although there are only $2K - 4$ independent dimensions). In the inverse of the deformation, the directions of the arrows would be reversed. The inverse of a translation is the same magnitude of translation in the opposite direction (negative X instead of positive X). Similarly, we can represent rotation as an angular displacement so its inverse is a *negative* angular displacement (counterclockwise instead of clockwise). Scaling is slightly different because it involves multiplication (whereas translations and rotations could be treated as additions). Scaling is multiplication by a factor F; its inverse is multiplication by the inverse of $F(1/F)$.

Unfortunately, the algebraic descriptions of the last two deformations and their inverses are not quite as simple (as we will see below). Graphically, we can see that the inverse of compression/dilation involves a reversal of which axis is compressed and which is dilated, and that the inverse of a shear is a shear of the same amount along the same axis in the opposite direction.

Calculating the shear and compression/dilation terms

Here we present the mathematical derivation of formulae for calculating the uniform components of a deformation that changes shape. Unlike the formulae for computing the non-uniform part of a shape change, which have been stable over the last decade, the formulae for computing the uniform part have changed repeatedly. Over the last several years, the uniform component has been computed using the formulae presented by Bookstein (1996). These, which are based on the Procrustes distance, are the ones we present here. We begin with a conceptual framework for Bookstein's derivation of the current formulae; then follow that with the full mathematical details.

Conceptual framework The goal of this derivation is to find a unit vector that describes the direction of deformation at each landmark due to shearing or compression/dilation, followed by a Procrustes generalized least squares (GLS) superimposition of the deformed shape back onto the original (undeformed) one. This represents what we measure in data: a deformation followed by a superimposition operation. Thus, both mappings must be taken into account. When we are done, we will have a set of unit vectors that describe the deformation under shearing or under compression/dilation. We can then take the dot product of the observed deformation with the unit vectors to obtain the component of the observed deformation lying along the shear or compression/dilation vectors. These are what we have been calling the explicit uniform components of the deformation.

Notice that we are taking a verbal description of the situation, turning the verbal statement into two mathematical operations or mappings (shear or compression/dilation, followed by the superimposition), then using those mappings to determine the direction of the vectors describing the deformation. That allows us to calculate components of any deformation along those desired directions. What might not be obvious yet is that vectors describing the uniform deformations depend on only one form – the one that we are modeling as deformed, which we will call the reference form (the other is the target). This terminology should be familiar – the reference form is the same one that we discussed in Chapter 4. If you do not wish to read further, you do not have to. You can go directly to the section on decomposing the non-uniform (non-affine) component.

Although we now have a general idea of the procedure, there are still a few ideas that need to be added. The first is the idea of complex number notation for landmark locations, which is often used in mathematical derivations (see Dryden and Mardia, 1998, for example). Consider a landmark configuration consisting of K landmarks in two dimensions, which we will call $\mathbf{Z}$, the reference form. Mathematically, we will say:

$$\mathbf{Z} = \left\{Z_j, Z_j = \left(X_j, Y_j\right)\right\}_{j=1}^{K} \qquad \textbf{(6.1)}$$

which means that $\mathbf{Z}$ is a set of K pairs of landmark positions Z_j, or (X_j, Y_j). It is a useful mathematical shortcut to think of Z_j as being a complex number $Z_j = X_j + iY_j$, where

i is the square root of minus one. Complex number notation is often used in texts on the statistics of shape, so understanding it is useful.

The next idea is to require that the reference form be rotated to a principal axis alignment, so that $\sum_j X_j Y_j = 0$, which will later simplify the mathematics (but may pose problems for aligning specimens in some software, discussed below). The summation $\sum_j$ is from $j = 1$ to $j = K$, and all the summations in the derivation are likewise over all K landmarks. We are also going to assume that the reference has a centroid size of one, so that $\sum_j (X_j^2 + Y_j^2) = 1$, and a centroid position of $(0, 0)$, so that $\sum_j X_j = 0$ and $\sum_j Y_j = 0$.

Mathematical derivations Let us consider the two functions of interest: shear, which we will call $\mathbf{S_1}(\lambda)$, and compression/dilation, which we will call $\mathbf{S_2}(\lambda)$ (λ describes the magnitude of the mapping). We will be taking the limit as $\lambda \to 0$ at the end of this derivation, so terms including λ^2 will be discarded. The mappings from a reference form $\mathbf{Z}$ to a target form $\mathbf{Z'}$ under these operations are as follows:

$$\mathbf{S_1}(\lambda)\colon \mathbf{Z} \to \mathbf{Z'}, \mathbf{Z'} = \left\{Z'_j = \left(X_j + \lambda Y_j, Y_j\right)\right\}_{j=1}^{K} \tag{6.2}$$

$$\mathbf{S_2}(\lambda)\colon \mathbf{Z} \to \mathbf{Z'}, \mathbf{Z'} = \left\{Z'_j = \left(X_j, Y_j + \lambda Y_j\right)\right\}_{j=1}^{K} \tag{6.3}$$

You can probably convince yourself that $\mathbf{S_1}$ describes a shear; the X-coordinates of each point are displaced a distance proportional to their Y-axis position relative to the centroid. Similarly, you should be able to recognize that $\mathbf{S_2}$ describes an expansion of the landmarks along the Y-axis. We do not need to worry about modeling the contraction along the X-axis, even though it must also be occurring, because the Procrustes GLS superimposition will take care of that.

If $\mathbf{Z}$ and $\mathbf{Z'}$ are both centered (i.e. have a centroid position of zero), then the Procrustes superimposition may be approximated as the multiplication of $\mathbf{Z'}$ by the complex factor $\mathbf{P_{Z'}}$, where:

$$\mathbf{P_{Z'}} = \frac{\mathbf{Z}\overline{\mathbf{Z'}}}{\mathbf{Z'}\overline{\mathbf{Z'}}} \tag{6.4}$$

and the expression $\overline{\mathbf{Z'}}$ refers to the complex conjugate of the complex vector $\mathbf{Z'}$ representing the landmark configuration after the compression/dilation. The Procrustes superimposition of $\mathbf{Z'}$ on $\mathbf{Z}$ is thus $\mathbf{P_{Z'}Z'}$. To get the vectors that describe the uniform deformation, we just subtract the starting position $\mathbf{Z}$ from $\mathbf{P_{Z'}Z'}$ and then divide through by the magnitude of the deformation λ, yielding $(\mathbf{P_{Z'}Z'} - \mathbf{Z})/\lambda$ as the set of vectors describing the deformation.

Further derivation of the uniform components To find $\mathbf{P_{Z'}}$ for the $\mathbf{S_2}(\lambda)$ mapping (compression/dilation), we note that the numerator of $\mathbf{P_{Z'}}$ is:

$$\mathbf{Z}\overline{\mathbf{Z'}} = \sum_j \left(X_j + iY_j\right) \times \left(X_j - i\left(Y_j + \lambda Y_j\right)\right) \tag{6.5}$$

which expands to:

$$= \sum_j \left(X_j^2 - iX_jY_j - iX_j\lambda Y_j + iX_jY_j - \left(iY_j\right)^2 - \left(iY_j\right)^2\lambda\right) \tag{6.6}$$

Because $i^2 = -1$ and the products of $X_j Y_j$ sum to zero (under the alignment specified earlier), we can simplify this to:

$$= \sum_j \left(X_j^2 + Y_j^2 + Y_j^2\lambda\right) \tag{6.7}$$

Now add the constraint that $\sum_j (X_j^2 + Y_j^2) = 1$ because we scaled the reference to unit centroid size, and we have:

$$\mathbf{Z}\overline{\mathbf{Z}'} = 1 + \lambda \sum_j Y_j^2 \tag{6.8}$$

Now we simplify the denominator of $\mathbf{P}_{\mathbf{Z}'}$:

$$\begin{aligned}
\mathbf{Z}'\overline{\mathbf{Z}'} &= \sum_j \left(X_j + iY_j\,(1+\lambda)\right) \times \left(X_j - iY_j(1+\lambda)\right) && (6.9)\\
&= \sum_j \left(X_j^2 + Y_j^2(1+\lambda)^2\right) = \sum_j X_j^2 + Y_j^2\left(1 + 2\lambda + \lambda^2\right) && (6.10)\\
&= \sum_j X_j^2 + Y_j^2 + 2Y_j^2\lambda + Y_j^2\lambda^2 && (6.11)
\end{aligned}$$

As mentioned before, $\sum_j (X_j^2 + Y_j^2) = 1$, and terms including λ^2 can be discarded in the limit of small λ, so that:

$$\mathbf{Z}'\overline{\mathbf{Z}'} \cong 1 + 2\lambda \sum_j Y_j^2 \tag{6.12}$$

This leaves us with:

$$\mathbf{P}_{\mathbf{Z}'} = \frac{\mathbf{Z}\overline{\mathbf{Z}'}}{\mathbf{Z}'\overline{\mathbf{Z}'}} = \frac{\left(1 + \lambda \sum_j Y_j^2\right)}{\left(1 + 2\lambda \sum_j Y_j^2\right)} \tag{6.13}$$

We can now expand the term $1/(1 + 2\lambda \sum_j Y_j^2)$ as $1 - 2\lambda \sum_j Y_j^2$, keeping only first order terms in λ for this power series expansion. This gives us:

$$\mathbf{P}_{\mathbf{Z}'} = \frac{\mathbf{Z}\overline{\mathbf{Z}'}}{\mathbf{Z}'\overline{\mathbf{Z}'}} \cong \left(1 + \lambda \sum_j Y_j^2\right)\left(1 - 2\lambda \sum_j Y_j^2\right) \cong 1 - \lambda \sum_j Y_j^2 \tag{6.14}$$

to first order in λ.

Now we can calculate the landmark coordinates after the operation of the compression/dilation ($\mathbf{S}_2(\lambda)$) and Procrustes superimposition (which is just a multiplication by $\mathbf{P}_{\mathbf{Z}'}$, since $\mathbf{Z}'$ is already centered):

$$\mathbf{P}_{\mathbf{Z}'}\mathbf{Z}' = \left(1 - \lambda \sum_j Y_j^2\right) \times Z' = \left\{ Z_j = \left(X_j \left(1 - \lambda \sum_j Y_j^2\right)\right), \right.$$
$$\left. \left(\left(Y_j + \lambda Y_j\right) \left(1 - \lambda \sum_j Y_j^2\right)\right)\right\}_{j=1}^{K} \tag{6.15}$$

The vector describing the displacement from $\mathbf{Z}$ to $\mathbf{P}_{\mathbf{Z}'}\mathbf{Z}'$ is then:

$$\mathbf{P}_{\mathbf{Z}'}\mathbf{Z}' - \mathbf{Z} = \left\{ \left(\left(X_j \left(1 - \lambda \sum_j Y_j^2\right) - X_j\right), \left(\left(Y_j + \lambda Y_j\right) \left(1 - \lambda \sum_j Y_j^2\right) - Y_j\right)\right)\right\}_{j=1}^{K} \tag{6.16}$$

$$= \left\{ \left(\left(-X_j \lambda \sum_j Y_j^2 \right), \left(\lambda Y_j - \lambda Y_j \sum_j Y_j^2 - \lambda^2 Y_j \sum_j Y_j^2 \right)\right)\right\}_{j=1}^{K} \tag{6.17}$$

Noting that $\lambda^2 \cong 0$, we can simplify this to:

$$= \left\{ \left(\left(-X_j \lambda \sum_j Y_j^2 \right), \left(\lambda Y_j - \lambda Y_j \sum_j Y_j^2 \right)\right)\right\}_{j=1}^{K} \tag{6.18}$$

$$= \lambda \left\{ \left(X_j \left(- \sum_j Y_j^2 \right), Y_j \left(1 - \sum_j Y_j^2 \right)\right)\right\}_{j=1}^{K} \tag{6.19}$$

We now define $\gamma = \sum_j Y_j^2$ and $\alpha = 1 - \sum_j Y_j^2 = \sum_j X_j^2$, so that $\gamma + \alpha = 1$. After making these substitutions and dividing through by λ, we have:

$$\mathbf{V}_2 = \frac{(\mathbf{P}_{Z'} - \mathbf{Z}')}{\lambda} = \left\{(-\gamma X_j, \alpha Y_j)\right\}_{j=1}^{K} \tag{6.20}$$

which is the vector of the displacements at each landmark point (X_j, Y_j) produced by the mapping S_2 per unit of λ. All we need to do now is to normalize this set so that the length of the vector is one.

The magnitude of this vector is:

$$\sqrt{\sum_j \left(\gamma^2 X_j^2 + \alpha^2 Y_j^2 \right)} \tag{6.21}$$

Using the definitions of α and γ to rearrange this and simplify it, we get:

$$= \sqrt{\gamma^2 \sum_j X_j^2 + \alpha^2 \sum_j Y_j^2} = \sqrt{\gamma^2 \alpha + \alpha^2 \gamma} = \sqrt{\alpha\gamma\,(\alpha + \gamma)} = \sqrt{\alpha\gamma} \tag{6.22}$$

So if we normalize $\mathbf{V}_2$, we get:

$$\mathbf{V}_2' = \frac{\mathbf{V}_2}{\sqrt{\alpha\gamma}} = \left\{\left(-\frac{\gamma}{\sqrt{\alpha\gamma}}X_j, \frac{\alpha}{\sqrt{\alpha\gamma}}Y_j\right)\right\}_{j=1}^{K} \tag{6.23}$$

$$= \left\{\left(-\sqrt{\frac{\gamma}{\alpha}}X_j, \sqrt{\frac{\alpha}{\gamma}}Y_j\right)\right\}_{j=1}^{K} \tag{6.24}$$

which is now a unit vector describing a compression/dilation operation followed by Procrustes superimposition.

Similarly, we start with a shearing operation, $\mathbf{S}_1(\lambda)$, and corresponding Procrustes superimposition, $\mathbf{P}_{\mathbf{Z}'}$, to find the unit vector corresponding to these operations. First we need to find $\mathbf{P}_{\mathbf{Z}'}$ for the $\mathbf{S}_1(\lambda)$ mapping:

$$\mathbf{Z}\overline{\mathbf{Z}'} = \sum_j \left(X_j + iY_j\right) \times \left(X_j + \lambda Y_j - iY_j\right) \tag{6.25}$$

$$= \sum_j \left(X_j^2 + Y_j^2 + X_jY_j\lambda + iY_j^2\lambda\right) \tag{6.26}$$

As before, $\sum_j \left(X_j^2 + Y_j^2\right) = 1$, $\sum_j X_jY_j = 0$ and $\sum_j Y_j^2 = \gamma$; therefore:

$$\mathbf{Z}\overline{\mathbf{Z}'} = 1 + i\gamma\lambda \tag{6.27}$$

Also:

$$\mathbf{Z}'\overline{\mathbf{Z}'} = \sum_j \left(X_j + \lambda Y_j + iY_j\right) \times \left(X_j + \lambda Y_j - iY_j\right) \tag{6.28}$$

$$\sum_j \left(X_j + \lambda Y_j\right)^2 + Y_j^2 = \sum_j \left(X_j^2 + 2\lambda X_jY_j + \lambda^2 Y_j^2 + Y_j^2\right) = 1 \tag{6.29}$$

Therefore:

$$\mathbf{P}_{\mathbf{Z}'} = \frac{\mathbf{Z}\overline{\mathbf{Z}'}}{\mathbf{Z}'\overline{\mathbf{Z}'}} = \frac{\mathbf{Z}\overline{\mathbf{Z}'}}{1} = 1 + i\gamma\lambda \tag{6.30}$$

Now we can simplify:

$$\mathbf{V}_1 = \frac{\mathbf{P}_{\mathbf{Z}'}\mathbf{Z}' - \mathbf{Z}}{\lambda} = \frac{(1 + i\gamma\lambda)(X_j + \lambda Y_j + iY_j) - (X_j + iY_j)}{\lambda} \tag{6.31}$$

$$= \frac{X_j + \lambda Y_j + iY_j + i\gamma\lambda X_j + i\gamma\lambda^2 Y_j + i^2\gamma\lambda Y_j - X_j - iY_j}{\lambda} \tag{6.32}$$

$$= \frac{\lambda Y_j + i\gamma\lambda X_j - \gamma\lambda Y_j}{\lambda} = Y_j + i\gamma X_j - \gamma Y_j \tag{6.33}$$

This leads to the series of coordinate pairs:

$$= (Y_j(1 - \gamma), \gamma X_j) \tag{6.34}$$

or

$$\mathbf{V}_1 = (\alpha Y_j, \gamma X_j) \tag{6.35}$$

Dividing this by its magnitude to normalize it yields:

$$\sqrt{\sum_j \left(\alpha^2 Y_j^2 + \gamma^2 X_j^2\right)} = \sqrt{\alpha^2 \sum_j Y_j^2 + \gamma^2 \sum_j X_j^2} = \sqrt{\alpha^2\gamma + \gamma^2\alpha} \tag{6.36}$$

$$= \sqrt{\alpha\gamma\,(\alpha+\gamma)} = \sqrt{\alpha\gamma} \tag{6.37}$$

so the unit vector $\mathbf{V}_1'$is:

$$\mathbf{V}_1' = \left\{\left(\frac{\alpha Y_j}{\sqrt{\alpha\gamma}}, \frac{\gamma X_j}{\sqrt{\alpha\gamma}}\right)\right\}_{j=1}^{K} = \left\{\left(\sqrt{\frac{\alpha}{\gamma}}Y_j, \sqrt{\frac{\gamma}{\alpha}}X_j\right)\right\}_{j=1}^{K} \tag{6.38}$$

which may now be used to determine the shear component of the uniform deformation.

Some software packages will give you α and γ as used in the calculation of the uniform component, others may give you the unit vectors instead. The expressions above (Equations 6.24 and 6.38) are for coordinates of the unit vectors for shear and compression/dilation for a reference form rotated to principal axis orientation. It turns out to be straightforward to rotate them to unit vectors to match any reference orientation preferred by a researcher, although some programs may not offer this option, meaning that the reference may be oddly oriented by the software.

Calculating uniform components based on other superimpositions

The approach taken in the above derivation was to determine the unit vectors that would result from a shear or compression/dilation of a reference form, followed by Procrustes superimposition back onto the reference form. It is also possible to determine the unit vectors produced by a shear or compression/dilation of a reference, followed by sliding baseline registration (SBR) or a two-point registration that yields Bookstein coordinates (BC). These unit vectors and specimens can then be used in SBR or BC to calculate the uniform components of the deformation, just as we did with those in Procrustes superimposition. Estimates of the explicit uniform components under SBR are identical to those derived from the Procrustes-based method presented here. This is not surprising, since the Procrustes superimposition differs from SBR only in the implicit uniform deformations (assuming that the Procrustes superimposition, like SBR, is performed with centroid size set to one, two superimpositions differ only in the rotation and translation terms). Thus, a deformation displayed by a Procrustes superimposition shows the same change in *shape* as the deformation displayed by SBR – the differences between them are due to the implicit deformations, and do not alter shape. Deformations shown by BC differ from those in Procrustes superimposition in scale as well as rotation and translation, but these are still implicit uniform terms. Likewise, RFTRA differs from the other superimpositions only in the implicit uniform terms.

Decomposing the non-uniform (non-affine) component

The non-uniform part of a deformation differs from the uniform in that it does not leave the sides of a square parallel. However, like the uniform part, the non-uniform can be further decomposed into a set of orthogonal components. The decomposition of the non-uniform deformation is based on the thin-plate spline interpolation function, and produces components called *partial warps*. We first describe an intuitive introduction to partial warps, then a more mathematical one.

An intuitive introduction to partial warps

The non-uniform component describes changes that have a location and spatial extent on the organism; they are not the same everywhere. They describe spatially graded phenomena such as anteroposterior growth gradients, and more highly localized changes such as the elongation of the snout relative to the eye. The notion of spatial scale is central to the analysis, so we need an intuitive notion of spatial scale. In general (but imprecise) terms, a change at small spatial scale is one confined to a small region of an organism. To refine that idea, and develop a firmer grasp of the concept, we show several components at progressively smaller spatial scales (Figure 6.5).

Figure 6.5A shows a component at large spatial scale that, while broadly distributed, is not the same everywhere (so it is not uniform). The particular example shown in Figure 6.5A is the elongation of the midbody relative to the more cranial and caudal regions. A more localized change, confined to the posterior region of the body, is shown in Figure 6.5B – a shortening of the region between the dorsal and adipose fins relative to the dorsal fin and caudal peduncle. Because more distant landmarks are not involved in the change, it is more localized than the one shown in Figure 6.5A. Another localized change is shown in Figure 6.5C, this time confined to the cranial region. This is a shortening of the postorbital region relative to the regions just anterior and posterior.

The components we have described above and depicted in Figure 6.5 are partial warps, but to draw them we had to specify their orientation (we drew them as oriented along the anteroposterior body axis). That orientation is not actually specified by the partial warps themselves; rather, it is provided by a two-dimensional vector, the *partial warp scores*. There is one two-dimensional vector per partial warp. These scores express the contribution that each partial warp makes to the total deformation. The scores have an *X*- and *Y*-component, and indicate the direction of the partial warp. The idea of direction or orientation should be familiar from previous chapters. In Figure 6.6 we show one partial warp (that depicted in Figure 6.5B) multiplied by three different vectors. It may be easiest to see the directions by looking at the orientation of the vectors at landmarks. Figure 6.6A shows the partial warp oriented horizontally, which in our case corresponds to the *X*-direction, so the coefficient of the *X*-component is large and that of the *Y*-component is negligible. In contrast, Figure 6.6B shows the vector with a negligible *X*-component and a large *Y*-component. Figure 6.6C shows the vector with *X*- and *Y*-components of equal magnitudes.

We have described partial warps one at a time, but a complete description (and interpretation) requires combining them all. Taken separately, partial warps are purely geometric constructs – a function of the location and spacing of the landmarks of the reference form. They are obtained by a geometric decomposition of the landmarks of the reference form

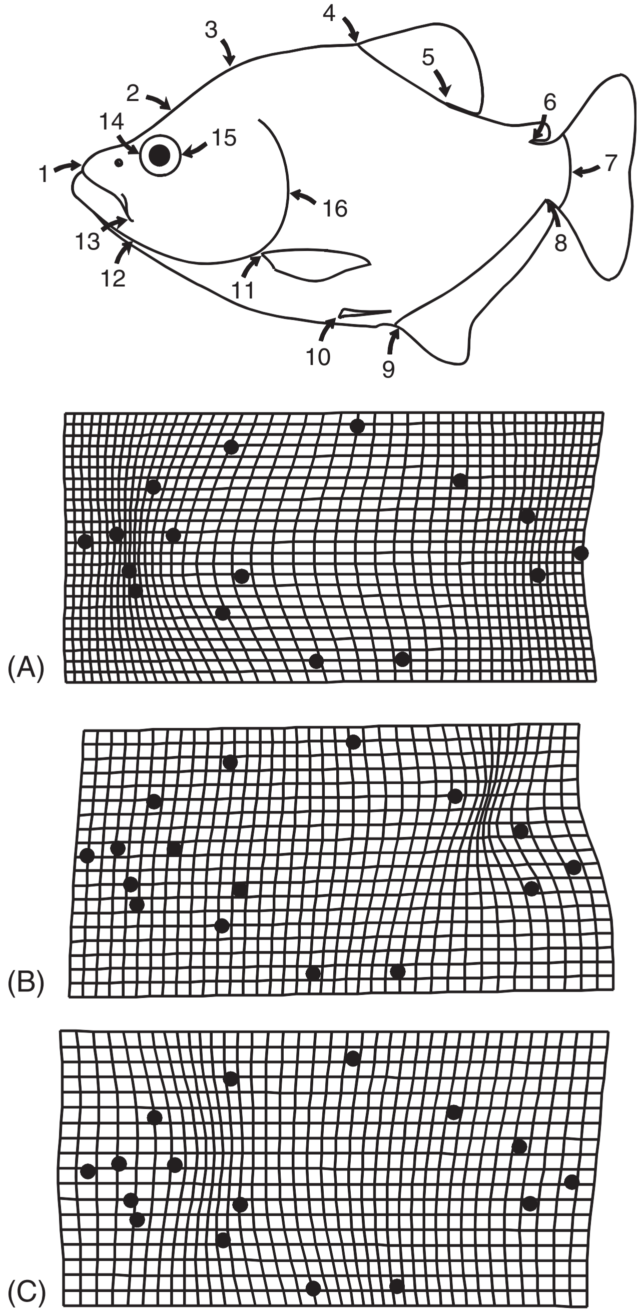


Figure 6.5 Three components of the non-uniform deformation, called partial warps. (A) A partial warp at large spatial scale, depicting an expansion of the midbody relative to the head and caudal body; (B) a partial warp at moderate to small spatial scale, depicting a contraction of the region between dorsal and adipose fins relative to the length of the dorsal fin and caudal peduncle; (C) another partial warp at moderate to small spatial scale, depicting a shortening of the postorbital region relative to the preorbital head and anterior postcranial body.

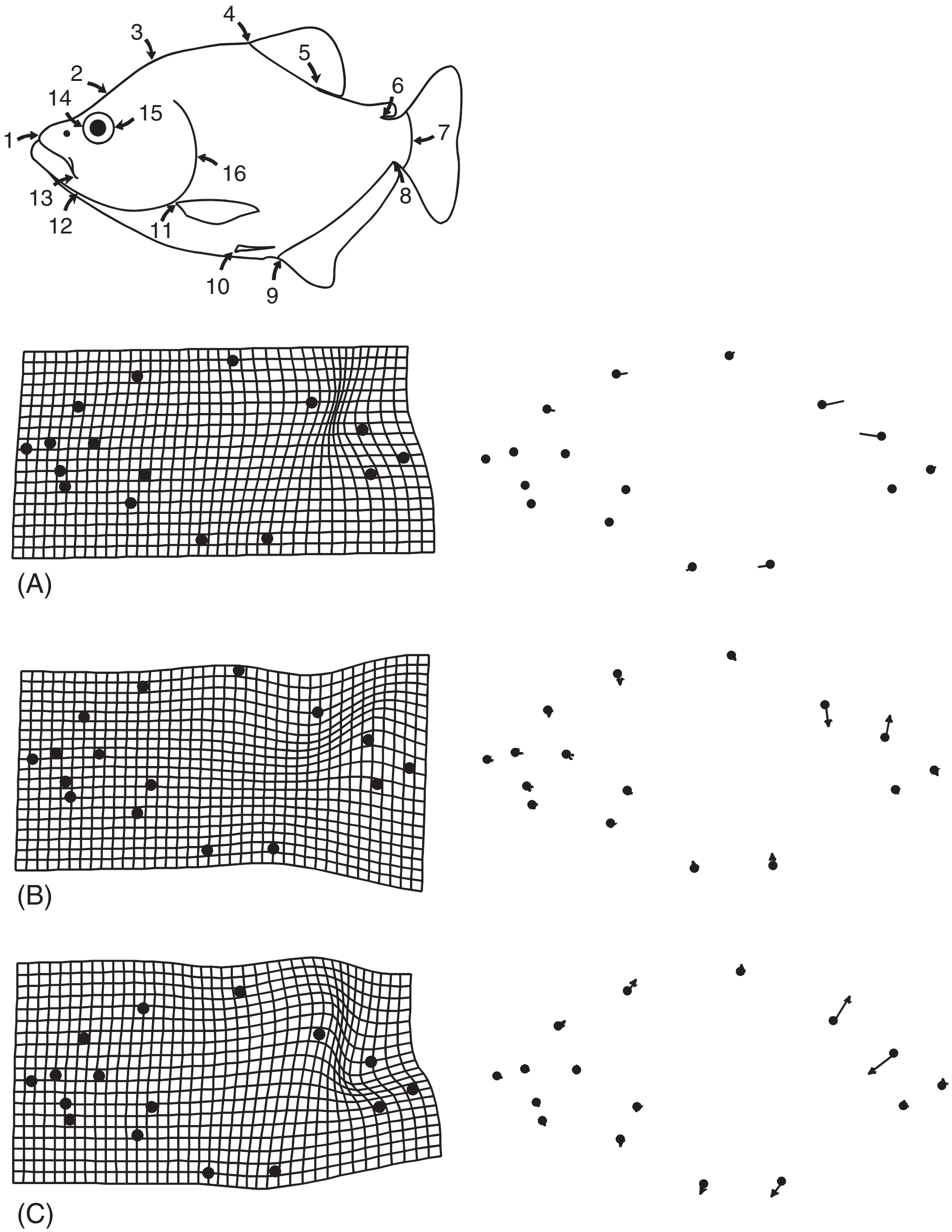


Figure 6.6 One partial warp, oriented in three directions: (A) along the *X*-axis; (B) along the *Y*-axis; (C) equally along *X*- and *Y*-axes. Due to the orientation of our landmark coordinates, the *X*-direction corresponds to the anteroposterior axis, and the *Y*-direction corresponds to the dorsoventral axis. To make it easier to see the direction in which the partial warps are oriented, we also display them by vectors of relative landmark displacements.

(as explained in detail in the next section). Although they provide a basis for the tangent space, they cannot be interpreted except in these abstract terms – we cannot, for example, say that one part of the change in the ontogeny of the fish is a shortening of the region between dorsal and adipose fins relative to the dorsal fin and caudal peduncle. That is a component of the deformation, not a component of an ontogeny. Only by looking at the total deformation can we say where change occurs. We have discussed them separately only to explain what they are (none of the currently available software draws them separately – to produce Figures 6.5 and 6.6 we had to use specialized software).

To summarize our intuitive presentation of spatial scale, we repeat our major points. First, any non-uniform deformation can be decomposed into a series of components (partial warps) at progressively smaller spatial scales. Each component describes a pattern of relative landmark displacements, based on the spacing and location of landmarks in the reference form. Each partial warp is multiplied by a two-dimensional vector (the partial warp scores) that measures the contribution made by the partial warp (in each direction) to the total deformation. We now present a more technical introduction to the thin-plate spline.

An algebraic introduction to partial warps

Algebraically, partial warps are obtained by eigenanalysis of the *bending energy matrix*. Eigenanalysis may be familiar from a quite different context, for example, principal components analysis, where it is used to extract eigenvectors (PCs) of the variance–covariance matrix of measurements. The exact same mathematics is involved in calculating the partial warps; the difference lies in the matrix being analyzed. Rather than extracting eigenvectors of a variance–covariance matrix, we instead extract them from the bending-energy matrix. (We will discuss eigenanalysis further in Chapter 7; here we focus on the derivation of the bending energy matrix.)

The idea behind the thin-plate spline is that it will approximate the observed deformation by a linear combination of a function that is the smoothest available and that fully describes the observed deformation. The function satisfying that pair of requirements has the form:

$$Z(X, Y) = \mathbf{U}(R) = -R^2 \ln R^2 \qquad (\mathbf{6.39})$$

where R is the distance between a pair of landmarks in the reference configuration (scaled to unit centroid size). This particular function satisfies the biharmonic equation:

$$\Delta^2\mathbf{U} = \left(\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{\mathrm{d}^2}{\mathrm{d}y^2}\right)^2 \mathbf{U}(R) \propto \delta_{(0,0)} \qquad (\mathbf{6.40})$$

where $\delta_{(0,0)}$ is the generalized function, or delta function, which is defined to be zero everywhere except at $X = 0$, $Y = 0$, with the odd requirement that:

$$\int \left(\delta_{(0,0)}\mathrm{d}x\,\mathrm{d}y\right) = 1 \qquad (\mathbf{6.41})$$

The delta function is oddly behaved, but mathematically tremendously useful. It is sometimes called a functional, rather than a function.

U is said to be the fundamental solution of the biharmonic equation, which is the equation for the shape of a thin steel plate lifted to a height $Z(X, Y)$ above the (X, Y)-plane. This is because the bending energy (BE) of the steel plate at a point (X, Y) is given by:

$$\left(\frac{\mathrm{d}^2\mathbf{U}}{\mathrm{d}x^2}\right)^2 + 2\left(\frac{\mathrm{d}^2\mathbf{U}}{\mathrm{d}x\,\mathrm{d}y}\right)^2 + \left(\frac{\mathrm{d}^2\mathbf{U}}{\mathrm{d}y^2}\right) = BE(X, Y) \tag{6.42}$$

and the total bending energy of the entire plate is:

$$\int \left(BE\,(X, Y)\,\mathrm{d}x\,\mathrm{d}y\right) \tag{6.43}$$

which is the bending energy at each point integrated over the entire surface. The choice of $\mathbf{U}(R)$ minimizes this total bending energy.

For biological purposes, we do not really care about the bending energy of a steel plate. Rather, we care about the connection between bending energy and the curvature of the plate (and their connection to spatial scale). Minimizing bending energy minimizes the curvature of the plate, so when we fit a linear combination of the $\mathbf{U}(R)$ function to our data, we are fitting a function that minimizes the amount of curvature needed to model the observed deformations.

Suppose we want a linear combination of $\mathbf{U}(R)$ values, centered on each of the K landmarks of our reference form (because we are describing a deformation, we are talking about changes relative to a reference). We need to describe deformations in the X and Y directions, so we form the following linear combinations:

$$f_X\,(X, Y) = A_{X1} + A_{XX}X + A_{XY}Y + \sum_{i=1}^{K} W_{Xi}\,\mathbf{U}(X - X_i, Y - Y_i) \tag{6.44}$$

$$f_Y\,(X, Y) = A_{Y1} + A_{YX}X + A_{YY}Y + \sum_{i=1}^{K} W_{Yi}\,\mathbf{U}(X - X_i, Y - Y_i) \tag{6.45}$$

where $f_X(X, Y)$ and $f_Y(X, Y)$ are the spline functions that describe the deformations along the X- and Y-directions relative to the reference form, and W_{Xi} and W_{Yi} are weights of the functions $\mathbf{U}(X - X_i, Y - Y_i)$, centered on the landmark locations of the reference (X_i, Y_i). The A terms describe uniform (or affine) deformations of the target, using what is known as the six-component uniform model. We need to include those A terms at this stage, but will discard them later in favor of the two uniform components discussed earlier (Equations 6.24 and 6.38).

Fitting the functions to the observed deformations is a standard problem in systems of linear equations; we can thus cast the problem into matrix form. We form a $(K + 3) \times 2$ matrix **V** of the observed deformations at each of the K landmarks, where the deformation at the ith landmark is denoted (X_i', Y_i'):

$$\mathbf{V} = \begin{bmatrix} X_1' & Y_1' \\ X_2' & Y_2' \\ \vdots & \vdots \\ X_K' & Y_K' \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} f_X(X_1, Y_1) & f_Y(X_1, Y_1) \\ f_X(X_2, Y_2) & f_Y(X_2, Y_2) \\ \vdots & \vdots \\ f_X(X_K, Y_K) & f_Y(X_K, Y_K) \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \mathbf{LW} \tag{6.46}$$

where **LW** is the product of two matrices **L** and **W**. **L** is the $(K+3) \times (K+3)$ matrix:

$$\mathbf{L} = \begin{bmatrix} \mathrm{U}(0) & \mathrm{U}(R_{1,2}) & \mathrm{U}(R_{1,3}) & \cdots & \mathrm{U}(R_{1,K}) & 1 & X_1 & Y_1 \\ \mathrm{U}(R_{2,1}) & \mathrm{U}(0) & \mathrm{U}(R_{2,3}) & \cdots & \mathrm{U}(R_{2,K}) & 1 & X_2 & Y_2 \\ \mathrm{U}(R_{3,1}) & \mathrm{U}(R_{3,2}) & \mathrm{U}(0) & \cdots & \mathrm{U}(R_{3,K}) & 1 & X_3 & Y_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{U}(R_{K,1}) & \mathrm{U}(R_{K,2}) & \mathrm{U}(R_{K,3}) & \cdots & \mathrm{U}(R_{K,K}) & 1 & X_K & Y_K \\ 1 & 1 & 1 & \cdots & 1 & 0 & 0 & 0 \\ X_1 & X_2 & X_3 & \cdots & X_K & 0 & 0 & 0 \\ Y_1 & Y_2 & Y_3 & \cdots & Y_K & 0 & 0 & 0 \end{bmatrix} \tag{6.47}$$

in which $\mathrm{U}(R)$ is the function appearing in Equations 6.44 and 6.45 evaluated at each landmark location (X_i, Y_i). **W** is the $(K+3) \times 2$ matrix of weights and uniform terms appearing in Equations 6.44 and 6.45:

$$\mathbf{W} = \begin{bmatrix} W_{X1} & W_{Y1} \\ W_{X2} & W_{Y2} \\ \vdots & \vdots \\ W_{XK} & W_{YK} \\ A_{X1} & A_{Y1} \\ A_{XX} & A_{YX} \\ A_{XY} & A_{YY} \end{bmatrix} \tag{6.48}$$

So we have the equation:

$$\mathbf{V} = \mathbf{LW} \tag{6.49}$$

in which **L** and **W** are the matrices just described. We wish to solve for **W**, the matrix of coefficients in our spline model, which gives us:

$$\mathbf{W} = \mathbf{L}^{-1}\mathbf{V} \tag{6.50}$$

We can use the weights in the matrix **W** in conjunction with the spline functions in Equations 6.44 and 6.45 to interpolate the observed deformation at the landmarks over the entire specimen. However, it turns out that we can make some further use of the matrix $\mathbf{L}^{-1}$. This matrix is $(K+3)$ by $(K+3)$; if we take the first K rows and the first K columns of $\mathbf{L}^{-1}$, we can form $\mathbf{L}_{\mathbf{K}}^{-1}$, which is called the *bending energy matrix*.

The bending energy matrix can be rearranged into a series of eigenvectors $\mathbf{E_i}$, and eigenvalues, λ_i, such that:

$$\mathbf{L_K^{-1}E_i} = \lambda_i \mathbf{E_i} \tag{6.51}$$

The eigenvectors $\mathbf{E_i}$ have the usual properties of eigenvectors, and consequently they are a basis (or a set of coordinate axes) of a space. In this case, the eigenvectors are the basis of the Euclidean space tangent to shape space at the reference shape. This means that we can express our matrix of observed deformations **V** as a linear combination of the eigenvectors of the bending energy matrix. The eigenvalues are the bending energies required to effect a change (of a given amount of shape difference, i.e. a unit of Procrustes distance) at that spatial scale.

Three of the eigenvalues of the bending energy matrix are zero, corresponding to the components with no bending (with X- and Y-coefficients, these eigenvectors account for the six uniform components of the deformation). The remaining $K-3$ eigenvectors are the explicitly localized components of a deformation. These eigenvectors are called the *partial warps*; the vector multipliers of the partial warps are called the *partial warp scores* (following Slice et al., 1996). They are "partial" because they describe part of a deformation. We should note that Bookstein (1991) called the eigenvectors of the bending energy matrix *principal* warps, analogous to principal components. By "partial warp," he meant the vector multiple of a principal warp. Slice and colleagues use the term *principal warp* to refer to a partial warp interpreted as a bent surface of the thin-plate spline, and because the latter terminology has become standard, we use it here.

As evident in the definition of $\mathbf{L_K^{-1}}$, only one matrix of landmarks enters into the calculation of bending energy; the coordinates of the form usually called the reference or starting form. Thus, the eigenvectors that give us a coordinate system for shape analyses are a function of one single form. This may be highly counterintuitive, because more familiar eigenvectors, such as principal components, are functions of an observed variance–covariance matrix. They are functions of variation (or differences) among observed forms. That is not the case for the eigenvectors of the bending energy matrix. The eigenvalues of the bending energy are the bending energies that would be required to modify a given shape by a single unit of shape difference at each spatial scale. Thus the partial warps are not themselves features of shape change, they are simply a coordinate system or basis for the space in which we analyze shape change.

The "A" coefficients in Equation 6.48 describe the uniform deformation of the shape. There are six of these coefficients, which is enough to describe the six components of the uniform deformation of shape. However, we know that the reference and the target do not differ by rotation, rescaling or translation, because those differences were removed by the superimposition process. Consequently, we do not need six parameters to describe the uniform component of the deformation, only the two components derived earlier in this chapter.

By convention, partial (or principal) warps are numbered from the lowest to highest bending energy; the one with the highest number corresponds to the one with greatest bending energy. The two uniform components are sometimes called the zero$^{\text{th}}$ principal warp. Thinking of the uniform components in those terms is useful because it emphasizes that the uniform components cannot be viewed separately from the non-uniform ones. Including the uniform terms also completes the tally of shape variables. The $K - 3$ partial warps contribute $2K - 6$ scores; adding the two uniform scores brings the count up to $2K - 4$.

Using the thin-plate spline to visualize shape change

The combination of the uniform and non-uniform components completely describes any shape change. The set of partial warp scores (including scores on the uniform component) can be used in any conventional statistical analysis and, like the coordinates obtained by GLS, the sum of their squares equals the squared Procrustes distance from the reference. Moreover, like Bookstein's shape coordinates, they have the correct degrees of freedom. Thus we can use partial warps in any statistical procedure, such as regression, and diagram the results as a deformation.

Interpreting changes depicted by the thin-plate spline

Interpretations should be presented in terms of the total deformation, not by detailing the separate uniform and non-uniform components (or the more finely subdivided components of them). Just as we cannot talk about individual landmarks as if they were separately moved, we cannot talk about components of the total deformation as if they were separate parts of the whole. It is important to remember that the changes depicted are based on an interpolation function – we do not actually know what occurs between landmarks. If we have sparsely sampled some regions of the body, we cannot assume that the spline provides a realistic picture of their changes; there might be many highly localized changes that cannot be detected in the absence of closely spaced landmarks. All we can say is that our data do not require any more localized changes.

We cannot show an example of a biological transformation depicted by the thin-plate spline until we have results to show, so we will borrow examples from a later chapter (Chapter 10) to discuss the description of shape change using the thin-plate spline. In Figure 6.7 we depict the ontogenetic changes in body shape of two species of piranhas: *S. gouldingi* (Figure 6.7A), which we used earlier in this chapter, and *Pygopristis denticulata* (Figure 6.7B). In both species the head (as a whole) grows less rapidly than the middle of the body, and the eye grows far more slowly than the head. In neither species does the shortening of the eye result solely from the generally lower cranial growth rates; rather, there is an abrupt (and localized) deceleration of growth rates in the orbital region. However, that does not, by itself, fully account for the apparent contraction of the grid in the head, especially in *S. gouldingi*. Part of the relative shortening of the head, supraorbitally, results from the displacement of the landmark at the epiphyseal bar (landmark 2) towards the anterior landmark of the eye (landmark 14). Suborbitally, the apparent shortening of the head results from the displacement of the posterior jaw landmark (landmark 13) towards the posterior eye landmark (landmark 15), as well as from the more general shortening of the snout and eye. These two species also differ in the ontogeny of posterior body shape. In *S. gouldingi*, the caudal peduncle (the region bounded by landmarks 6, 7, and 8) appears to contract, but no change appears to be localized there – the posterior body generally shortens (as does the head). Growth rates appear to decrease, moving posteriorly from the midbody to the tail. Because the caudal peduncle is the most posterior part of the body, the growth rates are lowest there. In *P. denticulata*, growth rates decrease more

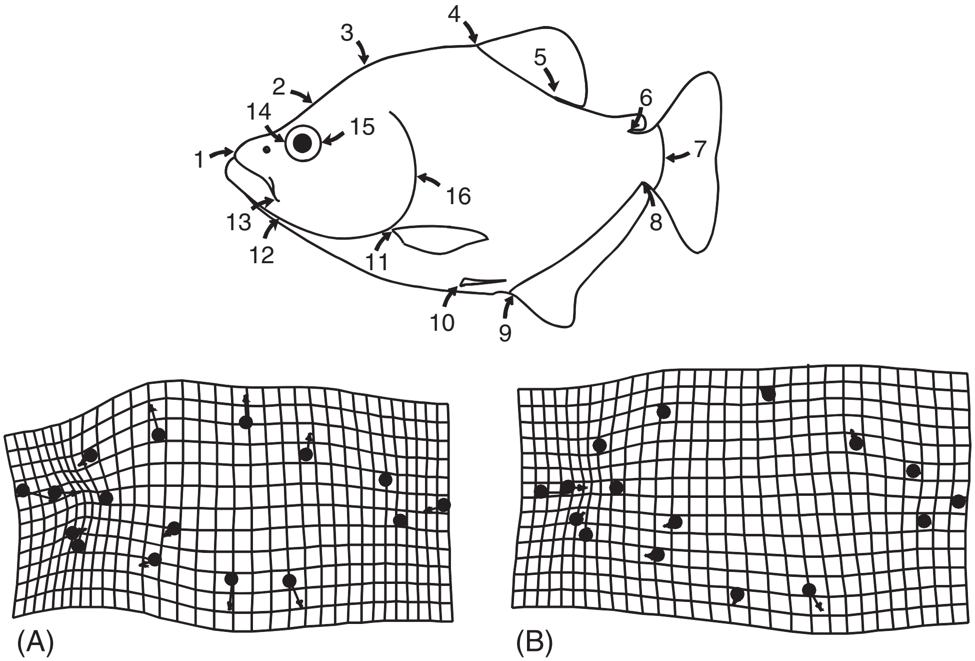


Figure 6.7 Ontogenetic shape change for two species of piranhas: (A) *Serrasalmus gouldingi*; (B) *Pygopristis denticulata*.

slightly, and most of the change in the posterior body seems to result from the posterior displacement (and relative shortening) of the anal fin. That increases the distance between the pelvic and anal fins (which expands the grid between them), but because that is not a part of the general expansion of the midbody (it is limited to the ventral region between the fins) the change is ventrally localized. Due to the sparse sampling of landmarks in the middle of the body, there is no abrupt contraction or expansion of the grid such as we see in the head. Sparse sampling of that region makes it difficult to detect localized changes because we cannot show what happens between landmarks when we have not sampled them (quoting Gertrude Stein, "there is no there there").

Software

Until we have results to depict, the spline serves the purpose of providing variables, with the correct degrees of freedom, for statistical analysis. A file of partial warps, along with the uniform components, can be computed by several programs in the IMP software series, all of which output the data in a form that can be input into statistical packages (i.e. they are in the X1,Y1,...CS format). They are perhaps most easily obtained from the Principal Components Analysis program (**PCAGen**, discussed in Chapter 7), which calculates partial warps relative to the mean shape (that is, the mean serves as the reference form). Within that software, as in all the others, the explicit uniform terms are always calculated using the partial Procrustes superimposition (meaning that centroid size is fixed to one). To draw the deformations in different registrations, the software simply calculates the implicit uniform deformations corresponding to the desired method of depicting the shape change. We will return to this when we have results to depict.

References

Bookstein, F. L. (1989). Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 567–585.

Bookstein, F. L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.

Bookstein, F. L. (1996). Standard formula for the uniform shape component in landmark data. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 153–168. Plenum Press.

Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*. John Wiley & Sons.

Myers, P., Lundrigan, B. L., Gillespie, B. W. and Zelditch, M. L. (1996). Phenotypic plasticity in skull and dental morphology in the prairie deer mouse (*Peromyscus maniculatus bairdii*). *Journal of Morphology*, **229**, 229–237.

Slice, D. E., Bookstein, F. L., Marcus, L. F. and Rohlf, F. J. (1996). Appendix I: A glossary for geometric morphometrics. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 531–551. Plenum Press.

Thompson, D'Arcy W. (1942). *On Growth and Form: A New Edition*. Cambridge University Press. (Reprinted in 1992 as *On Growth and Form: The Complete Revised Edition*, Dover Publications.)

PART II

Analyzing Shape Variables

7

Ordination methods

In this chapter, we discuss two methods for describing the diversity of shapes in a sample: principal components analysis (PCA) and canonical variates analysis (CVA). Our discussion of these methods draws heavily on expositions presented by Morrison (1967), Chatfield and Collins (1980), and Campbell and Atchley (1981). Both methods are used to simplify descriptions rather than to test hypotheses. PCA is a tool for simplifying descriptions of variation among individuals, whereas CVA is used for simplifying descriptions of differences between groups. Both analyses produce new sets of variables that are linear combinations of the original variables. They also produce scores for individuals on those variables, and these can be plotted and used to inspect patterns visually. Because the scores order specimens along the new variables, the methods are called "ordination methods." It is hoped that the ordering provides insight into patterns in the data, perhaps revealing patterns that are convenient for addressing biological questions. The most important difference between PCA and CVA is that PCA constructs variables that can be used to examine variation among individuals within a sample, whereas CVA constructs variables to describe the relative positions of groups (or subsets of individuals) in the sample.

We discuss PCA and CVA in the same chapter because both serve a similar purpose, and because the mathematical transformations performed in the two analyses are similar. We describe PCA first because it is somewhat simpler, and because it provides a foundation for understanding the transformations performed in CVA. We begin the description of PCA with some simple graphical examples, and then present a more formal exposition of the mathematical mechanics of PCA. This is followed by a presentation of an analysis of a real biological data set. The description of CVA follows a similar outline; the only difference is that we begin with a discussion of groups and grouping variables. CVA requires that the individuals be grouped, because the method analyzes the relative positions of groups in the sample. Consequently, the sample must be divided into groups before the analysis begins. The analysis of groups requires a few more computational steps than PCA, but none of the steps in CVA introduce new mathematical concepts. CVA will be just a new application of ideas you have already encountered in the discussion of PCA.

ISBN 0–12–77846–08

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

Principal components analysis

Geometric shape variables are neither biologically nor statistically independent. For example, the shape variables produced by the thin-plate spline describe variation in overlapping regions of an organism or structure. Because the regions overlap, they are under the influence of the same processes that produce variation; and therefore we expect them to be correlated. Even when they do not describe overlapping regions, morphometric variables (both geometric and traditional) are expected to be correlated because they describe features of the organism that are functionally, developmentally or genetically linked. Their patterns of variation and covariation are often complex and difficult to interpret. The purpose of PCA is to simplify those patterns and make them easier to interpret by replacing the original variables with new ones (principal components, PCs) that are linear combinations of the original variables and independent of each other.

One might wonder why it would be a worthwhile exercise to take simple variables that covary with each other and replace them with complex variables that do not covary. Part of the value of this exercise arises from the fact that the new complex variable is a function of the covariances among the original variables. It thus provides some insight into the covariances among variables that can direct future research into the identity of the causal factors underlying those covariances. Another useful purpose served by PCA is that most of the variation in the sample usually can be described with only a few PCs. Again this is useful, because it simplifies and clarifies what needs to be explained. Another important benefit of PCA is that the presentation of results is simplified. It is much easier to produce and explain plots of the three PCs that explain 90% of the variation than it is to separately plot and explain the variation on each of 30 original variables.

An indirect benefit of PCA that is useful (but often misused) is that it simplifies the description of differences among individuals. Clusters of individuals are often more apparent in plots of PCs than in plots of the original variables. Finding such clusters can be quite valuable, but those clusters do not represent evidence of statistically distinct entities. Legitimate methods for testing the hypothesis that *a priori* groups are statistically significantly different will be presented in Chapters 8 and 9 (computer-based statistical methods and multivariate analysis of variance, respectively).

Geometric description of PCA

Figure 7.1A shows the simple case in which there are two observed traits, $\mathbf{X}_1$ and $\mathbf{X}_2$. These traits might be two distance measurements or the coordinates of a single landmark in a two-dimensional shape analysis. Each point in the scatter plot represents the paired values observed for a single specimen. We expect that the values of each trait are normally distributed, and we expect that one trait is more variable than the other because one variable, (in this case, $\mathbf{X}_1$) has a larger range of observed values and a higher variance. In addition, the values of $\mathbf{X}_1$ and $\mathbf{X}_2$ are not independent; higher values of one are associated with higher values of the other. This distribution of values can be summarized by an ellipse that is tilted in the $\mathbf{X}_1$, $\mathbf{X}_2$ coordinate plane (Figure 7.1B). PCA solves for the axes of this ellipse, and uses those axes to describe the positions of individuals within that ellipse.

The first step of PCA is to find the direction through the scatter that describes the largest proportion of the total variance. This direction, the long axis of the ellipse, is the

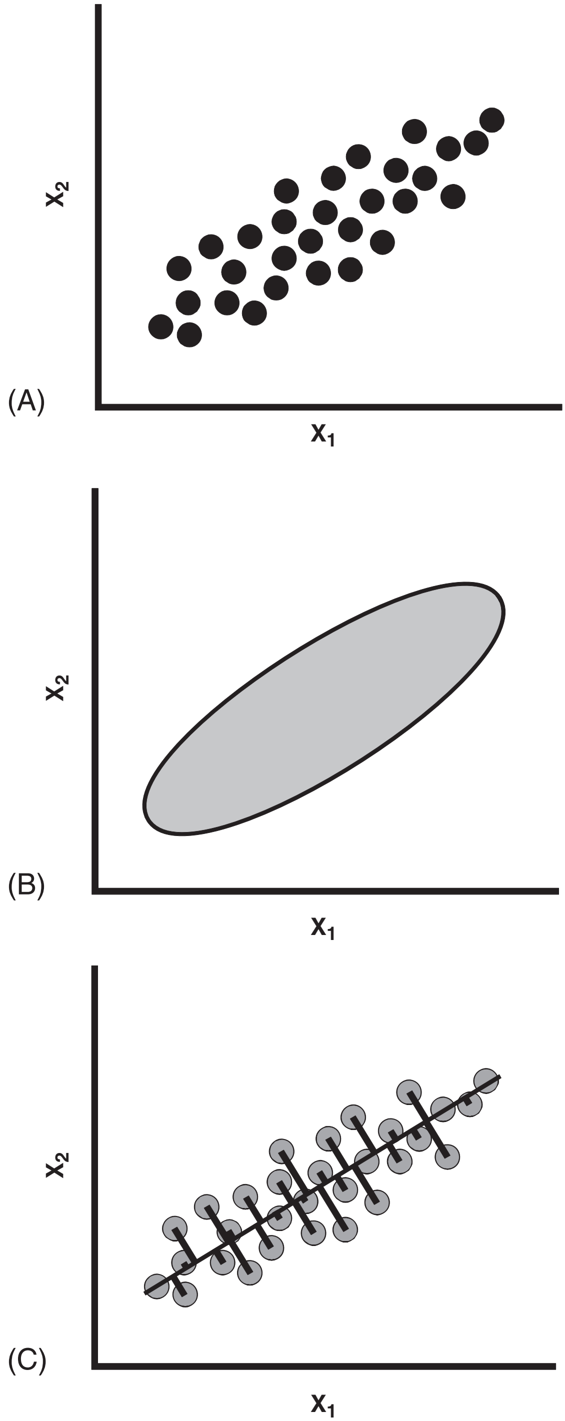


Figure 7.1 Graphical representation of the problem to be solved by PCA. (A) Scatter plot of individuals scored on two traits, $\mathbf{X_1}$ and $\mathbf{X_2}$; (B) an ellipse enclosing the scatter of points shown in part (A); (C) a line through the scatter and the perpendicular distances of the individuals from that axis. The goal of PCA is to find the line that minimizes the sum of those squared distances.

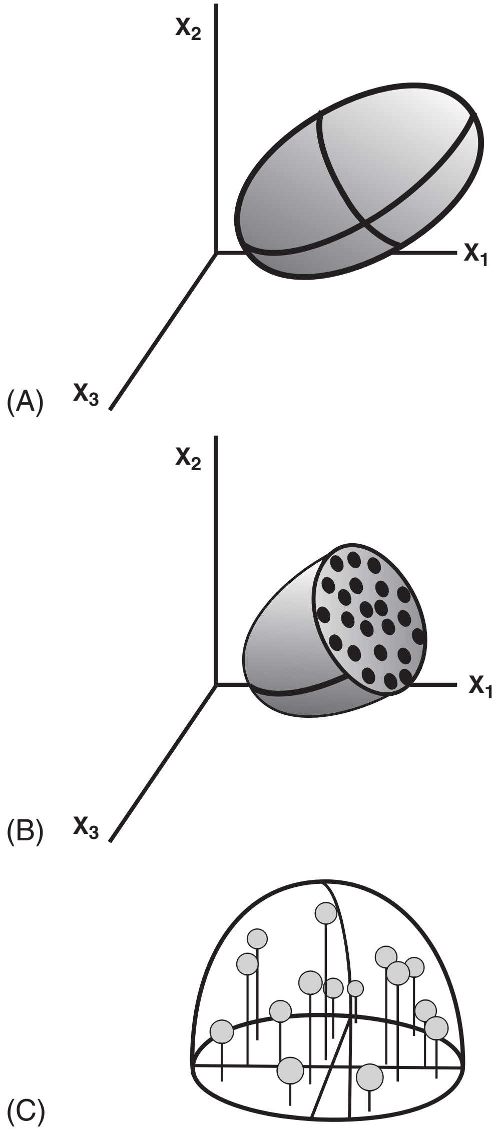


Figure 7.2 Graphical representation of PCA on three original variables ($\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$). (A) The distribution of individual specimens on the three original axes is summarized by a three-dimensional ellipsoid; (B) the three-dimensional ellipsoid is cut by a plane passing through the sample centroid and perpendicular to the longest axis (PC1) at its midpoint, showing the distribution of individuals around the longest axis in the plane of the section; (C) the upper half of the ellipsoid in B has been rotated so that the cross-section is in the horizontal plane. Perpendicular projections of all individuals (from both halves) onto this plane are used to solve for the second and third PCs.

first principal component (PC1). In an idealized case like that shown in Figure 7.1A, the line we seek is approximately the line through the two cases that have extreme values on both variables. Real data rarely have such convenient distributions, so we need a criterion that has more general utility. If we want to maximize the variance that the first axis describes, then we also want to minimize the variance that it does not describe – in other words, we want to minimize the sum of the squared distances of points away from the line (Figure 7.1C). (*Note*: the distances that are minimized by PCA are not the distances minimized in conventional least-squares regression analysis – see Chapter 10.)

The next step is to describe the variation that is not described by PC1. When there are only two original variables this is a trivial step; all of the variation that is not described by the first axis of the ellipse is completely described by the second axis. So, let us consider briefly the case in which there are three observed traits: $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$. This situation is unlikely to arise in optimally superimposed landmark data, but it illustrates a generalization that can be applied to more realistic situations. As in the previous example, all traits are normally distributed and no trait is independent of the others. In addition, $\mathbf{X}_1$ has the largest variance and $\mathbf{X}_3$ has the smallest variance. A three-dimensional model of this distribution would look like a partially flattened blimp or watermelon (Figure 7.2A). Again PC1 is the direction in which the sample has the largest variance (the long axis of the watermelon), but now a single line perpendicular to PC1 is not sufficient to describe the remaining variance. If we cut the watermelon in half perpendicular to PC1, the cross-section is another ellipse (Figure 7.2B). The individuals in the section (the seeds in the watermelon) lie in various directions around the central point, which is where PC1 passes through the section. Thus, the next step of the PCA is to describe the distribution of data points around PC1, not just for the central cross-section, but also for the entire length of the watermelon.

To describe the variation that is not represented by PC1, we need to map, or project, all of the points onto the central cross-section (Figure 7.2C). Imagine standing the halved watermelon on the cut end and instantly vaporizing the pulp so that all of the seeds drop vertically onto a sheet of wax paper, then repeating the process with the other half of the watermelon and the other side of the paper. The result of this mapping is a two-dimensional elliptical distribution similar to the first example. This ellipse represents the variance that is not described by PC1. Thus, the next step of the three-dimensional PCA is the first step of the two-dimensional PCA – namely, solving for the long axis of a two-dimensional ellipse, as outlined above. In the three-dimensional case, the long axis of the two-dimensional ellipse will be PC2. The short axis of this ellipse will be PC3, and will complete the description of the distribution of seeds in the watermelon. By logical extension, we can consider *N* variables measured on some set of individuals to represent an *N*-dimensional ellipsoid. The PCs of this data set will be the *N* axes of the ellipsoid.

After the variation in the original variables has been redescribed in terms of the PCs, we want to know the positions of the individual specimens relative to these new axes (Figure 7.3). As shown in Figure 7.3A, the values we want are determined by the orthogonal projections of the specimen onto the PCs. These new distances are called principal component *scores*. Because the PCs intersect at the sample mean, the values of the scores represent the distances of the specimen from the mean in the directions of the PCs. In effect, we are rotating and translating the ellipse into a more convenient orientation so we can use the PCs as the basis for a new coordinate system (Figure 7.3B). The PCs are the axes of that system. All this does is allow us to view the data from a different perspective; the positions of the data points relative to each other have not changed.

As suggested by Figure 7.4, we could compute an individual's score on a PC from the values of the original variables that were observed for that individual and the cosines of

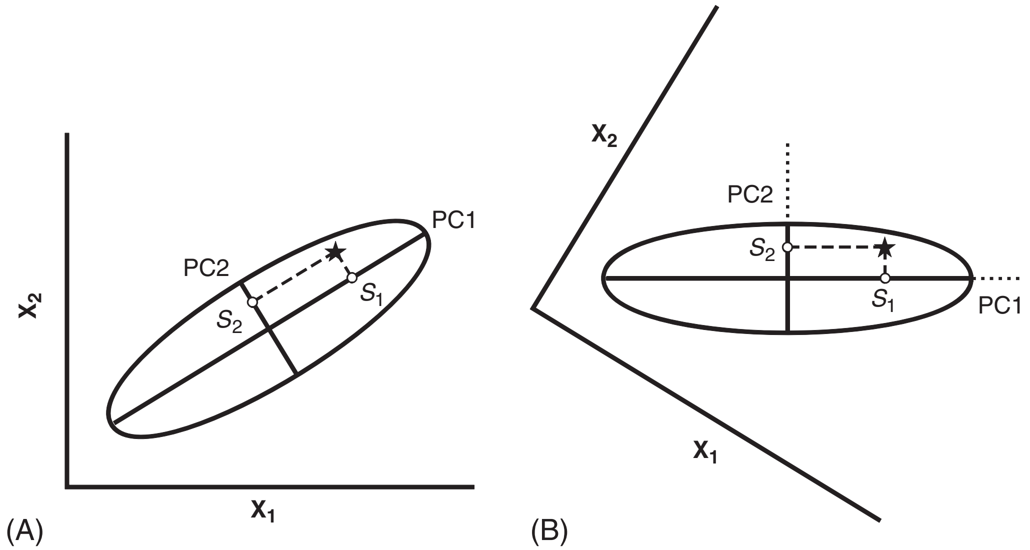


Figure 7.3 Graphical interpretation of PC scores. (A) The star is the location of an individual in the sample. Perpendiculars from the star to PCs indicate the location of the star with respect to those axes. The distances of points S_1 and S_2 from the sample centroid (intersection of PC1 and PC2) are the scores of the star on PC1 and PC2. (B) The figure in part A has been rotated so that PCs are aligned with the edges of the page. The PCs will now be used as the reference axes of a new coordinate system; the scores on these axes are the location of the individual in the new system. The relationships of the PC axes to the original axes has not changed, nor has the position of the star relative to either set of axes.

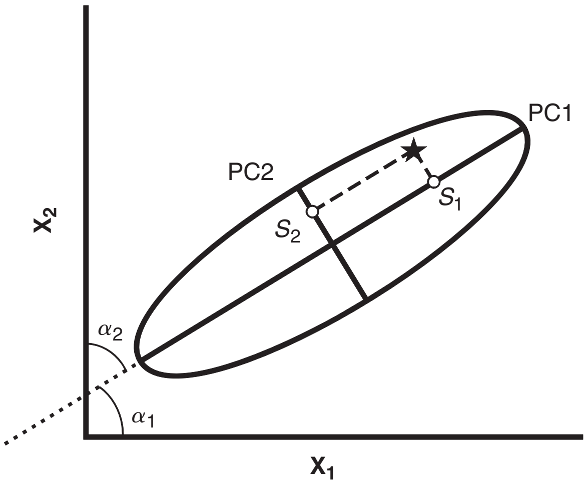


Figure 7.4 Graphical interpretation of PC scores, continued. The angles α_1 and α_2 indicate the relationship of PC1 to the original axes $\mathbf{X_1}$ and $\mathbf{X_2}$. Thus, S_1 can be computed from the coordinates of the star on $\mathbf{X_1}$ and $\mathbf{X_2}$ and the cosines of the angles between PC1 and the original axes. S_2 can be computed from the coordinates of the star on $\mathbf{X_1}$ and $\mathbf{X_2}$ and the cosines of the angles between PC2 and the original axes.

the angles between the original variables and the PCs. In our simple two-dimensional case, the new scores, Y, could be calculated as:

$$Y_1 = A_1X_1 + A_2X_2 \tag{7.1}$$

where A_1 and A_2 are the cosines of the angles α_1 and α_2 and the values of individuals on X_1 and X_2 are the differences between them and the mean, not the observed values of those variables.

It is important to bear in mind for our algebraic discussion that Equation 7.1 represents a straight line in a two-dimensional space. Later we will see equations that are expansions of this general form and represent straight lines in spaces of higher dimensionality. So, in case the form of Equation 7.1 is unfamiliar, the next few equations illustrate the simple conversion of this equation into a more familiar form. First, we rearrange the terms to solve for X_2:

$$Y_1 - A_1X_1 = A_2X_2 \tag{7.2}$$

$$A_2X_2 = -A_1X_1 + Y_1 \tag{7.3}$$

$$X_2 = \frac{-A_1X_1}{A_2} + \frac{Y_1}{A_2} \tag{7.4}$$

Then we make two substitutions ($M = -A_1/A_2$ and $B = Y_1/A_2$) to produce:

$$X_2 = MX_1 + B \tag{7.5}$$

Thus the formula for the PC is, indeed, the formula for a straight line.

Algebraic description of PCA

We begin this description of PCA by repeating the starting conditions and the constraints we want to impose on the new variable. We have a set of observations of P traits on N individuals, where P is the number of shape variables (not the number of landmarks). The data comprise P variances and $P(P-1)/2$ covariances in the sample. We want to compute a new set of P variables (PCs) with variances that sum to the same total as that computed from the variances and covariances of the original variables, and we also want the covariances of all the PCs to be zero. In addition, we want PC1 to describe the largest possible portion of variance, and we want each subsequent PC to describe the largest possible portion of the variation that was not described by the preceding components.

The full set of observations can be written as the matrix **X**:

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & X_{13} & \ldots & X_{1P} \\ X_{21} & X_{22} & X_{23} & \ldots & X_{2P} \\ X_{31} & X_{32} & X_{33} & \ldots & X_{3P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & X_{N3} & \ldots & X_{NP} \end{bmatrix} \tag{7.6}$$

where X_{NP} is the value of the Pth coordinate in the Nth individual. We can also think of this as a P-dimensional space with N points plotted in that space – just a multi-dimensional version of the simplistic examples presented in the previous section.

Our problem is to replace the original variables ($\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3, \ldots \mathbf{X}_\mathbf{P}$), which are the columns of the data matrix, with a new set of variables ($\mathbf{Y}_1$, $\mathbf{Y}_2$, $\mathbf{Y}_3, \ldots \mathbf{Y}_\mathbf{P}$), the PCs that meet the constraints outlined in the first paragraph of this section. Each PC will be a straight line through the original P-dimensional space, so we can write each $\mathbf{Y}_\mathrm{j}$ as a linear combination of the original variables:

$$\mathbf{Y}_\mathrm{j} = A_{1j}\mathbf{X}_1 + A_{2j}\mathbf{X}_2 + \cdots + A_{Pj}\mathbf{X}_\mathbf{P} \tag{7.7}$$

which can be expressed in matrix notation as:

$$\mathbf{Y}_\mathrm{j} = \mathbf{A}_\mathrm{j}^\mathrm{T}\mathbf{X} \tag{7.8}$$

where $\mathbf{A}_\mathrm{j}^\mathrm{T}$ is a vector of constants $\{A_{1j}, A_{2j}, A_{3j} \ldots A_{Pj}\}$. (The notation $\mathbf{A}_\mathrm{j}^\mathrm{T}$ refers to the *transpose*, or row form, of the column matrix $\mathbf{A}_\mathrm{j}$.) All this means is that the new values of the individuals, their PC scores, will be computed by multiplying their original values (listed in matrix $\mathbf{X}$) by the appropriate values $\mathbf{A}_\mathrm{j}^\mathrm{T}$ of and summing the appropriate combinations of multiples. Now we can see that our problem is to find the values of $\mathbf{A}_\mathrm{j}^\mathrm{T}$ that satisfy the constraints outlined above.

The first constraint we will address is the requirement that the total variance is not changed. Variance is the sum of the squared distances of individuals from the mean, so this is equivalent to requiring that distances in the new coordinate system are the same as distances in the original coordinate system. The total variance of a sample is given by the sample variance–covariance matrix $\mathbf{S}$:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & s_{13} & \cdots & s_{1P} \\ s_{21} & s_{22} & s_{23} & \cdots & s_{2P} \\ s_{31} & s_{32} & s_{33} & \cdots & s_{3P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{P1} & s_{P2} & s_{P3} & \cdots & s_{PP} \end{bmatrix} \tag{7.9}$$

in which s_{ii} is the sample variance observed in variable $\mathbf{X}_\mathrm{i}$, and s_{ij} (which is equal to s_{ji}) is the sample covariance observed in variables $\mathbf{X}_\mathrm{i}$ and $\mathbf{X}_\mathrm{j}$.

We can meet the requirement that the total variance is unchanged by requiring that each PC is a vector of length one. If we multiply matrix $\mathbf{X}$ by a vector of constants as indicated in Equation 7.8, the variance of the resulting vector $\mathbf{Y}_\mathrm{j}$ will be:

$$\mathrm{Var}(\mathbf{Y}_\mathrm{j}) = \mathrm{Var}\left(\mathbf{A}_\mathrm{j}^\mathrm{T}\mathbf{X}\right) = \mathbf{A}_\mathrm{j}^\mathrm{T}\mathbf{S}\mathbf{A}_\mathrm{j} \tag{7.10}$$

Thus the constraint that variance is unchanged can be formally stated as the requirement that the inner product or dot product of each vector $\mathbf{A}_\mathrm{j}^\mathrm{T}$ with itself must be one:

$$\mathbf{A}_\mathrm{j}^\mathrm{T}\mathbf{A}_\mathrm{j} = 1 = \sum_{k=1}^{p} A_{kj}^2 \tag{7.11}$$

This means that the sum of the squared coefficients will be equal to one for each PC. Substituting Equation 7.11 into Equation 7.10 yields $\mathrm{Var}(\mathbf{Y}_\mathrm{j}) = \mathbf{S}$, demonstrating that the constraint has been met.

The next constraint is the requirement that principal component axes have covariances of zero. This means that the axes must be *orthogonal*. More formally stated, this constraint is the requirement that the dot product of any two axes must be zero. For the first two PCs, the constraint is expressed as:

$$\mathbf{A}_{\mathrm{j}}^{\mathrm{T}}\mathbf{A}_2 = 0 = \sum A_{1i}A_{2i} \tag{7.12}$$

The general requirement that the products of corresponding coefficients must be zero for any pair of PCs is expressed as:

$$\mathbf{A}_{\mathrm{i}}^{\mathrm{T}}\mathbf{A}_{\mathrm{j}} = 0 \tag{7.13}$$

The requirements imposed by Equations 7.11 and 7.13 indicate that we are solving for an *orthonormal basis*. A basis is the smallest number of vectors necessary to describe a vector space (a matrix). An orthogonal basis is one in which each vector is orthogonal to every other, so that a change in the value of one does not necessarily imply a change in the value of another – in other words, all the variables are independent, or have zero covariance (Equation 7.13) in an orthogonal basis. An orthonormal basis is an orthogonal basis in which each axis has the same unit length. This very particular kind of normality was imposed by the first requirement (Equation 7.11). In an orthonormal basis, a distance or difference of one unit on one axis is equivalent to a difference of one unit on every other axis; consecutive steps of one unit on any two axes would describe two sides of a square.

So far, we have defined important relationships among the values of $\mathbf{A}_{\mathrm{i}}^{\mathrm{T}}$. There is an infinite number of possible orthonormal bases that we could construct to describe the original data. The third constraint imposed above defines the relationship of the new basis vectors to the original vector space of the data. Specifically, this constraint is the requirement that the variance of PC1 is maximized, and that the variance of each subsequent component is maximized within the first two constraints.

We begin with the variance of PC1. From Equation 7.10 we know that:

$$\mathrm{Var}(\mathbf{Y}_1) = \mathrm{Var}\big(\mathbf{A}_1^{\mathrm{T}}\mathbf{X}\big) = \mathbf{A}_1^{\mathrm{T}}\mathbf{S}\mathbf{A}_1 \tag{7.14}$$

The matrix $\mathbf{S}$ can be reduced to:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_p \end{bmatrix} \tag{7.15}$$

where each λ_i is an *eigenvalue*, a number that is a solution of the *characteristic equation*:

$$\mathbf{S} - \lambda_i\mathbf{I} = 0 \tag{7.16}$$

In the characteristic equation, **I** is the $P \times P$ identity matrix:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \tag{7.17}$$

If each original variable in the data matrix **X** has a unique variance (cannot be replaced by a linear combination of the other variables), then each λ_i has a unique value greater than zero. Furthermore, the sum of the eigenvalues is equal to the total variation in the original data.

For each eigenvalue, there is a corresponding vector $\mathbf{A_i}$, called an *eigenvector*, such that:

$$\mathbf{SA_i} = \lambda_i \mathbf{A_i} \tag{7.18}$$

This must be true, because we have already required:

$$\mathbf{S} - \lambda_i \mathbf{I} = 0 \tag{7.19}$$

Therefore:

$$(\mathbf{S} - \lambda_i \mathbf{I})\mathbf{A_i} = 0 \tag{7.20}$$

which can be rearranged to:

$$\mathbf{SA_i} = \lambda_i \mathbf{A_i} \tag{7.21}$$

Thus, the eigenvectors are a new set of variables with variances equal to their eigenvalues and covariances equal to zero. Because the covariances are zero, the eigenvectors satisfy the constraint of orthogonality. Eigenvectors usually do not meet the constraint of normality ($\mathbf{A_i^T A_i} = 1$), but this can be corrected simply by rescaling. Accordingly, the rescaled eigenvectors are the PCs, which comprise an orthonormal basis for the variance–covariance matrix **S**.

All that remains is to order the eigenvectors so that the eigenvalues are in sequence from largest to smallest. We can now show that the variance of PC1 is the first and largest eigenvalue. From Equation 7.10 we have $\mathrm{Var}(\mathbf{Y_j}) = \mathbf{A_j^T S A_j}$, and from Equation 7.18 we have $\mathbf{SA_i} = \lambda_i \mathbf{A_i}$. Putting these together, we get:

$$\mathrm{Var}(\mathbf{Y_1}) = \mathbf{A_1^T} \lambda_1 \mathbf{A_1} \tag{7.22}$$

We can rearrange this to:

$$\mathrm{Var}(\mathbf{Y_1}) = \lambda_1 \mathbf{A_1^T A_1} \tag{7.23}$$

which simplifies to λ_1 because we have already imposed the constraint that ($\mathbf{A_1^T A_1} = 1$).

A formal proof that principal components are eigenvectors of the variance–covariance matrix

This is the derivation as presented by Morrison (1990). Let us suppose that we have a set of measures or coordinates $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3} \dots \mathbf{X_P})$, and we want to find the vector $\mathbf{A_1} = (A_{11}, A_{21}, A_{31} \dots A_{P1})$ such that:

$$\mathbf{Y_1} = A_{11}\mathbf{X_1} + A_{21}\mathbf{X_2} + A_{31}\mathbf{X_3} + \cdots + A_{P1}\mathbf{X_P} \tag{7.24}$$

We would like to maximize the variance of $\mathbf{Y_1}$:

$$s_{Y_1}^2 = \sum_{i=1}^{P} \sum_{j=1}^{P} A_{i1} A_{j1} s_{ij} \tag{7.25}$$

where s_{ij} is the element on the ith row and jth column of the variance–covariance matrix $\mathbf{S}$ of the observed specimens. We can write the variance of $\mathbf{Y_1}$ in matrix form as:

$$s_{Y_1}^2 = \mathbf{A_1^T S A_1} \tag{7.26}$$

Now we seek to maximize $s_{Y_1}^2$ subject to the constraint that $\mathbf{A_1}$ has a magnitude of one, which means that $(\mathbf{A_1^T A_1} = 1)$. To do this, we introduce a term called a Lagrange multiplier λ_1, and use it to form the expression:

$$s_{Y_1}^2 + \lambda_1 \left(1 - \mathbf{A_1^T A_1}\right) \tag{7.27}$$

which we seek to maximize with respect to $\mathbf{A_1}$. Therefore, we take this new expression for the variance of $\mathbf{Y_1}$ and set its partial derivative with respect to $\mathbf{A_1}$ to zero:

$$\frac{\partial}{\partial A_1} \{ s_{Y_1}^2 + \lambda_1 (1 - \mathbf{A_1^T A_1}) \} = 0 \tag{7.28}$$

Using Equation 7.26, we can expand the expression for the partial derivative to:

$$\frac{\partial}{\partial A_1} \{ \mathbf{A_1^T S A_1} + \lambda_1 (1 - \mathbf{A_1^T A_1}) \} = 0 \tag{7.29}$$

which we now simplify to:

$$2(\mathbf{S} - \lambda_1 \mathbf{I})\mathbf{A_1} = 0 \tag{7.30}$$

where $\mathbf{I}$ is the $P \times P$ identity matrix. Because $\mathbf{A_1}$ cannot be zero, Equation 7.30 is a vector multiple of Equation 7.16, the characteristic equation. In Equation 7.30, λ_1 is the eigenvalue and $\mathbf{A_1}$ is the corresponding eigenvector.

Given Equation 7.30, we can also state that:

$$(\mathbf{S} - \lambda_1 \mathbf{I})\mathbf{A_1} = 0 \tag{7.31}$$

This can be rearranged as:

$$\mathbf{S A_1} - \lambda_1 \mathbf{I A_1} = 0 \tag{7.32}$$

and simplified to:

$$\mathbf{S A_1} - \lambda_1 \mathbf{A_1} = 0 \tag{7.33}$$

and further rearranged so that:

$$\mathbf{S A_1} = \lambda_1 \mathbf{A_1} \tag{7.34}$$

This leads to the following substitutions and rearrangements of Equation (7.26):

$$s_{Y_1}^2 = \mathbf{A_1^T S A_1} = \mathbf{A_1^T} \lambda_1 \mathbf{A_1} = \lambda_1 \mathbf{A_1^T A_1} = \lambda_1 \tag{7.35}$$

Thus, the eigenvalue λ_1 is the variance of $\mathbf{Y_1}$.

Interpretation of results

As we stated above, PCA is nothing more than a rotation of the original data; it is simply a descriptive tool. The utility of PCA lies in the fact that many (if not all) of the features measured in a study will exhibit covariances because they interact during, and are influenced by, common processes. Below, we use an analysis of jaw shape in a population of tree squirrels to demonstrate how PCA can be used to reveal relationships among traits.

Fifteen landmarks were digitized on the lower jaws of 31 squirrels (Figure 7.5). These landmarks capture information about the positions of the cheek teeth (2–5), the incisor

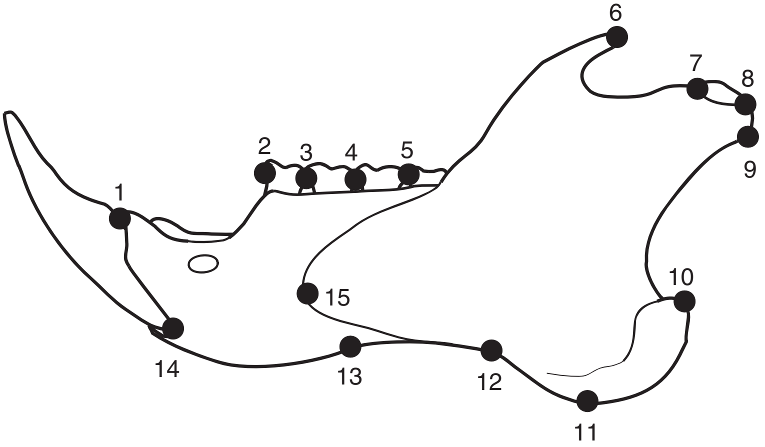


Figure 7.5 Outline drawing of the lower jaw of the fox squirrel, *Sciurus niger*, showing the locations of 15 landmarks.

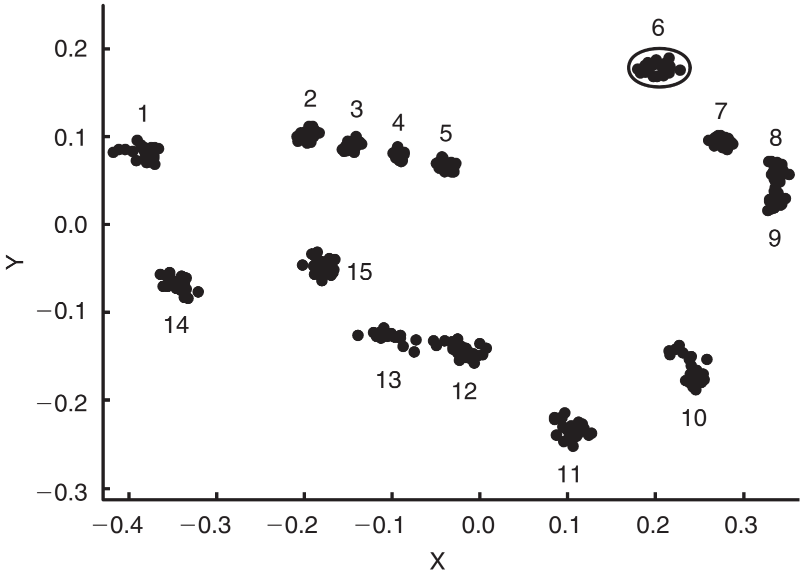


Figure 7.6 Plot of landmark coordinates of 31 *S. niger* jaws after partial Procrustes superimposition. The locations of landmark 6 in all 31 specimens are enclosed by an ellipse. Similar ellipses could be drawn for each landmark.

(1, 14 and 13), muscle attachment areas (6, 9–12, 15) and the articulation surface of the jaw joint (7 and 8). The 31 specimens include 23 adults and 8 juveniles (individuals lacking one or more of the adult teeth).

Figure 7.6 shows the landmark configurations of all 31 specimens, after partial Procrustes superimposition. This plot does not tell us much beyond the fact that there is shape variation in the sample. We can infer from the areas of the scatters for individual landmarks that there is not much variation in the relative positions of the cheek teeth. In contrast, many of the ventral landmarks have noticeably larger scatters, suggesting that their positions relative to the teeth are more variable.

To obtain more precise information about the pattern of shape variation, the 31 sets of landmark coordinates are converted into shape variables (see Chapter 6 for review), and these shape variables are subjected to PCA. The 15 landmarks yield 26 shape variables, so there are 26 PCs, and 26 scores for each specimen (its score on each component). The output from PCA consists of the list of coefficients describing the PCs, the variance of each component and its percentage of the total variance, and the scores of each specimen on each component.

As shown in Table 7.1, each PC has progressively less variance. Many of the components represent such a small proportion of the total variance that it is reasonable to ask whether

Table 7.1 Eigenvalues from PCA of squirrel jaws

PC	*Eigenvalues*	*% of total variance*
1	1.13×10^{-3}	51.56
2	2.15×10^{-4}	9.83
3	1.64×10^{-4}	7.49
4	1.36×10^{-4}	6.22
5	1.16×10^{-4}	5.32
6	9.52×10^{-5}	4.36
7	7.18×10^{-5}	3.28
8	5.45×10^{-5}	2.49
9	4.49×10^{-5}	2.05
10	3.58×10^{-5}	1.64
11	3.25×10^{-5}	1.49
12	2.36×10^{-5}	1.08
13	1.79×10^{-5}	0.82
14	1.37×10^{-5}	0.63
15	9.83×10^{-6}	0.45
16	9.31×10^{-6}	0.43
17	6.87×10^{-6}	0.31
18	3.72×10^{-6}	0.17
19	3.06×10^{-6}	0.14
20	2.17×10^{-6}	0.10
21	1.66×10^{-6}	0.08
22	7.04×10^{-7}	0.03
23	5.37×10^{-7}	0.02
24	3.62×10^{-7}	0.02
25	1.15×10^{-7}	0.01
26	5.02×10^{-8}	<0.01

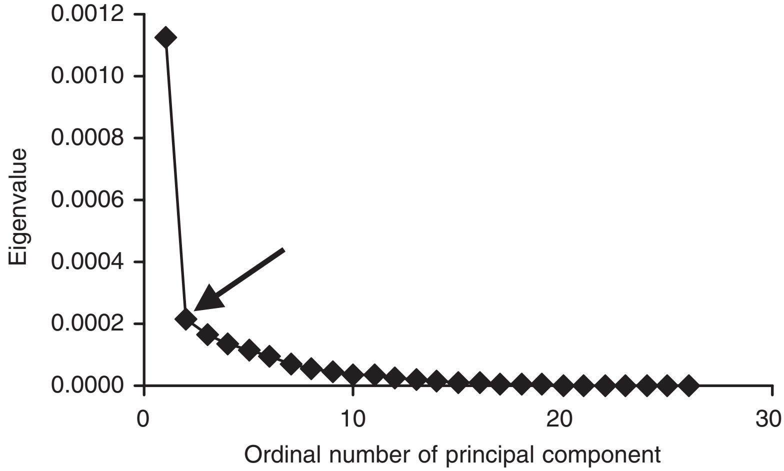


Figure 7.7 Scree plot of the proportion of variance described by each PC for the squirrel jaw data set. Arrow indicates the inflection point.

they describe anything biologically meaningful. One common rule of thumb is to interpret only those components that represent more than 5% of the variance. In the squirrel jaw example, PCs 1 through 5 meet this criterion. They account for a total of 80.4% of the variance in the sample, leaving 19.6% undescribed. This may seem like a large proportion of the variance to omit from further analysis, but it is doubtful that any one of the remaining 21 components describes a meaningful amount of variance.

The similarity in magnitudes of variances described by most components can be seen in a *scree* plot, in which the variance, or percentage of the total variance, is plotted against the ordinal number of the PCs (Figure 7.7). In this example, there is a large difference between the variances of the first two PCs, and much smaller differences between successive pairs of components. This difference is reflected in the scatter plot of scores in the two axes (Figure 7.8); the range of scores is much larger on PC1 than PC2, indicating that PC1 accounts for a much larger portion of the total variance. If two components have similar variances (e.g. if the distribution of scores in Figure 7.8 were closer to circular), then we have grounds to question whether either of them can be attributed to a distinct causal factor. Thus, an alternative rule of thumb is to find the inflection point on the scree plot and interpret only those components to the left of the inflection point (where the variance of each component is distinct from the variance of the following component). The main difficulty with applying this rule is that scree plots often do not have inflection points that are as obvious as the one in Figure 7.7.

Fortunately, there is a more rigorous approach to testing whether two successive PCs have distinct variances. This is an application of a test developed by Anderson (1958) and discussed in Morrison (1990). The null hypothesis is that some set of R consecutive eigenvalues are equal to each other. In other words, the variation described by these components cannot be distinguished from random variation. The eigenvalues are numbered from $Q+1$ to $Q+R$, where Q is a function of P (the total number of eigenvalues) and R (the number of the particular components of interest) such that $Q=P-R$. Anderson (1958) derived a χ^2 statistic based on the likelihood-ratio criterion to test the hypothesis

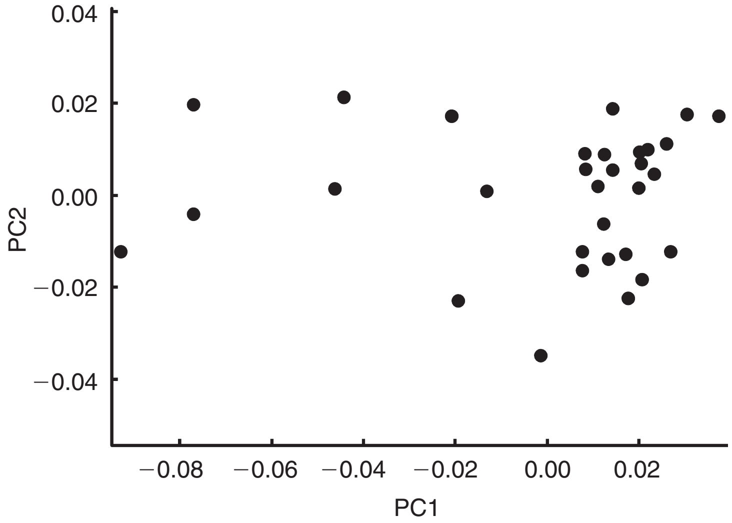


Figure 7.8 Scatter plot showing scores on the first two PCs for the sample of 31 squirrel jaws shown in Figure 7.6.

that the $Q+1$ eigenvalue is not distinct from the higher numbered eigenvalues:

$$\chi^2 = -N \sum_{j=Q+1}^{N} \ln \lambda_j + NR \ln\left(\frac{\sum_{j=Q+1}^{N} \lambda_j}{R}\right) \tag{7.36}$$

where N is the sample size minus one. When N is large, the degrees for freedom are $(\frac{1}{2} R(R+1) - 1)$ ($d.f. = 2$ when $R = 2$). In the special case where $Q + R = P$, the test evaluates whether variation in the last R eigenvectors is spherical. To test two successive eigenvalues, R is set to 2. For the squirrel jaw example, comparison of the first two eigenvalues yields $\chi^2 = 19.12$, which has a p-value less than 0.0001. Comparison of the second and third eigenvalues yields $\chi^2 = 0.55$, which has a p-value of 0.76. Thus, PC1 is the only one with a distinct eigenvalue, and the only one that can be regarded as biologically meaningful.

If you use several software packages to run PCAs, you may occasionally find the results differ in signs for the PCs (when that happens, the scores for individuals on those axes also differ by a sign). Reversed axes and scores can be disconcerting, but there is no need to worry – the sign of a PC is arbitrary. If $\mathbf{A}_1$ is an eigenvector corresponding to λ_1, then so is $-\mathbf{A}_1$. If we change the sign on $\mathbf{A}_1$, then the score of the jth specimen on the first axis will also change sign; $\mathbf{Y}_\mathrm{j} = \mathbf{A}_1^\mathrm{T}\mathbf{X}_\mathrm{j}$ so the product $\mathbf{Y}_1\mathbf{A}_1$ does not change sign. In other words, the eigenvectors $\mathbf{A}_1$ and $-\mathbf{A}_1$ are simply mirror images. The choice of sign has no effect on the interpretations of this component, and no effect on the computation of the subsequent component (a vector orthogonal to $\mathbf{A}_1$ will also be orthogonal to $-\mathbf{A}_1$).

To this point we have not discussed how to interpret the pattern of variation represented by a PC. That rests on the coefficients of the PC, which express the relationship between the PC and the original variables. Because our original variables are shape variables, we can generate a picture of shape variation along any PC by multiplying the original shape variables by the coefficients of the PC and summing them. Figure 7.9 shows the result of that computation for PC1 of shape variation in the sample of squirrel jaws.

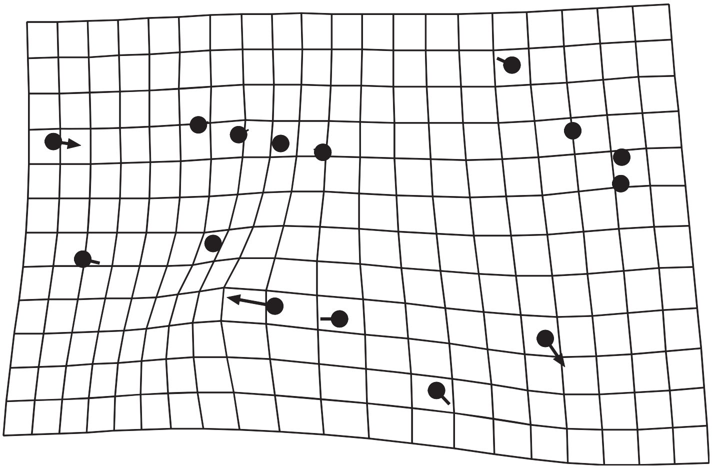


Figure 7.9 Pattern of shape change along PC1 for the 31 squirrel jaws shown in Figure 7.6. Circles indicate the locations of the landmarks in the mean shape of the sample; arrows indicate the changes in the relative positions of the landmarks as the score on PC1 increases. The deformed grid illustrates the thin-plate spline interpolation over the entire form.

We should note that many of the studies applying PCA to geometric data call the method "relative warps analysis" (RWA). PCA and RWA are not exactly equivalent, because the components of variance extracted by RWA are sometimes weighted by bending energy (originally, RWA was an analysis of components of variation relative to bending energy, hence the term "relative" in the name of the method). When variation is not weighted by bending energy, RWA is PCA. We prefer the more familiar term.

Canonical variates analysis

The purpose of CVA is to simplify the description of differences among *groups*. For example, CVA could be used to describe differences in mandible shape among queens, soldiers and workers in a colony of ants. It could also be used to describe differences in soldier morphology among colonies, species, or more inclusive categories. If individuals in a study can be sorted into mutually exclusive sets, CVA can be used to describe the differences among those sets. However, again we caution that CVA cannot be used to test the statistical significance of the differences among sets; for that, multivariate analysis of variance is needed.

There are many similarities between CVA and PCA. Like PCA, CVA constructs a new coordinate system (the canonical variates, CVs) and determines the scores on those axes for all individuals in a study. Also, the CVs are linear combinations of the original variables and are constrained to be mutually orthogonal. However, whereas PCA is used to describe differences among individuals, CVA is used to describe differences among group means. In this sense, CVA is analogous to a PCA of the group means. Another difference between CVA and PCA is that CVA uses the patterns of within-group variation to scale the axes of the new coordinate system. Because of this rescaling, CVs are not simply rotations of the original coordinate system, and distances in CV space are not equal to distances in the original coordinate system. (This is where the analogy breaks down.) Although rescaling may complicate interpretations of CVs, it also adds to their utility. As a result of the rescaling, CV1 is the direction in which groups are most effectively discriminated, which is not necessarily the direction in which the group means are most different.

Groups and grouping variables

A group is a set of individuals that share a particular state of a discontinuous trait. Examples of groups include sexes, color morphs, species, and supraspecific categories like guilds. To be analyzed by CVA the groups must be mutually exclusive, meaning that they cannot comprise nested or intersecting sets. The groups differ by a categorical variable, which is sometimes called a "qualitative trait" or a "grouping variable." The important characteristic of these variables is that they are not measured nor arrayed in a sequence; they do not have intrinsic numerical values, and nor do they have an inherent order or sequence.

Sometimes, features that can be scored on a continuous graded scale are treated as categorical variables. For example, the proportions of meat and vegetation in an animal's diet can be quantified and scored along a continuum. Nevertheless, it is a common practice to sort diets into a small number of categories (e.g. carnivore, herbivore, omnivore). Other traits that might be treated in a similar fashion include geographic location and age. There are several reasons for treating these kinds of traits as categorical variables. One is a lack of sufficient information to justify or support a more finely graded analysis – for example, a researcher may not have precise data on the proportions of food items in the diets of all species or individuals in a study. Another reason for treating a quantifiable trait as a categorical variable is that the investigator may not want to impose a hypothesis of ordering on the data, which is often a consideration when groups are not dispersed along a single straight line. Similarly, the investigator may not want to assume that all steps are of equal value (e.g. ontogenies often can be divided into discrete instars or age classes based on sequences of developmental events, but the sequentially numbered steps may represent different amounts of time or ontogenetic change). Under these circumstances, a quantifiable trait can be treated as a categorical variable and CVA would then be used to describe differences among the groups delineated by distinct states. However, the user should be aware that taking this approach also limits the inferences that can be drawn from the result – for example, an observation that age classes can be differentiated does not necessarily imply the kind of monotonic progression from age to age that can be inferred from a regression.

Geometric description of CVA

To develop a geometric intuition for CVA, we return to the metaphor of a slightly flattened watermelon. In PCA, we described the positions of seeds within the watermelon by finding its greatest dimensions. In CVA, we are not interested in the positions of seeds in the watermelon; instead, we want to describe the positions of the watermelons in the field (centroids of the ellipses in Figure 7.10). If all we want to know is the location of each melon, we could simply plot each melon's centroids; however, suppose we want to find the direction in which it is easiest to walk across the field without stepping on any of the melons (perhaps we want to spread fertilizer in the field). To solve this problem, we want to find the direction in which the melons are farthest apart. This requires that we know

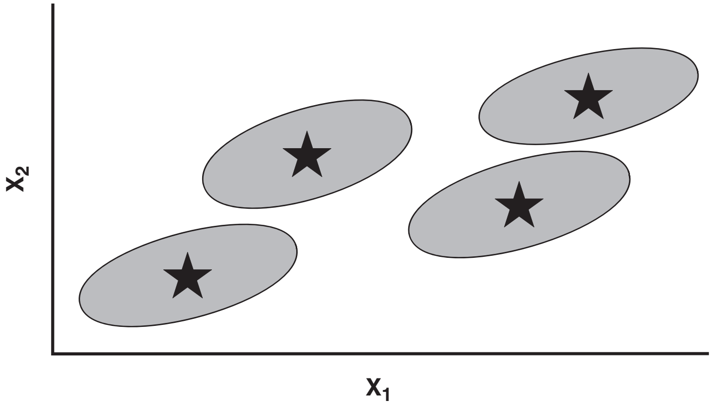


Figure 7.10 Ellipses of variation in two dimensions (X_1 and X_2) for four sample populations. Stars indicate locations of the means of each sample.

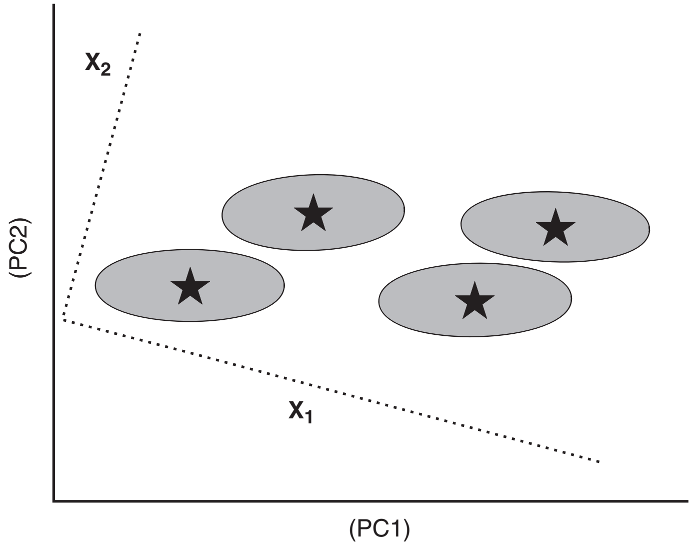


Figure 7.11 Graphical representation of the first step of CVA. The entire data set is rotated to a new coordinate system that is aligned with the PCs of the pooled variances. At this stage the relative positions of the four samples (and the individuals within groups) have not changed. The original coordinate system (Figure 7.10) is shown in the dotted lines. The axes of the new coordinate system are labeled in parentheses because we have not specified the location of the average sample, only the directions of its variances.

the average of the shapes and orientations of all the melons, not just the position of each melon's centroid.

Similarly, CVA begins with a PCA of the pooled (averaged) within-group variances. This gives us a new coordinate system in which we can describe the position of each group. In our field, we begin by defining a new coordinate system that would be aligned with the axes of the average melon (Figure 7.11).

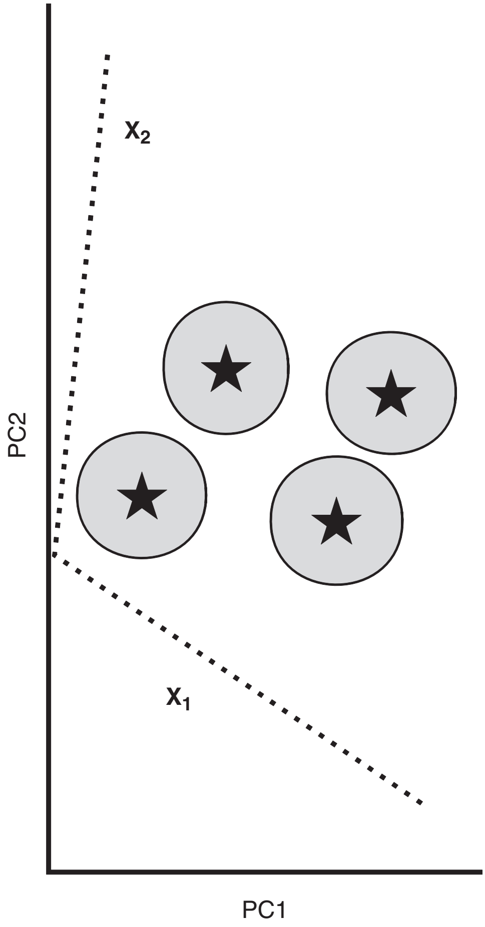


Figure 7.12 Graphical representation of the second step of CVA. The new coordinate system (solid lines) is rescaled in proportion to the pooled within-group variances in the original space. Variation within samples will be circular in the new space if the original variances were all identical. Note that the axes of the original coordinate system (dotted) are not orthogonal in the new space. Furthermore, distances in the new space are not equivalent to distances in the original space (Figure 7.10).

Now we can see that the melons overlap more in the direction defined by the long axis of the average melon. To take this into account, we rescale this axis proportionate to the elongation of the average melon. In effect, we distort our plot of the field until the average melon looks circular rather than elliptical (Figure 7.12).

Now we can solve for the direction in which melons tend to be farthest apart in the rescaled space by performing a PCA on the group centroids. The axes produced by this last computation are the CVs (Figure 7.13A). The scores of individuals on the CVs are the projections of the individuals onto these new coordinate axes (Figure 7.13B).

Because computation of the CVs involves a rescaling, interpretation of CV scores can be complex. If we undo the rescaling and rotation that were used to solve for the CVs (Figure 7.14), we see that each CV is a linear combination of the original variables.

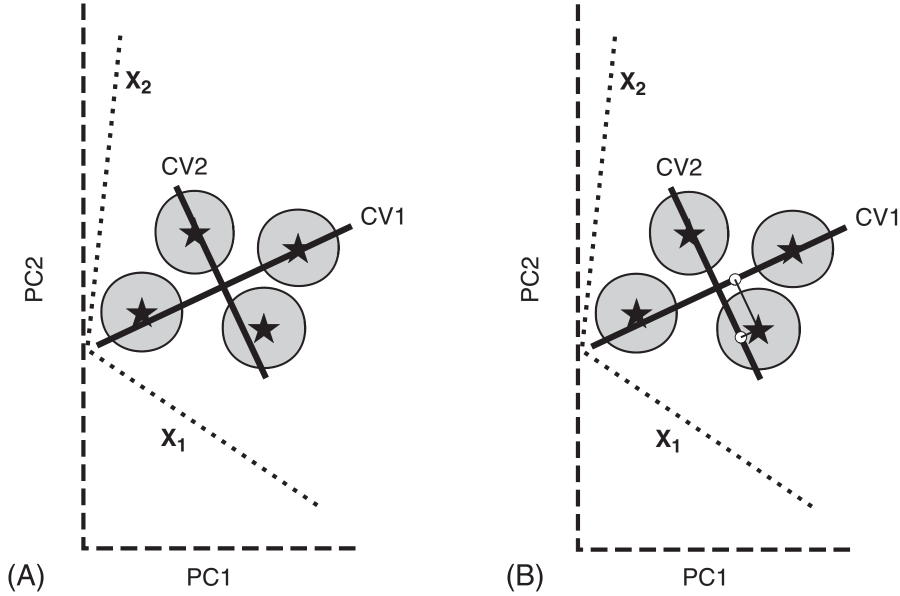


Figure 7.13 Graphical representation of the final steps in CVA. (A) CV1 is the direction through the rescaled space (outer, dashed axes) in which the group means are most different; CV2 is the direction orthogonal to CV1 in which the group means are most different. (B) Scores of individuals in the CV space are their projections onto the CVs. Circles represent the scores of one of the sample means.

However, we also see that the CVs are not orthogonal axes in the original coordinate space. Furthermore, distances on CVs are not equivalent to distances in the original space.

Note that in this example, there are more groups than variables in the original data set. In such cases, the number of CVs will be equal to the number of variables. Most studies will have fewer groups than variables, and in these cases the number of CVs will be one less than the number of groups. If there are three groups in a study, the differences among them can be summarized as a plane defined by two vectors, whether the original data included three variables or 300.

Algebraic description

In CVA, as in PCA, we begin with a set of measures or coordinates $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3} \ldots \mathbf{X_P})$, and we want to find the vector $\mathbf{A_1} = (A_{11}, A_{21}, A_{31} \ldots A_{P1})$ such that:

$$\mathbf{Y_1} = A_{11}\mathbf{X_1} + A_{21}\mathbf{X_2} + A_{31}\mathbf{X_3} + \cdots + A_{P1}\mathbf{X_P} \tag{7.37}$$

In PCA, we solved for the eigenvalues and eigenvectors of the variance–covariance matrix $\mathbf{S}$. In CVA, we are concerned with the ratio of two variance–covariance matrices: one is the pooled within-groups variance–covariance matrix, $\mathbf{S_W}$, which represents the deviations of individuals from their respective group means; the other is the between-groups variance–covariance matrix, $\mathbf{S_B}$, which represents the portion of the total variance (deviations from the grand mean) not explained by $\mathbf{S_W}$. In other words, $\mathbf{S_W}$ represents differences within

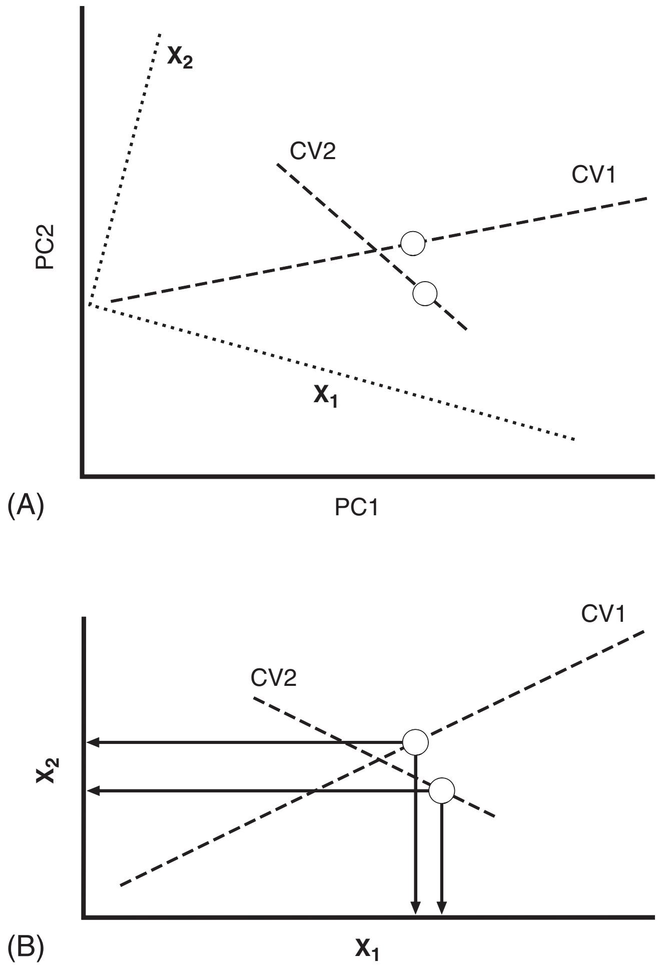


Figure 7.14 Interpretation of CV scores in terms of the original axes. (A) Rescaling the axes has been reversed, restoring orthogonality of $\mathbf{X}_1$ and $\mathbf{X}_2$. White circles represent the scores of one individual on each CV. (B) Rotation of the original axes is reversed, restoring the original orientation. Arrows show projections of the CV scores onto the original axes; each CV score represents a combination of scores on the original axes.

groups, and $\mathbf{S}_\mathbf{B}$ represents differences between the groups. So, in CVA we want to find the $\mathbf{Y}_1$ that maximizes the ratio of between-group variance to within-group variance. The within-group variance of $\mathbf{Y}_1$ is:

$$s^2_{\mathrm{Y}_1\mathrm{within}} = \mathbf{A}_1^\mathrm{T}\mathbf{S}_\mathbf{W}\mathbf{A}_1 \tag{7.38}$$

and the between-group variance of $\mathbf{Y}_1$ is:

$$s^2_{\mathrm{Y}_1\mathrm{between}} = \mathbf{A}_1^\mathrm{T}\mathbf{S}_\mathbf{B}\mathbf{A}_1 \tag{7.39}$$

The form of these expressions should be familiar from our discussion of PCA. As before, we use the Lagrange multiplier λ_1 to form the expression:

$$\left(\frac{s^2_{Y_1\text{between}}}{s^2_{Y_1\text{within}}}\right) - \lambda_1\left(1 - \mathbf{A}_1^T\mathbf{A}_1\right) \tag{7.40}$$

then make the substitutions indicated by Equations 7.38 and 7.39 to form:

$$\left(\frac{\mathbf{A}_1^T\mathbf{S}_B\mathbf{A}_1}{\mathbf{A}_1^T\mathbf{S}_W\mathbf{A}_1}\right) - \lambda_1\left(1 - \mathbf{A}_1^T\mathbf{A}_1\right) \tag{7.41}$$

This is the expression we will maximize relative to $\mathbf{A}_1$, under the constraint that $\mathbf{A}_1^T\mathbf{A}_1 = 1$. Taking the partial derivative of this expression again yields a characteristic equation that can be solved for the eigenvalues and corresponding eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$.

Interpretation of results

Results of CVA will look different from those of PCA for two crucial reasons. First, CVA is describing differences between groups, and the direction in which group means are most different is not necessarily the direction in which individuals are most different. Second, CVA does not simply rotate the original data to the axes that maximize the group differences (if it did, it would be exactly equivalent to a PCA on the group means). CVA finds the axes that optimize between-group differences *relative* to within-group variation and, in general, these axes will be different directions from the ones that maximize between-group differences. In addition, optimization also involves rescaling such that the new

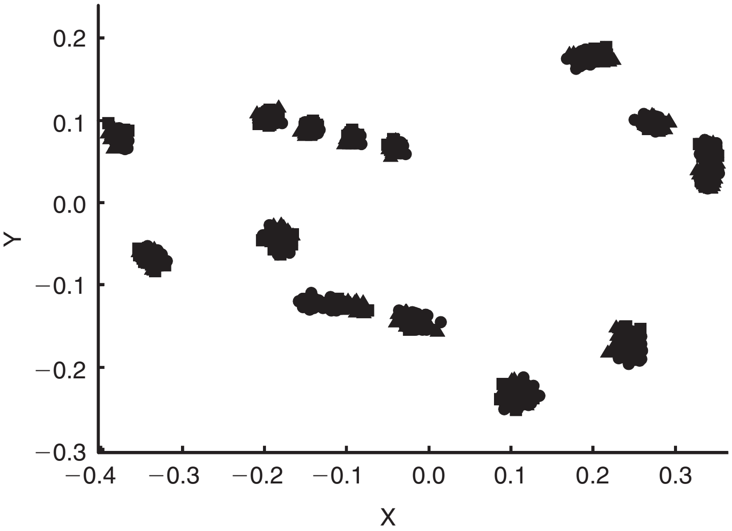


Figure 7.15 Landmark coordinates, in partial Procrustes superimposition, for 119 squirrels from three geographic samples. Circles = western Michigan, squares = eastern Michigan, triangles = southern states.

axes are scaled differently from the original axes and scaled differently from each other. Consequently, distances in CV space can be quite different from distances in the original data, and interpretations of results can be counterintuitive.

To illustrate the differences between PCA and CVA, we show results from performing each on the same data set. This data set is composed of 15 landmarks on the lower jaws of 119 squirrels from three geographic areas. As shown in Figure 7.15, the distribution of shapes in the three groups overlaps broadly, and this broad overlap is reflected in the plot of the first two PCs (Figure 7.16A). Clearly, the combination of shape variables

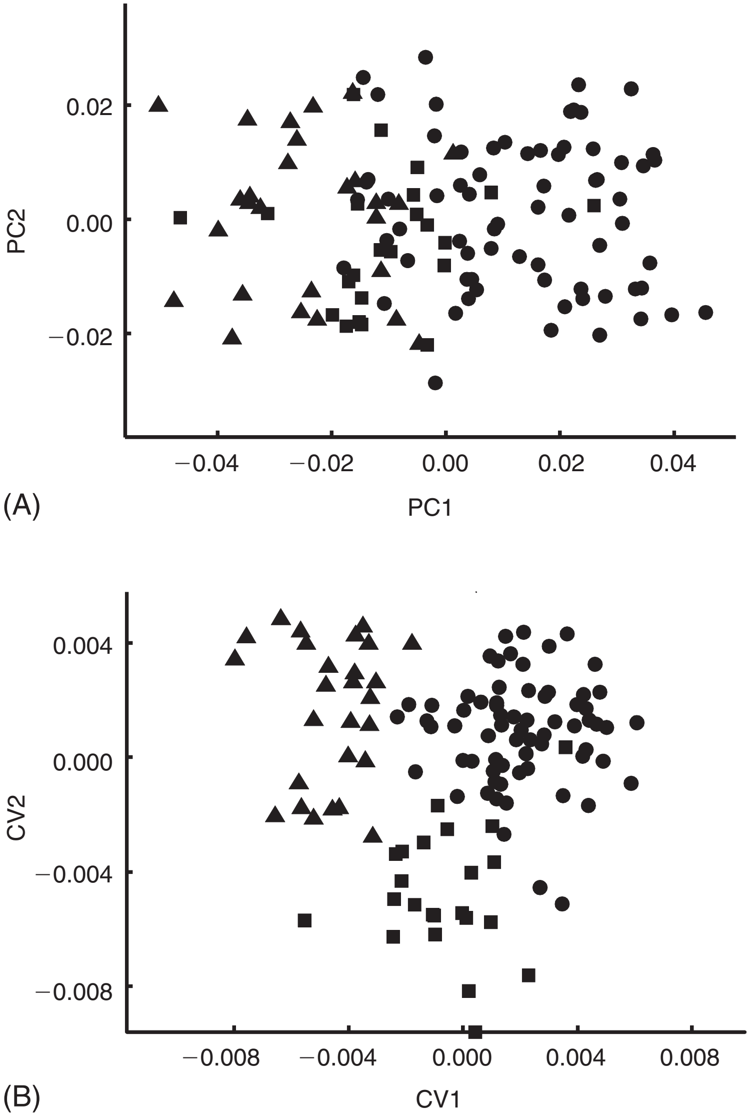


Figure 7.16 Scatter plots from PCA (A) and CVA (B) of 119 squirrel jaws from three geographic areas. Circles = western Michigan, squares = eastern Michigan, triangles = southern states.

represented by these components will not be useful for discriminating among these groups. In contrast, the plot of the two CVs (Figure 7.16B) shows much less overlap, indicating that other combinations of the shape variables are more effective discriminators than the PCs.

Like PCA, CVA will compute a set of axes under the specified constraints, regardless of whether the differences between groups are statistically significant. The optimal discriminator in a data set need not be an effective discriminator. To determine how many CVs are effective discriminators, we employ Bartlett's (1947) test for differences in the value of Wilk's lambda (Λ). Wilk's Λ is the within-groups sum of squares divided by the total sum of squares (within-plus between-groups):

$$\Lambda = \frac{\det(\mathbf{W})}{\det(\mathbf{T})} = \frac{\det(\mathbf{W})}{\det(\mathbf{W}+\mathbf{B})} \tag{7.42}$$

where det is the determinant of the matrix. Conveniently, Λ can be computed as the product of the eigenvalues of $\mathbf{W}(\mathbf{W}+\mathbf{B})^{-1}$. Bartlett's test uses the following formula to estimate a χ^2 test statistic:

$$\chi^2 = -(W - (P - B + 1)/2)\ln\Lambda \tag{7.43}$$

In this expression, P is the number of variables, $W = N - B - 1$ (where N is the total number of individuals) and $B = G - 1$ (where G is the number of groups). The degrees of freedom are determined by the product of P and B.

The testing procedure begins by computing the estimated χ^2 in which Λ is the product of the eigenvalues of all CVs. If this value is significantly greater than expected for the given degrees of freedom, it is safe to infer there are statistically significant differences among the groups. (We will discuss this implication further in Chapter 9.) In the squirrel jaw example, there are three groups and 26 shape variables, and the maximum possible number of meaningful CVs is two. Bartlett's test on both CVs yields a χ^2 of 206.6, with 52 degrees of freedom, for a p-value less than 0.000001. This result indicates that at least some of the groups in the study can be discriminated using scores on these two CVs.

We do not yet know whether both CVs contribute to discrimination of the groups, so the next step is to remove the eigenvalue for the first CV (the most efficient discriminator) and repeat the test. Reducing the number of CVs reduces the number of groups that can be discriminated, which reduces B by 1 and the degrees of freedom by P. These changes produce a χ^2 of 83.5 with 26 degrees of freedom for a p-value that is still less than 0.000001. Thus, the second CV also contributes to discriminating among the groups.

In general, the test is reiterated using the remaining R ($= B - i$) eigenvectors until $R = 0$ (all eigenvalues have been removed) or some set of R remaining eigenvectors fails the test. If R goes to zero, the analysis will have shown that some groups can be discriminated on the CV that is the least efficient discriminator. If a set of R eigenvectors fails the test, then only the first $B - R$ CVs contribute to discriminating among the groups. Note that the test cannot be taken to indicate that all groups can be discriminated, and it does not indicate which groups can be discriminated (see Chapter 9 for further discussion).

The utility of the CVs for discriminating among groups can also be evaluated using the Mahalanobis distances of specimens from the group mean. The means are computed using the *a priori* group assignments. The Mahalanobis distance between a specimen $\mathbf{X}$ and the

Table 7.2 CVA classification table for 119 squirrel jaws

A priori *assignments*	A posteriori *assignments*			*Total*
	Western Michigan	*Eastern Michigan*	*Southern states*	
Western Michigan	62	4	3	69
Eastern Michigan	1	22	0	23
Southern states	0	1	26	27

The *a priori* classifications are based on the geographic localities where specimens were collected. The *a posteriori* assignments are based on Mahalanobis distances of individuals from the means of the *a priori* groups. Total is the total number of specimens in each geographic sample. Thus, 62 specimens in the western Michigan sample were correctly classified using Mahalanobis distance, and 7 were misclassified as members of one of the other geographic samples.

mean **M** of a group, is given by:

$$\mathbf{D} = \sqrt{(\mathbf{X} - \mathbf{M})^{\mathrm{T}}\,\mathbf{S}^{-1}\,(\mathbf{X} - \mathbf{M})} \tag{7.44}$$

where $\mathbf{S}^{-1}$ is the inverse of the variance–covariance matrix of the CV scores of the specimens. The predicted group membership of each specimen based on the scores is determined by assigning each specimen to the group whose mean is closest (under the Mahalanobis distance) to the specimen. All of the CVs that pass Bartlett's test, and only those CVs, are used to compute the Mahalanobis distances and assign specimens to groups. As shown in the first row of Table 7.2, 62 of the 69 western Michigan squirrels have jaws that are closer to the mean of their sample than to the mean of another sample, based on the Mahalanobis distance. In contrast, only one specimen from each of the other samples is farther from the mean of its own sample than it is from the mean of another sample. Like the plot in Figure 7.16B, this result contributes to the general impression that the members of these three groups can usually be discriminated.

Having established that the CVs of the shape variables can be used to discriminate members of the three geographic samples, it would be useful to know what patterns of shape differentiation the CVs represent. So, as we did with the PCs, we multiply the original shape variables by the coefficients of the CVs and sum them. This produces a series of vectors of relative landmark displacement that illustrates the shape differentiation represented by the CVs. Figure 7.17A is a highly exaggerated picture of CV1 for the squirrel jaws. This figure shows that differences in the relative heights of the teeth are the most useful trait for discriminating among the groups. When the deformation is scaled to represent the actual magnitude of the difference between groups along this axis (Figure 7.17B), the amount of the shape difference is imperceptible. Figure 7.17C illustrates all of the other shape differences that are correlated with CV1. The comparison of the figures demonstrates an important point to bear in mind when using this method: the CV is not a complete description of the difference between groups, even when the group centroids lie on the axis. In fact, the CV may be only a small part of the difference between the groups. The CV is simply the part of the difference that is the most effective discriminator, the part that has the least variation within groups relative to the difference between groups.

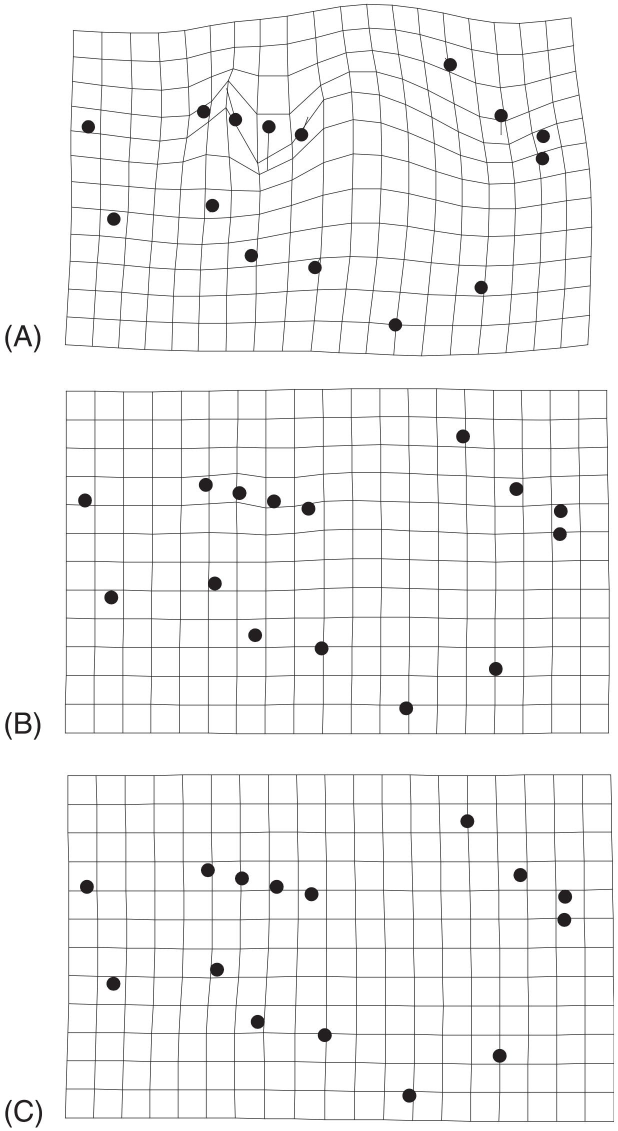


Figure 7.17 Shape differentiation associated with CV1 for the three geographic samples of squirrel jaws. (A) Transformation of the reference shape to the shape corresponding to a score of 0.1 on CV1; (B) transformation of the reference shape to the shape corresponding to a score of 0.01 on CV1, reflecting the actual magnitude of difference between the means of the eastern Michigan and southern samples; (C) deformation representing all of the shape change correlated with CV1, for an individual with a score of 0.01 on CV1.

Software

The IMP series includes a program that performs PCA, **PCAGen**, and one that performs CVA, **CVAGen**. The two programs have nearly identical interfaces with many of the same options. **CVAGen** requires you to execute a few extra steps and offers a few options that are

Table 7.3 Part of the group list file used in analyses of squirrel jaws

1
1
1
1
1
...
7
7
7
7
7
...
10
10
10
10
10

The ordinal position in this file corresponds to the ordinal position of the specimen in the data file generated by **CoordGen**. The numerical value of the code specifies the symbol used in the plot generated by **PCAGen** and **CVAGen**. 1 = black circle, 7 = yellow triangle, 10 = red square.

not available in **PCAGen**. Accordingly, we describe running **PCAGen** first, then describe the differences between **PCAGen** and **CVAGen**. In the last part of this section we describe the program **CCoder**, which can be used to define the symbols used in plots generated by **PCAGen** and **CVAGen**. Both programs read files in standard IMP format (X1, Y1, ... CS). When the file is loaded, both programs will display the superimposed landmarks in the visualization window.

PCAGen

PCAGen:

1. Performs a Procrustes GLS superimposition and provides a plot of the superimposed specimens (with groups color- or symbol-coded if desired)
2. Calculates partial warps and uniform component scores
3. Extracts the principal components of those scores and plots them as well as depicting the variables loading on the PCs
4. Determines the number of distinct eigenvalues based on Anderson's test.

Running PCAGen

Load a file; you may use coordinates obtained from any superimposition method. **PCAGen** will perform a Procrustes (GLS) superimposition before carrying out the analysis (the first plot you see will be a GLS superimposition of the data). You will also need to load another file if you want to color-code or symbol-code your plots (so you can visually distinguish among groups). This is the **GroupList** (described below). Loading a **GroupList** is optional. If your data come from a single homogenous population (or if you do not wish to plot by groups) you can select the **No group list** option. An example of a **GroupList** is shown in Table 7.3, part of the list for the squirrel jaws (the full file comprises 119 entries), in one column of one- or two-digit numbers. The number in each row is a code that corresponds to the *a priori* grouping, in this case, the geographic locations where the specimens were collected. (The key to the group list codes is given in the **PCAGen** manual; **CCoder** allows you to alter the default codes.)

After you load the data, and specify either of the **GroupList** options, the landmarks for all specimens will (eventually) appear in the visualization window. You can save the plot now, or reproduce it later by clicking on the **Show Landmarks** button. If you want to show the landmarks using a different superimposition method, select the desired option from the **Show Landmarks** menu (the PCA will be based on a GLS superimposition regardless of the method you select for displaying the landmarks). If you have a large number of individuals, you may wish to plot only the means; you can do that by going to the **More Plots** pull-down menu on the toolbar and selecting the desired superimposition for the means. If you do not load a **GroupList**, the mean that is shown will be the mean shape of the entire sample. If you do give a **GroupList**, the mean of each group will be plotted. You may wish to rotate the plot (e.g. the anteroposterior axis is oriented vertically or tilted, and you want it oriented horizontally); you can change this in either of two ways. One way to change the plot orientation is to type in the desired angle through which to rotate the image (if you know what it is) by typing it into **Default Ref Angle** box (on the lower left); otherwise click on the **Reference Rotation Active** radio button to find the orientation you prefer by trial-and-error. Activating the rotation option will cause a new window to pop up, where you can type in the angle through which you want to rotate your plot, and keep going until you find the one you prefer (at which point you either hit **Cancel** or type zero). You can type in "10" and keep going in increments of 10; on your next plot, the net rotation used in your first plot will appear in that window (e.g. if you type in 10 and you rotate by 10° three times, the next time the window appears in that session, the value of 30 will appear in it). You may save the plot either by clicking on the **Copy Image to Clipboard** option, or **Copy Image to EPS File** (an encapsulated postscript file that can be imported into a graphics program such as Adobe Illustrator). The plot can be edited before copying, using editing options available in the **Display Options** and **Axis Controls** pull-down menus on the toolbar. The **Display Options** allow you to adjust the line weight, symbol fill and symbol size; the **Axis Controls** allow you to remove the axes from your plot (and restore them later).

The PCA is finished as soon as the superimposed specimens appear in the visualization window. To find out how many distinct eigenvalues are in the data, go to the **Statistics** pull-down menu on the toolbar and click on **Significant Differences in PC Components**. This will generate a Window's window that tells you the number of distinct eigenvalues. *Before* you click the **OK** button (which closes the window), write down the number of distinct eigenvalues. If you really want the *p*-value of the test, save the eigenvalues as described above and plug the values into Equation 7.36. The same menu gives you two options for displaying a scree plot, either of the eigenvalues or of the percent of variance described by the eigenvalues. If you want to save the scree plot, click on one of the yellow **Copy Image to …** buttons below the plot. Depending on your graphics software, you might want to edit the image of the scree plot before copying it. Click on the **Display Options** pull-down menu. Use the options to adjust line weight, symbol fill and symbol size. The other options are only available for plots of deformations.

The eigenvalues are not displayed on the screen (only the percent variance explained is), but they have been computed and can be saved to a text file. To save them, select the **File** pull-down menu at the top of the screen, and then select **Save Eigenvalues**. You can also save the eigenvectors (select **Save Principal Component Vector Matrix**). You can also save the PCA scores, partial warp scores and the reference (consensus) form; to save these, click on the button on the interface. Use the pop-up window to choose the folder and filenames.

To plot the scores, click on the **Show PCA Plot** button in the purple field. This plots the scores of all individuals on the two PCs indicated in the windows below the buttons. To label the points (by specimen number), click the **Label Points** radio button (below the window), then click **Show PCA Plot** again. You can plot other pairs of PCs by clicking the **Up** and **Down** buttons or by entering the ordinal numbers of the desired PCs in the boxes on the left. If you have many specimens crowded together it may be difficult to see them; to zoom in, go to the **Axis Controls** menu and zoom in; to restore the original size of the plot, select **Original Plot Size**. The plot can be edited using the options available in the **Display Options** pull-down menu. You can alter the line weight (the lines being altered are those surrounding the symbols), you can fill the symbols (or remove the fill) and you can also adjust the size of the symbols. The plot can be copied to the clipboard or to an EPS file.

To generate a picture of the shape difference along a PC, select the PC and choose the superimposition you prefer for this display. If you want Bookstein coordinates (BC) or sliding baseline registration (SBR) superimpositions, and did not already specify the baseline, you must do this now. The deformations can be displayed using a variety of graphical methods; the default is the deformed grid. To select another, go to the **Deformation Display Format** menu; among the alternatives are a quiver plot, relative landmark displacements depicted as vectors on the landmarks of the reference form, and a combination of the deformed grid and vectors of relative displacements of landmarks. Each time you choose a different option, you will need to ask the program to **Display Deformation** again. The default is to show the deformation of the reference into a hypothetical specimen having a score of + 0.1 on PC1 and 0.0 on every other PC. To change the scale (such as to see the deformation to a specimen having a score of 0.2 on PC1), you can enter a number in the **Scaling Factor** window. If you want to show a specimen with a score of −0.1, you can type in the minus sign in the **Scaling Factor** window.

The image can be edited using the options available in the **Display Options** pull-down menu and on the interface. To edit the plot, go to the **Display Options** pull-down menu; you may alter line weight, line color, plot density (the number of lines used in drawing the grid), symbol type, whether arrowheads are used in drawing vectors of relative landmark displacement, whether symbols are filled, and the symbol size (this applies both to the size of the symbols in the scatter plot of scores and to the size of the symbols representing the landmarks).

If the grid does not fully encompass the specimen, you can increase the range of the grid using the **Adjust Grid Size for PW** in the blue-green field (below the **Deformation Display Format** menu). If the grid is too large, you can trim it by clicking on the **Grid Trimming Active**, which is centrally located at the bottom of the interface. The first step in trimming the grid is to define the lower left boundary of the plot, which is done by moving the red box right or left (this box appears when grid trimming is activated). To move it, left-click the mouse, walking the box across the grid, until it is positioned correctly, then right-click the mouse to set that value. Next you need to specify the bottom boundary of the grid; now move the red box vertically, left-clicking the mouse to walk the box vertically until you reach the desired location, then right-click the mouse. Next, set the upper right extent of the grid by left-clicking the mouse, moving the red box horizontally until you reach the desired location and set that value by right-clicking the mouse. Finally, to set the top of the grid, left-click the mouse and move it vertically until you reach the desired location, then right-click to set the value. The grid will be redrawn between those limits. When you have edited the image to your satisfaction, you can either copy it to the clipboard or save it to an encapsulated postscript (EPS) file. After trimming the grid, the axes you see on the next plot of scores may be compressed – turn off the grid-trimming option and redisplay a plot (such as the deformation). The axes should now be more conveniently scaled.

You may wish to display the deformation between a particular pair of specimens, or along a direction other than a PC. You can do that using the options in the **Show Deformation Implied by PCA** menu in the lower right. You will need to choose your desired superimposition method for this plot – the default is BC. You can either show the deformation from the consensus (located at the origin on the plot of the scores) to a single specimen, or the deformation between any two specimens. The two specimens are symbolized in the menu as **M1** and **M2.** If you want to display the deformation from the reference to **M1**, click on the **Place M1** button. The cursor is replaced by a pair of cross-hairs (which may be partly hidden by the various boxes on the interface). Move the cross-hairs to the plot window, center them over one of the specimen points, and click the left mouse button. A red diamond will appear on top of the specimen point. Select the superimposition type to be used in this display in the gray box *within* the pink field, and go to the **Deformation Display Format** in the blue field, as before. Now click the **Show M1** button. The picture will represent the sum of the deformations specified by the scores on both of the selected axes. It represents only a part of the total deformation of that individual from the consensus; differences that are not within this plane are not depicted.

If you now want to look at the deformation between two specimens, rather than between one and the consensus, redisplay the PCA scatter plot (go to the purple field again). If you want one of them to be the specimen you had already selected as M1, go to the pink field and click the **Restore Markers** button. The red diamond will reappear where you had placed it. Click the **Place M2** button, and use the cross-hairs to select a second specimen. A green diamond will appear on it. You can click on the **Show M2** button to display the deformation represented by that pair of scores, which you might want to compare to the deformation of the specimen under M1. To show the difference between M1 and M2, click the **Show M2-M1** button. This will produce a picture of the second specimen as a deformation of the first, using the differences between the specimens' scores on these two PCs. Again, this will only be the part of the difference between the specimens that lies in the plane of the two components. (The **Marker Exaggeration** box in the pink field functions like the **Scaling Factor** box in the gray field; the scores of the marked locations are multiplied by the indicated amount.) You do not have to place the markers on actual specimens – you can place them wherever you want.

CVAGen

CVAGen conducts a canonical variates analysis. **CVAGen:**

1. Performs a Procrustes GLS superimposition and provides a plot of the superimposed specimens (with groups color- or symbol-coded)

2. Calculates partial warps and uniform component scores
3. Extracts the canonical variates of those scores and plots them, as well as all the variables correlated with them
4. Performs Bartlett's test
5. Uses the discriminant function to classify specimens into groups
6. Does an assignment test, which determines the probability that the specimen is closer to the mean of the group to which it was assigned *a priori* than to the mean of another group.

Running CVAGen

When you start **CVAGen**, two windows will open. The main window is almost identical to the interface for **PCAGen**, the second is an **Auxiliary Results** window that you can minimize for now. Because of the similarity in interfaces between the two programs, we will focus herein on what differs (if you skipped the section describing **PCAGen**, read it now).

Unlike **PCAGen**, **CVAGen** requires a **GroupList** because the whole purpose of the analysis is to compute variables that optimally discriminate among groups. If you did not already read the description of the **GroupList** file given above, or look at the example **GroupList** file in Table 7.3, do so now. After the **GroupList** file has been loaded, **CVAGen** will compute the Procrustes superimposition of all the specimens, the partial warp and the uniform scores, the CVs and their eigenvalues, and perform Bartlett's test to determine the number of CVs that discriminate among the groups. When this is done, the Procrustes superimposed coordinates will appear in the plot window, with different symbols indicating group membership. Again, alternate superimpositions can be selected using the green **Show Landmarks** box, but, regardless of what you choose for displays, computation of CVs are based on the Procrustes superimposition of all specimens on the sample mean.

When the computations are completed, a pop-up window will appear, showing the summary results of Bartlett's test. This information is also written to the **Auxiliary Results** box window, along with additional details of the test. These results can be redisplayed at any time by going to the **Statistics** pull-down menu and selecting **Tests of significance**. In this same menu, select **Show groupings by CVA** to generate a table comparing *a priori* group assignments to the classification that would be based on the Mahalanobis distances of each specimens all of the group means. This table will be shown in the **Auxiliary Results** box window. These results can all be saved using options on the **File** menu. You can also copy the contents of the **Auxiliary Results** box and paste it into a file, by selecting the text, copying it (Ctrl-C) then pasting it (Ctrl-V). You can also see the results of the **Assignment Tests**, and save those or copy and paste them to a file. The assignment test determines the probability that a Mahalanobis' distance between an individual and the mean of the group is larger than expected under a null model of random variation around the mean of each group. The *p*-value for each specimen indicates the probability that it is a member of the group to which it is assigned (low values indicate that it is unlikely to be a member of that group). The particular method for determining that probability, and the particular implementation of the test in **CVAGen**, has yet to be subject to peer-review; for that reason, the test should now be regarded as useful more for heuristic purposes than as a rigorous or valid statistical test.

CVAGen provides the same options as **PCAGen** for generating plots. You can plot the superimposed specimens, scores of specimens on the canonical variates and the shape variables maximally discriminating among groups. The one unfamiliar (and very important) option is the **Regr?** radio button. This will direct **CVAGen** to regress all shape variables on the scores for the selected CV. The result will be a picture of all shape differences correlated with the CV. As discussed earlier in the text, this picture can be quite different from the picture of the CV itself (see Figures 7.16A, 7.16B, and related discussion). Use this option if you wish to display all of the differences between groups, and not just the most efficient discriminator.

All the options for editing and saving the plots are the same as for **PCAGen**.

CCoder

The programs **PCAGen** and **CVAGen** use a default set of 12 colored symbols to plot landmark coordinates and scores on scatter plots. **CCoder** (Color Coder) is a utility you can use to specify a different set of colors that are better suited to presentation graphics, to increase the number of symbols (if you have more than 12 groups), or use black-and-white or gray-tone symbols that are better suited to printed manuscripts.

To create a code file, go to the **File** pull-down menu and select **Start New Group Code File**. A pop-up window will appear asking if you want to change the default for the number of groups. If the default value of 12 groups is enough, click **No**. If you need more than 12, specify your desired value. The program will then proceed to load a set of default symbol and color codes. Use the **Up** button to the right of the plot window to scroll through the **Active Groups**. This will give you a preview of the symbols, each in a successively lighter gray tone. Note that higher numbers are associated with lighter tones.

To modify a group code, use the **Up** or **Down** button to make that group code active (e.g. if you wish to modify the code for Group 1, move up to activate Group 1). The size, shape and color of the default code will appear in the plot window. Use the options below the window to modify the symbol. None of the options will take effect until you click on the **Show Symbol and Set Values** button. When you click on the button, the new symbol will be displayed and the new code will be written to a temporary file.

When creating the codes it is important to understand that mixing screen colors is not the same as mixing paints. You can mix red and blue to make purple, but mixing red and green makes yellows and oranges. Rather than thinking about mixing paints, think of balancing lights of three pure wavelengths. If all three numbers are 0, the symbol will be black because there is no light of any wavelength. If you increase the intensity of color, you get a pure color tone and a progressively lighter symbol. In general, the symbol will be quite dark if the intensity is less than 50 because there is little light of any color being emitted. Similarly, if one color has a high intensity but the others are low, the color will be indistinguishable from the pure tone because little additional light is being added to the dominant light color.

When all three colors are at the same intensity, the symbol will be some shade of gray (you can control how dark the gray is by the values you give each color – the lower they are, the darker the gray). When all three numbers are 100, the symbol will be white. (*Caveat*: the color you see on one screen may not be the color you see on another screen or on a printed page. Translations to printers or graphics programs that use cyan–magenta–yellow codes can be especially tricky; so you might want to see a preview to avoid surprises.)

The changes you make are saved to a temporary file, so the original codes are not overwritten until you return to the **File** pull-down menu and select **Save Group Code File**.

Enter the name you wish to give the file in the pop-up window. Be sure you have saved changes to a file *before* you click **Exit**. The files produced by this program are text files, but you may want to use an extension like .cod instead of .txt to help you keep track of the files.

To instruct **PCAGen** and **CoordGen** to use your group code file instead of the default, you need to load this file into **PCAGen** and **CoordGen** before it plots the superimposed landmarks. This means you need to load the **Group Code File** after loading the data file and before entering your **GroupList**. The option to load the **Group Code File** is on the **File** menu on the toolbar up top.

References

Anderson, T. W. (1958). *An Introduction to Multivariate Analysis*. New York, Wiley.

Bartlett, M. S. (1947). Multivariate analysis. *Journal of the Royal Statistical Society*, *Series B*, **9**, 176–197.

Campbell, N. A. and Atchley, W. R. (1981). The geometry of Canonical Variates analysis. *Systematic Zoology*, **30**, 268–280.

Chatfield, C. and Collins, A. J. (1980). *Introduction to Multivariate Analysis*. Chapman and Hall.

Morrison, D. F. (1990). *Multivariate Statistical Methods*, 3rd edn. McGraw Hill.

8

Computer-based statistical methods

Most active scientists are probably familiar with some computer-based statistical methods, such as the bootstrap, jackknife, permutation or Monte Carlo simulations. The point of this chapter is to present their underlying principles in a coherent fashion. The basic ideas of all these methods appeared in the work of R. A. Fisher in the 1930s, but the ideas and techniques were neither developed extensively nor used widely until recently. Perhaps the best summary of the discipline is contained in the title of Efron's (1979) paper, *Computers and the theory of statistics, thinking the unthinkable*. The approach he outlined was indeed unthinkable prior to the advent of computers, and it could not be used widely until computers became inexpensive as well as fast, and generally available to researchers (which is why there is such a long time-lag between the original development of the ideas and their widespread application). Computer-based statistical methods are computationally intensive because they replace the complex analytic mathematical methods of classical statistics by an extensive use of randomization and repeated calculations. The enormous number of calculations required by these methods makes them unthinkable without inexpensive (and fast) computers.

Classical statistics is highly algebraic, relying on extended algebraic derivations of formulae that are based on a limited number of well studied distributions, particularly the normal (Gaussian), *F*-, gamma, chi-square, uniform, and Poisson distributions, among others. Before the advent of personal computers, extensive calculations were expensive (in both time or money) so statistical research relied primarily on analytic proofs and mathematical approximations to minimize the number of calculations needed. Currently, most calculations are done by computers even when analytical statistical methods are used; computers have altered even how classical statistical methods are used and taught.

In this chapter, we present a brief discussion of some of the basic statistical concepts that are needed to understand statistical methods in general (such as confidence intervals and hypothesis testing), as well as the more specialized concepts that are needed to understand computer-based statistical methods. We present four classes of these methods, including the bootstrap, jackknife, and permutation tests, and Monte Carlo simulations. To illustrate these methods, we will focus on a few univariate statistical tests. The extension to

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

multivariate statistics is not difficult, but it seems useful to focus on univariate statistics to develop an intuitive understanding of how computer-based methods work. More complete discussions of the topics presented in this chapter can be found in the texts by Efron and Tibshirani (1993) and Manly (1997).

Basic statistical concepts

Human beings look for general patterns, such as connections between events. Statistics provides tools to help identify these patterns as well as to verify that they are reliable, given the information we have. Two statistical tasks related to studying patterns are the estimation of confidence intervals, and hypothesis testing. To talk about these in any detail, we need the basic concepts of:

- samples
- populations
- variables
- probability distributions
- statistics
- parameters.

A *sample* is a collection of individual observations made on a limited number of individuals representing a *population*. An individual observation is the smallest sampling unit in the study, which might be an individual organism or one of its parts, or even a collection of organisms such as a species or bacterial colony. The sample is drawn from a *population*, which is the set of all individuals of a specific type, such as all members of a species, or all teleosts in a given river basin, or all the leaves on an individual tree. In general it is not possible to observe (much less measure) all the individuals in the population, so we rely on the sample, from which we generalize to the population.

A *variable* is a measurement (or observation) made on individuals within the sample. *Univariate* indicates that a single measurement was made on each individual; *bivariate* indicates that two measurements were made, and *multivariate* indicates that three or more measurements were made. Variables can be subdivided into two distinct types: *discrete* and *continuous*. Discrete variables comprise integer-valued variables, which are ordered (e.g. $1 < 2 < 3$), and categorical variables, which are not ordered (e.g. males versus females). Categorical variables are sometimes represented by integers, but no ordering is thereby implied. Continuous variables are real-valued, meaning that they are measured on an infinitely divisible scale.

A *probability distribution* is a mathematical function that describes the probability of a measurement taking on a particular value or a range of values, depending on whether the variable is discrete or continuous, respectively. Since landmark coordinates and partial warp scores are always continuous variables, we will focus on probability distributions for continuous variables. If X represents a specific value of a measured variable and $f(X)$ is a *probability distribution function*, then the following relationship holds:

$$\text{Probability } (A \leq X \leq B) = \int_A^B f(X)\mathrm{d}x \qquad \textbf{(8.1)}$$

which means that the probability that X lies between A and B is the sum of the probabilities that X is equal to each value between A and B. Given the probability distribution of a variable, or a function of a variable, we can determine (in principle) a large number of useful results, such as the mean and the standard deviation of the variable, as well as the probability of the different outcomes of different types of measurements.

Analytic statistical methods use a number of well-known, well-characterized mathematical models of populations, including the normal (or Gaussian) distribution:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(X-X_0)^2}{2\sigma^2}} \qquad (8.2)$$

where X_0 is the mean of the distribution and σ is the standard deviation. Analytic statistical methods start by assuming a specific distribution function for the variable, then use that distribution function to derive results (analytically) about the probabilities of different events. A commonly used result, based on the normal distribution, is that if X_0 is the mean of the distribution and σ is the standard deviation, then the chance that X measured on an individual drawn randomly from the sample exceeds $X_0 + 1.96\sigma$ is 2.5%. This result is derived by integrating the Gaussian distribution function from 1.96σ to infinity, which yields a probability of 0.025. As you might expect, the results based on analytic statistical models are valid only when the mathematical model of the distribution matches the actual distribution of the measured variables.

We would like to make statements about the population from which our sample is drawn so we can generalize from our limited observations to the population as a whole. That would not be necessary if we could measure every individual in a population, because then we could determine the exact value of any *statistic* of interest. A *statistic* is any mathematical function calculated from all measured individuals, such as the mean, standard deviation, variance, maximum, minimum, or range. Because we cannot measure all the individuals in a population we must rely on our representative sample, which leads to uncertainty about our conclusions. The true value of the statistic in the population is the *parameter*, which is what we are trying to estimate from our sample.

Confidence intervals

The confidence interval tells us the range of values that a given statistic might have in light of the uncertainties due to sampling. Whenever we present results based on samples, such as the value of a mean, we must also include a statement about how accurately that value is known. Otherwise, we cannot evaluate the meaning of the difference between two measurements. For example, suppose you are told that John is 185 cm tall and Bob is 190 cm tall: is it possible to say that Bob is taller than John? In this case, it would seem easy to say that he is because these are two individuals – they are not samples representing populations. However, even in this case, the height of an individual is a sample of heights that we could measure for these individuals. Perhaps their heights were measured with their shoes off, with a very precise measuring device, by a single technician practiced in such measurements. Under those conditions, we might anticipate a small error, such as 1 cm or less. Considering that expected error, we could say that John's height is 185 cm, with a *confidence interval* of 184 cm to 186 cm, whereas Bob's height is 190 cm, with a confidence interval of 189 cm to 191 cm. Given that information, we can conclude that Bob is taller than John. Now imagine measuring their heights using a much coarser device, such as marks placed every 10 cm on a pillar in the hallway, and imagine replacing the expert technician by an observer sitting on a chair on the opposite side of the hall who notes the height of people passing by the marks. Furthermore, suppose that two different observers estimate John's and Bob's heights on different days, sitting on chairs of different heights. Under these conditions, we might expect substantial measurement errors for a large variety of reasons, including differences in posture between John and Bob, the difficulty of measuring heights of moving objects, differences in the thickness of the soles of their shoes, as well as differences between observers and the parallax introduced by changing chair height, and so on. The measured heights might be off by as much as 5 cm, so John's height should be given as 185 cm with a confidence interval ranging from 180 cm to 190 cm, while Bob's height should be given as 190 cm with a confidence interval ranging from 185 cm to 195 cm. Each estimate would be within the confidence interval of the other estimate, so we could not say that Bob and John are *significantly different* in height.

We are all used to thinking about numbers mathematically, so it may seem obvious that 190 is greater than 185. However, we need to bear in mind that scientific measurements are not pure mathematical objects; rather, they result from human attempts to assign numbers to physical quantities. Not all attempts are of equal quality, so we need to take that quality into account when comparing measurements. That is part of what enters into confidence intervals.

Most often, we are comparing statistics calculated over samples drawn from populations, rather than comparing observations on two individuals. Sampling introduces another possible source of error. We wish to make statements about the population, so our confidence interval must reflect not only the quality of the individual measurements but also how well the sample is expected to reflect the population from which it is drawn. Statistics calculated from large samples drawn from homogeneous populations will tend to yield accurate estimates of the population's parameters, whereas those calculated from small samples drawn from heterogeneous populations will not. The confidence interval expresses the uncertainty of sampling as well.

We need the confidence intervals as well as the estimated statistics to determine if populations differ. The confidence intervals for many statistics can be calculated from the probability distribution functions of the population. For example, if we assume that the population follows a normal distribution, we can state the uncertainty of a measurement by stating its standard deviation. In doing so, we imply that the measurement follows a normal distribution. Another approach to stating the uncertainty of the measurement is to give its *95% confidence interval*. Doing so implies that repeating the measurement process many times would result in 95% of the measured values falling within that interval (which is equivalent to saying that there is a 95% chance that any single repetition would fall within that interval), but it does not imply anything more about the distribution of values within those limits.

Just as the value of the mean is an estimate from a sample, so is the confidence interval. Different statistical approaches differ in how accurately they are able to estimate confidence intervals. We return to this issue below when we talk about using computer-based methods to estimate confidence intervals.

Hypothesis testing

Statistics works with limited samples, so it is difficult to prove that a statement is true. Statistics is more useful for proving that a statement is false within some level of confidence. Our measurements will always have some degree of uncertainty, so we cannot prove the truth of a statement like "the adult body mass of shallow water crabs is 40 g." Even if we could measure every crab in the population, we would still have some error in our measurements, so we would still have uncertainty in our estimate of the mean. The true mean body mass is likely to be somewhere within a confidence interval, but we cannot say exactly where within that range it is. At best, we can say that the center of the interval is the same as the hypothesized value. This is still not enough to claim that the hypothesis is actually true. We can, however, disprove a statement like the hypothesis with some level of confidence. Suppose that we found that the 95% confidence interval of mean adult body mass is 37–39 g; this means the chance that the mean adult body mass of the shallow water group is greater than 39 g is 2.5% ((100 – 95)/2). By convention, a hypothesis is rejected when there is only a 5% probability of being incorrect, so we would normally reject the hypothesis that the mean weight is 40 g.

In general, when using a statistical approach to hypothesis testing, we state the hypothesis we wish to disprove, which is called the *null hypothesis*. Usually, the null is the hypothesis of "no difference" or "no effect." The null hypothesis states that there is nothing to explain – the apparent signal in the data is an outcome of chance. In our example above, the null hypothesis is: "The adult body mass of the shallow water crabs is 40 g." We could state two alternatives: (1) "The adult body mass of the shallow water crabs is less than 40 g" and (2) "The adult body mass of the shallow water crabs is greater than 40 g." Based on the test of the null hypothesis, the chance that it is true is less than 2.5%. The data are consistent with only one of the two alternatives – that the adult body mass of the shallow water crabs is less than 40 g.

There are several subtleties to note about this approach to hypothesis testing. First, being unable to reject the null hypothesis does not mean that the null is actually true. Instead, our samples might be so small and the variability within a population so great that we do not have enough evidence to reject a false null. When we cannot reject the null, all we can say is that the data are consistent with it. Given more data (or more powerful tests), we might be able to reject it. Another point to note is that we might have more than two alternative hypotheses, and rejecting the null might not tell us which of them is most consistent with the data. Additional statistical tests may be needed to rule out alternative nulls.

When carrying out statistical hypothesis testing, we can make two kinds of errors. *Type I error* is rejecting a null hypothesis *when it is true*; the converse is *Type II error* – i.e. failing to reject the null *when it is not true*. Thus, Type I error means that we improperly rejected the null, and Type II error means that we have improperly failed to reject it (working through the double negatives can be difficult, but we do not ever *accept* the null, so we cannot avoid using the phrase "fail to reject"). The rate of Type I errors is controlled by setting the *alpha level* of the test, which is the chance that the null hypothesis is true. The alpha level typically favored is 5%, meaning that the null is rejected as "untrue" when the estimated probability of the observed value of the test statistic (under the null hypothesis) is 5% or less. For example, if the probability is less than 5% that two populations have the same mean, we reject the null hypothesis of equality of means. The rate of Type II error is more complex, being influenced by: (1) the nature of the null hypothesis and the alternatives; (2) the statistical test used; (3) sample size; and (4) the alpha level used to reject the null hypothesis. Higher alpha levels lower the rate of Type I error but increase the rate of Type II error (for a more extensive discussion of the relationship between these errors, see Sokal and Rohlf, 1995). Estimation and control of the Type II error rate is rather difficult, and many workers focus on Type I error, neglecting Type II error, when assessing results of statistical tests.

Why we need computer-based statistics: an example

To develop an intuition about the need for computer-based statistics, we will work through an example. Suppose X is a set of 31 observations of a length:

$$X = \{2, 2, 3, 4, 2, 5, 3, 2, 6, 2, 3, 4, 6, 2, 1, 4, 3, 7, 2, 3, 4, 4, 5, 8, 5, 2, 1, 3, 4, 4, 3\} \quad (8.3)$$

In this case, $N = 31$. We can compute the mean (denoted $<X>$ for "the expectation of X") by:

$$<X> = \sum_{i=1}^{N} \frac{X_i}{N} \quad (8.4)$$

where X_i is the ith element in the list. In our example, $<X> = 3.52$. Of course, we also need to quantify our uncertainty in this value. If we assume that the distribution of X fits the model of a normal distribution, then the standard error of the mean is given by the standard deviation σ divided by the square root N (the number of observed individuals). The standard deviation is:

$$\sigma = \left(\frac{\sum_{i=1}^{N} (X_0 - <X>)^2}{N - 1} \right)^{1/2} \quad (8.5)$$

so the standard error of the mean (SEM) is:

$$\text{SEM} = \frac{\sigma}{\sqrt{N}} = \left(\frac{\sum_{i=1}^{N} (X_0 - <X>)^2}{N(N - 1)} \right)^{1/2} \quad (8.6)$$

For our example, $\sigma = 1.69$ and $N = 31$, so $\text{SEM} = 1.69/(31)^{1/2} = 0.304$.

The 95% confidence interval for the mean, *assuming a normal distribution*, ranges from $<X> - 1.96(\text{SEM})$ to $<X> + 1.96(\text{SEM})$ because, for a normal or Gaussian distribution, 95% of the values in the distribution lie within 1.96 standard deviations of the mean. So, for our example, $1.96\ \text{SEM} = 0.304$ and $<X> = 3.52$, so the 95% confidence interval is from 2.92 to 4.12. Suppose that we want to claim that the average body length of this population is greater than 3.0 cm. Again, using the normal distribution, we can calculate that the chance of the mean being less than or equal to 3.0 is 0.049%, so we can reject the hypothesis that the mean is less than or equal to 3.0 at a 5% confidence level (meaning that we accept a 5% chance of rejecting the null model when it was true (Type I error)).

What difficulties arise in this example? First, we have assumed that the distribution is normal. This is important, even though statistics based on the normal distribution are known to be robust to violations of the assumptions of normality. Nevertheless, as the distribution departs further from normality, larger errors appear in the results, leading to increased error rates. The validity of the normal distribution for our example has not been determined. Is that assumption reasonable? If the distribution is normal, 1.9% of the measurements will be less than or equal to zero (that is the expectation under the model). Does that pose a problem? Yes, because we are measuring lengths, and *none* can be less than zero, under any circumstances – in fact, the lower bound may be substantially larger than zero (due to physiological constraints on the size of the organism). So we *know* that our distribution must deviate from the normal distribution. Perhaps that deviation has only a small effect on our estimate of SEM, but we are relying on the reputation of the normal distribution as a robust estimator to reassure ourselves about that. We really do not know what effect that lower bound has on our statistical inferences.

The other difficulty we face is the lack of an exact formula for the standard error of any statistic other than the mean. Suppose we want to know the standard error in the median of the distribution. We can calculate the median of our measurements of X, which equals 3.0, but can we actually conclude that the median of the population is greater than 2.0? We do not really know the range of values that the median might take on for this distribution, and the normal model provides no estimate of the uncertainty in the median. The standard deviation and variance of populations are also of tremendous biological interest, but how do we estimate the range of values for these statistics?

Resampling-based methods

Having noted that we can face serious difficulties when we assume a normal distribution and rely on the theory based on it, we now examine methods that allow us to make statistical inferences without assuming any distribution.

The bootstrap

We begin with the bootstrap because it is probably the easiest to understand. It was not the first computer-based statistical method developed; in fact it is one of the more recent (it was developed from jackknife and permutation methods). The term “bootstrapping” comes from the novel *Baron Münchausen’s Narrative of his Marvellous Travels and Campaigns in Russia*, by Rudolph Erich Raspé (1785), in which the Baron falls to the bottom of a deep lake. He cannot figure out what to do until, at the last moment, he thinks to pull himself up by his own bootstraps. This describes, fairly accurately, the approach used in a bootstrap procedure: the observed data themselves are used as a basis for resampling; we approximate the unknown statistical distribution from which the data were drawn by (randomly) resampling our data.

A bootstrap set is a set of data of the same sample size as the original data set, whose elements are *randomly drawn with replacement* from our original set of observations. To randomly draw them (with replacement) from a set of N elements, a uniformly distributed random number from 1 to N is generated by a random number generator. The corresponding element from the original set of observations then forms the first element in the bootstrap set. For example, given our 31 observations, we will construct a sample that also has 31 observations. The number provided by the random number generator is 8, so we take the value of the eighth individual of our sample as the first value in the bootstrap set. This procedure is repeated N times. Note that a single value from the original data set may appear multiple times in a bootstrap set (this is because we are sampling *with replacement*, meaning that we do not remove an individual from the sample after we have placed its value in the bootstrap set). Additionally, not all values in the original set need appear in the bootstrap set.

To develop an understanding of how a bootstrap set is formed, we'll consider an abstract, symbolic example. Suppose **C** contains five values:

$$\mathbf{C} = \{C_1, C_2, C_3, C_4, C_5\} \tag{8.7}$$

To form a bootstrap version of **C**, we generate a list of five random numbers, each independently chosen and ranging from 1 to 5 (because $N = 5$):

$$L = \{5 \;\; 2 \;\; 4 \;\; 3 \;\; 5\} \tag{8.8}$$

The numbers in L are the ordinal positions of the elements of **C**; $\mathbf{C}_{\mathbf{Bootstrap}}$ contains the corresponding values of **C** (e.g. $L_1 = 5$, so it corresponds to the fifth element of **C**, which is C_5). Thus:

$$\mathbf{C}_{\mathbf{Bootstrap}} = \{C_5, C_2, C_4, C_3, C_5\} \tag{8.9}$$

Note that C_1 does not appear in this bootstrap set, while C_5 appears twice.

Returning to the numerical example presented earlier:

$$\mathbf{X} = \{2,2,3,4,2,5,3,2,6,2,3,4,6,2,1,4,3,7,2,3,4,4,5,8,5,2,1,3,4,4,3\} \tag{8.10}$$

To form a bootstrap set, $\mathbf{X}_{\mathbf{Boot}}$, from **X**, we generate the list, **B**, of 31 random numbers:

$$\begin{aligned}\mathbf{B} = \{&30,8,19,16,28,24,15,1,26,14,20,25,29,23,6,13,29,29,\\ &13,28,2,11,26,1,5,7,7,19,9,7,1\}\end{aligned} \tag{8.11}$$

We then select the elements of **X** corresponding to those ordinal values:

$$\mathbf{X}_{\mathbf{Boot}} = \{4,2,2,4,3,8,1,2,2,2,3,5,4,5,5,6,4,4,6,3,2,3,2,2,2,3,3,2,6,3,2\} \tag{8.12}$$

The first element of $\mathbf{X}_{\mathbf{Boot}}$ is the 30th element of **X** (because 30 is the first element of **B**), and the seventh element of **X** appears three times in the bootstrap set (because 7 appears three times in **B**). We can now calculate the mean, standard deviation and median of $\mathbf{X}_{\mathbf{Boot}}$: $<\mathbf{X}_{\mathbf{Boot}}> = 3.39$, $\sigma_{\mathbf{X}_{\mathrm{Boot}}} = 1.62$, and median$(\mathbf{X}_{\mathbf{Boot}}) = 3$. These values are slightly different from those of the original distribution, $<\mathbf{X}> = 3.52$; $\sigma = 1.69$, and median$(\mathbf{X}) = 3.0$.

To arrive at an estimate of the confidence intervals for these statistics, we will compute a large number ($N_{\mathrm{Bootstrap}}$) of bootstrap sets. We will then determine the 95% confidence interval over the $N_{\mathrm{Bootstrap}}$ sets, forming a *bootstrap estimate* of the confidence intervals on the mean, standard deviation and the median. If we generate 200 bootstrap sets based on **X**, we find that the 95% confidence intervals for the mean is 3.00–4.10; for the standard deviation the confidence interval is 1.23–2.10, and for the median it is 3.00–4.00. The normal model predicted a 95% confidence interval for the mean, 2.91–4.12, so the two methods approximately agree. They appear to differ at the lower boundary (at small lengths), which is where we expect departures from the normal distribution, for the reasons discussed earlier.

The approach outlined here may be extended to virtually any statistic and to any function, univariate or multivariate. We now use it to perform t-tests.

Using the bootstrap to conduct t*-tests*

The t-test is used to compare the means of two samples, the F-test to compare means of two or more samples (both procedures are discussed at more length in the next chapter). The question is whether the mean of one group differs from the mean of another in a statistically significant way. It is possible that the observed difference in means is due to an arbitrary division of one group into two, such that the variability within the two groups gives rise to a difference between means solely by chance.

Let us look again at our sample of 31 measured lengths:

$$X = \{2,2,3,4,2,5,3,2,6,2,3,4,6,2,1,4,3,7,2,3,4,4,5,8,5,2,1,3,4,4,3\} \qquad \textbf{(8.13)}$$

and consider a second group of 18 lengths:

$$Y = \{2,2,3,2,4,2,3,2,8,9,2,9,3,2,3,3,3,9\} \qquad \textbf{(8.14)}$$

Using the normal model, we find that $<X> = 3.52$, $\sigma_X = 1.69$, and $<Y> = 3.94$ and $\sigma_Y = 2.71$. To test whether the means are different, we find the probability of statistic t:

$$t = \frac{(<Y> - <X>)}{\sqrt{\left(\dfrac{\sigma_X^2(N_X - 1) + \sigma_Y^2(N_Y - 1)}{N_X + N_Y - 2}\right)\left(\dfrac{N_X + N_Y}{N_X N_Y}\right)}} \qquad \textbf{(8.15)}$$

with degrees of freedom of $(N_X + N_Y - 2)$. For relatively large values of N_X and N_Y, the t-value will be normally distributed with a mean of zero and a standard deviation of one, *provided that* the null hypothesis of equal means is true. If the absolute value of t exceeds 1.96, we may assert that, under the normal model, there is only a 5% chance of the mean values being that different by chance. We can thus reject the null hypothesis at a 5% level of confidence.

The problem is that the list of lengths contained in **Y** is highly non-normal. Most values are close to 3, but there are several around 8 or 9, so **Y** appears to be rather bimodal. Also, in a normal distribution with a mean of $<Y> = 3.94$ and a standard deviation of $\sigma_Y = 2.71$, we would expect that 7.3% of the measured lengths would be less than zero. So the distribution of **Y** departs substantially from normality, more so than does the distribution of **X**.

To form a bootstrap version of the t-test, we will use the bootstrap approach *to simulate the null hypothesis we wish to reject*. This simple principle is the key to understanding how to form your own bootstrap tests when asking novel statistical questions. The null hypothesis of the t-test is that the means of the two groups are equal, which we can also phrase as the hypothesis that the two groups in question came from a single underlying distribution that was arbitrarily subdivided into two groups. If this were the case, any difference between the means would arise simply by chance. So to test this hypothesis, we assume that the null hypothesis is true – i.e. that **X** and **Y** were drawn from the same population. Therefore we merge the two sets of observations (**X** and **Y**) into a common pool of specimens (**Z**) and draw (with replacement) two bootstrap sets from **Z**, one of size N_X and one of size N_Y, and compute the differences in means between the two bootstrap sets. This is repeated $N_{Bootstrap}$ times. We can then determine the number of times in which the difference between the means of paired bootstrap sets exceeds the observed difference between the means of **X** and **Y**. Expressed as a proportion of the total, we get an estimate of the probability that the observed difference is due to chance; i.e. if the difference between means of pairs of bootstrap samples exceeds the observed differences in 5% (or fewer) of the total number of iterations, we can reject the null hypothesis that the means are equal. This is simply another way of phrasing the statement that the observed difference is statistically significant at a 5% confidence level if the observed difference between means exceeds the 95th percentile of differences between means of the bootstrap sets.

A symbolic example of this merging and subsequent formation of two bootstrap sets may help to develop an understanding of how the test operates. Suppose we have a set **C** of five elements, and a set **D** of four elements:

$$\mathbf{C} = \{C_1, C_2, C_3, C_4, C_5\} \tag{8.16}$$

$$\mathbf{D} = \{D_1, D_2, D_3, D_4\} \tag{8.17}$$

The merged set, **M**, would have nine elements:

$$\mathbf{M} = \{C_1, C_2, C_3, C_4, C_5, D_1, D_2, D_3, D_4\} \tag{8.18}$$

To draw two bootstrap sets out of **M**, we would form a list of five random integers (because there are five elements in **C**), and the elements in **M** corresponding to this list would be the elements in the bootstrap version of **C**:

$$L_1 = \{7 \;\; 5 \;\; 1 \;\; 8 \;\; 5\} \tag{8.19}$$

$$\mathbf{C_{Bootstrap}} = \{D_2, C_5, C_1, D_3, C_5\} \tag{8.20}$$

Note that two elements in $\mathbf{C_{Bootstrap}}$ come from **D**. A second list of four integers is used to form a bootstrap version of **D**:

$$L_2 = \{2 \;\; 4 \;\; 9 \;\; 9\} \tag{8.21}$$

$$\mathbf{D_{Bootstrap}} = \{C_2, C_4, D_4, D_4\} \tag{8.22}$$

The formation of the bootstrap versions of **C** and **D** reflects the null hypothesis that **C** and **D** come from a common underlying distribution. The elements of **C** and **D** are thus interchangeable.

The difference between means of the bootstrapped versions of **C** and **D** can be determined by many repetitions, developing a bootstrap estimate of the distribution of the differences between means produced by the null hypothesis (given the data). When we carry out this bootstrap *t*-test on our numerical example, sets **X** and **Y**, we find that 268 of 1000 bootstrap sets (26.8%) have a difference between means as large or larger than that between the means of **X** and **Y**. Thus, we cannot reject the null hypothesis that these samples were drawn from populations with equal means, the difference between them being due to chance. Using a *t*-test based on the normal distribution, we would have rejected that null hypothesis. Because both samples appear to have non-normal distributions, as discussed earlier, it seems reasonable to attribute the difference between results to violating the assumption of normality.

The bootstrap method is probably the most popular of the computer-based methods for estimating confidence intervals, and it is also one of the easiest to implement.

Permutation tests

Permutation tests pre-date the bootstrap test. They were introduced by R. A. Fisher in the 1930s as a basis for supporting the ideas of the Student's *t*-test rather than as a tool for computation. With the advent of computers, permutation methods could be used profitably for statistical inference. Permutation tests operate in much the same manner as bootstrap tests, but differ in that they resample groups *without* replacement. This makes permutation tests suitable for hypothesis testing, but not for the estimation of confidence intervals (Efron and Tibshirani, 1993).

Again, we can look at a simple, abstract example of how a permutation set is formed to get a sense of how the approach works, and how it differs from the bootstrap. Consider two data sets **C** and **D**:

$$\mathbf{C} = \{C_1, C_2, C_3, C_4, C_5\} \tag{8.23}$$

$$\mathbf{D} = \{D_1, D_2, D_3, D_4\} \tag{8.24}$$

with sample sizes of five and four respectively. We form the merged set **M** of nine elements:

$$\mathbf{M} = \{C_1, C_2, C_3, C_4, C_5, D_1, D_2, D_3, D_4\} \tag{8.25}$$

To produce permutation set versions of **C** and **D**, we want to resample **M** without replacement. To do this, write a list of nine integers, then randomly permute it to form a list **L**:

$$\mathbf{L} = \{5 \quad 2 \quad 6 \quad 8 \quad 7 \quad 3 \quad 9 \quad 4 \quad 1\} \tag{8.26}$$

The first five values in **L** are the ordinal values of the elements in **M**, placed in the permuted version of **C**:

$$\mathbf{C}_{permutation} = \{C_5, C_2, D_1, D_3, D_2\} \tag{8.27}$$

The last four values in the list are the ordinal values of the elements in **M** that are placed in the permuted version of **D**:

$$\mathbf{D}_{permutation} = \{C_3, D_4, C_4, C_1\} \tag{8.28}$$

Note the different way that the permutation sets (Equations 8.27, 8.28) and bootstrap sets (Equations 8.20, 8.22) are constructed from **C** and **D**.

To carry out a permutation test of the hypothesis that the means of the two groups **X** and **Y** (see Equations 8.13 and 8.14) are equal, we would first compute the difference between the means of the two groups, which have sample sizes of $N_X = 31$ and $N_Y = 18$. The second step is to merge the two data sets into a single larger one and form a series of paired permutation sets, each drawn from the merged data set. The first permutation set in each pair, containing N_X specimens, is drawn randomly without replacement from the merged set. The second permutation set of the pair contains the remaining N_Y elements of the merged data set. (No element of the original sets appears twice in the paired permutation sets, and none is omitted.) The difference between means of the two permutation sets is then calculated, and repeated for $N_{Permutation}$ sets. The proportion of times in which the difference between the means of the paired permutation sets exceeds that between the original data sets is taken as the probability that the observed value could have arisen by a random splitting of a single underlying distribution.

The permutation test of the difference between the means of sets **C** and **D** indicates that 21.3% of the permuted sets had a difference in means equal to or greater than the observed difference of 0.428, so we cannot reject the null hypothesis that the means are equal at a 5% level of confidence. The permutation test has produced results agreeing with the bootstrap test (in which 26.8% of the bootstrap sets had a difference between means as large or larger than the observed data set).

It is possible to form permutation tests for a wide variety of statistical hypotheses in a manner similar to the bootstrap (see Efron and Tibshirani, 1993; Good, 1994). However, there is an important difference between the permutation and bootstrapping approaches due to fundamental differences in how they operate. Permutation tests are not suited to the estimation of confidence intervals because the standard deviation of the estimates of a parameter (such as a mean or median) is not a reliable estimate of the standard error in that parameter. Rather, the permutation test yields an estimate of the range of parameter values possible under the null model simulated by the test. In contrast, the standard deviation of the bootstrap estimates of the same parameter yields a reliable estimate of its standard error because the bootstrap resampling simulates a repetition of the process of selecting specimens from the population (Efron and Tibshirani, 1993). When used for hypothesis testing, both methods tend to give very similar results, so it is difficult (and perhaps unnecessary) to determine which approach is preferable in most cases. To some extent, the choice between them appears to be a matter of preference among writers of software. There are some reasons to think that permutation tests may yield a more exact achieved significance level (ASL) than bootstrap approaches (Efron and Tibshirani, 1993), but this is at the cost of precluding estimates of confidence intervals (or standard errors) on the statistics involved.

The jackknife

Jackknife methods (Quenouille, 1949; Tukey, 1958) also preceded bootstrap methods, and, to some extent, have been supplanted by them. Jackknife estimates are obtained by resampling such that one element is left out at a time (hence the name – to use a jackknife, you have to leave one out, either one blade or one specimen). If there are N specimens in a sample, then it is possible to form N jackknife data sets, each with $N - 1$ specimens. If we again look at the set **C**:

$$\mathbf{C} = \{C_1, C_2, C_3, C_4, C_5\} \tag{8.29}$$

The five possible jackknife versions of **C** are:

$$C_{J1} = \{C_2, C_3, C_4, C_5\} \tag{8.30}$$

$$C_{J2} = \{C_1, C_3, C_4, C_5\} \tag{8.31}$$

$$C_{J3} = \{C_1, C_2, C_4, C_5\} \tag{8.32}$$

$$C_{J4} = \{C_1, C_2, C_3, C_5\} \tag{8.33}$$

$$C_{J5} = \{C_1, C_2, C_3, C_4\} \tag{8.34}$$

Jackknife data sets will always be more similar to the original data set than bootstrap sets are because the bootstrap offers a greater variety of ways of resampling the data. The jackknife may be viewed as an approximation to the bootstrap (Efron and Tibshirani, 1993), and it is a good approximation when the changes in the statistic are smooth or linear with respect to changes in the data. The mean is a linear statistic, but the median is not (because the median may change abruptly as observations are added or subtracted from the sample); therefore the jackknife estimate of the mean will not differ much from the bootstrap estimate of the mean, but their estimates of the median may differ considerably.

There are some approaches to combining the bootstrap and the jackknife (see particularly Efron, 1992; Efron and Tibshirani, 1993, Chapter 19, on assessing the error of bootstrap estimates), but otherwise the jackknife appears to offer few advantages over the bootstrap.

Monte Carlo methods

Monte Carlo methods compare the value of an observed statistic to the range of values expected under a given null hypothesis, assuming a model of the populations involved. Like analytical statistical methods, Monte Carlo methods require making assumptions about the nature of the distribution from which populations are drawn. They then fit parameters of the distributional models to the observed samples. In contrast, analytic statistical approaches use algebraic derivations to estimate the values of statistics (and standard errors in those statistics) based on the nature of the underlying distributions. The distinction is that Monte Carlo approaches generate random data sets based on the parameters and distribution of the model; those random data sets are drawn from model distributions having the same sample size as the original one. The distribution of the statistic of interest (estimated over many computer-generated Monte Carlo sets) is used to estimate the mean and standard deviation of that statistic, under the null model and the model distribution used. Monte Carlo methods can be used both for hypothesis testing and for generating confidence intervals.

Monte Carlo methods use numerical simulations to avoid the need for extensive algebraic computations and approximations. It may often be easier to program a Monte Carlo simulation than to determine analytically the distribution of an intricate statistical function, particularly when the statistic is not a linear function. Because it is necessary to assume a model of the distributions of the samples, the Monte Carlo method shares most of the primary weaknesses of analytic statistics; if the observed distribution departs substantially from the model, the Monte Carlo sets will not represent the actual system of interest. One useful feature of the Monte Carlo method is the ability to determine the effect of different distributional models (the ones typically used are the uniform, normal or Gaussian, and Poisson) on the range of values estimated by the Monte Carlo sets. The comparison of observed distributions to those produced by Monte Carlo methods is a powerful approach to hypothesis testing.

For example, if we wish to determine the significance of the observed difference in the means of sets **X** and **Y**:

$$\mathbf{X} = \{2,2,3,4,2,5,3,2,6,2,3,4,6,2,1,4,3,7,2,3,4,4,5,8,5,2,1,3,4,4,3\} \quad (\mathbf{8.35})$$

$$\mathbf{Y} = \{2,2,3,2,4,2,3,2,8,9,2,9,3,2,3,3,3,9\} \quad (\mathbf{8.36})$$

we will test the null hypothesis that the two sets (**X** and **Y**) came from the same underlying distribution, with the observed difference between them being due to a random assignment of specimens into groups. To form the Monte Carlo set, we will assume that the single underlying distribution is normal. We then estimate the mean and standard deviation of this underlying distribution by merging the data sets into a single group. The mean of the single distribution is 3.67 and the standard deviation is 2.1. To determine the significance of the observed difference in the means of the two groups, we generate a series of paired Monte Carlo sets, one with a sample size $N_X = 31$, one with a sample size $N_Y = 18$, and we determine the difference between the two means. We then determine the proportion of $N_{Monte\ Carlo}$ sets in which the difference between the means of the paired Monte Carlo sets exceeds that observed between the means of the original data sets.

For the sets **X** and **Y** above, the Monte Carlo sets were generated under the assumption that both samples were drawn from the same normal distribution, with a mean of 3.67 and a standard deviation of 2.1 (the mean and standard deviation of the combined data sets). In 480 of 1000 pairs of Monte Carlo sets (48%), the difference between the means of the paired Monte Carlo sets exceeds the observed difference between the means of the original data sets, thus the null hypothesis of a single underlying normal distribution cannot be rejected. It should be noted that the combined data set (of all specimens in **X** and **Y**) is probably not normally distributed, so we might want to repeat the Monte Carlo test using other models of the underlying distribution.

Monte Carlo simulations are particularly useful for testing different hypothetical situations when the underlying distributions are believed to be well known. Monte Carlo methods can be used in cases when bootstrap methods cannot, such as to estimate the effect of increasing the sample size on the estimated variance; Monte Carlo simulations are not limited by the observed sample sizes (as bootstrap methods are).

Example: computer-based tests and regression models

To this point, we have focused on *t*-tests, but computer-based methods are useful for a wide variety of tests. To develop a more general understanding of these methods, we now show how bootstrap and permutation methods can be used in regression analysis (the subject of Chapter 10). Both approaches can be used to determine if one set of measured variables Y (the dependent variable) has a statistically significant dependence on a second set of measured variables X (the independent variable). If we have N observations, each of a pair of measurements (X_i, Y_i), then the typical linear regression model is:

$$Y_i = A + BX_i + \varepsilon_i \tag{8.37}$$

The regression slope, B, is given by:

$$B = \frac{s_{XY}}{s_{XX}} \tag{8.38}$$

The intercept term, A, is given by:

$$A = <Y> - B<X> \tag{8.39}$$

where $<X>$ and $<Y>$ are the expected values (means) of the X_i and Y_i values, and

$$s_{XX} = \sum_{i=1}^{N} (X_i - <X>)^2 \tag{8.40}$$

$$s_{XY} = \sum_{i=1}^{N} (X_i - <X>)(Y_i - <Y>) \tag{8.41}$$

are the values of A and B which minimized the summed square residuals (ε_i). This sum of squared error terms is:

$$\text{Error} = \sum_{i=1}^{N} (Y_i - A - BX_i)^2 = \sum_{i=1}^{N} (\varepsilon_i)^2 \tag{8.42}$$

under the assumption that the residuals are independently and identically normally distributed.

To show that there is a statistically significant dependence of Y on X, it is sufficient to show that the confidence interval on the slope excludes zero. This is equivalent to showing that there is a non-zero correlation between Y and X, which may be tested using the squared value of the correlation coefficient (R^2) between X and Y, which indicates the fraction of the variance in the dependent variable (Y) that is explained by the independent variable (X). The expression for R^2 is:

$$R^2 = \frac{s_{XY}^2}{s_{XX}s_{YY}} \tag{8.43}$$

where

$$s_{YY} = \sum_{i=1}^{N} (Y_i - <Y>)^2 \tag{8.44}$$

It is very common to interpret high R^2 values as being indicative of high explanatory power in a regression model. There is a method of testing whether an R^2 value is statistically significant (under the assumption of normality of the residuals), by the expression:

$$\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) \tag{8.45}$$

which is a normally distributed variable, with variance equal to $1/(N-3)$, where N is the sample size.

The significance of the slope can be assessed by a permutation test. The objective is to determine the range of slopes that could be generated by random permutations of the associations among X and Y values. Thus, we again adopt the strategy of assuming that the null hypothesis is true (which, in this case, is that the associations among X and Y values is random). The associations of the X_i values with the Y_i are then randomized, generating a permutation set of paired X and Y values with the same distribution of X and Y values as in the data, but with randomized combinations of X and Y. The regression model is then fitted to each permutation set, and the slope (or correlation coefficient) is calculated. The distribution of the regression slopes (or the correlation coefficients) generated by the permutation sets can be used to determine if the observed regression slope (or correlation coefficient) could have been produced by a random association among X and Y variables. If the observed slope (or correlation coefficient) is outside the 95% confidence interval of the permutation sets, then we can reject the null hypothesis that the slope (or correlation coefficient) does not differ from zero. Note that the permutation test estimates the range of slopes (or correlation coefficients) *produced by the null model*, not by the observed data. Thus we reject the null hypothesis by showing that the observed statistic lies outside the range of the values predicted by the null model.

To carry out a bootstrap test of the significance of the regression line, two approaches are available: one is to bootstrap (resample with replacement) the *paired observations* (X_i, Y_i); the other is to bootstrap the *residuals* from the regression. When bootstrapping specimens, we form bootstrap sets by sampling (with replacement) from the paired specimen values (X_i, Y_i) to form a bootstrap set. The regression model is fitted and the slope (or correlation coefficient) is determined for each bootstrap set, forming a bootstrap estimate of the confidence intervals for the slope (or correlation coefficient). This yields a confidence interval on the slope itself, so that if it excludes zero, we can reject a null hypothesis that the regression slope (or correlation) is zero.

The alternative is to bootstrap the residuals, by first determining the residuals to the bootstrap, and the Y values that are predicted by the regression model for each X value:

$$Y_{predicted} = A + BX \tag{8.46}$$

Then the residuals are randomly combined with the paired X_i and $Y_{predicted}$ values, both of which are resampled (with replacement). This approach produces a wider variety of possible paired values of X_i and Y_i; it can be thought of as bootstrapping the variable part of the distribution, independently of the portion that is dependent on X. The range of slopes (or correlation coefficients) is determined over many bootstrap sets; if the 95% confidence interval for the slope (or correlation coefficient) excludes zero, we can infer that there is a statistically significant dependence of Y on X at a 5% confidence level.

The discussion of how a permutation test is used to determine the statistical significance of a regression slope serves as a useful illustration of the differences in approach between bootstrap and permutation methods. In the permutation method, the approach is to estimate the confidence interval under the *null model*, given the distribution of observed data. Thus, if the observed statistic is outside the confidence interval of the null, the observed statistic is judged to be significant. In contrast, the bootstrap approach estimates the range of the statistic on the *observed data* (rather than the range under the null). Permutation tests almost always focus on estimating distributions under the assumption that the null model is true, whereas bootstrap methods can be used to estimate the distribution of a statistic either over the observed data or under an assumption that the null is true.

Issues common to all computer-based methods

Statistical power

When evaluating the utility of statistical tests we are faced with Type I error (i.e. falsely rejecting the null hypothesis when it is true), which is controlled by setting the *alpha* level of the test. Because that is under control, statistical tests cannot be said to differ in their rates of Type I error. In contrast, statistical tests can differ in their rates of Type II error (i.e. failure to reject the null hypothesis when it is false and an alternative is true). The rate of Type II error depends on the nature of the test, the null hypothesis and the alternative hypotheses used. The *power* of a statistical test is its ability to distinguish between the false null hypothesis and the true alternative, and it is sometimes expressed as 1 minus the rate of Type II error.

Estimating the power of statistical tests turns out to be both difficult, and neglected by many researchers. Some work indicates that permutation, bootstrap and analytic tests have equivalent statistical power when the data meet the requirements of the analytic tests (Hoeffding, 1952; Robinson, 1973; Romano, 1989; Manly, 1997). Edgington (1995) reports higher statistical power for randomization tests when there are violations of the assumptions of the analytic statistical tests. Efron and Tibshirani (1993) present an approach to estimating power, given a specific sample size. The approach offered by Sheets and Mitchell (2001) is to use Monte Carlo methods to estimate the rates of Type II error under several plausible alternatives to the null hypothesis. Despite the attendant difficulty in estimating the statistical power of different tests, computer-based tests seem to have at least as much statistical power as the more familiar analytical tests.

How many repetitions?

Regardless of the method used, the researcher is always faced with the question of how many replications or repetitions should be made. We want a small bias and standard deviation, but it is not clear how many replications are required to achieve this end. The number of independent bootstrap samples that one may form out of N specimens is $(2N-1)!/N!(N-1)$ (Efron and Tibshirani, 1993), which is over 90,000 for $N=10$ specimens. In most cases, even thousands of bootstrap replicates will not come close to exhausting all possible bootstrap sets. Typically, a modest subset of all possible sets is adequate for most statistical questions. Estimates of standard errors can usually be produced using only 100 or fewer bootstrap sets (Efron and Tibshirani, 1993), but reliable estimates of confidence intervals may require using many more. It does not appear that there is complete consensus on this issue (see Efron, 1992; Efron and Tibshirani, 1993; Jackson and Somers, 1989; Manly, 1997), but it does seem that more repetitions are necessary for estimating confidence intervals, where we must estimate a specific percentile point value, than either for hypothesis testing (see Manly, 1997) or for estimating of standard errors (Efron and Tibshirani, 1993). If computer time is not an issue, a range of 1000 to 2000 bootstrap tests is recommended for estimating a 95% confidence interval on a parameter (Efron, 1987; Efron and Tibshirani, 1993). When the time necessary to complete a calculation is a factor, one approach is to increase the sample size steadily until arriving at a value that is stable with respect to further increases in sample size. The stability criterion is perhaps most applicable to hypothesis testing, where we may not need to know the exact confidence level of the observed statistic – only that we can (or cannot) reject the null hypothesis at a 5% confidence level.

For example, if we run a bootstrap *t*-test and find that in 100 bootstrap tests the difference in means exceeds the observed difference 40 times (yielding $p = 0.40$), it is probably safe to state that we cannot reject the null at a 5% confidence level. A repetition of the bootstrap procedure might yield a slightly different confidence level, even changing by several percentage points, but it is highly unlikely to yield $p < 0.05$. Similarly, in such a bootstrap *t*-test, if the difference in bootstrap means never exceeds the observed difference in means (in 100 bootstrap sets), a single repetition of the bootstrap calculations at 100 bootstrap sets confirms that $p < 0.05$ appears to be reasonable. The difficulty arises when the bootstrap estimate of the *p*-value is very close to the desired confidence level ($p = 0.05$ in this example). In such a case, a large number of bootstrap sets may be warranted.

It is worth remembering that for $N_{Bootstrap}$ sets, the smallest confidence level we could possibly estimate is $1/N_{Bootstrap}$ – e.g. for 1000 bootstraps, the smallest confidence level we could ever hope to estimate is $1/1000 = 0.001$. The estimate of the confidence interval at 0.001, using 1000 bootstrap sets, is essentially based on the value obtained from a single bootstrap set (the one producing the largest or smallest value out of the 1000 sets examined). This suggests that it would be more appropriate to use 10,000 to 20,000 sets to obtain an estimate of the confidence interval at 0.001, so that the estimate is based on the results of 10 to 20 bootstrap sets (the 10 or 20 most extreme values out of the 10,000 or 20,000 total sets). In most cases it is not necessary to estimate confidence intervals at 0.1% (0.001); 5% confidence intervals are the standard, and are achievable with lower numbers of bootstraps.

When in doubt about the number of bootstrap sets that should be used to establish a particular confidence interval, the safest approach is to repeat the analysis after doubling the number of bootstrap sets (to determine whether that doubling alters the confidence level). This doubling should be repeated until the estimate stabilizes; the iterative approach may be time-consuming, but it is preferable to a blind reliance on a rule of thumb.

Summary

Computer-based statistics provide a useful alternative to the more familiar analytical statistical approaches, particularly when the observed distribution departs substantially from the assumptions of analytic models, or when no analytic estimate is available for the confidence interval of a specific statistic needed for the analysis. The performance of computer-based methods appears to be equal to that of analytic methods, although the greater flexibility of computer-based methods comes at the cost of increased computational time (and the need to produce specialized software for specific tests).

References

Edgington, E. S. (1995). *Randomization Tests*. Marcel Dekker.

Efron, B. (1979). Computers and the theory of statistics, thinking the unthinkable. *Society for Industrial and Applied Mathematics Review*, **21**, 460–480.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, **82**, 171–185.

Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society Series B, Methodological*, **54**, 83–127.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Good, P. (1994). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag.

Hoeffding, W. (1952). The large-sample power of tests based on permutation of observations. *Annals of Mathematical Statistics*, **23**, 169–192.

Jackson, D. A. and Somers, K. M. (1989). Are probability estimates from the permutation models of Mantel's test stable? *Canadian Journal of Zoology*, **67**, 766–779.

Manly, B. F. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society B*, **11**, 18–44.

Raspé, R. E. 1785. Baron Münchhausen's narrative of his Marvelous Travels and Campaigns in Russia.

Robinson, J. (1973). Large-sample power of permutation tests for randomization models. *Annals of Statistics*, **1**, 291–296.

Romano, J. P. (1989). Bootstrap and randomization tests of some non-parametric hypotheses. *Annals of Statistics*, **17**, 141–159.

Sheets, H. D. and Mitchell, C. E. (2001). Why the null matters: statistical tests, random walks and evolution. *Genetica*, **112**, 105–125.

Sokal, R. R. and Rohlf, F. J. (1995). *Biometry: The Principals and Practice of Statistics in Biological Research*, 3rd edn. Freeman.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. (Abstract) *Annals Mathematical Statistics*, **29**, 614.

9

Multivariate analysis of variance

In Chapter 7 we discussed ordination methods, including one used to discriminate among groups defined *a priori*, canonical variates analysis. In Chapter 8 we discussed several computer-based statistical methods that use resampling techniques to test explicitly whether an observed difference among groups is statistically significant. In this chapter we present a second set of methods that take a somewhat different approach to testing whether the observed difference is statistically significant. The approach used here is to compare the observed value of a test statistic (e.g. the values of t for a difference between two sample means) with the probability distribution of expected values for that statistic under a particular theoretical model for the distribution of variation in a population. This analytic approach is quite flexible, so the range of questions that can be addressed by using it is broad. For example, an investigator might want to know whether males and females differ in height. This is the simplest possible kind of question about differences between groups, because there is just one continuous variable (height), which is a simple one-dimensional trait (i.e. it is a scalar), and there is just one categorical variable (sex) with only two classes (male and female). Often the question is more complex, as when the investigator is comparing shape differences among several species. This is more complex because the continuous variable (shape) is multivariate (i.e. it is a vector) and the categorical variable (species) has more than two classes (one for each species in the study). The question can be made even more complex by considering multiple categorical variables (e.g. sexual dimorphism in several populations). Below we present the analytic tests for answering both simple questions and more complex ones. Much of this presentation follows expositions presented by Snedecor and Cochran (1967), Chatfield and Collins (1980), and Morrison (1990); readers requiring further details are referred to those works.

We begin this chapter with a brief review of groups and grouping variables. We then present the simplest case, the test for a difference in one trait between two groups, and the methods that would be used in such cases. We follow this with a series of more complex analyses, and the more generalized methods that would be applied to them. In the final section, we present instructions for performing the analyses discussed in this chapter.

Geometric Morphometrics for Biologists
ISBN 0–12–77846–08

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

Groups revisited

A group is a set of individuals (a class) defined as sharing a state of a discontinuous trait. In mammals and birds, "sex" is an example of a discontinuous trait that has two classes – "male" and "female." An individual is either one or the other as a consequence of having one set of chromosomes or the other. Such traits may be called grouping variables, qualitative traits or categorical variables. All these names refer to the fact that the states of the trait do not have intrinsic numerical values or an inherent order, but they can nonetheless be used to sort individuals into groups or categories.

Frequently, traits that could be quantified are treated as categorical variables. "Diet" and "locality" are examples of these kinds of traits. There are several reasons for treating these traits as categorical variables: first, the available information may not be sufficiently detailed to support a more finely graded analysis; second, the investigator may not want to impose a hypothesis of ordering on the data; or third, the investigator may not want to assume that all steps are of equal value. Under these circumstances, a quantifiable trait may be treated legitimately as a categorical variable. The only requirement is that the states of a particular variable are mutually exclusive – that is, each individual can belong to only one group.

Analytic techniques

One simple trait, two groups

We begin with a simple case – a test for a difference in jaw size between male and female adult squirrels collected at a single locality in western Michigan. Centroid size was computed using the landmarks shown in Figure 9.1, and the observed values of jaw size and their natural logs are given in Table 9.1. In this example, there is one continuous variable (centroid size of the jaw) and one categorical variable (sex) with two categories or classes (male and female). The question to be answered is whether jaw size differs between males and females.

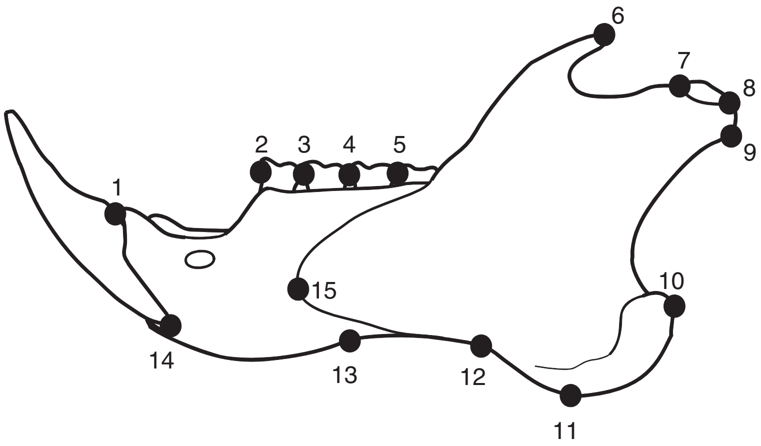


Figure 9.1 Outline of squirrel jaw, with landmarks.

The answer to this question can be obtained from a *t*-test. In this test, a variable known as Student's *t* (usually just '*t*') is computed as a function of the difference between the means of the two classes and variances around those means. The statistical model is the case in which two samples of equal size are drawn from the same normal distribution. Under this model, the difference between the sample means is expected to be zero and variance of the difference is a function of the population variance and size of the samples. Thus:

$$t = \frac{(<X_1> - <X_2>)}{\sqrt{\frac{2\sigma^2}{N}}} \qquad \textbf{(9.1)}$$

Table 9.1 Jaw size variation in 58 squirrels from Allegan County, Michigan

Sex	*Centroid size*	*ln centroid size*	*Sex*	*Centroid size*	*ln centroid size*
Female	53.0	3.97	Male	52.7	3.96
Female	51.8	3.95	Male	51.6	3.94
Female	51.5	3.94	Male	52.2	3.95
Female	48.6	3.88	Male	52.4	3.96
Female	50.7	3.93	Male	51.5	3.94
Female	51.4	3.94	Male	51.8	3.95
Female	52.0	3.95	Male	53.9	3.99
Female	50.3	3.92	Male	53.0	3.97
Female	51.7	3.95	Male	51.5	3.94
Female	52.2	3.96	Male	51.2	3.94
Female	50.6	3.92	Male	51.9	3.95
Female	51.8	3.95	Male	52.8	3.97
Female	51.3	3.94	Male	53.4	3.98
Female	52.7	3.96	Male	53.9	3.99
Female	50.6	3.92	Male	53.1	3.97
Female	52.6	3.96	Male	52.6	3.96
Female	51.1	3.93	Male	51.6	3.94
Female	50.4	3.92	Male	51.5	3.94
Female	51.0	3.93	Male	52.2	3.96
Female	51.4	3.94	Male	51.4	3.94
Female	52.0	3.95	Male	51.8	3.95
Female	52.0	3.95	Male	52.6	3.96
Female	50.4	3.92	Male	52.4	3.96
Female	51.9	3.95	Male	51.7	3.95
Female	53.0	3.97			
Female	52.9	3.97			
Female	51.0	3.93			
Female	52.4	3.96			
Female	52.4	3.96			
Female	51.8	3.95			
Female	53.0	3.97			
Female	51.7	3.95			
Female	52.8	3.97			
Female	51.5	3.94			

where $<X_1>$ is the expected value (mean) of group 1, $<X_2>$ is the expected value of group 2, σ^2 is the variance, and N is the sample size. Under the conditions of the model (a single population with a normal distribution), t has a known probability distribution, which approaches the normal distribution as N increases. Thus, the t-test evaluates the probability that two samples with means differing by the observed amount could be drawn by random sampling from a single population with the given variance.

In most studies (as in the squirrel jaw example) the two classes are represented by samples that have different variances and different numbers of individuals. This creates the problem of deciding what values to use to compute the standard error. The usual solution to this problem is to treat the two sample variances as estimates of one population variance, which is consistent with the model underlying the t-test. Accordingly, the sample variances (s_i^2) are weighted by their respective sample sizes (N_i) to compute a "pooled" estimate of the standard error, as shown in Equation 9.2:

$$t = \frac{(<X_1> - <X_2>)}{\sqrt{\left(\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2}\right)\left(\frac{N_1+N_2}{N_1N_2}\right)}} \tag{9.2}$$

This is just a generalized version of Equation 9.1. In the special case where variances are equal, the denominator of Equation 9.2 simplifies to:

$$\sqrt{s^2\left(\frac{N_1+N_2}{N_1N_2}\right)} \tag{9.3}$$

and when sample sizes are also equal, this simplifies further to:

$$\sqrt{s^2\left(\frac{2}{N}\right)} \tag{9.4}$$

as in Equation 9.1.

For the natural logs of jaw centroid sizes in Table 9.1, the 34 females have a mean of 3.94 and a variance of 3.9×10^{-4}, while the 24 males have a mean of 3.96 and a variance of 2.4×10^{-4}. Putting the sample sizes and variances into the denominator of Equation 9.2 yields a value of 0.0048. The difference between means is 0.02, so $t = 2.7$. The degrees of freedom are one less than the number of individuals, which is 57. With 57 degrees of freedom, the probability that t could be greater than or equal to 2.7 is 0.0091, which is usually considered statistically significant. Thus the difference between means is small, but the variances are even smaller, and so we can infer that males and females in this squirrel population do differ in jaw size.

The question of whether there is a significant difference between groups can also be answered without computing the difference between means. Instead, the variance explained by the categorical variable is compared to the variance it does not explain (which is the basis of the term *analysis of variance*, or ANOVA). The ratio of these two variances (explained divided by unexplained) is the test statistic F. Like t, F has a known probability distribution for pairs of samples drawn from the same normal distribution. Consequently,

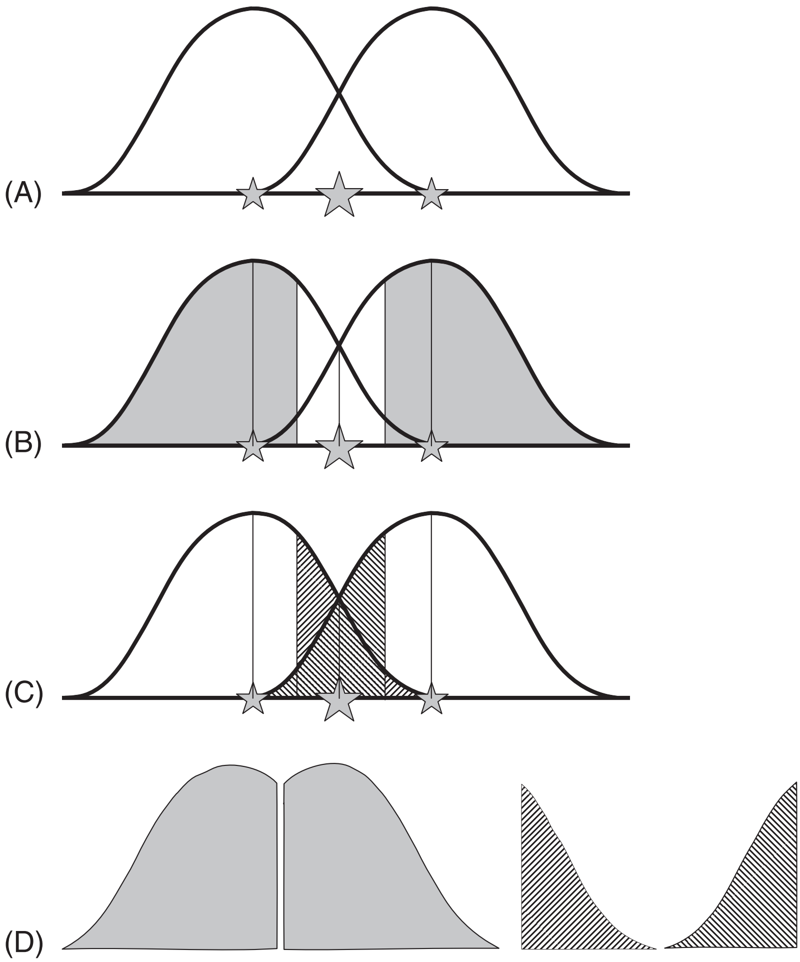


Figure 9.2 Graphic representation of ANOVA. Bell curves for two distributions are shown; the small stars are the group means of the respective distributions; the large star is the grand mean. (A) Both overlapping distributions are shown; (B) regions occupied by individuals closer to their group mean than the grand mean are shaded; (C) regions occupied by individuals closer to the grand mean than their group mean are indicated by diagonal hatching; (D) regions defined in B and C are shown side-by-side to compare their areas.

the probability reported for the *F*-test is the probability of an equal or larger *F*-ratio for two samples drawn randomly from the same distribution.

The logic underlying the *F*-test can also be explained graphically (Figure 9.2). We can represent the variation in each sample as the area under a curve (Figure 9.2A). In each group, some individuals are closer to their group mean than they are to the grand mean (Figure 9.2B); other individuals are closer to the grand mean than they are to their group mean (Figure 9.2C). The areas under the curves represented by these sets are shown in Figure 9.2D. If the means are far apart, a large proportion of individuals will be closer to their respective group means than to the grand mean, the value of *F* will be high, and the classes will be judged significantly different.

Computation of the F-ratio is complicated by the fact that the variance explained by the categorical variable cannot be calculated directly from the data but must be computed indirectly as the difference between the total variance and the variance that is *not* explained. If this seems strangely convoluted, look again at Figure 9.2. We can compute the deviations of each individual from the grand mean and use them to compute the total variance. We can also compute the deviations of each individual from its respective group mean and use them to compute the unexplained variance (the variance within the groups cannot be attributed to the factor responsible for the difference between the groups). Subtracting the unexplained variance from the total variance leaves the explained variance. For the natural logs of jaw size, the sum of squared deviations from the grand mean is 0.0207. The sums of squared deviations from the group means are 0.0055 for males and 0.0128 for females for a total within-groups sum of squares of 0.0183. The difference between the total and within-groups sums of squares is 0.0024; this is the between-groups sum of squares that can be used to compute the variance explained by the categorical variable, sex.

Variance is the sum of the squared deviations divided by the number of degrees of freedom. The total degrees of freedom are $N - 1$. The number of degrees of freedom attributed to the categorical variable is $G - 1$, where G is the number of groups or categories. (There are fewer degrees of freedom than classes, because an individual that does not belong to the first $G - 1$ groups necessarily belongs to the last one.) Subtracting the degrees of freedom allotted to the explained variance leaves $N - G$ degrees of freedom for the unexplained variance. Returning to our example, $N = 58$ and $G = 2$. The explained variance (due to sexual dimorphism) is $2.4 \times 10^{-3}/1$, and the unexplained variance is $0.0183/56 = 0.33 \times 10^{-3}$. The explained variance divided by the unexplained is 7.3. This F-ratio, with 1 and 57 degrees of freedom, has a p-value of 0.0091, which is identical to the p-value that was obtained from the t-test. Thus, despite taking different approaches, the F-test and t-test lead to the same result – the same conclusion regarding the significance of the difference between the two groups.

It is important to remember that both the t-test and the F-test assume that the variances within the groups are the same. Furthermore, both tests are asking whether samples as different as yours could have been drawn from a single sample with a specific known variance. Fortunately, both tests are fairly robust to violation of the assumption of equal variances.

One simple trait, more than two groups

Because ANOVA compares variation within groups to variation between groups, it can also be applied to analyses that examine more than two groups. For example, Table 9.2 illustrates an analysis of geographic variation in jaw size (the rows in the lower half of this table are not in the order usually reported, but in an order that corresponds more closely to the sequence of calculations). The categorical variable is geographic location, which has three classes referring to three collecting areas (eastern Michigan, western Michigan, and southern states). To test whether there is significant geographic variation in jaw size (to test whether there are significant differences among the three populations), we follow exactly the same procedure as for two groups. The sum of squared distances of individuals from the grand mean is the total sum of squares (SSQ), and the sum of squared distances of the individuals from their class means is the unexplained sum of squares. The difference

Table 9.2 ANOVA of jaw size with three groups

	Western Michigan	*Eastern Michigan*	*Southern States*	*All*
N	69	23	27	119
Mean	3.95	3.96	4.01	3.97
SSQ	0.034	0.017	0.019	0.131

	SSQ	*DF*	*MSQ*	F	p
Total	0.131	118			
Within-groups	0.070	116	<0.001		
Between-groups	0.061	2	0.031	>50	<0.0001

between these quantities is the between-groups sum of squares. The variances are the mean squares (MSQ), computed by dividing the sums of squares by the appropriate numbers of degrees of freedom (DF). As in the previous example, these values are the number of groups (localities) minus one, and the number of individuals minus the number of groups. Again, the value of *F* is the ratio of the explained and unexplained variances. In this particular example, the ratio is enormous (and imprecise due to rounding error) and the *p*-value is miniscule, indicating that there is a highly significant difference in jaw size among the three localities.

The conclusion that there is a difference among three or more classes does not imply that the mean of each class is significantly different from the mean of every other class. To determine whether that is the case, *t*-tests must be performed for all possible pair-wise comparisons. In the squirrel example, there are three pairs of localities, and the three tests indicate that the means of the two Michigan localities are not significantly different from each other but both are significantly different from the mean of the southern locality.

When performing multiple unplanned tests to determine which groups are different, it is important to remember that each additional test increases the chance that we might falsely reject the null hypothesis that the two groups are not different. If we perform a test and accept that the null hypothesis is rejected when the *p*-value is less than or equal to 0.05, we are also accepting the possibility that the test could be erroneous 5% of the time (based on the model of a normal distribution). Each time the test is repeated, we run the risk that the particular result will be erroneous, and increase the cumulative probability of at least one erroneous result. The way to correct this problem is to require a lower *p*-value before accepting that a test result supports rejection of the null hypothesis. One common adjustment is the table-wide Bonferroni adjustment, in which the desired *p*-value is divided by the number of tests. In the example above, we would require $p < 0.05/3$ for all tests.

Two or more categorical variables

The examples above have only a single categorical variable, but it is common to have multiple categorical variables representing independent classifications of the specimens in the data set. In the data analyzed in the last example, there are male and female squirrels in all three localities. This suggests that we should divide the variance in jaw shape into three components: variance explained by sex, variance explained by geography, and variance not explained by either sex or geography (Table 9.3). Two *F*-ratios are computed, one for each categorical variable. As before, the variance explained by the categorical variable is compared to the variance that is not explained, but now the unexplained variance is the variance that is not explained by *any* categorical variable. This is a smaller quantity than the variance that is not explained by that particular categorical variable.

The ANOVA can also be used to test whether sexual dimorphism differs among localities. In other words, we can test whether there is an interaction between the two categorical variables. This requires computing the variance explained by a third categorical variable in which the classes are defined by all possible combinations of the classes in the two original categorical variables. This test can only be used if we have specimens representing all of the possible classes in this new variable. If we had only males for one locality, we would not be able to test for an interaction across all three localities (although we would be able to test for an interaction over the two localities for which we do have both sexes). In the jaw size example (Table 9.4), the interaction between sex and location explains very little of the variation – albeit more than sex by itself. The lack of a significant interaction means that sexual dimorphism is not significantly different among the three locations. (Separate analyses on the three locations indicate that none of them exhibits significant sexual dimorphism in jaw size.)

Notice that as we add explanatory variables to the model, the unexplained SSQ gets progressively smaller as more of the total SSQ is attributed to the explanatory variables. In addition, the unexplained SSQ is attributed to fewer degrees of freedom. Both of these changes influence the *F*-ratios for the tests of specific hypotheses. In the squirrel jaw data, the differences among localities are so large and the differences due to the other variables so small that the decision to include additional variables did not alter conclusions. However,

Table 9.3 ANOVA of jaw size with two categorical variables

	SSQ	*DF*	*MSQ*	F	p
Locality	0.061	2	0.031	>50	<0.0001
Sex	<0.001	1	<0.001	<0.15	>0.69
Unexplained	0.070	115	<0.001		
Total	0.131	118			

Table 9.4 ANOVA of jaw size with two categorical variables and their interaction

	SSQ	*DF*	*MSQ*	F	p
Locality	0.060	2	0.031	>50	<0.0001
Sex	<0.001	1	<0.001	<0.03	>0.86
Locality × sex	<0.001	2	<0.001	<0.11	>0.89
Unexplained	0.070	113	<0.001		
Total	0.131	118			

if the effects of sex or locality were marginal (if the *F*-ratios were close to the cut-off point for an α of 0.05), then the conclusions could have been altered by the inclusion of the interaction effect in the analysis. For this reason, all explanatory variables (including all of their interaction terms) should be included in an ANOVA, and the explanatory variables should be tested simultaneously, not one by one. If your data do not include specimens for every possible combination of states of the categorical variables (e.g. you do not have both sexes from every location), you should reduce the data set so that every possible combination of the remaining states is included (e.g. include only the locations for which you have both sexes). This will at least tell you whether there are interactions in that subset of the data.

A categorical variable and a continuous variable

In some cases, one of the explanatory variables may be continuous rather than discontinuous. For example, we might anticipate that differences in shape between sexes are partly due to differences in size between the sexes. Thus, we would want to account for the differences in shape caused by differences in size so we can test for shape differences independent of size differences. The continuous explanatory variable in this type of analysis is called a "covariate." The variance due to it is explained by the regression of the dependent variable on the covariate. (Regression and related analyses, including alternative methods of accounting for variation explained by the covariate, are discussed in Chapter 10.) Analyses of variance that include a covariate are called ANCOVA, or MANCOVA in the multivariate case. Briefly, these methods use regression to control for the covariate. This is done by estimating the expected value for the dependent variable(s) at a given value of the covariate (usually, the *Y*-intercept). The value of the covariate does not matter when the different groups have the same slope of the dependent variable on the covariate, because their regression lines are then either parallel or coincident. If they are parallel, the groups differ in shape even after adjusting for the covariate and they will differ by the same amount and in the same direction over all possible values of the covariate. If the slopes are coincident, meaning the lines are actually the same, the difference between groups will be zero for all values of the covariate. Thus, the first step of the analysis is to test for a significant interaction between the covariate and the categorical variable. In Chapter 10 we present a method for comparing shapes across groups when the interaction term is significant.

In Table 9.5 we show results for an analysis to determine whether there are differences in jaw length among localities after accounting for variation in jaw length explained by variation in jaw size (the covariate). This result shows a significant effect of locality on jaw length, so there are differences in jaw length among localities that are independent of the differences in jaw length that are associated with differences in jaw size. (The independence

Table 9.5 ANOVA of jaw length with jaw size as a covariate

	SSQ	*DF*	*MSQ*	F	p
Locality	0.003	2	0.002	>24	<0.0001
Jaw size	0.052	1	0.052	>740	<0.0001
Unexplained	0.008	115	<0.0001		

of length and size suggests differences in shape; in the next section we present a more direct test for shape differences.) We also find that most of the variation in jaw length is explained by the regression on jaw size, not by differences among localities (indicated by difference in mean squares).

A complex trait and a categorical variable

So far in this chapter, we have discussed tests in which the dependent variable is a simple, one-dimensional, continuous trait such as size. In such cases, there is a single total variance to parse into its explained and unexplained components. That total was divided into more explained components when more categorical variables were added, but the test still only evaluated the relative magnitudes of the explained and unexplained components. In contrast, shape is a single, complex trait described by several continuous components. To parse the variance of this multidimensional trait, we use the same technique that would be used to parse the variances and covariances of multiple, separately measured traits, namely multivariate analysis of variance (MANOVA).

In the simplest MANOVA, we have a multivariate dependent variable and just one categorical variable with two classes. Our question is whether there is a difference between classes in the dependent variable, so we want to perform a multivariate equivalent of the *t*-test. In other words, we want to evaluate the difference between the two means on all measured variables simultaneously. The multivariate generalization of the *t*-test is Hotellings T^2. T^2 can be derived from the univariate *t*, using the formula introduced earlier for the case in which samples have equal variances but different numbers of individuals (from Equations 9.2 and 9.3):

$$t = \frac{(<X_1> - <X_2>)}{\sqrt{s^2\left(\dfrac{N_1+N_2}{N_1N_2}\right)}} \tag{9.5}$$

Squaring this expression produces:

$$t^2 = \frac{(<X_1> - <X_2>)^2}{s^2\left(\dfrac{N_1+N_2}{N_1N_2}\right)} \tag{9.6}$$

which can be rearranged to:

$$t^2 = \left(\frac{N_1N_2}{N_1+N_2}\right)\left(\frac{(<X_1> - <X_2>)^2}{s^2}\right) \tag{9.7}$$

Now, we replace the univariate difference between means $<X_1> - <X_2>$ with the vector of mean differences on all variables $(<\mathbf{X_1}> - <\mathbf{X_2}>)^{\mathrm{T}}$, and we replace the pooled within-group variance s^2 with the pooled within-group variance–covariance matrix $\mathbf{S_w}$. This yields:

$$T^2 = \left(\frac{N_1N_2}{N_1+N_2}\right)(<X_1> - <X_2>)^{\mathrm{T}}\mathbf{S}_{\mathbf{w}}^{-1}(<X_1> - <X_2>) \tag{9.8}$$

which is distributed approximately as an F-distribution. The degrees of freedom are given by the number of variables (V) and $N_1 + N_2 - 1 - V$.

When we have multiple groups, each described by a multivariate dependent variable, we need a multivariate generalization of the F-test used in ANOVA. Recall that the univariate F-test is a function of the variances within and between groups. Accordingly, the multivariate F-test should also be a function of the within-groups and between-groups variance–covariance matrices (**W** and **B**). Although this implies that the test statistic for the multivariate F-test should be a function of the eigenvalues of **W** and **B**, it is not clear what that function should be. One simple solution is the Hotelling–Lawley trace, which is the trace of $\mathbf{BW}^{-1}$ (the trace of a matrix is the sum of its eigenvalues). Several alternatives to the Hotelling–Lawley trace have been proposed; all can be equated with functions of the eigenvalues of $\mathbf{BW}^{-1}$. Each of these test statistics can be converted to a value that is distributed approximately as an F-distribution, making it possible to determine the p-value for the hypothesis that the samples were drawn from the same multivariate normal distribution. For example, a commonly used statistic is Wilks' Λ, formally defined as the determinant of $\mathbf{W}(\mathbf{B}+\mathbf{W})^{-1}$ (i.e. the product of the eigenvalues of that matrix), which is equivalent to the product $\prod \frac{1}{1+\theta_i}$ where θ_i are the non-zero eigenvalues of $\mathbf{BW}^{-1}$. Wilks' Λ can be converted to functions that approximate either the χ^2 ($-\ln \Lambda$ weighted by a function of the degrees of freedom) or F distribution ($W \frac{1-\Lambda^{1/H}}{\Lambda^{1/H}}$, where H is 1 less than the number of groups and W is a function of the degrees of freedom). Chatfield and Collins (1980) cite several studies that compare the performance of these and other tests, and conclude that the comparisons are indecisive. They also note that the two tests mentioned above and a third test, Pillai's trace (trace of $\mathbf{B}(\mathbf{B}+\mathbf{W})^{-1}$), are asymptotically equivalent, meaning that they approach the same value at large sample sizes and differ little in power at small sample sizes. Furthermore, the three tests are exactly equivalent when there is only one independent variable.

The total number of degrees of freedom differs slightly among the three test statistics, but in all cases it is approximately the product of the number of groups (G) and the total number of individuals in all groups (N). In all three tests, the number of degrees of freedom for the between-groups variance is $V(G-1)$ where V is the number of variables and G is the number of groups. The number of degrees of freedom for the within-groups variance is the total degrees of freedom minus the between-groups degrees of freedom. For this difference (approximately $NG - VG$) to be greater than zero, the number of individuals must be greater than the number of variables. This is consistent with the algebraic requirement that the number of equations must be greater than the number of variables in order to have more equations than there are variables in those equations, so N must be greater than V.

Although MANOVA can be used to test for shape differences among groups, there are constraints on the kind of shape data that can be used. One constraint is due to the fact that MANOVA assumes that the measurement space is Euclidean (as discussed in Chapter 4, shape space is not Euclidean). Another constraint is due to the fact that the number of shape variables is smaller than the number of variable coordinates produced by most methods of superimposition (as described in Chapter 5). By specifying the "variable coordinates" we mean to exclude the fixed baseline endpoints produced by the two-point

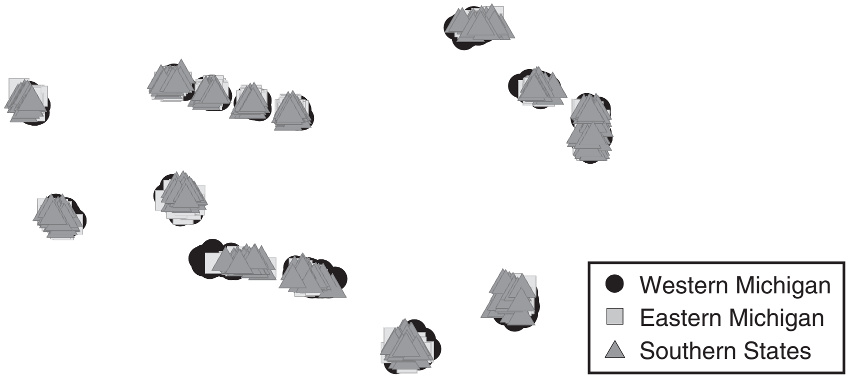


Figure 9.3 Superimposed landmarks of 119 squirrel jaws from three localities, eastern and western Michigan, and southern states.

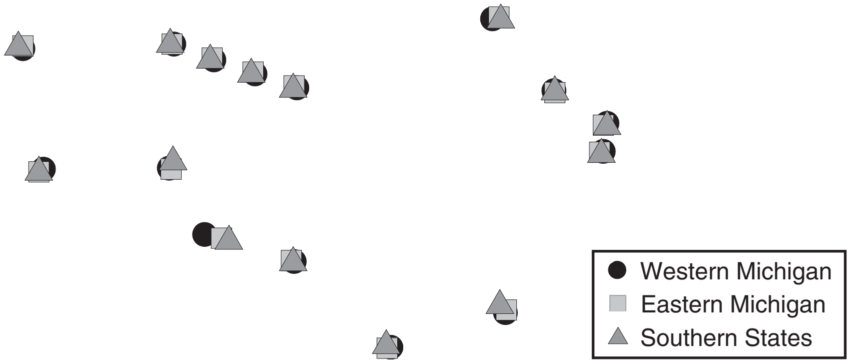


Figure 9.4 Superimposed landmarks of the mean jaw shapes for the three geographic samples shown in Figure 9.3.

registration (Chapter 2). For landmarks taken on two-dimensional images, the number of shape variables is $2K - 4$, where K is the number of landmarks. If the number of variable coordinates is greater than the number of shape variables, then the number of variable coordinates overestimates the true degrees of freedom.

Fortunately, we have two ways to project shapes in shape space onto a Euclidean tangent space, thereby converting shape information to a form that satisfies these two assumptions of MANOVA. One option is to compute Bookstein shape coordinates, which produces the same number of variable coordinates as there are dimensions in the shape space. The other option is to use partial warp scores for configurations obtained by Procrustes (GLS) superimposition. The number of partial warps (including the uniform components) is the same as the number of dimensions in the shape space.

Figure 9.3 shows the Procrustes superimposed landmarks for the squirrel jaws. At each landmark the distributions of the position of that landmark in the three geographic samples overlap broadly, suggesting there is little if any difference in jaw shape among localities. Comparison to the picture of the three mean shapes (Figure 9.4) suggests that the differences among groups are small relative to the variability within each group. Table 9.6 lists results

Table 9.6 MANOVA of jaw shape among localities using thin-plate spline coefficients

Test statistic	*Value*	F	*DF*	p
Wilks' Λ	0.136	5.98	52, 182	<0.000001
Pillai trace	1.25	5.87	52, 184	<0.000001
Hotelling–Lawley trace	3.54	6.08	52, 180	<0.000001

Table 9.7 Goodall's *F* test for shape differences among locations

Localities	*PPd*	F	*DF*	p
W and E	0.0265	9.83	26, 2340	<0.000001
W and S	0.0384	22.2	26, 2444	<0.000001
E and S	0.0204	3.76	26, 1248	<0.000001
All three	n.a.	13.5	52, 3016	<0.000001

W = western Michigan, E = eastern Michigan, S = southern states. PPd is the partial Procrustes distance between pairs of mean shapes, which does not apply to the three-sample case.

of three tests for differences in jaw shape among the three squirrel populations. The data are the scores of the thin-plate spline components, with the mean of all individuals (computed by Procrustes superimposition) as the reference shape. The *F*-values and degrees of freedom differ slightly, reflecting differences in the computations performed for each test; when *p*-values are close to the preferred α-level, these differences can lead to different conclusions.

An alternative to performing a MANOVA on shape variables in the tangent space is to perform the analysis in shape space, measuring deviations from means as sums of squared Procrustes distances. The test statistic for this analysis is Goodall's *F*, which is the ratio of explained and unexplained variation in those distances. As in conventional MANOVA, the degrees of freedom for the explained variance are given by $V(G-1)$, where V is the number of dimensions (variables) in shape space and G is the number of groups. However, the total number of degrees of freedom, $V(N-1)$, and the within-groups degrees of freedom, $V(N-G)$, are much higher than in a conventional MANOVA. This difference is due to the fact that we are computing Procrustes distances for N individuals in V dimensions, but do not need to estimate all the variances and covariances of the shape variables.

Table 9.7 shows the *F*-values and corresponding *p*-values obtained for all three pair-wise comparisons of the three populations, and for a simultaneous analysis of all three groups. The partial Procrustes distances between the means are also shown for the two-group comparisons. Because the dimensionality of shape space is a factor in the total number of degrees of freedom, the *p*-values can be quite small even when the categorical variable explains very little of the shape variation in the data set. As might be surmised from Figures 9.3 and 9.4, differences among the three populations explain only a small fraction of the variation in the Procrustes distances. Those differences can be judged significant because a large amount of information (resulting in large numbers of degrees of freedom) was used to estimate what those differences are.

Other uses of the *t*-test

Earlier in this chapter we discussed a typical use of the *t*-test, namely to determine whether the difference between two sample means is statistically significant. The *t*-test can also be used to evaluate the uncertainty of a derived quantity – that is, a quantity that is not computed for each specimen but is computed after analysis of several specimens. In geometric morphometric studies, we often want to know whether the partial Procrustes distance between one pair of mean shapes is significantly different from the distance between another pair of mean shapes. For example, the distance between mean shapes of squirrel jaws in the western Michigan and southern samples appears to be much greater than the distance between mean shapes of the eastern Michigan and southern samples (0.0384 vs 0.0204). The partial Procrustes distances computed from the original data sets are the expected values $<X_1>$ and $<X_2>$ in the formula for t (Equation 9.2, reproduced here):

$$t = \frac{(<X_1> - <X_2>)}{\sqrt{\left(\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2}\right)\left(\frac{N_1+N_2}{N_1N_2}\right)}} \tag{9.2}$$

The sample sizes N_1 and N_2 are the combined sizes of the paired data sets $(69+27)$ and $(23+27)$ corresponding to the distances $<X_1>$ and $<X_2>$. Although these replacements are straightforward, the values of s_1 and s_2 are less clear. Given only the original data sets, we have only one estimate for each of the distances; they are single observations, not means of multiple observations. Without multiple observations, we have no direct way to estimate the uncertainty of the expected value. One solution to this dilemma is to use bootstrap resampling (Chapter 8) to estimate the standard error of each distance (another solution is to bootstrap the difference between the distances – see the next section). Because $SE^2 = s^2/N$, we must substitute $N \cdot SE^2$ into Equation 9.2, producing:

$$t = \frac{(<X_1> - <X_2>)}{\sqrt{\left(\frac{(N_1-1)N_1SE_1^2 + (N_2-1)N_2SE_2^2}{N_1+N_2-2}\right)\left(\frac{N_1+N_2}{N_1N_2}\right)}} \tag{9.9}$$

Substituting the distances (0.0384 and 0.0204), standard errors (0.0030 and 0.0032) and sample sizes (96 and 50) into this equation yields $t = 3.60$ with 144 degrees of freedom for $p < 0.0005$. Thus, the difference between the two distances is statistically significant.

Whenever using the *t*-test, whether comparing means of observed variables or computed values of derived variables, it is important to remember that the test assumes that deviations within the groups are normally distributed. Fortunately, the test is fairly robust to violations of this assumption, but care should be taken when the *p*-value is close to 0.05 (or whatever α is chosen as the criterion for statistical significance).

Resampling-based tests

All of the statistical tests discussed so far in this chapter make assumptions about the distribution of variation around the means that are being compared. As discussed in Chapter 8, deviations of the populations from the model can lead to erroneous conclusions. Resampling procedures can be used, allowing biological samples to deviate from ideal theoretical distributions, but implementing a test using these procedures often requires the user to write the program for that specific test. Below, we briefly demonstrate two readily available bootstrap tests.

In the first example, we use the bootstrap to determine whether jaw shape differs between the two samples of squirrels from Michigan. The null hypothesis is that the observed difference could arise by chance when sampling from a single population, so all 96 specimens were combined into a single pool. In each iteration, two bootstrap sets of the original sample sizes (69 and 27) were drawn with replacement from the pool and the *F*-ratio was computed for that pair. After the chosen number of iterations was completed (400 in this case), the bootstrap sets with an *F* at least as large as the original set (9.83) were counted. This number divided by the number of iterations is the probability of obtaining the original samples under the null hypothesis. In this case, only the original samples had an *F* as large as 9.83, so the *p*-value is 1/400 (0.25%).

In the second example, we use the bootstrap resampling procedure to evaluate the uncertainty of a derived quantity: the partial Procrustes distance between the mean shapes of the two samples. Here, the question is about the uncertainty of the distance between sample means, which is a function of the sampling of each source population. To simulate this by bootstrapping, we keep the samples separate, and in each iteration draw separate bootstrap sets of each sample and compute the distance between the means of the bootstrap sets. This set of distances is used to estimate the 95% range interval around the distance between the population means. One use of this range interval is testing whether the distance between one pair of samples differs from the distance between another pair of samples. In the analysis of squirrel jaws, the partial Procrustes distances between mean shapes suggest that the southern sample is much farther from the western Michigan sample than it is from the eastern Michigan sample (0.0384 vs 0.0204). This is supported by the 95% range intervals, which do not overlap (0.0338 – 0.0456 vs 0.0172 – 0.0295).

We can also answer the question about the difference between the two distances by bootstrapping that difference (now the *difference* is the derived trait rather than one of the distances). In this case, we draw bootstrap sets of all four samples in each iteration, compute the two distances between pairs of means and the difference between those distances. These results are used to determine the 95% range of the difference between distances over the series of iterations. For the squirrel jaws, the distance between mean shapes of the western Michigan and southern samples (0.0384) is greater than the distance between mean shapes of the eastern Michigan and southern samples (0.0204). The difference between the distances is 0.0184. After 400 bootstrap iterations, the 95% range of the difference is estimated to be 0.0070 – 0.0246. This range does not include zero, so we can infer that the observed difference between distances is statistically significant.

Software

Two programs in the IMP series are available for performing the statistical analyses discussed in this chapter, **CVAGen** and **TwoGroup**. Both perform simple analyses; more complex ones will require a commercial statistical package, or **TPSRegress** (at the end of this section we provide general instructions for using commercial packages to conduct these analyses).

MANOVA of shape using CVAGen

As discussed in Chapter 7, **CVAGen** can be used to describe differences among groups. The output for **CVAGen** includes a test of the hypothesis that differences among the groups are statistically significant (for detailed instructions on running this program, look at the discussion of **CVAGen** in Chapter 7). Here we discuss only the results of a statistical test reported in the **Auxiliary Results** box window. These are the results of Bartlett's test for the number of informative CVs. Bartlett's test is based on a MANOVA testing the hypothesis that there are differences among the groups; the MANOVA is repeated with progressively fewer CVs to determine how many of them are informative. Accordingly, each row in the window shows the value of Wilk's Λ, the corresponding χ^2, the degrees of freedom and the p-value for the test that there are differences among groups in a progressively smaller subset of the data. The first iteration, using all of the CVs, is the one that is relevant here; it is the test for differences among groups using all of the available data. The null hypothesis is that there are no differences among the groups; that the differences among the samples are no greater than would be expected if they had all been drawn from the same multivariate normal population. Remember, rejection of this null does not mean that each group is different from every other. The plots of CVA scores and the classification table (produced by selecting **Show Grouping By CVA** in the **Statistics** menu) can suggest reasonable hypotheses of which groups differ, but these are not definitive tests.

It is possible to use **CVAGen** explicitly to test whether two particular groups are significantly different by including only those two groups in the data file. Then there is only one CV (the axis maximally discriminating between the two groups). If they can be differentiated on this axis, then they are significantly different. Doing this requires you to construct a separate file for each pair of groups, so we recommend using **TwoGroup,** which requires fewer files to perform the same set of tests.

Running TwoGroup

Each group must be in a separate file, in standard (X1,Y1,...CS) format. The files are loaded separately by clicking **Load Data Set 1**, finding the file, then clicking **Load Data Set 2** and finding that file. As usual, you can display the data in various superimpositions, by clicking on your choice in the **Show Data** field below the visualization window. Be sure to select the correct baseline points before choosing plots or analyses that use a baseline superimposition (i.e. BC or SBR). After the files are loaded, all of the test options are active, so be sure that you have selected the right number of bootstraps before clicking one of these buttons.

To test the significance of the difference between samples using Bookstein coordinates, choose **Hotelling's T^2 (BC)**; this is the only test available for these coordinates. The results window will report values for F, the degrees of freedom, and p. The results will also include the distance between means. This distance is the sum of the squared distances between corresponding landmarks, but it is *not* the minimized Procrustes distance because the landmark configurations are not in Procrustes (GLS) superimposition.

To test the significance of the difference between samples using the partial Procrustes distances, choose **Goodall's F (Procrustes)**. For this test, the coordinates are superimposed using GLS with the specimens rescaled to unit centroid size (see Chapter 5). Again, values of F, the degrees of freedom, and p will appear in the results window. The distance between means is also reported, which is the partial Procrustes distance.

Next to the buttons for the analytic tests are buttons for two tests that use a bootstrap resampling procedure (**F-test, SBR** and **F-test, Procrustes**). The **F-test, SBR** is a resampling-based F-test for coordinates in the SBR superimposition; **F-test, Procrustes** is a resampling-based version of Goodall's F-test. Before choosing either option, select the right number of iterations in the **No. of Bootstraps** box on the far left. The results, which will appear in the results window, will include the F-value computed for the original data set. After this is a "Significance level:.." which is the fraction of iterations (in decimal format) in which F is greater than or equal to the value reported for the original data. The output also includes a distance between the means of the original data sets. Again, if you selected the Procrustes test, this is the minimized partial Procrustes distance; if you selected SBR, it is not the minimized distance because the specimens are not in the partial Procrustes superimposition.

The three buttons in the box labeled **Bootstrapped Distances Between Means** invoke analyses in which bootstrap resampling is used to estimate the standard error and 95% range of estimates for the distance between the group means under the indicated superimposition. The observed distance and the bootstrapped standard error of that difference can be used to test whether the distance between one pair of samples is different from the distance between another pair of samples (using **TBox**, described below).

Like the standard error, the 95% range is a measure of the uncertainty of the observed distance between the two means. However, this range should *not* be used to test the hypothesis that this distance is significantly greater than zero. A distance *cannot* be less than zero. Even if the groups have identical means, it is unlikely that bootstrap sets will be drawn in which the difference is exactly zero. You can demonstrate this by loading the same file twice (i.e. using the same file as data sets 1 and 2); the lower end of the confidence interval will be a small number, but still greater than zero.

The value of the 95% range is that it can be used to evaluate whether the distance between one pair of groups is different from the distance between a second pair of groups. If the ranges for the distances do not overlap, then the difference between the distances is statistically significant. The limitation of this test is that it may be too stringent: the probability that both distances are more similar than the adjacent ends of the ranges is considerably less than 0.05 (in fact, it is less than 0.05^2 if the normal model applies). If you have ranges that overlap, you may want to consider using the standard errors in an analytic test (in **TBox**), as mentioned before.

Another option for comparing the difference between two distances is to bootstrap that difference. Use **Load Data Set 1** and **Load Data Set 2** to load the first pair of samples, then go to the **File** pull-down menu and use **Load Group 3** and **Load Group 4** to load the second pair of samples. Now go to the **More Stats** pull-down menu and select **Bootstrap Distances 1+2 vs 3+4**. When the iterations are completed, the results window will show the partial Procrustes distance between means 1 and 2, the 95% range and the standard error. Scrolling through the results will reveal the same information for the distance between means 3 and 4. At the end, you will come to the 95% range for the difference between the distances. If this range includes zero, then the distances are not significantly different; sometimes one distance is larger, sometimes the other distance is larger.

In addition to the statistical tests, **TwoGroup** can plot the superimposed landmarks and the superimposed means (but only for the data sets 1 and 2). These plots can be modified using the **Symbols Control** pull-down menu, which allows you to change the red and blue symbols to black or gray, fill the symbols, and increase their size. The plots of the differences between means can be edited using the options located on the **Difference Plot Options** pull-down menu. As usual, you can select from a variety of superimposition methods and types of displays, trim the grid and rotate the reference.

Using TBox

This program can be used to perform *t*-tests on means of directly measured traits or on expected values of derived quantities like the Procrustes distances between groups. Type the *values* you want to compare into the boxes labeled **Mean of Group 1** and **Mean of Group 2**. These values could be the means of a univariate trait, like centroid size. In that case, simply type the means, sample sizes and standard errors into the appropriate boxes. The group with the larger mean should be entered as group 1. When you click the big green **Run Calculation** button, the program will compute the *t*-value for the difference between the means and determine the corresponding *p*-value. If the values are Procrustes distances between samples, enter the distance between the first pair of samples as the **Mean of Group 1** and the distance between the second pair of samples as the **Mean of Group 2**. The sample size for the group is the sum of the two sample sizes. Standard errors for the distances can be obtained from the resampling tests in **TwoGroup**, as described above.

Note the caveats for **TBox**. You must have standard errors for the quantities you are comparing; if you have variances or standard deviations, you must convert them to standard errors. The more important caveat is the assumption of normality. **TBox** reports the value of p for t, when t is normally distributed, which is only expected when sample sizes are large (>60). If your sample sizes are smaller, you may prefer to use the more conservative *p*-values reported in *t*-tables of more conventional statistics texts and programs. However, even these *p*-values are computed under the assumption that deviations within samples are normally distributed. If even this assumption of normality is doubtful, you should probably use a resampling-based test.

Conducting ANOVAs/MANOVAs using other programs

Simple ANOVAs, like the test for differences in centroid size between two samples, can often be performed using a hand-held calculator or a spreadsheet program. The hard part is sorting the data correctly and keeping track of the number of entries. If you use a spreadsheet program, be sure you choose the correct ANOVA or *t*-test options for equal or unequal sample sizes, and for equal or unequal variances.

Complex MANOVAs, analyses in which there is more than one categorical variable (locality and sex), or analyses in which there is a categorical variable and a covariate (sex and size), usually must be performed in a computer program package specifically designed for multivariate analysis. Although these analyses are not algebraically difficult, they can be computationally intensive. Below we present some general guidelines for performing a MANOVA on shape variables in commercial statistical packages.

The first step is figuring out how to get your data into the program you intend to use to analyze the data. To do this correctly, you will need to understand how the analytic program expects the data to be formatted. One part of this is determining whether the analytic program can accept data from your database, spreadsheet or text file; another part is determining whether the program requires particular symbols to delimit fields (e.g. space, tab, or comma), or types of variables (e.g. "$" as the last character in the name of a categorical variable). You should also determine whether it will be easier to add the categorical variables to the data before or after they are read into the analytic program; this will be a function of how easy it is to edit the data file after it has been read.

The next step is deciding whether you want to analyze the shape variables from the thin-plate spline decomposition (scores on the partial warps and uniform components) or the coordinates of the landmarks. If you decide to use landmark coordinates, you will need to use **CoordGen** to compute the superimposition *before* you import the data. Whichever you choose, make sure you import *all* of the shape data. For spline components, this includes all of the partial warps and uniform components. (IMP programs output partial warps in order of *increasing* spatial scale, followed by the scores for the uniform component, with centroid size or ln(centroid size) in the last column.) If you use landmark coordinates registered to a baseline (Bookstein shape coordinates, or sliding baseline registration), remember to omit the invariant coordinates of the baseline points. If you do not omit them from the input file, you will have to remember to omit them when you select the variables to be included in the analysis. If you do not use Bookstein shape coordinates or the scores on the spline components, you must remember that the correct number of degrees of freedom is less than the number of variables (−4 if Procrustes superimposition, −2 if sliding baseline registration). You must also remember that Procrustes superimposition and sliding baseline registration do not project the specimens onto the same space as the thin-plate spline or the Bookstein shape coordinates (see Chapter 5). Under many circumstances, these choices will not alter conclusions about the significance of differences between groups. One situation in which the choice can influence conclusions is when the range of shape variation across all groups is so large that the difference between the shape space and the tangent space become noticeable (you can check whether that is the case for your data using **TPSsmall**). The other situation in which the choice can influence conclusions is when the differences among groups are so small relative to the variance within groups that a small difference in the number of degrees of freedom shifts the p-value of the test statistic across the preferred α-level. In such cases, it is better to be cautious about claiming significance.

After you have resolved all of the issues relating to formatting and entering data into your program, the next step is navigating through the program to select the right analysis. If your program does not have a giant button labeled MANOVA, look for a menu item referring to "linear models" or "linear hypotheses;" MANOVA will be one of the options within that category. (You could also try searching the help menu or the index of the manual.) After you have started the MANOVA module, you will probably be asked to select variables. Again, remember to include all of the relevant variables in the list of dependent variables (all of the spline components or all of the variable landmark coordinates). The independent variable will be the categorical variable or covariate hypothesized to explain shape variation. Your program may also require an extra step to indicate that the explanatory variable is a categorical variable. Add the interactions between the explanatory variables if they are not automatically included in the model. As discussed earlier in this chapter (see discussion of simple and complex ANOVAs, Tables 9.2–9.4), exclusion of the interaction terms from the analysis is not advised because it alters all of the sums of squares, which may influence conclusions if any of the explanatory variables has a marginal effect.

Now punch the "go" button, sit back, and wait for the output to stop scrolling (if you have a lot of landmarks and more than a couple of explanatory variables, it may take a while for the program to work through all of the tests). If your program generates a series of univariate tests for each landmark coordinate or spline component, ignore them. As discussed in Chapters 4 and 6, these variables are not independent in the sense that is relevant to these tests. If your results include a test for the MANOVA constant, this result can also usually be ignored. However, this component of the test cannot be excluded; it is analogous to evaluating whether a regression has a non-zero intercept. Excluding the constant is equivalent to forcing a regression through the origin. In the case of regression, this can affect the estimation of the slope, conclusions regarding the deviation of the slope from zero, and inferences about the proportion of variation explained by the regression. For similar reasons, excluding the constant from the MANOVA is not recommended under most circumstances. (Effects of constants are also computed in ANOVA, but are often not included in the output.) Eventually, scrolling through the results will reach the multivariate results for each categorical variable and any interactions. As mentioned earlier, there are several possible test statistics that could be reported. The differences among them only matter when the effects are marginal. Each test statistic should have a corresponding F or χ^2, degrees of freedom, and p-value. These are the numbers on which you will base conclusions regarding the significance of effects.

References

Chatfield, C. and Collins, A. J. (1980). *Introduction to Multivariate Analysis*. Chapman and Hall.

Morrison, D. F. (1990). *Multivariate Statistical Methods*, 3rd edn. McGraw Hill.

Snedecor, G. W. and Cochran, W. G. (1967). *Statistical Methods*, 6th edn. Iowa State University Press.

10

Regression

Chapter 9 covered methods for testing hypotheses about samples that differ categorically. This chapter covers methods for testing hypotheses about samples that vary along a continuously valued factor – a factor measured on an infinitely divisible scale. Size is an example of such a continuously valued factor because there is always a size between any two others; similarly, latitude is continuously valued because there is a latitude between any two others. When we hypothesize that a continuously valued factor affects shape, we use regression to test the hypothesis. Additionally, when we want to control for the effects of such a factor so that we can distinguish between groups defined by a categorical variable, we use regression to control for those effects. Finally, we would use regression when our hypotheses concern the particular nature of an effect, i.e. the direction of the shape variable covarying with the factor of interest. For example, if our hypothesis is that two species follow a common ontogeny of shape, we use regression to describe each ontogeny, then we compare the two vectors, asking if they point in the same direction.

The chief aim of regression is to explain the variation in one variable (shape, in our case) by another. For example, we might suspect that several factors account for the variation in our data, including: age or size; geographic variables such as latitude, longitude or temperature; ecological variables such as the size of predators and the density of the canopy; or even clinically important characteristics such as health status. So long as the candidate factor is measured, and measured along a continuously valued scale, we can test the hypothesis that it affects shape. The strategy for testing the hypothesis is simple and straightforward: (1) formulate a mathematical model that predicts shape as a function of the presumed explanatory variable; (2) fit the model to the data; (3) evaluate the fit. However, the analysis is somewhat more delicate than it seems, for two reasons. First, and perhaps most obvious, the mathematical model might not be simple (either to devise or to fit). Second, what we actually are doing is predicting shape, not explaining it – and prediction is not quite the same thing as explanation, just as a mathematical model is not quite the same as a biological model. The distinction is important to keep in mind because, when we test our model statistically, the hypothesis we are actually testing is that the mathematical model predicts shape, which is not the same as testing the hypothesis that the independent variable actually causes the variation in shape. It is common to make the distinction between

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

causation and correlation, which is often done by pointing to trends that are accidentally related; but sometimes the trends are biologically related and yet there still is not a direct, causal relationship.

To clarify the distinction between prediction and explanation, and also between mathematical and biological models, we can consider one common predictor of shape: size. Often, much (or most) of the variation in shape is predicted by size. Based on the good fit of our model to the data, we might conclude that size predicts shape, and so it might seem that size *explains* shape. However, size is not a process. In the context of developmental biology, we can explain size in terms of the proliferation of cells that add tissue to a structure. Because growth rates vary over the organism, cell proliferation (in conjunction with cell death, cell differentiation, deposition of an extracellular matrix, etc.) produces changes in shape. In this context, saying that size "explains" shape does not mean that size itself causes shape; rather, it means that we are using "size" as shorthand for all those developmental processes that jointly alter size and shape. Also, we are modeling this process by a simple mathematical function, which is the model that is actually tested. In the context of functional morphology, "size" is also shorthand, but it is shorthand for a more complex argument. The underlying causal hypothesis is biomechanical; the idea is that shape covaries with size because the mechanically optimal shape for one size differs from that for another size. However, in correlating shape to size we are not demonstrating that selection molds shape, nor even that shape affects performance; instead, we are demonstrating that the relationship between size and shape is predicted by a particular mathematical model.

Most often, that mathematical model is the equation of a straight line, hence the term "linear regression." We are fitting the equation of a straight line to the data to find the coefficients that best predict shape from values of the independent variable (e.g. size). More specifically, we are trying to find the best estimates of the coefficients m and b of the equation:

$$Y = mX + b + \varepsilon \tag{10.1}$$

where Y is the dependent variable (shape in our case), m is the slope of the line, b is the Y-intercept of the line, and ε is "error" (the variation in Y not explained by X). To predict Y from X we need to find the values for m and b. Having obtained the best estimates for them (using the approach described below), we can then ask whether they are statistically different from zero.

The approach we use to find the values for those coefficients *assumes* a linear relationship between X and Y. The reason for emphasizing this assumption is that a strong but nonlinear relationship might look like a weak linear one. Consequently, we end up rejecting our biological model because the statistical analysis suggests a weak relationship between variables, but the relation is actually strong but not linear. When the assumption of linearity holds, our statistical analysis can tell us if Y is only weakly dependent on X – meaning that knowledge about X does not enable us to predict Y. It is also possible that the relationship of the two variables is statistically significant, but that m is such a small number that the effect of X on Y is biologically trivial. It may be a *statistically significant* relationship, in that it is stronger than expected by chance, but it might not be *biologically significant*. Recognizing this distinction is important, because statistical significance is a matter of sample size and the power of a test. With very large samples, or very powerful tests, we might have little difficulty rejecting the null hypothesis. However, if X accounts for very little of the variation in Y, X provides little biological insight into Y. We therefore need to pay as much attention to the explanatory power of X and to the magnitude of its impact on Y as to the statistical results. The fraction of the variance in Y explained by X provides the needed information about explanatory power; the magnitude of the effect is evident primarily in the depiction of the regression as a deformation, although we can also estimate it from the Procrustes distance between the shapes at the lowest and highest values of the independent variable. If that distance is small, the impact of X on Y is slight.

To this point, we have talked about the relationship between Y and X as if it is the primary focus of a study. Often it is, but sometimes their relationship is a complicating factor. For example, we might want to know whether males and females differ in shape, and to make that determination we might need to take into account that they also differ in size. We then have two questions to address: (1) do males and females differ in shape? and (2) do males and females differ in shape solely because they differ in size? Even if we find that they differ in shape, we might suspect that they would be the same shape were we able to compare them at the same size. Making such comparisons is another purpose of regression. Using the regression model, we can control for differences in X when comparing groups. That is done by (multivariate) analysis of covariance ({M}ANCOVA). The independent variable, X, is treated as a covariate whose effects are controlled statistically when comparing two or more means. So long as the relationship between X and Y is linear, and the groups have the same value for m, we can statistically control for the effect of X on Y when comparing the groups. This analysis depends on two assumptions: (1) linearity, and (2) equality of m. If either is false the results can be seriously misleading; we might manufacture differences between groups, or fail to detect the ones that exist.

Prediction and control are the two main uses of regression, but there is a third: testing the equality of the regression equations. This third use is important when we wish to know whether populations evince the same response to a particular factor. For example, we might want to know if species follow the same ontogeny of shape. Each ontogeny is described by the linear model relating shape to size or age, so to compare the ontogenies we compare the vectors of regression coefficients. Similarly, we might be interested in whether two or more populations respond to variation in temperature in the same way – which we can determine by testing the hypothesis that they undergo the same changes in shape as a function of temperature.

In this chapter we begin with a general overview of regression, starting with simple bivariate regression then generalizing to multivariate regression. We first discuss estimating the parameters m and b, evaluating the strength of the relationship between X and Y, and testing the statistical significance of the regression model. We then discuss the use of regression to control for variation in X when we want to compare values of Y between groups. Finally, we discuss comparative analyses.

An overview of regression

We presume that most readers are familiar with simple bivariate regression, but we discuss it in some detail, both as a review of the general idea of regression and as preparation for moving from the bivariate to the multivariate case.

Bivariate regression

The simplest possible use of regression is to analyze the relationship between two variables, Y and X, both of which are single numbers (meaning that neither is a vector). Many mathematical models could be used to analyze the dependence of Y on X, but the simplest and most popular is a straight line, hence the model uses the formula for a line ($Y = mX + b$). We must add another term to the model, ε, representing "error," not only because measurements are made with error, but also because individuals within populations vary. "Error" in this context refers not only to measurement error, but also to *any* source of random variation in Y that is independent of X.

For the simple bivariate case, we have one *dependent variable* (Y) and one *independent variable* (X), each of which is measured on N individuals. The relationship between the two variables was given in Equation 10.1, but we repeat it here so you do not have to turn back to that page:

$$Y = mX + b + \varepsilon \qquad \textbf{(10.1)}$$

Our objective is to estimate m and b.

In a moment we will present the equations that provide the best estimates of m and b, but to explain why they are considered "best" we first need to consider how that decision could be made, in general. The standard approach for deriving the best estimator is to choose an *error function*. By minimizing that error, we find the optimal values for the parameters. A least squares analysis, as the term suggests, uses the sum of squared residuals as the error function, so that is the function minimized. We then express the relationship between that error term and the regression model:

$$\sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - mx_i - b)^2 \qquad \textbf{(10.2)}$$

where $x_i = X_i - <X>$ (the difference between an observed value of X_i and its expected value $<X>$, which is the sample mean) and $y_i = Y_i - <Y>$ (the difference between an observed value of Y_i and its expected value $<Y>$). Thus, we are summing residuals, or deviations from expected values, over all N individuals in a population. By minimizing this function, we will obtain the best estimates for m and b.

To find the values of m and b that minimize the sum of squared residuals, we set the derivative to zero (for both m and b). As you recall from calculus, the derivative of a function is zero at the maximum and minimum. We then solve for m and b. Using this optimization method, the equation for the slope, m, can be written as:

$$m = \frac{\sum xy}{\sum x^2} \qquad \textbf{(10.3)}$$

which is the sum of the products of the deviations divided by the sum of the squared deviations of the X values (each sum is taken over all individuals). In other words, the slope is the ratio of the deviations of Y to the corresponding deviations of X. When the corresponding deviations are identical, the slope is one; when the deviations of Y are a consistent multiple of the deviations of X, the slope will be that multiple.

Substituting the $X_i - {<}X{>}$ for x_i and $Y_i - {<}Y{>}$ for y_i allows us to compute m directly from the observed values. The sum of the products can be written as:

$$\sum xy = \sum (X_i - {<}X{>})\,(Y_i - {<}Y{>}) \tag{10.4}$$

which can be simplified to:

$$N \sum X_i Y_i - \sum X_i \sum Y_i \tag{10.5}$$

After applying a similar substitution and simplification to the sum of the squared deviations, we can write:

$$M = \frac{\left(N \sum_{i=1}^{N} X_i Y_i\right) - \left(\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i\right)}{\left(N \sum_{i=1}^{N} X_i^2\right) - \left(\sum_{i=1}^{N} X_i\right)^2} \tag{10.6}$$

Now that we have an expression for the slope, we can solve for the intercept, b, and complete the equation for the regression. When $b = 0$, ${<}Y{>} = m{<}X{>}$, so we can calculate b from the observed values, X_i and Y_i, and the sample size, N:

$$b = {<}Y{>} - m{<}X{>} = \frac{\sum_{i=1}^{N} Y_i - m \sum_{i=1}^{N} X_i}{N} \tag{10.7}$$

In addition to an estimate of the value of m, we will also need measures of the uncertainty of that estimate. These measures will be used to test whether m is significantly different from zero (because if we cannot say that, we cannot claim that Y depends on X), and to test whether the value of m differs between samples (whether the relationship between X and Y is different).

Before we derive the measures of uncertainty, it will be useful to introduce some shorthand notation. The sums of squares of the deviations x_i and y_i will be:

$$s_{xx} = \sum_{i=1}^{N} x_i^2 \tag{10.8}$$

and

$$s_{yy} = \sum_{i=1}^{N} y_i^2 \tag{10.9}$$

Similarly, the sum of the products of the deviations will be:

$$s_{xy} = \sum_{i=1}^{N} x_i y_i \tag{10.10}$$

In testing whether the regression is significant, it is important to keep in mind that we are asking whether the relationship between X and Y explains a significant proportion of the variance in Y. If we knew the values of the error terms, ε_i, we could compute their variance and use that to determine the proportion of the variance in Y explained by the regression of Y on X. More often than not, ε_i are unknown, so we need a different approach. Following the logic of ANOVA (Chapter 9), we can compute an F-ratio from the information we do have. F is a ratio of variances or mean squared deviations, which are sums of squared deviations divided by the appropriate degrees of freedom. The sum of squared deviations explained by the regression is s_{XY}^2/s_{XX}. This has one degree of freedom, so the variance explained is also s_{XY}^2/s_{XX}. Recall that the slope is s_{XY}/s_{XX}, so the explained variance can also be written as $m \cdot s_{XY}$. The unexplained or residual sum of squared deviations is $s_{YY} - m \cdot s_{XY}$, which has $N-2$ degrees of freedom, so the unexplained variance is $(s_{YY} - m \cdot s_{XY})/(N-2)$. F is the explained variance divided by the unexplained, so $(N-2)m \cdot s_{XY}/(s_{YY} - m \cdot s_{XY})$ with 1 and $N-2$ degrees of freedom; the corresponding p-value indicates the likelihood that such a high F (such a large proportion of the variance in Y explained by regression on X) is due to chance.

We can also use the explained variance to calculate an estimate of the variance of the slope:

$$s_m^2 = \frac{\left(\dfrac{s_{YY} - ms_{XY}}{N-2}\right)}{s_{XX}} \tag{10.11}$$

The square root of this quantity is the standard error of the slope, which can be used in conjunction with the t distribution to test whether the slope deviates from a specific value and to construct confidence intervals around the slope. To test whether the slope differs from zero, compute $t = (m-0)/s_m$ and look up the p-value associated with that t and $N-2$ degrees of freedom. To construct a confidence interval around m, select an appropriate value of α, which is the critical value for the test statistic. This value is chosen according to the rate at which we are willing to make a Type I error (which is the error of rejecting a true null hypothesis). Usually α is chosen to be 0.05, which means that we are willing to risk an error rate of 5%. To have a total error rate of 0.05, we usually want the value of t that allows on 2.5% error on either side of the estimate ($t_{\alpha/2,\,N-2}$, i.e. the critical value of the t distribution for the confidence level of α, with $N-2$ degrees of freedom). The width of the confidence interval is $2t \cdot s_m$; its upper and lower bounds are given by $m \pm t \cdot s_m$. To show that Y depends on X, m must be significantly different from zero. When N is large (>60 or so) the t-distribution approximates the normal one, so that $t_{0.05/2,\,N-2}$ is 1.96 (the 2.5% upper and lower bounds for the normal distribution).

In some circumstances it is desirable to estimate the variance in the intercept. This can be computed as:

$$\sigma_b^2 = \frac{\left(s_{XX} + N{<}X{>}^2\right)\sigma^2}{Ns_{XX}} \tag{10.12}$$

in which σ^2 is the unexplained variance (above). Again, there are $N-2$ degrees of freedom. Then the confidence interval for b can be determined using either the t or normal distribution, depending on N.

The correlation coefficient

The correlation coefficient (R), which ranges from minus one to one, expresses the strength of the linear relationship between X and Y. Its squared value (R^2), which ranges from zero to one, indicates the fraction of the variance in Y explained by X. The expression for R^2 is:

$$R^2 = \frac{s_{XY}^2}{s_{XX}s_{YY}} \tag{10.13}$$

It is very common to regard high R^2 values as if they indicate high explanatory power of the model. However, even high values of R^2 need not be statistically significantly greater than zero. For that reason we need to test the statistical significance of R^2, which we can do (assuming normality of the residuals) by the expression:

$$\frac{1}{2}\ln\left[\frac{(1+R)}{(1-R)}\right] \tag{10.14}$$

which is a normally distributed variable, with variance equal to $1/(N-3)$, where N is the sample size.

Multivariate regression

To apply this theory to shape we need to extend it to the multivariate case, because shape is multidimensional. Our dependent variable is a vector with $2K-4$ components (where K is the number of landmarks and each landmark has two coordinates). The statistics are much easier to handle if we use partial warp scores instead of the coordinates obtained by a Procrustes (GLS) superimposition, because partial warp scores have the correct degrees of freedom. Thus, throughout the remainder of this chapter, the dependent variable is a vector of partial warp scores (including the scores on the uniform component – rather than saying this repeatedly, assume that the uniform component is included whenever we refer to the vector of partial warp scores).

To regress shape on an independent (scalar) variable, we regress the full set of partial warp scores on the independent variable. For example, suppose we have P partial warp and uniform components, which we can write as a row vector $\{Y_1, Y_2, Y_3, \ldots Y_P\}$. Then the (linear) model for the regression of that vector on a scalar (X) is:

$$\{Y_1, Y_2, Y_3, \ldots Y_P\} = \{m_1, m_2, m_3, \ldots m_P\}X + \{b_1, b_2, b_3, \ldots b_P\} + \{\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots \varepsilon_P\} \tag{10.15}$$

where $\{m_1, m_2, m_3, \ldots m_P\}$, $\{b_1, b_2, b_3, \ldots b_P\}$ and $\{\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots \varepsilon_P\}$ are vectors of slope and intercept coefficients and residuals, respectively. Although this expression looks far more complicated than that for a bivariate regression, it actually is not. In fact, we can determine the ith component of the slope and intercept terms using the same m_i and b_i values that minimize the residuals in the corresponding bivariate model.

Mathematically, the regression of P components on a scalar X is identical to doing P separate simple bivariate regressions of each Y on X. The parameters m_i and b_i are determined by the equations for the bivariate case, given above. However, the test for the significance of the regression is different from that for the bivariate case because we are dealing with a multivariate system. One approach is to use the Wilks' Lambda, Λ, which is:

$$\Lambda = \frac{det(\Sigma_R)}{det(\Sigma)} \tag{10.16}$$

where Σ_R is the variance–covariance matrix of the predicted values of Y at a value of X in the data set, *det* is the determinant of the matrix, and Σ is the variance–covariance matrix for the original set of variables (i.e. partial warp scores in our case); this is the same statistic discussed in Chapter 9 (MANOVA). Several other conventional multivariate test criteria can be used that all give the same results when there is only one independent variable. Additionally, we can use a generalized form of Goodall's F-statistic to test the significance of the regression of geometric shape data on size (this statistic was also introduced in Chapter 9).

To determine the proportion of the shape variance that is a function of the independent variable we should not use the standard multivariate version of R^2 because that is a function of two determinants, one of which is the determinant of the sample variance–covariance matrix (the other is the determinant of the matrix of predicted values). Because R^2 is partly a function of the correlations among the *dependent* variables, it does not measure the correlation between dependent and independent variables. As an alternative measure of the explanatory power of the regression, we can use one that depends on the Procrustes distance between each specimen and its expected shape (given its value of X). Squaring and summing those distances gives a measure of the variance in shape *not* explained by X (because the distances are the deviations from the regression, so the model does not explain them). This metric corresponds to what we would normally regard as the variance not explained by the regression, i.e. $1 - R^2$, and has the advantage of being in the familiar (and meaningful) units of Procrustes distance.

The assumption of linearity

When we fit a straight line to the data we are assuming that the relationship between shape and the independent variable is linear. Sometimes it is not. Fortunately, in some of those cases, it is easy to transform the independent variable to make the relationship linear. For example, a number of studies of ontogenetic allometry use the logarithm of centroid size, rather than centroid size itself, as the independent variable. That transformation is useful when most of the shape change occurs over small values of X, such as when most shape change occurs early in ontogeny (as it often does). In other cases, other transformations of X (such as other trigonometric functions, for example) might do a better job of linearizing the relationship between variables. We should note that it does not matter whether the logarithm is taken to base 10 (log) or base e (ln) because these differ only by a constant, i.e. $\log(X) = \log(e)\ln(X) = 0.4329\ln(X)$.

The assumption of linearity should *always* be checked before using a linear model (and before taking any statistical test at face-value). There are at least two ways to check this assumption, although neither is ideal. One is to look at the relationship between each individual component of shape and the independent variable, such as by regressing each partial warp on size. If one or more evinces a highly non-linear relationship, such as shown in Figure 10.1A, then it is unlikely that shape and size are linearly related. This method is not ideal, because it falls back on bivariate regression when it is multivariate linearity that

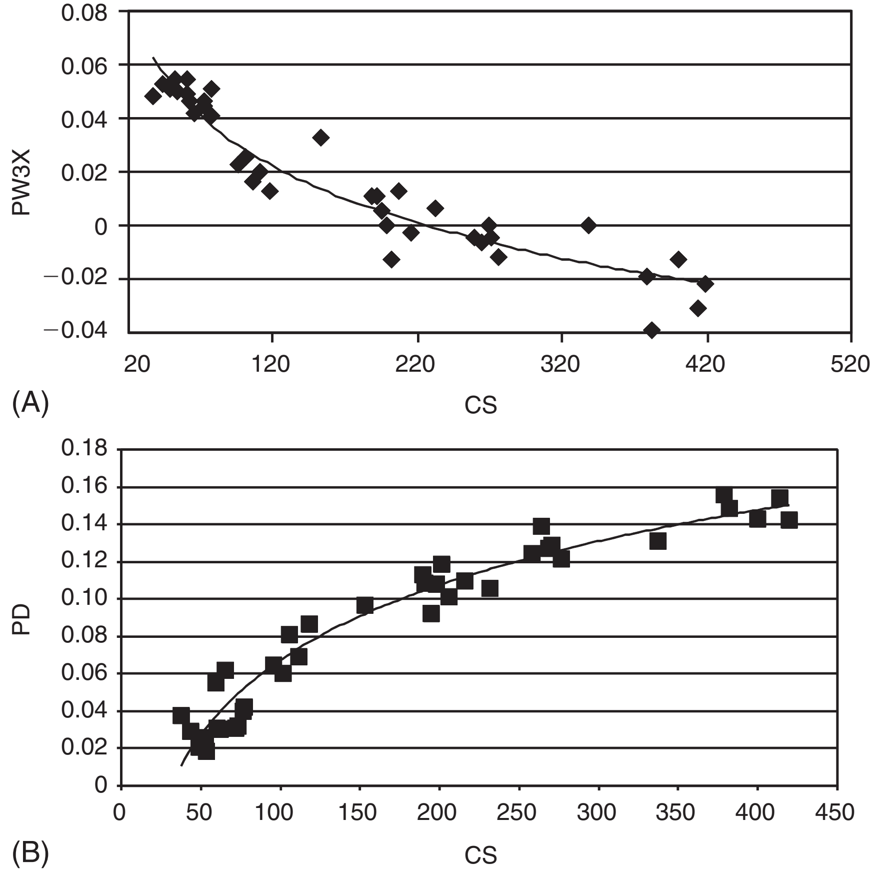


Figure 10.1 Checking the assumption of a linear relationship between shape and the independent variable: (A) using a single variable plotted on centroid size; (B) using the Procrustes distance of each specimen from the shape having the smallest size, plotted on centroid size.

matters. Another approach is to estimate the Procrustes distance between each specimen and the shape at the lowest value on the independent variable. Regressing that distance on the independent variable may show if *that* relationship is non-linear (as in Figure 10.1B). If not, it is unlikely that shape and size are linearly related. This method is again not ideal, because the Procrustes distance measures only the magnitude of the difference between each specimen and the reference, not its direction. Two specimens that differ a great deal from each other in shape may be equally distant from the reference.

Nevertheless, we can use the results from these two less than ideal methods to determine if it is unlikely that shape is linearly related to size. The results shown in Figure 10.1 both indicate a non-linear relationship, and both also suggest that shape might be linearly related to the log of centroid size (that suggestion is in the shape of the curves, which indicate a very rapid change in shape relative to size over the smaller values of size). So we can try a log transform of centroid size, then repeat the analyses to check for linearity again (Figure 10.2). Both plots now suggest a nearly linear relationship between shape and log centroid size. Thus, we would use log centroid size as our independent variable.

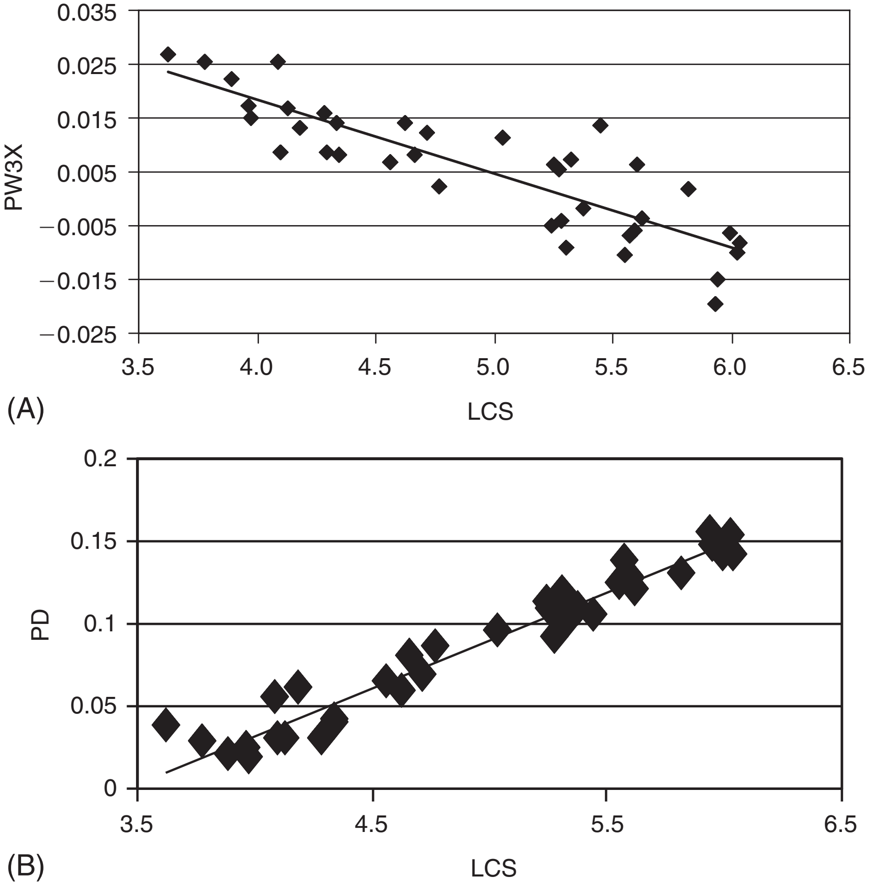


Figure 10.2 Checking the assumption of a linear relationship between shape and the independent variable: (A) using a single variable plotted on ln centroid size; (B) using the Procrustes distance of each specimen from the shape having the smallest size, plotted on ln centroid size.

To this point, we have talked about the assumption of linearity as it is usually stated in bivariate regression. However, in multivariate studies there is another assumption of linearity – the mutual linearity of all the components of the dependent variable. In other words, we are assuming that all the components of shape are linearly related to each other. This assumption will not hold if some components of shape are linearly related to the independent variable, but others are non-linearly related. The components of shape cannot be linearly related to each other if different ones fit differently shaped curves. Because this departure from the assumption of non-linearity is specific to multivariate data, it does not arise at all in bivariate studies, so it may not be intuitively obvious what the assumption means. What it means is that the slope of the relationship between shape and the independent variable is constant – the values $\{m_1, m_2, m_3, \ldots m_P\}$ are not functions of the independent variable.

In some cases, such as in studies of ontogeny, the shape variable correlated with age changes from age to age. If so, we cannot model the ontogeny of shape by a single vector of

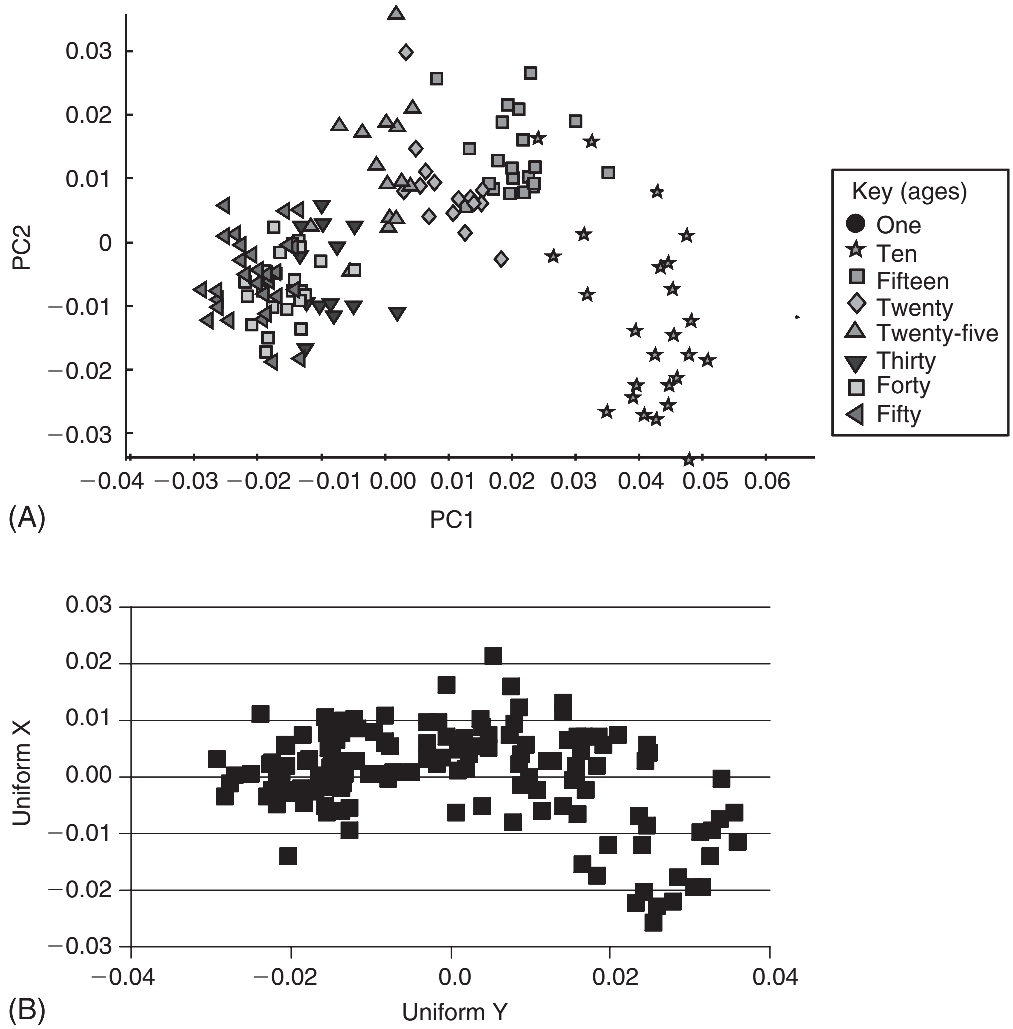


Figure 10.3 Checking the assumption of multivariate linearity of the dependent variable: (A) using principal components analysis; (B) using two shape variables on each other (the two uniform components).

slope coefficients because that vector changes with time. This means that the ontogenetic trajectory of shape is a curving path in shape space, not a straight line. The assumption of multivariate linearity can be checked in two ways, although again neither method is ideal. One, shown in Figure 10.3, is to conduct a principal components analysis (PCA) of the data, and check for a statistical relationship between multiple PCs and the independent variable. In the example shown in Figure 10.3A, there is a substantial deviation from linearity – not only is PC1 correlated with age (which we would expect), but PC2 and PC3 are also. PC2 and PC3 describe the deviations from the linear trend represented by PC1. The assumption can also be checked by regressing several shape variables on each other (Figure 10.3B). If the relationship among these variables is non-linear, we must reject the assumption of multivariate linearity.

When shape data violate the assumption of multivariate linearity, there is no easy way to transform them. They are not individual variables that can be individually transformed; they all, taken together, represent a single variable – shape. If we log transform some of the components, we thereby alter the meaning of "shape." Also, whenever the dependent variable is transformed, the error structure of the data is also affected (which does not happen when only the independent variable is transformed, because that variable is presumed to be measured without error). Moreover, and perhaps most important, the non-linear dynamics of the shape variable are not just a nuisance, they are biologically interesting (but they do complicate statistical analyses).

Testing the null hypothesis of isometry for **S. gouldingi**

We checked the assumption that shape is linearly related to size over the ontogeny of *S. gouldingi* in Figures 10.1 and 10.2, determining that shape is nearly linearly related to log centroid size. We can now test the null hypothesis that the relationship between shape and size is no greater than we would expect by chance. Our null hypothesis is *isometric* growth, meaning that shape does not change as a function of size. If we can reject that null hypothesis, we can say that shape is a function of size. Regressing the full set of partial warps on the natural log of centroid size yields a value of Wilk's Λ of 0.0109, corresponding to an F-statistic of 29.1 with 28 and 9 degrees of freedom ($p = 6.07 \times 10^{-6}$). Thus, it is highly improbable that the null hypothesis is true. We can therefore reject it in favor of the alternative hypothesis – that shape is allometric (meaning it changes as a function of size). To determine the proportion of the shape variation predicted by size, we will sum the squared Procrustes distances between the observed shape and the shape predicted for that individual given its size. From that sum, we conclude that 28.1% of the shape variance is *not* explained by the regression. Thus, 100% − 28.1% = 71.9% of the shape variance *is* explained by size.

Using regression to compare group means

To this point we have used regression to examine the relationship between continuous variables, but regression can also be used for comparing populations that differ categorically if the categories are viewed as discrete points along an inherently continuous scale. This application of regression requires transforming what was measured as a categorical variable into a variable on a continuous scale, a procedure that can be tricky (and even unjustified). We first show how it is done, then discuss when it might be justified (or not).

We will begin with a simple case, in which we are actually doing a simple two-group analysis of variance. We assign numerical values (called "dummy codes") to our two groups. Typically, one is assigned a value of 1 and the other a value of −1 or 0. Then, shape is regressed on these coded values. If the dummy codes are 1 and 0, the regression describes the difference between groups, and the intercept is equivalent to the mean of the group coded "0" (because the intercept, by definition, is the value for Y when X equals zero). Alternatively, if the groups are coded by −1 and 1, the Y-intercept is located at the mean over both groups, and the regression will show half the difference between them. The statistical significance of the regression indicates whether there is a significant difference between the two groups, and the test is equivalent to a generalized Hotelling's T^2 test.

This simple case of using regression to compare two groups raises no problems, either conceptual or statistical. It does not matter that we have transformed categorical variables A and B into the ordinal variables −1 and 0, or 0 and 1, and used a method that presumes these are continuous variables. However, real problems can arise when we are analyzing more than two groups because then the codes (e.g. −1, 0, and 1) are treated as if the distance between the integers is meaningful on a continuous scale. When we use regression, we are calibrating the effect of a change of a given amount in X on Y; if that amount of change in X is arbitrary, the calibration does not make sense. Thus, whether this approach is justified or not depends on whether it makes sense to translate the categorical variable into a continuous one.

In some cases that translation might seem reasonable, even more appropriate than leaving the variable categorical, because the categories are arbitrary subdivisions of an underlying continuum. For example, perhaps we subdivided a continuum of ages into classes such as juveniles, subadults and adults. Age is a continuously valued variable and the age classes are ordered from youngest to oldest, and we would like to take that ordering into account when analyzing the data. Using analysis of variance we cannot take the ordering into account, so regression might seem a superior approach to the data. However, our ordered classes might not be separated by equal increments of time (either chronological or developmental) – the distance from juvenile to subadult on a temporal scale might not correspond to the distance from subadult to adult on that same scale. If that is the case our X-axis is not meaningful, so it does not make sense to calibrate the change in shape by the change along a meaningless scale. That objection might be subdued by finding a strong linear relationship between shape and age class. However, it is far more difficult to justify using this approach when we cannot view the coded variables as representing a progression from least to most along an underlying factor.

The most problematic cases are those in which the categorical variable is complex but we single out one component as the independent variable. For example, suppose we wish to know whether diet has an effect on shape. To that end, we subdivide diets into "herbivore," "carnivore" or "omnivore," perhaps ordering them by percentage of meat in the diet. This approach might seem reasonable at first, but we could also order diets by hardness of food (or even by the energy required to find it, capture it, or process it, or the net energy required by all those activities). Hardness might be a reasonable choice, because carnivores that crush bone might be more similar to herbivores that crush nuts than either is to carnivores that shear flesh or to frugivores. The energy required to capture and process prey is also a reasonable choice, because shape may matter when the costs of energetically expensive activities can be reduced by optimizing shape. Considering that all three characterizations of the independent variable can yield different results, we cannot equate one of them to "diet."

Another, more technical, issue that also makes this approach problematic is that we are modeling the relationship between shape and the categorical variable by a straight line. That assumes that the relationship between them is linear, and the form of the relationship might depend largely on how we have subdivided the categories. It also assumes that there is a meaningful distance between the classes that we have quantified with some arbitrariness. For the distance to be meaningful, the change from 0.0 to −1.0 should be of the same magnitude, and in the opposite direction, as the change from 0.0 to 1.0. Moreover, the change from 1.0 to 3.0 should be in the same direction, and equal to twice the magnitude, as the change from 0.0 to 1.0. To calibrate the effect of the coded states on shape, which is what we are doing when we estimate a slope, we must have good reason to assume that there is a linear and regular relationship between coded states. For example, if we subdivide diets into herbivory, omnivory, and carnivory, coding them as 1, 2, and 3, we are assuming that the difference between herbivory and ominivory is equal to the difference between omnivory and carnivory, which implies that frugivory and granivory are equivalent (as types of herbivory). Additionally, it implies either that insectivory and molluscivory do not occur, or that they can be classified with carnivory (on physiological grounds). The problems we face are due to the complexity of "diet." Like shape, it is a multidimensional variable, so we face both the problem of transforming a complex multidimensional variable into a simple scalar, and also the challenging task of creating meaningful distances along the continuum. It may be more appropriate to treat complex multidimensional factors as exactly that and use a different method (such as partial least squares, Chapter 12) rather than regression.

The reason for raising these issues is that regression may seem like an attractive approach to analyze ordered variables. Unlike MANOVA, it does not merely ask if discrete classes differ. When the classes are ordered, regression may be a more appealing method. However, it is not always an appropriate one, even when the underlying factor is continuous. To decide whether regression is appropriate, consider whether it makes sense to treat the dependent variable as one-dimensional, and if there is a meaningful metric along that dimension.

Standardization

To this point, we have used regression to study the phenomenon of interest – the dependence of shape on another variable (size in our example). Often that relationship is *not* of primary interest, but is a nuisance that may be obscuring something more interesting. For example, we might want to ask if two species differ in shape, when we already know that they differ in size. We also know that size affects shape, so before comparing the species, we want to remove the effect of size on shape. Specifically, we want to ask if they differ in shape when the effect of size is controlled. To take a concrete example, suppose we wish to compare the shapes of *S. gouldingi* and *S. manueli*; and, as we have already shown, the shape of *S. gouldingi* depends highly on size (the same is true for *S. manueli*). If we were fortunate enough to have large samples of comparably sized specimens of both species (so that the mean size is the same for both), we could compare their mean shapes directly. However, the average body size in our sample of *S. gouldingi* is 177.99 mm whereas it is 108.95 mm in our sample of *S. manueli*, so it is possible that any differences we might find between their shapes is due to the impact of size on shape.

One common approach to solving this problem is to include a covariate in the analysis of variance (thus doing an analysis of *covariance*, ANCOVA, or MANCOVA in the multivariate case). The null hypothesis of an ANCOVA is that the groups (the two species in our case) do not differ after we take the covariate into account. In effect, we will remove the shape variance predicted by the covariate (size, in our example) and ask if the mean shapes differ. That is done by fitting both species to the same regression line (meaning the slopes are the same for both species) and comparing their values of $\{b_1, b_2, b_3, \dots b_P\}$

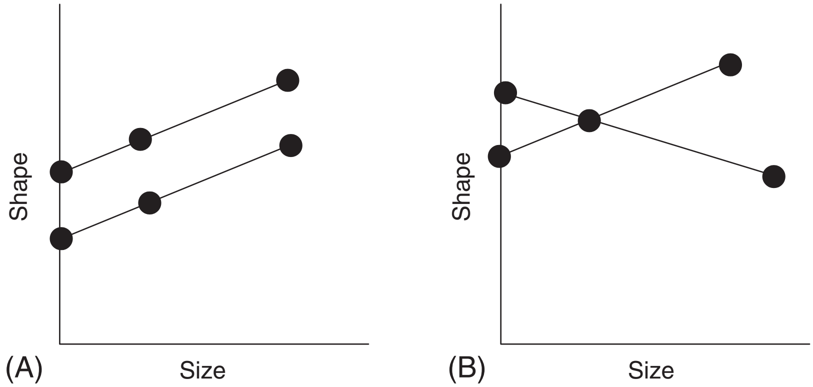


Figure 10.4 Why the assumption of common slopes matters. (A) When the slopes are the same, the same results are obtained regardless of the value of the independent variable at which samples are compared. (B) When the slopes differ, the results are a function of the values of the independent variable at which samples are compared. The points on the lines indicate the shapes being compared. Note that in (A) the same distance separates the points on the two lines over all sizes, whereas in (B) the distance between points increases then decreases as size increases.

which gives the expected shape when the independent variable is zero. Actually, it does not matter what size we use for the comparison – if the assumption of a common slope holds, we will always find the same difference between the two species. The rationale for this is shown in Figure 10.4: when the slopes are the same, the difference between the two regression lines is constant (Figure 10.4A). It is not a function of the size at which we compare them. In contrast, when the slopes are different, the difference between the two groups is a function of the independent variable; the difference between the shapes of these two groups depends on the size at which they are compared (Figure 10.4B). Because the two lines intersect, there is a value at which their shapes are the same; but because they intersect at just one point, there is only one point at which their shapes are the same.

The first step in any analysis of covariance (ANCOVA or MANCOVA) is to test the null hypothesis that the slopes are the same. This null is rejected when there is a significant interaction between the covariate (size in this case) and the factor of interest (species in this case). In the comparison between *S. gouldingi* and *S. manueli*, the null hypothesis of a common slope is unequivocally rejected; Wilks' $\Lambda = 0.151$, corresponding to an F-statistic of 9.46 with 28 and 47 degrees of freedom ($p \ll 1 \times 10^{-6}$). Clearly, it is highly improbable that the null model is true.

At this point it might not seem necessary to pursue the analysis any further, because we have already shown that the regression lines intersect so the difference in shape between these two groups cannot be a simple consequence of their difference in size. Even if their shapes are not different at some value of size, they will be at another. Nevertheless, we might still wish to pursue the analysis further because we might want to know if they differ at a *particular* size. For example, we might want to know if they are different early in development, or just later. Just because we know that the difference in shape between the two species depends on the size at which we compare them does not mean that we no longer are interested in their differences at particular sizes. In ruling out the simple hypothesis that the interspecific difference in shape is purely a function of the interspecific difference in size, we have not exhausted our questions.

When slopes differ, we can still use regression to remove the effects of size but we must remove those effects separately, group by group. Now, we have to choose the value(s) of size at which we will compare them, because the results will depend on that choice. We also need to decide whether to compare them at the same size or at a biologically comparable size. If we want to interpret the shape differences in functional terms, it makes sense to compare them at the same size; all the theories we are considering relate shape to size. However, if we want to interpret shape differences in developmental terms, we might prefer to compare groups at developmentally comparable stages. Different groups may reach the same developmental stage at different sizes (and/or ages), so comparing them at a comparable stage may require comparing them at different sizes.

Whatever size(s) we pick, the procedure is the same. We fit the data to the linear model, predict the expected shape at a particular size, and use that expected shape in our comparisons. The expectation is for the mean, and if we want to know whether the difference between species is statistically significant, we need more than the estimate of the mean. We also need to know the variation around the mean for each species. We can estimate that from the variation around the regression line – each individual deviates from the shape expected for its size. The residuals from the regression line are the deviations of an individual from the mean shape expected for its size, so we can use those residuals to estimate the variation around the expected shape at one particular size. We add those residuals to the expected shape at a given size, creating a "model population." The model population has the mean shape predicted by the regression equation, and the variance obtained from the residuals from the regression. In producing this model population we are assuming linearity of the relationship between size and shape, and even small departures from linearity can become important because the residuals will not be randomly distributed around the regression line (hence they are not randomly distributed around the mean). Also, in using the regression equation to remove the effect of one factor on shape, we are assuming that this factor does not interact with any others.

We need to take a cautious approach to size standardization (and to standardization by any other variable), but it is often useful when we want to know whether samples differ in another variable, taking into account their differences in size. To exemplify both the rationale for size standardization, as well as its impact on comparative studies, we will compare the shapes of *S. elongatus*, *S. gouldingi* and *S. manueli*, first without controlling for the effects of size, and then after standardizing them to two different sizes.

Comparing shapes of *S. elongatus*, *S. gouldingi* and *S. manueli*

We first ask whether these three species differ significantly in shape, and if so how and by how much. We will use MANOVA to test the hypothesis that they do not differ in shape (see Chapter 9), then use canonical variates analysis (CVA) to find the dimensions along which they are optimally discriminated (see Chapter 7), and then measure the Procrustes distance between their shapes to determine by how much they differ. Based on MANOVA, the three species are unquestionably different in shape, Wilks' $\Lambda = 0.0095$ corresponding to a χ^2 of 500, with 56 degrees of freedom ($p \ll 1 \times 10^{-6}$). The discriminant function misclassifies only three of the 124 specimens (one individual of *S. gouldingi* is misclassified

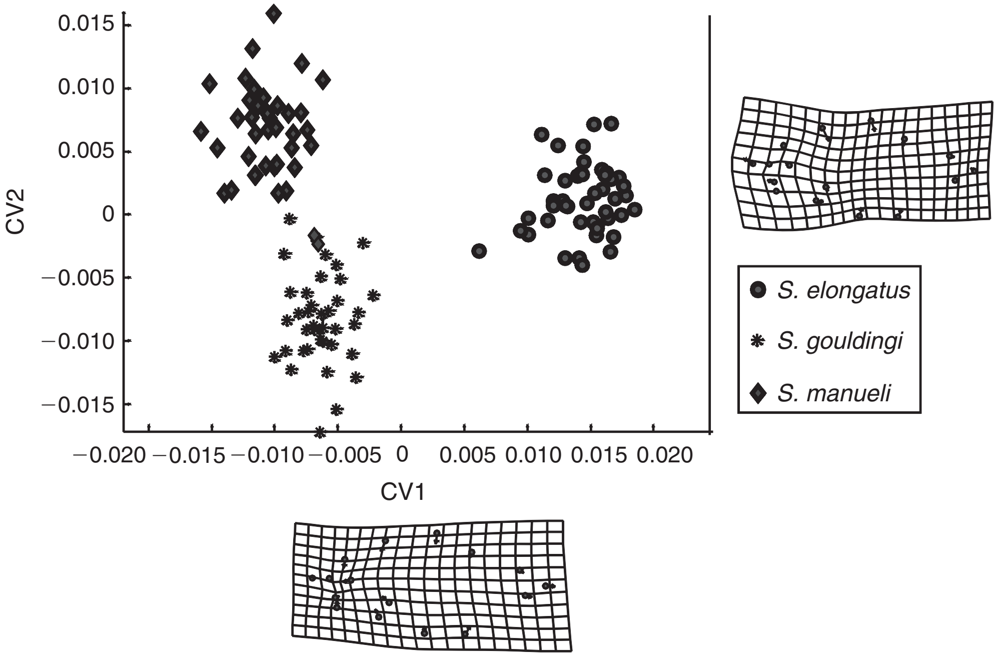


Figure 10.5 Canonical variates analysis of three species of piranhas; data are from an ontogenetic series and are not standardized to remove the variation related to size.

Table 10.1 Procrustes distances between species, the unstandardized values are calculated without removing the variation due to size from the data

Species	*Unstandardized*	*Standardized juveniles*	*Standardized adults*
S. elongatus vs *S. gouldingi*	0.079	0.060	0.126
S. gouldingi vs *S. manueli*	0.046	0.080	0.051
S. elongatus vs *S. manueli*	0.076	0.071	0.128

The distances between juvenile shapes and adult shapes are calculated after removing the variation due to size from the data. Juvenile shapes are standardized to the size at which each species undergoes the transition from larval to juvenile growth; adult shapes are standardized to the maximum adult body size for each species.

as *S. manueli*, and two *S. manueli* as *S. gouldingi*). *A posteriori* tests of the pairwise differences find that all are distinct from all others at ($p < 0.001$). The two dimensions maximally distinguishing among species are shown in Figure 10.5, and the Procrustes distances are given in Table 10.1.

We will now compare them at 20 mm standard length (SL), the size at which all three species undergo the transition from larval to juvenile growth. We do not have specimens of that size for either *S. gouldingi* or *S. manueli* because we have been unable, to date, to distinguish between them at those sizes. However, analyses of other species show that the regression of shape on size is nearly linear and that it does not tend to depart from linearity at small values, so we will extrapolate the regression to 20 mm even though that

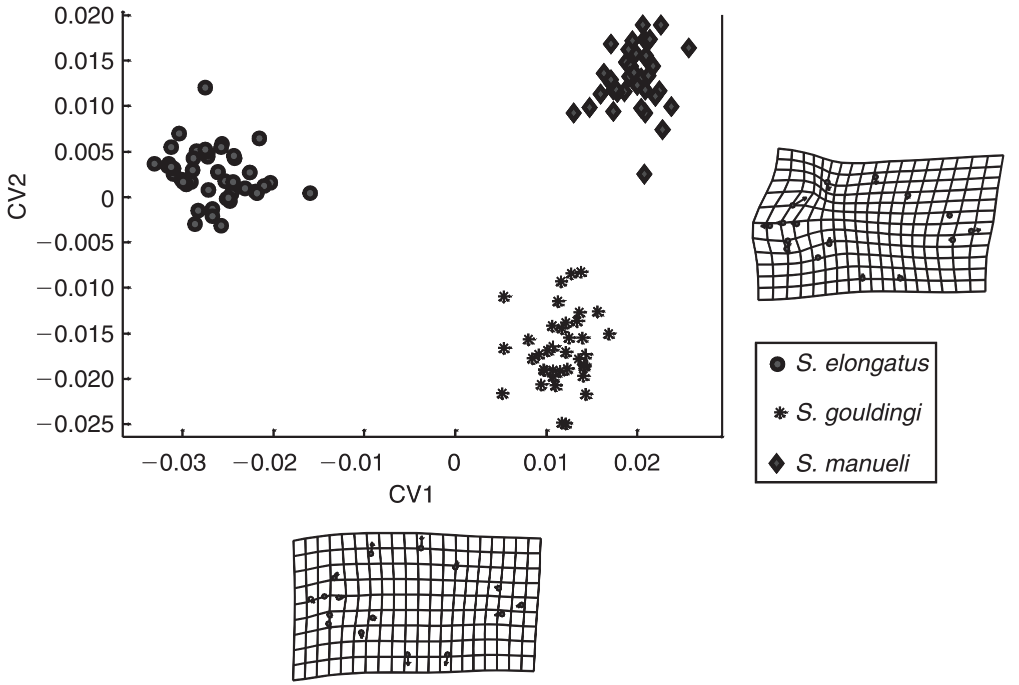


Figure 10.6 Canonical variates analysis of three species of piranhas; data are from ontogenetic series and are standardized to remove the variation related to size. Comparisons are made at the transition from larval to juvenile phases.

is beyond the range of the data for these two species. Having adjusted the data for size, we will repeat the same three analyses. Once again, we find that the three species are unquestionably different in shape, Wilks' $\Lambda = 0.0013$ corresponding to a χ^2 of 718 with 56 degrees of freedom ($p \ll 1 \times 10^{-6}$). This time, no specimens are misclassified. The *a posteriori* pairwise tests again show that all three species differ significantly from all others in mean shapes ($p < 0.001$). To this point, standardization might seem to have had little effect, other than to inform us that the differences we found above are not an artifact of the distribution of body sizes in our samples. However, the Procrustes distances are clearly affected by the removal of the size-related variation (Table 10.1). Based on the unstandardized data we would conclude that *S. elongatus* is strikingly different from the other two species, but based on the standardized data it appears that *S. manueli* and *S. gouldingi* are actually more different from each other than either is from *S. elongatus*. Also, the directions along which the species are optimally discriminated change when the data are standardized (Figure. 10.6), and so do the directions in which pairs of species differ (Figure 10.7).

The comparison between these three species conducted at their maximum body sizes reveals a very different pattern. CVA at this size still reveals unequivocally significant differentiation, with Wilk's $\Lambda = 0.0014$ corresponding to a χ^2 of 703 with 56 degrees of freedom ($p \ll 1 \times 10^{-6}$) and no specimens misclassified. However, the optimal discriminator is very different from that determined for both the unstandardized data and the data

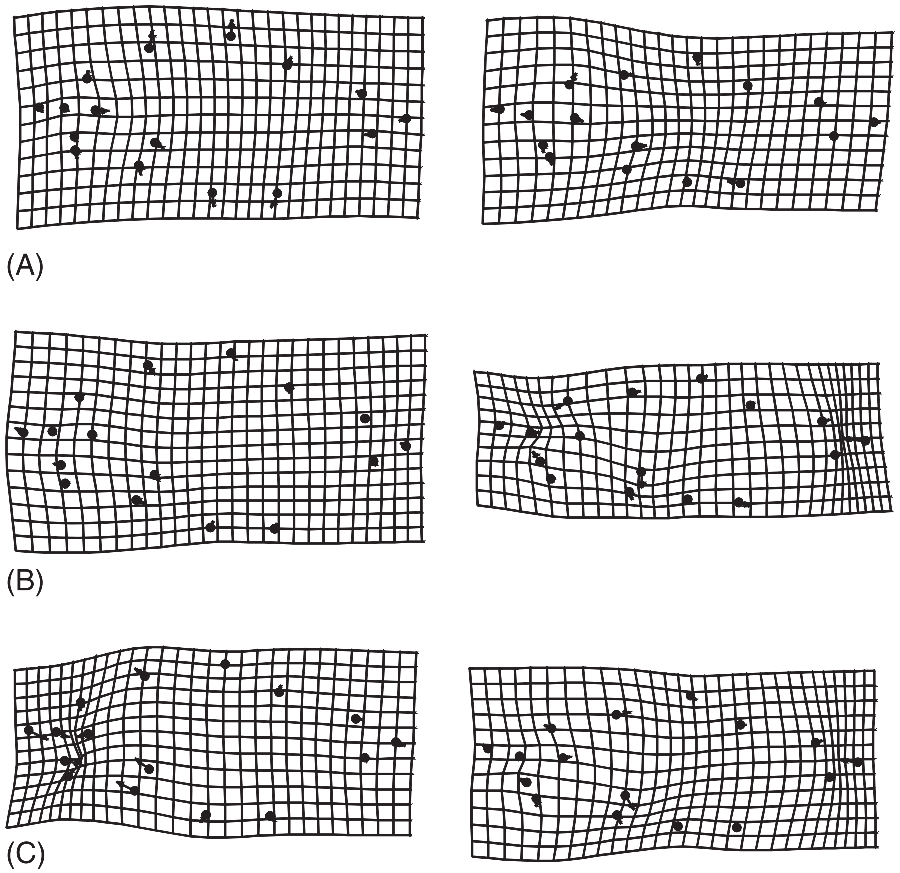


Figure 10.7 Pairwise differences between means of unstandardized data (on the left) and the data standardized to 20 mm SL (on the right): (A) *S. elongatus* vs *S. gouldingi*; (B) *S. gouldingi* vs *S. manueli*; (C) *S. elongatus* vs *S. manueli*.

standardized to a small juvenile's size (Figure 10.8), and the pairwise differences do not resemble those found in either the standardized data (Figure 10.9) or in the comparisons at small juvenile sizes (Figure 10.10). Also, the interspecific distances are affected both by standardization and by the size/age at which they are compared (Table 10.1). The distance from *S. elongatus* to both *S. gouldingi* and *S. manueli* is far greater when compared at the small juvenile size than when compared at maximum adult size, and the distance between *S. gouldingi* and *S. manueli* is far less. Although that second distance is not much greater than that found in the unstandardized data, the ratio of the distance between *S. gouldingi* and *S. manueli* relative to the distance between *S. elongatus* and each of those two species differs dramatically.

This simple example makes two important points: one is that standardization can have a large impact on both the magnitude and direction of a difference between groups, and the second is that the inter-group differences are sensitive to the value of the independent variable. This sensitivity is due to violating the assumption of a common slope. When regression lines are parallel, it does not matter what value of *X* is chosen – the same

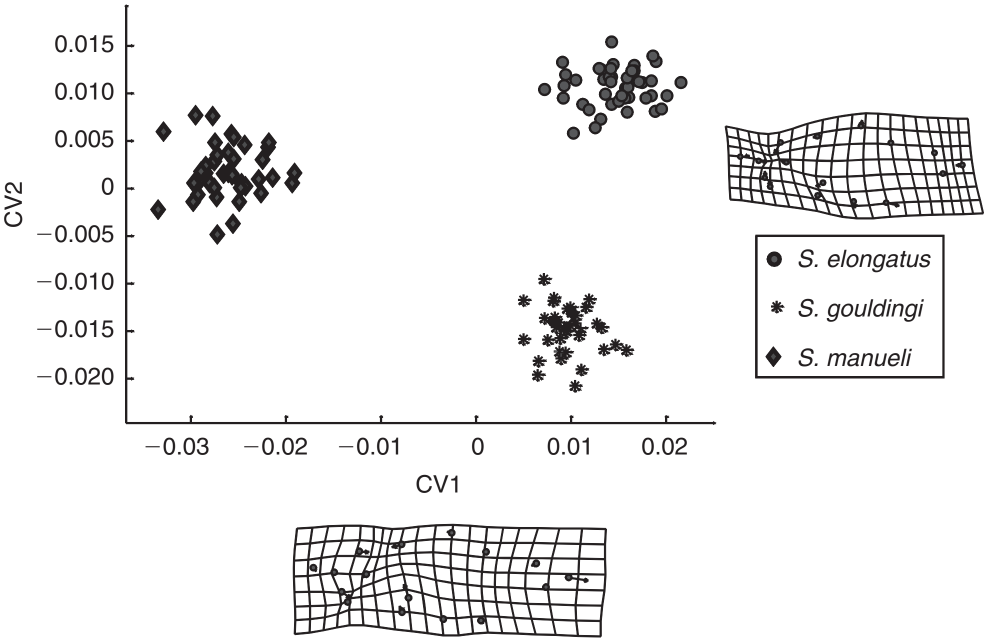


Figure 10.8 Canonical variates analysis of three species of piranhas; data are from ontogenetic series and are standardized to remove the variation related to size. Comparisons are made at maximum adult body size.

differences are found over all values, so we can arbitrarily pick any point (usually, they are compared at $X = 0$). However, when regression lines are not parallel, the results depend upon the value chosen for the standardization. This is important because it means that we cannot find "*the* difference" between species – the differences are a function of the value of X at which they are compared. Consequently, the value of X must be chosen carefully, and that choice is a biological matter, not a technical, statistical one.

It can be difficult to make a choice, especially when we are using one variable as a proxy for another, such as when using size as a proxy for age. When species differ in the relationship between size, age and developmental stage, the results can depend on whether they are standardized to a common size, age or developmental stage. However, one advantage of separately standardizing each sample is that we can pick different values of the independent variable to make the samples comparable by age or developmental stage rather than by size. Of course, if we are actually interested in size, we would make them of comparable sizes. Because our biological conclusions will be affected by these choices, alternatives must be weighed carefully.

Comparing dynamics of shape

Often we are interested in comparing groups according to how they respond to a common factor. For example, we might want to know if two populations follow a common

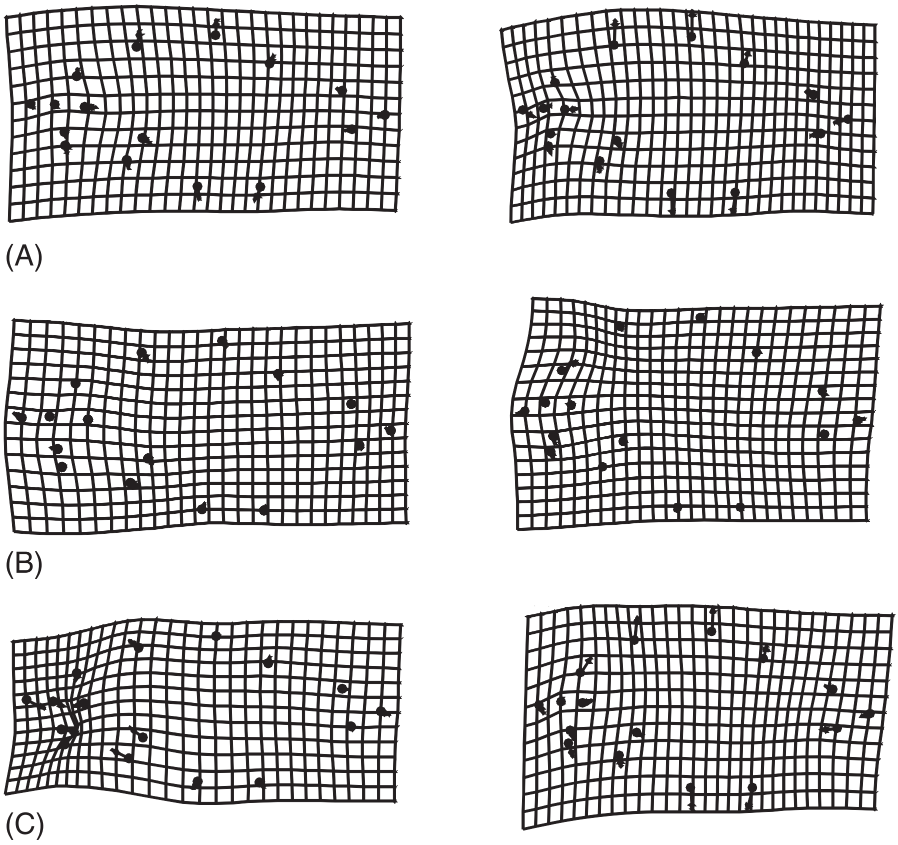


Figure 10.9 Pairwise differences between means of unstandardized data (on the left) and the data standardized to maximum body size (on the right): (A) *S. elongatus* vs *S. gouldingi*; (B) *S. gouldingi* vs *S. manueli*; (C) *S. elongatus* vs *S. manueli*.

latitudinal gradient, or if their ontogenies are the same. These questions are complex because there are two components to them:

1. Do the groups respond at the same rate to that factor?
2. Do the groups undergo the same changes in shape in response to that factor?

In the case of ontogeny, for instance, we might want to know whether species have the same developmental rate and we might also want to know whether they undergo the same ontogenetic changes in shape. Such questions are answered by comparing the multivariate regression equations.

Comparing regression equations (directions of change)

The question addressed in this section is whether two (or more) samples undergo the same change in shape in response to the same independent variable. For example, is the change in shape over one phase of ontogeny the same as that over another phase, or do two species

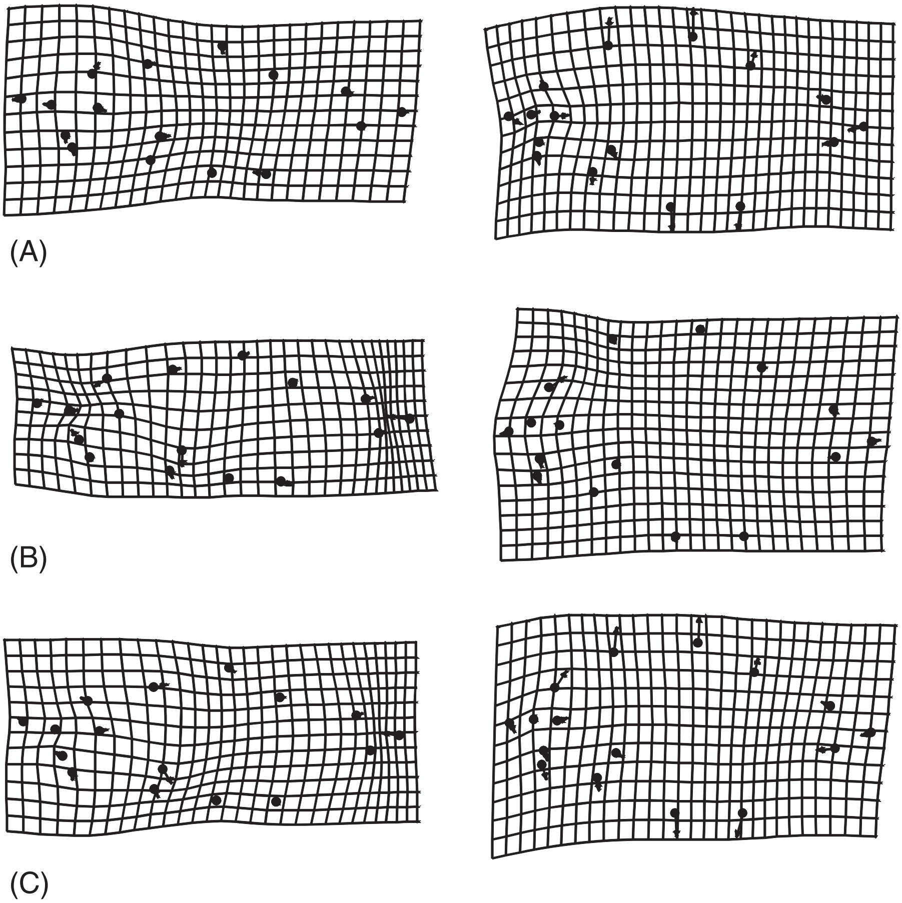


Figure 10.10 Pairwise differences between means of data standardized to 20 mm SL (on the left) and the data standardized to maximum body size (on the right): (A) *S. elongatus* vs *S. gouldingi*; (B) *S. gouldingi* vs *S. manueli*; (C) *S. elongatus* vs *S. manueli*.

share the same ontogeny of shape? This question is answered by comparing the multivariate regression vectors, which is done by measuring the angle between them. From geometry, when the regression vectors are the same line they point in the same direction, so the angle between them is 0°. Thus, the angle between the two vectors is a measure of their similarity in direction. The cosine of the angle is the vector correlation (R_V), so we can also use this as a metric of similarity. Using one or both of these measures we can quantify the similarity between vectors, and ask if the angle between them differs statistically significantly from 0°, meaning by more than expected by chance. This is equivalent to asking whether the correlation between them is significantly lower than 1.0. We can also ask if the samples differ by *less* than expected under the null hypothesis that the two vectors are independent, which is equivalent to asking if the angle between them is less than 90° and if the correlation between them is greater than 0.0. We first explain how to calculate that angle, then how to test it statistically.

Calculating the angle between two vectors

The angle between any two vectors **A** and **B**, each with P components, may be computed by taking the dot product (also called the "inner product") of the two vectors. The dot product is calculated by multiplying the corresponding components of the two vectors together, then summing those products. For example, if we have two vectors, **A** and **B**, with $\mathbf{A} = \{A_1, A_2, A_3, \dots A_P\}$ and $\mathbf{B} = \{B_1, B_2, B_3, \dots B_P\}$, the dot product is:

$$\mathbf{A} \cdot \mathbf{B} = A_1B_1 + A_2B_2 + A_3B_3 + \cdots\ A_PB_P \tag{10.17}$$

To calculate the angle between two vectors of shape variables, we would first estimate the regression coefficients for each such component, such as the regression coefficients for partial warps. We would then calculate the dot product by multiplying the coefficient of PW1X in one species by the coefficient of PW1X in the other, and then multiply the coefficient of PW1Y in the first species by the coefficient of PW1Y in the other, continuing the same process for all coefficients. Finally, we would sum all those products. When the two vectors are both normalized to unit length (meaning that the square root of their summed squared coefficients equals one), the dot product is the vector correlation, R_V. Because a correlation is a cosine of an angle, we can also write the equation for the dot product as:

$$\mathbf{A} \cdot \mathbf{B} = |A|\,|B| \cos\theta \tag{10.18}$$

where $|\boldsymbol{A}|$ is the magnitude (length) of **A**, which is calculated by $(A_1^2 + A_2^2 + \dots A_P^2)^{1/2}$, and similarly $|\boldsymbol{B}|$ is the length of **B**, calculated by $(B_1^2 + B_2^2 + \dots B_P^2)^{1/2}$, and θ is the angle between them. If **A** and **B** are unit vectors the two lengths $|\boldsymbol{A}|$ and $|\boldsymbol{B}|$ are both one, so to find the angle between the two vectors we solve for θ by:

$$\theta = \arccos \frac{(A \cdot B)}{(\,|\mathbf{A}|\,|\mathbf{B}|\,)} \tag{10.19}$$

When two vectors are parallel, the angle between them is 0° and the vector correlation between them is 1.0; in contrast, when two vectors point in exactly the opposite directions (which is termed being anti-parallel), the angle between them is 180° and the vector correlation between them is −1.0. The angle between perpendicular (orthogonal) vectors is 90°, and the correlation between them is 0.0.

Testing the statistical significance of the angle

Once we have computed an angle between two regression vectors, we are left with the question of whether it is statistically significant. Rather than attempt to find an analytic test of significance, we can rely on a bootstrap procedure (see Chapter 8 and references cited therein). The approach used will be to determine a confidence interval for the range of angles between regression vectors that can be produced by random variation within each group. At issue is whether the uncertainty of our estimate of each vector (due to sampling) is so large that we cannot reject the null hypothesis of no difference.

To estimate the range of angles within each species, we estimate the residuals from the regression of shape on the independent variable. Each individual gives a multidimensional set of residuals that describe the deviation of that individual from its expected shape. We then form a pair of bootstrap sets for each group that will be used to calculate the angle between the vectors. These pairs are constructed by resampling the residuals (with replacement) and randomly assigning them to expected values of shape (derived from the original regression model) at the values of size observed in the original data. This procedure preserves the covariance structure among variables, and is a multivariate extension of the standard approach to estimation of uncertainties of regression slopes by resampling.

From the paired samples we calculate the angles between the vectors, reiterating this procedure to generate a distribution of within-group angles. Because sample sizes can differ for different groups, the two bootstrap sets formed from the group with the larger sample size match the sample sizes of the two groups (that is, one of the bootstrap sets will have a sample size equal to that of the group with more observations, and one will have the sample size equal to that of the group with fewer observations). Both bootstrap sets formed from the data of the group with the smaller sample size have that group's smaller sample size because we ought not to form bootstrap sets larger than the original data set.

We then determine the statistical significance of the inter-group angle by comparing it to the 95th percentile of the range of both within-group angles. Should it be larger, the inter-group difference is judged to be statistically significant. Under those conditions, the inter-group angle is judged to be statistically significant at a 5% level.

Comparing ontogenies of shape of **S. elongatus, S. gouldingi** *and* **S. manueli**

We will compare the ontogenies of *S. elongatus*, *S. gouldingi* and *S. manueli* using the method described above. Each data set comprises an ontogenetic series; because we do not have information on their ages, we will regress shape on size and compare those regressions. Their ontogenies of shape are shown in Figure 10.11. They visibly differ, but part of that difference might lie in a difference in developmental rate – the species might differ in *how much* change they undergo, and hence in the lengths of the vectors rather than in their directions. However, the angle between the ontogenetic vectors of *S. gouldingi* and *S. elongatus* is 38.8° (corresponding to $R_V = 0.779$) and the 95th percentile of the ranges of the within-species angles are 10.2° for *S. gouldingi* and 34.4° for *S. elongatus*. The interspecific angle exceeds both those within-species ranges, so we can conclude that the two species differ significantly in their ontogenies of shape.

Similarly, the angle between *S. gouldingi* and *S. manueli* of 35.0° (corresponding to $R_V = 0.819$) exceeds the 95th percentile of the range obtained by resampling within *S. gouldingi* (11.0°) and within *S. manueli* (16.6°). Also, the angle between *S. manueli* and *S. elongatus* is 46.0° (corresponding to $R_V = 0.695$) in comparison to the range of angles obtained by resampling within *S. manueli* (13.7°) and *S. elongatus* (32.1°). All three comparisons demonstrate statistically significant differences between ontogenetic trajectories of shape.

These comparisons tell us that the vectors are significantly different, but they may still be far more similar than expected by chance. We have tested the null hypothesis that the vectors do not differ by more than expected by chance, but now we want to test the null hypothesis that they are no more similar than expected by chance (meaning that the angle is significantly smaller than 90°). This second null hypothesis is tested by a permutation test, comparing the observed vector to randomly permuted versions of it. These permutations preserve the range of values in the original data, as well as the relative frequencies of positive, negative, and high and low coefficients. Permuting of the coefficients numerous

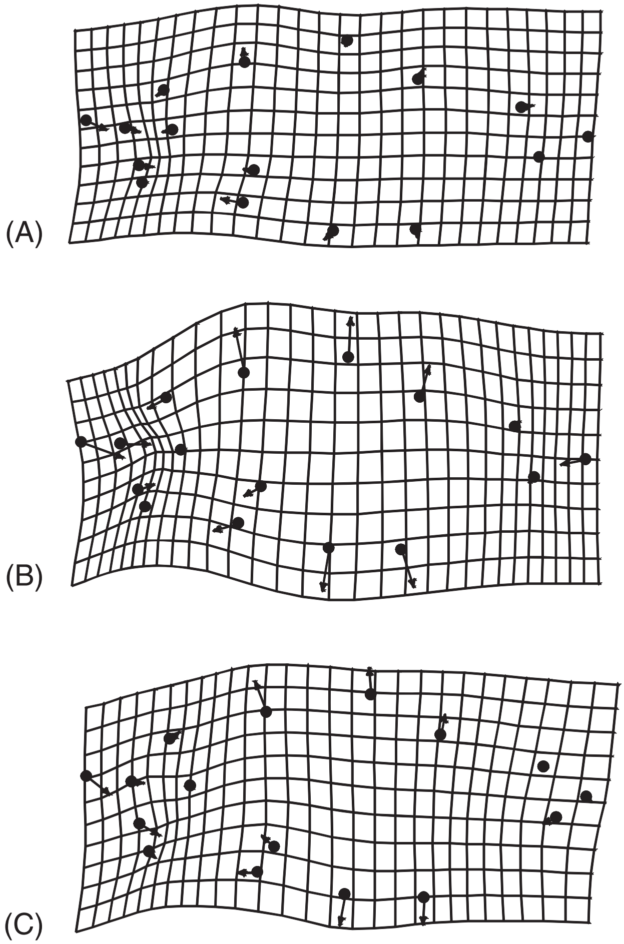


Figure 10.11 Ontogenies of shape: (A) *S. elongatus*; (B) *S. gouldingi*; (C) *S. manueli*.

times (e.g. 400) gives us a measure of the average angle among randomized vectors as well as a confidence interval for the correlation. Permuting the vector of regression coefficients of *S. gouldingi* 400 times, we find an average correlation among the randomized vectors of 0.011, with a confidence interval ranging from −0.296 to 0.345 (corresponding to an average angle of 89.3° within a confidence interval of 107.2° to 69.8°). For *S. manueli* the mean correlation among the randomized vectors is 0.020, with a confidence interval ranging from −0.288 to 0.318, and for *S. elongatus* that mean is 0.024, with a confidence interval ranging from −0.276 to 0.418. Therefore, the three species are all more similar to each other than expected by chance.

Comparing two angles

In some cases, we might want to know whether the angle between one pair of regressions is significantly different from the angle between another pair of regressions. For example, we might want to know if one ontogenetic trajectory departs more than another from the putative primitive trajectory. This question can be addressed by bootstrapping the difference between angles, just as we would bootstrap the difference between Procrustes distances. We begin by computing the angles between trajectories and the difference between those angles, then we resample each data set with replacement and repeat the calculation of the angles and the difference between them. After a sufficient number of bootstraps, we can determine the 95% interval for the range of differences. If this range excludes zero, we can conclude that the observed difference is significant at the 95% level.

Comparing rates of response to the independent variable

The simplest method for estimating the rate of response takes the approach used above for checking the assumption of linearity between shape and independent variable – calculating the Procrustes distance between each shape and the shape having the lowest score on the independent variable. By regressing that distance on the independent variable (e.g. size) we estimate the rate of response of shape to the independent variable, and can also put confidence intervals on the rate. These confidence intervals, however, do not take into account the uncertainty of the estimate of the shape that will be used as the reference. Although that uncertainty does not normally matter much (because we are not treating the reference as a statistic), it does in this case. An alternative and somewhat more complex procedure is to measure the distance between the average shapes separated by one unit of the independent variable. This tells us how much shape has changed over that single unit of change in the independent variable, which gives us the distance traveled (for shape) relative to a unit change in the predictor. For example, if we want to estimate a rate of development, we could measure the distance between the mean shape at 3 mm and 4 mm; this would tell us how much of a difference occurs per mm change in size. Because the function we are using is linear, that rate is a constant. Confidence intervals can be constructed for both methods; the first is the confidence interval around the slope of the regression, the other is the confidence interval around the distance.

The two approaches can give different results, partly because of the different sources of uncertainty that enter into the estimates of the rates, and partly because of the impact of non-linearities on the relationship between shape and the independent variable.

Comparing rates of ontogenetic shape change among S. elongatus, S. gouldingi *and* S. manueli

Continuing the comparison between *S. elongatus*, *S. gouldingi*, and *S. manueli*, we can now ask if they differ in their rate of change in shape relative to size. We use the two methods for comparing those rates (Table 10.2), and it is clear that the species do differ regardless of the fact that the two methods yield somewhat different values. Despite those differences, it is clear that there is a striking difference between *S. elongatus* and the other two species – *S. elongatus* undergoes about half as much shape change per unit change in centroid size. The other two do not differ significantly in rate, although this is because of the large confidence intervals surrounding the estimates.

Table 10.2 Estimating the rate of response to a common independent variable by the slope of the regression and by the distance traveled over a unit change in the independent variable

Species	*Slope*	*Distance traveled*
S. elongatus	0.0215 (0.0172–0.0253)	0.0245 (0.0224–0.040)
S. gouldingi	0.0579 (0.0530–0.0627)	0.0510 (0.0436–0.0583)
S. manueli	0.0557 (0.0511–0.0603)	0.0469 (0.0412–0.0566)

Software

In the IMP series, four programs are currently available to implement the methods introduced in this chapter: **Regress** performs a regression of shape on an independent variable; **Standard** removes the variation in shape due to variation in the independent variable; **VecCompare** estimates the angle between vectors of regression coefficients and statistically tests whether that angle exceeds those that can be obtained by resampling within each population; and **ShuffleAllometry** performs the permutation test (randomly reshuffling vectors of regression coefficients) to test the hypothesis that the angle between two vectors is no more similar than expected by chance. One other program, **VecDisplay**, performs no calculations but is useful for visualizing similarities and differences among vectors. In addition to programs in the IMP series, **TPSRegress** fits a variety of linear models to the data (including models more complex than those that can be analyzed using the IMP software or discussed in this book).

Running Regress6

As with other programs in the IMP series, the input data should be in X1, Y1, ... CS format. The program will regress shape on the last column of your file; if you do not want to regress shape on size, replace the column of centroid sizes with the independent variable of your choice. The input coordinates X1, Y1, ... can be obtained by any superimposition method (**Regress6** does a GLS Procrustes superimposition, and the analysis is based on the Procrustes coordinates). The program asks you to set the superimposition type, but this is for the display of results – you can input any coordinates you want and display the results using any that you want (of the options listed). Specify your choice before loading the data – you can change it later.

To run the program, load the data (clicking on **Load Data**) and then specify whether you wish to use the untransformed values found in the last column of your data, or log transform it instead. You will also need to specify the reference form. Usually, you will choose the mean when performing the statistical analysis (where it asks you to **Set N**, enter the sample size). However, you might want to depict the change away from the specimens with the lowest values on the independent variable rather than away from the mean. Although the change that is depicted will be the same regardless of your choice of reference, the landmarks will be placed where they are in the reference – so, to show the change away from a small juvenile to a large adult, you might want to place the landmarks where they are found in a small juvenile. If so, **Set N** to the number of individuals in your sample that will allow you to estimate the desired mean shape. You will also need to use that reference to plot the Procrustes distance away from that reference regressed on the independent variable (if you use the mean you will probably see a U-shaped curve, because the specimens will likely approach the mean shape then depart from it). Occasionally you might want to load a file to use as a reference form, such as when you want to compare vectors of partial warp coefficients regressed on size (in **VecCompare**) or if you want to input the vectors into **VecDisplay**. For the regression vectors of partial warp scores to be comparable, all must be calculated from the same reference form.

To run the program, click on **Compute Partial Warps**. This not only computes the PWs, it also carries out the regression. You can now check that there is a significant relationship between shape and the dependent variable (using the options on the **Regression Statistics** pull-down menu), display the Procrustes distance of each specimen from the reference on the independent variable (using **Display Distance vs CS/LCS**), and display the relationship between shape and the independent variable as a deformation (using **Display Regression (Deformation)**).

If your sample size is large enough for a multivariate analysis, you can select the option on the **Regression Statistics** menu: **PW+Uniform vs CS/LCS**. This will give you Wilk's Λ, Rao's F, the degrees of freedom and the p-value (although it may tell you that $p=0$ if the value is smaller than the program calculates). At present, Goodall's F-test is not available (but it is in **TPSRegress**). If your sample size is too small, use the univariate test of **Procrustes Distance vs CS/LCS**. However, the null hypothesis being tested is that the Procrustes distance from the smallest value is significantly related to size – this is not the same as the hypothesis being tested by the multivariate test. Still, it does give an estimate of the rate of change, and you may wish to plot that relationship and save the plot. If so, use **Auxiliary Copy** on the toolbar at the top. Copying the image directly to the clipboard will not work because of the different ways the two copy functions treat the aspect ratio of the plot.

If the statistical test indicates a significant association between shape and the independent variable, you may display it using a variety of options – including vectors of relative landmark displacements, the deformed grid, deformed grid plus vectors, a quiver plot, contour plots, and contouring the absolute values of the partial warps. The plots can be edited as described in Chapter 7 (see especially the discussion of the grid trimming options and reference rotation options). The images can be copied to the clipboard or saved to an Encapsulated Postscript (EPS) file.

As well as saving the pictures, you can save several files, including partial warp scores, the growth vector (i.e. the vector of regression coefficients for the partial warps normalized to unit length) or the deformation vector (which is the same as the growth vector except it is not normalized), the reference form, the Procrustes distance between each specimen and the reference. Normally there is no reason to save the files of partial warp scores or the reference, but if you are planning on running **VecCompare** (see below) you will need a file of partial warp scores for each group (all of which must be calculated from the same reference). In addition, you may also save a file of regression information, including the name of the file analyzed, its sample size, the reference (the name of the file and the coordinates of the reference), the regression coefficients of the partial warps (the uniform components are listed as PW0x, y), and results of the univariate test of Procrustes distance on CS/LCS.

Running Standard6

This program is somewhat different from several others in the IMP series in that it requires you actively to accept the defaults. As usual, you load the data in the standard X1,Y1…CS format. The data set may be landmarks or partial warp scores. The data will be plotted in the visualization window (but when they are partial warp scores, the plot will be meaningless). You need to say whether the independent variable is in the last column of your data file (where CS is usually located) or instead is contained in another file. To say that it is in the last column, select **Use x = CS** (the independent variable need not be centroid size; selecting this option means that the variable is located where CS usually is). Alternatively, you can select **Load x-List**, which allows you to input a file containing the values of the independent variable. This file must be a single column of data, with one entry per specimen (in the same order as the specimens are in the input data file). The values must be numerical, no letters or formatting codes can be read (although they can be included in the file by placing them to the right of a % sign).

An error message will appear if the number of specimens and number of values in the independent variable do not match. Also, you will get an error message if there is a hard-return after the last value in the independent variable list (if you get an error message that doesn't make sense, check for this possibility).

You can choose to regress on the values in the last column of your data file (or on the input independent variable file) or you can transform it, either to ln(x) or log(x). Once you have made your choice (where you are asked to **Accept: Regression Function**), click **Accept.** The last choice you need to make is the value of the independent variable; the default is the minimum value of x, but you can choose the mean or the largest value, or type in one of your choosing. Once you have decided whether to **Accept: Standardize on x =**, click **Accept**.

Clicking on the **Do Regression** button runs the program. You can then show the standardized data (they are plotted in red if you ask to **Show Standardized Data**). As usual, you can copy the image to the clipboard (with or without axes on the plot) using the **Copy Image to Clipboard** button, and you can save the standardized data in the X1, Y1 … CS file format by asking to **Save Standardized Data.**

Before loading your next file, click on **Clear Data**.

Running VecCompare

The purpose of this program is to determine whether two vectors differ significantly in direction. It takes the input data, calculates the regression of all the dependent variables on the independent variable (i.e. the last column), normalizes the vector (to unit length), estimates the angle between the vectors of two groups, and tests that angle for its statistical significance by bootstrapping. The program was designed to be very flexible – it can be used to compare vectors of traditional morphometric measurements (or of any other set of variables) as well as geometric shape data. Because it was designed to be flexible, the program does not superimpose the coordinates; nor does it calculate partial warps from them. If you want to regress the GLS coordinates or partial warp scores on the last column in the file, you will need to input files containing GLS coordinates or partial warp scores (make sure that the same reference was used in computing them).

Each group must be in a separate file. When you load them, the data will be plotted in the visualization window (the plots will not make sense if the data are traditional morphometric measurements or partial warp scores). Before doing the analysis, choose whether to regress on the values in the last column (where CS is usually located) or its log (LCS). Also before doing the analysis, determine the number of bootstrap sets you wish to use; the default is 100. To perform the analysis, click on **Compute Growth Vector**. The results will be displayed in the results window; it will give the angle between vectors, and the range of angles that can be obtained by resampling each. They can be saved to a file using the **Save File** option.

To test whether the angles between pairs of vectors are significantly different, load the first pair of data sets using the **Load File 1** and **Load File 2** buttons. Load the second pair of data sets by going to the **File** pull-down menu and selecting **Load Data Set 3** then **Load Data Set 4** (for a three-way comparison, A-B vs A-C, load A as data sets 1 and 3, load B as 2 and C as 4). Next, select **Regression Function** (CS or LCS) and **Number of Bootstrap Sets Used** in those control windows. Now start the calculation by going to the **More Stats** pull-down menu and selecting **Bootstrap Test of Difference in Angle**. The results will be displayed in the results window; they can be saved to a file using the **Save File** option.

Running ShuffleAllometry

Unlike all other programs in the IMP series, this one reads data in column vector format. The input data should be a single column of regression coefficients. These can be regression coefficients from any sort of data, such as allometric coefficients of traditional data. The original purpose of the program was to shuffle allometric vectors of traditional morphometric data, so the program can read these as well as column vectors of geometric shape variables. The reason for formatting the data as column vectors is because most journals publish vectors of allometric coefficients (or principal component coefficients) as column vectors. If your vectors are in rows, you can transpose them by opening the file in Excel, copying the vector, then, pasting it, using the **Paste Special** menu to select the option "transpose."

Once you have loaded the data, you need to select the desired number of bootstraps; the default is 400 and can be altered by typing the desired number in the box. Clicking on **Shuffle** runs the program. The results will appear in the results window, where you will find the mean value for the correlation between the input vector and 400 randomly permuted versions of it, along with the upper and lower 2.5 and 5 percentiles of the distribution. You will also obtain the maximum value found over the chosen number of bootstraps.

Running VecDisplay

This program is intended for purely graphical purposes. Its function is to provide pictures of deformations that cannot be obtained by the conventional software but can be expressed in terms of partial warp scores. You can use it to display a hypothetical vector, such as the expected change in shape under a model (so long as you can frame the model in terms of partial warp scores). You can also use it to display the difference between two deformations obtained from other software, such as the difference between two regressions (those files, properly formatted for input into **VecDisplay**, are produced by **Regress6** and can be saved either in normalized (the Growth vector) or unnormalized (the Deformation vector) form). The program will display each vector individually, and the sums and differences between them, using all the standard display options. In addition to the usual displays, there is also an option to show the difference between two deformations by pairs of vectors at each landmark.

The input files each contain a row vector of partial warp scores, ordered from greatest to least bending energy, including the two uniform components (the last column, as usual, is centroid size). For **VecDisplay** to be able to interpret the partial warp scores, it needs the reference form that was used in calculating them. The same reference must have been used in calculating the partial warp scores for both files. If you want to input a hypothetical vector, you need to produce the partial warps describing it. One method you could use for this purpose is to draw expected changes in landmark locations (as vectors of landmark displacement from a given shape, which will serve as the reference). You can then digitize the coordinates located at the endpoints of the vectors and use those coordinates in any program that outputs partial warp scores. In some cases a biological model might correspond to a particular partial warp – for example, one partial warp might describe a global growth gradient. If that is the case, you can input a vector of zeros for all scores except for that one.

After loading the reference form and selecting the superimposition method, load the one or two vector files. If using only one file, select the **Zero Vector** option for the second. You may display (1) each vector separately (these are the first two options), or (2) the sum of the two vectors (which represents an average between them, omitting only the step of dividing the scores by two), or (3) the difference between them, either by subtracting the second from the first or the first from the second.

The options for drawing the deformations, as well as for editing them, are as described in Chapter 7, with the exception that you can also draw the difference between the two groups as pairs of vectors at each landmark (the final option). We should note that if the input vectors are the normalized growth vectors produced by **Regress6**, you probably will need to rescale them because the changes will be so large that they will be virtually unreadable. To reduce the scale, use the **Deformation Multiplier** function.

11

Partial least squares analysis

Partial least squares (PLS) is a method for exploring patterns of covariation between two (and potentially more) blocks of variables. It can be used to study covariation between two blocks of shape variables, making it potentially useful for studies of morphological integration (e.g. Bookstein et al., 2003; Bastir et al., 2004), and also for synthesizing information about three-dimensional morphologies sampled by two two-dimensional views (e.g. Rohlf and Corti, 2000). The method can also be used for analyzing the relationship between shape and other variables, including traditional morphometric or meristic variables, or measures of biomechanical, ecological factors or behaviors (e.g. Corti et al., 1996; Lundrigan, 1996; Rüber and Adams, 2001). A particularly creative use of the method examines the relationship between patterns of fluctuating asymmetry and variation (Klingenberg et al., 2001).

An important feature of the method is that the blocks are taken as a given – the data are partitioned into blocks before the analysis begins. For that reason, PLS is not useful for finding the blocks in the first place, which means it is not suited for dissecting complex morphologies into modular units. Even so, it may be useful for building integrated units out of the modules, if we accept that the blocks are indeed modules. We can ask whether one block is more highly correlated with another than either of those is with a third block (e.g. Bastir et al., 2004). Although that is part of the question normally addressed in studies of integration, the other part is the identity of the blocks. PLS may eventually prove useful for that second question as well, but no current implementation of the method has attempted to address it. Thus, when using PLS to study morphological integration, it is important to remember that the method is useful for analyzing correlations between units but not for subdividing the whole into units.

PLS is probably unfamiliar to many biologists, although it has been used extensively in the social sciences (see Bookstein, 1982; Jöreskog and Wold, 1982) and in clinical studies (e.g. Sampson et al., 1989; Streissguth et al., 1993; Lowe et al., 1997). Thus, a large part of this chapter discusses similarities and differences between PLS and more familiar methods, including regression, principal components analysis (PCA) and canonical correlation analysis (CCA). Because of the potentially wide range of applications of this method, it is important to understand how it is related to other methods that address similar questions. As well as understanding the relationships of PLS to other methods it

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

is also important to understand the limitations of PLS, especially when another method might be equally appropriate (both conceptually and mathematically).

Before exploring the relationships among methods, we should note that partial least squares analysis employs a mathematical technique called singular value decomposition (SVD), which has not been introduced yet in this text. SVD is related to the more familiar decomposition by eigenanalysis used to extract principal components (from the variance–covariance matrix) and partial warps (from the bending-energy matrix). Because PLS uses SVD, the vectors generated by PLS are often called *singular axes* (SAs); in studies of covariances among geometric shapes, they are also sometimes called *singular warps*.

PLS compared to regression

Both regression and PLS examine the relationship between two sets of variables, but they differ in that the (Model I) regression model casts one set of variables as dependent on the other whereas PLS treats them symmetrically. That is, PLS does not assume that one set of variables causes the other, but rather views both sets as jointly (and linearly) related to the same underlying causes. Linear combinations of measured variables that are thought to reflect responses to underlying (unobserved) variables sometimes are called "latent variables," and the terminology of latent variables often is used in PLS (in this, PLS resembles factor analysis).

PLS seeks the latent variables in one block that are maximally correlated with the latent variables of another block. Thus, there is assumed to be at least one latent variable within each block. These linear combinations are constructed to account for as much of the covariation as possible between the two blocks of variables. The method yields two sets of vectors, one consisting of the coefficients of the latent variable underlying the first block, the other of the coefficients of the latent variable of the second block. The number of interblock linear combinations is equal to that of the smaller block, P_{min}. So, for example, if we have two blocks, one of which has four variables and the other of which has three, the first linear combination will have four coefficients, the other will have three, and there will be three interblock linear combinations.

Linear regression is based on a linear statistical model, which assumes that the independent variable is measured without error (hence all the error is ascribed to the dependent variable). No model underlies PLS, and no error is ascribed to any variables (in either block). For this reason alone, we would not expect to obtain the same coefficients from PLS as we obtain from regression. However, there is a more important reason for expecting that the two methods will yield different results when there is more than one independent variable. In that case, we would compare PLS to multiple regression, and the coefficients of PLS and multiple regression have very different interpretations. The vectors obtained by multiple regression express the dependence of the dependent variables on just one of the independent variables, *with all others held constant*. The coefficients produced by multiple regression do not assess the covariance between the two blocks of variables, a distinction that becomes particularly important when several independent variables are highly correlated. Under those conditions, most of the variance in the dependent variables will be associated with one independent variable, leaving little to be explained by the others. Even though *all* the independent variables might affect the dependent variables, only one would be accorded a high weight, so the others might appear to have trivial explanatory power.

That is because they are explaining the *residual* variance – i.e. the variance not already explained by the one with the large coefficient. Consequently, interpreting the coefficients of multiple regression can be difficult, and the method is poorly suited to cases in which the independent variables are uncorrelated with each other. In those cases, no latent variable accounts for covariances among the independent variables because they do not covary.

PLS treats the variables of both blocks symmetrically, and therefore we obtain variables within one block most relevant for predicting the variables in the other block and *vice versa*. These coefficients are called *saliences* because they indicate which variables in one block are most relevant (salient) for explaining covariation with the other block.

PLS compared to PCA

PLS and PCA greatly resemble each other in the definition of axes. As we saw in Chapter 7, PCs are extracted from a variance–covariance matrix (by eigenanalysis), producing a set of mutually orthogonal dimensions (eigenvectors) ordered according to the amount of variance each one explains. Similarly, PLS decomposes a matrix into mutually orthogonal axes, but the matrix is an *interblock* variance–covariance matrix and the components are ordered according to the amount of *covariance* between blocks explained by each one. The mathematical difference is that, instead of using an eigenvalue decomposition of the variance–covariance matrix, PLS uses a SVD of the interblock variance–covariance matrix. The reason for using SVD instead of an eigenvalue decomposition is that the covariance matrix between blocks need not be a square, symmetric matrix (a square, symmetric matrix is one in which the number of rows equals the number of columns, the first row of the matrix is the same as its first column, and every other row is also the same as its corresponding column). Variance–covariance matrices are always square and symmetric, but interblock variance–covariance matrices need not be (because the numbers of variables in each block can differ). Therefore, we need a different method to decompose the matrix. SVD yields pairs of singular axes (SAs), one per block. Each pair is associated with a *singular value* (SV), which is a relative measure of the covariance explained by the paired axes (we should note that singular axes “explain” covariance in the same sense that principal components “explain” variance). Consequently, one of the primary differences between PCs and SAs is that SAs come in pairs. For each singular value there is a pair of axes that, taken together, accounts for the patterns of covariances between blocks.

Just as we can calculate scores for individuals on PCs, and explore the patterns of variance in their plots, we can also calculate scores for individuals on SAs and explore the patterns of covariance between blocks in their plots. Scores on SVs are computed the same way as scores on PCs (i.e. by the dot product between either the PC or SA and the data for an individual). Also, SAs, like PCs, can be depicted graphically by the deformation along an axis, aiding the interpretation of their biological meaning. In the case of SAs, we would plot SV1 for one block against SV1 of the other block. When one of the blocks is not a set of landmark coordinates, no deformation can be drawn for the associated SA, but we can still interpret the axis using the loadings of the variables on it. Rather than having a picture, we will have a list of numerical values expressing the correlation between the variable and the SA. The plots of the scores, as well as the depictions of shape transformations or numerical values, provide the information about the nature of the covariance between blocks.

Unlike the situation for PCA, there is no analytic statistical test of the significance of SAs, nor for the significance of the correlation between blocks. However, resampling-based approaches can be applied to test these hypotheses. A permutation test, discussed by Rohlf and Corti (2000), determines if the singular values are larger than could be produced by a random permutation of associations among variables between blocks (keeping within-block associations intact). We can ask whether the covariances between blocks exceed those we would expect by chance. We can also ask if the correlation between singular axes is significant using a permutation test – this determines whether the correlation between the scores for each block exceeds what we would expect by chance. Both tests indicate whether the observed patterns of covariance between blocks are statistically significant.

The similarity between PLS and PCA is important to understand because both impose a similar constraint on the analysis: both define axes to be mutually orthogonal. SV2 is defined to be orthogonal to SV1, just as PC2 is defined to be orthogonal to PC1. This becomes important when biological factors are not orthogonal, which may be the general rule. Even though the axes (both PCs and SAs) provide a useful, simplified space in which to explore patterns in the data, the axes themselves need not correspond to any biological factors. It is likely that PC1 and SA1 have a biological interpretation when they account for a very large proportion of the variance or covariance, but the remaining axes are, by definition, constrained to be orthogonal to them, making their interpretation more dubious. This same issue arises when using PCA for explanatory or even comparative purposes (see Rohlf and Corti, 2000, pp. 747–748; Houle et al., 2002). It is possible that no useful (interpretable) axes will emerge from the PLS analysis, and that no significant correlations between blocks will be found, particularly when the structure of the variation–covariation *within* each block is especially complex.

Another important similarity between the methods, which should also inspire a cautious approach to interpreting results, is that PLS extracts *linear* combinations of variables (like PCA), but the relationship between blocks may be non-linear. In such cases, the first dimension may represent the dominant linear trend, and others represent orthogonal deviations from linearity. Thus, we would need to interpret SV1 together with SV2 to understand the relationship between the two blocks, recognizing that a single non-linear factor accounts for both. Of course, the issue of linearity is also important whether we are analyzing the data by PCA/PLS, by regression, or by the method discussed in the following section, CCA. However, most workers recognize that linearity is an important assumption of regression; non-linearity might not seem so important in studies using PCA or PLS because neither method is explicitly based on a linear model, so the impact of non-linear relationships among variables might not seem to violate assumptions of the method.

PLS compared to CCA

Canonical correlation analysis, like multiple regression and PLS, examines the correlation between blocks of variables. CCA closely resembles multiple regression, although, like PLS, both blocks are treated symmetrically (there is no presumption that one block of variables comprises causes and the other comprises effects). Nevertheless, the coefficients produced by a CCA are interpreted like partial regression coefficients. This means that each coefficient indicates the contribution made by an independent variable *when the effects of* *the others are held constant*, as discussed above in the context of regression. In this way, CCA, like multiple regression, differs from PLS.

CCA and PLS also differ in the quantity being maximized by the procedure. CCA seeks pairs of axes (canonical axes) that are maximally correlated *with each other*; in contrast, PLS seeks axes that maximally account for the covariance between the blocks (for a more detailed comparison between CCA and PLS, see Rohlf and Corti, 2000).

Multigroup PLS: using PLS to compare patterns of covariance

PLS is usually used to examine patterns of covariances between blocks of variables measured in a single sample, but we can also use it to compare those covariances between samples, as in a comparative analysis of morphological integration. Such comparisons rely on the same logic (and methods) used in comparative analyses of regression equations or PCs. In all of these, we are asking if the biologically corresponding vectors point in the same direction. To answer that question, we can compute the angle between comparable SAs, then test it statistically (using, for example, a bootstrapping procedure). In a similar fashion, we can also compare SAs to PCs, asking whether the major dimension of covariance between blocks is equivalent to the dominant dimension of variation within blocks. For example, when our data come from an ontogenetic series, the major dimension of covariance between blocks of variables may be their developmental correlations and the major dimension of variance within each block might also be explained by ontogeny. Comparing SAs to PCs can be especially useful for understanding causes of variance when PLS indicates a significant relationship between morphology and some collection of environmental variables; that same relationship may also explain the within-sample variation.

Mathematical details of two-block PLS

Suppose we have a matrix $\mathbf{Y}$ of P variables measured on N specimens, and that these observations can be split into two blocks, $\mathbf{Y}_1$ and $\mathbf{Y}_2$, which have P_1 and P_2 variables respectively. We can now compute the variance–covariance matrix, $\mathbf{R}$, which can be thought of as comprising the variance–covariance matrices within blocks $\mathbf{Y}_1$ and $\mathbf{Y}_2$ ($\mathbf{R}_1$ and $\mathbf{R}_{2,}$ respectively) and the covariance matrix between the two blocks $\mathbf{R}_{12}$, giving:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 & \mathbf{R}_{12} \\ \mathbf{R}_{12}^{\mathbf{t}} & \mathbf{R}_2 \end{bmatrix} \tag{11.1}$$

where $\mathbf{R}_{12}^{\mathbf{t}}$ is the transpose of $\mathbf{R}_{12}$. We then perform a SVD of $\mathbf{R}_{12}$:

$$\mathbf{R}_{12} = \mathbf{USV}^{\mathbf{t}} \tag{11.2}$$

which produces a $P_1 \times P_2$ diagonal matrix $\mathbf{S}$, whose diagonal entries are the P_{min} singular values, λ_i. As mentioned above, there are as many singular values as there are variables in the smaller block (P_{min}). The matrices $\mathbf{U}$ and $\mathbf{V}$ have dimensions $P_1 \times P_{min}$ and $P_2 \times P_{min}$, respectively; their columns are called the singular axes (SAs). The first columns of $\mathbf{U}$ and $\mathbf{V}$ comprise the paired SAs corresponding to the first singular value λ_1, just as the first principal component (PC1) is the axis corresponding to the first eigenvalue of the variance–covariance matrix. The SAs are ordered by decreasing singular values, just as PCs are ordered by decreasing eigenvalues. Also, as mentioned above, scores on SAs are calculated just like scores on PCs, i.e. by taking the dot product between an SA and the data for a specimen (scores are calculated for each block separately). The fraction of the total covariance of the two blocks expressed by the *i*th pair of singular axes is given by:

$$\frac{\lambda^2}{\sum_{j=1}^{P_{min}} \lambda_j^2} \tag{11.3}$$

Whether a singular value is larger than we would expect from randomly related blocks is determined by comparing the observed singular value to the distribution produced by randomly permuting the covariance structure between blocks. In such a permutation test, the vectors of P_1 observations, each representing a specimen in the first block, are randomly associated with vectors of observations from the second, thereby randomizing the covariance structure between blocks without altering the variance–covariance structure within the blocks. If the observed singular value lies outside the 95% confidence interval obtained from the permuted data sets, the observed SA is judged to be statistically significant. The correlation between the scores on the two blocks on the *i*th SA is also a measure of the statistical significance of the axis, and this correlation may also be tested via a permutation test in exactly the same manner.

Using PLS to examine ontogenetic integration between cranial and postcranial regions

One of the most promising applications of PLS is in studies of morphological integration, although, for the reasons discussed in the introduction to this chapter, it does not fully address one of the basic questions asked in such studies: which parts comprise integrated units or modules? PLS takes the blocks as a given rather than testing the hypothesis that a given block is, in fact, a coherent unit. The value of PLS lies in its ability to test a different sort of hypothesis – the integration *between* blocks. Granting its limitations, the method can be useful for analyzing relationships among morphological parts that are chosen *a priori*.

Studies of morphological integration usually focus on covariances among variables measured in a single homogenous sample (for studies of this sort using PLS, see Hingst-Zaher et al., 2000; Bastir et al., 2004). It also has been used to analyze the relationship between blocks in heterogeneous samples such as ontogenetic series and purported evolutionary lineages (e.g. Bookstein et al., 2003). Used in that second way, PLS examines the covariance between blocks over age, or “evolutionary divergence” or “geological time.”

When applied to studies of heterogeneous samples, regression is often a feasible alternative to PLS – so long as the factor explaining the heterogeneity of the sample is separately measured. For example, if the sample comprises an ontogenetic series, the dominant factor explaining the heterogeneity of the sample is age. Age is not a factor in a causal sense, but it explains the variation in shape within the sample. Because studies of integration are usually concerned with hypotheses of causality, age is not an explanation for morphological integration. However, often we want to know whether two or more structures are integrated through their ontogenies, and this is the sort of question we can answer using either regression or PLS. Using PLS to explore the integration of ontogeny is an alternative to describing that same ontogeny by multivariate regression. Regression does not require breaking up the data into multiple blocks *a priori* (and that subdivision is one of the more problematic features of PLS), but it does require that we measure age (or a proxy for it, like size). Furthermore, regression does not fully answer the question about integration because it only tells us whether shape covaries with age; it does not tell us which blocks most covary with each other. The two questions, while at least subtly different, are also interrelated – the block that covaries least with the others also covaries least with the measured factor. Considering that regression is a feasible alternative to PLS for some questions, it is important to understand how the two methods differ when describing the same ontogeny. Additionally, because age might account for most of the variance in the sample, it is also important to understand how PLS might differ from PCA.

We will use PLS to address three sorts of questions about the ontogenetic integration between cranial and postcranial morphology:

1. How are cranial and postcranial regions integrated over ontogeny?
2. Do species differ in these patterns of ontogenetic integration?
3. Are cranial landmarks more highly integrated with those that measure the position of median fins, or with those of the caudal peduncle?

A major objective of the first two analyses is to compare results based on PLS to those based on regression and PCA. The primary objective of the third is to explore the use of PLS to test explicit and conflicting hypotheses about integration.

Ontogenetic integration between cranial and postcranial shapes: results from PLS, PCA and regression

The landmark configuration and its subdivision into two blocks are shown for the data of *Pygopristis denticulata* (Figure 11.1). The two blocks do not correspond to halves of the body, because the pectoral fin landmark (landmark 11) is topographically part of an "anterior" block but is included in the postcranial block because it is not a cranial feature. Thus, the cranial and postcranial blocks partly overlap on the body. Analyzed by PLS, we find that the covariance between blocks is substantial; the first singular value (0.0536) explains 67.1% of the total covariance, significantly more than expected by chance ($p < 0.01$). The correlation between the two blocks is a very high 0.862, and that correlation is also significant ($p < 0.01$). No other singular value is statistically significant, so there is only one dimension of covariance between the blocks. The plot of the scores of postcranial SA1 on cranial SA1 shows the pattern we expect from such a high correlation, although the relationship between blocks is not strictly linear (Figure 11.2).

Having found that there is a dimension of significant covariation, we need to display and interpret it (Figure 11.3). To aid in this interpretation, we first ask whether the SA1 of each block is correlated with size as expected – and they are; both cranial SA1 and postcranial SA1 are indeed highly correlated with size ($R = 0.86$, $R = 0.72$, respectively). Size thus appears to be a plausible interpretation of the biological factor responsible for the covariance between regions. Given that result, we might expect to find a similar depiction

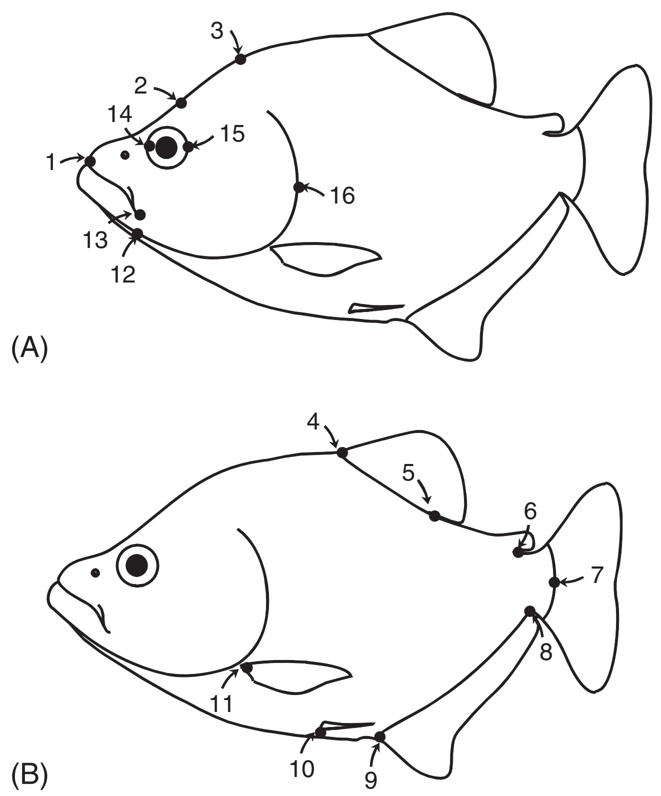


Figure 11.1 Subdivision of landmarks into two blocks: (A) cranial configuration; (B) postcranial configuration.

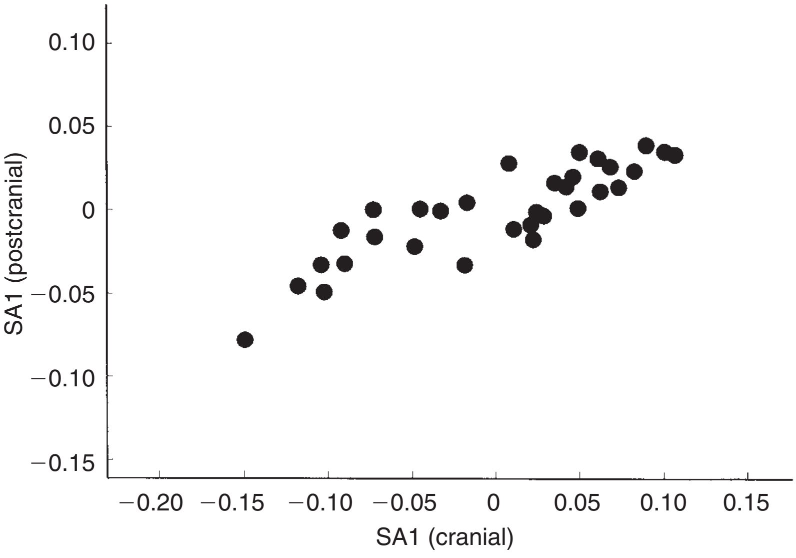


Figure 11.2 Scores of postcranial SA1 on cranial SA1 for *P. denticulata*.

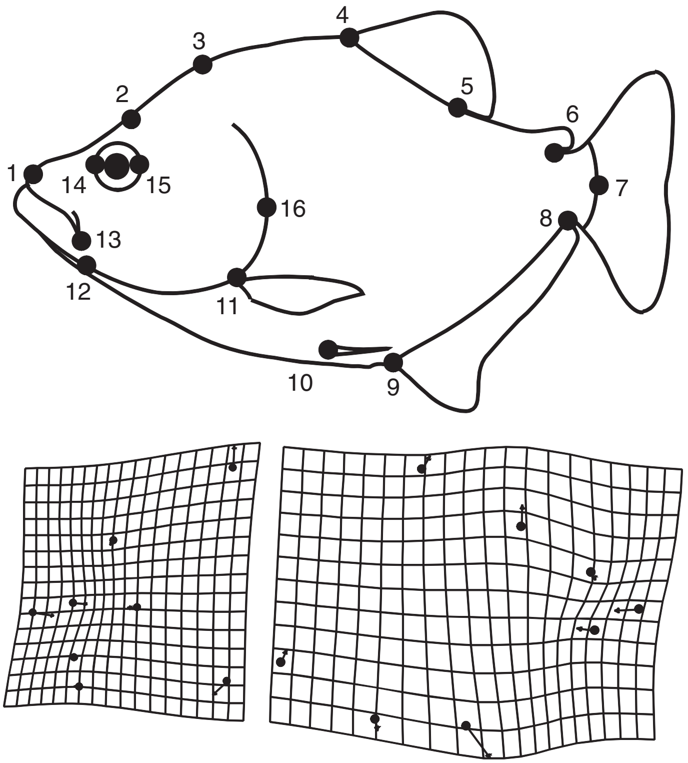


Figure 11.3 Singular axis expressing the covariance between cranial and postcranial landmarks of *P. denticulata*.

of ontogenetic integration using regression because that technique allows us to find the shape variable correlated with size, which ought to explain the covariance between cranial and postcranial landmarks. We would also expect PCA to yield the same result because size is the dominant component of the variance in an ontogenetic series.

Despite such apparently realistic expectations, we find that regression and PCA give similar results (Figures 11.4A, 11.4B), but they appear to differ from those yielded by PLS (Figure 11.4C). However, the visual comparison of graphics is complicated for three reasons. First, the analysis by PLS examines two partly overlapping parts rather than a single whole, in contrast to PCA and regression. Second, when analyzed by PLS, each half is separately scaled to unit centroid size rather than the whole, so the relative sizes of each block are portrayed differently in the two analyses. Third, when analyzed by PLS, the magnitude of change within each part is calculated and scaled separately rather than in relation to the whole as done by methods that analyze the entire configuration of landmarks comprising both blocks.

Proceeding with the visual comparison nonetheless, one apparent difference is the balance between cranial and postcranial changes. Regression and PCA (of the whole fish) both suggest that there is a large change in the orbital region relative to the change of the posterior body, but PLS suggests that they are more equally balanced. A more troubling difference is that the posterior covariate of size found by regression or PCA looks quite

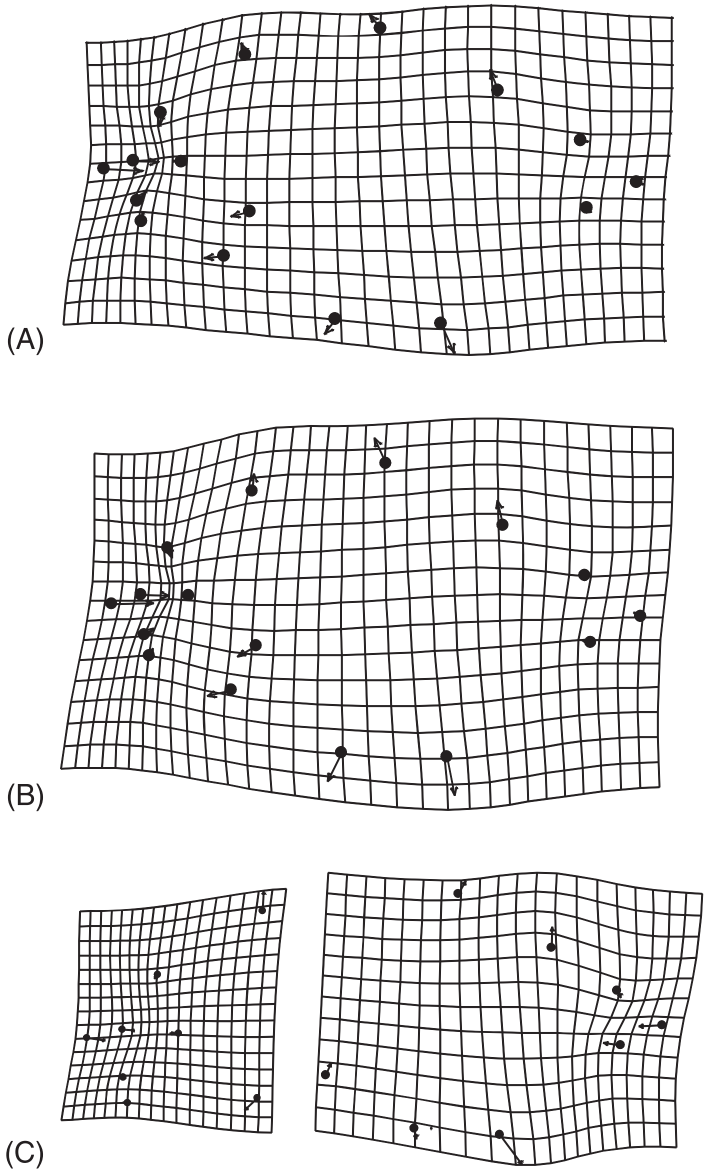


Figure 11.4 Comparing the direction of ontogenetic change in *P. denticulata* as determined by: (A) regression of whole body shape on size; (B) PC1 of whole body shape; (C) SA1 of the axis of covariance between cranial and postcranial shape extracted by PLS.

different than posterior SA1 – it appears that the most caudal region (the caudal peduncle) is integrated with the posterior anal fin, and more change is localized here (relative to the remainder of the posterior body).

The consequence of analyzing each half separately can be appreciated by regressing each block separately on size, and also by using PCA to extract the dominant component of variance of each separate half. Regression of each part separately on size yields a result

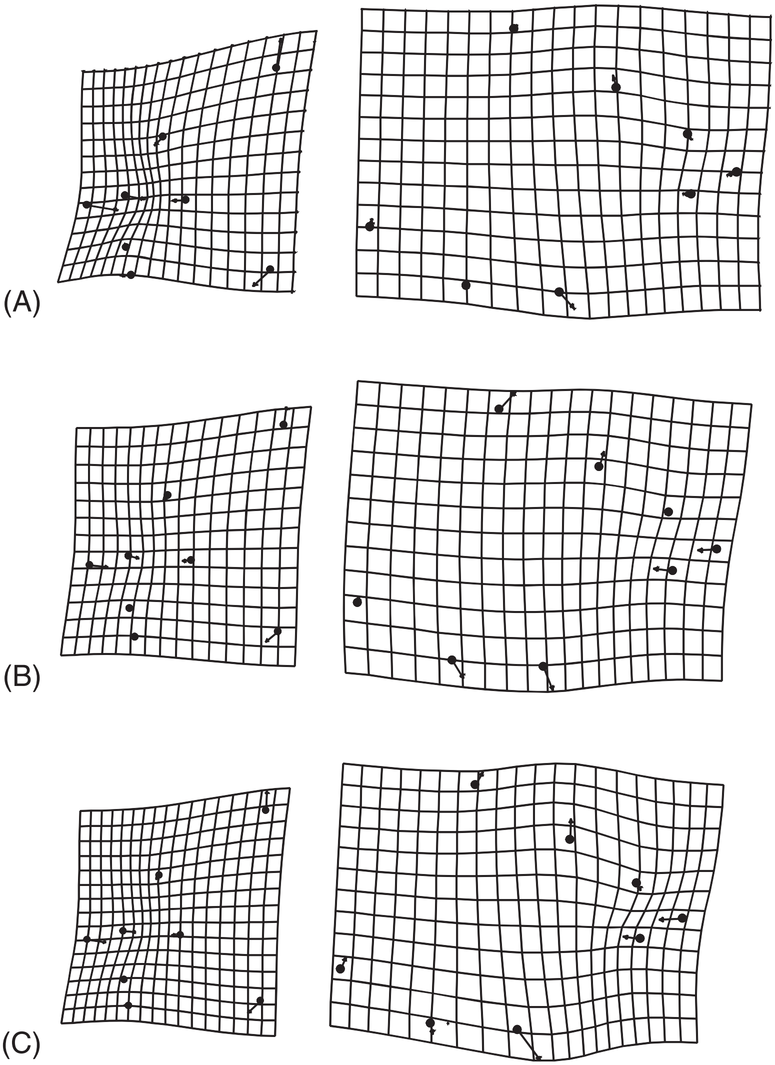


Figure 11.5 Comparing results of three methods for analyzing the two blocks of landmarks: (A) regression of *P. denticulata* cranial and postcranial landmarks separately on size; (B) PC1 of cranial and postcranial landmarks analyzed separately; (C) cranial and postcranial SA1.

very similar to that of regression of the whole (compare Figures 11.4A and 11.5A). Each block appears to be nearly linearly related to size, and the correlations between each block and size equal the correlations between each SA1 and size (i.e. 0.89, 0.72 for cranial and postcranial blocks, respectively). There is, however, a notable difference between the two blocks in the magnitude of change in relation to size. Measuring the rate at which the Procrustes distance away from the smallest specimen increases with size gives strikingly different estimates for the two blocks: 0.08 for the cranial block and only 0.03 for the postcranial block (Figure 11.6). Thus, over a given change in size, the cranial region undergoes far more change than the postcranial body. Such information is captured by

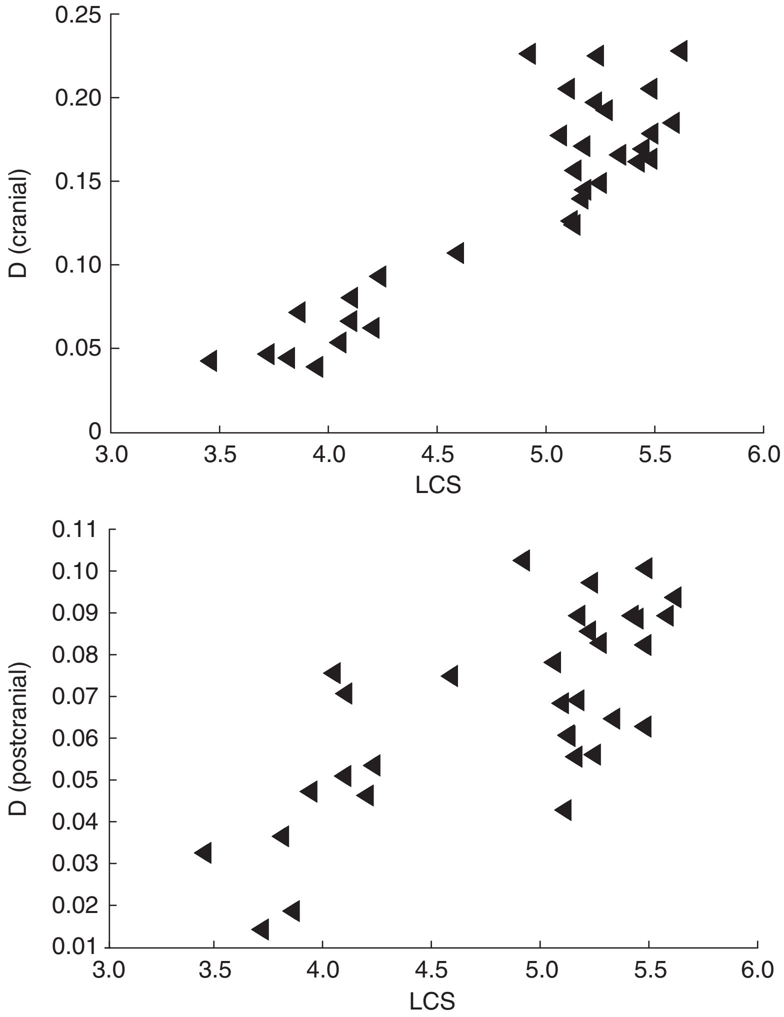


Figure 11.6 Relative rates of cranial and postcranial development analyzed by regressing the Procrustes distance away from the average of the smallest specimens (D) against log centroid size (LCS).

regression even when the parts are analyzed separately, because the changes are calibrated in relation to size. Going back to the plot of the SA1 scores (Figure 11.2), it is now possible to appreciate the effect of these different magnitudes of response by noting the larger range of scores for cranial SA1 scores compared to postcranial SA1 scores.

PCA of the separate blocks yields one distinct eigenvalue for the cranial landmarks and none for the postcranial block, which is not surprising in light of their different magnitudes of change. Because of the larger ontogenetic change in the cranial region there is a larger distance between the shapes of the smallest and largest specimens, and consequently variation is more elliptical because ontogeny produces a long axis of variation (correlated with size). In contrast, the ontogenetic change of the postcranial region is subtle, so the distance between the smallest and largest specimen is not as large, and the postcranial variation is

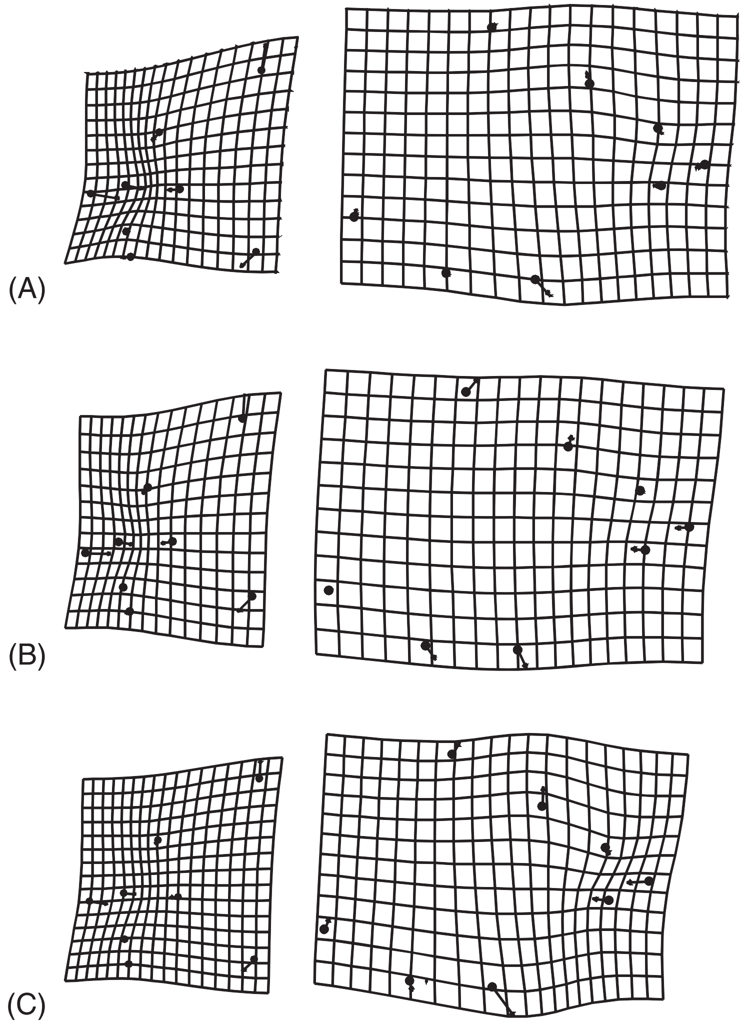


Figure 11.7 Comparing results of three methods for analyzing the two blocks of landmarks, with PLS results rescaled in light of relative rates of development. (A) Regression of *P. denticulata* cranial and postcranial landmarks separately on size; (B) PC1 of cranial and postcranial landmarks analyzed separately; (C) SA1 of cranial and postcranial shape.

more nearly spherical. There is no dominant size axis, although size is still a factor explaining variation within the postcranium. When analyzed as part of the whole, we can see the impact of size on the postcranium in the context of its effect on the cranium, an effect that becomes more ambiguous when the blocks are analyzed separately (compare Figures 11.4B, 11.5B). However, that ambiguity is again partly due to the greater magnitude of change undergone by the head.

We can rescale the plots of SA1, amplifying the deformation of the head to reflect its greater developmental rate (Figure 11.7C). That does not fully reconcile the graphical results of regression, PLS and PCA, but it removes one major discrepancy among them. The

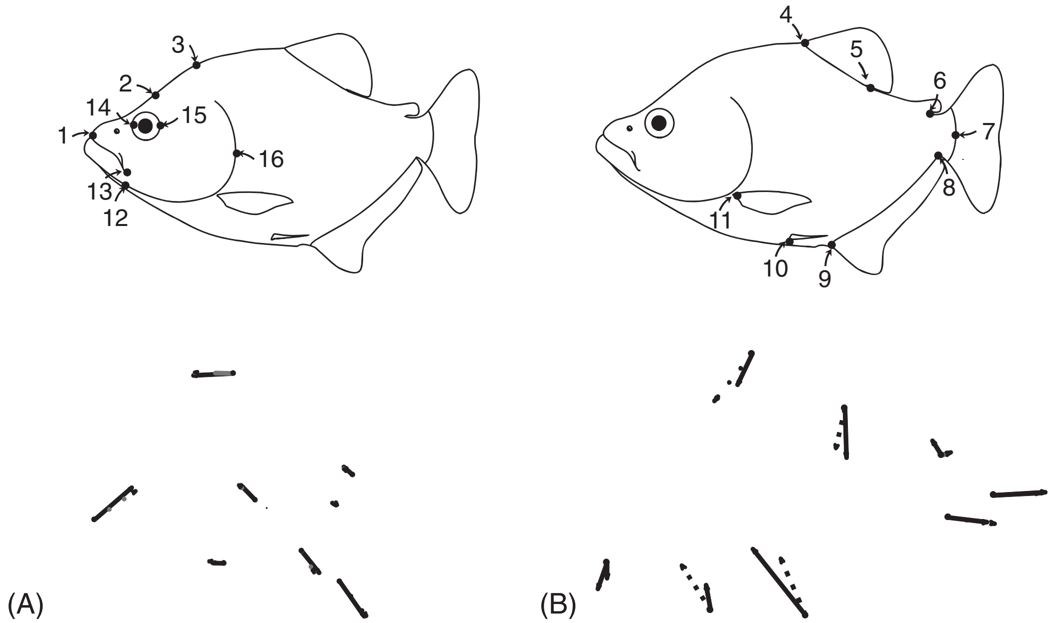


Figure 11.8 Comparing SA1 (solid lines) to PC1 (dotted lines).

changes within the head appear to be similar now, regardless of method. The similarity between PCA and PLS is even more evident if we diagram PC1 and SA1 on the same plot (Figure 11.8). The difference is most striking for the posterior block, where it is most notable in the orientation of the vectors at the two anal fin landmarks (landmarks 8 and 9).

To check the generality of the conclusions we drew from the analysis of *P. denticulata*, we can (more briefly) analyze two other species. One, *S. gouldingi*, differs from *P. denticulata* in that there are two distinct eigenvalues for the variation of the head (rather than one) and there is one for the variation of the postcranial landmarks (rather than none). In the analysis of *S. gouldingi* the first singular value is very high (0.2379), accounting for 89% of the covariance between the two blocks, and the correlation is also remarkably high (0.968). Not surprisingly, the covariance explained by the paired SA1 axes is significant ($p < 0.01$), as is the correlation ($p < 0.01$). No other SA explains more covariance than expected by chance, so these data, like those of *P. denticulata*, produce a one-dimensional solution. Having already found that differences in relative rates of development between the blocks can produce apparent discrepancies among the results of the three methods, we show the results taking the difference in cranial and postcranial developmental rates into account. The primary discrepancy among the results of the three methods is in their descriptions of posterior head deepening (Figure 11.9). There is virtually no difference between the results of PCA and PLS, to the point that we cannot visually compare them by superimposing the two sets of vectors on the same plot – they entirely overlap each other. The results of PCA and PLS do differ from those of regression, albeit subtly.

The analysis of *S. manueli*, like that of the other two species, yields a single significant dimension of covariation between blocks. SA1 accounts for 76% of the covariance between

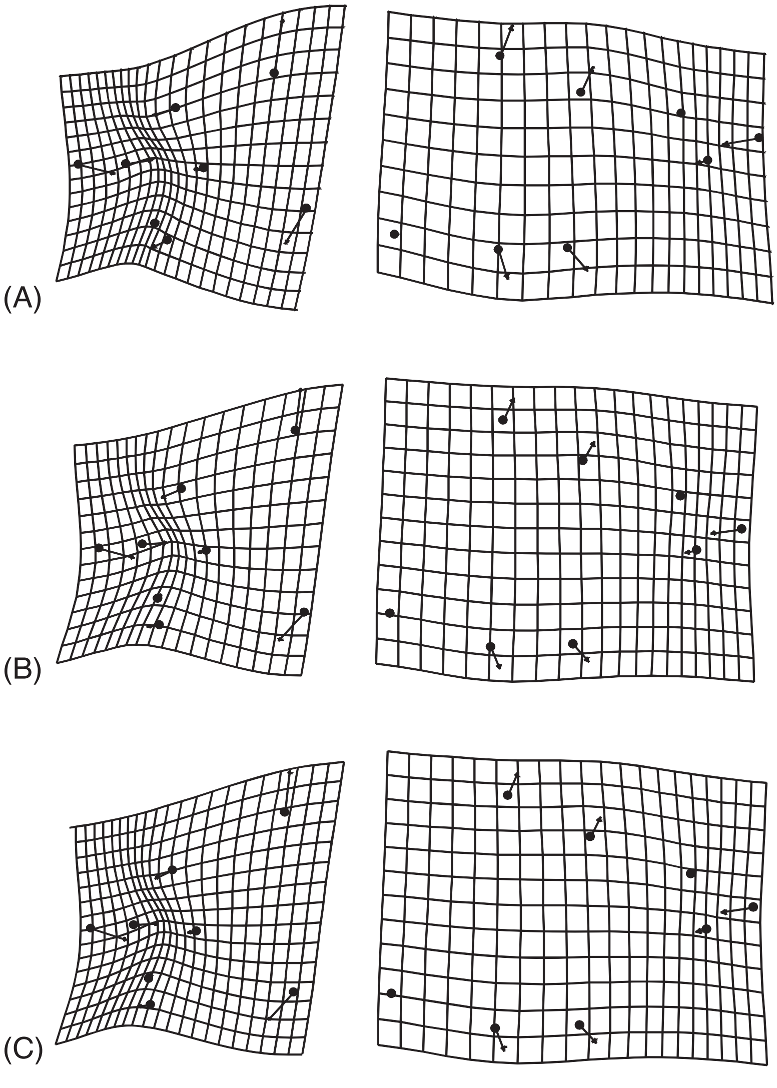


Figure 11.9 Comparing results of three methods for analyzing the two blocks of landmarks after rescaling plots to reflect the magnitude of the change of each block relative to a unit change in size. (A) Regression of *S. gouldingi* cranial and postcranial landmarks separately on size; (B) PC1 of cranial and postcranial landmarks analyzed separately; (C) cranial and postcranial SA1.

blocks, which is significantly greater than expected by chance ($p < 0.01$). The correlation between blocks is 0.92, which is also significant ($p < 0.01$). In drawing the SAs, we again scale the magnitudes of the cranial and postcranial blocks in accordance with their relative rates of development (Figure 11.10). As in the case of *S. gouldingi*, regression provides a somewhat different picture of the ontogenetic change in shape than do PLS and PCA, but the results of PCA and PLS are consistent with each other.

The possibility that the three methods can give different results underscores the importance of deciding which methods *ought* to be used.

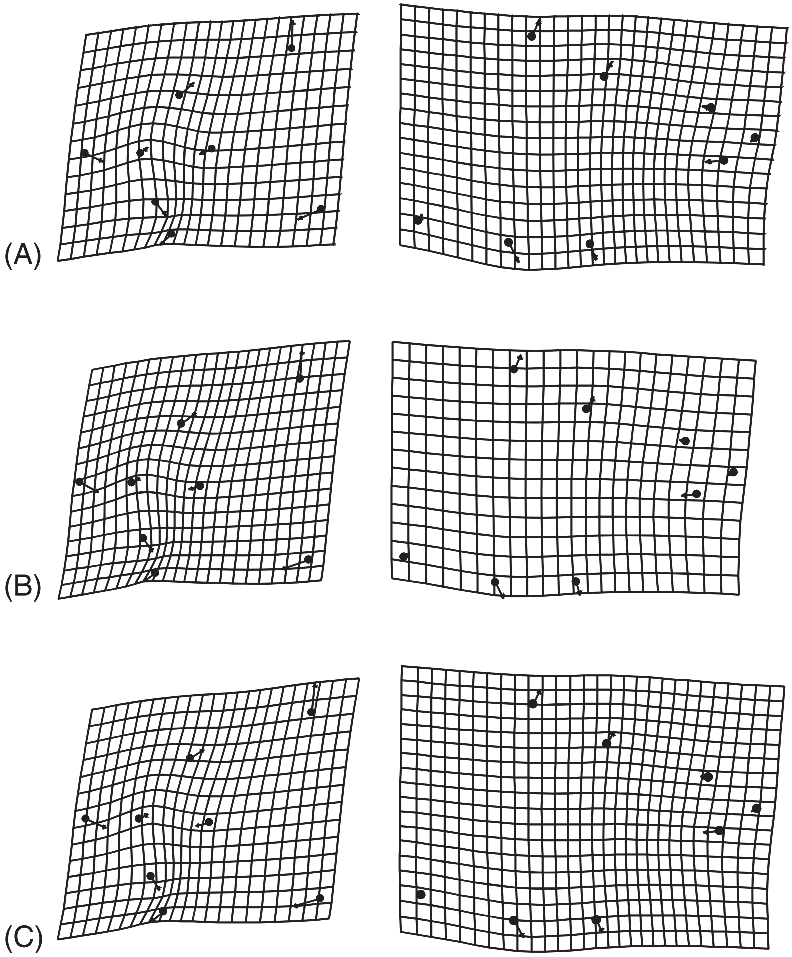


Figure 11.10 Comparing results of three methods for analyzing the two blocks of landmarks after rescaling plots to reflect the magnitude of the change of each block relative to a unit change in size. (A) Regression of *S. manueli* cranial and postcranial landmarks separately on size; (B) PC1 of cranial and postcranial landmarks analyzed separately; (C) cranial and postcranial SA1.

Interspecific comparisons of ontogenetic integration

We can use PLS to compare ontogenetic integration among species. This is another case where we could also use a regression-based approach (as we did in Chapter 10), so we focus on the distinction between the results of the two methods. To make the analyses as similar as possible, we subdivide the landmarks into cranial and postcranial landmarks in the analyses based on regression, just as we do for the analyses based on PLS.

In the comparison between *P. denticulata* and *S. gouldingi*, the interspecific angle between cranial SV1s is small (13.8°), suggesting that these species are virtually indistinguishable in cranial ontogeny; not surprisingly, that angle is not statistically significant. More surprisingly, the larger angle of 30.9° between postcranial SA1s is also not significant (owing to the larger range of angles obtained by resampling within *S. gouldingi* – 32.8°). Comparing the cranial regressions also yields a modest angle (12.8°) that is not statistically significant. However, the interspecific angle between postcranial ontogenies is larger (44.0°), and this is statistically significant. Thus we can conclude that the two species share a common cranial ontogeny, but the results for the postcranial landmarks are more ambiguous. Based on PLS, we cannot say that the patterns of postcranial integration differ between species. However, if we analyze the relationship between the postcranial landmarks and size, we do find a significant difference between species. In this case, it is important to decide whether the hypothesis ought to be formulated in terms of PLS or regression.

In the comparison between *P. denticulata* and *S. manueli*, we find a relatively large angle of 39.2° between SA1 of the cranial landmarks, which is statistically significant, and an equally large angle for the postcranial landmarks of 40.2°, which is also significant. Turning to the comparison of their cranial ontogenetic allometries based on regression, we find an interspecific angle of 45.4° between cranial landmarks, and an angle of 51.1° for the postcranial landmarks. In this case, both methods detect statistically significant differences between species in both blocks of landmarks. Similarly, in the comparison between *S. manueli* and *S. gouldingi*, the results from both methods are consistent. Comparing cranial SV1s between species yields an angle of 43.4°, which is statistically significant, and a comparably large angle between postcranial SV1s of 30.3°, which is also significant. The analysis based on vectors of allometric coefficients yields an interspecific cranial angle of 46.8° and postcranial angle of 32.7°, both of which are statistically significant. Unfortunately we cannot assume that the results will always be consistent, as they were not in the comparison between *P. denticulata* and *S. gouldingi*. Thus, as in the analyses of intraspecific integration, it is important to decide whether the analysis ought to be based on regression or on PLS.

Using PLS to test competing hypotheses of integration

Our objective now is to formulate competing hypotheses of integration and use PLS to test them. Specifically, we ask whether the integration between the cranial and caudalmost landmarks is greater than that between the cranial and median fin landmarks (the three blocks are depicted in Figure 11.11). We might expect that this would be the case, because the head and caudal body usually develop earlier than the midbody, and the deepening of the midbody occurs fairly late in development. Any factors, both genetic and environmental, that affect larval development are likely to affect both these cranial and caudal regions, but might have little impact on midbody depth (or anteroposterior locations of the median fins). Therefore, if the timing of development explains integration, we might anticipate a greater correlation between parts that develop at the same time. We will test this hypothesis using two species. In one, *S. gouldingi*, the general expectations appear to be met (for reasons that will not be evident until Chapter 13, when we discuss the relationship between allometric coefficients and developmental timing). In the other, the caudalmost part of the body seems to develop unusually late in relation to the head, so we would expect that this species would *not* evince greater integration between head and tail than between head and fin/midbody landmarks.

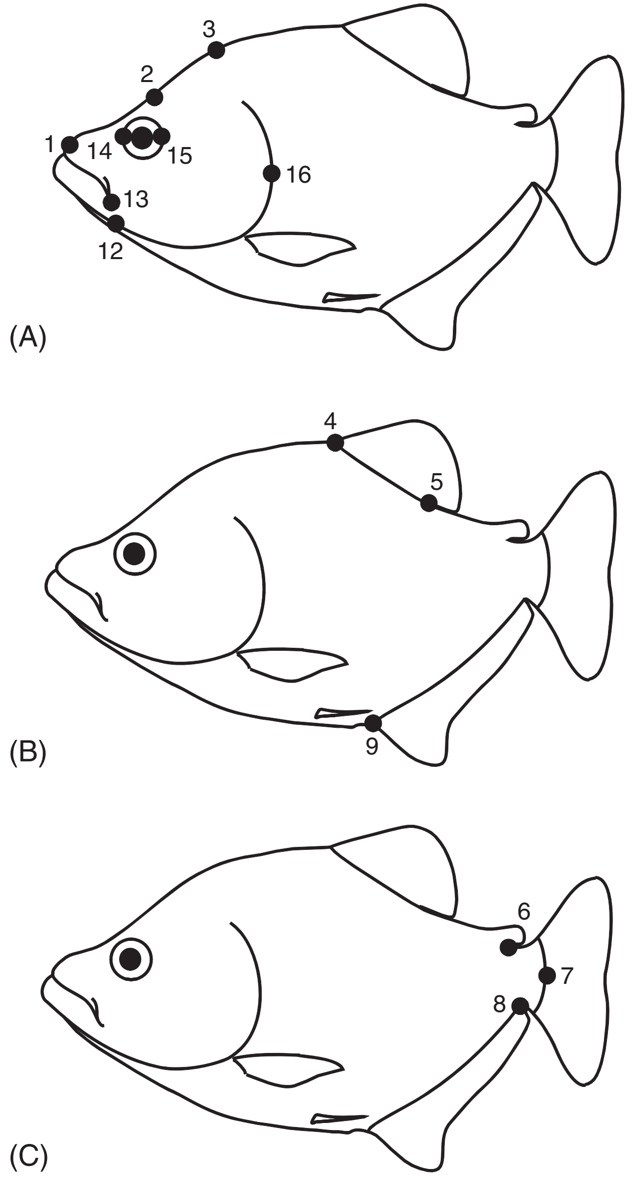


Figure 11.11 Landmarks subdivided into three blocks: (A) cranial; (B) midbody median fins; (C) caudal.

We thus separate the landmarks into three blocks: (1) cranial, comprising landmarks 1, 2, 3, 12, 13, 14, 15, 16; (2) midbody median fins, comprising landmarks 4, 5, 9; and (3) caudal, comprising landmarks 6, 7, 8 (Figure 11.11). Landmark 8 is the posterior base of the anal fin, so it might seem appropriate to include it in the median fin block; however, this landmark does not provide any information about body depth, distinguishing it from the landmarks included in that block. Landmarks 10 and 11 are on the paired fins; because they do not belong to any of the blocks singled out by this hypothesis, they are excluded from the analysis. We will examine the integration between the cranial landmarks and each of the other two blocks, producing two pairs of blocks with the cranial landmarks being included in both pairs.

The correlation between cranial and tail landmarks is very high in *S. gouldingi* ($R = 0.888$), as is the correlation between cranial and fin landmarks ($R = 0.751$). To determine if one correlation is higher, we need to determine whether the difference between the two correlations is larger than we would expect by chance. There are two ways to test this hypothesis: one is to compare the difference between the correlations to the standard error of the difference; the other is to bootstrap the difference between correlations and ask if the 95% range of the difference includes zero. Based on the first (analytic) approach we would reject the null hypothesis that the correlations are equal ($p = 0.011$), but, because the test presumes normality, we ought to check the result using a resampling-based method. The 95% range for the difference between correlations is −0.262 to −0.0342, which excludes zero and also leads us to reject the null hypothesis of equal correlations. Thus we conclude that cranial and tail landmarks are more highly integrated with each other than are cranial and median fin landmarks.

The analysis of *S. manueli* reveals a different pattern. In this species, both correlations are weak: that between cranial and tail landmarks is $R = 0.443$, and that between cranial and median fin landmarks is $R = 0.599$. The analytic test of the difference between correlations indicates that the two correlations differ by no more than expected by chance ($p = 0.251$), as does the resampling-based test (the 95% interval of R is −0.366 to +0.134, which includes zero). This contrast between integration patterns of *S. manueli* and *S. gouldingi* is expected in light of their different ontogenetic allometries; we would need to devise a timing hypothesis for *S. manueli* that does not follow the expected spatiotemporal pattern. An interesting subject to pursue further is whether we can use studies of allometry to infer patterns of timing that accurately predict developmental correlations among blocks.

Using PLS to relate shape to ecological factors

We now use PLS to examine the relationship between shape and a block of non-shape variables, specifically latitude and longitude. We will examine geographic variation in adult body shape in a widely distributed piranha species, *Pygocentrus nattereri*. Geographic variation is often analyzed using conventional ordination methods like PCA, so we will analyze the same data using both PCA and PLS. To simplify the analysis we restrict the sample to the northern populations, because the southernmost ones might belong to a different species (an inference difficult to make from morphology without a detailed analysis of geographic variation). Also, we exclude the smallest specimens to avoid confounding geography with ontogeny. We could conceivably include all the specimens, standardizing shape statistically (using the regression equation for shape on size), but that procedure assumes that deviations from the regression are equal across the entire size range. Violating that assumption could distort the covariance structure of shape, which could complicate the analysis of geographic variation. Therefore we limit this analysis to the 48 largest specimens of *P. nattereri*, which still includes considerable variation in size: individuals range from 102 mm to 225 mm standard length. Before we can interpret

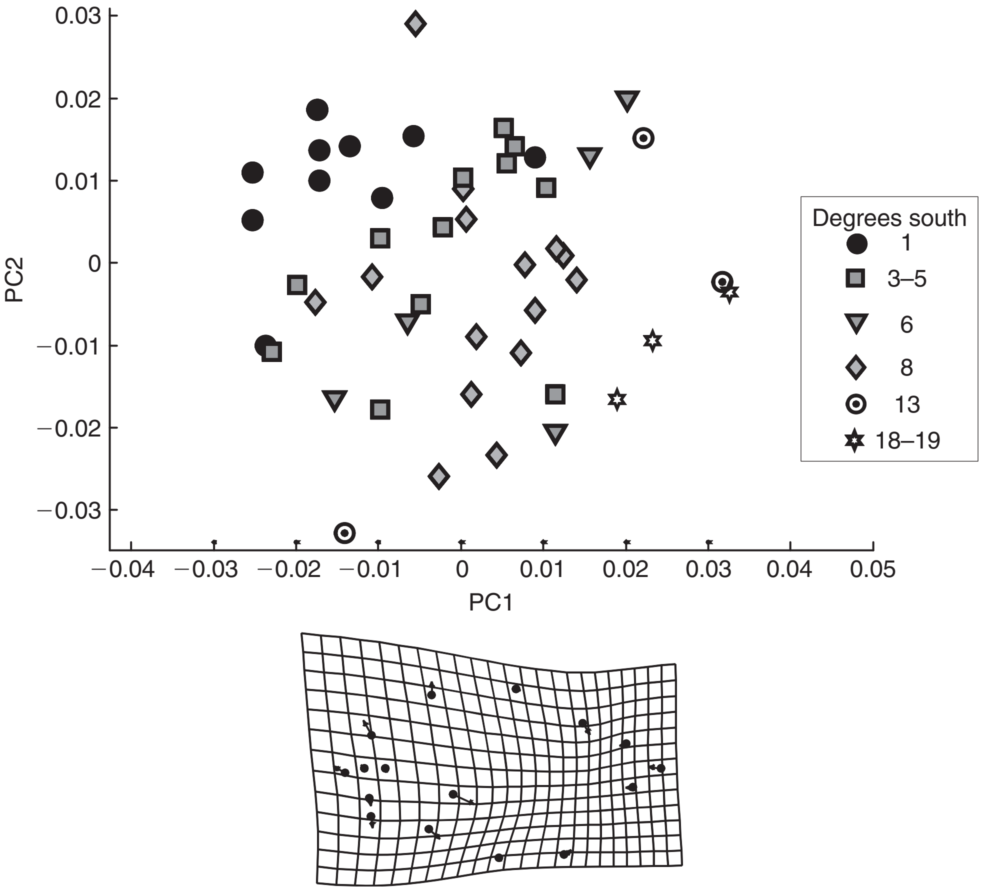


Figure 11.12 Principal components of shape variation of *P. nattereri*, specimens coded to indicate latitude where collected. The deformation grid depicts the shape transformation in the direction of higher scores on PC1, which is towards the more southerly localities.

the covariance between shape and geography, we will have to determine whether that covariance could instead be due to a covariance between size, shape and geography.

PCA of *P. nattereri* shape

The analysis of whole body shape by PCA yields no distinct eigenvalues: PC1 accounts for just 18.1% of the variance, PC2 accounts for 15.1%. Although the variation has no dominant direction, there is a slight hint of a geographic pattern in the scores on PC1 (Figure 11.12). Although the evidence is hardly compelling, the specimens collected in the northern localities are generally towards the left of the plot, whereas those collected in the most southern localities are generally towards the right. We thus tentatively explain PC1 as a latitudinal gradient in shape: southern *P. nattereri* tend to have larger heads for their bodies (especially relative to the most posterior body). However, the evidence is not compelling. The analysis of the cranial landmarks provides even less evidence for

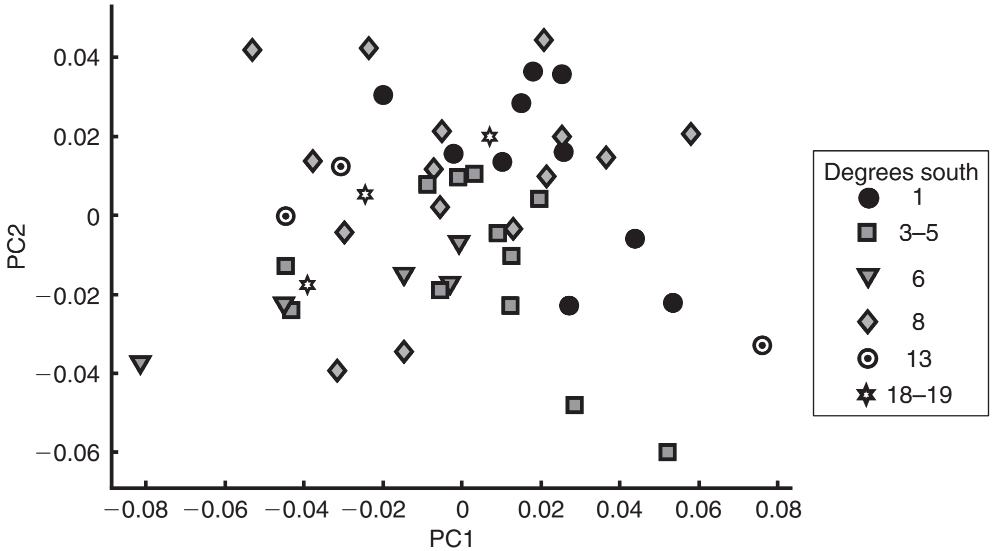


Figure 11.13 Principal components of cranial shape variation of *P. nattereri*, specimens coded to indicate latitude of the locality where the individual was collected.

any geographic trends (Figure 11.13). Again there are no statistically distinct eigenvalues, although PC1 accounts for 31% of the variance and PC2 for only 18.4%. Individuals from the northernmost locality appear to be scattered fairly evenly over the entire plane. Four individuals from the southernmost localities are concentrated to the left, but two other specimens from southern localities are towards the right. The variation in postcranial shape also lacks a dominant direction, although PC1 explains 28.9% of the variance and PC2 explains 16.2% of the variance (Figure 11.14).

Compared to cranial shape, postcranial shape offers a stronger hint of a geographic pattern because individuals from the northernmost localities are generally concentrated to the right of the plot of PC2 on PC1 whereas those from the southernmost localities are generally to the left. Based on Figure 11.13, we might thus interpret PC1 tentatively as indicating a latitudinal gradient in *postcranial* shape; more southern *P. nattereri* tend to have a longer dorsal fin base relative to the length of the posterior body.

PLS: the covariance between *P. nattereri* shape and geography

When we explicitly examine the covariation between shape and geography for whole body shape, PLS extracts one dimension of covariance (with a singular value of 0.0493) explaining 46.6% of the covariance between the two blocks. Interestingly, the one statistically significant singular value is the second; the first, with a value of 0.0564, does not explain more covariance than expected by chance. The loadings of the two geometric variables suggests that this is a longitudinal factor, and the reason why this may not be significant is that there is a small number of eastern Brazilian specimens that might differ from the more easterly populations (for reasons unrelated to a longitudinal trend). SA2 does account for significantly more covariance than expected by chance ($p < 0.01$), and the

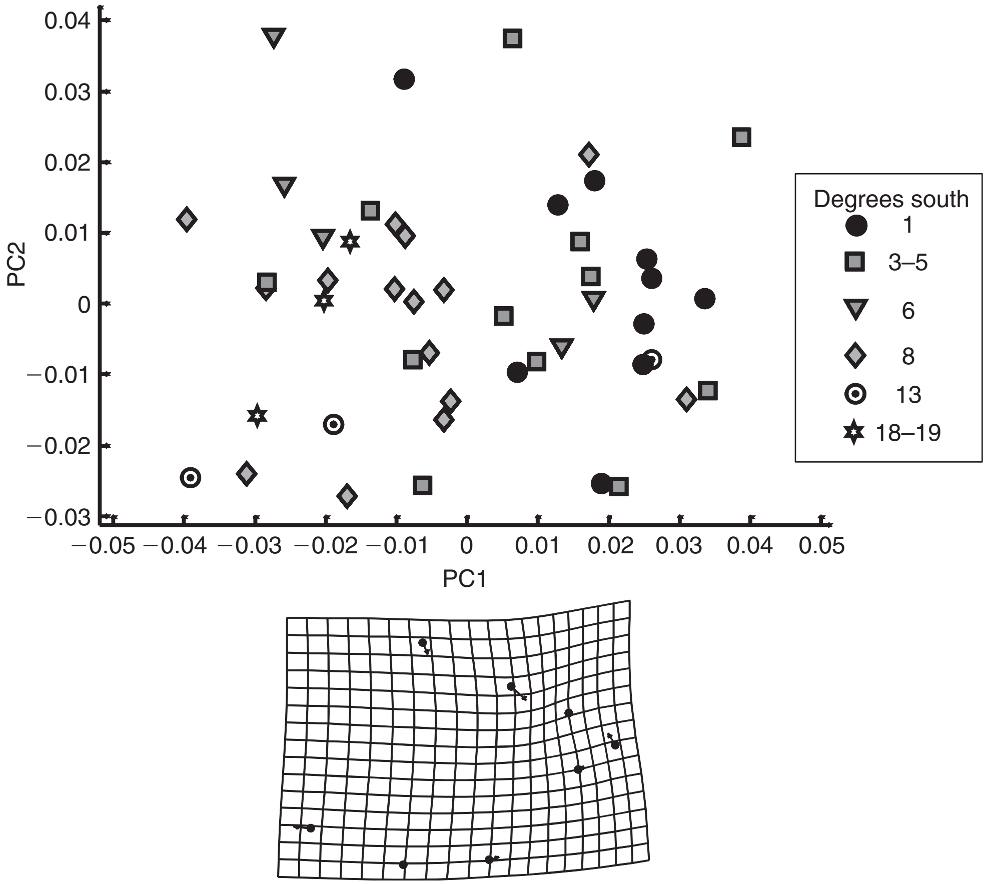


Figure 11.14 Principal components of postcranial shape variation of *P. nattereri*, specimens coded to indicate latitude where collected. The deformation grid depicts the shape transformation in the direction of lower scores on PC1, which is towards the more southerly localities.

correlation between the blocks (of 0.77) is also statistically significant. Before going any further, we need to check whether this correlation between geography and shape might be confounded by a correlation between shape and size. The evidence against this hypothesis is in the weak (non-significant) correlation between scores on SA2 and size ($R = -0.093$). Thus we can proceed to interpret SA2 as a possible geographic factor, bearing in mind that SA2 is defined to be independent of SA1, and SA1 is not significant (and may be heavily influenced by the easternmost specimens).

Only latitude makes a large contribution to SA2; its loading is 0.993 whereas that of longitude is only 0.117. Figure 11.15 shows the shape covariate of latitude (depicting the transformation in the whole shape moving southerly): a marked increase in head size relative to that of mid- and posterior body (especially relative to the region between dorsal and adipose fins), with steepening of the anterior head profile and shallowing of the more posterior head profile (producing a generally blunter head), and an expansion of the postorbital region.

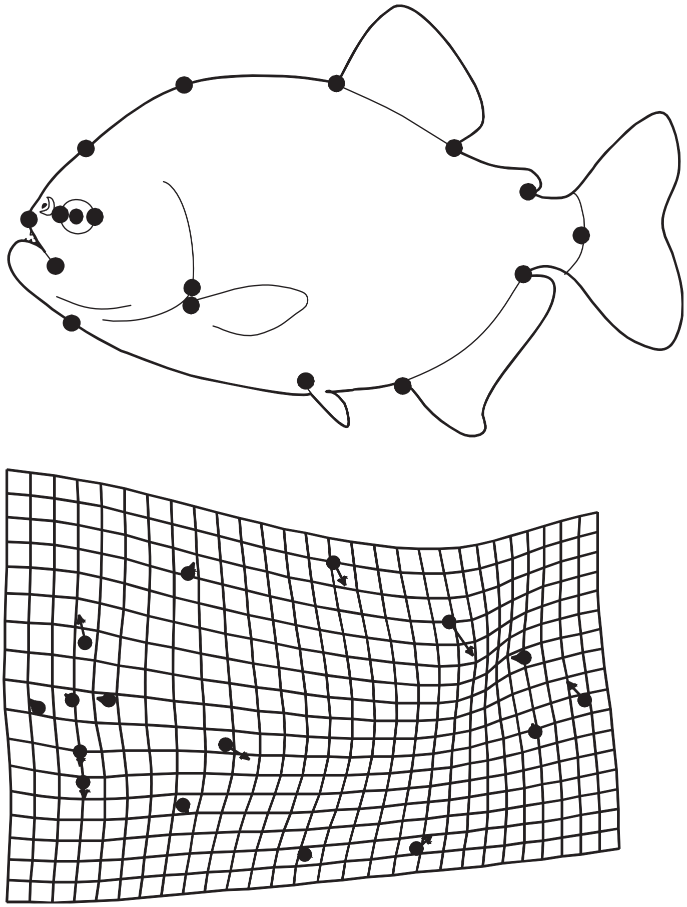


Figure 11.15 The covariance between *P. nattereri* shape and geography, depicted as a transformation in the southerly direction.

In the analysis restricted to cranial landmarks we again find that the second singular value, 0.0606, is significant. SA2 explains 31% of the covariance between geography and shape. The correlation of 0.59 is moderately high and statistically significant ($p < 0.01$). SA2 does not appear to be confounding size and geography, because the correlation between scores on SA2 and size is non-significant (-0.152; $p < 0.05$). Again the geographic factor is dominated by latitude: the loading of latitude on SA2 is 0.99, whereas that of longitude is only 0.16. Figure 11.16 shows the shape covariate of latitude (again depicting the transformation in shape when moving southerly). As in the analysis of whole body form, there is steepening of the anterior head profile and shallowing of the more posterior head profile (producing a generally blunter head), and also a slight lengthening of the postorbital head relative to eye diameter. Because the analysis is restricted to the cranial landmarks, we cannot see the general increase in head length relative to the mid- and posterior body.

In the analysis restricted to postcranial landmarks we find a strikingly different pattern, although the salience of geography for shape is still apparent. As before, there is only one

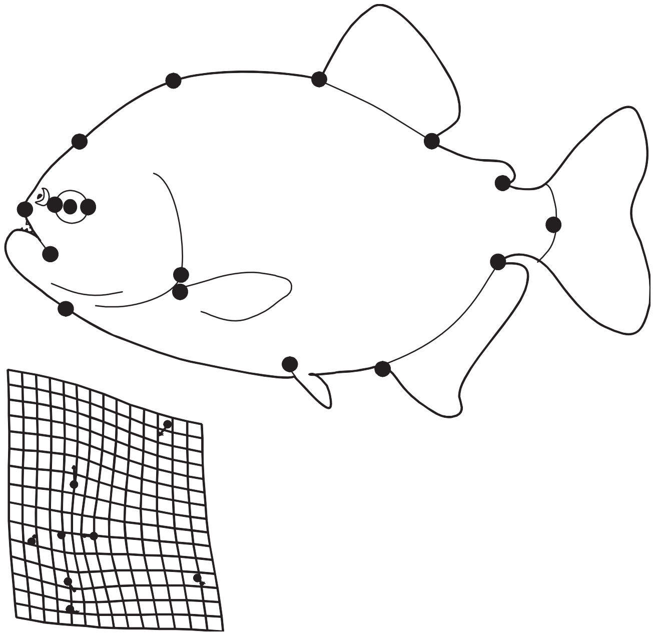


Figure 11.16 The covariance between *P. nattereri* cranial shape and geography, depicted as a transformation in the southerly direction.

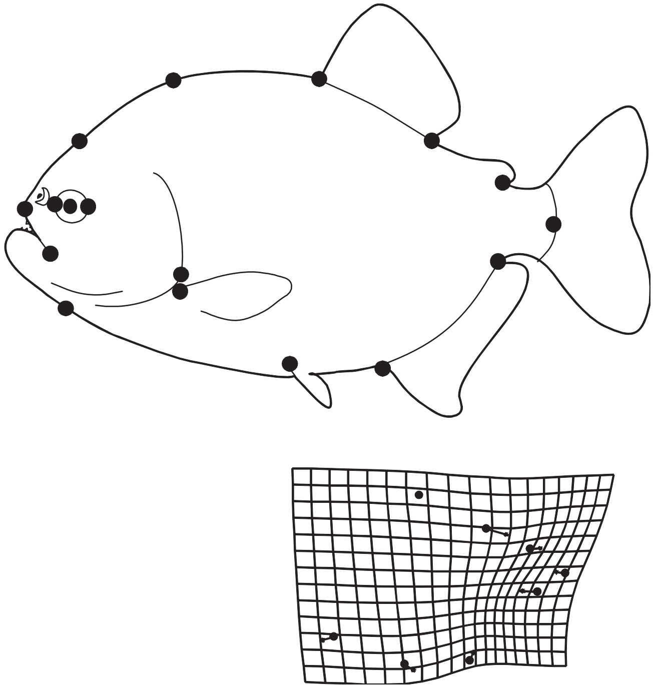


Figure 11.17 The covariance between *P. nattereri* postcranial shape and geography, depicted as a transformation in the southeasterly direction.

significant singular value; the second (0.050), which accounts for 46.4% of the covariance between shape and geography. The correlation between postcranial shape and geography is relatively low ($R = 0.40$), but it is an interesting composite of latitude and longitude. The loading for latitude is 0.617 and that of longitude is −0.717, so the two geographic directions are nearly equal and inversely related. Figure 11.17 shows the transformation in postcranial shape moving southeasterly: a shortening of the body posterior to the dorsal fin (especially of the anal fin), which is in contrast to a slight elongation of the region between pectoral and pelvic fins.

Software

Two programs in the IMP series perform partial least squares analysis. One is designed to analyze a single population (**PLSMaker**), as in the analyses of developmental integration within species and in the analysis of geographic variation of *P. nattereri*; the other (**PLSAngle**) is used for comparative studies, including comparisons between species and comparisons between correlations of blocks (as in the analysis of correlations between cranial and tail versus cranial and median fin landmarks in *S. gouldingi* and *S. manueli*).

Both programs take input files of landmark coordinates in standard X1,Y1...CS format. One block can be non-landmark data, which should be formatted so that all measurements for each specimen are in the same row and the specimens are in the same order as they are in the file of landmark data. Both programs perform a GLS superimposition prior to computing SAs, so the superimposition used to produce the input file of landmark coordinates does not matter.

Running PLSMaker

To load the first block of data, which must be landmarks, click on the **Load Data** button. The second block can comprise any non-landmark data; if that is what you are loading, click on the radio button in the second block next to **Landmark Data** (which will turn off the default). You will notice that there is a third field, to allow for loading a third block of data; this option is not yet enabled (check for upgrades). When the data are loaded, they will appear in the visualization window to the left. You can see those plots again by clicking on the **Show Data** buttons located within the field for each block below **Load Data**.

To perform a two-block PLS, click on the **2Block SVD** button below the **Load Data** fields. The numerical results will appear in the orange field at the bottom, although only the results for the first singular value and axis will initially be displayed. To look at those for the second (and subsequent) axes, move **Up** or **Down** the **Active SVD Axes**. You will see the singular value (SVD score), the percent covariance explained (SVD percentage), and the correlation between blocks explained by that axis.

To determine which, if any, of the singular values and correlations are significant, use the **Statistics** pull-down menu on the toolbar. At present, there is only one option (**Permutation Test**). The default is to do 100 permutations, so if you wish to do more, type in the number in the box under **# of Permutations** (located in the purple field of the display options). The results will appear in the auxiliary window. The first results state the singular value, and the number of times that a value equal to or higher than this was obtained in the chosen number of random permutations; the final column is the *p*-value for the null hypothesis (that this frequency can be explained by chance). The second set of results, printed below, gives the correlations between the scores of the first and second blocks for each singular axis, and the number of times that an equal or higher value was obtained in the chosen number of random permutations; again, the final column is the *p*-value for the null hypothesis (that this frequency can be explained by chance). It is entirely possible that the singular value is not significant but the correlation is. This occurs when the axis explains a trivial part of the covariance. The results seen in the auxiliary window can be copied from the window (by selecting the text and copying it using ^C) and pasted into a text file (using ^V) or appended to a file by clicking on the option **Append Results to File**. As usual, you can safely ignore the caution about overwriting the file.

You can see the relationship between the scores of Block 1 and Block 2 by clicking on **Show Scores** just below **2Block SVD** (both are below the **Load Data** fields). If both data sets are blocks of shape data the plot can be copied to the clipboard, but you will need to use the auxiliary copy function (because the copy function that preserves the aspect ratio in plots of the shape transformations interferes with copying the plots of the scores). Alternatively, you can save the scores to files and use the plotting options in Excel (or another program) by going to the **File** menu on the toolbar up top and selecting **Save Scores for Block 1** then **Save Scores for Block 2**.

To depict the singular axes as shape deformations (for landmark data) or as loadings of the non-shape variables, click on **Plot Axis** (located in the field for each block, beneath the **Load Data** and **Plot Data** options). You have the usual options for displaying the shape transformations; some are in the purple field below the visualization window, the remainder are listed in the **Image** pull-down menu on the toolbar up top. In the purple field you may select **Plot Style**, the **Superimposition** method to use when depicting the deformation (if you select either **Bookstein Coordinates** (BC) or **Sliding Baseline Registration** (SBR), make sure to type in the endpoints of your baseline in the boxes that are provided on the right side of the purple field). You can multiply the deformation by a factor by typing that factor into the **Exaggeration** box, you can alter the range of the grid (if it is too large or small for your landmarks) by typing the desired range in the **Range** box, and you can also alter the **Density** of the plot (the number of grid lines in the deformed grid plots). To alter line weights, symbol sizes (and whether empty or filled) and to remove the axes from the plot, go to the **Image** menu on the toolbar. As usual, you can trim the grid if it extends too far beyond the landmarks by clicking on the **Grid Trimming Active** radio button, located on the right, and you can rotate the plots by clicking on the **Reference Rotation Active**. Because it may be difficult to see how the specimens are oriented when looking at an unfamiliar subset of landmarks, you have the option of printing the number for each landmark on the plot. This should help you determine the angle through which you will need to rotate the plots.

You can save the scores for each SA for each block, and the singular value decomposition information (the singular values, S-Value, the percentage of the covariance between blocks explained, Percentage, and the **U** and **V** matrices). The files of scores are ordered so that SA1 is in the leftmost column, SA2 in the one to the right of it, etc.

Running PLSAngle

Two sorts of comparisons can be made using **PLSAngle**. The first is between SAs of the same blocks of landmarks belonging to different groups (i.e. species). In this kind of analysis, the hypothesis being tested is that the corresponding SAs are the same between groups (for example, that the SA1 of the cranial landmarks is the same for two species). The second kind of comparison is between correlations. It is easiest to follow the logic of the instructions by thinking of the groups as competing hypotheses of integration; one "group" is the hypothesis that Block 1 and Block 2A are most highly correlated, the other "group" is the alternative hypothesis that Block 1 and Block 2B are most highly correlated. As well as asking if one correlation exceeds the other, we can also ask whether SA1 is the same when Block1 is constant but Block 2 varies (i.e. whether the dominant axis of covariance between the cranial and tail landmarks is the same dimension as the dominant axis of covariance between the cranial and median fin landmarks). Finally, **PLSAngle** also allows you to visualize PCs and SAs, so you can see if the dimensions of variance are equivalent to the dimensions of covariance. You can also visually compare the PCs between the two groups, as well as look at the PCs for each group separately. **PLSAngle** also displays the SAs for each block in each group. The program is still under development, so check for upgrades that offer statistical tests between PCs and SAs.

Comparing singular axes of homologous blocks between groups

The default is that both blocks are homologous. Load the blocks; the landmarks will appear in the visualization window as each file is loaded. Clicking on the **Do SVD (2Block)** button will calculate the angles between the corresponding SAs of the two groups. The results will appear in the **Auxiliary Results** box (another window). You will see a list of results, beginning with the line "SVD 1 Block 1 = 21.525 Block 2 = 30.534." This means that the angle between the SA1 of Block 1 in the two groups is 21.525°, whereas the angle between the SA1 of Block 2 in the two groups is 30.534°. The next line reports the angles between SA2 for each block, the third for SA3, and so forth.

Asking to do an SVD will not provide confidence intervals on the angles; to get those, you will need to click on **Bootstrap SVD Angle**. The default is to run 100 bootstraps, and to test the null hypothesis at an α level of 0.05. If you want to increase the number of bootstraps or lower the alpha level, type in your preferences before clicking on **Bootstrap SVD Angle** (Remember that α cannot be lower than $1/N_{bootstraps}$, where $N_{bootstraps}$ is the number of bootstraps. For example, if you run 100 bootstraps, the smallest value of α you can request is 1/100, or 0.01.)

You can save the SAs for Block 1 and Block 2 for each group, along with the reference forms used to compute the partial warp scores. You can also save PC1 for each of the blocks of each group. The output files are row vectors, the format required by **VecDisplay**. If you want to save the SA scores for each specimen, use **PLSMaker**.

A long list of options for graphical displays is presented in the **Display Item** menu in the purple field below the visualization window (although some of the options are not available if Block 2 is not homologous):

1. **SVD Block1, Group1+2** shows SA1 for Block 1 for both groups simultaneously. The PCs are displayed by pairs of vectors of relative landmark displacements. The vectors

for Group 1 are shown in black, those for Group 2 are shown in red. The position of the landmarks is determined by the coordinates of the reference form. You can edit the plot using the options on the **Image Control** pull-down menu (the options are to alter line width, the size of the symbols for the landmarks, and to fill the symbols). To remove the axes surrounding the plots, use the **Axis Controls** pull-down menu (also located on the toolbar up top). You may need to rotate the plots if the orientation is not interpretable. If so, use the **Reference Rotation Active** radio button at the bottom center of the interface to rotate the reference interactively. Alternatively, if you know the angle of rotation you need, you can type it into the **Default Ref Angle** window, in the yellow field below the red **Exit** button.

If you would like to plot the difference between the two SA1s using a different plotting style, such as a deformation grid, you can save the SAs (along with the reference form) by going to the **File** pull-down menu located on the toolbar. These vectors can be input into the program **VecDisplay** (described in Chapter 10), which shows the difference between two vectors or their sum using a variety of display options.

2. **SVD Block2, Group1+2** shows SA1 for the second block for both groups simultaneously. The plotting styles and editing options are as described for (1). This option is not available if Block 2 is not homologous between groups.
3. **SVD Block1, Group1** shows SA1 for Block 1 of Group 1; the available plotting styles and editing options are the same as for **PLSMaker**, described above.
4. **SVD Block1, Group2** shows SA1 for Block 1 of Group 2. The graphical options are as described in (3).
5. **SVD Block2, Group1** shows SA1 for Block 2 of Group 1. The graphical options are as described in (3).
6. **SVD Block2, Group2** shows SA1 for Block 2 of Group 2. The graphical options are as described in (3).
7. **Data Block1, Group1+2** shows the landmarks for the first block for both groups; the data for Group 1 are in blue, those for Group 2 are in red. The Procrustes GLS superimposition is the only option. Editing options are as given in (1).
8. **Data Block2, Group1+2** is the same as (7), except that Block 2 is shown (this option is not available if the Block 2 is not homologous between groups).
9. **PCA Block1, Group1** shows PC1 for Block 1 in Group 1. Display and editing options are as given in (3).
10. **PCA Block2, Group1** shows PC1 for Block 2 in Group 1. Display and editing options are as given in (3).
11. **PCA Block1, Group2** shows PC1 for Block 1 in Group 2. Display and editing options are as given in (3).
12. **PCA Block2, Group2** shows PC1 for Block 2 in Group 2. Display and editing options are as given in (3).
13. **PCA + SVD Block1, Group1** shows PC1 and SA1 for Block 1 of Group 1. SA1 and PC1 are displayed by vectors of relative landmark displacements (the position of the landmarks is determined by the coordinates of the reference form). Those for Group 1 are shown in black, those for Group 2 in red. Editing options are as described above for (1).
14. **PCA + SVD Block2, Group 1** is the same as (13), except that the plot shows PC1 and SA1 for Block 2.

15. **PCA + SVD Block1, Group2** is the same as (13), except that the plot shows PC1 and SA1 for Group 2.
16. **PCA + SVD Block2, Group2** is the same as (14), except that the plot shows Group 2.
17. **−PCA + SVD Block1, Group1** allows you to reverse the direction of the PC. The signs of PCs and SAs are arbitrary; you may find that PC1 and SA1 look nearly identical except that the arrows point in opposite directions.
18. **−PCA + SVD Block2, Group1** is the same as (17), except that the plot shows Block 2.
19. **−PCA + SVD Block1, Group2** is the same as (17), except that the plot shows Group 2.
20. **−PCA + SVD Block2, Group2** is the same as (18), except that the plot shows Group 2.
21. **PCA Block1, Group 1+2** shows PC1 for Block 1 for both groups simultaneously. The PCs are displayed by pairs of vectors of relative landmark displacements. Those for Group 1 are shown in black, those for Group 2 in red. The positions of the landmarks are determined by the coordinates of the reference form. Editing options are as given for (1).
22. **PCA Block2, Group 1+2** is the same as (21), except that Block 2 replaces Block 1.
23. **−PCA Block1, Group 1+2** is the same as (21), but reverses the sign of one of the PCs.
24. **−PCA Block2, Group 1+2** is the same as (22), but reverses the sign of one of the PCs.

Comparing correlations between different pairs of blocks of a single group

In this analysis, the two "groups" are competing hypotheses of integration. The same set of landmarks represents Block 1 for both hypotheses because we are asking if that block is more highly correlated with one Block 2 than with another Block 2. Therefore, the first step is to turn off the default option that Block 2 is homologous between groups. The same file is input as Block 1 of both groups, and two different files are input as the two Block 2s. To do the analysis, click on **Do SVD 2Block** to see the preliminary results (the statistical analysis will be done when you click on **Bootstrap SVD Angle**). If you want more than 100 bootstraps, or an α level other than 0.05, type your preferences in the boxes provided. The results will appear in the **Auxiliary Results** box (another window). The first results are the angles between Block 1 and each of the two "groups". Below that are the correlations between Block 1 and each Block 2, giving the observed correlation, its confidence interval, and its standard error (which can be used in analytic tests of the difference between correlations). The final three lines are the results of the resampling-based test of the equality of correlations.

The variety of output files that can be saved, and the options for graphical displays, are detailed above.

References

Bastir, M., Rosas, A. and Sheets, H. D. (2004). The morphological integration of the hominoid skull: a partial least squares and PC analysis with morphogenetic implications for European Mid-Pleistocene mandibles. In *Developments in Primatology: Progress and Prospects* (D. Slice, ed.), in press. Kluwer Academic/Plenum Press.

Bookstein, F. L. (1982). The geometric meaning of soft modeling, with some generalizations. In *Systems Under Indirect Observation: Causality–Structure–Prediction* (K. G. Jöreskog and H. Wold, eds) pp. 55–74. North Holland Publishing Co.

Bookstein, F. L., Gunz, P., Ingeborg, H. et al. (2003). Cranial integration in *Homo*: singular warps analysis of the midsagittal plane in ontogeny and evolution. *Journal of Human Evolution*, **44**, 167–187.

Corti, M., Fadda, C., Simson, S. and Nevo, E. (1996). Size and shape variation in the mandible of the fossorial rodent *Spalax ehrenbergi*. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 303–320. Plenum.

Hingst-Zaher, E., Marcus, L. F. and Cerqueria, R. (2000). Application of geometric morphometrics to the study of postnatal size and shape changes in the skull of *Callomys expulsus*. *Hystrix*, **11**, 99–113.

Houle, D., Mezey, J. and Galpern, P. (2002). Interpretation of the results of Common Principal Components Analysis. *Evolution*, **56**, 433–440.

Jöreskog, K. G. and Wold, H. (eds). *Systems Under Indirect Observation: Causality–Structure–Prediction*. North Holland Publishing Co.

Klingenberg, C. P., Badyaev, A. V., Sowry, S. M. and Beckwith, N. J. (2001). Inferring developmental modularity from morphological integration: analysis of individual variation and asymmetry in bumblebee wings. *American Naturalist*, **157**, 11–23.

Lowe, A. A., Özbeck, M. M., Miyamoto, K. and Fleetham, J. A. (1997). Cephalometric and demographic characteristics of obstructive sleep apnea: an evaluation with partial least squares analysis. *The Angle Orthodontist*, **67**, 143–154.

Lundrigan, B. (1996). Morphology of horns and fighting behavior in the family Bovidae. *Journal of Mammalogy*, **77**, 462–475.

Rohlf, F. J. and Corti, M. (2000). Use of two-block partial least squares to study covariation in shape. *Systematic Biology*, **49**, 740–753.

Rüber, L. and Adams, D. C. (2001). Evolutionary convergence of body shape and trophic morphology in cichlids of Lake Tanganyika. *Journal of Evolutionary Biology*, **14**, 325–332.

Sampson, P. D., Streissguth, A. P., Barr, H. M. and Bookstein, F. L. (1989). Neurobehavioral effects of prenatal alcohol: part II. Partial least squares analysis. *Neurotoxicology and Teratology*, **11**, 477–491.

Streissguth, A. P., Bookstein, F. L., Sampson, P. D. and Barr, H. M. (1993). The Enduring Effects of Prenatal Alcohol Exposure on Child Development: Birth through Seven Years, A Partial Least Squares Solution. *International Academy for Research in Learning Disabilities monograph series no. 8*. University of Michigan Press.

PART III

Applications of Morphometric Methods to Complex Hypotheses

12

Disparity and variation

Disparity and variation are closely allied concepts – both refer to the general idea of "variety." Disparity usually signifies the variety of a group of species and is the outcome of evolutionary processes; variation, on the other hand, refers to the variety of individuals within a single (homogeneous) population and is the raw material necessary for evolution. In light of the theoretical distinction between the two concepts, it may seem difficult to cover both in a single chapter. However, the distinction between the concepts lies in the processes that produce them and the theories that predict them. The metric (or formula) for measuring disparity among species is the same as that used to measure variation within a species. Because the same metric is used to measure both, we cover them both in the same chapter. Even so, to avoid confounding concepts that have little in common aside from a metric, we begin by reviewing their biological meanings, then turn to the issue of measurement.

Disparity

Disparity may be an unfamiliar term to many biologists, but it has emerged as a major theme in the paleobiological literature. The term was introduced to clarify the distinction between two notions of diversity that were often confounded: (1) phenotypic variety (often but not always morphological), and (2) taxonomic richness. Over the past decade, owing largely to work by Foote (especially Foote, 1990, 1993a, 1993b) the distinction between them has been clarified – a major step towards increasing both conceptual clarity and methodological rigor. In the early literature the number of taxa was often used as a measure of "disparity," but, as Foote showed (1993b), and as many other studies have confirmed, the number of taxa increases even as their morphological variety decreases.

To date, most studies of disparity have focused on its temporal dynamics over a geological time scale. The chief questions addressed by such studies are:

1. What is the temporal pattern of disparity?
2. What evolutionary processes explain those patterns?

Such studies are almost invariably based on fossils because they require sampling disparity at multiple times in the geological record. Some groups studied in this way include Cambrian marine arthropods (Foote and Gould, 1992; Wills et al., 1994), Paleozoic blastozoans

Geometric Morphometrics for Biologists
ISBN 0–12–77846–08

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

(e.g. Foote, 1992), stenolaemate bryozoans (Anstey and Pachut, 1995), crinoids (e.g. Foote, 1994; Ciampaglio, 2002), gastropods (Wagner, 1995) and Ordovician trilobites (Miller and Foote, 1996). The growing empirical literature on disparity repeatedly documents a surprising historical pattern: disparity initially increases and then stabilizes or even decreases while the number of taxa increases.

Efforts to explain this pattern have focused on two classes of hypotheses: ecological and developmental. Ecological hypotheses postulate that ecological space is initially open and then becomes saturated; limits on disparity are thought to arise from the structure of the ecological space. In contrast, developmental hypotheses propose an intrinsic explanation for limits on disparity – the acquisition of developmental constraints that stabilize morphology (see Wagner, 1995 and Ciampaglio, 2002 for reviews of hypotheses and approaches to testing them). Whether any explanation is even needed has been questioned in a profound (if difficult) theoretical analysis (Gavrilets, 1999). At present it is not clear what we ought to expect from disparity under plausible models; nor is it clear what role artifacts might play in the patterns detected by empirical analyses. It is also difficult to isolate causal factors that might explain the temporal dynamics of disparity because of the multiplicity of uncontrollable factors that can influence those dynamics, including rates of speciation and extinction, selectivity of extinction or speciation that is non-random with respect to morphology, the magnitude of change within a lineage, and factors potentially limiting that magnitude (such as developmental and selective constraints).

Of the various factors that can influence disparity, constraints may be the least understood – partly because they are rarely documented prior to analyzing disparity. Instead, constraints are inferred from the data, even though it is not clear how either developmental or selective constraints ought to influence disparity. Both sorts of constraints are thought to limit disparity, which may seem intuitively obvious; however, like many intuitions, it may be faulty. We know little about the impact of either sort of constraint on disparity, and determining their impacts will require studies that document constraints independently of such supposed effects. We cannot simply infer constraints from decreases in disparity when we do not know if they generally decrease disparity. Instead, we need to determine whether development is constrained or not, and then ask how those constraints affect disparity. In at least one case, developmental constraints are inferred to *increase* disparity (Zelditch et al., 2003).

Studies of disparity of living taxa are still relatively rare, but they have been used to address basic issues in evolutionary biology – such as whether decoupling of integrated parts increases disparity (Schaefer and Lauder, 1996), whether biomechanical and morphological disparity are related to each other (Hulsey and Wainwright, 2002), and whether developmental constraints might limit disparity (Zelditch et al., 2003). Surprisingly few studies have tried to relate ecological heterogeneity and morphological disparity, an obviously important direction for future research (Roy and Foote, 1997).

Any biological explanation for an empirically documented pattern rests on the assumption that the pattern is real. Whether it is real or an artifact depends partly on how disparity is measured, and also on the sampling design. Both metrics and sampling designs have been foci of critical reviews. In particular, a number of critics have taken issue with the phenetic approach to disparity implicit in the use of a variance as its metric (e.g. Wills et al., 1994). Alternative metrics, which measure change along branches of a phylogeny, have been recommended, but they are still in their infancy. Such metrics are difficult to apply when ancestors have not been sampled (or are unknown), and they also pose an interpretative challenge because they redefine disparity, replacing the idea of variety (around an average) with that of directed change away from the ancestor (see Wills et al., 1994; Wagner, 1997; Smith and Lieberman, 1999). A second criticism is that measures of disparity typically do not consider the biological significance of the contributing variables. It is conceivable that large morphological changes could have few biological consequences, and some small changes affecting just a few morphological details could have profound consequences for function. In that light, weighted measures of disparity that take the biological significance of the changes into account might seem more justified than measures of disparity *per se* (see Wagner, 1995).

For recent reviews of the literature, including critical discussions of metrics and methods, and summaries of empirical studies, see Foote (1997), Ciampaglio et al. (2001) and Wills (2001).

Variation

Variation within populations is a major theme in evolutionary biology because it is so fundamental to evolution – phenotypic variation provides the opportunity for selection to act, and genetic variation enables selection to effect change. Variation is the raw material on which selection acts, and its structure can influence the outcome of selection. Because evolution can be constrained by limited or biased variance, the variance–covariance matrix is sometimes viewed as an intrinsic constraint on evolution; such limits or biases arising from developmental processes are developmental constraints (see Maynard Smith et al., 1984). Although that view of variation emphasizes its role as a potential constraint, the structure of (co)variation itself may be molded by selection. Theoretical models predict that phenotypic and genetic (co)variance structures evolve to match patterns of developmental and functional integration (e.g. Lande, 1980; Cheverud 1982, 1984; Wagner, 1988; Wagner and Altenberg, 1996). This matching is expected to result from differential elimination of pleiotropic effects between members of different functional complexes, combined with the maintenance (or augmentation) of pleiotropic effects within a complex. There is much empirical evidence that phenotypic and/or genetic covariances reflect developmental and functional relationships among traits, a conclusion based on many exploratory studies (Olson and Miller, 1958; Berg, 1960; Van Valen, 1962, 1970; Gould and Garwood, 1969). In addition, many studies have deduced the structure of (co)variation among measurements from developmental and functional theories (e.g. Cheverud, 1982, 1995; Zelditch and Carmichael, 1989; Kingsolver and Wiernasz, 1991; Marroig and Cheverud, 2001). Most studies concentrate on a single developmental stage, but a few have examined the ontogenetic dynamics of variance (e.g. Foote, 1986; Zelditch, 1988; Zelditch and Carmichael, 1989; Zelditch et al., 1993).

The concept of variation is also central to systematic studies, both because systematists study evolutionary processes and also because the systematic value of a character is partly a function of its variability. In the systematics literature the term "variation" is sometimes used very broadly, such as when talking about "ontogenetic variation." In that context the "variation" results from the mixture of ages in the sample; because individuals differ in age, they differ in everything that changes with age. Ontogeny is thus the factor explaining the variation within the sample, but that is *not* the variance on which selection acts (unless we seriously entertain the idea that selection favors adults over juveniles, which is unlikely in the first place and would not have any evolutionary consequences in the second). To study the variance on which selection could act, we would first need to remove the variation resulting from the heterogeneity of the sample. Should removing that variation strike you as an improper manipulation of the data, ask yourself whether it is reasonable to imagine that selection acts on it.

A classic hypothesis linking variance to disparity is often called the "Kluge–Kerfoot" phenomenon: traits that vary the most (within populations) are also the ones that most differentiate populations (Kluge and Kerfoot, 1973). The original empirical support for the hypothesis was harshly criticized on methodological grounds (e.g. Sokal, 1976; Rohlf et al., 1983), but the hypothesis has re-emerged in the recent literature with more impressive empirical support; the dimension of greatest (genetic) variance is sometimes regarded as the evolutionary line of least resistance (e.g. Schluter, 1996).

Metrics for disparity and variance

As mentioned above, there is no universally accepted metric for disparity (there is for variation, so we will focus on disparity throughout this section). One major distinction among available metrics is whether they measure the variety of forms in a sample or the diversification along branches of a cladogram. The first could be viewed as a static measure of disparity, the second as a dynamic measure of diversification. We will focus on the first approach for two reasons: the first is that we define disparity in terms of variety rather than in terms of magnitudes or rates of diversification; the second is that ancestral morphologies are rarely observed and known to be ancestral. Without direct observations of known ancestors, ancestral morphologies must be inferred, and the methods for inferring ancestral morphologies are still a matter of dispute.

Metrics for the variety of observed forms can be subdivided into two broad classes: (1) those applied to continuously valued variables (such as size and shape) and (2) those applied to ordinal or categorical data. The distinction (which is based on the type of data) is important, because continuously valued variables are measured on an unambiguous scale, which is not the case for ordinal or categorical data. For example, if we want to know how different two organisms are, and one is 10 mm while the other is 12 mm, we can say that their difference is 2 mm. Given a third, which is 14 mm, we would say that the difference between the first and third is 4 mm, and the difference between the second and third is 2 mm. Because 2 mm is equal to 2 mm, we can say that the difference between the first and second organisms is equal to that between the second and third. We might choose a scale that takes proportions into account, so that 2 mm counts for more when organisms are near 1 mm than when they are near 100 mm, but still the scale is unambiguous and measurements are mathematically commensurable. In contrast, if we classify morphologies into three types – "one," "two" and "three" – "one" and "two" are taken to be one unit apart, as are "two" and "three," but we cannot say that the difference between "one" and "two" is equal to the difference between "two" and "three." Perhaps the first two types differ by the presence or absence of a notochord, whereas the second two differ by the presence or absence of a tubercle on the tibia. The problem faced here does not arise when coding discrete classes for phylogenetic analyses because the characters may be equally informative in that context.

However, weighting them equally in studies of disparity implies that they contribute equally to morphological variety. Fortunately size and shape data are continuously valued variables, so we will concentrate on metrics of disparity suited to continuously valued variables.

The metrics for continuously valued variables can be either Euclidean or non-Euclidean distances, although most workers use Euclidean distances. We can also distinguish among metrics by whether the measures are of: (1) linear distances between forms (corresponding to a standard deviation); (2) squared distances between forms (corresponding to a variance); or (3) volumes. Measures of volume might seem most desirable because they could appear to capture the most information about the size of the occupied morphospace. Unfortunately no satisfactory measure of volumes is available yet, because measuring them involves multiplication rather than addition. When distances along dimensions are multiplied, a trivial distance along one deflates the size of the space. For example, if we multiply distances along several dimensions, such as 0.4, 0.3 and 0.2, we get a volume of 0.024, and if we multiply that product by 0.002, we get 0.000048 – therefore, adding information about that fourth dimension reduces the size of the space to nearly zero. Logically, we would expect that the additional information would only increase the size of the space. Another disturbing feature of this volume-based approach to disparity is that the volume of several slightly disparate variables can be far larger than the volume of three very disparate variables and one nearly invariant variable. For example, above we considered a case of three disparate variables and one that is nearly invariant. We might have another case in which there are also four variables, each with a disparity of 0.1; the product of (0.1)(0.1) (0.1)(0.1) = 0.0001, which is more than twice the volume of the first case (0.000048). In contrast, if we restrict our analysis to only the first three variables, the disparity would be (0.1)(0.1)(0.1) = 0.001 – substantially less than that of the first case (0.024).

If we had an objective and non-arbitrary method for ignoring some dimensions (so that their low levels of disparity do not deflate the space), we could circumvent these problems. However, all methods for deciding whether to exclude a variable depend on subjective arguments, and the decision about whether to exclude a variable can have an enormous impact on the results. For that reason, we prefer metrics based on standard deviations and variances. Both standard deviations and variances are equally useful metrics, and there is no reason to debate which of them is preferable because one is easily derived from the other. The major reason for using a variance is that variances are additive. Because of that property, we can calculate the overall disparity of a group, then partition it into the contribution made by each taxon (the partial disparity of that taxon; Foote, 1993a). The additivity of variances means that the sum of partial disparities equals the overall disparity. However, it is worth noting that the two measures weigh outliers differently, and consequently their results can differ. Standard deviations and variances are not linearly related, and a highly distinctive taxon has a much greater impact on a variance than on a standard deviation.

Measuring disparity

To measure morphological disparity (*MD*) by a variance, we calculate:

$$MD = \frac{\sum_{j=1}^{N} D_j^2}{(N-1)} \qquad (\mathbf{12.1})$$

where D_j is the distance of species j from the overall centroid (which is the grand mean calculated over the n species or other groups being analyzed). We can use Equation 12.1 to calculate both size and shape disparity. For *size* data, D_j is the difference between the centroid size of an individual species and the grand mean centroid size. For *shape* data, D_j is the Procrustes distance between the average shape of an individual species and the grand mean shape. We can compute shape disparity directly by estimating those Procrustes distances, or we can calculate the variances of coordinates obtained by a generalized least squares Procrustes superimposition (GLS) or variances of partial warp scores (including scores on the uniform component). All three approaches yield the same results because the sum of squared coordinates obtained by GLS equals the squared Procrustes distance to the mean, as does the sum of squared partial warp scores. In those analyses the grand mean shape is the consensus, so if we are using partial warps we can use the formula:

$$MD = \frac{\sum_{j=1}^{N} PW_j^2}{(N-1)} \qquad \mathbf{(12.2)}$$

where PW represents the partial warp scores for an individual, so the formula tells us to sum all the squared partial warp scores for each individual over all individuals. Because the grand mean shape is the consensus, its partial warp scores are all zeros, so Equation 12.2 is equivalent to Equation 12.1.

Both are also equivalent to:

$$MD = \mathbf{Tr}\{\mathbf{S}\} \qquad \mathbf{(12.3)}$$

where **Tr** is the trace of a matrix (the sum of its diagonal elements) and **S** is the variance–covariance matrix of the partial warp scores (including the uniform component, and computed using the grand mean as the consensus). The diagonal elements of a variance–covariance matrix are the variances, so this formula tells us to sum the variances of the variables, which takes us back to the squared distances from the consensus.

To exemplify the analysis of disparity, we will measure the disparity of adult body shape of nine species of piranhas sampled at the 16 landmarks shown in Figure 12.1. Before doing this analysis, we remove the shape variance within each species that is due to ontogeny, allowing us to estimate the shape of an average adult (this is done by standardizing each species to its maximum adult size, as explained in Chapter 10). Each species is represented by a single data point, the mean shape for that species. There are nine species, so $N = 9$. The result of the analysis is that $MD = 0.00398$. Of course, we cannot yet interpret this number – we cannot say if that value is large or small, or how uncertain it is. Before we can go any farther, we need to deal with the issue of uncertainty.

Placing confidence intervals on morphological disparity (*MD*)

To construct the confidence interval, we need first to consider the various parameters being estimated. In general, there is uncertainty in the estimate of the mean shape of each species, and in the estimate of the consensus. Both uncertainties must be taken into account when putting confidence intervals around *MD*. Additionally, when the mean shape of each species is calculated by removing the variance due to ontogeny (or some other factor) we must also account for the uncertainty of the regression model used to standardize the

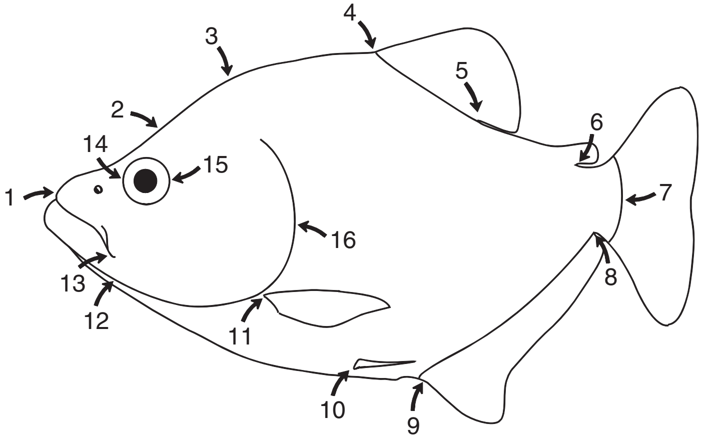


Figure 12.1 Landmarks sampled on the external body form of piranhas.

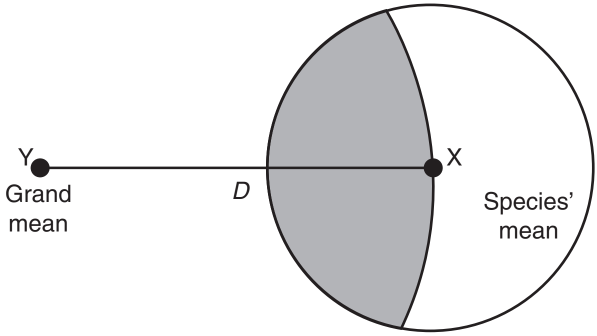


Figure 12.2 The line joining a species' mean to the grand mean; random variation in the position of the mean only rarely lies along the line within the shaded region. Changes in the position of shapes orthogonal to that line or within the unshaded region increase the distance to the mean.

shapes. We may also need to take a further source of uncertainty into account – the sampling of species, because unless we have measured them all we must consider the uncertainty of the grand mean that arises from our sampling of species. If we do not consider this particular source of uncertainty, we cannot generalize from our sample of species to the larger group that includes them, although we can make statements about our particular sample of species that takes the uncertainty of our sampling of them into account.

The confidence intervals might look odd because they frequently are not symmetric about the mean, even when the distribution of shapes around the GLS consensus *is* symmetric. That symmetric distribution of shapes implies that the uncertainty in the estimate of the mean is roughly equal in all directions (i.e. it is a hyperspherical solid). Turning to the estimates of disparity, we can see why the uncertainty in the distance of a species from grand mean is not symmetric about the mean distance even then. The hyperspherical distribution of uncertainty in the mean yields a non-symmetric distribution of distances – there are many more possible locations of a species' mean that increase the distance than there are that decrease it. As we can see in Figure 12.2, the line joining the grand mean to a species' mean is in a single direction in a high dimensional space; random variation in the position of the sample mean rarely lies along the line between the species' mean and grand mean. In Figure 12.2, *D* is the distance from the species' mean to the grand mean shape, and the circle around X represents the range of uncertainty about the species' mean. The region within the circle that is a distance *D* or less from the grand mean is shaded, and this region is clearly smaller than the unshaded region that is farther than *D* from the grand mean. This effect is even more pronounced in higher dimensions.

We can construct confidence intervals and standard errors for *MD* by bootstrapping. When we need to take the uncertainty of the regression into account, we first fit a regression model to the data, then use the procedure described in Chapter 10 – determining the residuals, predicting the shape expected for each size, bootstrapping the residuals and randomly allocating them to each predicted shape, then refitting the regression model to the data to generate a standardized data set for the bootstrap set. This is iterated *N* times (where *N* is the number of bootstrap sets). If we do not need to take the uncertainty of the regression into account, we simply resample (with replacement) from each of the samples. For each bootstrap set of standardized values, we calculate the disparity of that sample using the formula for *MD* above. In the case of the adult piranhas discussed above, the estimate of $MD = 0.00398$; the 95th percentile over the bootstrap sets gives us the two-tailed confidence interval on that estimate, 0.00377 to 0.00440.

We still do not know if that value is large or small because we have still not compared it to the disparity of anything else. We will thus continue the analysis, comparing the levels of adult disparity to that of juveniles, and comparing the disparities of several piranha clades (Figure 12.3).

Example: ontogenetic and interclade comparisons of disparity

Table 12.1 gives the disparities (*MD*) of juvenile and adult shapes, as well as the standard errors (*SE*) for the estimates. As explained in Chapter 9, we can use a *t*-test to determine whether derived traits like mean disparities are significantly different:

$$t = \frac{MD_1 - MD_2}{\sqrt{\left(\frac{(N_1 - 1)N_1 SE_1^2 + (N_2 - 1)N_2 SE_2^2}{N_1 + N_2 - 2}\right)\left(\frac{N_1 + N_2}{N_1 N_2}\right)}} \tag{12.4}$$

with $(N_1 + N_2 - 2)$ degrees of freedom. Because *MD* is computed from the mean shapes of species, N_1 and N_2 are the numbers of species in the respective clades. We can also use a bootstrap procedure like that used to test whether two Procrustes distances are different. We begin by computing the disparities of the two groups and the difference between those disparities, then we resample each data set with replacement and repeat the calculation of the disparities and the difference between them. After a sufficient number of bootstraps, we can determine the 95% interval for the range of differences. If this range excludes zero, we can conclude that the observed difference is significant at the 95% level.

For the most inclusive piranha group (Clade 1), disparity decreases significantly over ontogeny, as it does in Clade 2. In Clade 3, disparity increases statistically significantly, but the change is slight – in contrast to the dramatic increase in Clade 4. In Clades 5 and 6,

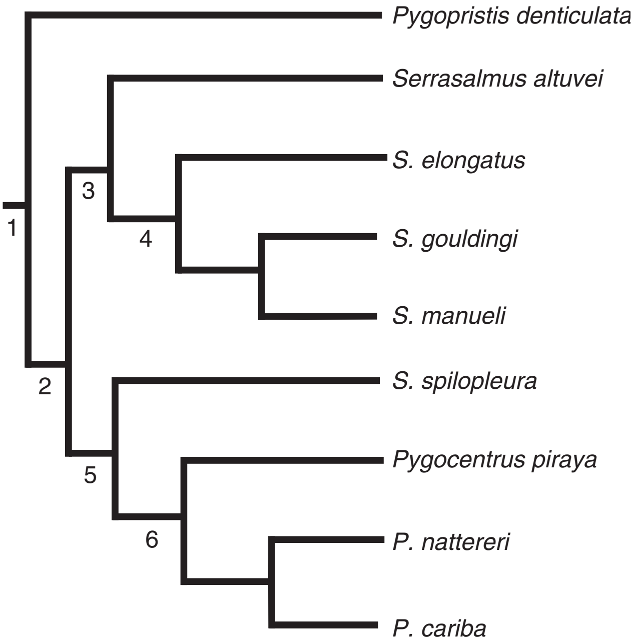


Figure 12.3 Cladogram of the piranhas analyzed in this chapter; nodes are numbered to designate clades.

Table 12.1 Disparities of clades (numbered as in Figure 12.3), measured at two ontogenetic stages (disparities of juveniles are measured at the transition from larval to juvenile growth; those of adults are measured at maximum body size attained by each species)

Taxon	*Juvenile disparity*	*Standard error*	*Adult disparity*	*Standard error*
Clade 1	0.00543	0.0003	0.00398	0.0002
Clade 2	0.00575	0.0003	0.00405	0.0002
Clade 3	0.00431	0.0004	0.00550	0.0003
Clade 4	0.00229	0.0002	0.00603	0.0004
Clade 5	0.00116	0.0002	0.00151	0.0001
Clade 6	0.00073	0.0002	0.00051	0.0002

disparity is constant throughout ontogeny. A perhaps counterintuitive result is that adult disparities of Clades 3 and 4 are significantly greater than that of the group as a whole (Clade 1), which may seem impossible, but disparities measured this way are not additive. In these analyses, we are measuring the disparity of each clade relative to that clade's own mean – hence a low disparity indicates that few species differ by much from the mean of that clade. Consequently, a group comprising three or four species that differ a great deal from each other (and from the group mean) can have a much higher disparity than a larger group that includes those species. That is because the additional species in the larger group may all be much closer to the grand mean. Consequently, their values of D_j are small and contribute relatively less to $\sum D_j^2$, whereas the addition of each species increases $N-1$ by one.

Table 12.2 Partial disparities (*PD*) of adults, and the standard errors of *PD*

Species	PD	% MD	*Standard error*
P. denticulata	0.00039	9.82	0.00032
S. elongatus	0.00144	36.27	0.00029
S. gouldingi	0.00026	6.55	0.00031
S. manueli	0.00033	8.31	0.00032
S. altuvei	0.00014	3.53	0.00032
S. spilopleura	0.00023	5.79	0.00032
P. cariba	0.00036	9.07	0.00028
P. nattereri	0.00039	9.82	0.00027
P. piraya	0.00043	10.83	0.00031

The net effect is that *MD* decreases. For that reason, a large group containing only a few species that are far from the grand mean can be less disparate than a small group with the same number of species far from the mean. That is one reason why morphological disparity can decrease while taxonomic diversity increases.

Partial disparity

When we want to quantify the contribution that a particular taxon makes to the overall disparity of a larger group, we want a metric that allows us to partition disparity additively. Therefore, we need an alternative to the method discussed above. The alternative does allow us to estimate partial disparity (*PD*) of the species, and the partial disparities sum to the total disparity. We estimate partial disparities (*PD*), following the procedure outlined by Foote (1993a), in terms of the variance contributed by each individual species:

$$PD = \frac{D_i^2}{N - 1} \qquad (\mathbf{12.5})$$

where D_i is the distance of the *i*th species from the grand mean and N is the total number of species (or other groups). If we wish to calculate the partial disparity of several species (e.g. a subclade in a larger clade) we can sum their individual partial disparities, yielding the partial disparity of that group.

We can see the difference between the two approaches by comparing results (for adults) in Tables 12.1 and 12.2. The total disparity over all nine species (Clade 1) is the same for both. By estimating the partial disparities for all the species, we can determine that the partial disparity of Clade 4 is 0.00203, which is 52.6% of the total. The partial disparity of a single species, *S. elongatus*, accounts for 36.3% of the total disparity of adults of these nine species. Quantifying partial disparities is one method for estimating the phenotypic distinctness of a particular taxon, which may have a practical application in conservation biology.

Variation

Studies of variation, like those of disparity, use a variance as a metric. The major computational difference between analyses of disparity and variance are that (1) studies of variance use the mean of a single homogeneous population as the grand mean, and (2) individuals (rather than mean shapes of species) are the data points in studies of variance. One quick method for estimating the variance in shape is to calculate the variance for all the coordinates obtained by a GLS superimposition and sum those variances over all landmarks (this is exactly the same as calculating the trace of the variance–covariance matrix, and can be done in any spreadsheet). This method, while quick and intuitive, will not provide confidence intervals. It can also be risky if it leads to thinking of variances as being *at* landmarks (recall that changes in relative landmark positions are distributed across landmarks, a topic discussed in context of superimposition methods, Chapter 3). Just as change is not located *at* a landmark, neither is variance.

We exemplify an analysis of the ontogeny of variation by comparing the variance of skull shape across four ages, 10-, 15-, 20-, and 25-days postnatal, of the house mouse (*Mus musculus domesticus*). The superimposed landmarks for each sample are shown in Figure 12.4; the estimates for the variance in shape at each age, and standard errors of the estimate, are given in Table 12.3. To compare the levels of variance between successive ages, we again use the *t*-test to evaluate the difference between variances relative to the pooled standard errors of those variances (the same procedure discussed above for comparing levels of disparity). Over the initial 5-day interval variance is halved, but it is subsequently stable. The loss of variance, in the absence of any selective deaths in the colony, indicates that variation is developmentally regulated and the later stability of the variance also suggests canalization because we would expect continued production of variation by the ongoing process of skeletal development.

Analyzing the structure of disparity

To this point we have talked solely about the magnitudes of disparity and variance; in this section, we discuss methods for analyzing their structure. We address two questions about that structure:

1. Are shapes randomly distributed throughout the morphospace?
2. Do two samples occupy the same subspace?

The first question is answered using nearest-neighbor analysis, the second by comparing occupied subspaces or variance–covariance matrices.

Nearest-neighbor analysis

Nearest-neighbor analysis, as the term implies, examines the smallest distances between shapes. From those distances, we can ask whether shapes are more (or less) similar than expected by chance. If they are closer than expected by chance, we would reject the null hypothesis in favor of one of clustering; conversely, if they are further apart than expected by chance, we would reject the null model in favor of a hypothesis of "over-dispersion" (or "repulsion"). Because the null model is the distribution expected by chance, it is important to consider what the reasonable null model might be. One reasonable null model is that the probability of being at any location in the morphospace is equal (uniform) over the entire space, and is independent of the shape of any other species. Another reasonable null model

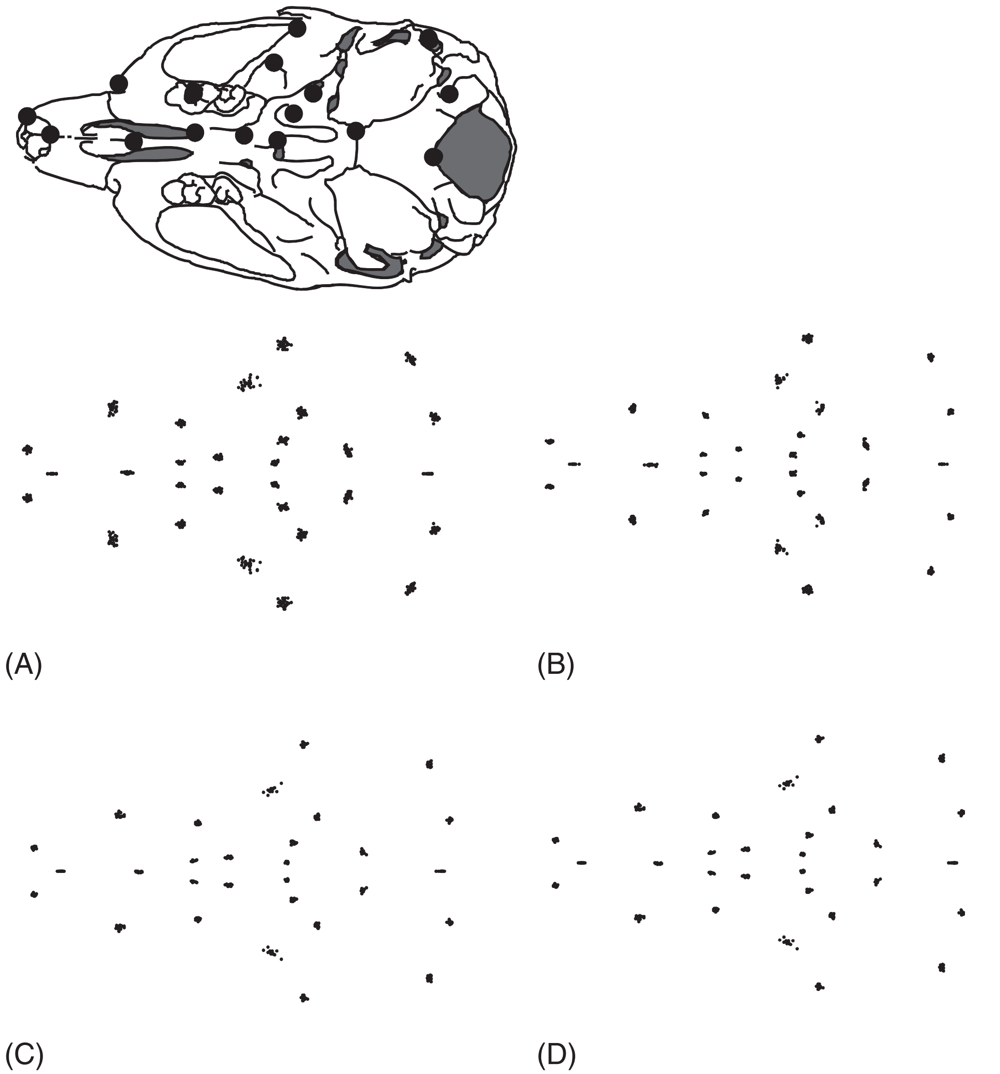


Figure 12.4 Superimposed landmarks of *M. m. domesticus*: (A) 10-day-olds; (B) 15-day-olds; (C) 20-day-olds; (D) 25-day-olds. Analyses are based on the 16 landmarks of the half-skull.

Table 12.3 Skull shape variance of *M. m. domesticus* sampled at four ages (given in days after birth), and the standard errors of the variance (the superimposed landmarks are shown in Figure 12.4)

Age	*Variance*	*Standard error*
10	0.000628	0.0001
15	0.000349	0.00005
20	0.000316	0.0001
25	0.000410	0.0001

is that shapes follow a normal (Gaussian) distribution. The uniform model is a reasonable null for comparisons among species, whereas the Gaussian model is more reasonable when analyzing distributions of individuals around the mean of a homogeneous sample. Having two null models allows us to guard against accepting a hypothesis of a *particular* random distribution.

Nearest-neighbor analysis is another method pioneered by Foote (1990), so we begin by reviewing his approach, and then we extend it to geometric shape data.

Foote's approach to nearest-neighbor analysis The first step in a nearest-neighbor analysis is to compute the nearest-neighbor distance D_i for each of the N species (or other groups) in the study. For the sake of brevity, we will refer to "species" as the units of analysis, but the analysis follows the same protocol even when the units are individual specimens. The next step is to construct a second data set using Monte Carlo simulations. That is done by estimating the mean and range of each variable; from the data, $N - 1$ simulated specimens are generated with values randomly drawn from the observed range. Monte Carlo simulations are similar to bootstraps in that they simulate data based on a given null model and an observed set of data, but they differ in that bootstrapping is carried out using a non-parametric resampling procedure whereas Monte Carlo simulations are based on a distributional model. The distribution of the original data set is parameterized, and those parameters are used to generate a simulated dataset having the distribution of the observations (see Chapter 8). Given the simulated data, a second nearest-neighbor distance, R_i, is computed between each observed specimen and the one closest to it in the Monte Carlo set (note that R_i is not a nearest-neighbor distance between Monte Carlo specimens, but rather the distance between an *observed specimen* and the *nearest Monte Carlo simulated specimen*).

Foote provides a measure that allows us to compare the fit of the simulated distances to the observed ones, the proportional distance P_i for the *i*th specimen. This is a ratio whose numerator is the difference between the two distances (D_i, the observed nearest neighbor distance, and R_i, the Monte Carlo nearest neighbor distance) and whose denominator is the Monte Carlo nearest neighbor difference:

$$P_i = \frac{D_i - R_i}{R_i} \tag{12.6}$$

If the random model fits the data, we would expect that, on average, D_i would equal R_i, and hence the mean P_i over all specimens (P_{mean}) is zero. When P_{mean} is less than zero the observed specimens are more clustered than expected by chance; conversely, if P_{mean} is greater than zero they are further apart than expected by chance. To determine whether zero lies within the confidence interval, we estimate the range of P_{mean} by running the Monte Carlo simulation many times.

To generate a Monte Carlo set under a multivariate normal (Gaussian) model, we must estimate the mean and standard deviation of each variable; to generate a Monte Carlo set under a uniform distribution model, we must estimate the upper and lower bounds of the range for each variable. It can be difficult to estimate the range accurately when sample sizes are small because, at small sample sizes, the observed minimum and maximum will underestimate the "true" range. Thus, rather than using the observed minimum and maximum values to estimate the range, Foote uses estimators developed by Strauss and

Sadler (1989) for the "true" minimum (Y) and the "true" maximum (Z) of a distribution:

$$Y = \frac{NA - B}{N - 1} \tag{12.7}$$

$$Z = \frac{NB - A}{N - 1} \tag{12.8}$$

where A is the lowest observed value and B is the highest observed value in N specimens. Rather than use the *observed* minimum and maximum values, Foote determines the mean and the standard deviation of a normal distribution fitted to the data. He uses normal theory (citing Feller, 1968) to predict the mean and standard deviation:

$$X_{mean} = Y + \frac{(Z - Y)}{2} \tag{12.9}$$

$$SD_X = \left\{ \frac{(Z - Y)^2}{12} \right\}^{1/2} \tag{12.10}$$

and he uses those to estimate the range parameters:

$$Y = X_{mean} - 3^{1/2} SD_X \tag{12.11}$$

$$Z = X_{mean} + 3^{1/2} SD_X \tag{12.12}$$

The geometric approach to nearest-neighbor analysis Extending nearest-neighbor analysis to geometric data is straightforward. Distances D_i and R_i are measured by Procrustes distance; estimates of means, standard deviations or ranges used in the Monte Carlo simulation are obtained by calculating the statistics from the coordinates of each landmark. The rest is straightforward: a Monte Carlo data set is generated and R_i is calculated for each specimen, and these are used to estimate P_{mean}. The simulation is reiterated numerous times, yielding the distribution of P_{mean} values over the Monte Carlo sets. It is then possible to carry out all the usual statistical tests using this distribution.

Nearest-neighbor analysis of piranha disparity

We will test two hypotheses:

1. Piranha body shapes, both juvenile and adult, are further apart than expected.
2. Those shapes are more clumped than expected.

The reason for testing these hypotheses separately is that a conservative test of one is a liberal test of the other. For the hypothesis of over-dispersion, the conservative approach uses the Strauss and Sadler estimator of the range – the estimator enlarges the range so that large distances between points will not necessarily be further apart than expected. However, that expansion of the range can lead to a liberal test of clumping because, within that expanded range, observations may be closer than expected. To be conservative, we would test the hypothesis of over-dispersion using the enlarged range, but we would use parameters of the observed range to test a hypothesis of clustering. Each hypothesis will be tested using two null models, one uniform and the other Gaussian, because we have no good reason to view one as a more plausible random model.

Testing over-dispersion Using the uniform model, the average P_{mean} of the juveniles is -0.2810 and the 95% range of P_{mean} is from -0.3551 to -0.1792, an interval that excludes zero. This result suggests a non-random distribution, with distances being smaller than expected under a random uniform model. Using the Gaussian model, the average $P_{mean} = -0.2758$ and its range is from -0.3450 to -0.1950, an interval that again excludes zero. Both results thus argue against the hypothesis of a random distribution and also against over-dispersion. Instead they suggest clustering, the hypothesis we will explicitly test after we have tested the hypothesis of over-dispersion for adults.

Using the uniform null model, the average P_{mean} of the adults is -0.267 and the range is from -0.3365 to -0.1689, an interval that excludes zero. This result also suggests a non-random distribution, with distances being smaller than expected under a random uniform model. Using the Gaussian model, the average $P_{mean} = -0.2636$ and the range is from -0.3312 to -0.2036, an interval that also excludes zero. As we found for the juveniles, the data argue against the null hypothesis of a random distribution, and also against over-dispersion. Therefore, we now explicitly test the hypothesis of clustering.

Testing clustering We now test the hypothesis of clustering using the narrower estimate of the range. For the juveniles, based on the uniform model, the average $P_{mean} = -0.3172$ with a range from -0.3813 to -0.2247, an interval that excludes zero and supports the hypothesis of clustering. Analyzing the data under the null Gaussian model, the average $P_{mean} = -0.3006$ with a range from -0.3700 to -0.2372, an interval that again excludes zero. Taking these results altogether, they suggest that juvenile piranha body shapes are more tightly clustered than expected under either null model.

For the adults, using the uniform null model, the average $P_{mean} = -0.2537$ with a range from -0.3092 to -0.1788, an interval that excludes zero. These results again support the inference of clustering. Analyzing the data under the Gaussian null model, the average $P_{mean} = -0.2388$ with a range from -0.3091 to -0.1598, an interval that once again excludes zero. Taking these results altogether, they suggest that adult piranha body shapes are more tightly clustered than expected under either null model.

While both developmental stages seem to exhibit clustering, that does not mean that they are otherwise similar in their patterns of disparity. Later in this chapter we will compare the subspaces of morphospace they occupy to determine if they are the same.

Nearest-neighbor analysis of 10-day-old house mouse skull shape variation

Nearest-neighbor analysis can be used to examine patterns of variation as well as disparity. To exemplify this, we will analyze the variation in 10-, 15-, 20- and 25-day-old mouse skulls. The superimposed landmarks for these ages were shown in Figure 12.4. Considering that each sample comprises individuals from a single homogeneous population, we would expect random variation to follow a Gaussian distribution. Results of analyses based on both range estimators (i.e. the parameter values estimated using the Strauss–Sadler estimate of the range (SS), and those estimated from the data (DP)) are given in Table 12.4. It is difficult to argue that the data suggest a departure from random variation. When the parameter estimates are based on an expanded range, the two youngest samples seem to be more clustered than expected under the null hypothesis of a Gaussian distribution. That expansion seems appropriate in light of the small sample sizes, but using it could be

Table 12.4 Nearest-neighbor analysis of skull shape variation in *M. m. domesticus*, sampled at five-day intervals (average P_{mean} and the range of P_{mean} obtained from 100 Monte Carlo simulations)

Age	*SS* ($\mathrm{P_{mean}}$)		*DP* ($\mathrm{P_{mean}}$)	
	Average	*Range*	*Average*	*Range*
10	−0.0929	(−0.1356)–(−0.0276)	−0.0028	(−0.0425)–(0.0377)
15	−0.0944	(−0.1503)–(−0.0334)	0.0153	(−0.0326)–(0.0598)
20	−0.0409	(−0.0963)–(−0.0178)	0.0126	(−0.0313)–(0.0658)
25	−0.0745	(−0.1343)–(−0.0051)	0.0122	(−0.0495)–(0.0654)

Parameter estimates are based either on the Strauss–Sadler estimators (SS) or on the parameters of the data (DP).

considered an overly liberal test of clustering. When estimates are based on the observed values, the range of P_{mean} invariably includes zero, and for that reason we cannot rule out the Gaussian null model.

Comparing patterns of (co)variance

The structures of disparity in different groups can be compared by comparing variance–covariance matrices. Several methods are available, which differ in both underlying mathematical models and statistical approaches. One currently favored method is common principal components analysis, which tests a series of hypotheses ordered according to what is often termed "the Flury hierarchy," based on the sequencing of hypotheses established by Flury (1988). The highest level of similarity is complete matrix equality, the next is matrix proportionality (they differ only by multiplication by a constant), the next is common PCs, and the lower levels range from all but one common PC to only one common PC; at the lowest level is complete inequality. A number of studies have used the method to compare genetic and/or phenotypic covariance matrices (e.g. Steppan, 1997; Arnold and Phillips, 1999; Phillips and Arnold, 1999; Marroig and Cheverud, 2001). An alternative, based on a factor-analytic rather than principal component model, is confirmatory factor analysis – a method which requires having a causal theory that predicts the factor structure *a priori*, and asks whether two or more samples are randomly drawn from a single homogeneous population with the predicted factor structure (e.g. Zelditch, 1988; Zelditch and Carmichael, 1989). Both CPCA and confirmatory factor analysis require large samples (it is usually recommended that $N > 100$ for CPCA, and samples that are large may also be required for comparing parameters estimated by confirmatory factor analysis).

CPCA can be applied to geometric data just as easily as to traditional data (see Polly, 2000). Confirmatory factor analysis has never been applied to geometric data, and it may prove difficult to do so; the difficulty lies in devising *a priori* hypotheses that predict the variance–covariance matrix of geometric shape variables from theory (either developmental or biomechanical, for example). Such models are most readily devised for variables that are individually meaningful. The method we highlight in the remainder of this chapter is an innovative approach developed by J. Mezey and D. Houle (unpublished manuscript). It is also based on PCA, the method widely used to reduce the dimensionality of a space. The objective is to calculate the angles between subspaces, just as we earlier computed an angle between vectors of regression coefficients (Chapter 10). Using that method, we can compute the angle between two-dimensional planes or extend the analysis to higher dimensions, calculating the angles between *hyperplanes* ("flat" surfaces of more than two dimensions embedded in higher dimensional spaces). The angle between two subspaces embedded in a common higher dimensional space is the angle through which one subspace must be rotated to match the other; this relationship applies whether the subspaces are two-dimensional planes or hyperplanes. We first discuss the largest possible angle between two hyperplanes, then how to calculate them, how to determine if the observed angle is larger than that between random resamplings of a single group, and how to compare the angles between hyperplanes of different groups.

What is the largest possible angle between two hyperplanes?

It is relatively easy to intuit the largest possible angle between two vectors in a plane – we can rely on our physical intuition. To apply those intuitions to PCs we need to recall that the sign of a PC is arbitrary – rotating a PC by 180° actually brings us back to where we started; the rotated axis differs from the original by only its sign and, because that sign is arbitrary, the two vectors do not differ at all. Another important point to remember is that PCs always pass through the origin $(0, 0)$ (the importance of this fact will become apparent when we need to determine whether lines or planes intersect). In the simplest possible case, PCs have been extracted from an analysis of two variables in two samples, so we are comparing the subspace defined by PC1 between samples. The maximum possible angle is 90° (because an angle of 180° corresponds to an angle of 0°). If we now extract PCs from three variables, and still compare PC1s between groups, we are comparing two lines embedded in a three-dimensional space. We still cannot get an angle greater than 90°; nor could we get a larger one if we embedded them in a higher dimensional space. That is because the two PCs define a plane, and a single rotation about the axis perpendicular to the plane will always align the two PCs. The maximum possible angle depends on the number of rotations required, each of which can range from 0° to 90°.

We can see how understanding the number of rotations requires aids in determining the maximal angle of rotation for the next simplest case: two PCs, still in a two-dimensional space. Because we have two axes and our space is still a two-dimensional plane, the pair of PCs must define the entire space (two orthogonal lines define a plane, and PCs are orthogonal). Any point within that space can be located relative to the coordinate system defined by the two PCs. Because the two PCs span the entire space, there cannot be an angle between the spaces defined by the PCs.

We can now place those two PCs (from each of two groups) in a higher dimensional space, meaning that we have analyzed three measurements in both groups (so the space in which the PCs are embedded is three-dimensional). We are still measuring the angle between the subspaces defined by two PCs in each group, so we are measuring the angle between two two-dimensional subspaces embedded in a three-dimensional space. Both planes pass through the origin, so the planes defined by each pair of PCs must intersect along a line. If we use that line as the fixed axis of rotation (i.e. a "hinge"), we can superimpose one plane on the other by a rotation ranging from 0° to 90°. Again, the angle cannot exceed 90°. At this point it may seem that the angle cannot ever exceed 90°, but that is not the case.

That the angle can exceed 90° becomes apparent when we consider two planes embedded in a four-dimensional space. That means we have measured four variables in each of two groups, and are comparing the subspaces defined by the first two PCs of each group. The dimensionality of the subspace is two, and that of the space in which they are embedded is four. Now there are three possibilities: (1) the two planes are identical; (2) the two planes share a common line; or (3) the two planes are completely independent, intersecting only at the origin. The latter case may be difficult to imagine, because it can only arise in a space higher than three dimensions. If the two planes are identical, the angle between them is 0°. If they share a common line, then a single rotation around that line, which can range from 0° to 90°, will align them. However, if the two planes are entirely disjunct, then we need to rotate them around two distinct axes, and *each* rotation ranges from 0° to 90° – although that does not mean that the maximal angle is 180°. We need to think of the rotations as vectors along orthogonal axes. To add the two rotations, we add the lengths of the vectors. Therefore, expressing the rotations in radians, the total (net) rotation is the square root of the sum of the squared rotations around each axis. That can be calculated just like we compute net displacements along perpendicular axes – as the square root of the summed squared rotations around each axis. Because the maximal rotation about any axis is $\pi/2$ radians (90°), the maximum possible angle of rotation is $\sqrt{(\pi/2)^2 + (\pi/2)^2} = \pi/\sqrt{2}$ radians (~127°).

The maximal angle of rotation depends on two things: (1) the number of dimensions of the hyperplanes (the number of PCs defining each subspace), and (2) the number of perpendicular vectors shared by the two spaces. The maximum number of distinct axes is equal to the difference between these numbers (the number of hyperplane dimensions minus the number of shared perpendicular axes). Because the maximal angle of rotation around a single axis is $\pi/2$, if there are Y distinct (unshared) axes in each hyperplane, the maximal angle between them is $\sqrt{Y\,(\pi/2)^2} = \sqrt{Y}\,(\pi/2)$. If we are comparing pairs of two-dimensional subspaces embedded in a high dimensional space, Y would still be 0, 1 or 2 because there cannot be more than two perpendicular (unshared) axes in two-dimensional spaces. As the dimensionality of the subspaces under comparison increases, the maximal value of Y increases.

Calculating the angle between two subspaces Using an algorithm generously provided by Jason Mezey, we begin the calculation of an angle between two subspaces by calculating M PCs for two groups, **A** and **B**. Our objective is to determine the angle between subspaces defined by the first K PCs of each group (within the total shape space of all PCs). To compute that angle, we construct the matrix $\mathbf{V_A}$ such that its N columns are the PCs (eigenvectors) of the variance–covariance matrix for group **A**. Next we compute a similar matrix, $\mathbf{V_B}$, based on the data set **B**. We will extract the first K vectors from the $M \times M$ matrices, $\mathbf{V_A}$ and $\mathbf{V_B}$, creating the $M \times M$ matrix **P**, such that:

$$\mathbf{P} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{(12.13)}$$

where **I** is a $K \times K$ identity matrix (i.e. a $K \times K$ matrix with ones on the diagonal and zeros everywhere else). The zeros in Equation 12.13 indicate that all other elements in **P** are zeros (these other elements are necessary to make **P** an $M \times M$ matrix).

We then calculate the following projection matrices:

$$\mathbf{Q} = \mathbf{V_A P V_A^{-1}} \tag{12.14}$$

$$\mathbf{R} = \mathbf{V_B P V_B^{-1}} \tag{12.15}$$

The matrices **Q** and **R** are operators that project an arbitrary vector **X** in the original $M \times M$ variable space onto the subspace defined by the first K eigenvectors of **A** and **B**, respectively. We can then define an operator **J**, which is the difference between **Q** and **R**:

$$\mathbf{J} = \mathbf{Q} - \mathbf{R} \tag{12.16}$$

and do an eigenvector decomposition to determine the angular change implied by **J**. The eigenvalues of **J** are paired positive and negative values, having the form $(J_1, -J_1, J_2, -J_2, J_3, -J_3 \ldots)$. There will be several pairs of positive and negative values, and a number of roughly zero eigenvalues. The J_i values $(J_1, J_2, J_3 \ldots)$ express the angles of rotation in orthogonal two-dimensional subspaces that produce the smallest rotation of one K-dimensional subspace into another. To compute the total angular distance, we compute the square root of the summed squared angles of rotation:

$$A_{\text{Distance}} = \sqrt{\arcsin(J_1)^2 + \arcsin(J_2)^2 + \arcsin(J_3)^2 \cdots + \arcsin(J_K)} \tag{12.17}$$

where all angles are in radians (in reporting the angles, we convert them into degrees).

Evaluating the statistical significance of the angle To determine whether the observed angle is larger than expected by chance, we need to compare it to the range of angles expected under the null hypothesis. That null hypothesis is that the observed angle arises by a random subdivision of either group into two. Thus, the null hypothesis states that the between-group angle is no greater than the within-group range (of either sample). To determine if it is, we can use bootstrapping; each group is randomly partitioned into two groups, and a pair of bootstrap sets is formed by resampling (with replacement). The 95th percentile of the range of angles between the data sets (drawn from a single group) can be compared to the angle between the two groups. When the two groups differ in size, the bootstrap sets of the group with the larger sample size are the sample sizes of the (1) larger and (2) smaller data sets. The two bootstrap sets drawn from the smaller of the two groups both have the sample size of the smaller group, because we cannot create a bootstrap set larger than that of the dataset from which it is drawn.

The entire PCA is carried out using the two bootstrap sets drawn from one group, and the angle between hyperplanes is determined for these pairs (in the same manner as it was for the original dataset). Then the same procedure is done for the other groups. Reiterating the procedure for both groups numerous times yields the bootstrap distribution of within-group angles. If the observed angle between the hyperplanes exceeds the 95% range of the within-group angles (generated by the bootstrap procedure), we can conclude that the observed angle could not have been arisen by a random subdivision of a single group. In those cases, the observed angle between hyperplanes is statistically significant.

Implementing this procedure for geometric data is straightforward: we simply compute the angle between hyperplanes defined by the first K PCs (which are calculated from the partial warp scores, including those of the uniform component following a GLS superimposition).

Comparing angles between hyperplanes Above, we asked whether the angle between two planes exceeds what we might expect by chance. We might also want to compare the angle to that found in another comparative analysis. Suppose we are working with three groups, **A**, **B** and **C**, and wish to know whether the subspaces of **A** and **B** differ by more than those of **A** and **C**. To make that comparison, we follow a bootstrap procedure like that used to test whether two disparities are different. We begin by computing the angles between hyperplanes and the difference between those angles, then we resample each data set with replacement and repeat the calculation of the angles and the difference between them. After a sufficient number of bootstraps, we can determine the 95% interval for the range of differences. If this range excludes zero, we can conclude that the observed difference is significant at the 95% level.

Comparing occupied morphospaces across developmental stages

For the comparison between morphospaces occupied by juvenile and adult piranhas, we first estimate the angle between subspaces defined by the first two PCs, and then by the first five PCs. We compare two-dimensional subspaces because the variance–covariance matrix of adult body shapes has two distinct eigenvalues (the variance–covariance matrix of juvenile body shapes has none). We also compare the five-dimensional subspaces because approximately 85% of the variance within each stage is explained by the first five PCs. Looking at the distribution of shapes in the plane of the first two PCs (Figure 12.5) allows us to anticipate the results: a significant difference between the subspaces. Indeed, for the comparison between juvenile and adult two-dimensional morphospaces, the between-stage angle is 83.83° and the 95% confidence intervals are 30.98° (juveniles) and 9.97° (adults). Increasing the dimensionality to five PCs yields results consistent with the conclusion based on two; the between-stage angle is 91.53° and the within-stage ranges are 49.23° (juveniles) and 85.19° (adults), so the two samples occupy different subspaces.

Comparing mouse skull shape hyperplanes between ages

Comparisons among subspaces of successive age-classes of *M. m. domesticus*, sampled at 5-day intervals, are more complex, because none of the variance–covariance matrices (through 25 days) have distinct eigenvalues. The first three PCs, taken together, account for only 50–60% of the variance of the two youngest age classes; it takes as many as five PCs to explain just 75% of the variance of the two youngest ages. We thus need at least four components to capture most of the variance of the younger stages, and five would be preferable. Therefore, we will compare both four- and five-dimensional subspaces.

There is an enormous range of within-age variation (Table 12.5). This is expected, because the within-age variation may be random (or nearly so), and under those conditions variation is nearly spherical so PCs are nearly arbitrary. Consequently, PCs of the resampled data may change a great deal from one iteration to the next, producing a very large range of within-age angles. Nevertheless, two of the comparisons indicate a significant difference between morphospaces; those between (1) 15- and 20-day-olds, and (2) 20- and 25-day-olds.

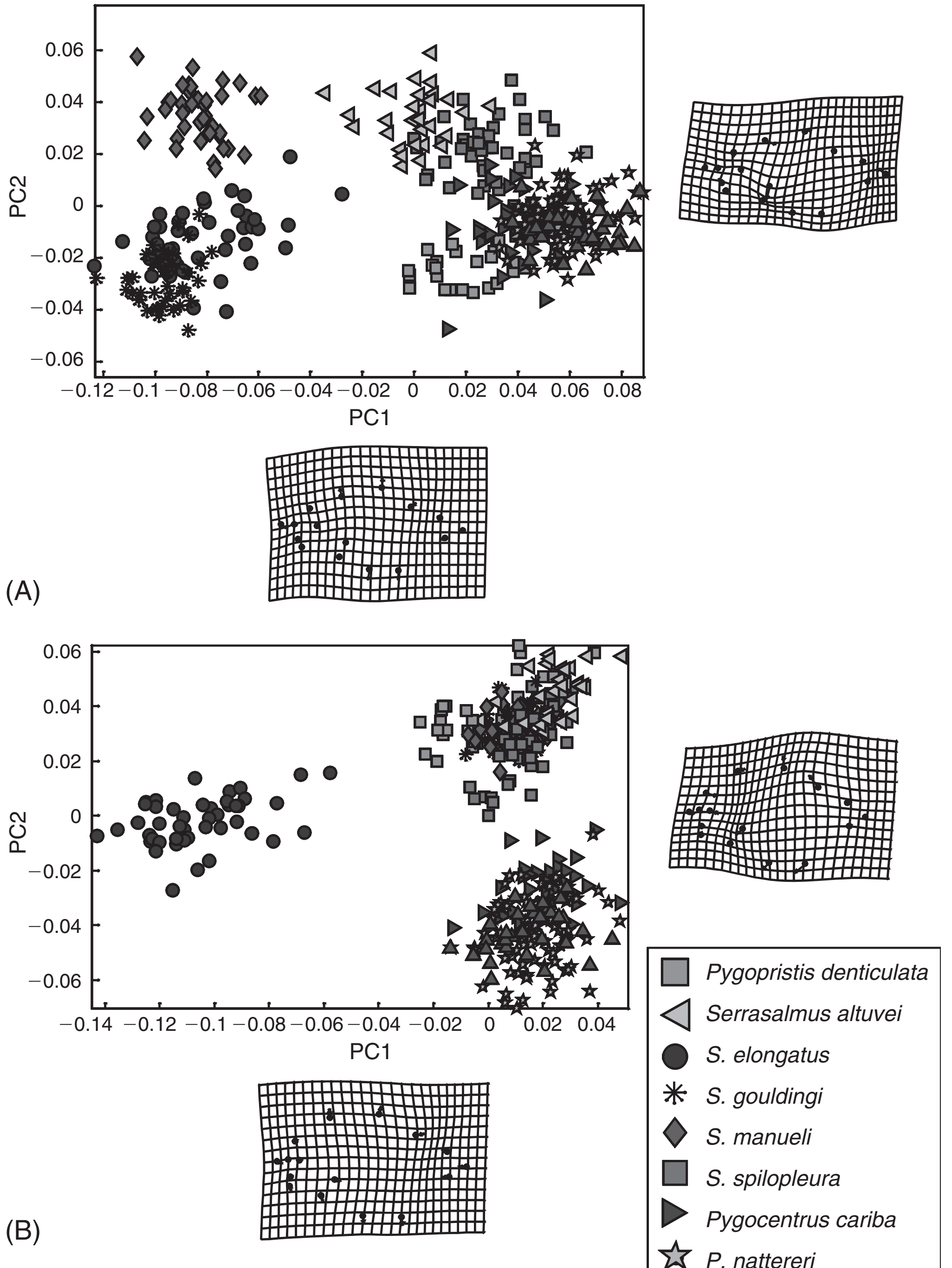


Figure 12.5 Principal components of piranha body shape: (A) juveniles; (B) adults.

Table 12.5 Comparing hyperplanes of skull shape variation between successive ages of the house mouse *M. m. domesticus* (the within-age ranges are calculated over 500 bootstraps)

Ages	*Four dimensions*			*Five dimensions*		
	Between-ages	*Within-age (younger)*	*Within-age (older)*	*Angle between*	*Within-age (younger)*	*Within-age (older)*
10–15	101.58	110.56	123.70	124.46	128.77	123.92
15–20	129.52	123.04	115.30	132.51	127.07	124.65
20–25	125.99	114.73	115.66	129.51	122.17	125.69

Software

Two programs in the IMP series are designed to implement analyses of disparity: **DisparityBox** (which calculates morphological disparity, *MD*, and partial disparity, *PD*, as well as within-group variance), and **SpaceAngle** (which calculates the angles between hyperplanes). To compare values of *MD* and variance (that is, to test the significance of the difference between two values of *MD* or two variances), use **T-Box** (described in the context of MANOVA in Chapter 9).

DisparityBox

DisparityBox takes input data, in standard (X1, Y1, ... CS) format, estimates the disparity or variance for geometric shape and also for traditional measurements (calculated from the landmark coordinates), and provides confidence intervals for the estimates. Estimates can be based on the input data, or the data can be standardized by regression (on the last variable in the data, usually CS). Two sorts of analyses are available, but the distinction between them does not correspond precisely to the distinction between within-population variance and between-group disparity. That is because a data set comprising the means of multiple species would correspond to a single group analysis (disparity is calculated as the variance over those individuals, even though the individuals are species' means). The basis for choosing the type of analysis is the kind of resampling design you wish to employ. If you want to remove individuals from species, thereby putting confidence intervals on disparity taking into account the effects of sampling each species, you are removing specimens from individual groups (not removing whole groups). This is the resampling scheme used in a "multi-group" analysis. In contrast, if you want to remove entire species from the analysis, thereby constructing confidence intervals that take into account the effect of sampling from the population of species, you are doing the kind of analysis that **DisparityBox** terms a "1-group analysis."

The two analyses are logistically very different, and some methods (or tests) are available only for one, so we explain how to conduct each. Features common to both the analyses are discussed in the context of multi-group analysis.

Multi-group analysis

Each group should be in a separate file (in standard X1, Y1, ... CS format). Load them one after another, by clicking on the **Load Data Set** button, loading the file, then clicking again on the **Load Data Set** button and opening the next file. As each is loaded, its superimposed landmarks will appear in the small visualization window on the upper right (the GLS superimposition is used). It is a good idea to ask for the list of files loaded (so you can keep track of the order in which they are loaded) by clicking on the **List Loaded Sets** button. To save that list, go to the **File** menu on the toolbar and click on **Save Results Box**. Before doing an analysis, you need to calculate the mean. A multi-group analysis uses the grand mean across groups, so click on **Find the Grand Consensus Mean (Groups)**. The other option is for analyses of within-sample variance.

If you want to analyze the disparity of size-standardized data, you can load the list of target sizes to which you wish to standardize shapes for each group (different values can be used for different groups). This requires preparing the target size list, which is the list of desired sizes, ordered in the same sequence as the species were loaded (i.e. the first size on the list is the target size for the first group loaded). The sizes should be in units of log transformed centroid size (to either base *e* or base 10). For example, the following list says to standardize the first group to 3.1 LCS, the second to 2.3 LCS, and the third to 3.4 LCS:

```
3.1
2.3
3.4
```

Load the list either by clicking on the **Load Log Size Targets** button, or go to the **File** menu on the toolbar and select the **Load Log Size Target** option.

If you wish to analyze the disparity of traditional morphometric measurements, you need to load a measurement protocol. This consists of a three-column list; the first column is the number of the measurement, the second is the number of the landmark that will serve as one endpoint of the measurement, and the third is the number of the landmark that serves as the other endpoint. For example, the measurement protocol for lengths measured between landmarks 1 and 7, between landmarks 2 and 4, and between landmarks 4 and 5, is:

```
1 1 7
2 2 4
3 4 5
```

(This is the same protocol used in the program **TradMorphGen**, which calculates traditional morphometric measurements given the landmarks and protocol – see Chapter 13). Load the list either by clicking on the **Load Length Protocol** button, or by going to the **File** menu on the toolbar and selecting the **Load Length Protocol** option.

A variety of analyses are available, listed on the **Multi-Group** pull-down menu on the toolbar. Your choice depends on whether you want to analyze geometric shape or traditional morphometric data, and on whether you wish to analyze untransformed data or size-standardized data. Additionally, you can ask for estimates of morphological disparity (*MD*) or for both *MD* and partial disparity (*PD*) of each group. Selecting your choice starts the program (it will take a long time).

The results will appear in the **Results Box** window. The confidence intervals are obtained by resampling (with replacement) within each group. To save the results to the same file in which you listed the loaded data sets, go to the **File** pull-down menu and select **Append Results Box to File**, then select the file in which you saved the list of loaded files. The program will warn you that the file will be overwritten, but it won't be. Alternatively, you can save them to a new file by selecting **Save Results Box**. Finally, you can copy the **Results Box** window by selecting all the text then copying it (Ctrl-C) and pasting that text into a text file (Ctrl-V).

In addition to the **Results** file you can save several others, including one that concatenates the separately loaded data files, a **GroupList** (to be used in **PCAGen** or **CVAGen**, see Chapter 7), and the size-standardized traditional morphometric data (if you input a protocol to obtain these measures and also a log target size list).

Multi-group nearest-neighbor analysis To do a nearest-neighbor analysis, you need to specify both the null model (uniform or Gaussian) and the range estimator (the parameters of the observed data or the Strauss–Sadler estimator). The default null is a uniform distribution, but you can ask for a Gaussian model instead, using the radio button labeled **NN Model Gaussian.** The default method for estimating the range is to use the parameters obtained from the data, but you can select the Strauss and Sadler method instead with the radio button labeled **Sadler Style Range**, just below the one for selecting the Gaussian null model.

To run the analysis, go to the **NN Model** pull-down menu, select either **Foote NN model** or **Size-Standardized Foote NN Model** (depending on whether you wish to size-standardize the data). Selecting either of these starts the program running. The results will appear in the **Results** window. They can be saved by going to the **File** pull-down menu and selecting **Append Results Box to File** if you wish to save the results to the same file in which you saved the list of loaded files and the results of the previous analysis. Again, the program will warn you that the file will be overwritten, but it won't be. You can save them to a different file by selecting the **Save Results** window. Finally, you can also copy the **Results** window (using Ctrl-C) and paste it into a text file (using Ctrl-V).

Single-group analysis of disparity/variance

To do the analysis of disparity within a single group, all specimens must be in the same file (this file is the unit of analysis). The file may comprise multiple individuals of a single population (in which case the analysis is of variance, not disparity), or each "individual" could be the mean shapes of a species (in which case the analysis is of disparity, not variance). If you plan to analyze several such files, you can load them all now, then specify the one you wish to analyze in the **Active Set** window. You can move up and down in that window, thereby progressing through a series of analyses. Before you can do an analysis, you need to calculate the mean; click on **Find the Grand Consensus Mean (Specimens)**. The other option is for analyses of among-group disparity when each group is in a separate file (comprising multiple observations of the same group). If you wish to size-standardize the data, or to analyze the variance of traditional morphometric data, follow the procedures for constructing and loading the target size files and length protocol files (see above).

A variety of analyses are possible; these are listed on the **1-Group Analysis** pull-down menu on the toolbar up top. You can choose to analyze the disparity of geometric shape within the group (**Bootstrap Disparity within Group**), the disparity of traditional measurements within the group (**Trace of the Trad Measures Var/Cov Matrix**) or the disparity of size-standardized geometric shape (**Bootstrap Size Corrected, Within Group Disparity**). Selecting an option starts the analysis.

The results will appear in the **Results Box** window, which gives the within-group disparity accompanied by the 95th percentile range. They can be saved either by going to the **File** pull-down menu and selecting the **Save Results Box**, or you can copy the window by selecting the text then copying it (Ctrl-C) and pasting it into a text file (Ctrl-V).

Single-group nearest-neighbor analysis As in the case of the multi-group analysis, you first need to choose your null model and range estimator. To perform the analysis, go to the **NN Model** pull-down menu on the toolbar and select **Foote NN Model (within Active Group)**. The results will appear in the **Results** window, and can be saved/copied as described above.

SpaceAngle

This program uses the algorithm by Jason Mezey to estimate the angle between hyperplanes. The program is not limited to analyses of geometric morphometric data, although that is the default. When landmark coordinates are loaded, the first step is to calculate partial warp scores; if your data are not coordinates of landmarks, you need to turn off the option to **Compute PW scores** (click on the radio button). Each sample must be in a separate file, in standard (X1, Y1, ... CS) format.

Before beginning the analysis, determine how many dimensions you wish to include in the comparison and type in that number where asked for the number of axes. To estimate the angle (without testing it), click on **Calculate Angle Between PC Planes**. To estimate that angle *and* test it for its statistical significance, click on **Calculate Range of Angles Within Groups**. If you want to place confidence intervals on the between-group angle click on **Calculate Confidence Int. on Angle**. Selecting one of these options runs the program.

The results will appear in the **Results** window, and can be saved/copied as described above.

To test whether the angles between pairs of hyperplanes are significantly different, load the first pair of data sets using the **Load Data Set 1** and **Load Data Set 2** buttons. Load the second pair of data sets by going to the **File** pull-down menu and selecting **Load Data Set 3** then **Load Data Set 4** (for a three-way comparison, A-B vs A-C, load A as data sets 1 and 3, load B as 2 and load C as 4). Next, set the number of bootstraps using the bootstrap control window. Now, start the calculation by going to the **More Stats** pull-down menu and selecting **Bootstrap Test of Difference in Angle**.

References

Anstey, R. L. and Pachut, J. F. (1995). Phylogeny, diversity history, and speciation in Paleozoic Bryozoans. In *New Approaches to Speciation in the Fossil Record* (D. H. Erwin and R. L. Anstey, eds) pp. 239–284. Columbia University Press.

Arnold, S. J. and Phillips, P. C. (1999). Hierarchical comparison of genetic variance–covariance matrices. II. Coastal-island divergence in the garter snake, *Thamnophis elegans*. *Evolution*, **53**, 1516–1527.

Berg, R. L. (1960). The ecological significance of correlation pleiades. *Evolution*, **14**, 171–180.

Cheverud, J. M. (1982). Phenotypic, genetic and environmental integration in the cranium. *Evolution*, **36**, 499–512.

Cheverud, J. M. (1984). Quantitative genetics and developmental constraints on evolution by selection. *Journal of Theoretical Biology*, **110**, 155–172.

Cheverud, J. M. (1995). Morphological integration in the saddle-back tamarin (*Saguinus fuscicollis*) cranium. *American Naturalist*, **145**, 63–89.

Ciampaglio, C. N. (2002). Determining the role that ecological and developmental constraints play in controlling disparity: examples from the crinoid and blastozoan fossil record. *Evolution & Development*, **4**, 170–188.

Ciampaglio, C. N., Kemp, M. and McShea, D. W. (2001). Detecting changes in morphospace occupation patterns in the fossil record: characterization and analysis of measures of disparity: *Paleobiology*, **27**, 695–715.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications*. John Wiley and Sons.

Flury, B. (1988). *Common Principal Components and Related Multivariate Methods*. John Wiley and Sons.

Foote, M. (1986). Developmental buffering as a mechanism for stasis. *Evolution*, **42**, 396–399.

Foote, M. (1990). Nearest-neighbor analysis of trilobite morphospace. *Systematic Zoology*, **39**, 371–382.

Foote, M. (1992). Paleozoic record of morphological diversity in blastozoan echinoderms. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 7325–7329.

Foote, M. (1993a). Contributions of individual taxa to overall morphological disparity. *Paleobiology*, **19**, 403–419.

Foote, M. (1993b). Discordance and concordance between morphological and taxonomic diversity. *Paleobiology*, **19**, 185–204.

Foote, M. (1994). Morphological disparity in Ordovician-Devonian crinoids and the early saturation of morphological space. *Paleobiology*, **20**, 320–344.

Foote, M. (1997). The evolution of morphological diversity. *Annual Review of Ecology and Systematics*, **28**, 129–152.

Foote, M. and Gould, S. J. (1992). Cambrian and Recent morphological disparity. *Science*, **258**, 1816.

Gavrilets, S. (1999). Dynamics of morphological diversification on the morphological hypercube. *Proceedings of the Royal Society of London, Series B*, **266**, 817–824.

Gould, S. J. and Garwood, R. A. (1969). Levels of integration in mammalian dentitions: an analysis of correlations in *Nesophantes micrus* (Insectivora) and *Oryzomys couesi* (Rodentia). *Evolution*, **23**, 276–300.

Hulsey, C. D. and Wainwright, P. C. (2002). Projecting mechanics into morphospace: disparity in the feeding mechanics of labrid fishes. *Proceedings of the Royal Society of London, Series B*, **269**, 317–326.

Kingsolver, J. G. and Wiernasz, D. C. (1991). Development, function, and the quantitative genetics of wing melanin pattern in *Pieris* butterflies. *Evolution*, **45**, 1480–1492.

Kluge, A. G. and Kerfoot, C. (1973). The predictability and regularity of character divergence. *American Naturalist*, **107**, 426–464.

Lande, R. (1980). The genetic covariance between characters maintained by pleiotropic mutations. *Genetics*, **94**, 314–334.

Marroig, G. and Cheverud, J. M. (2001). A comparison of phenotypic variation and covariation patterns and the role of phylogeny, ecology and ontogeny during cranial evolution of New World monkeys. *Evolution*, **55**, 2576–2600.

Maynard Smith, J. M., Burian, R., Kauffman, S. et al. (1984). Developmental constraints and evolution. *Quarterly Review of Biology*, **60**, 265–287.

Miller, A. I. and Foote, M. (1996). Calibrating the Ordovician radiation of marine life: implications for Phanerozoic diversity trends. *Paleobiology*, **22**, 304–309.

Olson, E. C. and Miller, R. L. (1958). *Morphological Integration*. University of Chicago Press.

Phillips, P. C. and Arnold, S. J. (1999). Hierarchical comparison of genetic variance–covariance matrices. I. Using the Flury hierarchy. *Evolution*, **53**, 1506–1515.

Polly, P. D. 2000. Geography and sample size in **P** matrix evolution: molar shape evolution in island populations of *Sorex araneus*. *Journal of Evolutionary Biology*.

Rohlf, F. J., Gilmartin, A. J. and Hart, G. (1983). The Kluge–Kerfoot phenomenon: a statistical artifact? *Evolution*, **37**, 180–202.

Roy, K and Foote, M. (1997). Morphological approaches to measuring biodiversity. *Trends in Ecology & Evolution*, **12**, 277–281.

Schaefer, S. A. and Lauder, G. V. (1996). Testing historical hypotheses of morphological change: biomechanical decoupling in loricariod catfishes. *Evolution*, **50**, 1661–1675.

Schluter, D. (1996). Adaptive radiation along genetic lines of least resistance. *Evolution*, **50**, 1766–1774.

Smith, L. H. and Lieberman, B. S. (1999). Disparity and constraint in olenelloid trilobites and the Cambrian radiation. *Paleobiology*, **25**, 248–272.

Sokal, R. R. (1976). The Kluge–Kerfoot phenomenon reexamined. *American Naturalist*, **110**, 1077–1091.

Steppan, S. J. (1997). Phylogenetic analysis of phenotypic covariance structure. I. Contrasting results from matrix correlation and common principal component analysis. *Evolution*, **51**, 571–586.

Strauss, D. and Sadler, P. M. (1989). Classical confidence intervals and Bayesian probability estimates for ends of local taxon ranges. *Mathematical Geology*, **21**, 411–427.

Van Valen, L. (1962). Developmental gradients in the dentition of *Peromyscus*. *Evolution*, **16**, 272–277.

Van Valen, L. (1970). An analysis of developmental fields. *Developmental Biology*, **23**, 456–477.

Wagner, G. P. (1988). The influence of variation and of developmental constraints on the rate of multivariate phenotypic evolution. *Journal of Evolutionary Biology*, **1**, 45–66.

Wagner, G. P. and Altenberg, L. (1996). Complex adaptations and the evolution of evolvability. *Evolution*, **50**, 967–976.

Wagner, P. J. (1995). Testing evolutionary constraint hypotheses with early Paleozoic gastropods. *Paleobiology*, **21**, 459–470.

Wagner, P. J. (1997). Patterns of morphologic diversification among the Rostroconchia. *Paleobiology*, **23**, 115–150.

Wills, M. A. (2001). Morphological disparity: a primer. In *Fossils, Phylogeny, and Form: An Analytical Approach* (J. M. Adrain, G. D. Edgecombe and B. S. Lieberman, eds) pp. 55–144. Kluwer Academic/Plenum Publishers.

Wills, M. A., Briggs, D. E. G. and Fortey, R. A. (1994). Disparity as an evolutionary index – a comparison of Cambrian and Recent arthropods. *Paleobiology*, **20**, 93–130.

Zelditch, M. L. (1988). Ontogenetic variation in patterns of phenotypic integration in the laboratory rat. *Evolution*, **42**, 28–41.

Zelditch, M. L. and Carmichael, A. C. (1989). Ontogenetic variation in patterns of developmental and functional integration in skulls of *Sigmodon fulviventer*. *Evolution*, **43**, 814–824.

Zelditch, M. L., Bookstein, F. L. and Lundrigan, B. L. (1993). The ontogenetic complexity of developmental constraints. *Journal of Evolutionary Biology*, **6**, 121–141.

Zelditch, M. L., Sheets, H. D. and Fink, W. L. (2003). The ontogenetic dynamics of shape disparity. *Paleobiology*, **29**, 139–156.

13

The relationship between ontogeny and phylogeny

According to one of the best-known aphorisms in evolutionary developmental biology, "ontogeny recapitulates phylogeny." The statement has been discredited numerous times, but evolutionary biologists remain intrigued by the idea that ontogeny and phylogeny might be related. Gould proposed an alternative to the theory of recapitulation, one that allows for the converse phenomenon: rather than a descendant completing and going beyond the ancestral ontogeny, some fail to complete it (Gould, 1977). In those cases, descendant adults resemble ancestral juveniles, in direct contradiction to the theory of recapitulation. However, a broader theory encompasses both cases: *parallelism* between ontogeny and phylogeny. According to the theory of parallelism, the descendant's adult morphology can be found either *in* the ancestral ontogeny or by extrapolation of it (Figure 13.1). Either the descendant adult looks like a subadult ancestor, or it looks like an overgrown one. Although termed parallelism, the direction of evolutionary change (Figure 13.1, the path labeled E) actually coincides with the direction of the ancestral ontogeny (Fig. 13.1, the path labeled O), it does not merely parallel it.

The explanation for parallelism is that species evolve in rates or timings of development, but otherwise retain the ancestral ontogeny. If development speeds up, or lasts for longer, the descendant will go further than the ancestor along the ancestor's ontogeny. Conversely, if development slows down, or lasts for less time, the descendant will not reach the ancestral endpoint. Such changes in developmental rate or timing are called heterochrony, and Gould (1977) devoted his entire book to the phenomenon. His work stimulated hundreds of studies of heterochrony, but Gould never claimed that heterochrony is a common phenomenon. Rather, he characterized it as a part of a broader subject, that broader subject being the relationships between ontogeny and phylogeny. Nevertheless, he found heterochrony especially interesting, and thus worthy of special attention, for two major reasons. First, it has an intriguing implication, which is that morphology evolves as an indirect effect of selection on life-history parameters. Should selection favor a younger age at sexual maturity, the ancestral ontogeny might be truncated so the descendant adult reaches sexual maturity with a larval morphology. Even though morphology itself is not the target

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

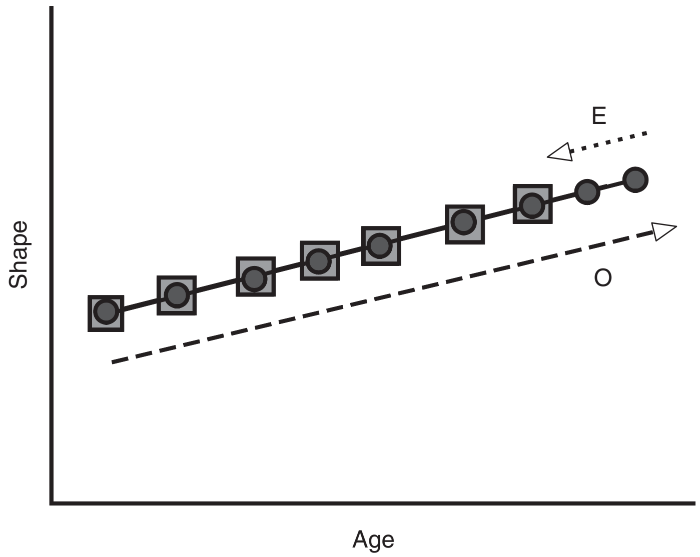


Figure 13.1 Parallelism. The ontogeny of the ancestor (indicated by the circles) and the descendant (indicated by the squares) begin development having the same shape and follow the same ontogeny of shape. They differ only in that the descendant's ontogeny is a truncated version of the ancestral ontogeny. Consequently, the descendant adult resembles an ancestral juvenile. In cases of parallelism, the direction of evolutionary change in adult morphology (E) coincides with the direction of the ancestral ontogeny (O).

of selection, it is nonetheless modified, sometimes dramatically, and those modifications require no adaptive explanation in their own right. Recognizing that they are simply a correlated effect of an evolutionary change in a life-history parameter can temper an excessive enthusiasm for adaptive explanations. Gould's logic follows the same line taken by Huxley (1932) in his discussion of evolutionary allometry – that shape evolves as an indirect effect of selection on size, so the changes in shape are explained by the changes in size. This connection between heterochrony and allometry is seemingly obvious, but Gould was the first to recognize and emphasize it.

A second reason for concentrating on the phenomenon of parallelism is that Gould sought to rehabilitate the concept of recapitulation. Even though he denied that recapitulation is a general rule, he thought the idea had been dismissed unfairly, and for reasons unrelated to the failure of the theory. Other workers had also attempted to rescue the idea of recapitulation, especially Cope (e.g. 1887), by applying it to individual parts (or measurements). Unlike them, Gould took an organismal, multivariate view of parallelism. He strongly opposed the trait-by-trait approach to morphology, whereby each individual organ (or measurement) is accorded its own explanation. Instead of that approach, which he called "atomistic," he favored viewing organisms as integrated entities, bound together by developmental correlations. Accordingly, rather than analyzing changes in growth rates of individual measurements, he modeled changes in properties of whole organisms, such as life-history parameters. His models are admittedly informal and verbal rather than mathematical, so they cannot be considered multivariate in a technical sense, but formalizing them requires a multivariate mathematical model. Gould did

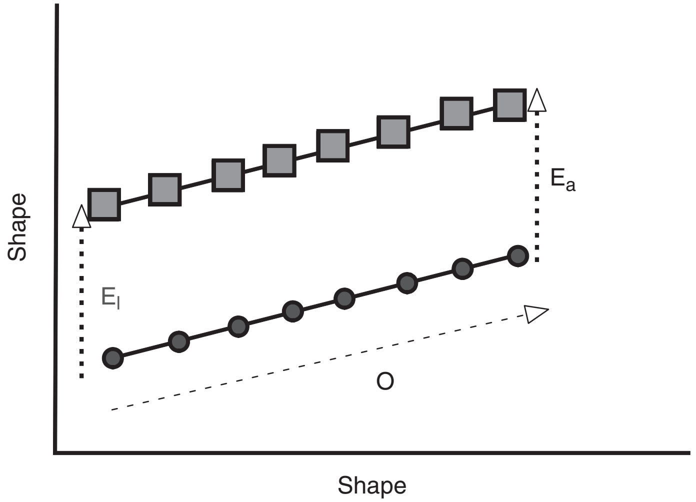


Figure 13.2 Changes confined to early morphogenesis. This pattern requires using two shape axes to depict the change; one oriented in the direction of the divergence in early morphogenesis, the other depicting the shared direction of subsequent ontogenetic change. The ancestral ontogeny is depicted by circles, the descendant ontogeny by squares. The two lines differ in elevation, a difference that is constant throughout ontogeny. Consequently, the direction of evolutionary change in larval morphology (E_l) parallels the direction of evolutionary change in adult morphology (E_a). These evolutionary transformations do not parallel the direction of the ancestral ontogeny.

not provide one, so the assumptions of his theory cannot be examined rigorously, but a more rigorous model for evolutionary allometry grounded in evolutionary theory was developed by Lande (1979) and extended to the more general multivariate case by Lande and Arnold (1983). The important point made by these models is that selection on body size, or any other life-history trait, will likely have indirect effects, but not necessarily in the direction anticipated by either Gould or Huxley. Nevertheless, despite that important flaw in Gould's model, he offered a multivariate, organismal view of allometry and heterochrony.

Gould acknowledged that heterochrony is just a part of a broader subject, deserving special attention because of its intriguing implications. However, parallelism is neither the only intriguing nor the only theoretically significant possibility. Another, which is equally remarkable, is shown in Figure 13.2 – the descendant's ontogeny is a vertically transposed version of the ancestral ontogeny. This pattern differs from parallelism because the descendant adult does not lie *along* the ancestral ontogeny – it lies above or below it. This pattern is difficult to draw (and perhaps also to read) because we need two shape axes to represent it: one, the *X*-axis, is in the direction of their shared ontogeny; the other, the *Y*-axis, is in the direction of the difference arising early in development. That the two lines differ solely in elevation means that the species diverge early in development but thereafter follow the same ontogeny. Consequently, as adults they differ in precisely the same features that distinguished them very early in development. This pattern is at least as intriguing as parallelism because of what it implies about the dynamics of evolving ontogenies.

It means that late development is more conservative than early. Conventionally, embryos are thought to pass through a "phylotypic period" – the stage at which embryos of all members of a phylum look the same (Seidl, 1960; Sander, 1983; Slack et al., 1993). Several studies have challenged this "hour-glass" model (e.g. Richardson et al., 1997; Bininda-Emonds et al., 2003) on the grounds that the phylotypic period is not as conservative as generally thought, but even these critiques do not go so far as to say that morphologies are more disparate early rather than later. Yet that is precisely the expectation implied by the two parallel lines. That is not to say that finding this pattern in morphometric data contradicts the theory of the phylotypic period, because most morphometric studies encompass far older stages (e.g. postnatal, postlarval). The phylotypic period begins at onset of neurulation and ends with somitogenesis (see Kimmel et al., 1995), so that is the phase that must be compared to later stages to test the theory. Nonetheless, it would still be surprising to find out that *all* divergence occurs by changes in late embryonic/early larval development.

Clearly, parallelism is not the only theoretically interesting or counterintuitive pattern that can be imagined. However, rather than seeking out counterintuitive possibilities, we might do better to look for the most general patterns. Focusing on a rare (albeit interesting) pattern can foster a highly biased view of the evolution of ontogeny. If our aim is to generalize about the evolution of ontogeny, we need to devote as much attention to common patterns as to rare ones. Given the enormous literature on the subject, heterochrony might seem to be a very common phenomenon, but that impression results partly from the broadened definition of heterochrony, which obviously increases the number of cases that satisfy it (see, for example, McKinney and McNamara, 1991). Sometimes, heterochrony is defined so broadly that it means nothing more than that ontogeny evolves. The theoretical significance of the concept is thereby diluted, and by classifying all the possibilities into a single category we lose the ability to recognize distinctions among them.

Our objective in this chapter is to describe a variety of patterns that can be found in comparative studies of ontogeny. To that end, we focus on the patterns amenable to discovery by comparative studies of ontogenetic allometry. Studies of allometry are sometimes viewed as a poor substitute for studies of heterochrony, but allometry is not just something we study when we have no information about age. Rather, comparative studies of allometry allow for a richer formalism than is feasible in studies of heterochrony because the formalisms for heterochrony were designed for cases of parallelism. They cannot be applied more generally without sacrificing a multivariate approach to the evolution of ontogeny.

Comparative analyses of allometry not only rely on a richer formalism; they also analyze a phenomenon that is interesting in its own right. Allometry is no less interesting than heterochrony, an argument we develop below. After motivating the study of ontogenetic allometry, we introduce the formalism for analyzing it and discuss the meaning (both formal and biological) of the coefficients obtained by that formalism. We next discuss methods for discerning patterns in the relationship between ontogeny and phylogeny, focusing on the significance of those patterns for our understanding of the evolution of development. The first part of this chapter focuses on traditional morphometric data because most comparative studies of allometry have relied on them. We then briefly review the geometric analysis of ontogenetic allometry (the subject of Chapter 10) and revisit the patterns introduced in context of traditional allometry, describing how these would appear in studies of geometric shape data.

Why allometry is interesting in its own right

For purely biomechanical reasons, we expect allometry because we expect organisms to change shape as they grow. Were they to grow without changing shape, they would likely decrease their ability to perform such vital functions as respiration, locomotion and feeding. Allometric scaling maintains functional equivalence. However, only certain very basic physical properties (such as surface area: volume relationships) might be expected to scale predictably over an entire ontogenetic series because young (small) organisms are often ecologically very different than older (larger) members of their own species. For that reason, they do not face the same functional demands. Nonetheless, they grow allometrically. That by itself is interesting; over an individual's life-time, it is increasing in size, changing shape, and also experiencing transitions in functional demands. At every age the organism must be competent, but it is continually changing. How these transformations in size, shape and function are interrelated is a central question in studies of ontogenetic allometry.

This focus on related transformations in size, shape and function arises from the perception that an organism's morphology is a continuum of shapes. To understand morphology we must not only understand the relationship between form and function at one age, we must also understand the entire continuum. That continuum can be described by a vector that extends from the form of the smallest to that of the largest individual. The optimal form for any one age might not lie along that vector, because the direction in which the vector points may represent a compromise among age-specific optima. The adult morphology may be especially important for determining the direction of the vector; even though ontogeny is not just the vector pointing towards an optimal adult, adult morphology could conceivably matter as much as the pathway. That is because the adult shape may be stable once it is attained; in organisms that have determinate morphogenesis (whether or not they also have determinate growth) the adult form persists for much of the life cycle. Yet that adult shape will never be reached if organisms do not first survive the especially vulnerable pre-adult phases. Thus, the direction in which the vector points, as well as the rate at which it develops, may be related, in part, to age-specific mortality rates and to the processes responsible for mortality. When mortality is largely a result of predation, and predation rates are particularly high early in life, the ontogenetic vector may be oriented towards development of anti-predator defenses early in life, with consequences for morphology at later ages.

Allometry is interesting not only because of its implications for form–function relationships, but also because of the insight it offers into growth and development. Those processes cause the changes in size and shape recorded in studies of allometry. Often we know much about the processes causing those changes, especially in studies of vertebrate skeletal form. In this case, the processes we are studying are the spatiotemporal dynamics of bone growth. From the coefficients we obtain in studies of allometry, we can learn about the spatial distribution of relative growth rates. Evolutionary changes in the spatiotemporal dynamics of growth can be discovered by comparative studies of allometry.

Formalisms for the analysis of ontogenetic allometry: traditional morphometric data

The traditional formalism

The traditional formalism for the study of allometry relates the increase in size of one part (Y) to that of another (X). Often, X is intended to represent the size of the whole organism. To make our discussion of allometry as concrete as possible, and to ease the transition from geometric to traditional morphometric data, we will focus on the case of the piranha, *Serrasalmus gouldingi*, which we used as our model for the geometric analyses of ontogenetic allometry. To analyze its ontogenetic allometry we measure a variety of lengths and depths (Figure 13.3). For our measure of body size we will use the measurement extending from landmark 1 to landmark 7, which is termed *standard length* (SL) and is frequently used as the measurement of body size in studies of teleosts – so, for our example, $X = SL$. The other 29 measurements are the measures 2–30, which we will represent by the vector $\{Y_1, Y_2, Y_3, \ldots Y_{29}\}$. We first discuss the mathematical analysis

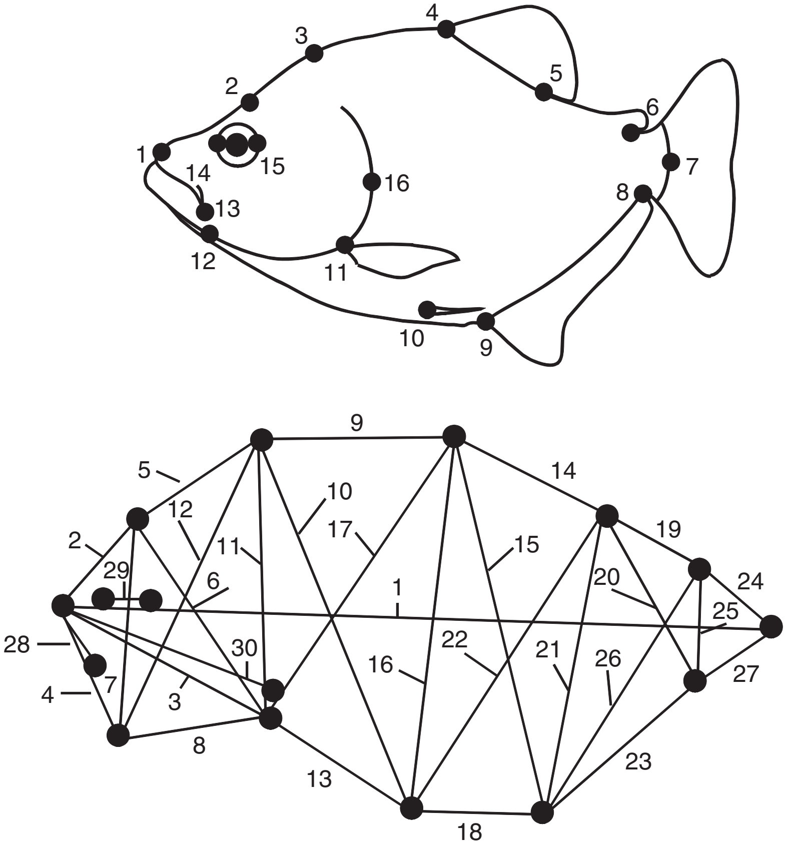


Figure 13.3 Landmarks sampled on *Serrasalmus gouldingi*, and the traditional morphometric measurement scheme based on those landmarks.

of allometry, then follow this with an interpretation of the coefficients obtained by the analysis, and then consider their developmental significance.

The mathematical analysis of allometry

The relationship between X and Y often fits a model, the power law (Huxley, 1932):

$$Y = bX^k \qquad (\mathbf{13.1})$$

where k is the rate of growth of part Y relative to X, and b is the size of Y when X is at unit size. To ease fitting the model to data, it is often rewritten in a linear form:

$$\log(Y) = \log(b) + k\log(X) \qquad (\mathbf{13.2})$$

Expressed in this form, we can use linear regression to estimate the parameters b and k; they are the intercept and slope (respectively) of a linear regression of log(Y) on log(X). Table 13.1 gives the regression coefficients, b and k, of the variables shown in Figure 13.3

Table 13.1 Allometric coefficients for *Serrasalmus gouldingi*; b is the intercept term, k is the slope (measurements are shown in Figure 13.3)

Variable	*b*	*k*
v2	−0.939	0.806
v3	−0.613	0.885
v4	−1.163	0.850
v5	−2.643	1.225
v6	−1.396	1.042
v7	−1.512	1.002
v8	−1.383	0.928
v9	−1.974	1.104
v10	−1.761	1.220
v11	−1.931	1.198
v12	−1.595	1.136
v13	−2.378	1.186
v14	−2.228	1.116
v15	−1.681	1.210
v16	−1.781	1.225
v17	−1.551	1.171
v18	−2.572	1.170
v19	−1.938	1.085
v20	−1.685	1.104
v21	−1.991	1.225
v22	−1.834	1.217
v23	−1.473	1.072
v24	−1.674	0.939
v25	−2.686	1.120
v26	−1.558	1.129
v27	−1.815	0.898
v28	−1.231	0.811
v29	−1.160	0.751
v30	−0.635	0.903

Table 13.2 Allometric coefficients, *k*, computed by multivariate regression (R) and PCA (variable 1 is the independent variable in the multivariate regression; its value of *k* must be included to make these vectors comparable)

Variable	*R*	*PCA*
v1	1.000	0.694
v2	0.806	0.559
v3	0.885	0.614
v4	0.850	0.589
v5	1.225	0.851
v6	1.042	0.723
v7	1.002	0.696
v8	0.928	0.644
v9	1.104	0.766
v10	1.220	0.847
v11	1.198	0.832
v12	1.136	0.789
v13	1.186	0.823
v14	1.116	0.775
v15	1.210	0.840
v16	1.225	0.850
v17	1.171	0.813
v18	1.170	0.812
v19	1.085	0.753
v20	1.104	0.766
v21	1.225	0.850
v22	1.217	0.845
v23	1.072	0.744
v24	0.939	0.651
v25	1.120	0.777
v26	1.129	0.783
v27	0.898	0.623
v28	0.811	0.563
v29	0.751	0.522
v30	0.903	0.627

regressed on *SL*. We should note that the literature is inconsistent on the symbols used for these two coefficients.

Usually, the coefficients are estimated by simple bivariate regression (both ordinary least squares, OLS, and reduced major axis, RMA, regression). As shown in Chapter 10, multivariate least squares regression yields the same estimates as obtained from bivariate analysis so we can treat the bivariate estimates of *k* as components of the vector $\{k_1, k_2, k_3, \ldots k_P\}$ (where *P* is the number of measurements) and those of log(*b*) as components of the vector $\{\log(b_1), \log(b_2), \log(b_3), \ldots \log(b_P)\}$. Another common multivariate approach is principal components analysis (PCA); Jolicoeur (1963) first proposed that PC1 is a multivariate allometry vector when PC1 is extracted from a variance–covariance matrix of log-transformed measurements. Conceptually, multivariate regression and PCA differ in that PCA does not single out one variable as independent. Instead of treating one of

Table 13.3 Ratios between allometric coefficients, *k*, of each variable and standard length (v1), for the coefficients computed by multivariate regression (R) and PCA

Variable	*R*	*PCA*
v2	0.81	0.81
v3	0.89	0.88
v4	0.85	0.85
v5	1.23	1.23
v6	1.04	1.04
v7	1.00	1.00
v8	0.93	0.93
v9	1.10	1.10
v10	1.22	1.22
v11	1.20	1.20
v12	1.14	1.14
v13	1.19	1.19
v14	1.12	1.12
v15	1.21	1.21
v16	1.23	1.22
v17	1.17	1.17
v18	1.17	1.17
v19	1.09	1.09
v20	1.10	1.10
v21	1.23	1.22
v22	1.22	1.22
v23	1.07	1.07
v24	0.94	0.94
v25	1.12	1.12
v26	1.13	1.13
v27	0.90	0.90
v28	0.81	0.81
v29	0.75	0.75
v30	0.90	0.90

the observed variables as the measure of size, PCA constructs a multivariate size measure from the observed variables; scores on PC1 are the measures of size. Also, OLS multivariate regression presumes that the independent variable is measured without error, whereas PCA does not.

Regression (both OLS and RMA) and PCA tend to give very similar results when measurements are highly correlated, which they usually are in studies of ontogenetic series. For example, Table 13.2 shows the estimates of the slope for the measurements of *S. gouldingi* obtained by Model I regression (OLS) and PCA. The numbers may appear to be quite different, but these differences disappear when the coefficients are rescaled so that each is the ratio between the *k* for one variable and the *k* for *SL*. Because *SL* is the independent variable, $k_{SL} = 1$.

Rescaling the coefficients of PC1 by dividing each coefficient by that for *SL* gives the values shown in Table 13.3. The estimates obtained by multivariate regression and PCA are identical. Perhaps the most important distinction between regression and PCA is that PCA does not provide estimates of b.

Interpreting allometric coefficients

The interpretation of k is straightforward – it is the growth rate of one measurement relative to that of a standard, the growth rate of X. When k is 1.0, the growth of the part keeps pace with that of X, which we will take as the whole body – i.e. their proportions are constant throughout growth. Such measurements are termed "*isometric*." When k is greater than 1.0, the part increases its size relative to overall body size; these parts are termed "*positively allometric*." When k is less than 1.0, the part decreases in its size relative to body size; these measurements are termed "*negatively allometric*." Only one measurement in Table 13.1 is isometric, but many coefficients are equal to each other, so we could view them as isometric *relative to each other*. We are not constrained to think of k_i solely in terms of growth rates of each part relative to body size – all ratios among the k_s are relative growth rates as well. Several measurements are isometric relative to each other, including the four measurements of body depth (measured from landmarks 4 and 5, which are at the anterior and posterior bases of the dorsal fin). These four (v15, v16, v21 and v22) grow at equal rates relative to each other, so *relative to each other* their proportions do not change over ontogeny. All four are positively allometric relative to body length, so the body (in that region) deepens relative to its length. Among the negatively allometric measurements are the most anterior lengths (v2, v3, v4, v8, v28, v29, v30) and the two most posterior ones (v24, v27). This means that measurements in the anterior head and caudal regions shorten relative to the whole body (of course they do not actually shorten – they lengthen in an absolute sense, it is just that they shorten relative to the length of the body). Consequently, the head and caudal region form a relatively smaller fraction of body length in adults than in juveniles.

The interpretation of b is less straightforward, and there has been some controversy about its biological meaning. One reason for doubting that b has any general biological significance is that its value depends on the units of measurement; unlike k, b is not a dimensionless quantity. However, a more important one is that $\log(b)$ is the value of $\log(Y)$ when $\log(X)$ is zero, a size at which Y might not yet exist. For example, when the body is 1 mm long, the dorsal fin might not have developed yet so it cannot have a meaningful size. Additionally, $\log(b)$ is estimated under the assumption that k is constant from $\log(X) = 0$, not just that it is constant over the range of values actually sampled.

Under one condition, b does have a simple interpretation. When populations do not differ in k, a difference in b does have a meaning because the difference in b will persist throughout the entire ontogeny. Even if we hesitate to infer a value for $\log(Y)$ when $\log(X) = 0$, under the condition that both species have the same value of k, the regression lines would differ only in elevation. So *at any point* in ontogeny, they will differ in elevation according to their difference in b. We might reasonably hesitate to claim that they have diverged by the time that $\log(X) = 0$, but we could claim that the difference between the species arose prior to the stage when we first observe them, and it is invariant throughout the rest of ontogeny. That difference in b says how those populations will differ *at any given value of* X. For example, if we are comparing the brains of several species that do not differ in k, we can determine their relative brain size *at all body sizes* by comparing

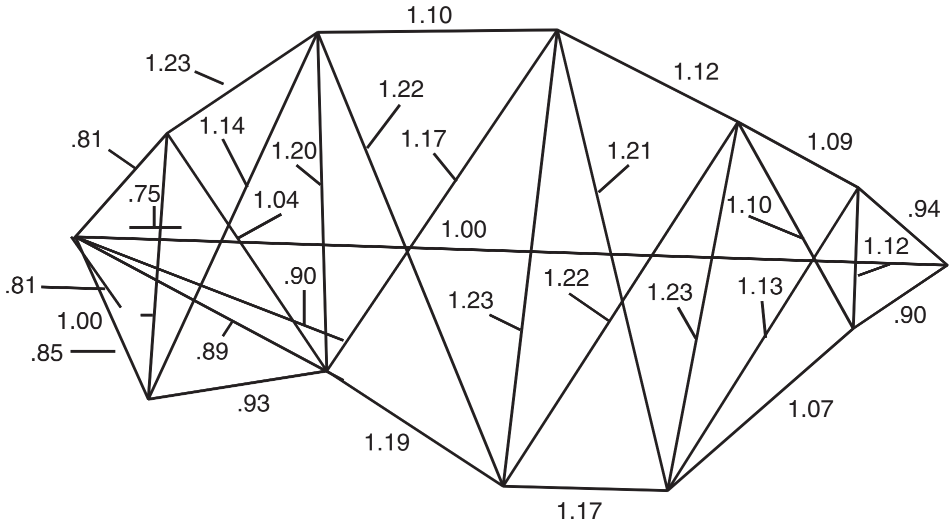


Figure 13.4 Allometric coefficients of *Serrasalmus gouldingi*. Coefficients significantly higher than 1.0 indicate positive allometry, coefficients significantly lower than 1.0 indicate negative allometry, and those between that cannot be distinguished statistically from 1.0 indicate isometry.

b; in this specific context, *b* has been termed an "index of cephalization" (see White and Gould, 1965). When *k* does not differ, *b* can be viewed as a scaling parameter. Under other conditions, *b* is just a parameter needed to predict *Y* at a given value of *X*.

The developmental meaning of b *and* k

Having defined *b* and *k* mathematically, and discussed how to estimate and interpret them, we can now consider their developmental meaning. Most of the theoretical literature has focused on *k* because *b* is static – it is not a descriptor of development, just of where the regression line intersects the *Y*-axis. At the heart of the literature is the view of growth as a multiplicative process. This was the rationale given by Huxley (1932) for the power law, and it is the basis for cellular models of allometric growth (e.g. Katz, 1980). Within that context, the meaning of *k* has been viewed from both spatial and temporal perspectives.

Huxley (1932) emphasized the spatial interpretation of *k*, proposing that changes in *k* over the organism indicate spatially organized "growth intensities." He noted that values of *k* tend to be spatially coherent, rising and falling in organized patterns across the body. To help visualize spatial patterns in *k*, we can first put the coefficients on the organism rather than in a table (Figure 13.4). We can see that they increase from the head to the middle of the body, then fall towards the tail, although not to a level as low as found in the head. This is (approximately) an inverted U-shaped gradient, which is interesting because it is the inverse of the gradient found in several teleost larvae (Fuiman, 1983). This suggests that the allometry of juvenile growth, in effect, compensates for that of larval growth: the head and caudal body initially grow very rapidly, and the middle of the body catches up during juvenile growth. Growth rates can also suggest anteroposterior gradients, which means that they fall off linearly from the head to the tail. To analyze

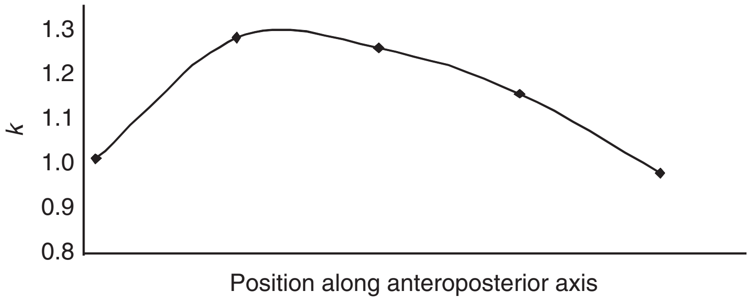


Figure 13.5 Growth profile for relative growth rates along the anteroposterior axis of *S. gouldingi*. Allometric coefficients for measurements of growth along the anteroposterior axis are plotted as a function of the ordinal position of the measurement (anterior is to the left of the plot, posterior to the right).

these patterns more rigorously Huxley constructed "growth profiles," which are plots of allometric coefficients as a function of their position along body axes. Unfortunately, few studies record the position of a measurement along a body axis (they were not recorded for the measurements shown in Figure 13.3, for example). As a crude approximation we can order the measurements in a linear sequence, ordered from anterior to posterior, and plot allometric coefficients at each ordinal position (Figure 13.5). Although it would clearly be better to have more accurate estimates of position along the body, the plot is nonetheless intriguing because it suggests a spatial ordering to the coefficients. Remarkably few recent studies have used growth profiles to understand developmental spatial patterning or its evolution (one exception is the analysis by Fuiman mentioned above, another is Zelditch et al., 2001).

Laird and colleagues have stressed the temporal significance of allometric coefficients (e.g. Laird 1965; Laird et al., 1968). Even though time is not explicitly incorporated in studies of allometry, it is nonetheless implicit. This becomes evident when considering why the power law holds in the first place. As mentioned before, the primary biological explanation for allometry is that growth is a multiplicative process. When analyzing the relationship between size and time, the best-fitting models are usually not linear but rather are sigmoidal in form. An important feature of these models is that growth rates decay over time. Similarities in decay rates are interpreted by Laird (1965) as the explanation for the linear relationship among log-transformed measurements. In effect, all measurements follow the same growth curve; their differing values of k tell us how they are displaced relative to each other in time – different parts of the body reach the same point on their growth curves at different times. Laird et al. (1968) elaborated on this theory, stating the relationship between k and lag time (ΔT) as:

$$\Delta T = -\frac{1}{\alpha}\ln(k) \tag{13.3}$$

where α is the decay rate and k is an allometric coefficient.

We cannot measure decay rates without information on age, but we can use Equation 13.3 to understand the temporal relationships among growth curves so long as we are willing to assume that decay rates are the same for all measurements whose logs are linearly related. Because growth rates decay over time, we would intuit that a more negatively allometric part has decayed over a longer time, and that it has decayed for longer because it began growing earlier. The increment of time by which we need to shift one curve to match another that starts growing later is ΔT. Based on this interpretation of allometric coefficients, we would conclude that the head and caudal peduncle develop before the midbody, that the eye is the first structure to develop, and that the body elongates before it deepens.

The spatial and temporal perspectives on allometric coefficients are not antagonistic. The spatial coherence noted by Huxley, interpreted within the temporal framework of Laird, suggests that growth is spatiotemporally organized. There is no reason to think that either space or time is primary. We do not need to adopt one view over the other – they are mutually consistent, and help explain each other. With increasing information about the spatial determination of development, in conjunction with that on its temporal organization, we can relate allometric coefficients to the underlying developmental processes that explain them. Because these theories of developmental controls over the spatiotemporal organization of relative growth may be most easily expressed in terms of traditional morphometric measurements, studies of allometry using traditional morphometric measurements will remain an important part of evolutionary developmental biology.

Of course, allometric coefficients are also informative about the relationship between form and function. The literature on biomechanics is filled with theories that predict scaling relationships among measurements. Applied to ontogenetic series, such theories may explain ontogenetic allometry in terms of the ontogeny of function. For example, in many larval teleosts the head and caudal region are highly positively allometric, which is due to the early demands imposed by swimming, feeding and respiratory systems (see, for example, van Snik et al., 1997). The converse allometric pattern is seen later, in juvenile growth, as exemplified by the coefficients of *S. gouldingi*. These patterns are hardly surprising, which is reassuring if our aim is to make sense of ontogenetic allometry in functional and ecological terms.

Comparative analysis of ontogenetic allometry: traditional morphometric data

Comparative studies of ontogenetic allometries serve two primary purposes. First, they test general theories about the relationship between form, function and development. For example, considerations of function suggest that, generally, teleosts ought to share the positive allometry of the head and caudal body during larval growth, with postlarval growth being characterized by positive allometry of the region between. Comparative studies can test that general hypothesis, and pursue more refined theories of the ontogeny of function should the expected pattern be less general than anticipated. Second, comparative studies test theories about the evolution of development and the impact of evolving developmental systems on morphological diversity (= disparity; see Chapter 12). For example, we might anticipate that early development is more conservative than later development because modifications of early development are likely to have dramatic consequences for later phases. Of course, we would not anticipate that this would be the case for organisms that have distinct metamorphic phases, because metamorphosis can decouple phases of development. Consequently, each could evolve independently. Therefore we could state the theory as comprising two parts: (1) in organisms that have continuous development, early stages are more conservative than later ones; and (2) early development is more conservative in organisms that have continuous development than in those with distinct metamorphoses. Testing such general theories about the evolution of growth and morphogenesis, including theories about heterochrony, is another major rationale for comparative allometric studies.

Below we have classified a variety of evolutionary patterns according to the changes in development producing them and the resultant relationship between ontogeny and phylogeny. We first discuss parallelism or channeling (for the remainder of this chapter we will use the term "channeling" rather than parallelism, because we are talking about coincident rather than parallel lines). This pattern is characterized by species that have the same shape at the outset of differential growth, as well as a common direction of allometry. Biologically, channeling results from changes in the rates or timings of development along a conservative ancestral ontogeny. The second class includes modifications of the ontogenetic trajectory itself – not just its relationship to growth or age. These involve changes in the spatiotemporal organization of development, and we can subdivide them according to the phases affected: (1) changes confined to early morphogenesis (which we typically infer only by their affects on proportions at the outset of allometric growth); (2) changes in the spatiotemporal organization of growth from the outset to end of allometric growth; and (3) changes in spatiotemporal organization of growth confined to late development. There is an additional, heterogeneous category: complex changes in multiple processes and stages.

Below, we describe several patterns in more detail, focusing first on the relationship between ontogeny and phylogeny engendered by each and their consequences for disparity, and finally on the criteria whereby they can be distinguished empirically.

Channeling

Channeling refers to the case in which the descendant adult morphology lies along the ancestral ontogeny because the ancestral and descendant ontogenies begin at the same starting point and proceed in the same direction. Representing this pattern in terms of a plot of log(Y) on log(X), channeling occurs when the descendant merely extends or truncates the ancestral ontogeny; graphically, scaling can be seen in plots of log(Y) on log(X) – both ontogenetic series start at the same point and lie on the same line, it is just that one species extends that line further (Figure 13.6A). Consequently, the direction of evolutionary change lies along the vector describing the ancestral ontogeny (Figure 13.6B). This is the pattern that Gould calls "parallelism" (Gould, 1977), and it is expected when the species differ solely in rate or timing of either growth or development. It implies that rate or timing evolve, but the ancestral pattern of relative growth rates (both larval and postlarval) is conserved.

We subdivide channeling into two types because growth and development can either be associated or decoupled. If they are associated, the descendant has the shape expected for its size – e.g. it looks young for its age because it is small for its age. Organisms that look young for their age are called paedomorphic (from "child-like"). Descendants might also look old for their age because they are large (they are called peramorphic because

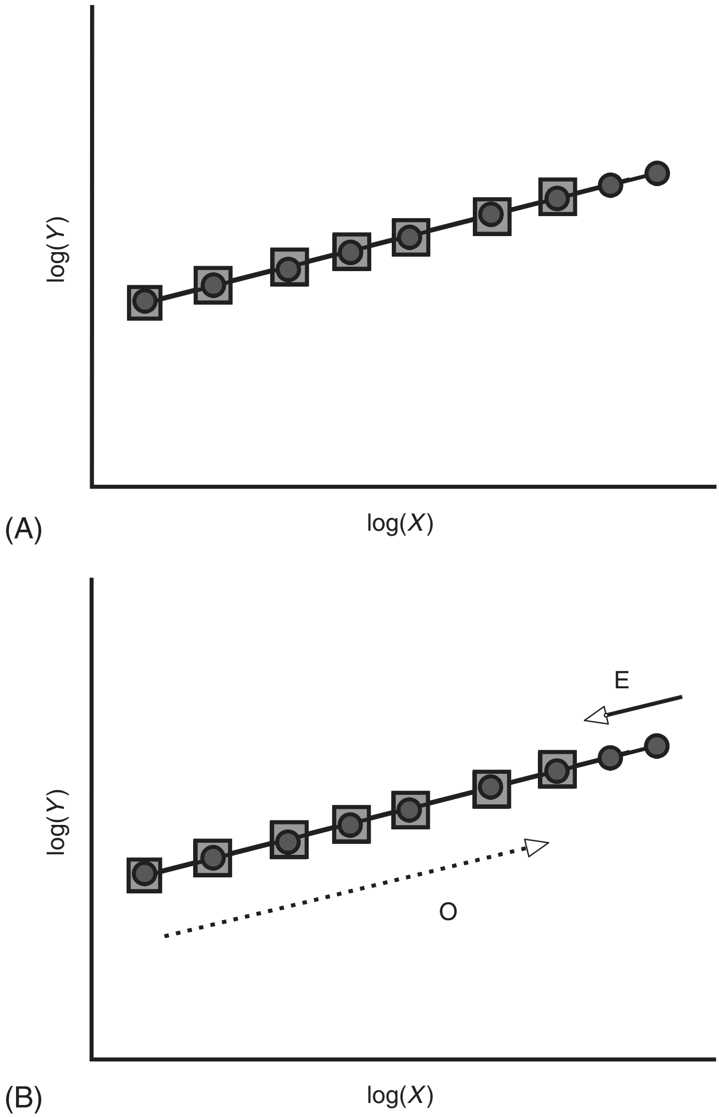


Figure 13.6 Channeling by ontogenetic scaling, depicted by a bivariate plot of log(Y) on log(X). The same pattern will be found for all the measured Y variables, thus the pattern can be represented by a single bivariate plot. (A) Ontogenetic allometries of ancestor (circles) and descendant (squares); (B) the directions of ontogenetic (O) and evolutionary change (E).

they go *through* and beyond the endpoint of the ancestral ontogeny). In the latter case, the descendant adult morphology does not actually appear *in* the ancestral ontogeny, but it results from extending it. However, it is possible to be paedomorphic and large, and also to be peramorphic and small. These are cases in which growth and development are decoupled – for example, developmental rates might be reduced without a corresponding reduction in growth rates, so the descendant reaches maturity before it attains the ancestor's adult morphology even though it reaches the ancestor's adult body size. The empirical criteria for documenting ontogenetic scaling differ in an important respect from those that document channeling via decoupling of growth from development. For that reason, we discuss the empirical criteria separately.

Empirical criteria for documenting ontogenetic scaling

Of all the various modifications of development, ontogenetic scaling is the easiest to detect, at least in principle. The biological hypothesis predicts that species differ only in adult body size, not in either $\{b_1, b_2, b_3, \ldots b_P\}$ or $\{k_1, k_2, k_3, \ldots k_P\}$. For the remainder of this chapter, we will refer to $\{b_1, b_2, b_3, \ldots b_P\}$ rather than $\{\log(b_1), \log(b_2), \log(b_3), \ldots \log(b_P)\}$ to simplify the presentation. Given this hypothesis, we can use MANCOVA to test it. However, given our expectations of no evolutionary change in either vector, using MANCOVA poses a problem – the substantive biological hypothesis is equivalent to the statistical null. Normally we think of the null as the hypothesis we would like to reject, and we use various strategies to ensure that we do not reject it too readily. However, the hypothesis of scaling is the one we wish to *accept*, so we are put in an odd position. One consequence is that all the factors that normally can prevent us from rejecting a false null hypothesis, such as lack of statistical power/small sample size, favor accepting a false hypothesis of scaling.

A second problem that arises in many studies is that many of the variables imply scaling, but a few others do not. Viewed from a bivariate perspective, this means that many traits undergo scaling but a few do not. However, viewed from a multivariate perspective, it means that the organism is not a scaled up or down version of its ancestor because not all traits are scaled up or down. It also means that either early or late development is altered, depending on whether the departures from scaling involve $\{b_1, b_2, b_3, \ldots b_P\}$ or $\{k_1, k_2, k_3, \ldots k_P\}$ (and they could both be involved). Whether those modifications are trivial or consequential depends on how far they deviate from scaling, and what their consequences are for morphology. We therefore need measures of the deviations from the expectations under the hypothesis, as well as measures of the impact of those deviations on $\{y_1, y_2, y_3, \ldots y_P\}$.

Because the hypothesis predicts that both species will have the same vector $\{k_1, k_2, k_3, \ldots k_P\}$, we would predict that the interspecific angle between those vectors will be 0.0°; we can test that hypothesis using the methods introduced in Chapter 10. However, we also need to have a feel for the magnitude of the angle. In analyses based on traditional morphometric data, angles will rarely be large, even if the vectors are no more similar than expected by chance. Being no more similar than expected by chance does *not* mean that the angle will be 90°. To understand why that is the case, it is necessary to recall how those angles are calculated, as well as to appreciate the meaning of k. The angles are computed by taking the dot product between the two vectors, which means we multiply k_1 of one species by k_1 of the other, and add that to the product of k_2 in one species by k_2 of the other, and so forth; so we are summing products of corresponding allometric coefficients. These coefficients are the power to which body size is raised in the power law (Equation 13.1). So long as structures grow over ontogeny, k is invariably a positive number. The coefficients rarely differ by much – the most extreme values for *S. gouldingi* are 0.75 (for eye diameter) and 1.23 (for midbody depth and posterodorsal head length). That difference may seem very large because one is highly negatively allometric whereas the other is highly positively allometric, but the difference is still numerically very small (in this case, it is less than 0.5).

Because all the elements of both vectors are positive numbers, the sum of their products cannot be zero – much less negative. To produce an angle of 90°, the sum would have to be zero (because the cosine of 90° is zero). The angle is necessarily smaller than that – often very much smaller. To appreciate how small it can be when vectors are independent, we can compare a vector of ontogenetic allometric coefficients to 400 random permutations of it. In the case of *S. gouldingi* the average vector correlation over those 400 random permutations is 0.9812 with a confidence interval of 0.9754–0.9874, corresponding to an angle of 11.13° with a confidence interval of 9.10°–12.74°. Therefore, only correlations higher than 0.9874 (or, equivalently, angles smaller than 9.10°) indicate any greater similarity than expected by chance. It may seem unreasonably strict to insist that the angle cannot be much greater than 0°, but in light of the large differences implied by very small angles, we cannot document scaling if angles much exceed 0°.

If the vectors of allometric coefficients are different, we cannot compare the vectors of b by MANCOVA. However, we do not actually need to compare them because we have already rejected the hypothesis of scaling by finding a significant difference in $\{k_1, k_2, k_3, \ldots k_P\}$.

Empirical criteria for documenting a decoupling of growth from development

When growth and development are decoupled, we anticipate changes in $\{k_1, k_2, k_3, \ldots k_P\}$, but only those consistent with the hypothesis that development is conserved. The specific expectations may seem strikingly counterintuitive; rather than simply stating them, we will derive them from the meanings of paedomorphosis and peramorphosis. Paedomorphosis means that the descendant adult resembles the ancestral juvenile morphology – over ontogeny it does not depart as far from the shared juvenile morphology. The extreme case is an adult that does not depart *at all* from the juvenile morphology, meaning its growth is isometric (it does not change shape as it grows). In less extreme cases, we expect the descendant ontogeny to be more nearly isometric than the ancestor's. Based on that reasoning, we can predict that positively allometric coefficients will *decrease* in slope whereas negatively allometric coefficients will *increase* because positively and negatively allometric coefficients approach isometry from opposing directions. An example of a pattern expected under this hypothesis is shown in Figure 13.7.

Of course, the coefficients must all change by the appropriate amount. Considering that paedomorphosis results from truncating the ancestral ontogeny, and peramorphosis from extending it, we would anticipate *only* a change in the length of the ontogenetic trajectory – the two vectors of allometric coefficients should otherwise be the same. Thus, we anticipate an angle of 0.0° between vectors. Framing the hypothesis of a conserved larval morphology is more complicated because we cannot test the coefficients $\{b_1, b_2, b_3, \ldots b_P\}$ owing to the differences in $\{k_1, k_2, k_3, \ldots k_P\}$. Above, we determined that we do not need to test the hypothesis that $\{b_1, b_2, b_3, \ldots b_P\}$ does not differ because the differences in $\{k_1, k_2, k_3, \ldots k_P\}$ rule out the hypothesis of scaling. However, we are no longer asking whether size predicts shape. It is possible that species do not differ in proportions at the smallest observed stage (even if they differ in proportions at any given size). Thus, instead of comparing $\{b_1, b_2, b_3, \ldots b_P\}$ we compare the expected shapes at a comparable

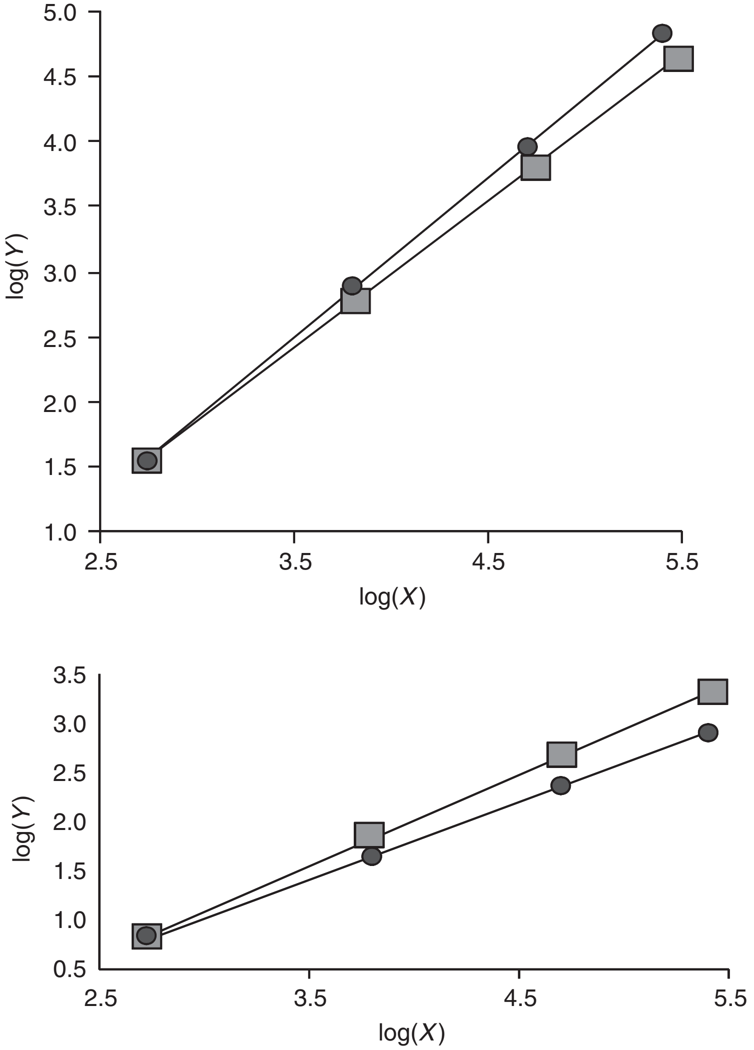


Figure 13.7 Channeling when growth and development are decoupled, depicted by a bivariate plot of log(Y) on log(X). Two plots are required to depict such cases, such as that of neoteny (a decrease in developmental rate) because positively and negatively allometric coefficients are differently modified. The figured case is an example of neoteny; compared to the ancestor (circles), the descendant (squares) has allometric coefficients that are closer to isometry – the descendant's adult proportions more closely resemble those of the shared juvenile morphology. Positively allometric coefficients decrease (towards 1.0) whereas negatively allometric coefficients increase (towards 1.0).

developmental stage, i.e. $\{Y_1, Y_2, Y_3, \ldots Y_P\}$ at that stage. That comparison is made by predicting $\{Y_1, Y_2, Y_3, \ldots Y_P\}$ at the appropriate values of X for each species, using species-specific regression equations. The residuals from each regression are then added to the species-specific mean, and the two samples are compared. That is precisely what we did when size-standardizing geometric data in Chapter 10.

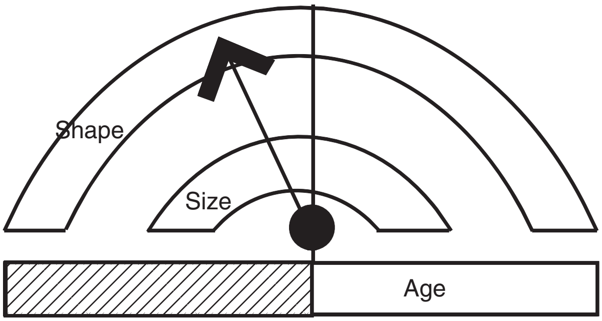


Figure 13.8 Gould's (1977) clock model. The descendant's size, shape and age at one developmental stage are compared to the ancestor's ontogeny of shape, size and age. The hands of the clock show the change from ancestral to descendant values, pointing from the descendant's age-specific shape and size to the corresponding ancestral values (no hand is shown if there is no change). The shape hand points to the left, so the descendant adult has the morphology of a younger stage in the ancestral ontogeny.

When channeling is found and age data are available (or collected), it is possible to go further and identify the changes in developmental rate and/or timing. Both Gould (1977) and Alberch et al. (1979) provide analytic formalisms for that purpose, and it is worth understanding them because they are so widely used in the literature.

Formalisms for heterochrony

There are two formalisms for the study of heterochrony: Gould's (1977) "clock-model" and Alberch et al's. (1979) scheme. Although the clock model is rarely used in the modern literature, understanding it is important because it supplied the context for Alberch et al.'s scheme. Alberch et al. retained the meaning of the concepts and terms defined by Gould (excepting those they explicitly redefined).

The face of the clock contains two arcs and one bar (Figure 13.8). One arc is a shape axis. The values of the ancestral shape are plotted along the arc, with the values for the youngest age on the left. The second arc is the size axis; values of the ancestral size are plotted on this axis so that the size and shape for each age match up. Age is represented by the bar at the bottom. Although the entire ontogeny of the ancestral shape is represented on the clock, the descendant is analyzed at a single (static) stage. Not surprisingly, the need to single out one stage for comparison prompted much discussion about the appropriate standard for comparison. That standard could be a chronological age, a developmental age, or even a size. Whatever standard is used, the objective is to find the matching ancestral size and shape at that point. When found, the hands of the clock are arranged to point to it; if the matching shape occurs at an earlier stage in the ancestor, the "shape hand" of the clock will point to the left. Similarly, if the matching size occurs at an earlier stage in the ancestor, the "size hand" also points to the left. Differences between ancestor and descendant in chronological age at the developmentally comparable stages are indicated by cross-hatching on the age bar.

The clock provides diagnostic tests for types of heterochrony, which Gould defines both verbally and in terms of the patterns revealed by the clock model. For example,

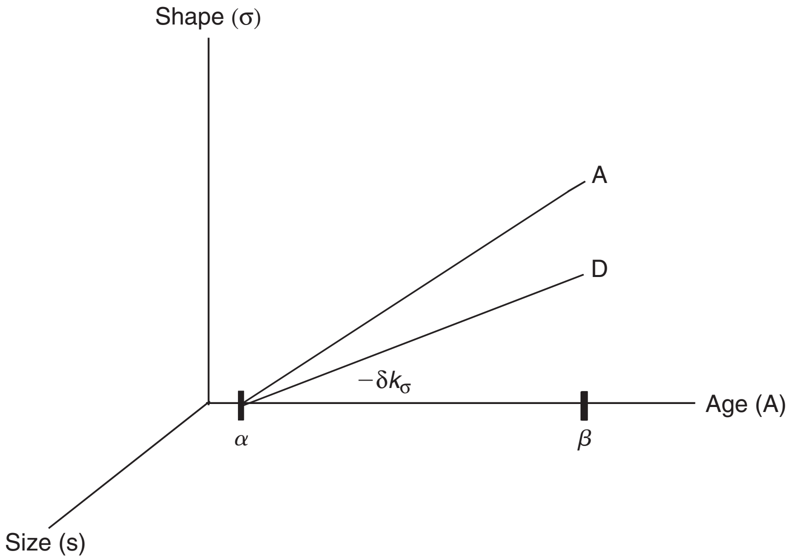


Figure 13.9 Alberch et al. formalism. The clock is redrawn by representing the ancestral shape, size and age as three mutually orthogonal axes. Species are compared with respect to the age at onset of development (α), rate of development (k_σ), rate of growth (k_s), and age at termination of development (β). See Table 13.4 for the names of the heterochronic perturbations defined by changes in these three parameters.

neoteny is retardation in the development of shape, and it is evident from the shape hand pointing to the left. However, Gould's classification was soon replaced by the one devised by Alberch et al., and the two schemes are not completely consistent. Thus, we will detail the types of heterochrony based on the Alberch et al. scheme.

Alberch et al. (1979) redesigned Gould's formalism, using a more conventional representation of a three-dimensional space: three mutually orthogonal axes (Figure 13.9). They also replaced Gould's static comparative framework by a dynamic one; the descendant ontogeny (not just one point along it) is analyzed in conjunction with the ancestral ontogeny. Each ontogeny is represented as a vector in the three-dimensional space defined by the ancestral values of size and shape. The comparisons are made with respect to four parameters: (1) α, the age at the onset of development; (2) β, the age at offset of development, (3) k_σ, the rate of development (i.e. the rate of change in shape); and (4) k_s the rate of growth (i.e. the rate of change in size). Each parameter can differ in two directions, yielding the eight pure heterochronic perturbations (Table 13.4). Two of them, proportional giantism and dwarfism, are not usually considered to be heterochronic perturbations because they do not yield either paedomorphic or peramorphic descendants, but they are usually included for the sake of completeness. Of course, combinations of these pure cases are also possible. However, if we found a combination of $+k_s$ and $-k_\sigma$, for example, we would not construct a compound name from the labels for each one – that would result in "proportional giantism plus neoteny" when, by definition, a proportional giant is a giant replica of the ancestral morphology, and neoteny necessarily signifies a difference in shape.

Table 13.4 Definitions of the eight pure heterochronic perturbations and their morphological expression, as defined by Alberch et al. (1979)

Control parameter	*Incremental change*	*Process*	*Morphological expression*
α	$-\delta\alpha$	Predisplacement	Peramorphosis
	$+\delta\alpha$	Postdisplacement	Paedomorphosis
β	$-\delta\beta$	Progenesis	Paedomorphosis
	$+\delta\beta$	Hypermorphosis	Peramorphosis
K_σ	$-\delta k_\sigma$	Neoteny	Paedomorphosis
	$+\delta k_\sigma$	Acceleration	Peramorphosis
k_s	$-\delta k_s$	Proportional giantism	
	$+\delta k_s$	Proportional dwarfism	

Changes in ontogenetic trajectories

Changes in ontogenetic trajectories produce novel morphologies, not just descendants that resemble ancestors at an older or younger developmental stage. Although all involve changes in the spatiotemporal organization of development, they produce different relationships between ontogeny and phylogeny as well as different ontogenetic patterns in disparity.

Changes confined to early morphogenesis

When species diverge very early in development they will differ in proportions at the outset of allometric growth, but they subsequently follow the same ontogeny. Consequently, the descendant allometric vector parallels the ancestral one but starts at a different shape (Figure 13.10A). This pattern is often termed "transpositional allometry" because, as is evident in Figure 13.10A, the descendant ontogeny is merely translated up or down the Y-axis (when reading the picture, it is important to recall that it is a log–log plot). The evolutionary direction of change is the same whether we look at juveniles or adults (Figure 13.10B). For example, if the descendant ends larval development with a head twice as long for its body as the ancestor's, it will be twice as long for its body (compared to an ancestor at the same size) throughout the whole of ontogeny. Neither the magnitude nor the structure of disparity changes over ontogeny.

Transpositional allometry means that early development is *more* labile than later. It also means that divergence occurs *very* early – the differences are evident at $X = 1$ (after which the species follow the same ontogeny).

Empirical criteria for documenting changes confined to early morphogenesis

Detecting this pattern is relatively straightforward: the expectation is that species do not differ in $\{k_1, k_2, k_3, \dots k_P\}$ but do in $\{b_1, b_2, b_3, \dots b_P\}$, which is easily tested by MANCOVA. Because the differences are in $\{b_1, b_2, b_3, \dots b_P\}$, which is the Y-intercept, the divergence in shape is manifest when $X = 1$, which is when $\log(X) = 0$. That all change must occur prior to that point is why transpositional allometry indicates a divergence in, and solely in, very early development.

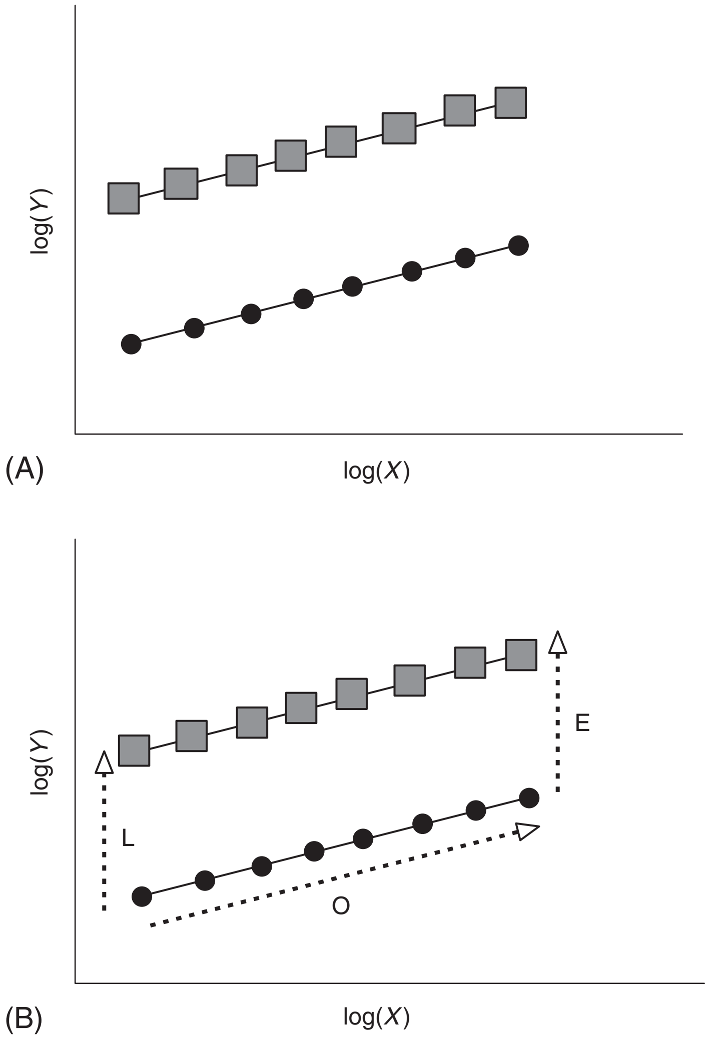


Figure 13.10 Changes confined to early morphogenesis, depicted by a bivariate plot of log(Y) on log(X). Species differ in proportions early in development (shown by differences in the Y-intercept) but subsequently follow the same allometric vector. A plot of a single measurement suffices to show the general case, even though some might not differ in the Y-intercept. No measurements differ in slope. (A) Ontogenetic allometries of ancestor (circles) and descendant (squares); (B) comparison of directions of ontogenetic change (O), and evolutionary change (E), to divergence of shape during larval (i.e. early) morphogenesis (L).

Changes in the spatiotemporal organization of development from the outset to the end of allometric growth

If species differ in their ontogenetic allometries from the outset of larval growth, they undergo different ontogenetic transformations in shape. Whether they diverge progressively as they develop, or instead converge, is not specified by the hypothesis; it simply claims that the difference results from a particular kind of change in development. To

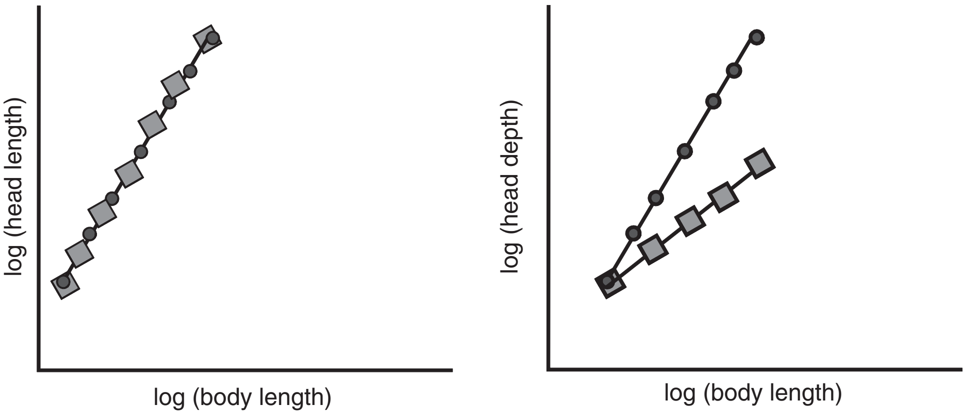


Figure 13.11 Changes in the spatiotemporal organization of development from the outset to end of allometric growth, depicted by two bivariate plots of log(Y) on log(X). Both the ancestor (circles) and the descendant (squares) begin development with the same shape, but undergo different changes in proportion. To show this pattern we need at least two plots because it involves a change in the ratios among allometric coefficients. Directions of ontogenetic change and evolutionary change cannot be compared in either two-dimensional space; such a comparison requires a multidimensional space.

consider a simple example, we can use head length : head depth proportions in two species (Figure 13. 11). The ancestral head is positively allometric in both length and depth, and the two allometric coefficients are nearly equal, so head shape is isometric (but the head enlarges relative to the body). The descendant head, like the ancestral head, is positively allometric in length, but depth is more nearly isometric. Therefore head depth is negatively allometric *relative to head length*, even if it is isometric relative to body length. Over ontogeny, head length increases relative to body length, and also relative to head depth.

This hypothesis cannot be depicted by a single bivariate relationship because it concerns a change in the ratios of two or more k's, so we need at least two bivariate plots to represent it graphically. Accordingly, we cannot draw the directions of ontogenetic and evolutionary change on these plots. There are two different directions of ontogenetic change, and the direction of evolutionary change will vary over ontogeny. Moreover, the hypothesis does not explicitly state whether species resemble each other at the outset of the measured phase, or at the outset of allometric growth (i.e. at log(X) = 0) or at a later developmental stage (meaning that the regression lines would intersect on the plot). To draw the plots, we have incorporated an assumption not required by the hypothesis – that species resemble each other when we first observe them.

Empirical criteria for documenting changes in spatiotemporal organization of development from the outset to the end of allometric growth

The hypothesis predicts that species will differ in $\{k_1, k_2, k_3, \ldots k_P\}$ and, like the hypothesis of transpositional allometry, this is easily tested by MANCOVA. However, channeling can also produce changes in k, so we need to document a change in the direction of $\{k_1, k_2, k_3, \ldots k_P\}$, not just a change in overall rate of development. To test that hypothesis, we can show that the interspecific angle between vectors of allometric coefficients is no larger than

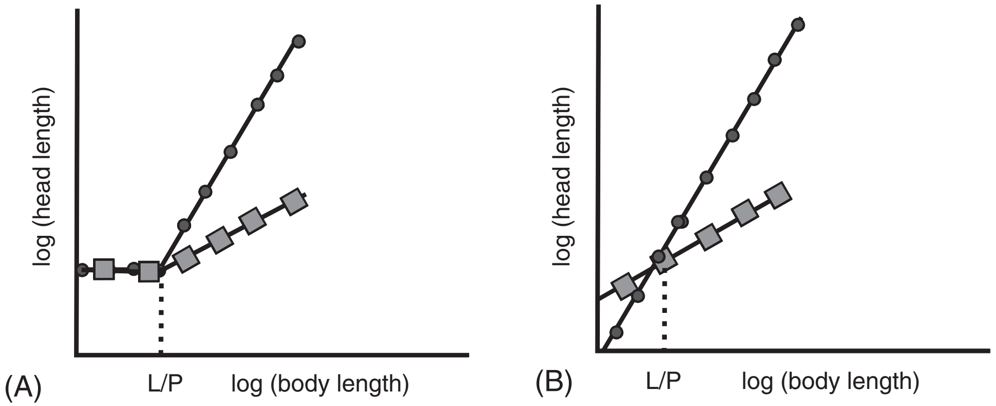


Figure 13.12 Changes in the spatiotemporal organization of development, confined to late stages of development. The hypothesis predicts (A): ancestor (indicated by circles) and descendant (indicated by squares) follow the same ontogeny up to the transition from one stage to the next (indicated as a transition from larval (L) to postlarval (P) phases), after which they diverge. The similarity in proportions at the transition point could, however, have a different explanation, shown by (B): Ancestor (indicated by circles) and descendant (indicated by squares) follow divergent ontogenies that intersect at the transitional point. The diagrams represent the two cases for a single measurement, but the same explanation would have to hold for all measurements.

anticipated by chance (it is important to test them and not just measure them, because very small angles are expected even when ontogenies differ considerably). We can also show that species do not differ in developmental rate/timing, which means documenting that neither extension nor truncation occurs. Both alter the length of the ontogenetic vector, so we can compare those lengths (which are estimated by the square root of the summed squared allometric coefficients).

Changes in spatiotemporal organization of development, confined to late development

This hypothesis differs from the one described above because the modifications are specific to late stages of ontogeny. Accordingly, species initially resemble each other and diverge progressively as they grow. Figure 13.12 depicts such a case. This hypothesis tacitly assumes that the allometric model of constant relative growth rates does not hold for the whole of ontogeny, because it claims that species share a common larval allometric pattern but diverge later. This hypothesis thus predicts that species coincide in their regression lines during early development and later diverge, when the regression line for at least one of the species changes its slope (Figure 13.12A). The alternative explanation, which is not consistent with the biological hypothesis, is that both regression lines are straight and extend all the way back to the *Y*-intercept, and happen to intersect at the youngest observed age (Figure 13.12B).

This hypothesis of conserved early development and divergent later development is consistent with von Baer's second law, so we might anticipate that the pattern is common. Although progressive divergence could occur by other kinds of modifications of ontogeny as well (see below), it appears that progressive divergence is not common – or at least it is rarely reported.

Empirical criteria for documenting changes in the spatiotemporal organization of development, confined to late development

The hypothesis predicts that species will differ in $\{k_1, k_2, k_3, \ldots k_P\}$ but will not differ in shape at the outset of the measured developmental phase. That is not equivalent to predicting that they do not differ in $\{b_1, b_2, b_3, \ldots b_P\}$, because b is the value for Y when $X = 1$, which is much smaller than a larval length. Instead, the hypothesis predicts that species do not differ in proportions at the earliest stage observed (or at the youngest stage relevant to the hypothesis); their divergence begins after that point. The first step in testing the hypothesis is the same as discussed above: documenting interspecific differences in $\{k_1, k_2, k_3, \ldots k_P\}$, which can be done by MANCOVA and by showing that the interspecific angles between the vectors are larger than anticipated by chance. The second step is to show that species are very similar in shape early in development (and diverge as they grow). This can be done by estimating the proportions expected for each species at that stage, which is done by predicting $\{Y_1, Y_2, Y_3, \ldots Y_P\}$ for each. Given that we find differences in $\{k_1, k_2, k_3, \ldots k_P\}$, we need to base these predictions on the regression models fitted separately to each species. We can then assess whether the expected values differ significantly, which is done by adding the residuals from the regression model to the expected value (as calculated for each species) and then comparing the expected shapes between species statistically. We can also compare the lengths of the ontogenetic vectors (as described above), which are predicted not to differ.

Complex changes in multiple parameters and stages

Having considered several simple cases, we can begin to explore the more interesting combinations of two or more modifications. We will consider four possibilities:

1. That both early and late morphogenesis are modified but developmental rate/timing is not (Figure 13.13)
2. That one phase is modified in both morphogenesis and developmental rate, whereas the other is not modified in either (Figure 13.14)
3. That early morphogenesis is modified but later development is modified solely in developmental rate/timing (Figure 13.15)
4. That both stages are modified in morphogenesis and developmental rate/timing is also altered (Figure 13.16).

Although all these cases are similar in that multiple developmental parameters differ, and also all predict a complex relationship between ontogeny and phylogeny; they differ considerably in their biological implications. The first implies that morphogenesis is more labile than developmental rate/timing – rate/timing is conserved although morphogenesis evolves. The next two imply that one developmental stage is more labile than the other and that it is labile in both morphogenesis and developmental rate/timing. The fourth implies that development is highly labile in general – everything that *can* change *does*. The most interesting consequences of these complex modifications, aside from what they tell us about the lability of development, are their potential impacts on disparity. The interactions among the multiple novelties may result in greater disparity than expected from the impact of each one, taken separately, or in less disparity than expected from a single modification. Interactions among multiple novelties might either amplify or counterbalance each other.

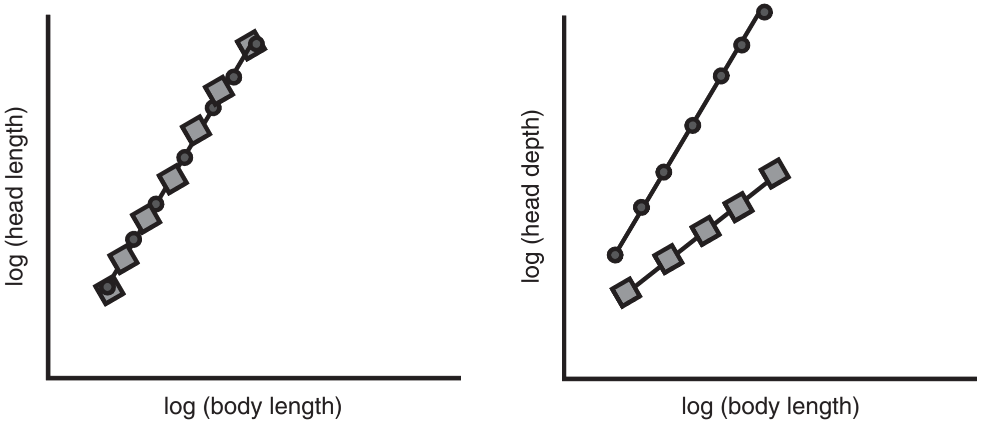


Figure 13.13 A complex change involving a modification of early morphogenesis combined with a change in the spatiotemporal organization of later development. The ancestral ontogeny is indicated by circles, descendent ontogeny by squares. Like the case shown in Figure 13.11, this one cannot be depicted by a single bivariate plot because it involves changes in the ratios of allometric coefficients.

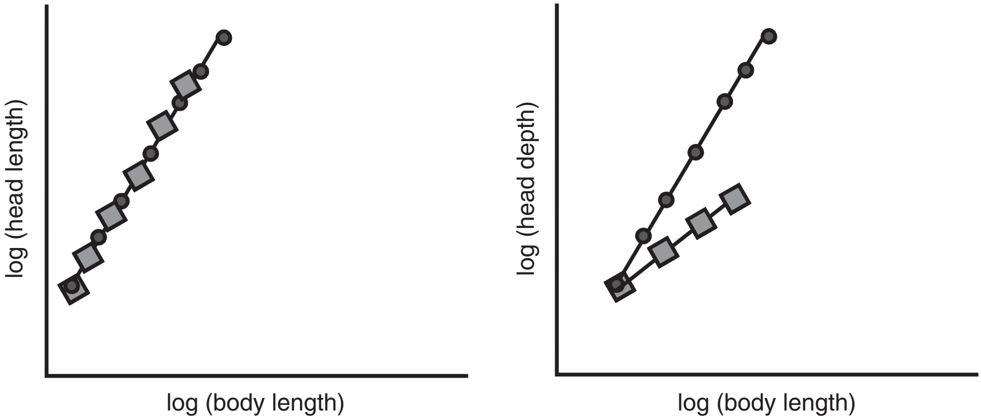


Figure 13.14 A complex change involving both a modification in the spatiotemporal organization of development late in ontogeny and a modification of developmental rate/timing. The ancestral ontogeny is indicated by circles, the descendent ontogeny by squares. The change in late development is evident in the translation of the descendant's ontogeny along the *Y*-axis; the change in developmental rate/timing is due to an increase in adult body size. To show a change in developmental rate without concomitant change in body size, two bivariate plots would be needed to depict the contrasting changes in positively and negatively allometric coefficients. Like all cases involving a modification of late morphogenesis, and hence a change in direction, at least two bivariate plots are needed to depict the pattern.

Empirical criteria for documenting changes in multiple parameters and stages

To identify the parameters that differ, we need to compare: (1) shapes at the youngest comparable stage; (2) vectors of allometric coefficients; and (3) lengths of the ontogenetic vectors. To determine the impact of these changes on disparity we need to measure disparity at two or more stages, and also, ideally, to measure the impact of each modification

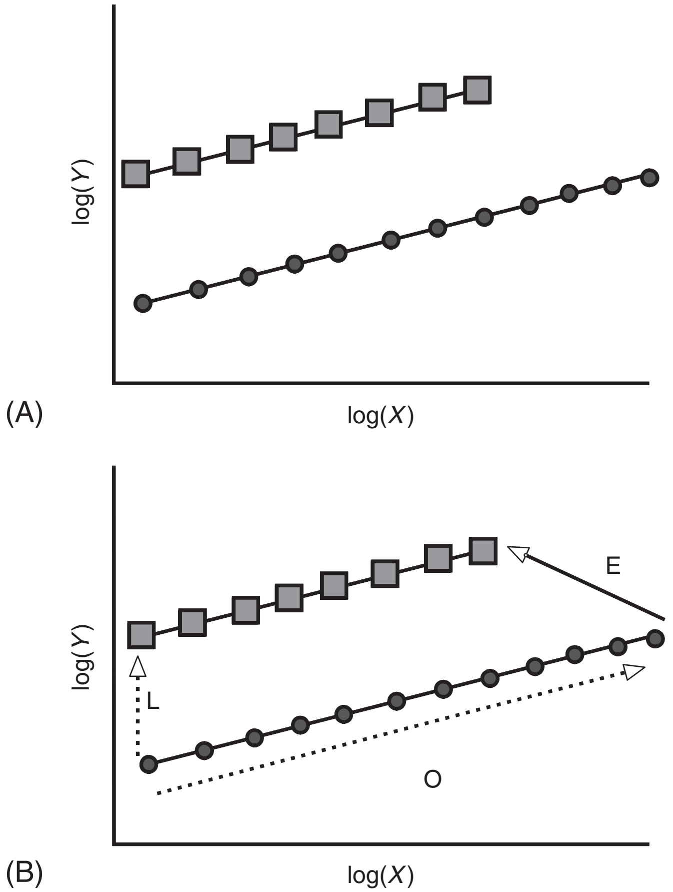


Figure 13.15 A complex change involving a modification of early morphogenesis and a change in developmental rate/timing. The ancestral ontogeny is indicated by circles, the descendent ontogeny by squares. As in Figure 13.13, the change in developmental rate/timing is due to an increase in adult body size. To show a change in developmental rate without concomitant change in body size, additional bivariate plots would be needed to depict the contrasting changes in positively and negatively allometric coefficients. Like all cases involving a modification of spatiotemporal patterns of development, at least two bivariate plots are needed to depict the pattern.

separately. In addition we can determine if ontogenies diverge over time, remain at a constant distance apart, or converge towards a similar endpoint by combining the ontogenetic series of multiple species and analyzing them by PCA.

Applying these criteria to an empirical case: comparing ontogenies of *S. gouldingi* and *S. manueli*

To conclude our discussion of comparing ontogenetic allometries based on traditional morphometric data, we will compare the ontogenetic allometries of two sister species – *S. gouldingi* and *S. manueli*. The allometric coefficients *b* and *k* are given in Table 13.5, and the *k*s are plotted on the measurements to give a better appreciation of where the species

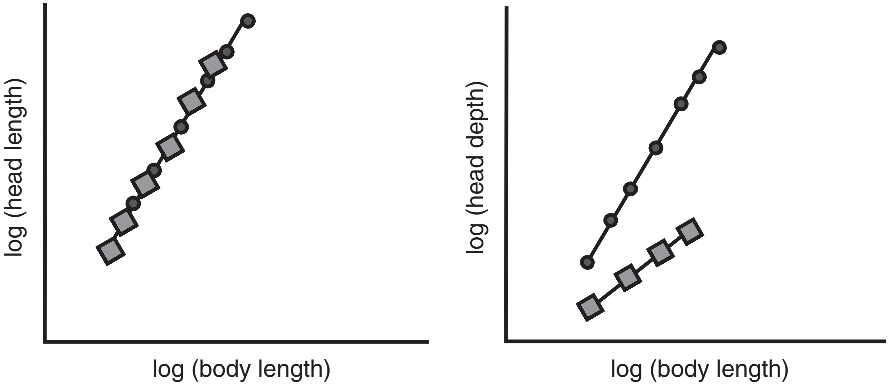


Figure 13.16 A complex change in three parameters: early morphogenesis; modification of later morphogenesis; and change in developmental rate and timing (due to increased body size). The ancestor ontogeny is indicated by circles, the descendent ontogeny by squares.

Table 13.5 Allometric coefficients *b* and *k* and their standard errors for *S. gouldingi* and *S. manueli* (see Figure 13.1 for the definition of the variables)

Variable	S. gouldingi		S. manueli	
	b (std error)	*k* (std error)	*b* (std error)	*k* (std error)
v2	−0.94 (0.13)	0.81 (0.03)	−1.81 (0.09)	1.03 (0.02)
v3	−0.61 (0.05)	0.89 (0.01)	−0.47 (0.06)	0.87 (0.01)
v4	−1.16 (0.07)	0.85 (0.01)	−1.27 (0.07)	0.90 (0.02)
v5	−2.64 (0.13)	1.23 (0.03)	−1.45 (0.11)	0.96 (0.03)
v6	−1.40 (0.04)	1.04 (0.01)	−1.14 (0.05)	1.00 (0.01)
v7	−1.51 (0.06)	1.00 (0.01)	−1.80 (0.05)	1.09 (0.01)
v8	−1.38 (0.09)	0.93 (0.02)	−1.00 (0.11)	0.85 (0.03)
v9	−1.97 (0.08)	1.10 (0.02)	−1.81 (0.08)	1.06 (0.02)
v10	−1.76 (0.04)	1.22 (0.01)	−1.56 (0.04)	1.17 (0.01)
v11	−1.93 (0.04)	1.20 (0.01)	−1.59 (0.04)	1.13 (0.01)
v12	−1.60 (0.04)	1.14 (0.01)	−1.32 (0.04)	1.08 (0.01)
v13	−2.38 (0.09)	1.19 (0.02)	−2.27 (0.07)	1.15 (0.02)
v14	−2.23 (0.07)	1.12 (0.02)	−2.10 (0.09)	1.08 (0.02)
v15	−1.68 (0.04)	1.21 (0.01)	−1.41 (0.05)	1.15 (0.01)
v16	−1.78 (0.04)	1.23 (0.01)	−1.52 (0.05)	1.17 (0.01)
v17	−1.55 (0.03)	1.17 (0.01)	−1.24 (0.05)	1.10 (0.01)
v18	−2.57 (0.08)	1.17 (0.02)	−1.98 (0.10)	1.05 (0.02)
v19	−1.94 (0.05)	1.09 (0.01)	−1.62 (0.06)	1.02 (0.02)
v20	−1.69 (0.05)	1.10 (0.01)	−1.40 (0.04)	1.04 (0.01)
v21	−1.99 (0.04)	1.23 (0.01)	−1.74 (0.05)	1.17 (0.01)
v22	−1.83 (0.04)	1.22 (0.01)	−1.56 (0.05)	1.16 (0.01)
v23	−1.47 (0.04)	1.07 (0.01)	−1.30 (0.05)	1.03 (0.01)
v24	−1.67 (0.08)	0.94 (0.02)	−2.70 (0.13)	1.13 (0.03)
v25	−2.69 (0.07)	1.12 (0.01)	−2.57 (0.06)	1.09 (0.01)
v26	−1.56 (0.03)	1.13 (0.01)	−1.37 (0.04)	1.09 (0.01)
v27	−1.82 (0.11)	0.90 (0.02)	−3.41 (0.16)	1.23 (0.04)
v28	−1.23 (0.09)	0.81 (0.02)	−1.89 (0.10)	0.97 (0.02)
v29	−1.16 (0.08)	0.75 (0.02)	−1.50 (0.08)	0.82 (0.02)
v30	−0.64 (0.04)	0.90 (0.01)	−0.51 (0.04)	0.89 (0.01)

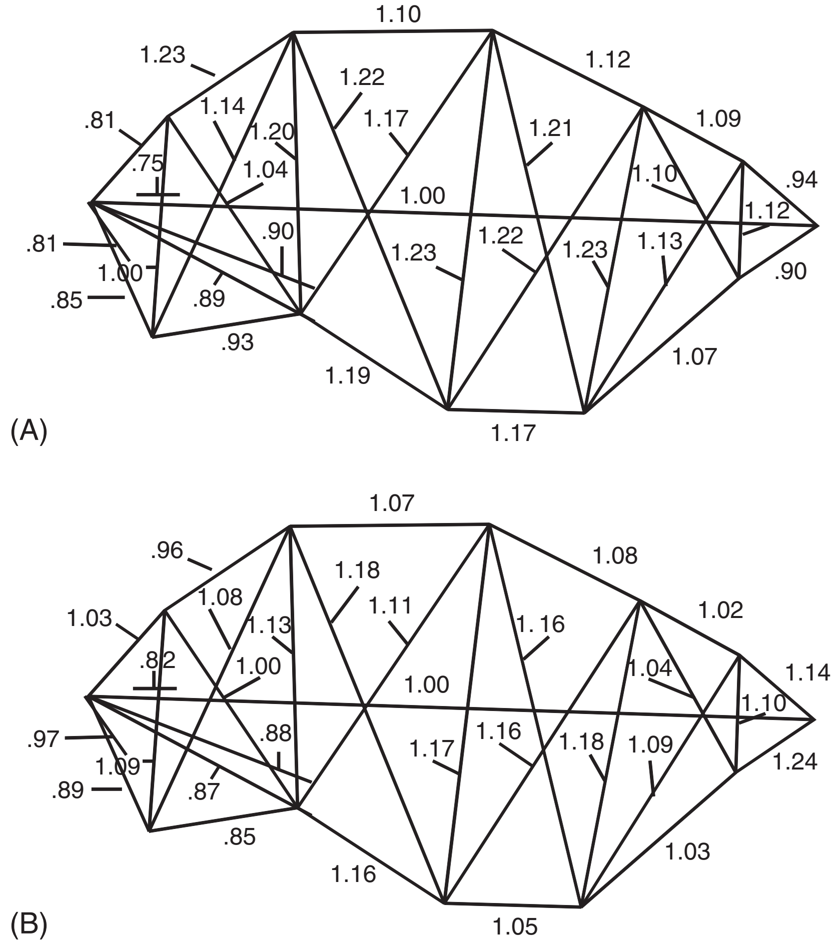


Figure 13.17 Allometric coefficients of *S. gouldingi* and *S. manueli.*

differ (Figure 13.17). Statistically, they differ in *k* for nearly all variables; the exceptions are two in the head (v3 and v30), two anterior postcranial measurement (v9 and v13) and the depth measurement of the caudal peduncle (v25). The angle between the vectors is small (6.27°), but is nonetheless significantly greater than 0.0° ($p < 0.05$). Considering that an angle of only just over 12.74° indicates greater *dissimilarity* than expected by chance, 6.27° does not seem so modest. There is some evidence, albeit slight, for an interaction between modifications of early and late ontogeny that reduce disparity over ontogeny; this pattern of counterbalancing modifications is suggested by the plot of PC2 on PC1 (Figure 13.18). Scores on PC1 increase with age, it also looks as though they are more differentiated on PC2 earlier rather than later, and that they gradually reach a more similar form. Yet the scale of PC2 is tremendously exaggerated in the plot; that axis accounts for only 0.4% of the variance (PC1 accounts for 98.9%). So it is not clear whether this really is a case of counterbalancing, an issue we will return to in the analysis of these species using a geometric approach.

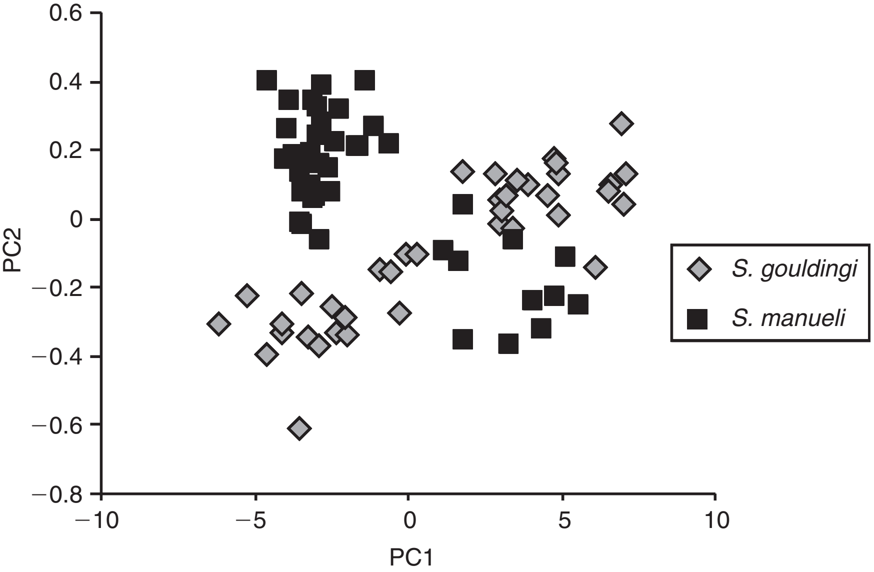


Figure 13.18 PCA of the pooled ontogenetic data of *S. gouldingi* and *S. manueli*. The scale of the *Y*-axis (PC2) is greatly exaggerated to make the separation between species more visible; older specimens are to the right of the plot. The two species appear to differ, albeit subtly, in shape at the outset of the measured stage and in their ontogenies. The modification of juvenile growth appears to counterbalance the change in larval morphogenesis such that adults are more similar than juveniles.

Exploring evolutionary patterns of evolving ontogenies: geometric morphometric analyses

Geometric studies of ontogenetic allometry are formally similar to traditional morphometric ones, but there are important differences in the meanings of the parameters. Revisiting the regression of geometric shape on size (covered in Chapter 10), the general bivariate linear regression model is:

$$Y_i = mX_i + b_i + \varepsilon_i \tag{13.4}$$

where Y is a shape variable, X is body size (measured by centroid size), b is the Y-intercept and ε is the error term. Because we invariably analyze shape data multivariately, the model we actually use is:

$$\{Y_1,\ Y_2,\ Y_3,\ldots\ Y_P\} = \{m_1, m_2,\ m_3,\ \ldots m_P\}X + \{b_1,\ b_2,\ b_3,\ \ldots b_P\} + \{\varepsilon_1,\ \varepsilon_2,\ \varepsilon_3,\ \ldots \varepsilon_P\} \tag{13.5}$$

where $\{Y_1, Y_2, Y_3 \ldots Y_P\}$ is the vector of shape variables (i.e. landmark coordinates obtained by GLS, or partial warp scores including the uniform component), X is centroid size, and $\{m_1,\ m_2,\ m_3,\ \ldots m_P\}$, $\{b_1,\ b_2,\ b_3,\ \ldots b_P\}$ and $\{\varepsilon_1,\ \varepsilon_2,\ \varepsilon_3,\ \ldots \varepsilon_P\}$ are vectors of slope coefficients, intercepts and residuals, respectively.

One consequence of analyzing allometry using geometric data instead of traditional data is that the allometric coefficients are no longer meaningful in terms of a specific growth model – either a power law (Equation 13.1) or a sigmoidal curve of size or shape relative to time. Instead, they indicate that shape changes with size. The underlying causes of those changes are the same ones that account for allometric coefficients of traditional measurements, but we cannot treat the coefficients as if they were rates of growth of individually meaningful variables, nor can we use them to estimate time-lags between growth curves of organs. Unlike the allometric coefficients of traditional measurements, those of geometric shape variables have no individual biological meaning. For that reason, comparisons of ontogenetic allometries using geometric variables are rarely (if ever) done bivariately. We would not plot one shape variable at a time on size, except to check for linearity. In a geometric analysis we are not comparing coefficients measurement-by-measurement; rather, we are comparing whole sets of coefficients describing the ontogeny of an entire landmark configuration. However, the difference in the meanings of the coefficients does not impede our ability to recognize the patterns discussed above in the context of traditional measurements. None of these patterns were defined in terms of particular coefficients; hence they are not functions of a particular measurement scheme. We can thus examine the evidence for them in geometric as well as traditional data.

Channeling

To depict the pattern of channeling, we can consider a hypothetical case in which species have the same shape at the outset of development and follow the same ontogeny of shape, but differ in the overall rate or timing of development. Graphical evidence of channeling is shown in Figure 13.19, where we see that the coordinates of the juveniles of the two species are the same (Figure 13.19A), as are the two ontogenies of shape (Figure 13.19B), but the trajectories differ in length (Figure 13.19C). Perhaps the most compelling visual evidence is shown in Figure 13.20 – the descendant adult morphology lies at a subadult position on the ancestral ontogeny. We can see the coordinates for the descendant's landmarks in an intermediate position along the ancestral ontogeny.

The graphical evidence is corroborated by statistical analysis. In a statistical test of channeling, we would not expect to find a significant difference between the two ontogenies of shape or in the shape at the youngest comparable phase but we would anticipate a difference in the length of the ontogenetic vector (the parameter that measures the total amount of change undergone in each ontogeny over the observed phase). Carrying out these tests for the hypothetical species depicted in Figure 13.18, we would measure the similarity between ontogenies of shape by the angle between the (normalized) vector of allometric coefficients $\{m_1, m_2, m_3, \ldots m_P\}$. We find no significant difference between them; the tiny angle of only 1.8° is not significant compared to the ranges that can be obtained by resampling within them (4.2°, 7.5°). We also find no significant interspecific difference between shapes at the outset of the measured phase of development using Goodall's *F*-test ($p > 0.999$); and the magnitude of the difference between them is a Procrustes distance of 0.0. We do, however, find a significant difference in the length of their ontogenetic vectors as estimated by the Procrustes distance between youngest and oldest comparable stages; for the ancestral species that distance is 0.1999 (0.1961–0.2030), whereas for the descendant it is only 0.109 (0.1055–0.1129). Thus, the descendant's ontogeny is a truncated version of the ancestor's.

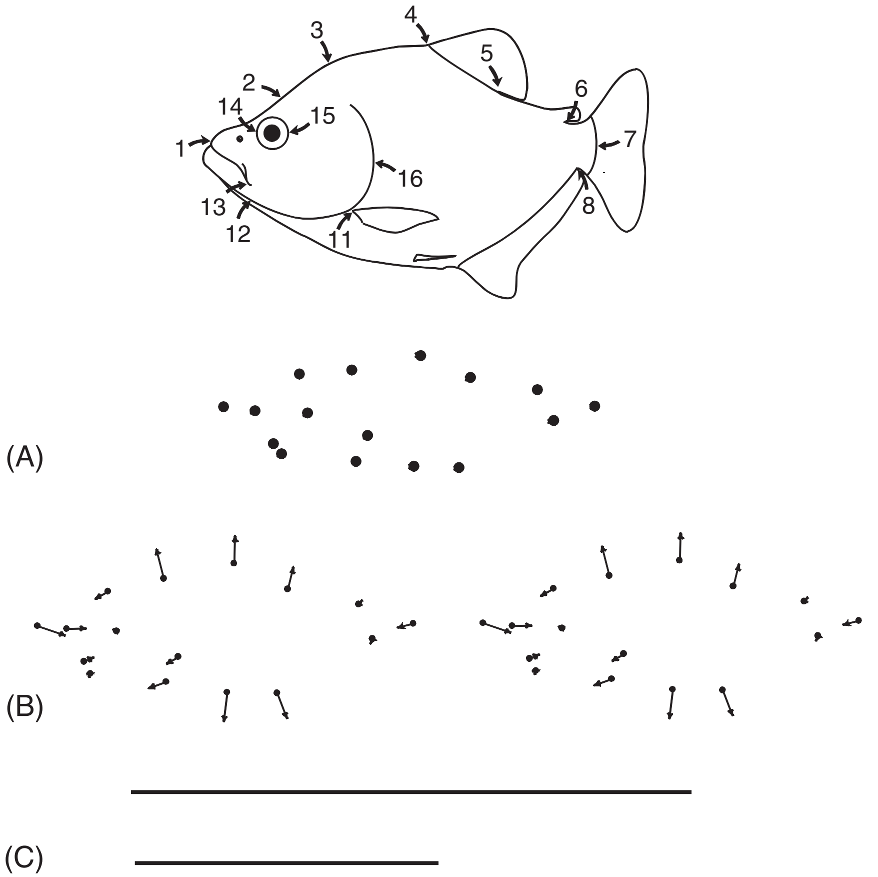


Figure 13.19 Channeling, depicted geometrically (see Figure 13.1 for the same pattern depicted for traditional measurement data). (A) Superimposed coordinates of juvenile shapes; (B) ontogenies of shape; (C) lengths of ontogenetic vectors of shape. The two species have the same shape at the outset of the measured phase, follow the same ontogeny of shape, but differ in the length of their ontogenetic vectors; the descendant has a truncated version of the ancestral ontogeny.

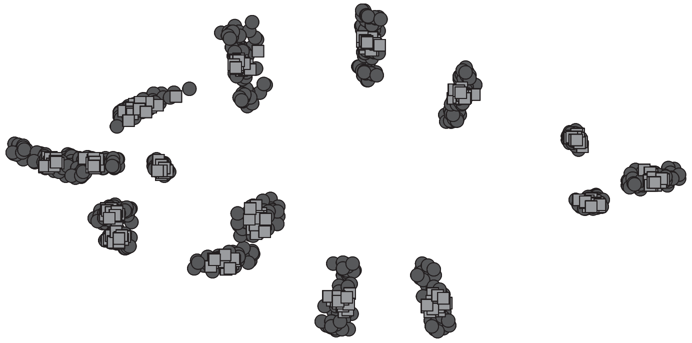


Figure 13.20 Superimposed coordinates for showing the ontogenetic transformation of ancestral shape (black circles) and the descendant adult shape (gray squares). The descendant adult shape is at an intermediate position along the ancestral ontogeny.

Changes in the ontogenetic trajectory

Changes confined to early morphogenesis

If changes occur solely in early morphogenesis, we would expect that species would be shaped differently at the youngest comparable stage (Figure 13.21A) but would subsequently follow the same ontogeny of form (Figure 13.21B), and to the same extent (Figure 13.21C). To test the hypothesis that only early development is labile, we can show that there is a significant difference in shape at the outset of the measured phase, but that any differences in later development are neither significant nor large. For the hypothetical species shown in Figure 13.21, the difference between their shapes at the transition

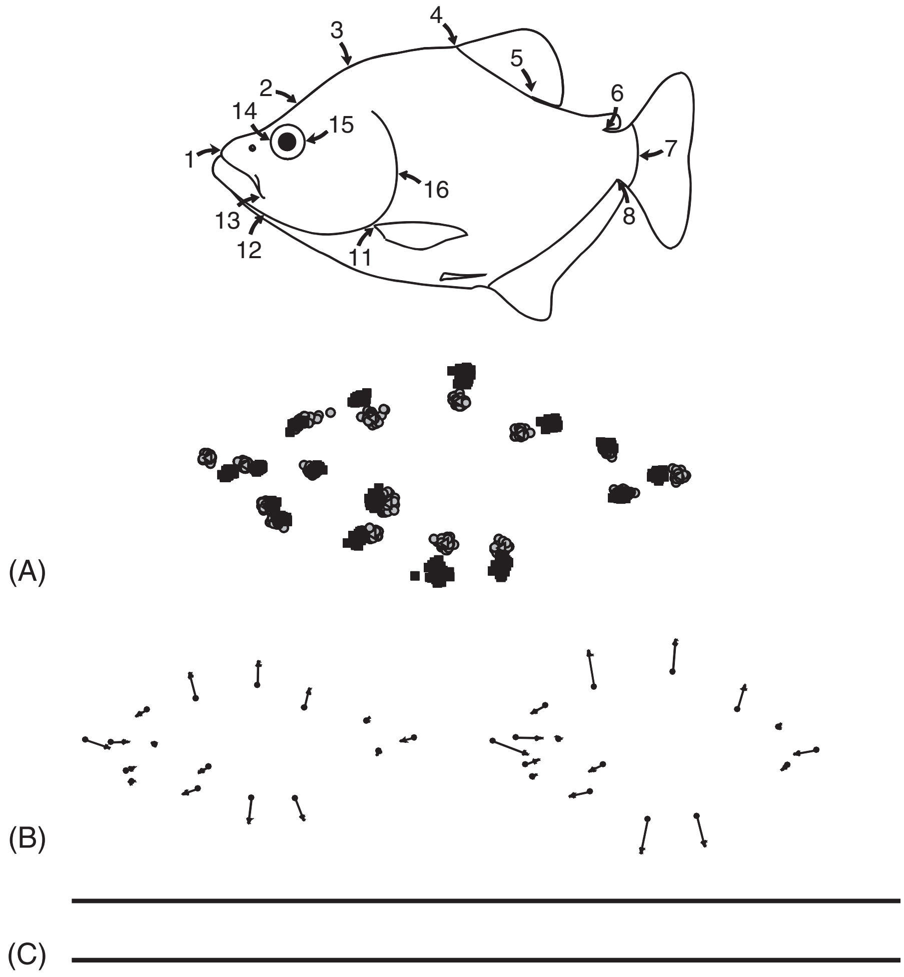


Figure 13.21 Change confined to early morphogenesis, depicted geometrically (see Figure 13.2 for the same pattern depicted for traditional measurement data). (A) Superimposed coordinates of juvenile shapes; (B) ontogenies of shape; (C) lengths of ontogenetic vectors of shape. The two species differ in shape at the outset of the measured phase, but subsequently follow the same ontogeny of shape and do not differ in the length of their ontogenetic vectors.

from larval to juvenile phases is highly significant ($p < 0.0001$) and the Procrustes distance between their means is large: 0.1247 (0.1196–0.1317). The contrast may seem particularly striking when we look at the superimposed coordinates and find so little overlap between species in several of them (Figure 13.21A). As anticipated, there is no significant difference in their ontogenies of form; the angle between the two vectors is a tiny 1.9° (compared to the within-species angles of 4.0° and 3.7°). Also as anticipated, the lengths of the ontogenetic vectors are the same for both species: 0.1999 (0.1961–0.2030) for the ancestor, and 0.2040 (0.1978–0.2095) for the descendant.

Changes in the spatiotemporal organization of development from the outset to the end of allometric growth

Should species differ in the spatiotemporal organization of morphogenesis, their ontogenetic vectors of shape will differ. It says nothing about shape at the youngest comparable stage, so the only expectation is that the ontogenies of shape will differ. Rather than discussing this simple case any further, we will consider the more complex hypotheses that include it.

Changes in the spatiotemporal organization of development, confined to late development

Should all change be confined to late morphogenesis, we would expect to see no difference between the species at the youngest comparable developmental stage (Figure 13.22A), and a visible difference in their ontogenies of shape (Figure 13.22B). This hypothesis says nothing about the length of the trajectory of late development; however, changes in length are due to differences in rate/timing and not to differences in morphogenesis, so we have drawn trajectories of the same length (Figure 13.22C). For this hypothetical case, juvenile shapes do not differ ($p > 0.999$, Procrustes distance = 0.0) and neither do the lengths of the ontogenetic vector: that of the ancestor is 0.1999 (0.1961–0.2030) and that of the descendant is 0.2011 (0.1977–0.2046). However, as anticipated, the ontogenies of shape differ significantly; the angle between the two vectors is 32° compared to the range of within-species angles that can be obtained by resampling (3.7° for both).

Complex changes in multiple parameters and stages

It is difficult to construct hypothetical ontogenies to demonstrate interactions among multiple parameters, because that requires modeling perturbations of the ontogeny of entire landmark configurations. Also, the most interesting distinctions among the possibilities lies in their conflicting predictions about the ontogenetic dynamics of disparity. In particular, we would wish to distinguish between amplification (which predicts increasing disparity over ontogenetic time) and counterbalancing (which predicts decreasing disparity over ontogenetic time). The contrasting patterns are not self-evident in the regression parameters; distinguishing them requires measuring disparity. Rather than considering another hypothetical case we will return to the analysis of an empirical case, *S. gouldingi* and *S. manueli*, because our earlier analysis suggested that these species might exemplify counterbalancing.

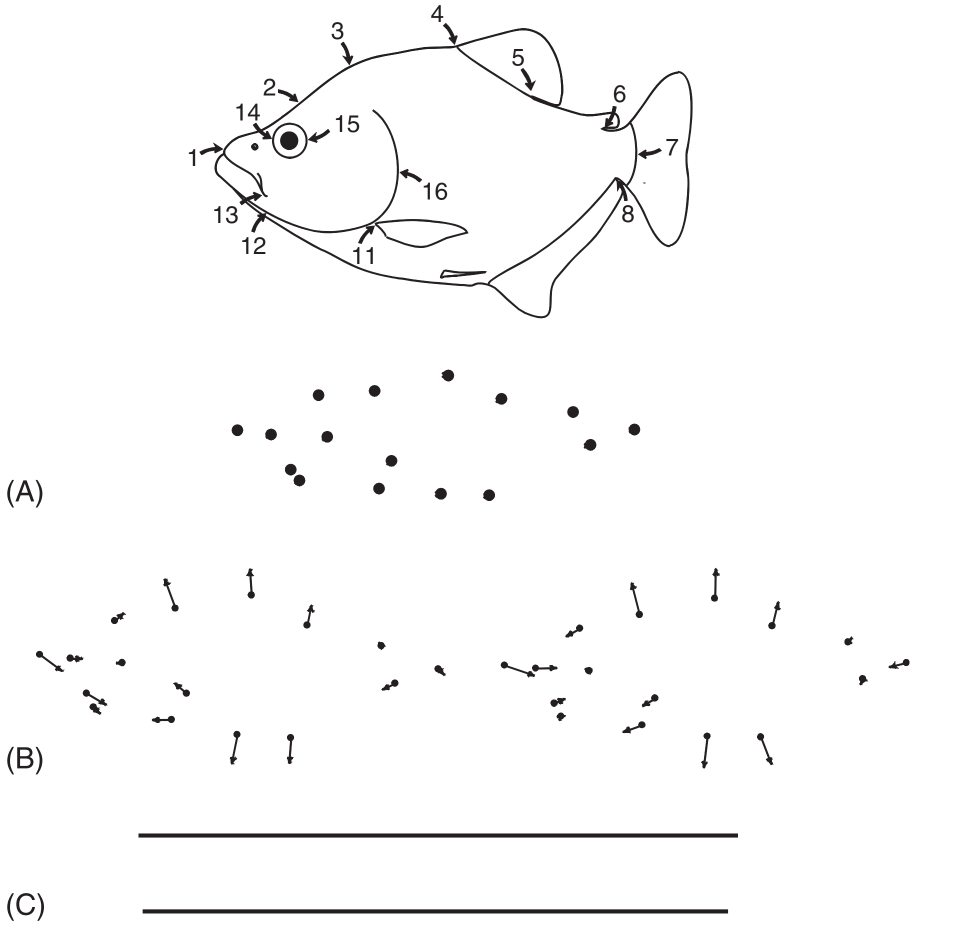


Figure 13.22 Change in the spatiotemporal pattern of development confined to late development, depicted geometrically. (A) Superimposed coordinates of juvenile shapes; (B) ontogenies of shape; (C) lengths of ontogenetic vectors of shape. The two species have the same shape at the outset of the measured phase, but subsequently follow different ontogenies of shape; they do not differ in the length of their ontogenetic vectors.

An empirical case: comparing ontogenies of *S. gouldingi* and *S. manueli*

Comparing these two species using geometric methods provides compelling graphical evidence that the species differ in morphology at the transition from larval to juvenile development (Figure 13.23A), in juvenile morphogenesis (Figure 13.23B), and in length of their ontogenetic trajectories (Figure 13.23C). The difference in shape at the transition from larval to juvenile development is statistically significant ($p < 0.0001$). Moreover, this difference is not just significant, it is large; the Procrustes distance between mean shapes at that stage is 0.080 (0.077–0.084). Ontogenies of shape are visibly different and the difference is statistically significant; the angle between their ontogenetic vectors is 34.9° (compared to the within-species ranges of 11.0° and 16.6° for *S. gouldingi* and *S. manueli*, respectively). The ontogenetic trajectories also differ significantly in length; that of *S. gouldingi*

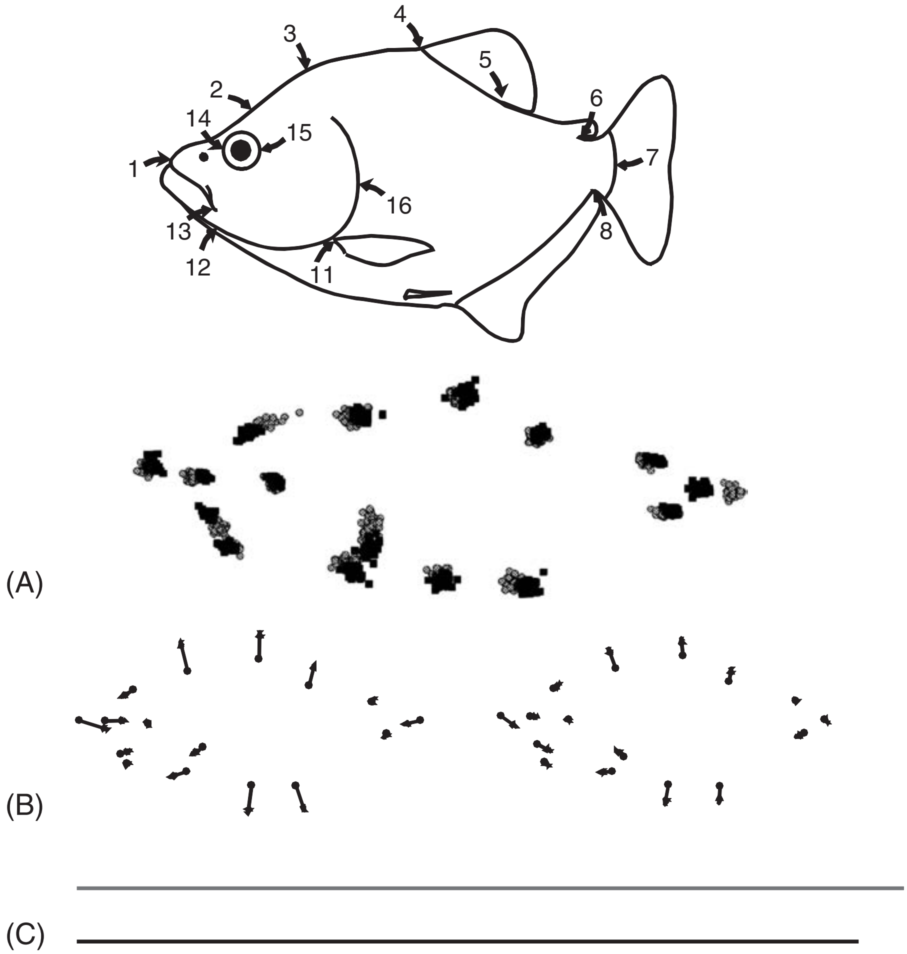


Figure 13.23 Ontogenies of *S. gouldingi* and *S. manueli*, depicted geometrically. (A) Superimposed coordinates of juvenile shapes; (B) ontogenies of shape; (C) lengths of ontogenetic vectors of shape. The two species differ in shape at the outset of the measured phase (the transition from larval to juvenile phases), subsequently follow different ontogenies of shape and differ in the length of their ontogenetic vectors.

is 0.2095 (0.2065–0.2128) and that of *S. manueli* is 0.1864 (0.01823–0.1906). Therefore, these two species differ in all three parameters. However, the adults are far less different than the young juveniles. The Procrustes distance between mean adult shapes is 0.051 (0.047–0.054) – a substantial drop from 0.080. That decrease can be seen in the analysis of the pooled ontogenies of shape by PCA (Figure 13.24); with increasing age, scores on PC1 increase, and the two species more closely resemble each other. We can now clearly see the pattern hinted at in the analysis of traditional morphometric data. The evidence based on geometric data is stronger, because now there are two distinct eigenvalues, and PC2 explains 9.5% rather than 0.4% of the variance. Taken together, these results all demonstrate that the modifications of juvenile development counterbalance those of larval development, thereby stabilizing adult morphology.

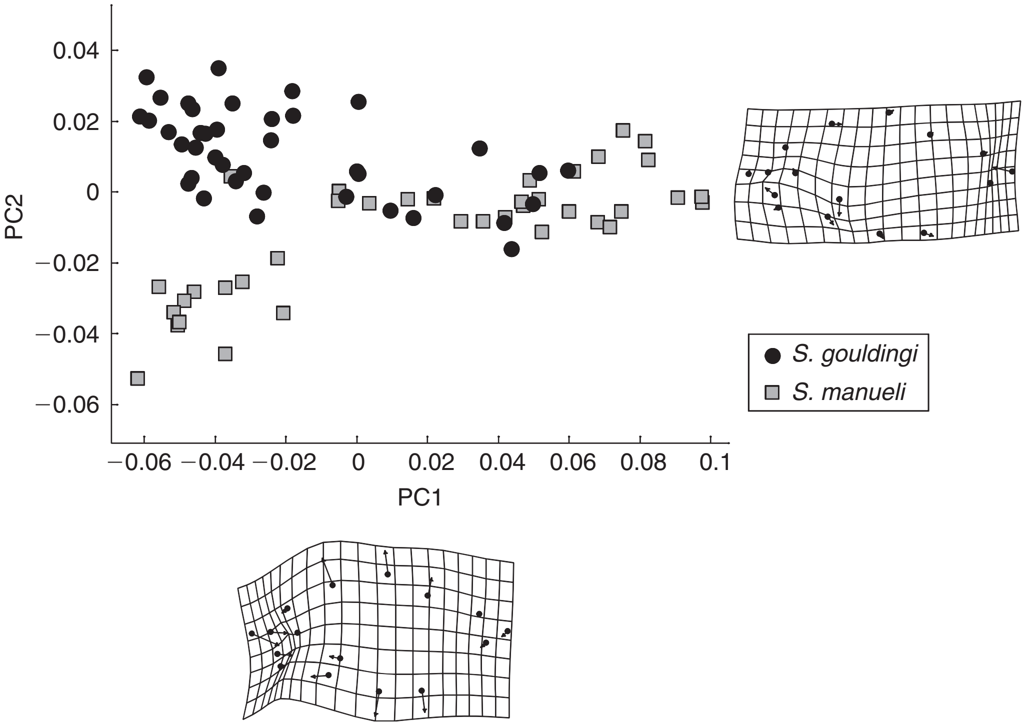


Figure 13.24 PCA of the pooled ontogenetic data of *S. gouldingi* and *S. manueli*. The two species differ in shape at the outset of the measured stage (the transition from larval to juvenile phases) and also in their ontogenies; older specimens are to the right of the plot. The modifications of juvenile growth appear to counterbalance the change in larval morphogenesis such that the adults are more similar than the juveniles.

Open questions

Having characterized a variety of evolutionary transformations of ontogeny, obvious questions are:

1. Which is (are) most likely?
2. Which occur(s) most often?

To answer the first question, we can derive expectations from developmental theory or functional morphology, or we can even turn to the empirical literature for insights into what expectations are biologically reasonable. Developmental biology offers few general theories, but there is a general set of principles that can be useful for framing hypotheses. One such general principle is that development is likely to be conservative when modifications disrupt the sequence of epigenetic interactions on which later development depends. Hall (1992) discusses the importance of epigenetic cascades, and these may resist modification either because modifications are lethal or because the cascades are internally stabilized. Stabilizing selection may play an important role in both cases, not only by eliminating deviants but also by building internal mechanisms that resist such disruptions (internal stabilizing selection, see Cheverud, 1984). A related general principle is that development is likely to be conservative when multiple (contemporaneous) processes are interdependent. An argument for the relative stability of embryogenesis around the time of neurulation is that this phase is the most highly integrated (Raff, 1992; Galis and Metz, 2001; Galis et al., 2001). A similar argument has been used to predict which types of transformations in ontogenetic allometries are most likely, based on the number of dissociations each requires (e.g. Shea, 2002).

Applying such general theories can be difficult in light of how little we know about development and its integration. Therefore, we might prefer to rely on the empirical literature, basing our expectations on the patterns most often reported. Klingenberg (1998) reviews the literatures on allometry and heterochrony, concluding that studies having enough statistical power to compare ontogenetic allometries statistically find significant differences in all allometric parameters, but changes in directions of allometry are subtle. The most common patterns detected in studies of vertebrates are (1) channeling (e.g. Gould, 1984; Strauss, 1984; Wayne, 1986; Shea, 1992); and (2) changes confined to early morphogenesis (e.g. Falsetti and Cole, 1992; Voss and Marcus, 1992; Klingenberg and Ekau, 1996). Apparently, early development is more labile than later (and changes in it are frequently responsible for adult disparity); and to the extent that later development evolves, it does so by channeling.

One notable exception to that pattern has been found in a study of (distantly related) sculpins: species diverge during larval development, and some continue to diverge further during postlarval development whereas others converge towards a similar adult shape (Strauss and Fuiman, 1985). Another exception is the group of piranhas that have served as the running example throughout this book. It is not known whether piranhas are at all exceptional. Their ontogenies do not seem to be exceptionally diverse; the angles between allometric vectors based on traditional measurement data do not indicate an exceptional degree of diversification. However, they may be unusual in that the modifications of postlarval development reduce the disparity generated during larval development. In this group, it appears that adult morphology is actively stabilized by modifications of postlarval ontogeny that compensate for modifications of larval ontogeny (Zelditch et al., 2003). Whether counterbalancing is a common phenomenon is another open question.

Some subjects in evolutionary developmental biology are not open to morphometric investigations, but many are. In addition to the relationship between ontogeny and phylogeny, many other subjects would benefit from morphometric studies. Among these are the causes of developmental integration, how integration evolves along with changes in ontogeny, the dynamics and causes of developmental regulation and its relationship to developmental buffering against random departures from bilateral symmetry, and the impact of mutants (or experimental treatments) on developmental pathways. All these are potentially rich subjects for morphometric studies.

Software

All the software required to implement the geometric analyses discussed in this chapter has been introduced in earlier chapters (see Chapter 10 for the software used to estimate regression coefficients, standardize shapes and compare ontogenies of shape, and Chapter 12 for software to analyze disparity). Thus, in this chapter we discuss only the software required by analyses of traditional morphometric data. One program in the IMP series, **TradMorphGen**, calculates traditional morphometric variables from a file of shape coordinates (in either X1,Y1, ... CS or TPS format). Another program, **VecCompare**, measures the angles between vectors of traditional (as well as geometric) measurements. For all the other analyses using traditional morphometric data, analyses can be done in commercially available software. Instructions for running **VecCompare** were given in Chapter 10; the only difference between using this program for the analysis of vectors of traditional measurements is that the input data file comprises (log-transformed) traditional measurements. Thus, to prepare a file of traditional measurements for analysis by **VecCompare**, use **TradMorphGen** and save the file of log-transformed traditional measurements.

Running TradMorphGen

To run **TradMorphGen**, you need a data file of landmark coordinates and a measurement protocol. You can load a file of shape coordinates (in any superimposition) produced by **CoordGen**, or you can load a file in TPS format, such as your file of digitized coordinates produced by TPSDig.exe. In addition to this file, you will need a measurement protocol file that gives the list of measurements you want to have calculated. This protocol is a three-column list; the first is the number of the measurement, the second is the number of the landmark at one endpoint of the measurement, and the third is the number of the landmark at the other endpoint. For example, the measurement protocol for lengths measured between landmarks 1 and 7, landmarks 2 and 4, and landmarks 4 and 5 is:

```
1  1  7
2  2  4
3  4  5
```

If you are loading a file in X1,Y1,... CS format, or a TPS file that includes a scale factor, **TradMorphGen** will have enough information to calculate the absolute distances between landmarks. If your file is a TPS file without a scale factor, your lengths will be calculated in terms of pixels, and you will need to convert them to your desired units. To do that, include the endpoints of your ruler in your protocol file, then divide the length of the ruler measurement by its absolute length (e.g. 10 mm); that quotient is the scaling factor you will need to rescale all the other measurements in your file. It is easier to produce a file of shape coordinates using **CoordGen**, which does the rescaling for you.

To load a file, click on the **Load** button for your file type. For example, if you are loading a file of shape coordinates in IMP format (X1,Y1,... CS), click the button that says **Load Data Set(XY ... CS format)**. The landmarks will appear in the visualization window. Then load the measurement protocol by clicking on **Load Measurement Protocol**. The protocol will now appear in the visualization window, along with the landmarks, so you can make sure it's correct. You can copy that picture to the clipboard by clicking on **Copy Image to Clipboard.** If you want to remove the landmarks from the picture so you see only the protocol, click on **Show Protocol Only**. Then click on **Calculate Traditional Length Set**.

You can save three output files: (1) the traditional length set; (2) geometrically scaled traditional measurements (="Constrained Length Set"); or (3) log-transformed traditional measurements (they are transformed to natural logarithms). Click on the button(s) for the files you wish to save.

You can now load another data file without reloading the protocol by clicking on **Clear Input Data, Retain Protocol**.

References

Alberch, P., Gould, S. J., Oster, G. F. and Wake, D. B. (1979). Size and shape in ontogeny and phylogeny. *Paleobiology*, **5**, 296–317.

Bininda-Emonds, O. R. P., Jeffrey, J. E. and Richardson, M. K. (2003). Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development. *Proceedings of the Royal Society of London, Series B*, **270**, 341–346.

Cheverud, J. M. (1984). Quantitative genetics and developmental constraints on evolution by selection. *Journal of Theoretical Biology*, **110**, 155–172.

Cope, E. D. (1887). *The Origin of the Fittest*. McMillan.

Falsetti, A. B. and Cole, T. M. (1992). Relative growth of the postcranial skeleton in callitrichines. *Journal of Human Evolution*, **23**, 79–92.

Fuiman, L. A. (1983). Growth gradients in fish larvae. *Journal of Fish Biology*, **23**, 117–123.

Galis, F. and Metz, J. A. (2001). Testing the vulnerability of the phylotypic stage: on modularity and evolutionary conservatism. *Journal of Experimental Zoology (Molecular and Developmental Evolution)*, **291**, 195–204.

Galis, F., van Alphen, J. M. and Metz, J. A. (2001). Why five fingers? Evolutionary constraints on digit numbers. *Trends in Ecology and Evolution*, **16**, 637–646.

Gould, S. J. (1977). *Ontogeny and Phylogeny*. Harvard University Press.

Gould, S. J. (1984). Morphological channeling by structural constraint: convergence in styles of dwarfing and gigantism in *Cerion*, with a description of two new fossil species and a report on the discovery of the largest *Cerion*. *Paleobiology*, **10**, 172–194.

Hall, B. K. (1992). *Evolutionary Developmental Biology*. Chapman and Hall.

Huxley, J. S. (1932). *Problems of Relative Growth*. MacVeagh.

Jolicoeur, P. (1963). The multivariate generalization of the allometry equation. *Biometrics*, **19**, 497–499.

Katz, J. M. (1980). Allometry formula: a cellular model. *Growth*, **44**, 89–96.

Kimmel, C. B., Ballard, W. W., Kimmel, S. R. et al. (1995). Stages of embryonic development of the zebrafish. *Developmental Dynamics*, **203**, 253–310.

Klingenberg, C. P. (1998). Heterochrony and allometry: the analysis of evolutionary change in ontogeny. *Biological Reviews*, **73**, 79–123.

Klingenberg, C. P. and Ekau, W. (1996). A combined morphometric and phylogenetic analysis of an ecomorphological trend: pelagization in Antarctic fishes (Perciformes: Nototheniidae). *Biological Journal of the Linnean Society*, **59**, 143–177.

Laird, A. K. (1965). Dynamics of relative growth. *Growth*, **29**, 249–263.

Laird, A. K., Barton, A. D. and Tyler, S. A. (1968). Growth and time: an interpretation of allometry. *Growth*, **32**, 347–354.

Lande, R. (1979). Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. *Evolution*, **33**, 402–416.

Lande, R. and Arnold, S. J. (1983). The measurement of selection on correlated characters. *Evolution*, **37**, 1210–1226.

McKinney, M. L. and McNamara, K. J. (1991). *Heterochrony: The Evolution of Ontogeny*. Plenum Press.

Raff, R. A. (1992). Direct-developing sea urchins and the evolutionary reorganization of early development. *BioEssays*, **14**, 211–218.

Richardson, M. K., Hanken, J., Gooneratne, M. L. et al. (1997). There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anatomy and Embryology*, **196**, 91–106.

Sander, K. (1983). The evolution of patterning mechanisms: gleanings from insect embryogenesis and spermatogenesis. In *Development and Evolution* (B. C. Goodwin, N. Holder and C. C. Wylie, eds) pp. 137–159. Cambridge University Press.

Seidl, F. (1960). Körpergrundgestalt und Keimstruktur. Eine Erörterung über die Gundlagen der vergleichenden und experimentellen Embryologie un deren Gültigkeit bei phylogeneticschen Überlegungen. *Zoologische Anzeiger*, **164**, 245–305.

Shea, B. T. (1992). Ontogenetic scaling of skeletal proportions in the talapoin monkey. *Journal of Human Evolution*, **23**, 283–307.

Shea, B. T. (2002) Are some heterochronic transformations likelier than others? In *Human Evolution through Developmental Change* (N. Minugh-Purvis and K. J. McNamara, eds) pp. 79–101. Johns Hopkins Press.

Slack, J. M., Holland, P. W. and Graham, C. F. (1993). The zootype and the phylotypic stage. *Nature*, **361**, 490–492.

Strauss, R. E. (1984). Allometry and functional feeding morphology in haplochromine cichlids. In *Evolution of Fish Species Flocks* (A. A. Echelle and I. Kornfield, eds) pp. 217–229. University of Maine.

Strauss, R. E. and Fuiman, L. A. (1985). Quantitative comparisons of body form and allometry in larval and adult Pacific sculpins (Teleostei: Cottidae). *Canadian Journal of Zoology*, **63**, 1582–1589.

van Snik, G. M. J., van den Boogaart, J. G. M. and Osse, W. M. (1997). Larval growth patterns in *Cyprinus carpio* and *Clarias gariepinus* with attention to the finfold. *Journal of Fish Biology*, **50**, 1339–1352.

Voss, R. S. and Marcus, L. F. (1992). Morphological evolution in muroid rodents. 2. Craniometric factor divergence in 7 neotropical genera, with experimental results from *Zygodontomys*. *Evolution*, **46**, 1918–1934.

Wayne, R. K. (1986). Cranial morphology of domestic and wild canids: the influence of development on morphological change. *Evolution*, **40**, 243–261.

White, J. F. and Gould, S. J. (1965). Interpretation of the coefficients in the allometric equation. *American Naturalist*, **99**, 5–18.

Zelditch, M. L., Sheets, H. D. and Fink, W. L. (2001). The spatial complexity and evolutionary dynamics of growth. In *Beyond Heterochrony: The Evolution of Development* (M. L. Zelditch, ed.) pp. 145–194. John Wiley & Sons.

Zelditch, M. L., Sheets, H. D. and Fink, W. L. (2003). The ontogenetic dynamics of shape disparity. *Paleobiology*, **29**, 139–156.

14

Morphometrics and systematics

Systematists use morphometrics to answer three types of questions. The first, which we label "taxonomic," asks whether populations are drawn from multiple species, and, if so, by what variable(s) they are most effectively discriminated. The second, which we label "phylogenetic," asks about phylogenetic relationships among taxa. Although they cannot be used to construct cladograms, morphometric analyses might nonetheless be useful for finding informative characters (for a more detailed discussion of characters, see below). The third, which we label "evolutionary," asks about the evolutionary history of the feature of interest – which, for our purposes, is shape. These are all interrelated issues, but there are important distinctions that bear on choosing the appropriate analytic method. Most importantly, taxonomic discriminators are often not equivalent to phylogenetically informative characters, so finding discriminators is not equivalent to finding characters. Also, characters usually comprise a subset of features that evolve, so tracing characters on a cladogram does not fully reconstruct the evolution of shape. Unfortunately, of the three types of questions, only those relating to taxonomic discrimination are so straightforward that they require nothing more than standard morphometric tools. This does not mean taxonomic discrimination is easy; on the contrary, it can be very difficult. However, compared to finding characters, or reconstructing the evolution of shape, the difficulties of taxonomic discrimination pale. At present we have no generally accepted method for finding characters, and it is not even clear what a method of character discovery would look like.

It might seem obvious that taxonomic discriminators are potential characters because there are differences among taxa, and characters are also features that differ among taxa. However, taxonomic discriminators are not characters because they describe the net difference between taxa; they are vectors extending between (or among) terminal taxa. The vector describes the direction in which the taxa can be distinguished from each other, regardless of whether the features distinguishing them are unique to one species, are shared by a group containing two species in the analysis, or are more broadly shared (with taxa not included in the analysis). All that matters is that the discriminator exists (telling us that the taxa are indeed different) and is successful (allowing us to identify unknowns correctly). In contrast to a discriminator, a character is a feature shared by members of a

ISBN 0–12–77846–08
Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

monophyletic group. In principle, a character is a feature that we place at the node of a cladogram. If we could measure shapes at successive nodes we could find a character by a simple pairwise comparison between them, but usually we do not have samples of taxa at successive nodes, and even if we did we would not know where to place them before reconstructing the cladogram.

It might also seem obvious that changes in shape characters, traced on a cladogram, reconstruct evolutionary transformations in shape. However, finding (and tracking) characters, and reconstructing the evolution of shape, are different exercises. When looking for characters we select particular features as informative, making no effort to provide a complete description of the changes in shape (or of the ancestral shape). For example, we might say that all members of a particular group have a shallow body compared to the other species, and "shallow body" is then selected as a character. However, we would not try to infer how much change occurred in body depth or how shallow their ancestor was, or to describe the ancestor's head shape. In contrast, when reconstructing the evolution of shape we need to infer the ancestral configuration of landmarks and the direction and magnitude of change along each branch.

Because taxonomic discrimination is a straightforward problem, we say little about it in this chapter, merely mentioning some conceptual issues that might arise before applying conventional morphometric methods to the data. We also say little about the third type of question, because the methods usually used to answer these questions raise issues that are outside the scope of morphometric theory. Those methods either (1) minimize a distance or squared distance over the cladogram (which in our case would be a Procrustes distance); or (2) use an explicit model of the evolutionary process and estimate values of the model's parameters that maximize the likelihood of the data, given the model (an accessible, general overview of these approaches can be found in Felsenstein, 2002, and a discussion of them in context of geometric shape data can be found in Rohlf, 2002). Although these procedures could be used to infer the cladogram, they have rarely been used for that purpose. The primary issue facing users of these methods is to choose (or develop) a realistic, justifiable model – a matter that involves considerations of evolutionary biology rather than morphometrics. Accordingly, we do not discuss this topic beyond listing the range of models that could be used and sources of information about them. In contrast, the second problem, that of finding characters in morphometric data that can be used to infer a cladogram (by standard cladistic approaches), raises profound methodological questions, with no satisfying answers.

For systematists, the lack of recommended methods for finding characters will make this a disappointing chapter. We had seriously considered the possibility of leaving out a chapter on systematics, but chose to include one for two major reasons. The first, and most important, is that we will never have a satisfactory method until we can tailor it to the question(s) at hand. Doing so requires stating the question(s) precisely enough to find a mathematical method for answering it. At present this is difficult, partly because some concepts (especially that of a "character") are so basic that they are difficult to articulate clearly. Systematists might not see any need to define that term because we all know, at least tacitly, what it means. However, we cannot tailor a method to find characters when we cannot say what the method should find – we need to define the problem before we can look for solutions. It is likely that this will be an iterative procedure – starting by stating one specific problem, proposing a method for solving it, realizing that a method is flawed (and why) and, based on that realization, revising the statement of the problem and then attempting another solution that may also fail, but for different reasons. In this chapter, we discuss one statement of the problem and one failed solution previously offered by us (Fink and Zelditch, 1995; Zelditch et al., 1995). Understanding *why* it fails is as important as realizing that it does when the aim is to avoid making the same kind of mistake again. Our second reason for including a chapter on systematics is to discourage readers from applying the method we previously suggested, giving two reasons why it should not be used (see also Adams and Rosenberg, 1998 and Rohlf, 1998).

We begin by discussing some issues that arise in taxonomic studies, then turn to the more difficult problem of finding characters in morphometric data.

Taxonomic discrimination

The taxonomic question can be divided into two parts:

1. Are the samples different enough to warrant judging them to be different species?
2. In what do they differ?

To answer the first, we must decide what would be "different enough." Once we state that criterion, we can ask whether the data meet it. We might say that "different enough" means that no more than 2% of the specimens are misclassified, or that the means of the samples differ statistically significantly, or even that the Procrustes distance between the means is minimally 0.03 (or any other favored value). Once we have chosen our criterion, which might be a combination of those, we can easily determine whether the data meet it. To that end, we could use CVA, or measure the Procrustes distances among means, or both.

Before applying either method, we need to consider what to do about geographic variation, ontogeny, sexual dimorphism and other factors that might complicate distinguishing species. Obviously, we do not want to claim that we have evidence for two species when the samples differ only in average developmental age or body size. If that might be the case, it would be useful to design the sampling scheme to ensure that the samples are homogenous and comparable, or else to standardize the data to a common age or size (using techniques discussed in Chapter 10). The results can be very different. For example, Figure 14.1 shows results from three analyses: (1) samples are compared without standardizing by ontogenetic stage (Figure 14.1A), (2) samples are compared at a common juvenile stage (Figure 14.1B), and (3) samples are compared at a common adult stage (Figure 14.1C). In all three analyses, all eight CVs are significant, and, with one exception (the unstandardized data), the misclassification rate is extremely low. For the unstandardized data, out of 390 specimens as many as 12 are misclassified, all of which are *Pygocentrus nattereri* that are classified either as *P. cariba* or *P. piraya*. However, for both standardized data sets no more than four individuals are misclassified (also *P. nattereri*). Not surprisingly, all species differ from all others significantly (in all pairwise comparisons, $p < 0.002$). In general, species differ by a Procrustes distance of more than 0.030, except for the three *Pygocentrus*, whose adults differ from each other by Procrustes distances as small as 0.027–0.028 (and by even less in comparisons of unstandardized specimens). Thus we would draw the same conclusion about the taxonomic status of these samples from all three analyses, but the results still differ because the variables discriminating among the species are different.

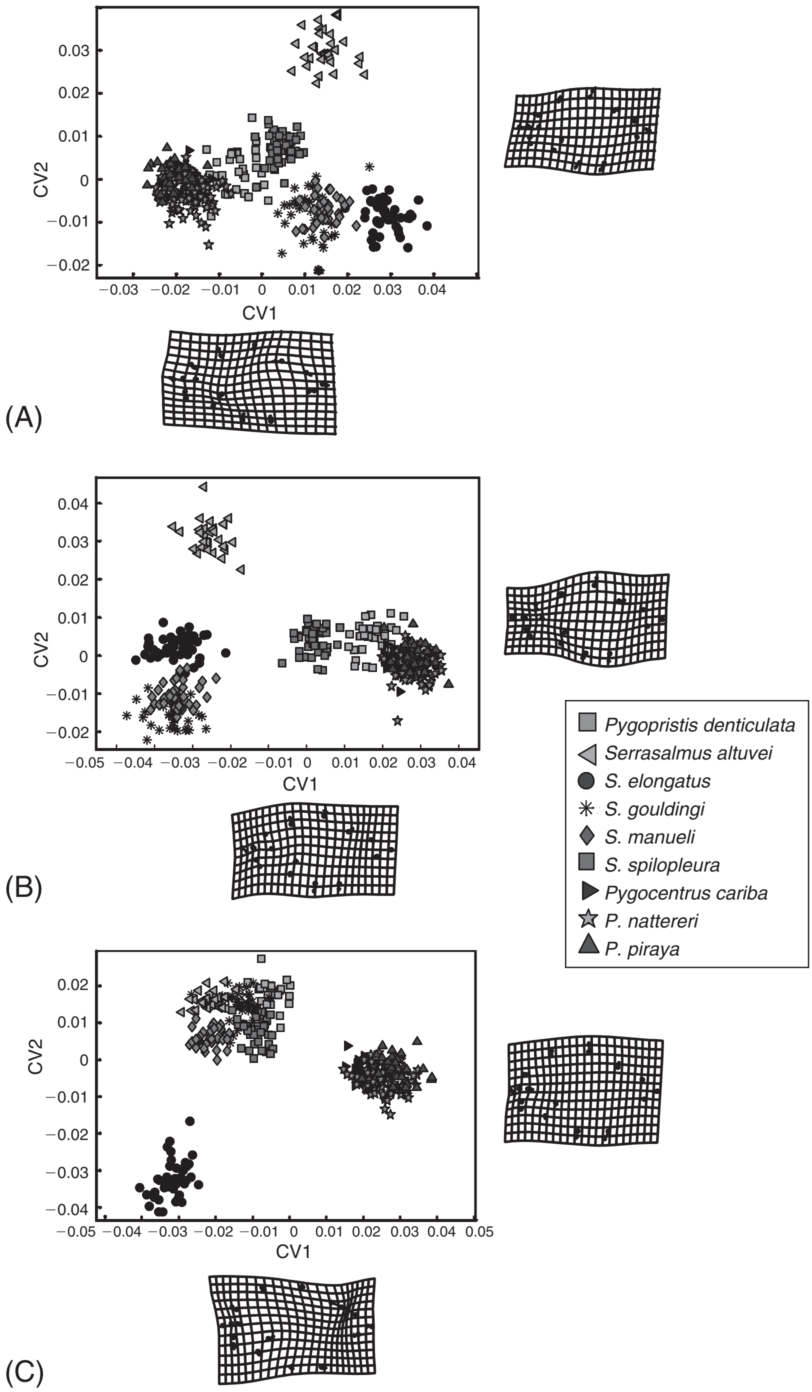


Figure 14.1 CVA of body shape of nine species of piranhas: (A) unstandardized data; (B) data are standardized and comparisons are made among juvenile shapes (at the transition from larval to juvenile phases); (C) data are standardized and comparisons are made among adult shapes (at the maximum body size regularly attained by each species).

After applying morphometric techniques to the data, we are still left with the problem of interpretation. Even if the data meet our criteria, the samples might not come from different species – they could come from geographically differentiated populations that were sampled only at the extremes of their range (e.g. the most northern and the most southern localities). Had they been sampled throughout the entire range, we might find that there is no statistically significant difference between geographically adjacent populations. Conversely, they might not meet our criterion but nonetheless be distinct species; it is just that the distinguishing features do not lie in shape. CVA provides a useful method for discrimination, but finding that samples can be discriminated is only part of the answer to the first taxonomic question.

The second taxonomic question (in what do they differ?) is also answered by CVA, which finds a mathematically optimal discriminator. That discriminator, however, might not be optimal for a biologist in the field. If it is intended to be useful for field biologists, no purpose is served by writing a key that requires digitizing specimens, entering their data in a CVA, and allocating them to species according to the discriminant function. Although that could be considered a merely technological limitation, a taxonomic key serves a pragmatic purpose and therefore must be useful. The key must be applicable to the specimens in hand, under the conditions when they are in hand. However, there are other matters that must also be considered when writing a key. In particular, the key characters that would allow a specimen to be identified correctly may depend both on the age (or size) of the specimen and on the age (or size) of those used in preparing the key. If the CVA is based on age- or size-standardized data but the specimen is not at the stage to which the data were standardized, it might not have the key characters. Conversely, if the key is based on ontogenetic series, the key characters might show enormous ranges within species.

Writing a useful key can be a challenging problem, but turning a geometric analysis into a useful key adds no further difficulties. That can be done by using geometric morphometrics to determine the shape variables that best discriminate, then translating them into terms of traditional morphometric variables that can be measured with calipers or rulers. If we find, for example, that relative body depth discriminates between species, we can calculate two lengths; one for the depth measured between two landmarks (such as anterior bases of the dorsal and anal fin), and standard length. That ratio does not fully describe the shape differences among species, but it suffices to identify unknown specimens.

Finding characters

The use of morphometric data in phylogenetic studies has long been controversial. Most often, debates among phylogenetic systematists have focused on two issues: (1) methods for coding variables that overlap, sometimes considerably; and (2) the reliability of the information obtained from the data for inferring phylogenies. Morphometric data have been viewed with suspicion partly because it is difficult to determine where to draw the line when there are no distinct gaps between the observed values. A wide variety of techniques have been proposed and debated heatedly (see, for example, Colless, 1980; Simon, 1983; Archie, 1985; Goldman, 1988; Chappill, 1989; Thiele, 1993; Swiderski et al., 1998). Only very recently has the discussion begun to focus on a more fundamental problem: what to code? What is it that we are extracting from the data and treating as a character? Clearly, this issue must be addressed before the first one is even relevant; coding becomes a moot issue if there are no characters to code, and if there are no characters, we cannot test the hypothesis that they are especially homoplastic.

It is clear that partial warps should not be used as characters (for the reasons discussed below), but it is not clear what ought to be used instead. It is not even clear that the problem has a solution. The major objective of the first part of this section is to define the problem we had hoped to solve using partial warps, then to explain why our approach was flawed. In the next section we discuss two alternatives, both of which rely on conventional multivariate methods, but neither is precisely tailored to the problem.

Defining the problem

The general problem we face is to find features that differ among taxa and are shared by a subset of them. The differences indicate evolutionary novelties and the similarities indicate common ancestry, although we will not be able to determine which are novelties until we have completed the phylogenetic analysis. We would not expect that an entire shape is a character because species rarely have exactly the same shape (whether we are comparing whole bodies or parts of them). If we think of the problem from the perspective of whole landmark configurations, we will not make any progress. On the other hand, if we do not think of the problem in terms of whole landmark configurations, we may be led to theoretically invalid solutions. Therefore the problem is to analyze entire configurations of landmarks, and find features that differ among taxa and are similar among a subset of taxa. Additionally, to say that we have a character we must be able to say *where* it is, and over how large a spatial expanse it extends. A primary objection to traditional morphometric variables is that they are lines, having no spatial extent as individual variables. As soon as we try to determine their spatial location and extent, even by multivariate analyses, we face one of the most severe limitations of traditional morphometric data – their poor ability to localize morphological differences.

When looking for these similarities and differences, we are not concerned with the magnitude of the difference, nor its degree of localization. Small differences (so long as they are large enough to be considered a difference at all) count as much as large ones, and spatially large-scale differences count as much as localized ones. Consequently, neither the Procrustes distance between taxa nor the bending energy of the transformation has any relevance to the problem. This is one of the reasons why it is so difficult to solve – neither of the metrics used in geometric morphometrics is germane to the problem, and if there is a relevant metric, it has yet to be defined.

When we first approached this problem, we focused on one major limitation of conventional (qualitative) approaches: that organisms are often dissected arbitrarily, along lines of convention. Conventional anatomical subdivisions are often not biologically meaningful except in the context of a particular problem. For example, if we are interested in locomotion and foraging, we can subdivide an organism into parts that are used in locomotion and parts that are used in foraging. Alternatively, if we are interested in development, we can subdivide the organism into parts that have a common germ-layer origin, or that develop from the same type of bone, or that undergo the same kinds of epigenetic interactions, etc. These subdivisions have long been regarded as arbitrary, except to the extent that they are useful in a particular investigation. These subunits are not suitable for dissecting an organism in systematic studies when a single character crosses several such divisions, or is partly within and partly without them. Our goal in using partial warps was to find a more objective basis for dissection. We did not succeed (for reasons discussed below) but the problem we defined remains a fundamental and unresolved difficulty for character analysis. Our method had fatal flaws, but so do others that require us to decompose the organism prior to measurement.

The approach we took is similar to one that is standard in cladistic studies using morphometric data. We defined a set of variables *a priori*, and compared taxa with respect to them. A similar tactic is applied to conventional morphometric variables, when a set of lengths or ratios is defined and measured on taxa, then the values of those lengths or ratios is compared among them. Most attention has focused on the problem of coding those variables, but coding is the least of the problems. Such variables do not solve the problems we had hoped to address, but share with them the flaw that we inadvertently introduced: they compare arbitrarily selected components of shape one at a time.

Why not to use partial warps as characters

Even though partial warps have a geometric scale, are a function of homologous landmarks, and do not emphasize differences of large magnitude at the expense of small ones, they cannot be used as characters for at least two reasons. The first is obvious (in hindsight at least): partial warps have a spatial scale, but an individual partial warp (PW) describes only part of a small-scale anatomical feature. Partitioning a change by PWs does not correspond to partitioning it by anatomy or by characters, because a single PW does not describe a single, spatially coherent change (although several, taken together, might). When comparing multiple taxa, a combination of several PWs taken together is usually needed to describe any change, even one that is anatomically local. Additionally, having a high score on a localized PW does not mean that there is a localized change. Instead, the change within that region may be partly described by a PW at a higher spatial scale, and the localized PW supplements that description. Taken out of context of the larger-scale PW, we cannot make anatomical sense of the one at smaller scale. Two taxa that have identical values for a small-scale PW might differ anatomically – differences that cannot be seen without looking at all PWs.

All that may be obvious to readers who have reached this chapter, but to clarify the point we can re-examine the example that we found most promising at the time – the ontogenetic change in scores on one PW (Figure 14.2). Two of the taxa, which were used as outgroups (*Pygopristis denticulata* and *Serrasalmus gouldingi*), have statistically significant ontogenetic change on that PW (in both *X* and *Y* directions), whereas the three *Pygocentrus* do not. We would not normally be concerned about similarities among outgroups, but this example shows that similarities implied by individual PWs are not found in complete descriptions. That *P. denticulata* and *S. gouldingi* have anything in common in their development of that region is not at all obvious when looking at more complete descriptions of the five ontogenies (Figure 14.3). They are similar to each other, and differ from the three *Pygocentrus*, only in that they undergo an ontogenetic change in the caudal peduncle region that is not fully described by PWs at higher spatial scales. However, *P. denticulata* and *S. gouldingi* are not similar to each other in the changes described by the higher spatial scales (and neither are the three *Pygocentrus*). Being similar

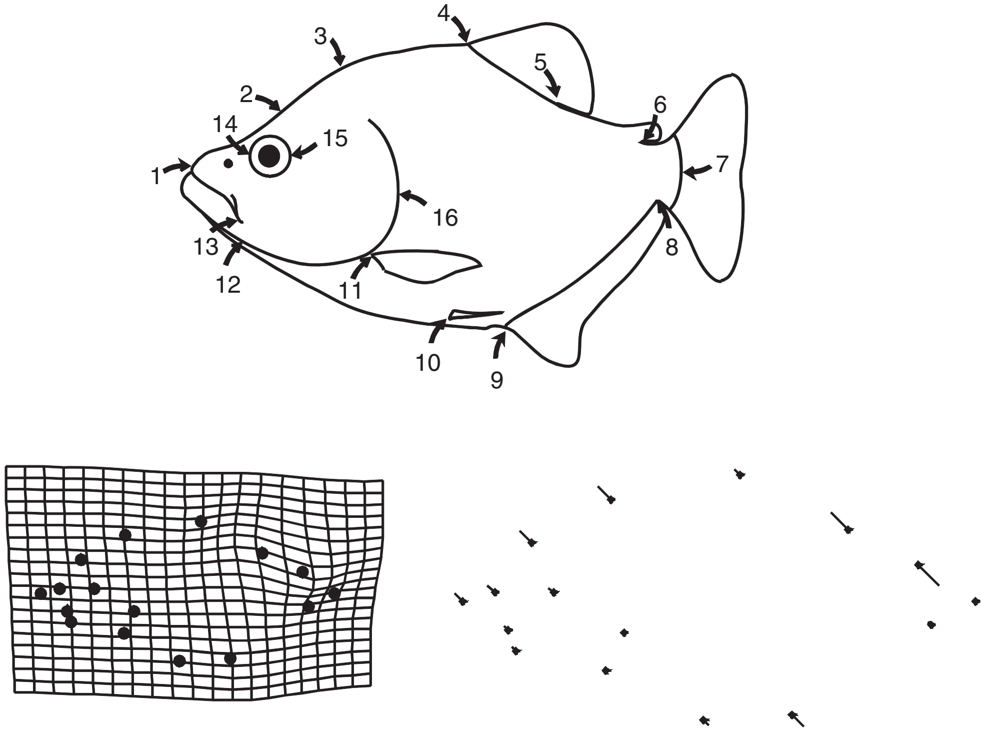


Figure 14.2 A single PW used to exemplify the procedure for finding systematic characters in ontogenies of shape in Fink and Zelditch, 1995.

in one PW does not mean being similar in shape (or ontogeny of shape) in a particular anatomical region. When looking at one PW we lose the context supplied by all the others, and PWs are all context-dependent. Therefore, we cannot describe what happens within any one region of the body without placing every PW in context of every other. Even judged by what the method was supposed to do, it fails; it does not provide an objective, non-arbitrary method for decomposing changes (except in a purely geometric sense).

The second issue, related to the one above but important in a broader context, is that interpretations based on individual variables violate the fundamental principles of geometric shape analysis – that results be invariant to the selection of variables. Obviously, a result that depends on using partial warps is invalid (even if the phylogenetic inference based on it happens to be valid). A partial warp score is a single variable, a one-dimensional projection onto a particular basis, and our results cannot depend on that choice. Adhering to that basic principle does not mean that our *phylogenetic results* will be invariant to our choice of *characters* – the results of a phylogenetic analysis always depend on the characters. Rather, it means that our recognition of characters must be invariant to the selection of variables – and for that reason, a morphometric variable cannot be a character in its own right.

The obvious question is: how *can* we discover characters when we cannot look at individual variables? If variables do not provide a legitimate basis for subdividing the organism, and if conventional anatomical lines of dissection are also viewed as biologically

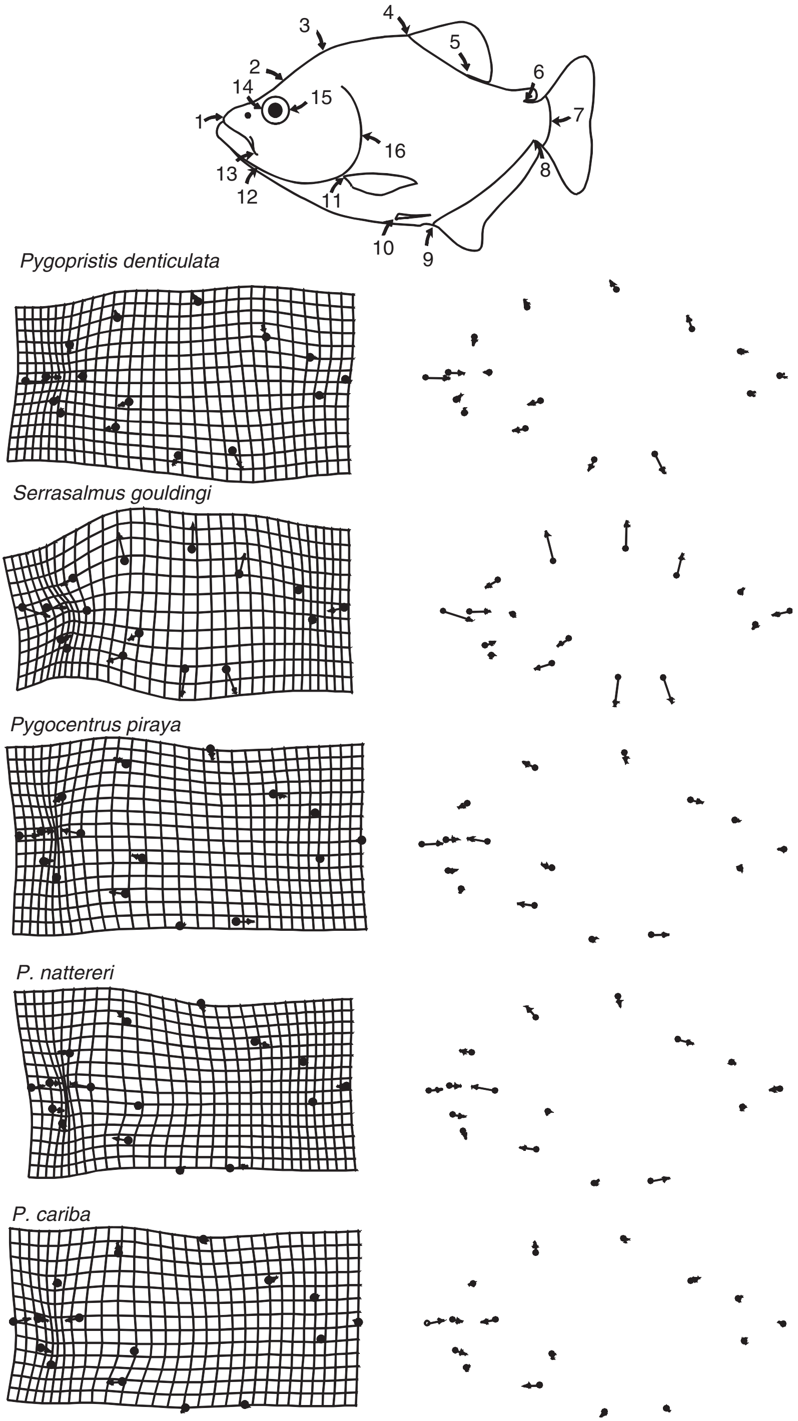


Figure 14.3 Ontogenies of shape for the species analyzed in Fink and Zelditch, 1995. The inference drawn from the PW shown in Figure 14.2 is that the outgroup species *P. denticulata* and *S. gouldingi* have a localized ontogenetic change in the length and depth of the caudal peduncle relative to the region between dorsal and adipose fins, whereas the three *Pygocentrus* do not.

arbitrary, then where can we look for characters? We end this section with that question because we have no satisfying answer. In the next section, we discuss two possible lines of attack. One uses a standard multivariate ordination method, principal components analysis (PCA), to explore similarities and differences, the other uses pairwise contrasts to find differences, which are then compared to find similarities among taxa in their differences from others. Neither method is tailored to the problem, but both represent feasible approaches that can be used in the interim, until we have a more satisfying method.

Using PCA to find characters

PCA provides a coordinate system for shape analysis, and may be useful for finding characters, but individual PCs (like individual PWs) cannot be viewed as characters in their own right. Just as a partial warp score is a projection onto a single axis, so is a principal component score, and just as a similarity on one PW does not indicate a similarity in shape, similarity on a single PC might not demonstrate a sufficiently general (or detailed) similarity.

Like PWs, PCs are context-dependent, and thus we would not expect a PC to be a character any more than a partial warp is. That is not to say that the two methods are strictly comparable – there is a major difference between them. Principal components are orthogonal directions of variation rather than orthogonal components of bending energy, and variation is biologically relevant to the problem at hand while bending energy is useful only in that it is used by the method for depicting the results. PCs have a biological meaning, as orthogonal dimensions of variance, even though that is not equivalent to the meaning of a character. They are not likely to be characters in their own right because they are directions of variation that are constrained to be orthogonal (by definition), not directions of evolutionary change. Directions of evolutionary change are likely to be oblique to the PCs – they are within the space spanned by the PCs, but they need not lie along an axis nor must they be orthogonal.

Although PCs are not likely to be characters, we may still find PCA useful for exploring similarities and differences. The scatter plots allow us to see the variation among taxa, and their overlap, and both are important for finding characters. However, just as we need to interpret partial warps in combination, so we also need to interpret PCs in combination. Just because two or more species overlap in their PC1 scores does not mean that they are similar with respect to all features described by PC1. They may differ in some, so that PC1 splits the difference between them and the other PCs describe what is specific to their deviations from PC1. Taxa located in different quadrants of a scatter plot may differ considerably in shape, depending on the proportion of the variance described by each PC and on how the PCs overlap in their descriptions of variation within the same regions. For example, we can look at a case that should be familiar by this point – the first two PCs of piranha juvenile body shape. The first, which accounts for 62% of the variance, clearly distinguishes three species (*S. manueli*, *S. elongatus* and *S. gouldingi*) from all others (Figure 14.4). Looking at the deformation that depicts the direction of greatest variance, we can see that body depth contributes heavily to it. However, PC1 is not only body depth; it also describes differences in proportions of the posterior body correlated with body depth. Species with high scores on this axis have relatively long

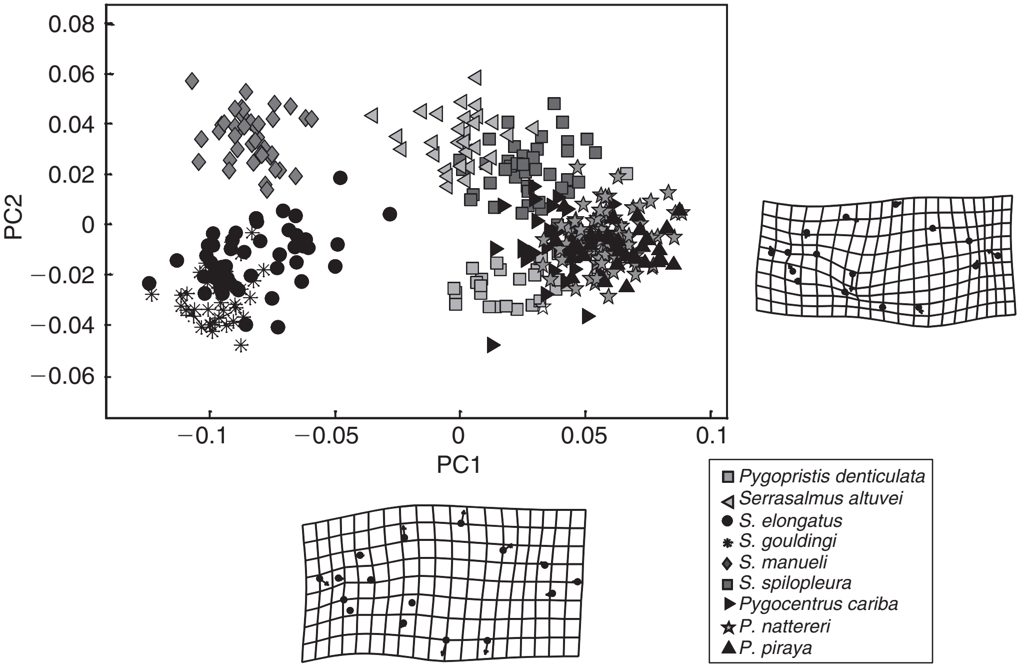


Figure 14.4 Principal components of body shape of nine species of piranhas; data were standardized and variation is examined among juvenile shapes (at the transition from larval to juvenile phases).

caudal peduncles compared to the region between dorsal and adipose fins, as well as deep bodies, but we cannot necessarily say that species with high scores on PC1 have long caudal peduncles if other PCs also describe variation in posterior body proportions and scores on those PCs differ among species with similar scores on PC1. PC2, which accounts for only 8.3% of the variance, also describes variation in caudal body proportions and, on this component, species with high scores have very short caudal peduncles relative to more anterior region. Consequently, species with high scores on both components have a short caudal peduncle relative to other species with equally high scores on PC1. In effect, PC2 partially "compensates" for PC1.

When interpreting PCs, it is also important to consider that they describe variation around an average shape. However, the average is not a "typical" piranha; rather, it is the shape of the consensus, the point having the coordinates 0, 0 (on all PCs). Obviously, the consensus is not a typical piranha since there are no specimens at the 0, 0 point. The outgroup (*P. denticulata*) is fairly near it, but if we want to describe differences between *P. denticulata* and other species (or to make any other comparisons among species) we cannot describe changes along one PC, then along another. The direction of the difference between particular species is often oblique to several PCs.

The importance of considering scores on several PCs becomes evident when comparing the three shallow-bodied species. All three have high scores on PC1, but they differ in scores on PC2. One of the three, *S. manueli*, has high scores on PC2 (as do *S. altuvei* and *S. spilopleura*). To see how *S. manueli* differs from *S. gouldingi* and *S. elongatus* with respect to their differences from other taxa, we can draw the vector extending from

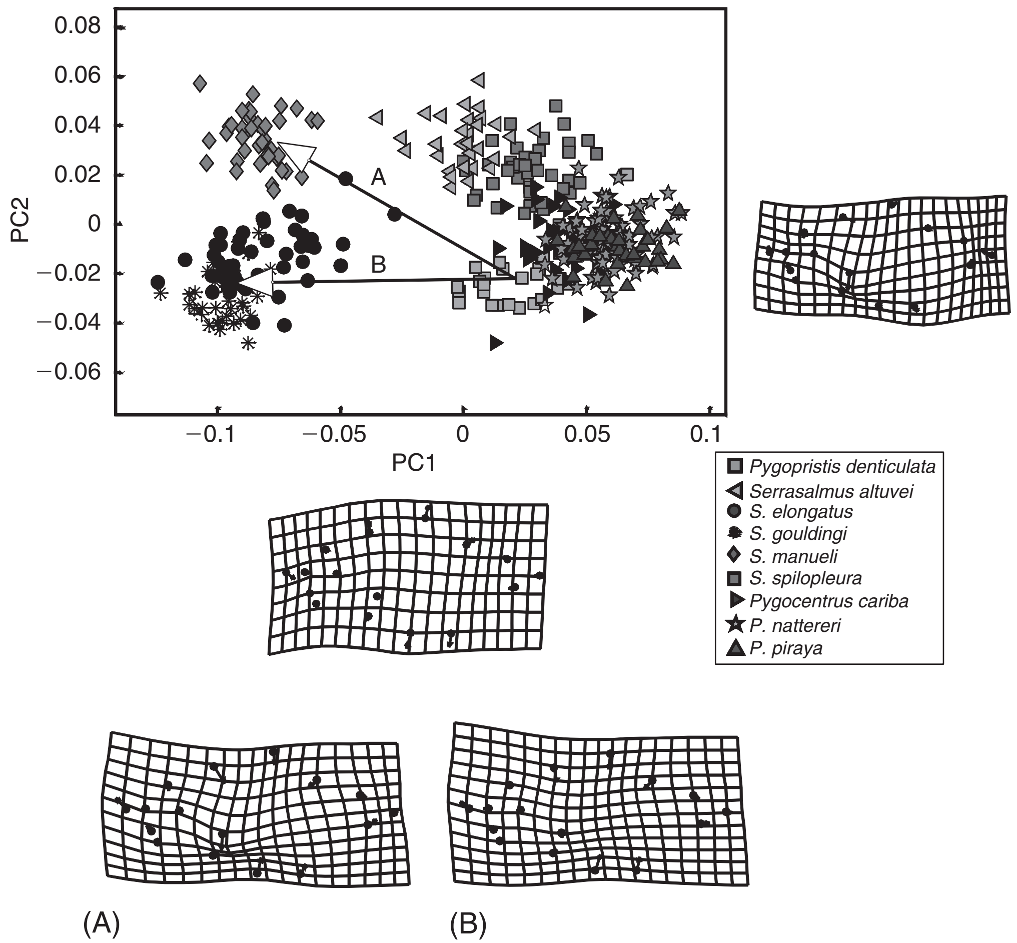


Figure 14.5 Analyzing direction in which species differ from *P. denticulata* in juvenile shape, to determine whether species with overlapping scores on PC1, but different scores on PC2, are similar with respect to features varying along PC1. (A) The direction of difference from *P. denticulata* to *S. manueli*; (B) the direction of difference from *P. denticulata* to *S. gouldingi* and *S. elongatus*.

those other taxa, e.g. *P. denticulata* to *S. manueli* (Figure 14.5A) and to *S. elongatus* and *S. gouldingi* (Figure 14.5B). The reason for doing this is to determine what differences from other taxa are shared by *S. elongatus*, *S. gouldingi* and *S. manueli*. We will then eliminate from the character the features peculiar to one species. Although we are making this comparison to the outgroup, we are not assuming that it has the primitive body shape. Comparisons to other species will also be necessary, and no decisions about polarity are made at this point. Based upon the similarities between the two vectors (Figures 14.5A, 14.5B), what all three taxa share is their shallow body. There also may be a second similarity not described by either PC – a shortening of the midbody relative to the head and posterior body. We could include that in the description of the character, but we would exclude the proportions of the caudal peduncle from that character description because *S. manueli* differs from the other two species in that. Clearly, the character is not equivalent to a PC.

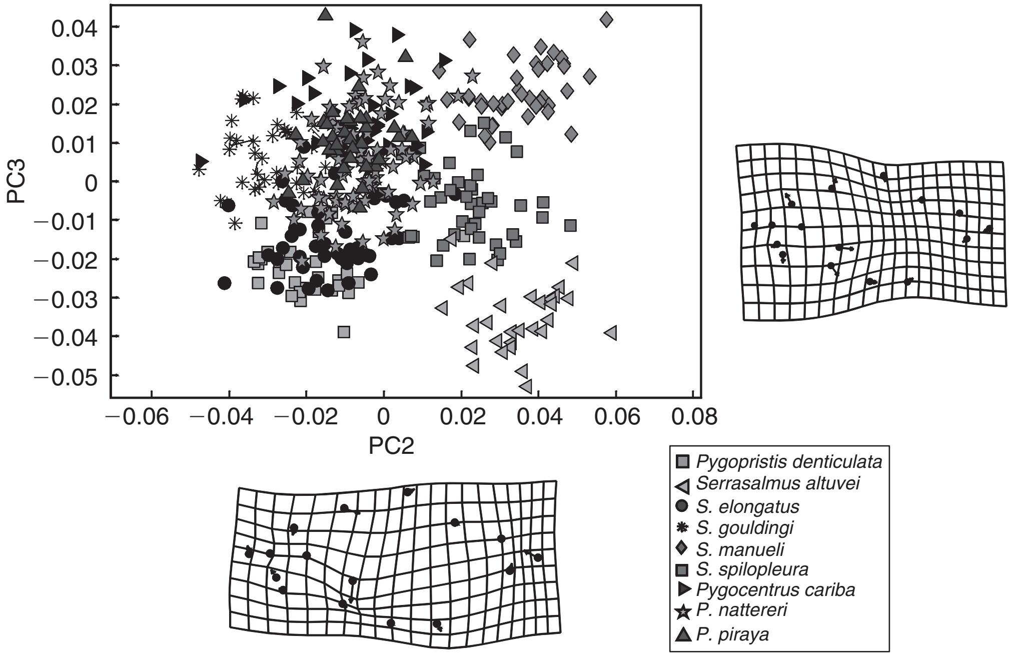


Figure 14.6 Scatter plot of PC3 on PC2, and the deformations depicting these two dimensions of variation.

The reason for not treating PC2 as a character in its own right is the same as the one we used to rule out treating individual PWs as characters. *S. manueli*, *S. altuvei* and also *S. spilopleura* have high scores on this one, which primarily describes a displacement of the opercle landmark towards the pectoral fin and a shortening of the caudal peduncle relative to the anal fin. However, *S. manueli* and *S. altuvei* differ along PC1 and are not similar in caudal peduncle proportions; they also differ along PC3 (Figure 14.6). Differences along PC3, as well as those along PC1, might belie the inference of morphological similarity implied by similar PC2 scores. Like PC2, PC3 accounts for only a small portion of the variance (5.6%), but like PC2 it describes a change in location of the pectoral fin relative to the opercle. *S. manueli* and *S. altuvei* have the highest and lowest scores on PC3, respectively, which means that their pectoral fins are displaced in opposite directions relative to the opercle, which needs to be taken into account when assessing their similarity on PC2. Despite their similar scores on PC2, a feature that might have been judged a morphological similarity might not be similar, by virtue of the differences along PC1 and PC3. Because of their different scores on PC1, we would also avoid construing their caudal peduncle proportions as similar, despite their similar values on PC2. Because species can be similar along one component and differ substantially along others, we cannot interpret one component at a time.

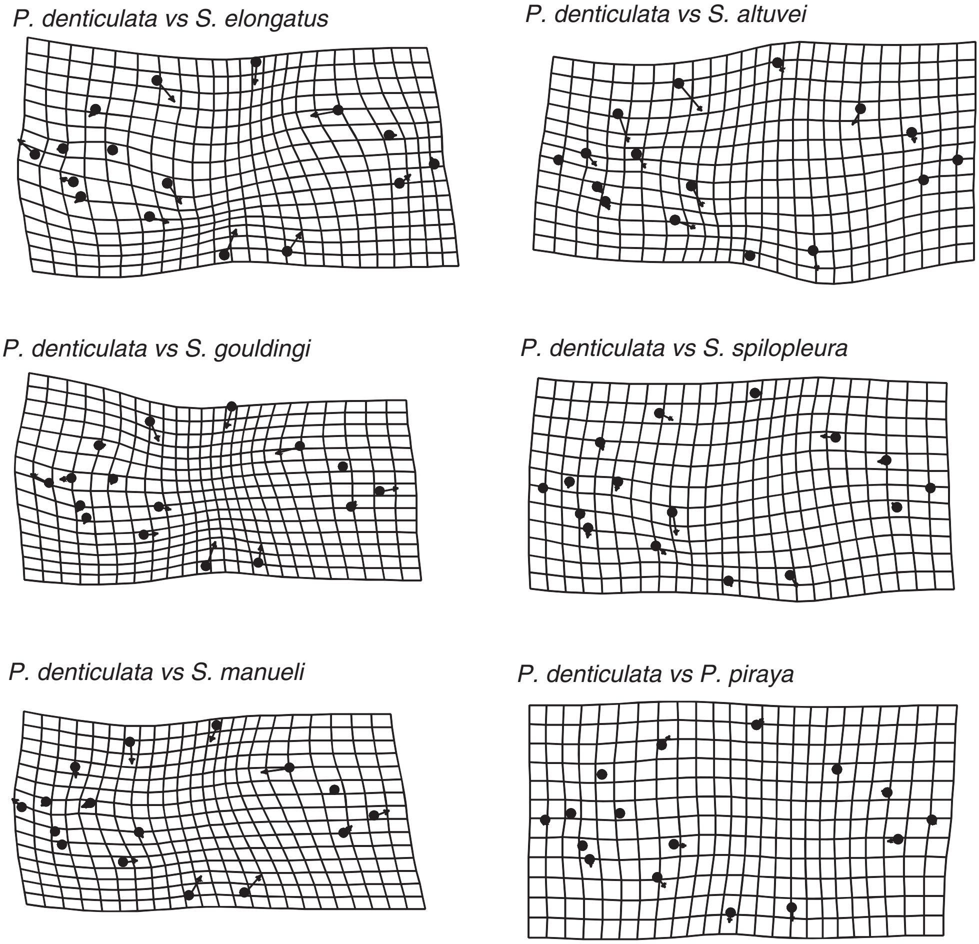
The strategy for combining PCs, outlined above, is undeniably tedious, but it might be successful at finding the features shared by two or more taxa. In cases like our example, when over 60% of the variation is along a single PC, two or more taxa have high scores and two or more have low ones, and there is virtually no overlap among the low and high scores, PC1 points to a character. When the variation is more evenly spread out across components, it will be necessary to combine many more of them because similarities on one may be outweighed by differences on the others. An obvious problem is that the comparison of the vectors in a single plane, such as we used to compare the similarities among the three shallow-bodied species with respect to their difference from the outgroup, examines only some of the differences among them in a single plane. We might prefer to look at *all* the differences between each species and the outgroup (or any other species taken as a standard), comparing *those* vectors among taxa.

Using comparisons between interspecific vectors to find characters

The basic idea of this approach is to compare all species to one other species (which is held constant). These pairwise contrasts can then be examined for similarities. By comparing the differences between one species and each of the others, we can then inspect the differences for similarities. The logic of the method is that we are looking for similarities in the differences – i.e. similarities among taxa in features specific to them. To exemplify this approach, we will continue the analysis of piranha juvenile body shape, comparing each species to the outgroup (Figure 14.7). Of course it is not necessary to use the outgroup in these comparisons; any species could be used as that "other," and it may be useful to use more than one before drawing conclusions.

From these comparisons, it is obvious that *S. elongatus*, *S. gouldingi* and *S. manueli* are shallow-bodied compared to all other piranhas. They differ profoundly from *P. denticulata* in this, whereas none of the other species do. This is the feature that dominated the PCA (and it is obvious by qualitative visual inspection as well). Additionally, these three species have a relatively short mid-body relative to the more posterior body, a feature hinted at but not so clearly presented when the vectors were drawn from *P. denticulata* to the species in the PC1–PC2 plane (Figure 14.5). This particular feature might reflect a decrease in the length of the dorsal fin relative to the posterior body (dorsally) and the posterior displacement of the pectoral fin and pelvic fins (rather than changes in proportions of body between them). The three shallow-bodied species appear to vary in the degree of "midbody contraction," but they appear to be consistently more contracted than the others. The possibility that these three species are similar in having a relatively shortened midbody is worth examining further because, unlike their shallow body, it is not obvious from a purely qualitative analysis.

To pursue that possibility further, we can compare the vectors of pairwise contrasts to each other, asking if a more contracted midbody (compared to that of *P. denticulata*) is characteristic of the shallow-bodied species but not of the others. This is done by subtracting one of the pairwise vectors from another; where species are identical to *S. gouldingi* (in the differences from *P. denticulata*) the grid is square (Figure 14.8). Large differences indicate that the direction of change from *P. denticulata* to *S. gouldingi* is not shared by another taxon. Subtracting each contrast from the contrast between *P. denticulata* to *S. gouldingi* shows that *S. gouldingi* is not much shallower or deeper than either *S. elongatus* or *S. manueli*. All three differ from *P. denticulata* by nearly the same degree, and in that same direction. Some differences are evident in the relative length of the midbody, however. The grid is slightly more contracted in that region, indicating that *S. gouldingi* is



P. denticulata vs S. elongatus

P. denticulata vs S. altuvei

P. denticulata vs S. gouldingi

P. denticulata vs S. spilopleura

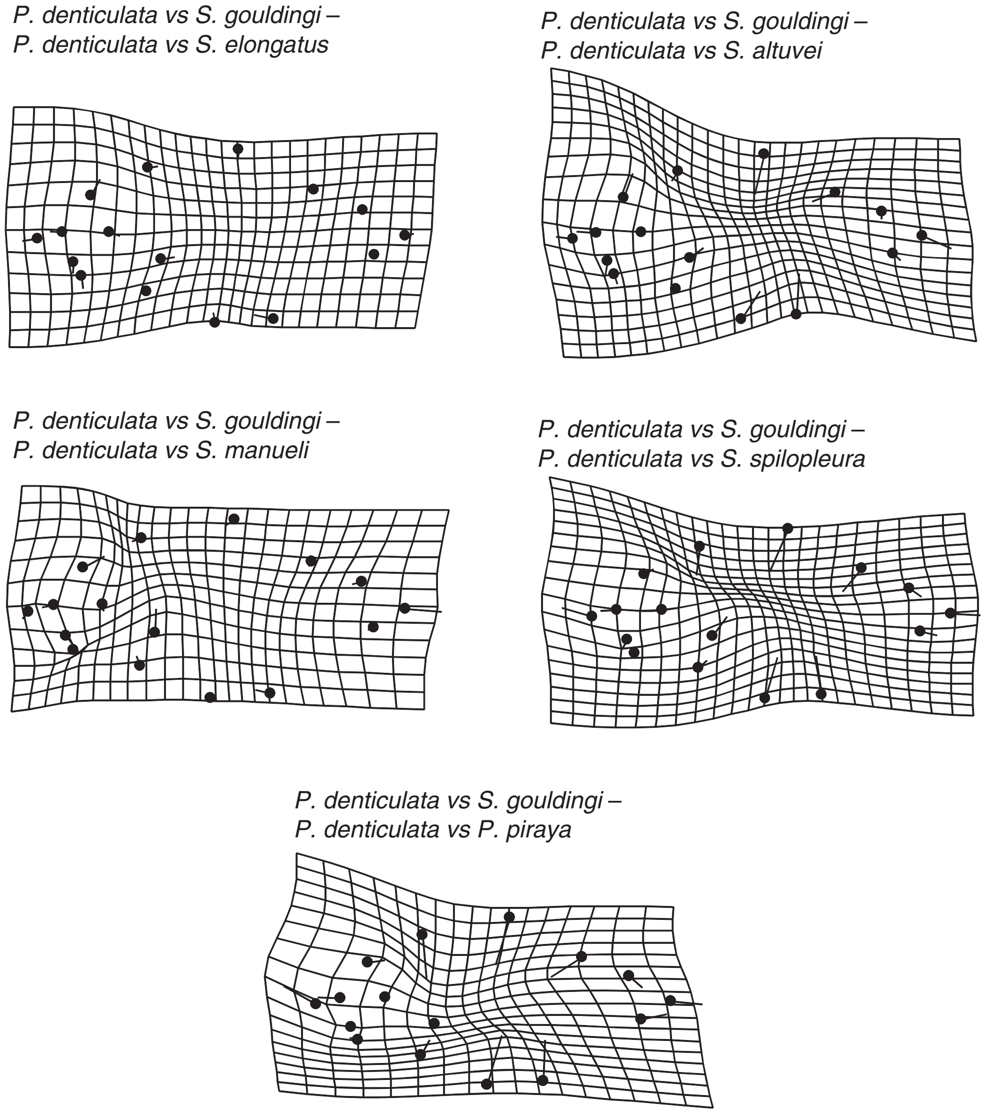
P. denticulata vs S. manueli

P. denticulata vs P. piraya

Figure 14.7 Pairwise comparisons between mean juvenile body shapes of *P. denticulata* and six other species. Comparisons to *P. nattereri* and *P. cariba* are not distinct from the comparison to *P. piraya*.

more extreme than the others in that feature. However, the differences are slight. In striking contrast, the comparisons to the other species indicate not only that *S. gouldingi* is far shallower than the others, but also that all differ from *S. gouldingi* in either the degree or the location of midbody contraction. We could either take these results to mean that *S. elongatus*, *S. gouldingi* and *S. manueli* are all shallow-bodied and contracted in the midbody compared to the other species, or we could continue the analysis, doing additional pairwise contrasts – this time between *P. denticulata* and *S. elongatus*, and also between *P. denticulata* and *S. manueli* – to determine that all three species are similarly different from the others. Of course we would need additional comparisons to find features that are more widely shared, or specific to some of the deeper-bodied species.

Unlike the shallow body, which is so evident visually that it requires no detailed quantitative study, the midbody contraction discerned in these comparisons is the kind of subtle feature that justifies the effort of a morphometric analysis.



P. denticulata vs S. gouldingi –
P. denticulata vs S. elongatus

P. denticulata vs S. gouldingi –
P. denticulata vs S. altuvei

P. denticulata vs S. gouldingi –
P. denticulata vs S. manueli

P. denticulata vs S. gouldingi –
P. denticulata vs S. spilopleura

P. denticulata vs S. gouldingi –
P. denticulata vs P. piraya

Figure 14.8 Comparisons among vectors describing the difference between *P. denticulata* and *S. gouldingi*, and the vector describing the difference between *P. denticulata* and each of the other six species shown in Figure 14.7. Each frame shows the contrast between the two vectors: where the squares of the grid are square, the two vectors are the same; where the grid shows large differences, the difference between that species and *P. denticulata* does not resemble the difference between *P. denticulata* and *S. gouldingi*.

Coding

Having found a character, we can treat it like any other. That is, if using conventional cladistic methods, we code the characters according to our preliminary judgments of homology, include it in the data matrix, and analyze that matrix by parsimony. Coding methods are a contentious subject; systematists vary considerably in their preferred criteria for coding. The debates have nothing to do with morphometrics except to the extent that the methods are applied to quantitative data, and that statistical methods are sometimes favored to decide whether species are different (and should therefore not be coded as having a homologous character). The literature on coding is large; interested readers can find general critiques of coding methods in several papers (e.g. Farris, 1990; Thiele, 1993; Gift and Stevens, 1997; Swiderski et al., 1998).

Any favored method can be applied to a variable that represents a character. In the case of "shallow body," PC1 is a reasonable proxy for the character, so we can apply our favored method to scores on PC1. It is not so easy to make decisions about characters that are combinations of several variables because we cannot easily examine the variation within species in more than two or three dimensions at a time, but doing so may be important for deciding whether species are similar enough to code their features as homologous.

Coding itself is a subject of debate. Not only are methods for coding controversial, even the idea of coding is. As mentioned earlier in this chapter, there are methods for inferring the evolution of shape that do not require coding characters. These methods use a very different approach to the problem. In particular, they do not make preliminary hypotheses of homology, then formalize them by codes, then infer the phylogeny that minimizes the net number of extra steps (a step is considered "extra" if it means that a putatively homologous character is reinterpreted as arising more than once). Instead, they use explicit models of the evolutionary process, among which are:

1. Randomly varying directions of natural selection in different lineages
2. Random genetic drift of species around a single, stable optimum
3. Randomly wandering optima
4. Constrained wandering of optima
5. Wandering optima whose paths are correlated, a correlation that diminishes over time
6. Bursts of change around the time of speciation with little or no change thereafter.

(For a more detailed synopsis of the models, see Felsenstein, 2002.) By using one of these approaches we could avoid the whole issue of coding, but we then have to confront the problem of deciding which model is reasonable and justifiable. Such models have not been widely used to infer cladograms from morphological data and, like the methods which minimize a net morphometric distance (linear or squared) over a tree, model-based methods might best be considered as methods for reconstructing the evolution of shape given a cladogram.

Summary

At present, no method is tailored to the problem of finding characters in morphometric data, and the available methods are cumbersome and involve an uncomfortable degree of subjectivity. Each could be improved by refining the part of the procedure that involves making linear combinations of variables, such as combining PC1, PC2 and PC3 to see whether similarities inferred from scores on one component are belied by scores on others. However, rather than improving methods that were devised to use standard morphometric techniques, it might be better to start at the beginning and develop a method tailored to our purposes. Doing so will require refining the statement of the problem. Currently we cannot state the problem in mathematical terms, and that is necessary before we can find a mathematical solution. Our original statement of the problem focused on one particular element of it: finding characters without having to dissect organisms arbitrarily into parts prior to the phylogenetic analysis. However, that dissection need not be an integral part of a method for finding characters. We could instead use partial least squares analysis (Chapter 11) to test the hypothesis that the blocks of landmarks do not covary; if they do not, we can analyze them separately. Even though PLS does not test the hypothesis that a block constitutes an integrated unit, it may provide a more informed dissection than one based purely on anatomical conventions.

Clearly, we need additional methodological research – we should not be limited to the methods currently available when others are feasible. We also need to complement the methodological investigations by a discussion of what our concepts mean. If we do not, we may find that we have a rich array of methods that all do something interesting, but none that do what we intended. It can be bewildering to read discussions about morphometric characters, because it sometimes appears that nearly every author has a different idea of the meaning of "character" (as well of "morphometric"). Until we can define "character" precisely, in terms just as comprehensible to mathematicians as to systematists, we will not make further progress towards a mathematical solution. We also need more than a definition of the term; we need to articulate more fully the process by which we find characters, in general. Most discussions of systematic methods focus on how to analyze the data, given the data matrix. Our problem is to get that matrix in the first place. One value of morphometric data is that we find them using mathematical methods, and these are necessarily explicit. By making our methods of character analysis explicit, just like our methods of phylogenetic inference, we will enhance the rigor of morphological systematics in general.

Software

None of the analyses used in this chapter require software beyond that introduced in earlier chapters. The CVA of the unstandardized data used **CVAGen** (introduced in Chapter 7); the analyses of the standardized data used **Standard6** to standardize specimens to a common developmental stage comparable across all taxa (introduced in Chapter 10); then **CVAGen** was used to analyze the standardized data. The PCA (of the standardized data) used **PCAGen** (also introduced in Chapter 7); to depict changes from one taxon to another within the plane of two PCs (rather than to depict the directions along the PCs) we placed a marker on each endpoint of the desired vector, using the **Place M1** and **Place M2** options, and depicted the difference between the two endpoints using the **Show M2-M1** option. To describe the differences between each species and another, we first combined the two files into one, replaced the centroid size by 0 for one species and 1 for the other, then regressed shape on those codes using **Regress6** (introduced in Chapter 10). To compare the regressions, we saved the vectors, using the option to save the deformation vector on the **File** pull-down menu, then input the vectors into **VecDisplay** (also introduced in Chapter 10).

References

Adams, D. C. and Rosenberg, M. S. (1998). Partial-warps, phylogeny, and ontogeny: a comment on Fink and Zelditch (1995). *Systematic Biology*, **47**, 167–172.

Archie, J. W. (1985). Methods for coding variable morphological features for numerical taxonomic analysis. *Systematic Zoology*, **34**, 236–345.

Chappill, J. A. (1989). Quantitative characters in phylogenetic analysis. *Cladistics*, **5**, 217–234.

Colless, D. H. (1980). Congruence between morphometric and allozyme data for *Menidia* species: a reappraisal. *Systematic Zoology*, **29**, 288–299.

Farris, J. S. (1990). Phenetics in camouflage. *Cladistics*, **6**, 91–100.

Felsenstein, J. (2002). Quantitative characters, phylogenies, and morphometrics. In *Morphology, Shape and Phylogeny* (N. MacLeod and P. L. Forey, eds) pp. 27–44. Taylor & Francis.

Fink, W. L. and Zelditch, M. L. (1995). Phylogenetic analysis of ontogenetic shape transformations: a reassessment of the piranha genus *Pygocentrus* (Teleostei). *Systematic Biology*, **44**, 343–360.

Gift, N. and Stevens, P. F. (1997). Vagaries in the delimitation of character states in quantitative variation – an experimental study. *Systematic Biology*, **46**, 112–125.

Goldman, N. (1988). Methods for discrete coding of morphological characters for numerical analysis. *Cladistics*, **4**, 59–71.

Rohlf, F. J. (1998). On applications of geometric morphometrics to studies of ontogeny and phylogeny. *Systematic Biology*, **47**, 147–158.

Rohlf, F. J. (2002). Geometric morphometrics and phylogeny. In *Morphology, Shape and Phylogeny* (N. MacLeod and P. L. Forey, eds) pp. 175–193. Taylor & Francis.

Simon, C. (1983). A new coding procedure for morphometric data with an example from periodical cicada wing veins. In *Numerical Taxonomy* (J. Felsenstein, ed.) pp. 378–382. Springer-Verlag.

Swiderski, D. L., Zelditch, M. L. and Fink, W. L. (1998). Why morphometrics isn't special: coding quantitative data for phylogenetic analysis. *Systematic Biology*, **47**, 508–519.

Thiele, K. (1993). The holy grail of the perfect character: the cladistic treatment of morphometric data. *Cladistics*, **9**, 275–304.

Zelditch, M. L., Fink, W. L. and Swiderski, D. L. (1995). Morphometrics, homology and phylogenetics: quantified characters as synapomorphies. *Systematic Biology*, **44**, 179–189.

PART IV

Last Things

15

Beyond two-dimensional configurations of landmarks

The focus of most of this book has been on the tools for comparing two-dimensional configurations of landmarks. However, many structures of interest to biologists are three-dimensional, or have few landmarks, or both. The skull of a marmot (Figure 15.1), like that of most mammals, is an example of “both.” The marmot skull is strongly curved anteroposteriorly and mediolaterally, making it highly three-dimensional (features on the same bone may be as far apart in the dorsoventral dimension as they are in the mediolateral or anteroposterior dimensions). In addition, the skull is composed of a small number of relatively large bony plates, so points that can be used as landmarks are sparsely distributed, occurring primarily at locations where at least three bones meet. In the first part of this chapter we examine methods that have been devised to analyze three-dimensional configurations of landmarks, and in the second we examine methods that have been devised to analyze curves and surfaces that lack landmarks. The methods discussed in both parts

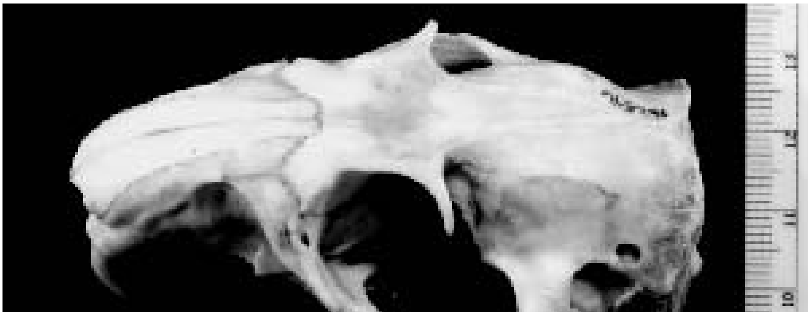


Figure 15.1 Skull of a yellow-bellied marmot (*Marmota flaviventris*), in dorsolateral view, illustrating curvature of the rostrum, braincase and zygomatic arch.

Geometric Morphometrics for Biologists
ISBN 0–12–77846–08

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

of this chapter have been presented and discussed elsewhere (Bookstein, 1996a, 1996b, 1997a, 1997b; Green, 1996; Sampson et al., 1996; Rohlf and Slice, 1990; Rohlf and Corti, 2000; Rohlf and Bookstein, 2003). Below, we discuss the general problems and the advantages and disadvantages of particular approaches. Because most of the software tools for applying these methods are in the early stages of development, we do not give detailed instructions for performing particular analyses.

Landmarks in three dimensions

Biological objects (organisms or parts of organisms) are inherently three-dimensional. Sometimes the third dimension can be ignored as a reasonable simplification – this is valid if the third dimension is unimportant relative to the other two. For example, distances between landmarks might be much smaller in the third dimension than in the other two, so that variation in this dimension contributes little to the description of overall shape variation (examples include the leaves of many plants, and the bodies of some fish). It is also possible that the third dimension simply is not relevant to the focus of a particular analysis. For example, studies of the shape of the lower jaw might focus on the proportions of lever arms associated with various muscles and teeth, and not be concerned with projections out of the plane of jaw action. However, there are also times when variation in the third dimension cannot be ignored without losing important information about the overall pattern of shape variation. As illustrated below, the analysis of landmarks digitized in three dimensions (X, Y and Z) only requires very simple extensions of the mathematics principles discussed in previous chapters of this book – *there are no new concepts*.

The principal obstacles to executing a complete three-dimensional study, from data collection to publication, are two problems that cannot be solved with mathematics: (1) the cost of the equipment needed to collect the data, and (2) the difficulty of illustrating three-dimensional shape differences on static two-dimensional media like the pages of this book. The solutions to these problems lie in the arts of grant-writing and illustration, so we do not address them in this book.

Some researchers have proposed clever alternatives to buying expensive equipment for three-dimensional digitizing (e.g. Spencer and Spencer, 1995; Fadda et al., 1997). Most of these alternatives involve collecting a series of overlapping images at different angles; some use mirrors, others rotate the specimen as if it were on a rotisserie. Landmarks are digitized in each two-dimensional image, and then the angle between two images is used to compute a set of three-dimensional coordinates from the two sets of two-dimensional coordinates for the landmarks present in both images. Similar triangulation schemes are used in some commercial digitizers.

The problem with a triangulation technique is that the uncertainty of the third coordinate (Z) is produced by a combination of three different potential errors: (1) error in digitizing the first two coordinates (X, Y) of the landmarks in the overlapping views; (2) error in measuring the angle between images; and (3) error in measuring the distance from the specimen to the camera lens. The compounding of these errors means that the computed third coordinate is likely to have a much larger error than the two directly observed coordinates; it also means that the error in the third coordinate is not independent of the errors in the first two coordinates. In commercially produced digitizers the confidence interval for the Z-coordinate is about twice as large as those for the X- and Y-coordinates, which is tolerable only because all of the confidence intervals are extremely small. In home-made equipment the difference between confidence intervals is likely to be larger, because the angle and distance cannot be controlled or measured with the same precision. These problems are exacerbated when more than two images are chained together to expand the coverage of the object. The progressive change in the orientation of the object (or camera) relative to a fixed coordinate system causes a progressive shift in the relative uncertainties of the coordinates. For example, the first pair of images might be used to infer the Z-coordinates of the landmarks from the X- and Y-coordinates, but the last pair of images might be used to estimate the Y-coordinates of the landmarks in those images from X- and Z-coordinates. Consequently, different landmarks will have different combinations of errors affecting estimation of coordinates on the same axis. For these reasons, we strongly recommend that you buy reliable commercial equipment designed for three-dimensional analysis if you think that such analyses will be necessary to fully describe shape variation in your data.

If the cost of equipment is absolutely out of reach, or you want to conduct a pilot project to investigate the need for such expenditure, we recommend using partial least squares (PLS). We have already presented a detailed discussion of PLS and illustrated its use to analyze the covariance of shape changes seen in different regions of the same lateral view of piranhas (Chapter 12). Rohlf and Corti (2000) present an example in which PLS is used to analyze the covariance of shape changes seen in dorsal and ventral views of skulls of the house mouse (*Mus musculus*). When the same procedure is used to analyze the covariance of shape changes seen in less divergent views (e.g. dorsal and lateral), the pictures of the correlated shape changes will be pictures of a three-dimensional change.

Superimposing configurations of landmarks

As in two-dimensional shape analyses, the first step after collecting the data is a generalized least squares Procrustes superimposition (GLS) to remove those differences between configurations that are not differences in shape. Differences in location, scale and orientation of three-dimensional configurations are removed by exactly the same operations that are used to superimpose two-dimensional configurations (translation, scaling and rotation); the only substantive difference is that superimposing three-dimensional configurations forces us to work with larger matrices. Just as the coordinates of K landmarks in two dimensions are represented as a $K \times 2$ matrix, the coordinates of K landmarks in three dimensions are represented as a $K \times 3$ matrix:

$$\mathbf{A} = \begin{bmatrix} X_1 & Y_1 & Z_1 \\ X_2 & Y_2 & Z_2 \\ X_3 & Y_3 & Z_3 \\ \vdots & \vdots & \vdots \\ X_K & Y_K & Z_K \end{bmatrix} \qquad (\mathbf{15.1})$$

The same operations will be performed on these matrices of three-dimensional landmarks as are performed on matrices of two-dimensional landmarks. In addition, most formulae used to perform these operations will have the same general form as those used in analyses of two-dimensional landmarks. The main consequence of having an extra column is that the computations are more tedious (especially for the programmer).

Below, we briefly review the operation (translation, scaling or rotation) that is performed to superimpose configurations of two-dimensional landmarks, then present the expanded form of the operation that applies to three-dimensional landmarks.

Centering

The first step of the superimposition is centering – translation of each configuration so that its centroid is located at the origin of the coordinate space. In two dimensions, the centroid (like the landmarks) has two original coordinates (X and Y). Each coordinate of the centroid is the average of the corresponding coordinates of the landmarks. Centering is accomplished by subtracting the coordinates of the centroid from the corresponding coordinates of each landmark, yielding new centroid coordinates of (0, 0). To superimpose three-dimensional configurations, we simply include the third (Z) coordinate in the same series of calculations. Thus, the original coordinates of the centroid are the averages of the corresponding coordinates of the landmarks:

$$\begin{aligned} X_C &= \frac{1}{K}(X_1 + X_2 + X_3 + \cdots + X_K) \\ Y_C &= \frac{1}{K}(Y_1 + Y_2 + Y_3 + \cdots + Y_K) \\ Z_C &= \frac{1}{K}(Z_1 + Z_2 + Z_3 + \cdots + Z_K) \end{aligned} \tag{15.2}$$

To center the configuration we again subtract the values of the centroid coordinates from the corresponding values of the landmark coordinates, but there are now three sets of subtractions:

$$\mathbf{A}_{\mathbf{centered}} = \begin{bmatrix} (X_1 - X_C) & (Y_1 - Y_C) & (Z_1 - Z_C) \\ (X_2 - X_C) & (Y_2 - Y_C) & (Z_2 - Z_C) \\ (X_3 - X_C) & (Y_3 - Y_C) & (Z_3 - Z_C) \\ \vdots & \vdots & \vdots \\ (X_K - X_C) & (Y_K - Y_C) & (Z_K - Z_C) \end{bmatrix} \tag{15.3}$$

When the operation is complete, the new coordinates of the centroid will be (0, 0, 0).

Scaling

The next step of the superimposition is scaling each centered configuration to unit centroid size. In three dimensions, as in two, centroid size is defined as the square root of the sum of the squared distances of the landmarks from the centroid. To compute the distance between two points in three dimensions, we simply include the difference in the Z-coordinates along with the differences in the X- and Y-coordinates:

$$D = \sqrt{(X_1 - X_C)^2 + (Y_1 - Y_C)^2 + (Z_1 - Z_C)^2} \tag{15.4}$$

Thus centroid size would be the square root of the sum of these squared distances. However, after centering the configuration ($X_C = Y_C = Z_C = 0$), we can simplify the computation to the square root of the sum of the squared coordinates:

$$CS = \sqrt{\sum_{i=1}^{K} X_i^2 + Y_i^2 + Z_i^2} \tag{15.5}$$

Then, to rescale the entire configuration to a centroid size of one, every coordinate in $\mathbf{A}_{\mathbf{centered}}$ is divided by CS.

Rotation

As in the two-dimensional case, we are going to rotate the three-dimensional configuration to the orientation that minimizes its partial Procrustes distance from a reference. At first glance this might seem to be a simple extension of what we have done before – we just include the third dimension in the calculations of the partial Procrustes distance and the angle that minimizes that distance. As explained in Chapter 4 (Equation 4.12), the partial Procrustes distance between two-dimensional configurations of K landmarks is:

$$D^2 = \sum_{j=1}^{k} (X_{Rj} - (X_{Tj}\cos\theta - Y_{Tj}\sin\theta))^2 + (Y_{Rj} - (X_{Tj}\sin\theta + Y_{Tj}\cos\theta))^2 \tag{15.6}$$

in which the coordinates of the target (X_{Tj}, Y_{Tj}) are related to the coordinates of the reference (X_{Rj}, Y_{Rj}) by the angle θ. On closer inspection the problem turns out to be a little more complex than just adding the difference in Z-coordinates to Equation 15.6, due to the fact that a three-dimensional object like the marmot skull can be rotated on three orthogonal axes (Figure 15.2). This means there are *three* angles involved in the computation of the partial Procrustes distance, and we have to solve for the particular combination of angles that minimizes that distance. Still, the solution remains conceptually simple, a singular value decomposition (SVD) of the matrix $\mathbf{X}_\mathbf{R}^t\mathbf{X}_\mathbf{T}$ in which $\mathbf{X}_\mathbf{R}$ and $\mathbf{X}_\mathbf{T}$ are the centered and scaled configuration matrices of the reference and target, respectively (Rohlf, 1990). As Rohlf points out, this is just one example of the general utility of SVD for finding the angular relationship between two matrices.

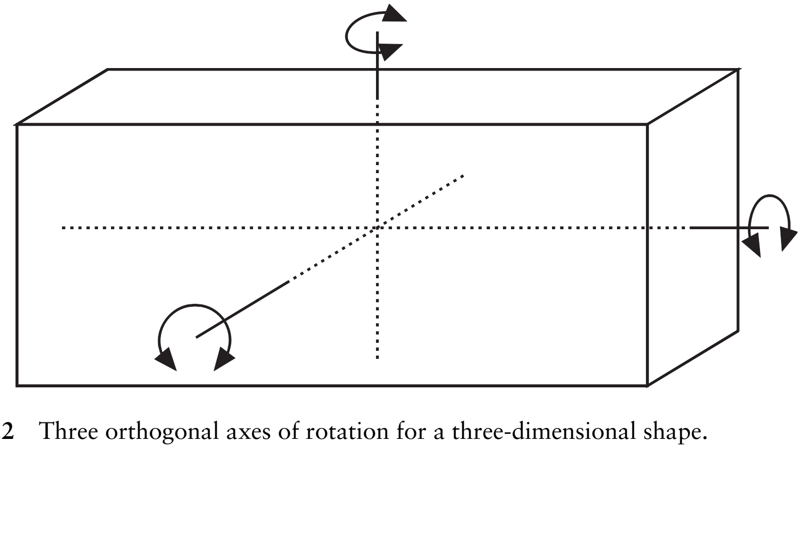


Figure 15.2 Three orthogonal axes of rotation for a three-dimensional shape.

The spaces of three-dimensional configurations

As discussed in Chapter 4, the set of all possible configurations of K landmarks with M coordinates is called a configuration space, and this space has $K \times M$ dimensions. Centering, scaling and rotating to a specific alignment all select subspaces with fewer dimensions. Because the same operations were used to select these subspaces, the same formulae can be used to determine their dimensions. Centering removes M dimensions because the centroid has M coordinates, so the space of centered coordinates has $KM-M$ dimensions, which is $3K-3$ when $M=3$. Scaling removes one dimension because we are still using centroid size, which is a one-dimensional scalar. Consequently, the space of centered and scaled configurations (pre-shapes) has $KM-M-1$ dimensions (Equation 4.9), which is $3K-4$ when $M=3$. Rotation to a standard orientation removes $M(M-1)/2$ dimensions (Equation 4.10), which are the number of orthogonal axes on which an M-dimensional configuration can be rotated. When $M=3$ there are three axes, and the space of aligned configurations (a shape space) has $3K-7$ dimensions.

When we impose on two-dimensional configurations of landmarks ($K \times 2$ matrices) the requirements of centering at the origin and scaling to unit centroid size, we generate a pre-shape space that has the form of the surface of a hypersphere with a radius of one, centered on the origin. When we impose the same requirements on three-dimensional configurations, we again get a pre-shape space that is the surface of a hypersphere with a radius of one, centered on the origin. Pre-shape spaces generated by these operations have the same general shape (differing only in the number of dimensions), regardless of the values of K and M.

The pre-shape spaces described above contain every possible rotation of every possible M-dimensional shape that can be formed of K landmarks. Each shape is represented by the set of all possible rotations of that shape, and the distance between shapes is the minimum distance between these sets. As mentioned in Chapter 4, the set of all possible rotations of a shape is called a *fiber*. This name seems apt when $M=2$; there is only one axis of rotation, so we can visualize a one-dimensional string lying in the pre-shape space. When $M=3$, calling the set of rotations a fiber may seem less appropriate because there are now three orthogonal axes of rotation, which does not fit our mental image of a one-dimensional string. However, the actual concept is still the same (the set of all possible rotations), and it is just as useful. Because different fibers represent different shapes, they do not intersect; and if they do not intersect, we can find the shortest distance between them. That distance is the difference between centered and rescaled configurations that is not due to the rotation of one relative to the other. Therefore, regardless of the values of K and M, the distance between two shapes in the same pre-shape space is the distance between two points on the surface of a hypersphere. Now that we are again on (relatively) familiar ground, we can see that we must solve for the rotation of the target that minimizes the partial Procrustes distance (the chord length), which can then be converted to the Procrustes distance (arc length) or the full Procrustes distance (the cosine of the angle subtended by the arc). Having a third set of coordinates makes the computation more tedious, but the procedure is the same.

The shape spaces we generate by the operations described above are hyperspheres tangent to their respective pre-shape spaces at the location of the reference shape. If centroid size is fixed at one, the space is the surface of a hypersphere of radius one. If centroid size is scaled to the cosine of the Procrustes distance, the space is Kendall's shape space, the surface of a hypersphere of radius one-half.

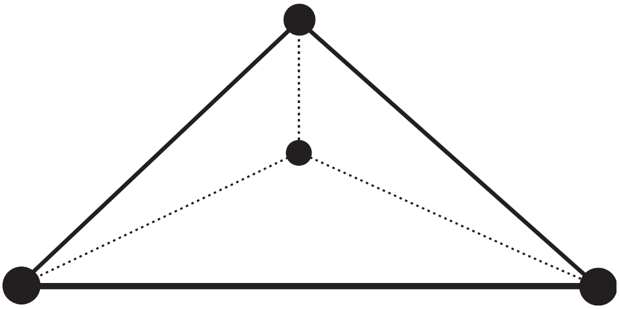


Figure 15.3 The simplest three-dimensional configuration of landmarks: a tetrahedron.

Decomposing the deformation

As in the two-dimensional case, the difference between three-dimensional configurations of landmarks can be described as a deformation of one shape (reference) into the other (target). This deformation can be decomposed into uniform and non-uniform parts (or affine and non-affine). The non-uniform part can be further decomposed into $3(K-4)$ independent components. The uniform part can be further decomposed into twelve independent components; but only five of these change shape.

The numbers of uniform and non-uniform components can be explained if we consider the possible deformations of the simplest three-dimensional shape, a tetrahedron of four landmarks (Figure 15.3). All deformations of a tetrahedron, like all deformations of a triangle, must be uniform; only when a fifth point is added can we detect non-uniform transformations (i.e. transformations that differ between regions of the tetrahedron). With just four landmarks a deformation can have twelve components, all of them uniform. Seven of the uniform components do not change shape – they are the ones removed by superimposition – which leaves five uniform components that do change shape. With each additional landmark beyond the fourth, there are three possible non-uniform components of deformation (because there are three directions in which that point might move relative to the others), hence $3(K-4)$.

The components of the non-uniform part of a three-dimensional deformation are defined in nearly the same terms as the components of the non-uniform part of a two-dimensional deformation. Again, we use the thin-plate spline model to describe the deformation at any point in space as f_X, f_Y and f_Z, which describe the X-, Y- and Z-components of the deformation:

$$f_X(X,Y,Z) = A_{X1} + A_{XX}X + A_{XY}Y + A_{XZ}Z + \sum_{i=1}^{K} W_{Xi}U(X-X_i, Y-Y_i, Z-Z_i)$$

$$f_Y(X,Y,Z) = A_{Y1} + A_{YX}X + A_{YY}Y + A_{YZ}Z + \sum_{i=1}^{K} W_{Yi}U(X-X_i, Y-Y_i, Z-Z_i)$$

$$f_Z(X,Y,Z) = A_{Z1} + A_{ZX}X + A_{ZY}Y + A_{ZZ}Z + \sum_{i=1}^{K} W_{Zi}U(X-X_i, Y-Y_i, Z-Z_i) \tag{15.7}$$

where $U(X - X_i, Y - Y_i, Z - Z_i)$ is a function of the interlandmark distances given by:

$$R_i = \sqrt{(X - X_i)^2 + (Y - Y_i)^2 + (Z - Z_i)^2} \tag{15.8}$$

Again, we have more columns to accommodate the third dimension. The more substantive difference is that $U = R$, in contrast to the two-dimensional case in which $U = R^2 \ln R^2$. As in the two-dimensional case (see Chapter 6), the next steps are to solve for the spline coefficients (the values of A and W) and the eigenvectors of the bending energy matrix (the partial warps).

In both the two-dimensional and three-dimensional cases, the thin-plate spline is only used to solve for the non-uniform components of the deformation; a different approach is taken to solve for the uniform components. Bookstein (1996b) shows that the approach he developed to construct a pair of basis vectors for the uniform part of a two-dimensional deformation can be extended to the three-dimensional case. This approach yields three pairs of vectors describing shear and compression/dilation in each of the three two-dimensional planes (XY, YZ and XZ). But remember, there are only five possible shape variables for the uniform part; therefore, the six vectors are not all completely independent. In fact, the problem lies in the three compression/dilation vectors; these three vectors actually describe a two-dimensional space. Bookstein suggests several methods to rectify this problem by constructing an orthonormal basis for this subspace (the current IMP software uses the Gram-Schmidt technique – *cf.* Axler 1996). These two vectors, combined with the three shear vectors, provide an orthonormal basis for the entire uniform subspace. More recently, Rohlf and Bookstein (2003) have presented two other methods, both using an SVD to compute an orthonormal basis for the entire uniform subspace (without dividing it into shear and compression/dilation subspaces). The methods differ in how they extract the uniform variation from the total variation. In one, a technique used to compute residuals from a regression is used to compute the uniform component as the residuals from the non-uniform (as the difference between the total deformation and the non-uniform part). In the other, a technique used by Rohlf and Slice (1990) to compute the uniform component directly from superimposed two-dimensional coordinates is extended to three-dimensional coordinates. The new methods differ from that proposed by Bookstein (1996b) only in the simplicity of the algorithms; all lead to the same conclusions regarding the differences among populations of shapes.

The result of the completed decomposition (of both uniform and non-uniform components) is an orthonormal basis for the Euclidean space that is tangent to the shape space at the location of the reference shape. Every configuration of landmarks in a data set can be described as a deformation of the reference shape; and that deformation is represented by the full set of scores on the five uniform components and $3(K - 4)$ non-uniform components. These scores preserve Procrustes distances and express shape differences as scores on the same number of orthogonal axes as there are dimensions of the shape space (which is equal to the number of statistical degrees of freedom). Consequently, these scores can be used in standard multivariate analyses.

Ordinations and statistics

All of the analytic techniques discussed in Chapters 7–12 can be performed on data from three-dimensional landmarks. This includes analyses performed on the partial Procrustes distances of individual specimens from a reference shape (e.g. Goodall's *F*-test) and analyses performed on the full set of scores over all of the axes of the tangent space (e.g. principal components analysis). None of these analyses is materially altered by the use of three-dimensional coordinates. The transition from univariate to multivariate requires changes in analytic tools, but after this transition has been made, no further methodological changes are required to accommodate further increases in the number of variables. The crucial thing to remember is that an analysis of three-dimensional shapes will have more shape variables than an analysis of two-dimensional shapes with the same number of landmarks. You will need larger sample sizes to perform comparable tests.

Illustrating shape differences

Although the mathematics of comparing three-dimensional shapes is well established, the difficulties of illustrating those comparisons on static two-dimensional media (like the pages of this book) have not been resolved to our satisfaction. The problem is not one of illustrating a single, solid three-dimensional object – a skilled artist or photographer can produce very convincing two-dimensional images. Instead, the problem is that the process of creating the illusion of three dimensions necessarily entails omission or distortion of some information. For example, lengths and angles are distorted (by foreshortening) to create the illusion of depth, and illustration of a fully rendered surface precludes illustration of internal details or the other side of the object.

The problems of illustrating three-dimensional objects are exacerbated when the objects are superimposed. If the surfaces are rendered, they will interpenetrate (i.e. only parts of each will be shown – the parts that are "in front"); consequently, the viewer does not see all of either object, and so cannot fully appreciate the shape difference. To illustrate differences at all landmarks, the images of the objects must be simplified in some way. Figure 15.4A shows a photograph of a skull with a selection of landmarks. In Figure 15.4B, the landmarks are projected onto the plane of the page and are connected by a wireframe, a set of straight lines chosen to approximate salient features of the skull. With judicious selection of line weights or colors, wireframes can be used to illustrate two superimposed configurations of landmarks (Figure 15.4C), but even with the most lurid color scheme a wireframe illustration of more than two configurations would be too confusing to be useful. Furthermore, a single view of the wireframe cannot convey differences in depth; at least one other view is needed, as in Figure 15.4D.

Illustrating a three-dimensional deformation on static two-dimensional media is even more difficult than illustrating two superimposed objects. A variety of devices can be used to draw the three-dimensional spline interpolation (e.g. the deformed grid or a series of arrows on the nodes of an undeformed grid), but all have the same limitation – any diagram displaying enough data to be useful contains too many data to be interpretable. The simplest useful approach appears to be placing arrows on the wireframe of the reference form (in two views) to indicate directions of relative landmark displacement (Figure 15.5). Although this approach requires viewers to do the interpolation in their heads it does

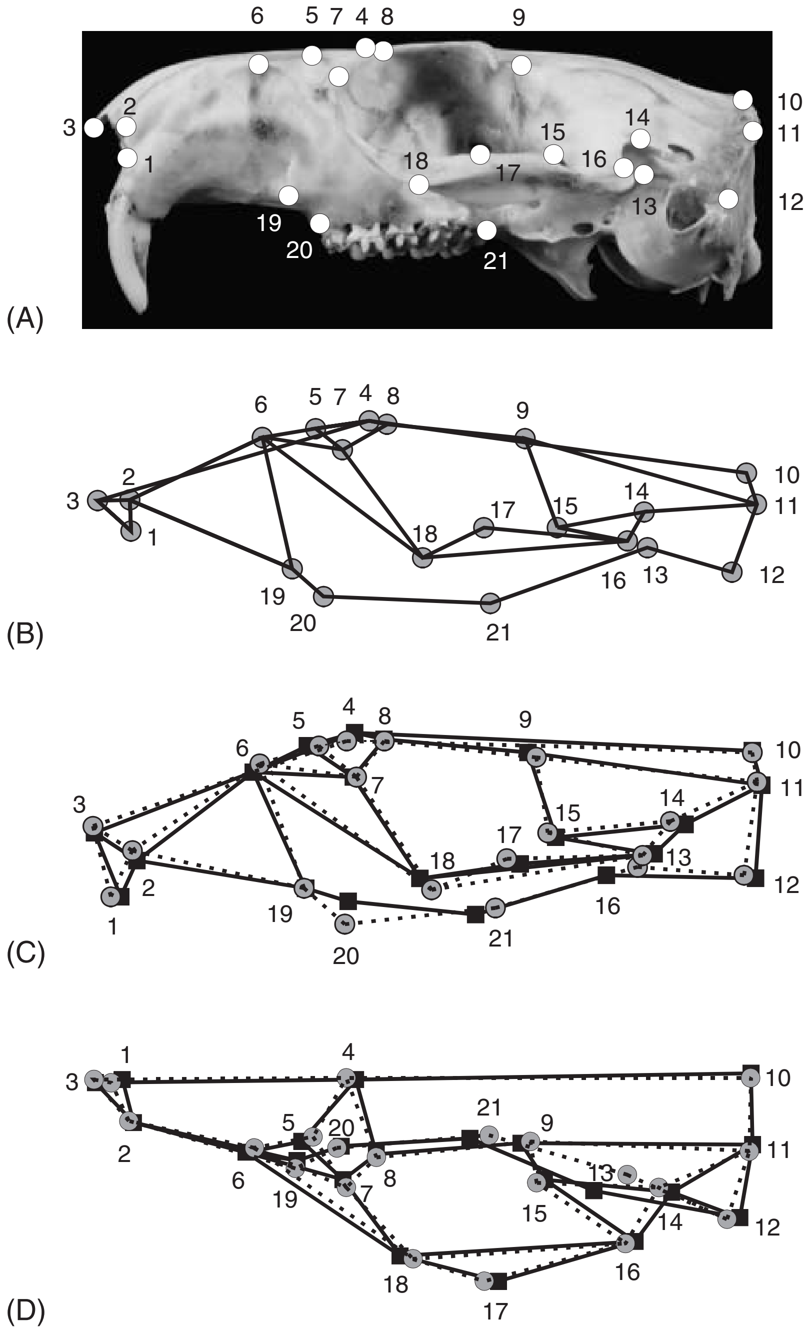


Figure 15.4 Illustration of three-dimensional shapes using wireframes: (A) lateral view of a marmot skull, with some possible landmarks; (B) the same landmarks in lateral view connected by a wireframe; (C) two superimposed configurations and their wireframes, in lateral view; (D) the same two configurations in dorsal view.

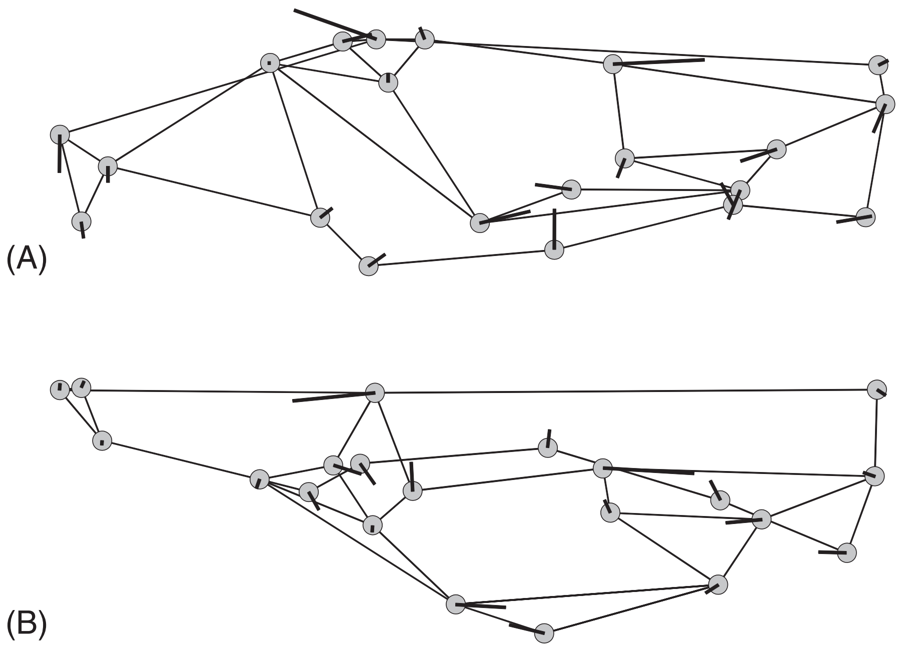


Figure 15.5 Illustration of a deformation of a three-dimensional shape, using vectors at landmarks connected by a wireframe: (A) lateral view; (B) dorsal view.

convey the localization of shape change, which may be the most important information that can be gleaned from the spline.

Curves without landmarks

Many features of interest to biologists are curves that lack landmarks, such as short segments of edges or ridges between landmarks, or outlines encompassing whole organisms. (Similar problems and methods apply to comparisons of three-dimensional surfaces like the cranial vault, but for the sake of simplicity we return here to two-dimensional shapes.) The lack of landmarks is a problem because the entire mathematical framework of geometric morphometrics rests on the comparability of landmarks from specimen to specimen. Without comparable landmarks there is no justification for the Procrustes distance metric, or for the superimpositions and shape spaces founded on that distance. Without landmarks, we cannot apply the mathematical theory of shape spaces.

There are methods of analysis that do not require comparable points along the curve of interest (Rohlf and Archie, 1984; Ferson et al., 1985; Lohman and Schweitzer, 1990; MacLeod and Rose, 1993). In these approaches, points digitized along the curve serve only as local estimates of the location of the curve. A function is fitted to the digitized points, producing a set of coefficients representing the shape of that curve, and these coefficients are used as variables in any subsequent comparative analysis. Although this appears to be a clever way to circumvent the lack of landmarks, it is important to remember that the configurations represented by the coefficients are not shapes as defined by Kendall.

The configurations may have been aligned and scaled by precise and rigorous methods (*cf.* Ferson et al., 1985), but a Procrustes superimposition is precluded by the lack of corresponding points, and perhaps by the nature of the curve-fitting algorithm. Consequently, the spaces occupied by these configurations are not the spaces covered by the theory of geometric morphometrics. In addition, the descriptions of curves produced by these methods are incommensurate with descriptions of shapes based on configurations of landmarks. This means that the coefficients of the curve-fitting function and shape variables computed from Procrustes superimposed landmarks cannot be combined into a single shape analysis. The best that can be done is to use PLS to look for correlations between the two sets of data.

Defining comparable points along a curve

To analyze curves in the same analytical framework as landmarks, and especially in a study simultaneously with landmarks, we need a way to identify points on the curve that can be treated as though they were landmarks. This means that we need to supply criteria for recognizing or selecting points that are not specific to the region immediately surrounding the curve (otherwise we would have landmarks). One such criterion would be to select points that are at equal intervals along the curve (e.g. 10% of the length of the curve). In Bookstein's (1991) typology of landmarks, Type 3 encompasses points that are defined by these sorts of extrinsic criteria. Subsequently, the term *semilandmarks* was used by Bookstein (1997a, 1997b) to refer to a series of points that are located along a curve using these kinds of criteria to define their positions along the curve. As Bookstein points out in all three cited references, semilandmarks and similarly defined points do not have as many degrees of freedom as the number of coordinates describing their location. The reduced degrees of freedom are a consequence of defining the semilandmark in terms of its position relative to other features. For example, a semilandmark defined to be halfway between the ends of a curve that connects two landmarks can only tell us one thing about the curve that we could not have inferred from the coordinates of the landmarks: the bowing of the curve (i.e. the amplitude of its deviation from a straight line). Consequently, the semilandmark has only one degree of freedom even though it has two coordinates.

There are a number of ways to delimit segments of the curve under analysis, and Figure 15.6 shows three possibilities: by increments along the length of the curve (Figure 15.6A); by increments along the length of a chord connecting the ends of the curve (Figure 15.6B); or by increments of an angle subtended by the curve (Figure 15.6C). In addition, the increments might all be equal, or they might vary in a way that reflects the complexity of the curve they are sampling. The combination of choices that produces the most satisfactory sampling of the curve will depend on the geometry of the curve and its relationship to other features represented by landmarks. This is not just a matter of aesthetics; the results of subsequent analyses can depend on the sampling of the curve, just as the results of landmark-based studies can depend on the selection of landmarks.

Superimposing configurations with semilandmarks

Choosing a general approach to selecting semilandmarks is only one of the decisions that must be made. Because semilandmarks are not locally defined and have reduced degrees of freedom, users must also decide how to adjust several steps in the analysis of shape

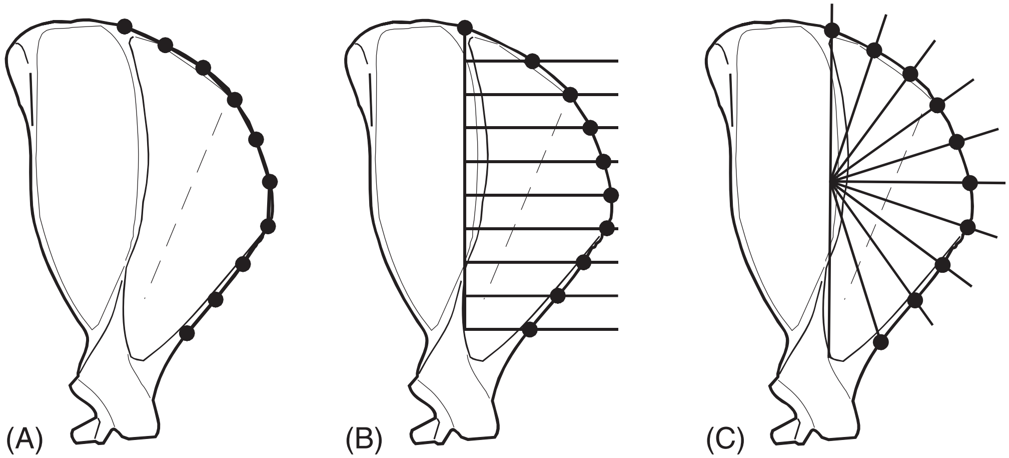


Figure 15.6 Some general approaches to selecting semilandmarks on a curve, illustrated on the anterior edge of a tree squirrel scapula: (A) increments of curve length; (B) increments of the chord; (C) increments of an angle subtended by the curve.

differences. In this section we discuss possible adjustments to the process of computing GLS Procrustes superimpositions of configurations of landmarks and semilandmarks (including the option of making no adjustment); in the next section we discuss possible adjustments to methods of ordination and statistical analysis. All of these issues are illustrated using an artificial data set designed to represent a hypothetical pattern of shape variation in the scapula of a tree squirrel (Figure 15.7).

No adjustment or weighting

In this simplistic approach, the semilandmarks are treated as equivalent to landmarks for the purpose of computing superimpositions. (Treating the two as equivalent at this point is independent of any differential weighting that might be applied subsequently in ordinations or statistical tests, and does not preclude such weighting.) In this superimposition, the semilandmarks of each specimen remain in the same positions relative to the landmarks of that specimen. Figure 15.8A shows a GLS superimposition of the artificial scapula data using this simplistic approach. The first principal component of variation (PC1) for these data is primarily a change in the curvature of the anterior edge (Figure 15.8B): as the anterodorsal corner becomes more squared, the ventral end becomes narrower. This change in the anterior edge is correlated with a general change in the relative height of the scapula and a rather small change in the anteroposterior lengths of the acromion and metacromion (landmarks 2–5).

Because landmarks and semilandmarks are treated as equivalent, the configurations that come out of the superimposition have the same shapes as the configurations that went into the superimposition (which is not true of some alternative methods). In addition, the configurations that come out represent shapes in Kendall's shape space. At first glance these appear to be clear advantages over any possible alternative method, but they could also be considered disadvantages because they reflect a disregard for the fact that semilandmarks are *not* equivalent to landmarks. Because landmarks and semilandmarks are treated as

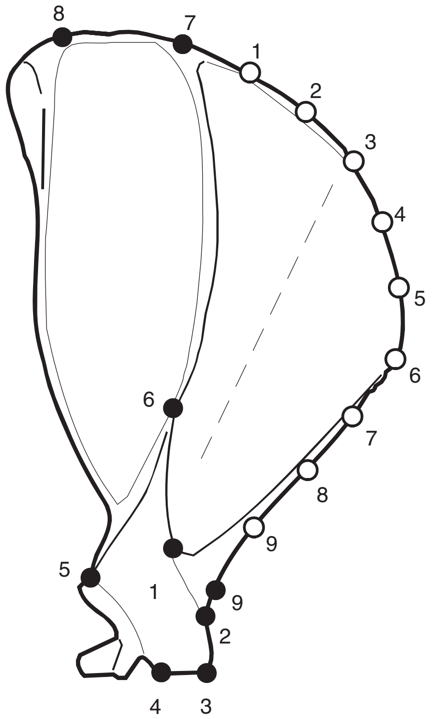


Figure 15.7 Points digitized for an analysis of squirrel scapula shape. Landmarks are indicated by black circles and semilandmarks by white circles. Semilandmarks are digitized in equal angular increments, as shown in Figure 15.6C.

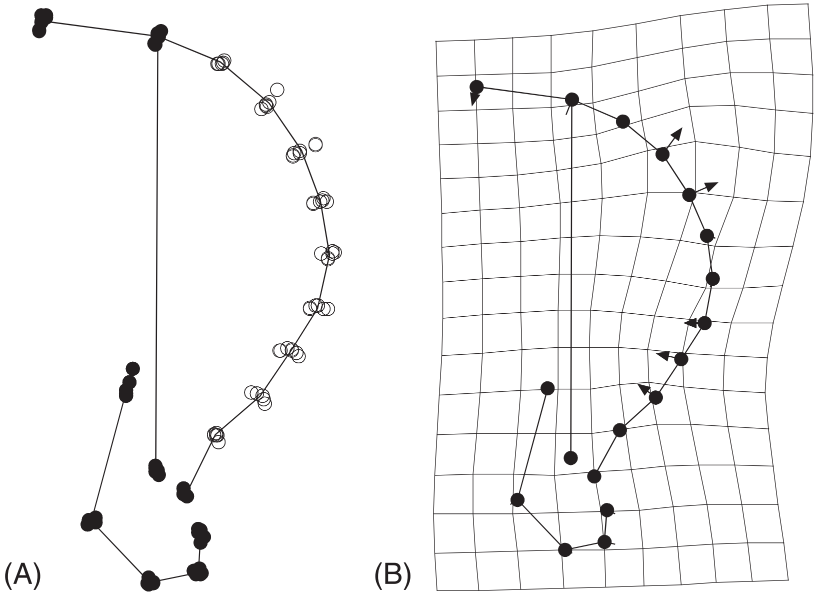


Figure 15.8 Analysis of scapula shape variation with semilandmarks treated as equivalent to landmarks: (A) GLS superimposition of all specimens; (B) shape change associated with positive scores on PC1.

equivalent, the semilandmarks have more influence on the result than is justified by the number of degrees of freedom they represent. In our example, each semilandmark on the anterodorsal corner plays as large a role in determining the optimal alignment as the landmark on the posterodorsal corner, which seems inappropriate because the semilandmarks represent less information than the landmarks. This raises the question: would the changes in curvature of the anterior edge be regarded as the dominant feature of variation if the semilandmarks had less influence on the result?

Differential weighting of landmarks and semilandmarks

One approach to reducing the influence of semilandmarks on the superimposition is to downweight them – i.e. construct a weighted Procrustes distance and use it as the criterion for optimal superimposition. This superimposition would be computed using essentially the same mathematics as the conventional superimposition, but applying scalar multiples to reduce the influence of semilandmarks on the computations of the superimposition and the shape differences.

This approach has the advantage of recognizing that there is a difference between landmarks and semilandmarks, but it also has the disadvantage that the steps of the superimposition do not lead to configurations in Kendall's shape space. At first this might not seem like such a bad thing. We do not want to ignore the semilandmarks, and we do not want to ignore the difference between landmarks and semilandmarks. This would suggest that we do not want the Kendall's shape space for just the landmarks, and we do not want the Kendall's shape space for configurations of landmarks plus semilandmarks. The disadvantage lies in the consequences of not having configurations in Kendall's shape space – namely a lack of information about the shape of the space that the configurations occupy or the properties of the distance metric for that space. An additional implication of these uncertainties is that statistical analysis cannot employ conventional parametric models; resampling-based methods must be used.

Another disadvantage of this approach is the lack of clear criteria for determining the appropriate weighting of semilandmarks. We can be fairly confident that semilandmarks represent less information than landmarks, especially if we know the rule used to select the semilandmarks. Rules like those described earlier imply that each semilandmark has only one degree of freedom. This may serve as a reasonable estimate of the information represented by each semilandmark, but it should be regarded as an estimate with a high degree of uncertainty. If curvature is simple (few reversals, as in the squirrel scapula example), a small subset of semilandmarks may be sufficient to characterize the curve, which implies that most of the semilandmarks do not contribute additional information. On the other hand, if the curvature has a high but consistent complexity (like an oak leaf with a fixed number of lobes), the information provided by landmarks and some semilandmarks might be nearly equivalent. Should you decide to take this approach, we recommend you try several different weighting schemes to insure that your conclusions are robust (i.e. not dependent on a particular scheme).

Sliding semilandmarks to minimize bending energy

In this approach, developed by Green (1996) and Bookstein (1997a), the first step is a conventional Procrustes superimposition (treating landmarks and semilandmarks as

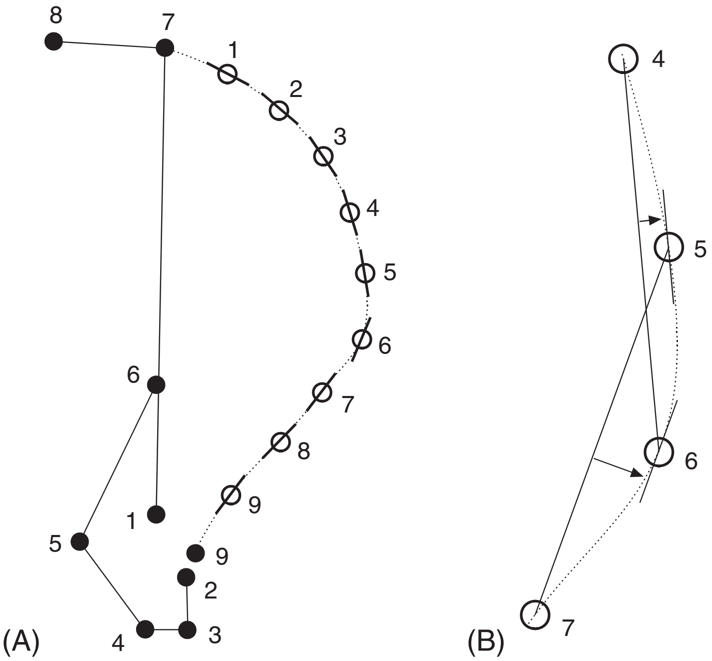


Figure 15.9 Tangents to the anterior edge at the locations of semilandmarks: (A) idealized estimates based on the curvature of the edge at the semilandmark; (B) enlarged view of part of the edge, showing estimation of the tangents at the semilandmarks from segments connecting adjacent semilandmarks. The line through the semilandmark is parallel to the line connecting the adjacent semilandmarks.

equivalent) to compute a mean configuration and align the targets to it. This is followed by moving the semilandmarks of each target to minimize the bending energy of the thin-plate spline describing the deformation of the reference to that target. The semilandmarks are not free to move in any direction; each is confined to “slide” along the line tangent to the curve at that semilandmark (Figure 15.9A). The shape of the curve is not actually known, so the tangent is estimated as the line parallel to the segment connecting adjacent landmarks or semilandmarks (Figure 15.9B). After sliding, the superimposition is recomputed; if the new mean configuration differs from the previous mean, the sliding and superimposition are reiterated until they converge on a solution. The justification for this sliding technique is that differences in relative positions of semilandmarks along the curve cannot be informative because this spacing was defined arbitrarily (that is, extrinsically). Thus, sliding to minimize the bending energy of the deformation adjusts the spacing of the semilandmarks to minimize the implication that there are shape changes due to differences in that spacing.

Figure 15.10A shows the same data set used earlier, superimposed after sliding to minimize bending energy. Several semilandmarks have ellipses of variation that imply displacements along the anterior edge, particularly the ones in the ventral half. PC1 indicates that these semilandmarks undergo correlated displacements toward the dorsal end as the anterodorsal corner is squared out (Figure 15.10B). Thus there is little change in

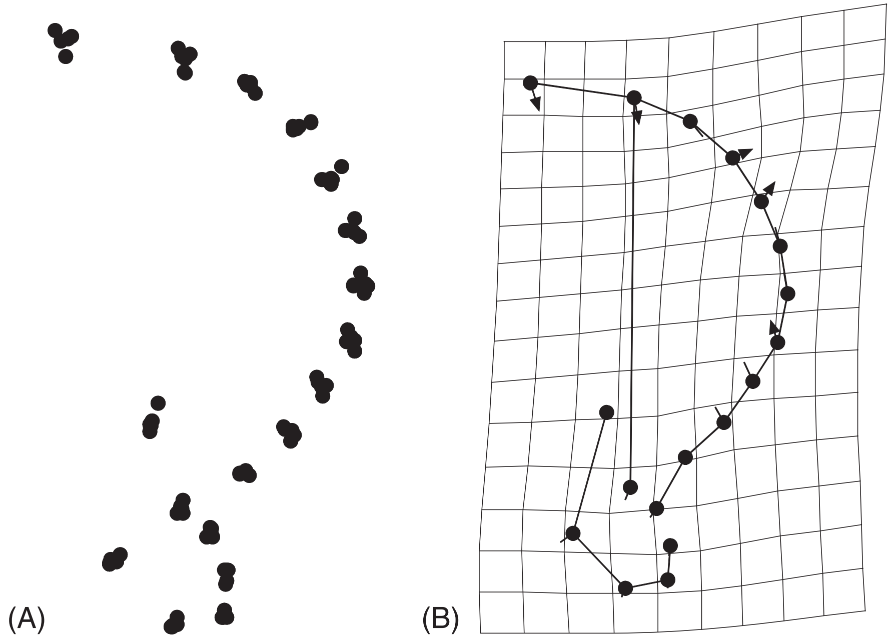


Figure 15.10 Analysis of scapula shape variation after sliding semilandmarks to minimize bending energy: (A) superimposition of all specimens; (B) shape change associated with positive scores on PC1.

the positions of these semilandmarks *relative to each other*, and therefore little localized change along the anterior edge. Most of the localized change in the anterior edge occurs at the corner. As before, the change in shape of the anterior edge is inferred to be the dominant component of shape change. Displacements of the landmarks are generally slight; the exceptions are the ventral displacements of the most dorsal landmarks, which are involved in the general flattening of the dorsal edge and squaring of the anterodorsal corner.

Compared to both options discussed above, sliding semilandmarks to minimize bending energy has the advantage that it does not ignore the difference between landmarks and semilandmarks. Compared to weighting, this sliding technique has the further advantage of having a clear criterion for the optimal superimposition. The principal disadvantage of this approach is that the semilandmarks are in new positions relative to the landmarks and the other semilandmarks. However, this may not be the devastating flaw that it seems to be. The underlying premise of sliding is that semilandmarks are not equal to landmarks. As pointed out above, semilandmarks do not represent the same amount of independent information as landmarks because semilandmarks are constrained to lie along the curve at arbitrary intervals. Put another way, if moving semilandmarks does not alter the information about the shape of the curve, then the configuration of landmarks and semilandmarks after sliding might be considered to have the same shape as the configuration before sliding.

Sliding semilandmarks to perpendicular alignment on the reference

This approach, suggested by Sampson et al. (1996), differs from the previous one only in the criterion used to determine how far the semilandmarks slide. The normal to the

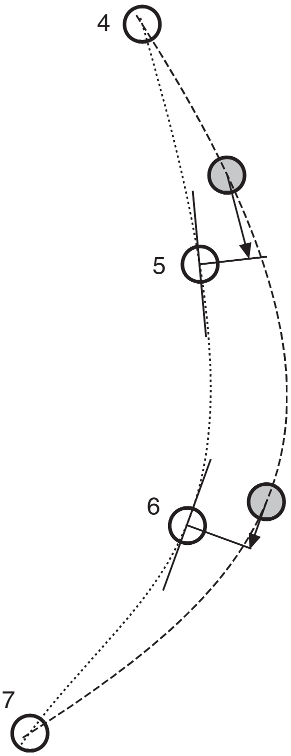


Figure 15.11 Perpendicular alignment sliding. This enlarged view of part of the anterior edge of the scapula shows the semilandmarks for a reference (white circles) and a target (gray circles) before sliding. The semilandmarks of the target are slid toward the lines that are perpendicular to the edge at the corresponding semilandmark of the reference.

curve (perpendicular to the tangent) is estimated for each semilandmark of the reference configuration (Figure 15.11). Each semilandmark of the target slides along its tangent to align with the perpendicular of the corresponding semilandmark of the reference. As in the other method, sliding is justified by the argument that positions of semilandmarks along the curve are uninformative because they are arbitrary, but here the conclusion from that argument is that semilandmarks can be informative only about the bowing of the curve. Sliding to perpendicular alignment extracts the information about bowing (displacement perpendicular to the tangent) and minimizes the inference that the semilandmarks moved along the curve (along the tangent).

Compared to the first two options discussed above (no adjustment and weighting), sliding to perpendicular alignment has the same advantages and disadvantages as sliding to minimize bending energy. In both cases, the final configurations do not have the same shapes as the original configurations, but they are in Kendall's shape space. Both sliding methods recognize the difference between landmarks and semilandmarks, and both imply that some displacements of semilandmarks do not constitute changes in shape. Because the two sliding methods have not been subject to extensive evaluation, it is not clear if one is better than the other.

Ordinations and statistical analyses

Regardless of the methods of selecting and superimposing semilandmarks, a curve will be represented by a large number of semilandmarks better than it will be by a smaller number (at least until the density of semilandmarks exceeds the resolution permitted by digitizing error). If either sliding method is used, it will also be improved by increasing the number of landmarks because that will tend to produce more accurate estimates of the tangents at the semilandmarks. However, increasing the number of semilandmarks on a curve means it will play a larger role in superimposition and comparison. In fact, it is easy to imagine cases in which the curve will exert a stronger influence on the results than all the features represented by landmarks. In addition, increasing the number of semilandmarks increases the number of coordinates in the data set, the number of spline coefficients computed from those coordinates, and the discrepancy between these numbers and the number of degrees of freedom.

One possible way to address the unbalanced representation of the curve and the other features is to reduce the number of points representing the curve. This will also reduce discrepancy in the number of degrees of freedom. Unfortunately, it is not clear how far to reduce the number of semilandmarks or what criteria might be used to select which semilandmarks to eliminate. In any event, the attendant reduction of the accuracy of the reconstruction makes the idea of cutting semilandmarks unpalatable. Making the density of semilandmarks inversely proportional to the general complexity of the curve can moderate the loss of accuracy. Tangents estimated from chords connecting adjacent points will deviate most from the true tangent in regions where the curvature changes most rapidly; closer spacing of semilandmarks will more accurately estimate the spatial extent of a sharp bend (Figure 15.12). Changes in the shape of the curve may include changes in the locations of sharp bends (as in the squirrel scapulae), so variability of the curvature should also be a consideration in determining the number of semilandmarks to digitize.

Another possible solution to unbalanced representation of curves and other features is downweighting the semilandmarks, as discussed above. This approach would reduce the influence of semilandmarks on computation of distances between shapes, so they would play a smaller role in testing for patterns of shape variation within groups, or differences in mean shape between groups. There would still be more coordinates and shape variables than degrees of freedom (for K landmarks and H semilandmarks there would still be $2(K+H)$ coordinates and $2(K+H)-4$ shape variables, but only about $2K-4+H$ degrees of freedom – depending on how degrees of freedom are attributed to semilandmarks).

A compromise between removing some semilandmarks and downweighting them is to use a large number of semilandmarks in the superimposition and sliding, then cull semilandmarks before performing ordinations or statistical tests. Having a large number of semilandmarks at the beginning maintains the accuracy of tangents and sliding; culling several of them after the final superimposition reduces discrepancies in weighting and degrees of freedom. Of course, the problem of choosing which points to cull is the same as the problem of selecting which points to exclude from the beginning.

Regardless of which approach you select, it is always possible to perform a sensitivity analysis to determine the effects of these choices. By comparing results obtained with different combinations of semilandmarks, you can determine whether the choices influence the conclusion. At the very least, this would provide reassuring information about the

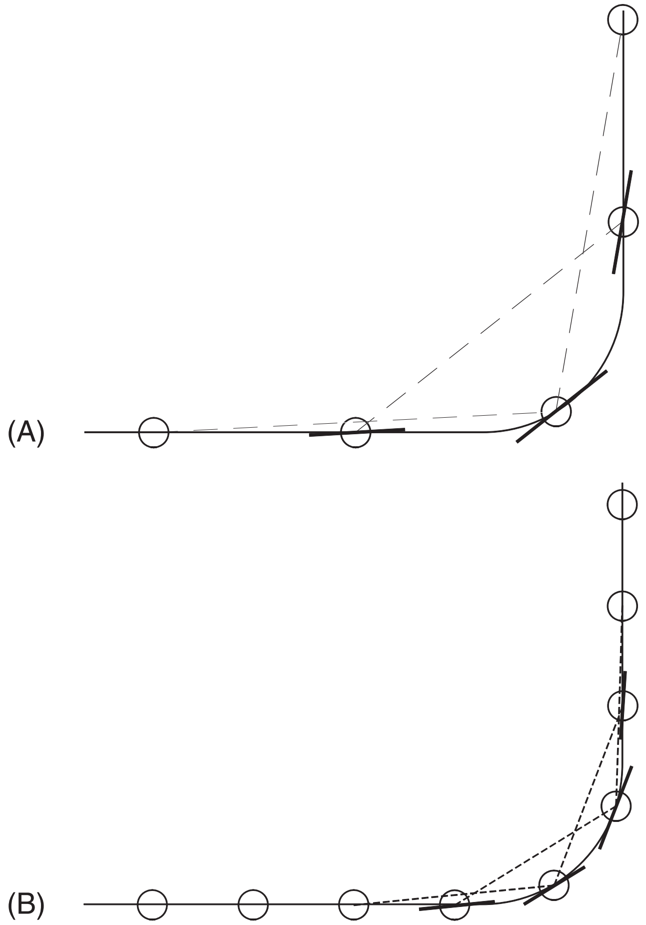


Figure 15.12 Effect of semilandmark density on estimation of tangents: (A) wide spacing; (B) closer spacing.

robustness of the conclusion. In addition, it may also reveal which parts of the curve are most relevant to the conclusion (for example, which semilandmarks covary with each other but are independent of the major pattern of landmark covariation).

Whatever method you use to evaluate or moderate the influence of semilandmarks on the results of an analysis, there remains the problem of determining the correct number of statistical degrees of freedom for conventional statistical tests and making the appropriate adjustments in sample size and criterion for rejecting the null hypothesis (α-level). If you digitize more points, you will have more coordinates and will need more specimens to maintain the same α-level. If you need to digitize a lot of semilandmarks to characterize a curve, you may need an unobtainable number of specimens just to have more individuals than coordinates (which is required for a conventional statistical analysis). Even if you can correct the p-value to reflect the correct number of degrees of freedom for a given combination of landmarks and semilandmarks, it may still be difficult to acquire enough specimens to permit as many semilandmarks as you might like. Furthermore, determining the correct number of degrees of freedom may not be as simple as adding the number of semilandmarks to twice the number of landmarks and subtracting the four degrees of freedom lost in the Procrustes superimposition; sliding may further reduce the independence of semilandmarks. Fortunately, all of these problems can be avoided by using the resampling methods discussed in Chapter 8. More specimens would still be better (as always), but the criteria for "enough" would no longer depend on the number of coordinates; more importantly, the risk of an erroneous conclusion due to a mistaken estimate of the number of degrees of freedom would be eliminated.

Summary

Analysis of three-dimensional configurations of landmarks can be performed within the same theoretical framework as analysis of two-dimensional configurations. The mathematics has been worked out and published. There is only a limited amount of readily available software, mostly due to lack of demand. The main obstacle to analyzing three-dimensional shapes is the difficulty of displaying the results on static two-dimensional media like journal and book pages.

Some techniques for analyzing objects without landmarks lie completely outside the theoretical framework used to analyze configurations of landmarks. These techniques do not employ the definitions of size and shape used in landmark-based analyses, so the configurations described by these methods do not lie in the shape space described by that theory. In addition, these descriptions of curves are incommensurate with shape variables computed from Procrustes superimposed landmarks. Therefore, PLS can be used to test for correlations, but the two kinds of description cannot be combined into a single shape analysis.

Techniques for analyzing features without landmarks in the same theoretical framework as configurations of landmarks are not fully developed. Methods have been developed to use extrinsic criteria to locate comparable points along a curve (semilandmarks) and then analyze the semilandmarks simultaneously with landmarks in the theoretical framework for landmarks. Criteria for choosing among these methods are not well established, so we recommend that prospective users try multiple methods and several options within those methods (including different combinations of points on the curve). The number of degrees of freedom associated with the semilandmarks is also not well established, although it is clear that each semilandmark can contribute no more than one additional degree of freedom (in contrast to two additional degrees of freedom for a landmark). Consequently, the number of degrees of freedom contributed by semilandmarks cannot be more than the number of semilandmarks. Because the number of degrees of freedom is unclear, but definitely much less than the number of coordinates, we recommend that resampling methods be used in statistical tests. This will alleviate demands for excessively large samples and avoid errors due to mistaken estimates of the number of degrees of freedom.

Software

As mentioned above, software tools for performing the analyses discussed in this chapter are not as well developed as the other software tools we have presented. Accordingly, we briefly review some of the tools that are available for three-dimensional configurations of landmarks and for semilandmarks, but do not discuss their operational details in depth.

Three-dimensional configurations of landmarks

The IMP package currently offers three programs for analysis of three-dimensional configurations of landmarks. The program **simple3D** takes raw data from a TPS-like format (one landmark per row), performs Procrustes superimposition, computes centroid size and produces output files in a more conventional datafile format (one specimen per row – $X_1, Y_1, Z_1 \ldots CS$); it can also perform Goodall's *F*-test and a bootstrapped *F*-test for significant differences between two groups. The other two programs perform regression (**ThreeDRegress6**) and principal components analysis (**ThreeDPCA6**) on files that have been processed by **simple3D**; both are modeled on the programs used for analyses of two-dimensional configuration. All three programs require a textfile describing a wireframe that will be used in illustrations (data analyses can proceed without the wireframe, but the only graphics that will function are the displays of the sample mean and the superimposed specimens). The wireframe file is simply a list of landmarks to be connected by each segment (e.g. 1 2 3, indicating that the first segment connects landmarks 2 and 3 with one row for each segment).

Semilandmarks

In the IMP package, two programs, **MakeFan** and **SemiLand**, have been designed for semilandmark processing. **MakeFan** provides several options for drawing fans (rays at equal angular intervals) or combs (perpendiculars from a reference line) that can be used as guides for digitizing semilandmarks. Pictures with fans can be digitized in **MakeFan**, or saved and opened for digitizing in **TPSDig**. **MakeFan** saves coordinates in the TPS format, so they must be converted in **CoordGen** before further processing. **SemiLand** takes the converted files, slides the semilandmarks to perpendicular alignment on the reference configuration, and then deletes the semilandmarks selected for culling (called *helpers*). Output from this program (superimposed configurations of landmarks and retained semilandmarks) can be input into any of the analytic programs in the IMP package.

In the TPS package, **tpsrelw** has been modified to permit analyses of semilandmarks and landmarks. This requires a "sliders" file in NTS format that identifies which points in the data file are semilandmarks. After loading both the data file and the sliders file, the option **Relax semilandmarks** will be available in the **Options** pull-down menu. Selecting **Relax semilandmarks** will invoke sliding to minimize bending energy.

References

Axler, S. (1996). *Linear Algebra Done Right*. Springer.

Bookstein, F. L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.

Bookstein, F. L. (1996a). Combining the tools of geometric morphometrics. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 131–152. Plenum.

Bookstein, F. L. (1996b). Standard formula for the uniform shape component in landmark data. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 153–168. Plenum.

Bookstein, F. L. (1997a). Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Medical Image Analysis*, **1**, 225–243.

Bookstein, F. L. (1997b). Shape and the information in medical images: a decade of the morphometric synthesis. *Computer Vision and Image Understanding*, **66**, 97–118.

Fadda, C., Faggiani, F. and Corti, M. (1997). A portable device for the three-dimensional landmark collection of skeletal elements of small mammals. *Mammalia*, **61**, 622–627.

Ferson, S., Rohlf, F. J. and Koehn, R. K. (1985). Measuring shape variation of two-dimensional outlines. *Systematic Zoology*, **34**, 59–68.

Green, W. D. K. (1996). The thin-plate spline and images with curving features. In *Proceedings in Image Fusion and Shape Variability Techniques* (K. V. Mardia, C. A. Gill and I. L. Dryden, eds) pp. 79–87. Leeds University Press.

Lohman, G. P. and Schweitzer, P. N. (1990). On eigenshape analysis. In *Proceedings of the Michigan Morphometrics Workshop* (F. J. Rohlf and F. L. Bookstein, eds) pp. 147–166. University of Michigan Museum of Zoology, Special Publication No. 2.

MacLeod, N. and Rose, K. D. (1993). Inferring locomotor behavior in Paleogene mammals via eigenshape analysis. *American Journal of Science*, **293A**, 300–355.

Rohlf, F. J. (1990). Rotational fit (Procrustes) methods. In *Proceedings of the Michigan Morphometrics Workshop* (F. J. Rohlf and F. L. Bookstein, eds) pp. 227–236. University of Michigan Museum of Zoology, Special Publication No. 2.

Rohlf, F. J. and Archie, J. W. (1984). A comparison of Fourier methods for the description of wing shape in mosquitoes (Diptera: Culicidae). *Systematic Zoology*, **33**, 302–317.

Rohlf, F. J. and Bookstein, F. L. (2003). Computing the uniform component of shape variation. *Systematic Biology*, **52**, 66–69.

Rohlf, F. J. and Corti, M. (2000). Use of two-block partial least-squares to study covariation in shape. *Systematic Biology*, **49**, 740–753.

Rohlf, F. J. and Slice, D. E. (1990). Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Zoology*, **39**, 40–59.

Sampson, P. D., Bookstein, F. L., Sheehan, F. H. and Bolson, E. L. (1996). Eigenshape analysis of left ventricular outlines from contrast ventriculograms. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 211–233. Plenum.

Spencer, M. A. and Spencer, G. S. (1995). Video-based three-dimensional morphometrics. *American Journal of Physical Anthropology*, **96**, 443–453.

Glossary

Affine transformation (Also called "uniform"). Transformation (or mapping) that leaves parallel lines parallel. The possible affine transformations include those that do not alter shape (scaling, translation, rotation) and those that do (shear and contraction/dilation). See also **Explicit uniform terms, Implicit uniform terms** (Chapter 6).

Allometry Shape change correlated with size change, sometimes more narrowly defined as a change in the size of a part according to the power law $Y = bX^k$, where Y is the size of the part, X is either the size of another part or overall body size, and k and b are constants. There are three distinct types of allometry: (1) ontogenetic, an ontogenetic change in shape correlated with an ontogenetic increase in size; (2) static, variation in shape correlated with variation size among individuals at a common developmental stage; and (3) evolutionary, an evolutionary change in shape correlated with evolutionary changes in size (Chapters 10, 13).

***Alpha* (α)** (1) The acceptable Type I error rate, typically 5%; (2) a factor multiplying partial warps before computing principal components of them; if $\alpha = 0$, principal components of partial warps are conventional principal components; when $\alpha \neq 0$, the partial warps are differentially weighted. Either those with lower bending energy are weighted more highly ($\alpha > 0$) or those with greater bending energy are weighted more highly ($\alpha < 0$). Typically, values of +1 or −1 are used. See also **Relative warps.**

ANCOVA Analysis of covariance. A method for testing the hypothesis that samples do not differ in their means when the effects of a covariate are taken into account. See also **ANOVA, MANOVA** and **MANCOVA** (Chapters 9, 10).

Anisotropic Not isotropic, having a preferred direction. In general, anisotropy is a measure of the degree to which variation in some parameter is a function of its direction relative to some axis. In geometric morphometrics, anisotropy usually refers to a measure of an affine transformation – either the ratio between principal strains, or a ratio of variances along principal axes. See also **Isotropic** (Chapter 3).

ANOVA Analysis of variance. A method for testing the hypothesis that samples do not differ in their means. ANOVA differs from MANOVA in that the means are unidimensional scalars. See also **ANCOVA, MANOVA** and **MANCOVA** (Chapter 9).

Baseline A line joining two landmarks, used in some superimposition methods to register shapes by assigning fixed values to one or more coordinates of those landmarks. See also **Baseline registration, Bookstein coordinates, Sliding baseline registration** (Chapters 3, 5).

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

Baseline registration A method of superimposing landmark configurations by assigning two landmarks fixed values (the two landmarks are the endpoints of the baseline). The most common method of baseline registration is the two-point registration developed by Bookstein, in which the ends of the baseline are fixed at (0, 0) and (1, 0), yielding Bookstein coordinates. Other methods of baseline registration fix the endpoints at different values (see Dryden and Mardia, 1998) or only fix one coordinate of each baseline point (see **Sliding baseline registration**) (Chapters 3, 5).

Basis A set of linearly independent vectors that span the entire vector space, also the smallest necessary set of vectors that span the space. The basis can serve as a coordinate system for the space because every vector in that space is a unique linear combination of the basis vectors. However, the basis itself is not unique; any vector space has infinitely many bases that differ by a rotation. An orthonormal basis is a set of mutually orthogonal axes, all of unit length. Partial warps and principal components are two common orthonormal bases used in shape analysis. See also **Eigenvectors** (Chapters 6, 7).

Bending energy (1) A measure of the amount of non-uniform shape difference based on the thin-plate spline metaphor. In this metaphor, bending energy is the amount of energy required to bend an ideal, infinite and infinitely thin steel plate by a given amplitude between chosen points. Applying this concept to the deformation of a two-dimensional configuration of landmarks involves modeling the displacements of landmarks in the X, Y plane as if they were displacements above or below the plane ($\pm Z$). (2) Eigenvalues of the bending-energy matrix, representing the amount of bending energy per unit deformation along a single principal warp (eigenvector of the bending-energy matrix). This concept of bending energy is useful because it provides a measure of spatial scale; it takes more energy to bend the plate by a given amount between closely spaced landmarks than between more distantly spaced landmarks. Thus, principal warps with large eigenvalues represent more localized components of deformation than principal warps with smaller eigenvalues. The total bending energy (definition 1) of an observed deformation is a sum of multiples of the eigenvalues, and accounts for the non-uniform deformation of the reference shape into the target shape. See also **Thin-plate spline, Principal warps, Partial warps** (Chapter 6).

Bending-energy matrix The matrix used to compute principal warps and their bending energies (eigenvectors and eigenvalues, respectively). This matrix is a function of the distances between landmarks in the reference shape. See also **Principal warps, Partial warps** (Chapter 6).

Biorthogonal directions Principal axes of a deformation; the term was used in Bookstein et al., 1985; more recently, workers refer to principal axes (Chapter 3).

Black Book Marcus, L. F., Bello, E. and Garcia-Valdcasas, A. (eds) (1993). *Contributions to Morphometrics*. Madrid, Monografias del Museo Nacional de Ciencias Naturales 8. (See also **Blue Book, Orange Book, Red Book** and **White Book**.)

Blue Book: Rohlf, F. J. and Bookstein, F. L. (eds) (1990). *Proceedings of the Michigan Morphometrics Workshop*. University of Michigan Museum of Zoology, Special Publication No. 2 (See also **Black Book, Orange Book, Red Book** and **White Book**.)

Bonferroni correction, Bonferroni adjustment An adjustment of the α-value to protect against inflating Type I error rate when testing multiple *a posteriori* hypotheses. The adjustment is done by dividing the acceptable Type I error rate (α) by the number of tests. That quotient is the adjusted α-value for each of the *a posteriori* hypotheses. For example, if the desired Type I error rate is 5%, and there are 10 *a posteriori* hypotheses to test, $0.05/10 = 0.005$ is the α-value for each of those 10 tests. A less conservative approach uses a sequential Bonferroni adjustment in which the desired α-value is divided by the number of remaining tests. Thus, the adjusted α for the first test would be 0.05/10; for the second it would be 0.05/9; for the third it would be 0.05/8, etc. To apply this sequential adjustment, hypotheses are ordered from lowest to highest p-value; the null hypothesis is rejected for each in turn until reaching one that cannot be rejected (the analysis stops at that point).

Bookstein coordinates (**BC**) The shape variables produced by the two-point registration, in which the configuration is translated to fix one end of the baseline at $(0, 0)$, and then rescaled and rigidly rotated to fix the other end of the baseline at $(1, 0)$. See also **Baseline registration** (Chapter 3).

Bookstein two-point registration (**BTR**) See **Two-point registration, Bookstein coordinates.**

Bootstrap test A statistical test based on random resampling (with replacement) of the data. Usually, the method is used to simulate the null model that one wishes to test. For example, if using a bootstrap test of the difference between means, the null hypothesis of no difference is simulated. Bootstrap tests are used when the data are expected to violate distributional assumptions of conventional analytic statistical tests. Rather than assuming that the data meet the distributional assumptions, bootstrapping produces an empirical distribution that can be used either for hypothesis testing or for generating confidence intervals. See also **Jackknife test, Permutation test** (Chapter 8).

Canonical variates analysis (**CVA**) A method for finding the axes along which groups are best discriminated. These axes (canonical variates) maximize the between-group variance relative to the within-group variance. Scores for individuals along these axes can be used to assign specimens (including unknowns) to the groups, and can be plotted to depict the distribution of specimens along the axes. CVA is an ordination rather than statistical method. See also **Ordination methods, Principal components analysis** (Chapter 7).

Cartesian coordinates Coordinates that specify the location of a point as displacements along fixed, mutually perpendicular axes. The axes intersect at the origin, or zero point, of all axes. Two Cartesian coordinates are needed to specify positions in a plane (flat surface); three are required to specify positions in a three-dimensional space. These coordinates are called "Cartesian" after the philosopher Descartes, a pioneer in the field of analytic geometry.

Centered A matrix is centered when its centroid is at the origin of a Cartesian coordinate system; i.e. at $(0, 0)$ of a two-dimensional system or at $(0, 0, 0)$ of a three-dimensional system (Chapter 6).

Centroid See **Centroid position.**

Centroid position The position of the averaged coordinates of a configuration of landmarks. The centroid position has the same number of coordinates as the landmarks. The X-component of the centroid position is the average of the X-coordinates of all landmarks of an individual configuration. Similarly, the Y-component is the average of the Y-coordinates of all landmarks of an individual configuration. It is common to place the centroid position at $(0, 0)$, because this often simplifies other computations (Chapter 6).

Centroid size (CS) A measure of geometric scale, calculated as the square root of the summed squared distances of each landmark from the centroid of the landmark configuration. This is the size measure used in geometric morphometrics. It is favored because centroid size is uncorrelated with shape in the absence of allometry, and also because centroid size is used in the definition of the Procrustes distance (Chapters 3, 4, 5).

Coefficient A number multiplying a function. For example, in the equation $Y = mX$, m is the coefficient for the slope, which is the function that relates X and Y.

Column vector A vector whose entries are arranged in a column. Contrast to a **Row vector**.

Complex numbers A number consisting of both a real and an imaginary part. An imaginary number is a real number multiplied by i, where i is $\sqrt{-1}$. A complex number is written as $Z = X + iY$, where X and Y are real numbers. In that notation, X is said to be the real part of Z and Y is the imaginary part. A complex number is often used to represent a vector in two dimensions. The mathematics of two-dimensional vectors and complex numbers are similar, so it is sometimes useful to perform calculations or derivations in complex number form.

Configuration see **Landmark configuration**.

Configuration matrix A matrix representing the configuration of K landmarks, each of which has M dimensions. A configuration matrix is a $K \times M$ matrix in which each row represents a landmark and each column represents one Cartesian coordinate of that landmark; $M = 2$ for landmarks of two-dimensional configurations (planar shapes), and $M = 3$ for landmarks of three-dimensional configurations. Two configuration matrices can differ in location, size and orientation, as well as shape (Chapter 4).

Configuration space The set of all possible configuration matrices describing all possible configurations of K landmarks with M coordinates (all with the same values of K and M). Because there are $K \times M$ elements in the configuration matrices, there are $K \times M$ dimensions in the configuration space. In statistical analyses, the configuration space accounts for $K \times M$ degrees of freedom because that is the number of independent pieces of information (e.g. landmark coordinates) needed to specify a particular configuration (Chapter 4).

Consensus configuration The mean (average) configuration of landmarks in a sample of configurations. Usually, this is calculated after superimposing coordinates. See also **Generalized Procrustes superimposition**, **Reference form** (Chapters 4, 5).

Contraction A mathematical mapping that "shrinks" a configuration along one axis. A contraction along the X-axis would map the point (X, Y) to the point (AX, Y), where A is less than one. A contraction along the Y-axis would map (X, Y) to (X, AY). **Expansion** or **dilation** is the opposite of contraction $(A > 1)$.

Coordinates The set of values that specify the location of a point along a set of axes (see **Cartesian coordinates**).

Correlation A measure of the association between two or more variables. In morphometrics, correlation is most often measured using Pearson's product-moment correlation, which is the covariance divided by the product of the variances:

$$\mathrm{R}_{XY} = \frac{\sum (X - X_{\mathrm{mean}})(Y - Y_{\mathrm{mean}})}{\sqrt{\sum (X - X_{\mathrm{mean}})^2 \sum (Y - Y_{\mathrm{mean}})^2}}$$

where the sums are taken over all specimens. When variables are highly correlated we can predict one from the other (e.g. Y from X), and the more highly correlated they are, the better our predictions will be. Uncorrelated variables are considered independent. See also **Covariance**.

Covariance Like correlation, a measure of the association between variables. The sample estimate of the covariance between X and Y is:

$$\mathrm{S}_{XY} = \left(\frac{1}{N - 1}\right) \sum (X - X_{\mathrm{mean}})(Y - Y_{\mathrm{mean}})$$

where the summation is over all N specimens.

Curved space A metric space in which the distance measure is not linear. The ordinary rules of Euclidean geometry do not apply in such spaces. The consequences of the curvature depend upon the distance between points; we can treat the surface of the earth as flat as long as the maps cover only small areas, but in long-distance navigation the curvature must be taken into account. Shape space is curved, so the rules of Euclidean geometry do not apply, which is why shapes are mapped onto a Euclidean space tangent to shape space.

D A generalized statistical distance between means of two groups (**X1** and **X2**) relative to the variance within the groups:

$$\mathrm{D} = \sqrt{(\mathbf{X1} - \mathbf{X2})^{\mathrm{T}} \mathbf{S}_{\mathrm{p}}^{-1} (\mathbf{X1} - \mathbf{X2})}$$

where $(\ )^{\mathrm{T}}$ refers to the transpose of the enclosed matrix, and $\mathbf{S}_{\mathrm{p}}^{-1}$ is the inverse of the pooled variance–covariance matrix. This distance takes into account the correlations among variables when computing the distance between means. The generalized distance is used in Hotelling's T^2-test. Also known as the **Mahalanobis' distance**.

D^2 The squared generalized distance, **D**. See **D**.

Deformation A smooth, continuous mapping or transformation; in morphometrics, it is usually the transformation of one shape into another. The deformation refers not only to the change in positions of landmarks, but also to the interpolated changes in locations of unanalyzed points between landmarks (Chapter 6).

Degrees of freedom In general, the number of independent pieces of information. In statistical analyses, the total degrees of freedom are approximately the product of the number of variables and the number of individuals (the total may be partitioned into separate components for some tests). If every measurement on every individual were completely independent, the degrees of freedom would be the product of the number of variables and the number of individuals, but if one statistic is known (or estimated), the number of degrees of freedom that remain to estimate a second statistic will be reduced. For example, the estimate of the mean height of N individuals in a sample will have $N \times 1 = N$ degrees of freedom, because all N measurements are needed and there is only one measured variable. In contrast, the estimate of the variance in height will have $N - 1$ degrees of freedom because only $N - 1$ deviations from mean height are independent (the deviation of the Nth individual can be calculated from the mean and the other $N - 1$ observed heights). In geometric morphometrics, when configurations of landmarks are superimposed, degrees of freedom are lost for a different reason; namely, information that is not relevant to comparison of shapes (location, scale and rotation) is removed from the coordinates.

Dilation Opposite of **Contraction**.

Discriminant function The linear combination of variables optimally discriminating between two groups. It is produced by discriminant function analysis. Scores on the discriminant function can be used to identify members of the groups (Chapter 7).

Discriminant function analysis A two-group **canonical variates analysis**. See **Canonical variates analysis** (Chapter 7).

Disparity, morphological disparity (*MD*) Phenotypic variety, usually morphological. Several metrics can be used to measure disparity, but the one most commonly used in studies of continuous variables is:

$$MD = \frac{\sum_{j=1}^{N} D_j^2}{(N-1)}$$

where D_j is the distance of species j from the overall centroid (i.e. the grand mean calculated over N groups, e.g. species) (Chapter 12).

Distance A function measuring the separation between points. Within any space there are multiple possible distances. For this reason, it is necessary to specify the type of distance used. See also **D**, **D^2, Euclidean distance, Generalized distance, Geodesic distance, Great circle distance, Partial Procrustes distance, Full Procrustes distance, Mahalanobis' distance** (Chapter 4).

Dot product (Also called inner product.) Given two vectors $\mathbf{A} = \{A_1, A_2, A_3 \ldots A_N\}$, $\mathbf{B} = \{B_1, B_2, B_3 \ldots B_N\}$, the dot product of **A** and **B** is:

$$\mathbf{A} \cdot \mathbf{B} = A_1B_1 + A_2B_2 + A_3B_3 + \ldots. + A_NB_N$$

and

$$\mathbf{A} \cdot \mathbf{B} = |A||B| \cos(\theta)$$

where $|A|$ is the magnitude of **A**, $|B|$ is the magnitude of **B**, and θ is the angle between **A** and **B**. If the magnitude of **A** is 1, then $\mathbf{A} \cdot \mathbf{B} = |B| \cos(\theta)$, which is the component of **B** along the direction specified by **A**. The dot product is used to calculate scores on coordinate axes, by projecting the data onto those axes (this is how partial warp scores and scores on principal components are calculated). It is also used to find the vector correlation, R_V, between two vectors (that correlation is the cosine of the angle between vectors).

Edge registration See **Baseline registration.**

Eigenvalues See **eigenvectors.**

Eigenvectors Eigenvectors are the non-zero vectors, **A**, satisfying the eigenvector equation:

$$(\mathbf{X} - \lambda\mathbf{I})\mathbf{A} = 0$$

The values of λ that satisfy this equation are eigenvalu*e*s of **X**. Eigenvectors are orthogonal to one another, and provide the smallest necessary set of axes for a vector space (i.e. they provide a basis for that space). The eigenvectors of a variance–covariance matrix are called principal components; the eigenvalue corresponding to each axis gives the variance associated with it. The eigenvectors of the bending-energy matrix are the principal warps; the eigenvalue corresponding to each axis gives the bending energy associated with it. See also **Basis** (Chapters 4, 6, 7).

Element of a matrix A number in a matrix, typically referenced by the symbol designating the matrix with subscripts indicating its row and column; for example, $X_{4,5}$ refers to the element on the fourth row and fifth column of the matrix **X**.

Euclidean distance The square root of the summed squared distances along all orthogonal axes. A Euclidean distance does not change when the axes of the space are rotated (in contrast to a Manhattan distance, which is simply the sum of the distances). See also **D, D^2, Distance, Generalized distance, Geodesic distance, Great circle distance, Procrustes distance, Full Procrustes distance, Partial Procrustes distance** (Chapter 4).

Euclidean space A coordinate space in which the metric is a Euclidean distance.

Explicit uniform term, explicit uniform component A uniform component describes affine or uniform deformations. Some of these do not alter shape (i.e. rotation, translation and rescaling) whereas others do (i.e. shear and dilation). Accordingly, we divide affine deformations into two sets: (1) implicit uniform terms, which do not alter shape and are used in superimposing forms but are not explicitly recorded; and (2) explicit uniform terms, which do alter shape and therefore are typically reported as components of the deformation. All uniform terms must be known to model a deformation correctly (Chapter 4).

Fiber In geometric morphometrics, the set of all points in pre-shape space representing all possible rigid rotations of a landmark configuration that has been centered and scaled to unit centroid size; in other words, the set of pre-shapes that have the same shape. Fibers are collapsed to a point in shape space (Chapter 4).

Form Size-plus-shape of an object; form includes all the geometric information not removed by rotation and translation. Form is also called **Size-and-shape**.

Full Procrustes distance (D_F) The distance between two landmark configurations in the linear space tangent to Kendall's shape space (i.e. the tangent space) when centroid size of one is allowed to vary to minimize the distance between the shapes rather than fixed to unit size. See also **Partial Procrustes distance** (Chapters 4, 5).

Full Procrustes superimposition A superimposition minimizing the full Procrustes distance. See also **Partial Procrustes distance** (Chapters 4, 5).

Generalized distance See **D**.

Generalized least squares superimposition A generalized superimposition method that uses a least squares fitting criterion, meaning that the parameters are estimated to minimize the sum of squared distances over all landmarks over all specimens. Usually, in geometric morphometrics, GLS refers specifically to a generalized least squares Procrustes superimposition – a different approach is used in generalized resistant-fit methods (Chapter 5).

Generalized least squares Procrustes superimposition (GLS) A generalized superimposition minimizing the partial Procrustes distance over all shapes in the sample, using a least squares fitting function. This is the method usually used in geometric morphometrics (Chapters 4, 5).

Generalized superimposition The superimposition of a set of specimens onto their mean. This involves an iterative approach because the mean cannot be calculated without superimposing specimens, which cannot be superimposed on the mean before the mean is calculated (an alternative approach is used in ordinary Procrustes analysis). See also **Consensus configuration** (Chapters 4, 5).

Geodesic distance The shortest distance between points in a space. On a flat planar surface, this is the length of the straight line joining the points – i.e. the Euclidean distance. On curved surfaces, this distance is the length of an arc.

Great circle The intersection of the surface of a sphere and a plane passing through its center. A great circle divides the surface of the sphere in half. On the surface of the sphere, the shortest distance between two points lies along the great circle that passes through those points. If the Earth were perfectly spherical, the equator and all lines of latitude would be great circles.

Great circle distance The arc length of the segment of the great circle connecting two points on the surface of a sphere; this is the geodesic distance between those points, the shortest distance between the points in the space of the surface of the sphere.

Homology (1) Similarity due to common evolutionary origin. In morphometrics, landmarks are considered homologous by virtue of the homology of the structures defining their locations. (2) Some morphometricians use the term for the correspondences between points that are imputed by a mathematical function, called a "homology function" (e.g. see Bookstein et al., 1985). Homology is the primary criterion for selecting landmarks (Chapter 2).

Hypersphere The generalization of a three-dimensional sphere to more than three dimensions. In three dimensions, points on the surface of a sphere of radius R that is centered at the origin satisfy the equation $X^2 + Y^2 + Z^2 = R^2$.

Implicit uniform terms See **Explicit uniform terms.**

Induced correlation A correlation induced by dividing two values by a third which is common to both. The induced correlation between the (rescaled) variable is not present in the original variables.

Inner product See **Dot product.**

Invariant A quantity is invariant under a mathematical operation or transformation when it is not changed by that operation. For example, centroid size is invariant under translation, centroid position is not.

Isometric In general, a transformation that leaves distances between points unaltered. In morphometrics, isometry usually means that shape is uncorrelated with size. In statistical tests of allometry, isometry is the null hypothesis (Chapters 10, 13).

Isotropic A property is said to be isotropic if it is uniform in all directions, i.e. if it does not differ as a function of direction. When an error is isotropic, it is equal in all directions, and there is no correlation among errors. Isotropic is the opposite of anisotropic.

Jackknife test An approach to statistical testing that involves resampling the original observations to generate an empirical distribution. Jackknifing is carried out by omitting one specimen at a time. See also **Bootstrap test, Permutation test** (Chapter 8).

Kendall's shape space The space in which the distance between landmark configurations is the Procrustes distance. This space is constructed by using operations that do not alter shape to minimize differences between all configurations of landmarks that have the same values of K (number of landmarks) and M (number of coordinates of a landmark). Kendall's shape space is the curved surface of a hypersphere, so conventional statistical analyses are conducted in a Euclidean tangent space (Chapter 4).

Landmark Biologically, landmarks are discrete, homologous anatomical loci; mathematically, landmarks are points of correspondence, matching within and between populations (Chapter 2).

Landmark configuration The positions (coordinates) of a set of landmarks representing a single object, containing information about size, shape, location and orientation. The number of landmarks is typically represented by K, and the dimensionality of the landmarks (number of coordinates) is typically represented by M. Therefore, if there are 16 landmarks, each with an X- and Y-coordinate, then $K = 16$ and $M = 2$ (Chapter 4).

Least squares A method of choosing parameters that minimizes the summed square differences over all individuals (and variables) (Chapter 10).

Linear A function $\mathrm{f}(X)$ is linear if it depends only on the first power of X; e.g. $\mathrm{f}(X) = 2(X)$ is linear, but $\mathrm{f}(X) = 2(X)^2$ is not.

Linear combination A vector produced by multiplying and summing coefficients of one or more vectors. For example, given the vector $\mathbf{X}^\mathrm{T} = \{X_1, X_2 \ldots X_N\}$ and $\mathbf{A}^\mathrm{T} = \{A_1, A_2 \ldots A_N\}$, then $\mathbf{Y} = A_1X_1 + A_2X_2 + \cdots A_NX_N$ is a linear combination of the vectors. We can write this as $\mathbf{Y} = \mathbf{A}^\mathrm{T}\mathbf{X}$.

Linear transformation A transformation producing a set of new vectors that are linear combinations of the original variables. See **Linear combination**.

Linear vector space The set of all linear combinations of a set of vectors. The space spans all possible linear combinations of the basis vectors, as well as all sums or differences of any linear combination of those basis vectors. The two-dimensional Cartesian plane is the linear vector space formed by the linear combinations of two vectors of unit length, one along the X-axis, the other along the Y-axis.

Mahalanobis' distance (D) The squared distance between two means divided by the pooled sample variance–covariance matrices. This is a generalized statistical distance, adjusting for correlations among variables. See also **D**, **Generalized distance**.

MANCOVA Multivariate analysis of covariance. A method for testing the hypothesis that samples do not differ in their means when the effects of a covariate are taken into account. See also **ANOVA**, **ANCOVA** and **MANOVA** (Chapters 9, 10).

MANOVA Multivariate analysis of variance. A method for testing the hypothesis that samples do not differ in their means; MANOVA differs from ANOVA in that the means are multidimensional vectors. See also **ANOVA**, **ANCOVA** and **MANCOVA** (Chapter 9).

Map A mathematical function relating **X** to **Y** by stating the correspondence between elements in **X** and **Y**. Each element in **X** is placed in correspondence with one element in **Y**. Multiple elements in **X** may map to the same element in **Y** (landmark configurations differing only in rotation for example would all map to the same shape). A map is written as: $f : \mathbf{X} \rightarrow \mathbf{Y}$ where f is the map from the set **X** to the set **Y**.

Matrix A rectangular array of numbers (real or complex). The numbers in a matrix are referred to as elements of the matrix. The size of a matrix is always given as the number of rows followed by the number of columns; e.g. a 4×2 matrix has four rows and two columns.

Mean Also known as the average; an estimate of the center of the distribution calculated by summing all observations and dividing by the sample size.

Median An estimate of the center of a distribution calculated such that half the observed values are above and the other half are below.

Metric A non-negative real-valued function, $D(X, Y)$, of the points X and Y in a space such that:

1. The only time that the function is zero is when X and Y are the same point, i.e. $D(X, Y) = 0$, if and only if $X = Y$
2. If we measure from X to Y, we get the same distance as when we measure from Y to X, so $D(X, Y) = D(Y, X)$ for all X and Y
3. The triangle inequality holds true. The triangle inequality states the distance between any two points, X and Y, is less than or equal to the sum of distances from each to a third point, Z, so $D(X, Y) \leq D(X, Z) + D(Y, Z)$, for all X, Y and Z.

Multiple regression Regression of a single (univariate) dependent variable on more than one independent variable. See also **Multivariate regression, Regression.**

Multivariate analysis of variance See **MANOVA.**

Multivariate multiple regression Regression of several dependent variables on more than one independent variable. In morphometrics, this method is used to regress shape (the dependent variables) onto multiple independent variables. See also **Multiple regression, Multivariate regression, Regression.**

Multivariate regression Regression of several dependent variables onto one independent variable. In morphometrics, this method is used to regress shape onto a single independent variable, such as size. The coefficients obtained by multivariate regression are the same as those estimated by simple bivariate regression of each dependent variable on the independent variable. However, the statistical test of the null hypothesis differs. See also **Multiple regression, Multivariate multiple regression, Regression** (Chapters 10, 13).

Non-uniform Not **Uniform; Non-affine.** See **Non-uniform deformation.**

Non-uniform deformation The component of a deformation that is not uniform. In contrast to a uniform deformation, which leaves parallel lines parallel and has the same effect everywhere across a form, a non-uniform deformation turns squares into trapezoids or diamonds (shapes that do not have parallel sides) and has different effects over different regions of the form. Most deformations comprise both uniform and non-uniform parts. The non-uniform component can be further subdivided, see **Partial warps** (Chapter 6).

Normalize To set the magnitude to one. Normalizing a vector sets the length of the vector to one; this is done by dividing each component of the vector by the length of the vector, calculated by taking the square root of the summed squared coefficients.

Null hypothesis, or **null model** Usually, the hypothesis that the factor of interest has no effect beyond that expected by chance. For example, in an analysis of allometry, the null hypothesis being tested by regression of shape on size is that shape does not depend on size (i.e. isometry). Similarly, in a comparison of two means using Hotelling's T^2-test, the null hypothesis is that the two groups do not differ beyond what is expected by chance.

Orange Book Bookstein, F. L. (1991). *Morphometric Tools for Landmark Data. Geometry and Biology*. Cambridge University Press. (See also **Black Book, Blue Book, Red Book** and **White Book**.)

Ordinary Procrustes analysis (OPA) An approach to superimposition in which one landmark configuration is fitted to another, differing from a **Generalized superimposition** in that it involves only two forms. This approach has rarely been used since iterative methods became available for generalized superimpositions. See also **Generalized superimposition, Consensus form** (Chapter 5).

Ordination Ordering specimens along one or more axes based on some criterion (e.g. from youngest to oldest, or shortest to tallest). Ordination methods include principal components analysis and canonical variates analysis; the scores on the axes provide a basis for ordering specimens (Chapter 7).

Orthogonal Perpendicular (at right angles to each other). Two vectors are orthogonal if the angle between them is 90°; when they are, their dot product is zero.

Orthonormal Perpendicular and of unit length; vectors are orthonormal if they are mutually orthogonal and of unit length.

Orthonormal basis See **Basis.**

Population The set of all possible individuals of a specific type, such as all members of a species, or all leaves on a particular kind of tree. See also **Sample** (Chapter 8).

Outline A curve around the perimeter of an object (or around a distinct part of it).

Partial least squares analysis A method of exploring patterns of covariance or correlation between two blocks of variables measured on the same set of specimens. A singular value decomposition is used to determine the pair of vectors (each a linear combination of variables within one of the blocks) that expresses the greatest proportion of the covariance between blocks. See also **Singular value decomposition**, **Singular warps** (Chapter 11).

Partial Procrustes distance (D_p) The distance between two landmark configurations in the linear tangent space to Kendall's shape space when both shapes are centered, fixed to unit centroid size, and rotated to minimize the sum of squared distances between their corresponding landmarks. See also **Full Procrustes distance**, **Procrustes distance** (Chapters 4, 5).

Partial Procrustes superimposition A superimposition that minimizes the partial Procrustes distance between shapes. See also **Full Procrustes distance, Procrustes distance** (Chapters 4, 5).

Partial warps The term partial warps sometimes refers solely to the components of the non-uniform deformation, which are computed as eigenvectors of the bending-energy matrix projected onto the X, Y-plane of the data (they are projections of principal warps), ordered from least to most bending energy. These eigenvectors provide an orthonormal basis for the non-uniform part of a deformation. Sometimes "partial warps" also includes the components of the uniform deformation, as the zeroth partial warp – in which case the scores on this component are included among the partial warp scores (Chapter 6).

Partial warp scores Coefficients indicating the position of an individual, relative to the reference, along partial warps. They are calculated by taking the dot product between the partial warps and the data for a specimen. When appropriate scores on the uniform component are also included among the partial warps scores, the sum of the squared scores equals the squared partial Procrustes distance of that specimen from the reference. This full set of scores can be used as shape variables in any conventional statistical analysis because they are based on the appropriate distance measure and have the same number of coordinates as degrees of freedom. See also **Non-uniform deformation**, **Partial warps, Principal warps, Uniform deformation** (Chapter 6).

Permutation test An approach to statistical testing that involves permuting (rather than randomly sampling) observed values. See also **Bootstrap test**, **Jackknife test**, **Monte Carlo simulations** (Chapter 8).

Pinocchio effect A large change concentrated at one landmark, with little or none at others; a highly localized change. In the presence of the Pinocchio effect, Procrustes superimpositions imply that the shape difference is distributed over all landmarks. Resistant-fit methods, such as RFTRA, were devised to avoid that implication (Chapter 5).

Position See **Centroid position.**

Pre-shape A centered landmark configuration, scaled to unit centroid size (Chapter 4).

Pre-shape space The set of all possible pre-shapes for a given number of landmarks with a given number of dimensions. This is the surface of a sphere of $KM - M - 1$ dimensions, where K is the number of landmarks and M is the number of dimensions of each landmark (Chapter 4).

Principal axes The set of orthogonal axes used in modeling the change of one shape into another as an affine transformation. This transformation can be parameterized by its effect on a circle or sphere (for two or three dimensional shapes, respectively). In two dimensions, an affine transformation takes a circle into an ellipse and the principal axes are the directions of the circle that undergo the greatest relative elongation or shortening mapped onto the major and minor axes of the ellipse. The ratio of the lengths of these axes is the anisotropy, a measure of the amount of affine shape change. Principal axes are invariant under a change in the coordinate system. See also **Principal strains** (Chapter 3).

Principal components analysis (PCA) A method for reducing the dimensionality of multivariate data, performed by extracting the eigenvectors of the variance–covariance matrix. These eigenvectors are called principal components. Their associated eigenvalues are the variance explained by each axis. Principal components provide an orthonormal basis. The position of a specimen along a principal component is represented as its principal component score, calculated by taking the dot product between that principal component and the data for that specimen (Chapter 7).

Principal strain In an affine deformation, the ratio of the length of a principal axis in the ellipse to the original diameter of the circle. See also **Principal axes** (Chapter 3).

Principal warp An eigenvector of the bending-energy matrix interpreted as a warped surface over the surface of the X, Y-plane of the landmark coordinates. Principal warps are ordered from least to most bending energy (smallest to largest eigenvalue), which corresponds to the least to most spatially localized deformation. Principal warps differ from partial warps in that partial warps are projections of principal warps onto the X, Y-plane of the data. See also **Bending energy, Bending-energy matrix, Orthonormal basis, Partial warp, Thin-plate spline** (Chapter 6).

Probability distribution A mathematical function that describes the probability of a measurement taking on either a particular value or a range of values, depending on whether the variable is discrete or continuous, respectively (Chapter 8).

Procrustes distance The distance between two landmark configurations in Kendall's shape space. It is approximately the square root of the summed squared distances between homologous landmarks when the configurations are in Procrustes superimposition. This distance is measured in the curved shape space (Chapter 4).

Procrustes methods A general term referring to the superimposition of matrices based on a least squares criterion. The term comes from the Greek mythological figure, Procrustes, who fitted visitors to a bed by stretching them or amputating overhanging parts (Chapter 5).

Procrustes residuals Coordinates of a landmark configuration obtained by a Procrustes superimposition. They are residuals in the sense that they indicate the deviation of each specimen from the mean (i.e. the consensus configuration) or other reference. See also **Consensus configuration, Procrustes superimposition, Reference** (Chapter 5).

Procrustes superimposition A superimposition of shapes that minimizes the Procrustes distances over the sample. The term is used whether the distance being minimized is the full or the partial Procrustes distance (Chapters 4, 5).

Red Book Bookstein, F. L., Chernoff, B., Elder, R. L. et al. (eds) (1985). *Morphometrics in Evolutionary Biology: The Geometry of Size and Shape Change, with Examples from Fishes*. Academy of Natural Sciences of Philadelphia, Special Publication No. 15. (See also **Black Book, Blue Book, Orange Book** and **White Book.**)

Reference, Reference form The shape to which all others are compared. It is the point of tangency between Kendall's shape space and the tangent space. Because the linear approximation to Kendall's shape space may be inaccurate when the point of tangency is far from the center of the distribution of specimens, the reference is usually chosen to minimize the distances between it and the other specimens – i.e. it is chosen to be the consensus shape (Chapters 4, 5).

Regression An analytic procedure for fitting a predictive model to data and assessing the validity of that model. One variable is expressed as a function of the other, e.g. $Y = mX + b$ expresses Y as a linear function of X. The predictor variable(s) are the independent variable(s), and those variables predicted by the model are the dependent variable(s). In the linear model above, X is the independent variable that predicts the dependent variable, Y. The term "regression" comes from Francis Galton (1889), who concluded that offspring tend towards (regress towards) the mean of the population. As stated by Galton in his law of universal regression, "each peculiarity in a man is shared by his kinsman, but *on the average*, in a less degree." Thus, the offspring of unusually tall fathers regress towards the mean height of the population (Chapters 10, 13).

Relative warps Principal components of partial warp scores, sometimes weighted to emphasize components of low or high bending energy (that weighting is done by setting the parameter α to a value other than 0). Originally, the term referred to an eigenanalysis of the variance–covariance matrix relative to the bending-energy matrix, hence a new term was coined for these components (Bookstein, 1991). Currently, the term usually refers to a conventional principal components analysis of partial warp scores. See also ***Alpha*** **($\boldsymbol{\alpha}$)**, **Bending energy, Partial warp scores, Principal components analysis** (Chapter 7).

Repeated median The median of medians, used in estimating the scaling factor and rotation angle by resistant-fit superimposition methods such as RFTRA. The repeated median is more robust to large deviations than the median or a least squares estimator. See also **Resistant-fit superimposition, RFTRA** (Chapter 5).

Resampling A method whereby a new data set is constructed by randomly selecting from the original data (either values recorded on specimens or residuals from a model).

Construction of a large series of resampled data sets can be used to simulate either the distribution of measured values or the distribution of a test statistic under the null model. Under some conditions, resampling can also be used to produce confidence intervals around the statistic. This approach permits hypothesis tests when the data are expected to deviate from the distributional assumptions of conventional analytic tests. Resampling may be done with replacement, meaning that each observation can appear more than once in a resampled data set; resampling without replacement means each observation appears only once in a set. See also **Bootstrap test, Jackknife test, Permutation test** (Chapter 8).

Rescale Multiply or divide by a scalar value; used in geometric morphometrics to change the centroid size of a configuration (Chapters 3, 4).

Residual Deviation of an observation from the expected value under a model. For example, a residual from a regression is the deviation between the observed and expected values of the dependent variable at a given value of the independent variable. The term is also used for the coordinates obtained by a Procrustes superimposition, the Procrustes residuals, which are deviations between individual specimens and the reference.

Resistant-fit superimposition A superimposition method that uses medians or repeated medians (rather than a least squares error criterion) to superimpose forms. The method is intended to be resistant to large localized shape differences, such as those produced by the Pinocchio effect. RFTRA is an example of this type of method. See also **Repeated medians, RFTRA** (Chapter 5).

RFTRA (Resistant fit theta-rho analysis) A resistant-fit superimposition method using the method of repeated medians to determine the scaling factor and rotational angle. See also **Resistant fit** and **Repeated median** (Chapter 5).

Rigid rotation A rotation of an entire vector or matrix by a single angle. Rigid rotations do not alter the size, shape or location of the object. Rotations are often represented by square matrices. The rotation matrix:

$$\mathrm{R} = \begin{vmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{vmatrix}$$

rotates a $2 \times N$ matrix through an angle θ. When different vectors are multiplied by different angles, the rotation is oblique, not rigid.

Row vector A vector with coefficients in a row. Contrast to a **Column vector**.

Sample The collection of observed individuals representing members of a population. An individual observation is the smallest sampling unit in the study, which might be an individual organism or one of its parts, or a collection of organisms such as a species or a bacterial colony (Chapter 8).

Scalar A real or complex number.

Scale (1) Noun – size of an object (given some definition of size); (2) verb – to change the size of an object (equivalent to rescale).

Scaling factor A constant which is used to change the scale or size of a matrix or vector. This is done by multiplying or dividing the matrix or vector by the constant.

Score In morphometrics, a coefficient locating a specimen along a vector, calculated by projecting the specimen onto an axis. Usually, scores locate the position of a specimen relative to the axes of a coordinate system. They are calculated by taking the dot product between an axis of the coordinate system and the data of a specimen. The scores are linear combinations of the original variables. Partial warp scores locate the position of an individual specimen relative to the coordinate system provided by the partial warps. Similarly, principal component scores locate the position of an individual specimen relative to the coordinate system provided by the principal components. Scores can be calculated relative to any basis of a vector space because each basis provides a coordinate system for that space. See **Dot product**.

Semilandmark A point on a geometric feature (curve, edge or surface) defined in terms of its position on that feature (e.g. at 10% of the length of the curve from one end). Semilandmarks are used to incorporate information about curvature in a geometric shape analysis. Because semilandmarks are defined in terms of other features, they represent less information (fewer degrees of freedom) than landmarks (Chapter 15).

Shape Shape has a variety of inconsistent definitions. In geometric morphometrics, the definition of shape is Kendall's: all the geometric information remaining in an object (such as a landmark configuration) after differences in location, scale and rotational effects are removed (Chapters 1, 4, 5).

Shape coordinates Within geometric morphometrics, coordinates of landmarks after superimposition, whether by a two-point registration (which yields Bookstein shape coordinates), or Procrustes superimposition (which yields Procrustes residuals) (Chapters 3, 5).

Shape space Within geometric morphometrics, shape space refers to Kendall's shape space. The term is more general, however, as it can apply to any space defined by a particular mathematical definition of shape. There are shape spaces for outline measurements, for example. There are also shape spaces based on different definitions of size. The characteristics of these various shape spaces are not necessarily the same as those of Kendall's shape space (Chapter 4).

Shape variable A general term for any variable expressing the shape of an object, including ratios, angles, shape coordinates obtained by a superimposition method, or vectors of coefficients obtained from partial warp analysis, principal components analysis, regression, etc. Shape variables are invariant under translation, scaling and rotation.

Shear An affine or uniform deformation that leaves the Y-coordinate fixed while the X-coordinate is displaced along the X-axis by a multiple of Y. Under a shear, the point

(X, Y) maps to $(X + AY, Y)$, where A is the magnitude of the shear. Visually, this looks like altering a square by sliding the top side to the left or right, without altering its height or the lengths of the top and bottom (Chapter 5, 6).

Singular axes Orthonormal vectors produced by singular value decomposition. See **Singular value decomposition** (Chapter 11).

Singular value In a singular value decomposition, a quantity expressing a relationship between two singular axes; an element λ_i of the diagonal matrix **S**. In partial least squares analysis, each singular value represents the covariance explained by the corresponding pair of singular axes. See **Singular value decomposition** (Chapter 11).

Singular value decomposition (SVD) A mathematical technique for taking an $M \times N$ matrix **A** (where N is greater than or equal to M) and decomposing it into three matrices:

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$$

where **U** is an $M \times N$ matrix whose columns are orthonormal vectors, **S** is an $N \times N$ diagonal matrix with on-diagonal elements λ_i, and **V** is an $N \times N$ matrix whose columns are orthonormal vectors. The values λ_i are called the *singular values* of the decomposition, and the columns of **U** and **V** are called the *singular vectors* or *singular axes* corresponding to a given singular value. In partial least squares analysis, **A** is the matrix of covariances between the two blocks, the columns of **U** are linear combinations of the variables in one of the two data sets, the columns of **V** are linear combinations of the variables in the other data set, and each λ_i is the portion of the total covariance explained by the corresponding pair of singular axes (Chapter 11).

Singular warps Singular axes computed from shape data (partial warp scores or residuals of a Procrustes superimposition), so that the singular axes describe patterns of differences in shape. See **Singular value decomposition** (Chapter 11).

Size Any positive real valued function $\mathrm{g}(\mathbf{X})$, where **X** is a configuration or set of points, such that $\mathrm{g}(A\mathbf{X}) = A\mathrm{g}(\mathbf{X})$, where A is any positive, real scalar value. In other words, multiplying every element in **X** by A multiplies $\mathrm{g}(\mathbf{X})$ by A. There are a wide variety of measures of size, including lengths measured between landmarks, sums or differences of interlandmark distances, square roots of area, etc. The size measure used in geometric morphometrics is centroid size. See also **Centroid size** (Chapters 3, 4).

Size-and-shape All the geometric information remaining in an object (such as a landmark configuration) after differences in location and rotational effects are removed. See **Form**.

Space A set of objects (or measurements thereof) that satisfies some definition. For example, a space might be defined as the set of all four-landmark configurations measured in two dimensions.

Statistic Any mathematical function based on an analysis of all measured individuals, e.g. the mean, standard deviation, variance, maximum, minimum, and range. The true value of the statistic in the population is called the parameter, which we are trying to estimate from our sample (Chapter 8).

Superimposition A method for matching two landmark configurations (or matrices) prior to further analysis. A number of different optimality criterion may be used. See also **Bookstein coordinates, Procrustes superimposition** (also **Full Procrustes superimposition** and **Partial Procrustes superimposition, RFTRA, Sliding baseline registration** (Chapters 3, 5).

Strain See **Principal strain.**

Tangent space The linear vector space tangent to a curved space. In geometric morphometrics, the Euclidean space tangent to Kendall's shape space. In the tangent space, distances between shapes are linear functions, which allows for analysis of shape variation by ordinary multivariate statistical methods. When the linear approximation to the curved surface is accurate (when all shapes in a study are close to the point of tangency), distances in the tangent space approximate distances in the curved space. The point of tangency between Kendall's shape space and the tangent space is the reference form. See also **Kendall's shape space, Reference form** (Chapter 4).

Target shape A shape being compared to the reference shape. See **Reference.**

Thin-plate spline An interpolation function used to predict the difference in shape between a reference and another shape over all points on the form, not just at landmarks. This interpolation function minimizes the bending energy of the deformation, which is equivalent to modeling that deformation as smoothly as possible given the observed landmarks (thus taking a parsimonious approach to interpolation). Thin-plate spline analysis produces scores for the non-uniform component of the deformation – scores for the uniform component are produced by a different analysis (Chapter 6).

Transformation See **Map.**

Two-point shape coordinates See **Bookstein coordinates.**

Type I, Type II error Type I error is invalidly rejecting a true null hypothesis. Type II error is failing to reject a false null hypothesis.

Type 1 landmark A landmark that can be defined in terms of local information, such as a landmark located at the junction of three bones or two bones and a muscle (i.e. anatomical features that meet at a point). There is no need to refer to any distant structures or maxima/minima of curvature. The typology of landmarks is based on Bookstein, 1991. See also **Type 2** and **Type 3 landmarks** (Chapter 2).

Type 2 landmark A landmark defined by a relatively local property, such as the maximum or minimum of curvature of a small bulge or at the endpoint of a structure. These are considered less useful than Type 1 landmarks because their evidence of homology is at least partly geometric rather than purely histological or osteological. See also **Type 1** and **Type 3 landmarks** (Chapter 2).

Type 3 landmark A landmark defined in terms of extremal points, such as the landmark on the rostrum *furthest away from* the foramen magnum. Such landmarks are regarded as deficient because they have one less degree of freedom than they have coordinates (the other degree of freedom is lost when specifying how to locate the landmark). Such landmarks can be used in geometric morphometric studies, but the loss of a degree of freedom must be taken into account when conducting statistical tests. See also **Type 1** and **Type 2 landmarks** (Chapter 2).

Uniform components The components describing the uniform deformation. For two-dimensional configurations, the uniform deformation is described by two components: compression/dilation and shear. The uniform deformation is sometimes considered the zero$^{\text{th}}$ partial warp (Chapter 6).

Uniform deformation A deformation that is purely uniform (or affine), or the purely uniform component of a deformation. The uniform deformations include only the uniform transformations that alter shape (compression/dilation and shear). They do not include transformations that do not alter shape (translation, scaling and rotation). See also **Uniform shape component** (Chapters 5, 6).

Uniform component scores Scores locating a specimen, relative to the reference, along the uniform components. The summed squared scores on the uniform components and partial warps equal the Procrustes distance between each specimen and the reference. Taken together, the uniform and non-uniform scores fully describe the shape difference between the reference and that specimen (Chapter 6).

Vector A set of *P* coordinates that specify the location of a point in *P* dimensions.

Vector space A set of vectors, together with rules for adding and multiplying them (thereby obtaining all permissible linear combinations of them). Addition and scalar multiplication are required to meet eight rules:

1. $\mathbf{X}+\mathbf{Y}=\mathbf{Y}+\mathbf{X}$
2. $\mathbf{X}+(\mathbf{Y}+\mathbf{Z})=(\mathbf{X}+\mathbf{Y})+\mathbf{Z}$
3. A unique zero vector exists such that $\mathbf{X}+\mathbf{0}=\mathbf{X}$, for all $\mathbf{X}$
4. For each $\mathbf{X}$ there exists a unique vector $-\mathbf{X}$ such that $\mathbf{X}+(-\mathbf{X})=\mathbf{0}$
5. $1\mathbf{X}=\mathbf{X}$
6. $(C_1C_2)\mathbf{X}=C_1(C_2\mathbf{X})$
7. $C(\mathbf{X}+\mathbf{Y})=C\mathbf{X}+C\mathbf{Y}$
8. $(C_1+C_2)\mathbf{X}=C_1\mathbf{X}+C_2\mathbf{X}$.

White Book Marcus, L. F., Corti, M., Loy, A. et al. (1996). *Advances in Morphometrics.* Plenum Press. (See also **Black Book, Blue Book, Orange Book** and **Red Book.**)

Bibliography

Alberch, P., Gould, S. J., Oster, G. F. and Wake, D. B. (1979). Size and shape in ontogeny and phylogeny. *Paleobiology*, **5**, 296–317.

Albrecht, G. (1978). Some comments on the use of ratios. *Systematic Zoology*, **27**, 67–71.

Anderson, T. W. (1958). *An Introduction to Multivariate Analysis*. Wiley.

Anstey, R. L. and Pachut, J. F. (1995). Phylogeny, diversity history, and speciation in Paleozoic Bryozoans. In *New Approaches to Speciation in the Fossil Record* (D. H. Erwin and R. L. Anstey, eds) pp. 239–284. Columbia University Press.

Archie, J. W. (1985). Methods for coding variable morphological features for numerical taxonomic analysis. *Systematic Zoology*, **34**, 236–345.

Arnold, S. J. and Phillips, P. C. (1999). Hierarchical comparison of genetic variance–covariance matrices. II. Coastal–island divergence in the garter snake, *Thamnophis elegans*. *Evolution*, **53**, 1516–1527.

Atchley, W. R. and Anderson, D. (1978). Ratios and the statistical analysis of biological data. *Systematic Zoology*, **27**, 71–78.

Atchley, W. R., Gaskins, C. T. and Anderson, D. (1976). Statistical properties of ratios. I. Empirical results. *Systematic Zoology*, **25**, 137–148.

Bartlett, M. S. (1947). Multivariate analysis. *Journal of the Royal Statistical Society, Series B*, **8**, 176–197.

Bastir, M., Rosas, A. and Sheets, H. D. (2004). The morphological integration of the hominoid skull: a partial least squares and PC analysis with morphogenitic implications for European Mid-Pleistocene mandibles. In *Developments in Primatology: Progress and Prospects* (D. Slice, ed.), in press. Kluwer Academic/Plenum Press.

Berg, R. L. (1960). The ecological significance of correlation pleiades. *Evolution*, **14**, 171–180.

Bininda-Emonds, O. R. P., Jeffrey, J. E. and Richardson, M. K. (2003). Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development. *Proceedings of the Royal Society of London, Series B*, **270**, 341–346.

Bookstein, F. L. (1982). The geometric meaning of soft modeling with some generalizations. In *Systems Under Indirect Observation: Causality–Structure–Prediction* (K. G. Jöreskog and H. Wold, eds) pp. 55–74. North Holland Publishing Company.

Bookstein, F. L. (1986). Size and shape spaces for landmark data in two dimensions. *Statistical Science*, **1**, 181–242.

Bookstein, F. L. (1989a). Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 567–585.

Bookstein, F. L. (1989b). "Size and shape": a comment on semantics. *Systematic Zoology*, **38**, 173–190.

Bookstein, F. L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.

Bookstein, F. L. (1996a). Combining the tools of geometric morphometrics. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 131–152. Plenum Press.

Bookstein, F. L. (1996b). Standard formula for the uniform shape component in landmark data. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 153–168. Plenum Press.

Bookstein, F. L. (1997a). Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Medical Image Analysis*, **1**, 225–243.

ISBN 0–12–77846–08

Copyright © 2004 Elsevier Ltd
All rights of reproduction in any form reserved

Bookstein, F. L. (1997b). Shape and the information in medical images: a decade of the morphometric synthesis. *Computer Vision and Image Understanding*, **66**, 97–118.

Bookstein, F. L., Chernoff, B., Elder, R. et al. (1985). *Morphometrics in Evolutionary Biology*. The Academy of Natural Sciences of Philadelphia.

Bookstein, F. L., Gunz, P., Ingeborg, H. et al. (2003). Cranial integration in *Homo*: singular warps analysis of the midsagittal plane in ontogeny and evolution. *Journal of Human Evolution*, **44**, 167–187.

Campbell, N. A. and Atchley, W. R. (1981). The geometry of canonical variates analysis. *Systematic Zoology*, **30**, 268–280.

Chapman, R. E. (1990). Conventional Procrustes methods. In *Proceedings of the Michigan Morphometrics Workshop* (F. J. Rohlf and F. L. Bookstein, eds) pp. 251–267. University of Michigan Museum of Zoology.

Chappill, J. A. (1989). Quantitative characters in phylogenetic analysis. *Cladistics*, **5**, 217–234.

Chatfield, C. and Collins, A. J. (1980). *Introduction to Multivariate Analysis*. Chapman & Hall.

Cheverud, J. M. (1982). Phenotypic, genetic and environmental integration in the cranium. *Evolution*, **36**, 499–512.

Cheverud, J. M. (1984). Quantitative genetics and developmental constraints on evolution by selection. *Journal of Theoretical Biology*, **110**, 155–172.

Cheverud, J. M. (1995). Morphological integration in the saddle-back tamarin (*Saguinus fuscicollis*) cranium. *American Naturalist*, **145**, 63–89.

Ciampaglio, C. N. (2002). Determining the role that ecological and developmental constraints play in controlling disparity: examples from the crinoid and blastozoan fossil record. *Evolution & Development*, **4**, 170–188.

Ciampaglio, C. N., Kemp, M. and McShea, D. W. (2001). Detecting changes in morphospace occupation patterns in the fossil record: characterization and analysis of measures of disparity. *Paleobiology*, **27**, 695–715.

Colless, D. H. (1980). Congruence between morphometric and allozyme data for *Menidia* species: a reappraisal. *Systematic Zoology*, **29**, 288–299.

Cope, E. D. (1887). *The Origin of the Fittest*. McMillan.

Corruccini, R. S. (1977). Correlation properties of morphometric ratios. *Systematic Zoology*, **26**, 211–214.

Corti, M. C., Fadda, C., Simson, S. and Nevo, E. (1996). Size and shape variation in the mandible of the fossorial rodent *Spalax ehrenbergi*. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 303–320. Plenum Press.

Dodson, P. (1975). Relative growth in two sympatric species of *Sceloperus*. *American Midland Naturalist*, **94**, 421–450.

Dodson, P. (1978). On the use of ratios in growth studies. *Systematic Zoology*, **27**, 62–67.

Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*. John Wiley & Sons.

Dryden, I. L. and Walker, G. (1999). Highly resistant regression and object matching. *Biometrics*, **55**, 820–825.

Edgington, E. S. (1995). *Randomization Tests*. Marcel Dekker.

Efron, B. (1979). Computers and the theory of statistics, thinking the unthinkable. *Society for Industrial and Applied Mathematics Review*, **21**, 460–480.

Efron, B. (1987). Better bootstrap confidence-intervals. *Journal of the American Statistical Association*, **82**, 171–185.

Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society, Series B: Methodological*, **54**, 83–127.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Fadda, C., Faggiani, F. and Corti, M. (1997). A portable device for the three-dimensional landmark collection of skeletal elements of small mammals. *Mammalia*, **61**, 622–627.

Falsetti, A. B. and Cole, T. M. (1992). Relative growth of the postcranial skeleton in callitrichines. *Journal of Human Evolution*, **23**, 79–92.

Farris, J. S. (1990). Phenetics in camouflage. *Cladistics*, **6**, 91–100.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications*. John Wiley & Sons.

Felsenstein, J. (2002). Quantitative characters, phylogenies, and morphometrics. In *Morphology, Shape and Phylogeny* (N. MacLeod and P. L. Forey, eds) pp. 27–44. Taylor & Francis.

Ferson, S., Rohlf, F. J. and Koehn, R. K. (1985). Measuring shape variation of two-dimensional outlines. *Systematic Zoology*, **34**, 59–68.

Fink, W. L. (1993). Revision of the piranha genus *Pygocentrus* (Teleostei, Characiformes). *Copeia*, **1993**, 665–687.

Fink, W. L. and Zelditch, M. L. (1995). Phylogenetic analysis of ontogenetic shape transformations: a reassessment of the piranha genus *Pygocentrus* (Teleostei). *Systematic Biology*, **44**, 343–360.

Flury, B. (1988). *Common Principal Components and Related Multivariate Methods*. John Wiley.

Foote, M. (1986). Developmental buffering as a mechanism for stasis. *Evolution*, **42**, 396–399.

Foote, M. (1990). Nearest-neighbor analysis of trilobite morphospace. *Systematic Zoology*, **39**, 371–382.

Foote, M. (1992). Paleozoic record of morphological diversity in blastozoan echinoderms. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 7325–7329.

Foote, M. (1993a). Contributions of individual taxa to overall morphological disparity. *Paleobiology*, **19**, 403–419.

Foote, M. (1993b). Discordance and concordance between morphological and taxonomic diversity. *Paleobiology*, **19**, 185–204.

Foote, M. (1994). Morphological disparity in Ordovician–Devonian crinoids and the early saturation of morphological space. *Paleobiology*, **20**, 320–344.

Foote, M. (1997). The evolution of morphological diversity. *Annual Reviews of Ecology and Systematics*, **28**, 129–152.

Foote, M. and Gould, S. J. (1992). Cambrian and Recent morphological disparity. *Science*, **258**, 816.

Fuiman, L. A. (1983). Growth gradients in fish larvae. *Journal of Fish Biology*, **23**, 117–123.

Galis, F. and Metz, J. A. (2001). Testing the vulnerability of the phylotypic stage: on modularity and evolutionary conservatism. *Journal of Experimental Zoology (Molecular and Developmental Evolution)*, **291**, 195–204.

Galis, F., van Alphen, J. M. and Metz, J. A. (2001). Why five fingers? Evolutionary constraints on digit numbers. *Trends in Ecology and Evolution*, **16**, 637–646.

Gavrilets, S. (1999). Dynamics of clade diversification on the morphological hypercube. *Proceedings of the Royal Society of London, Series B*, **266**, 817–824.

Gift, N. and Stevens, P. F. (1997). Vagaries in the delimitation of character states in quantitative variation – an experimental study. *Systematic Biology*, **46**, 112–125.

Goldman, N. (1988). Methods for discrete coding of morphological characters for numerical analysis. *Cladistics*, **4**, 59–71.

Good, P. (1994). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag.

Gould, S. J. (1977). *Ontogeny and Phylogeny*. Harvard University Press.

Gould, S. J. (1984). Morphological channeling by structural constraint: convergence in styles of dwarfing and gigantism in *Cerion*, with a description of two new fossil species and a report on the discovery of the largest *Cerion*. *Paleobiology*, **10**, 172–194.

Gould, S. J. and Garwood, R. A. (1969). Levels of integration in mammalian dentitions: an analysis of correlations in *Nesophantes micrus* (Insectivora) and *Oryzomys couesi* (Rodentia). *Evolution*, **23**, 276–300.

Green, W. D. K. (1996). The thin-plate spline and images with curving features. In *Proceedings in Image Fusion and Shape Variability Techniques* (K. V. Mardia, C. A. Gill and I. L. Dryden, eds) pp. 79–87. Leeds University Press.

Hall, B. K. (1992). *Evolutionary Developmental Biology*. Chapman & Hall.

Hills, M. (1978). On ratios – a response to Atchley, Gaskins and Anderson. *Systematic Zoology*, **27**, 61–62.

Hingst-Zaher, E., Marcus, L. F. and Cerqueria, R. (2000). Application of geometric morphometrics to the study of postnatal size and shape changes in the skull of *Callomys expulsus*. *Hystrix*, **11**, 99–113.

Hoeffding, W. (1952). The large-sample power of tests based on permutation of observations. *Annals of Mathematical Statistics*, **23**, 169–192.

Houle, D., Mezey, J. and Galpern, P. (2002). Interpretation of the results of common principal components analysis. *Evolution*, **56**, 433–440.

Hulsey, C. D. and Wainwright, P. C. (2002). Projecting mechanics into morphospace: disparity in the feeding mechanics of labrid fishes. *Proceedings of the Royal Society of London, Series B*, **269**, 317–326.

Huxley, J. S. (1932). *Problems of Relative Growth*. MacVeagh.

Jackson, D. A. and Somers, K. M. (1989). Are probability estimates from the permutation models of Mantel's test stable? *Canadian Journal of Zoology*, **67**, 766–779.

Jolicoeur, P. (1963). The multivariate generalization of the allometry equation. *Biometrics*, **19**, 497–499.

Jöreskog, K. G. and Wold, H. (1982). *Systems Under Indirect Observation: Causality–Structure–Prediction*. North Holland Publishing Company.

Katz, J. M. (1980). Allometry formula: a cellular model. *Growth*, **44**, 89–96.

Kendall, D. (1977). The diffusion of shape. *Advances in Applied Probability*, **9**, 428–430.

Kendall, D. G. and Kendall, W. S. (1980). Alignments in two-dimensional random sets of points. *Advances in Applied Probability*, **12**, 380–424.

Kim, K., Sheets, H. D., Haney, R. A. and Mitchell, C. E. (2002). Morphometric analysis of ontogeny and allometry of the Middle Ordovician trilobite, *Triarthrus becki*. *Paleobiology*, **28**, 364–377.

Kimmel, C. B., Ballard, W. W., Kimmel, S. R. et al. (1995). Stages of embryonic development of the zebrafish. *Developmental Dynamics*, **203**, 253–310.

Kingsolver, J. G. and Wiernasz, D. C. (1991). Development, function, and the quantitative genetics of wing melanin pattern in *Pieris* butterflies. *Evolution*, **45**, 1480–1492.

Klingenberg, C. P. (1998). Heterochrony and allometry: the analysis of evolutionary change in ontogeny. *Biological Reviews*, **73**, 79–123.

Klingenberg, C. P. and Ekau, W. (1996). A combined morphometric and phylogenetic analysis of an ecomorphological trend: Pelagization in Antarctic fishes (Perciformes: Nototheniidae). *Biological Journal of the Linnean Society*, **59**, 143–177.

Klingenberg, C. P. and Froese, R. (1991). A multivariate comparison of allometric growth patterns. *Systematic Zoology*, **40**, 410–419.

Klingenberg, C. P., Badayaev, A. V., Sowry, S. M. and Beckwith, N. J. (2001). Inferring developmental modularity from morphological integration: analysis of individual variation and asymmetry in bumblebee wings. *American Naturalist*, **157**, 11–23.

Kluge, A. G. and Kerfoot, C. (1973). The predictability and regularity of character divergence. *American Naturalist*, **107**, 426–464.

Lagler, K. F., Bardach, J. E. and Miller, R. R. (1962). *Ichthyology*. John Wiley & Sons.

Laird, A. K. (1965). Dynamics of relative growth. *Growth*, **29**, 249–263.

Laird, A. K., Barton, A. D. and Tyler, S. A. (1968). Growth and time: an interpretation of allometry. *Growth*, **32**, 347–354.

Lande, R. (1979). Quantitative genetic analysis of multivariate evolution, applied to brain : body size allometry. *Evolution*, **33**, 402–416.

Lande, R. (1980). The genetic covariance between characters maintained by pleiotropic mutations. *Genetics*, **94**, 314–334.

Lande, R. and Arnold, S. J. (1983). The measurement of selection on correlated characters. *Evolution*, **37**, 1210–1226.

Liebner, D. L. and Sheets, H. D. (2001). Superposer, available on the IMP website www.canisius,edu/~sheets/morphsoft.html

Lohman, G. P. and Schweitzer, P. N. (1990). On eigenshape analysis. In *Proceedings of the Michigan Morphometrics Workshop* (F. J. Rohlf and F. L. Bookstein, eds) pp. 147–166. University of Michigan Museum of Zoology, Special Publication No. 2.

Lowe, A. A., Özbeck, M. M., Miyamoto, K. and Fleetham, J. A. (1997). Cephalometric and demographic characteristics of sleep apnea: an evaluation with partial least squares analysis. *The Angle Orthodontist*, **67**, 143–154.

Lundrigan, B. (1996). Morphology of horns and fighting behavior in the family Bovidae. *Journal of Mammalogy*, **77**, 462–475.

MacLeod, N. and Rose, K. D. (1993). Inferring locomotor behavior in Paleogene mammals via eigenshape analysis. *American Journal of Science*, **293A**, 300–355.

Manly, B. F. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall.

Marcus, L. F., Hingst-Zaher, E. and Zaher, H. (2000). Applications of landmark morphometrics to skulls representing the orders of living mammals. *Hystrix* (n.s.), **11**, 24–48.

Marroig, G. and Cheverud, J. M. (2001). A comparison of phenotypic variation and covariation patterns and the role of phylogeny, ecology and ontogeny during cranial evolution of New World monkeys. *Evolution*, **55**, 2576–2600.

Maynard Smith, J. M., Burian, R., Kauffman, S. et al. (1984). Developmental constraints and evolution. *Quarterly Review of Biology*, **60**, 265–287.

McKinney, M. L. and McNamara, K. J. (1991). *Heterochrony: The Evolution of Ontogeny*. Plenum Press.

Miller, A. I. and Foote, M. (1996). Calibrating the Ordovician radiation of marine life: implications for Phanerozoic diversity trends. *Paleobiology*, **22**, 304–309.

Morrison, D. F. (1990). *Multivariate Statistical Methods*. McGraw-Hill.

Myers, P., Lundrigan, B. L., Gillespie, B. W. and Zelditch, M. L. (1996). Phenotypic plasticity in skull and dental morphology in the Prairie Deer Mouse (*Peromyscus maniculatus bairdii*). *Journal of Morphology*, **229**, 229–237.

Olson, E. C. and Miller, R. L. (1958). *Morphological Integration*. University of Chicago Press.

Oxnard, C. E. (1968). The architecture of the shoulder in some mammals. *Journal of Morphology*, **126**, 249–290.

Phillips, P. C. and Arnold, S. J. (1999). Hierarchical comparison of genetic variance–covariance matrices. I. Using the Flury hierarchy. *Evolution*, **53**, 1506–1515.

Polly, P. D. (2000). Geography and sample size in **P** matrix evolution: molar shape change in island populations of *Sorex araneus*. *Journal of Evolutionary Biology*.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1998). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society B*, **11**, 18–14.

Raff, R. A. (1992). Direct-developing sea urchins and the evolutionary reorganization of early development. *BioEssays*, **14**, 211–218.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. John Wiley & Sons.

Richardson, M. K., Hanken, J., Gooneratne, M. L. et al. (1997). There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anatomy and Embryology*, **196**, 91–106.

Robinson, J. (1973). Large-sample power of permutation tests for randomization models. *Annals of Statistics*, **1**, 291–296.

Rohlf, F. J. (1990). Rotational fit (Procrustes) methods. In *Proceedings of the Michigan Morphometrics Workshop* (F. J. Rohlf and F. L. Bookstein, eds) pp. 227–236. University of Michigan Museum of Zoology, Special Publication No. 2.

Rohlf, F. J. (1998). On applications of geometric morphometrics to studies of ontogeny and phylogeny. *Systematic Biology*, **47**, 147–158.

Rohlf, F. J. (2000). On the use of shape spaces to compare morphometric methods. *Hystrix* (n.s.), **11**, 8–24.

Rohlf, F. J. (2002). Geometric morphometrics and phylogeny. In *Morphology, Shape and Phylogeny* (N. MacLeod and P. L. Forey, eds) pp. 175–193. Taylor & Francis.

Rohlf, F. J. and Archie, J. W. (1984). A comparison of Fourier methods for the description of wing shape in mosquitoes (Diptera: Culicidae). *Systematic Zoology*, **33**, 302–317.

Rohlf, F. J. and Bookstein, F. L. (2003). Computing the uniform component of shape variation. *Systematic Biology*, **52**, 66–69.

Rohlf, F. J. and Corti, M. (2000). Use of two-block partial least squares to study covariation in shape. *Systematic Biology*, **49**, 740–753.

Rohlf, F. J. and Slice, D. E. (1990). Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Zoology*, **39**, 40–59.

Rohlf, F. J., Gilmartin, A. J. and Hart, G. (1983). The Kluge–Kerfoot phenomenon: a statistical artifact? *Evolution*, **37**, 180–202.

Romano, J. P. (1989). Bootstrap and randomization tests of some non-parametric hypotheses. *Annals of Statistics*, **17**, 141–159.

Roth, V. L. (1993). On three-dimensional morphometrics, and on the identification of landmark points. In *Contributions to Morphometrics* (L. F. Marcus, E. Bello and A. García-Valdecasas, eds), pp. 41–62. Museo Nacional de Ciencias Naturales, Madrid.

Roy, K. and Foote, M. (1997). Morphological approaches to measuring biodiversity. *Trends in Ecology & Evolution*, **12**, 277–281.

Rüber, L. and Adams, D. C. (2001). Evolutionary convergence of body shape and trophic morphology in cichlids of Lake Tanganyika. *Journal of Evolutionary Biology*, **14**, 325–332.

Sampson, P. D., Streissguth, A. P., Barr, H. M. and Bookstein, F. L. (1989). Neurobehavioral effects of prenatal alcohol: Part II. Partial least squares analysis. *Neurotoxicology and Teratology*, **11**, 477–491.

Sampson, P. D., Bookstein, F. L., Sheehan, F. H. and Bolson, E. L. (1996). Eigenshape analysis of left ventricular outlines from contrast ventriculograms. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 211–233. Plenum.

Sander, K. (1983). The evolution of patterning mechanisms: gleanings from insect embryogenesis and spermatogenesis. In *Development and Evolution* (B. C. Goodwin, N. Holder and C. C. Wylie, eds) pp. 137–159. Cambridge University Press.

Schaefer, S. A. and Lauder, G. V. (1996). Testing historical hypotheses of morphological change: biomechanical decoupling in loricariod catfishes. *Evolution*, **50**, 1661–1675.

Schluter, D. (1996). Adaptive radiation along genetic lines of least resistance. *Evolution*, **50**, 1766–1774.

Seidl, F. (1960). Körpergrundgestalt und Keimstruktur. Eine Erörterung über die Gundlagen der vergleichenden und experimentellen Embryologie un deren Gültigkeit bei phylogeneticschen Überlegungen. *Zoologische Anzeiger*, **164**, 245–305.

Shea, B. T. (1992). Ontogenetic scaling of skeletal proportions in the talapoin monkey. *Journal of Human Evolution*, **23**, 283–307.

Shea, B. T. (2002). Are some heterochronic transformations likelier than others? In *Human Evolution Through Developmental Change* (N. Minugh-Purvis and K. J. McNamara, eds) pp. 79–101. Johns Hopkins Press.

Sheets, H. D. and Mitchell, C. E. (2001). Why the null matters: statistical tests, random walks and evolution. *Genetica*, **112**, 105–125.

Siegel, A. F. and Benson, R. H. (1982). A robust comparison of biological shapes. *Biometrics*, **38**, 341–350.

Simon, C. (1983). A new coding procedure for morphometric data with an example from periodical cicada wing veins. In *Numerical Taxonomy* (J. Felsenstein, ed.) pp. 378–382. Springer-Verlag.

Slack, J. M., Holland, P. W. and Graham, C. F. (1993). The zootype and the phylotypic stage. *Nature*, **361**, 490–492.

Slice, D. E. (2001). Landmark coordinates aligned by Procrustes analysis do not lie in Kendall's shape space. *Systematic Biology*, **50**, 141–149.

Slice, D. E., Bookstein, F. L., Marcus, L. F. and Rohlf, F. J. (1996). Appendix I: A glossary for geometric morphometrics. In *Advances in Morphometrics* (L. F. Marcus, M. Corti, A. Loy et al., eds) pp. 531–551. Plenum Press.

Small, C. G. (1996). *The Statistical Theory of Shape*. Springer.

Smith, L. H. and Lieberman, B. S. (1999). Disparity and constraint in olenelloid trilobites and the Cambrian radiation. *Paleobiology*, **25**, 459–470.

Snedecor, G. W. and Cochran, W. G. (1967). *Statistical Methods*. Iowa State University Press.

Sokal, R. R. (1976). The Kluge–Kerfoot phenomenon re-examined. *American Naturalist*, **110**, 1077–1091.

Sokal, R. R. and Rohlf, F. J. (1995). *Biometry: The Principals and Practice of Statistics in Biological Research*, 3rd edn. Freeman.

Spencer, M. A. and Spencer, G. S. (1995). Video-based three-dimensional morphometrics. *American Journal of Physical Anthropology*, **96**, 443–453.

Stein, B. R. (1981). Comparative limb mycology of two opossums, *Didelphis* and *Chironectes*. *Journal of Morphology*, **169**, 113–140.

Steppan, S. J. (1997). Phylogenetic analysis of phenotypic covariance structure. I. Contrasting results from matrix correlation and common principal component analysis. *Evolution*, **51**, 571–586.

Strauss, D. and Sadler, P. M. (1989). Classical confidence intervals and Bayesian probability estimates for ends of local taxon ranges. *Mathematical Geology*, **21**, 411–427.

Strauss, R. E. (1984). Allometry and functional feeding morphology in haplochromine cichlids. In *Evolution of Fish Species Flocks* (A. A. Echelle and I. Kornfield, eds) pp. 217–229. University of Maine.

Strauss, R. E. and Bookstein, F. L. (1982). The truss – body form reconstructions in morphometrics. *Systematic Zoology*, **31**, 113–135.

Strauss, R. E. and Fuiman, L. A. (1985). Quantitative comparisons of body form and allometry in larval and adult Pacific sculpins (Teleostei: Cottidae). *Canadian Journal of Zoology*, **63**, 1582–1589.

Streissguth, A. P., Bookstein, F. L., Sampson, P. D. and Barr, H. M. (1993). *The Enduring Effects of Prenatal Alcohol Exposure on Child Development: Birth Through Seven Years, A Partial Least Squares Solution*. University of Michigan Press.

Swiderski, D. L. (1993). Morphological evolution of the scapula in tree squirrels, chipmunks, and ground squirrels (Sciuridae): an analysis using thin-plate splines. *Evolution*, **47**, 1854–1873.

Swiderski, D. L., Zelditch, M. L. and Fink, W. L. (1998). Why morphometrics isn't special: coding quantitative data for phylogenetic analysis. *Systematic Biology*, **47**, 508–519.

Taylor, M. E. (1974). The functional anatomy of the forelimb of some African Viverridae (Carnivora). *Journal of Morphology*, **143**, 307–336.

Thiele, K. (1993). The holy grail of the perfect character: the cladistic treatment of morphometric data. *Cladistics*, **9**, 275–304.

Thompson, D'Arcy W. (1942). *On Growth and Form: A New Edition*. Cambridge University Press. (Reprinted in 1992 as *On Growth and Form: The Complete Revised Edition*, Dover Publications.)

Tukey, J. W. (1958). Bias and confidence in not quite large samples (Abstract). *Annals of Mathematical Statistics*, **29**, 614.

van Snik, G. M. J., van den Boogaart, J. G. M. and Osse, W. M. (1997). Larval growth patterns in *Cyprinus carpio* and *Clarias gariepinus* with attention to the finfold. *Journal of Fish Biology*, **50**, 1339–1352.

Van Valen, L. (1962). Developmental gradients in the dentition of *Peromyscus*. *Evolution*, **16**, 272–277.

Van Valen, L. (1970). An analysis of developmental fields. *Developmental Biology*, **23**, 456–477.

Voss, R. S. and Marcus, L. F. (1992). Morphological evolution in muroid rodents.2. Craniometric factor divergence in 7 neotropical genera, with experimental results from *Zygodontomys*. *Evolution*, **46**, 1918–1934.

Wagner, G. P. (1988). The influence of variation and of developmental constraints on the rate of multivariate phenotypic evolution. *Journal of Evolutionary Biology*, **1**, 45–66.

Wagner, G. P. and Altenberg, L. (1996). Complex adaptations and the evolution of evolvability. *Evolution*, **50**, 967–976.

Wagner, P. J. (1995). Testing evolutionary constraint hypotheses with early Paleozoic gastropods. *Paleobiology*, **21**, 248–272.

Wagner, P. J. (1997). Patterns of morphologic diversification among the Rostroconchia. *Paleobiology*, **23**, 115–150.

Walker, J. A. (2000). Ability of geometric morphometric methods to estimate a known covariance matrix. *Systematic Biology*, **49**, 686–696.

Wayne, R. K. (1986). Cranial morphology of domestic and wild canids: the influence of development on morphological change. *Evolution*, **40**, 243–261.

Webster, M., Sheets, H. D. and Hughes, N. C. (2001). Allometric patterning in trilobite ontogeny: testing for heterochrony in *Nephrolenellus*. In *Beyond Heterochrony: The Evolution of Development* (M. L. Zelditch, ed.) pp. 105–142. John Wiley & Sons.

White, J. F. and Gould, S. J. (1965). Interpretation of the coefficients in the allometric equation. *American Naturalist*, **99**, 5–18.

Wills, M. A. (2001). Morphological disparity: a primer. In *Fossils, Phylogeny, and Form: An Analytical Approach* (J. M. Adrain, G. D. Edgecombe and B. S. Lieberman, eds) pp. 55–144. Kluwer Academic/Plenum Publishers.

Wills, M. A., Briggs, D. E. G. and Fortey, R. A. (1994). Disparity as an evolutionary index – a comparison of Cambrian and recent arthropods. *Paleobiology*, **20**, 93–130.

Zelditch, M. L. (1988). Ontogenetic variation in patterns of phenotypic integration in the laboratory rat. *Evolution*, **42**, 28–41.

Zelditch, M. L. and Carmichael, A. C. (1989). Ontogenetic variation in patterns of developmental and functional integration in skulls of *Sigmodon fulviventer*. *Evolution*, **43**, 814–824.

Zelditch, M. L., Bookstein, F. L. and Lundrigan, B. L. (1992). Ontogeny of integrated skull growth in the cotton rat *Sigmodon fulviventer*. *Evolution*, **46**, 1164–1180.

Zelditch, M. L., Bookstein, F. L. and Lundrigan, B. L. (1993). The ontogenetic complexity of developmental constraints. *Journal of Evolutionary Biology*, **6**, 121–141.

Zelditch, M. L., Fink, W. L. and Swiderski, D. L. (1995). Morphometrics, homology and phylogenetics: quantified characters as synapomorphies. *Systematic Biology*, **44**, 179–189.

Zelditch, M. L., Fink, W. L., Swiderski, D. L. and Lundrigan, B. L. (1998). On applications of geometric morphometrics to studies of ontogeny and phylogeny: a reply to Rohlf. *Systematic Biology*, **47**, 159–167.

Zelditch, M. L., Sheets, H. D. and Fink, W. L. (2000). Spatiotemporal reorganization of growth rates in the evolution of ontogeny. *Evolution*, **54**, 1363–1371.

Zelditch, M. L., Sheets, H. D. and Fink, W. L. (2001). The spatial complexity and evolutionary dynamics of growth. In *Beyond Heterochrony: The Evolution of Development* (M. L. Zelditch, ed.) pp. 145–194. John Wiley & Sons.

Zelditch, M. L., Sheets, H. D. and Fink, W. L. (2003b). The ontogenetic dynamics of shape disparity. *Paleobiology*, **29**, 139–156.

Zelditch, M. L., Lundrigan, B. L., Sheets, H. D. and Garland, T. Jr (2003b). Do precocial mammals develop at a faster rate? A comparison of rates of skull development in *Sigmodon fulviventer* and *Mus musculus domesticus*. *Journal of Evolutionary Biology*, **16**, 708–720.

Index

Affine transformations, *see* Uniform transformations
Allometry, 56, 58, 322, 324, 325–58
  allometric scaling, 325
  coefficients
    estimation of, from traditional morphometric data, 326–9
    estimation of, from geometric data, 350–1
    interpretation of, 330–3, 350–1
  comparing ontogenetic allometries
    traditional morphometric data, 333–347
    geometric data, 350–5
  evolutionary, 322, 323
  ontogenetic, 325, 326–33
  software, 255–6
  standardizing to remove effects of, *see* Standardization by regression
  transpositional, 341
  see also Regression
ANCOVA, 217
Anisotropy, 61, 62
ANOVA, 212–17
  graphic representation, 213
Aspect ratio of triangle, 59

Bartlett's test for differences in Wilk's lambda, 178
Baseline, 54–5, 55–7
  choice of, 56–7
  interpreting shape variables relative to, 58–60
  sliding, 109–113
Baseline registration, *see* Bookstein shape coordinates; sliding baseline registration
Basis of a vector space, 125
  orthonormal basis, 163–4
  see also Eigenanalysis
Bending energy, 133–4, 146–9, *see also* Non-uniform transformations, partial warps
Bending energy matrix, *see* Bending energy
BigFix program, 69, 71
Biorthogonal directions, 61
Bivariate analysis, 190, 232–4
BMP format for image files, 45
Bonferroni adjustment, 65, 215
Bookstein shape coordinates, 31, 51–72
  alternatives to, 105–9
  as shape variables, 51–5
  choice of baseline, 56–7
  for multiple landmarks, 65–8
  principal axes, 61
  shape variables from, 51–5
  shape differences described by, 58–60
  software, 69
  statistics, 57–8
  variables implied by principal axes, 62–5
Bootstrap estimate, 196
Bootstrapping, 195–9, 223, 225
Box truss, 3–5
Brightness, image, 47

Canonical correlation analysis, 264–5
Canonical variates analysis, 15, 155, 170–80
  algebraic description, 174–6
  classification of groups by, 179
  geometric description, 171–4
  groups and grouping variables, 170
  interpretation of results, 176–80
  relationship to principal components analysis, 155, 170–1
  testing the statistical significance of canonical variates, 178–9
  software, 184–6, 224
  using for taxonomic discrimination, 365–7
Cartesian coordinates, 76
CCA, *see* Canonical correlation analysis
CCoder program, 186–7
Centered configuration matrix, 77, 79–80, 83, 84, 86, 89, 90, 91, 93, 95, 98, 388, 390
Centering a configuration matrix, 89
Centroid position, 77, 109
Centroid size, 11, 13, 56, 78
  calculating, 56
  geometric depiction of the calculation, 56
  radial notion of scale, 56
  as uncorrelated with shape, 56
  as the independent variable in studies of allometry, 236–237, 240, 254
Characteristic equation, 163–4, 176, *see also* Eigenvalue
Characters (Phylogenetic), 367–8
  unsuitability of partial warps for, 369–72
  using comparisons between vectors to find, 376–8

Characters (*contd*)
  using PCA to find, 372–6
Circle construction of principal axes, 61–2
Clock model, Gould's for analysis of heterochrony, 339
Clustering, 307
Coding phylogenetic characters, 378–9
Comparing means, *see* ANOVA, MANOVA and specific test statistics
Comparing patterns of (co)variance, 308–14
Complex numbers, 137–8
Compression/dilation, 137–42
Computer-based statistical methods, 189–208
  basic statistical concepts, 190–1
  constructing confidence intervals, 191–2
  hypothesis testing, 193–4
  number of repetitions, 205–6
  power, 205
  resampling-based methods
    bootstrap, 195–9
    jackknife, 200–1
    Monte Carlo simulations, 201–2
    permutation tests, 199–200
Configuration of landmarks, 75
Configuration matrix, 76–9
  configuration space, 77
  position or location, 77–8
  size, 78–9
Configuration spaces, 77, 390–1
Consensus configuration, *see* Reference configuration
Consistency of relative position of landmarks, 27–8
Continuous functions and models of deformations, 132
Continuous variables, 190
Contraction, 413
Contrast, 47
CoordGen program, 69–70
Coplanarity of landmarks, 30–1
Correlation, 413
Correlation coefficient, 235
Correspondence of landmarks, 25
Covariance, 217, 413
Criteria for selecting landmarks, *see* Landmarks, criteria for choosing
Curves without landmarks, 395–6
CVA, *see* Canonical variates analysis
CVAGen program, 184–6, 224

Deformation, 131–2
  decomposition of, 134–150, 391–2
    see also Thin-plate spline, uniform transformations, non-uniform transformations, partial warps
Degrees of freedom, 88, 97–8, 214, 215
Depth of field, 43
Digitization of images, 48–9
Discontinuous functions and models of deformations, 132
Discontinuous variables, 210
Discrete variables, 190
Discriminant function, *see* Canonical variates analysis
Disparity, 293–5
  confidence intervals, 298–302
  hypotheses explaining temporal patterns, 294
  measurement, 297–8
  metrics, 296–7
  multi-group analysis, 316–17
  ontogenetic and interclade comparisons, 300–2
  partial, 302
  relationship to variation, 293
  software, 314–17
    DisparityBox, 314–17
    SpaceAngle, 317
  structure, analysis of
    nearest-neighbor analysis, 303–8
    patterns of (co)variance, 308–14
DisparityBox program, 314–17
Distance, 81, *see also* Euclidean distance, full Procrustes distance, Mahalanobis distance, partial Procrustes distance
Dot (inner) product, between two vectors, 251–2
Dummy codes, using in regression, 240

Edge registration, *see* Baseline registration
Eigenanalysis, 146
Eigenvalues, *see* Eigenvectors
Eigenvectors, 163–4, 167
  eigenvectors of the bending energy matrix (partial warps), 146, 149
  relationship to singular axes, 262, 263
  see also Principal components, partial warps
Electronically available resources, *see* Web resources
Elements of matrix, 77
Error, 193, 230, 232
Error function, 232
Euclidean approximation to Kendall's shape space, 87, 95–6
Euclidean distance, 97, 297
Euclidean space, 86, 87, 95, 96, 149, 392
Explicit uniform components, 136, 137
Explicit uniform deformations, *see* Explicit uniform components
Explicit uniform term, *see* Explicit uniform components

F-test, 197, 212–14, 219
Fiber, 81–3

Flury hierarchy, 308
Focus, 43
Full Procrustes distance, 85, 115, 390
Full Procrustes superimposition, 94, 96, 99, 115

Gaussian distribution, 191, 305
Generalized least squares Procrustes superimposition, 113–19, 298
Geodesic distance, *see* Partial Procrustes distance
Geometric shape, 11, 72–3
GLS, *see* Generalized least squares Procrustes superimposition
Goodall's *F*-test, 117, 122, 130, 221, 225, 236
Grand mean, 213, 298–9
Great circle distance, *see* Partial Procrustes distance
Grouping variable in MANOVA, 171
Growth:
  decoupling from development, 337–9
  describing changes in shape related to, *see* Ontogenetic trajectory
  profiles, 332

Heterochrony, 321, 323, 324
  formalisms for, 339–41
Homology of landmarks, 25–7
Hotelling-Lawley trace, 219
Hotelling's T^2-test, 58, 65, 112, 218, 224, 240
Hyperplanes, 309
  largest possible angle between, 309–12
Hypersphere, 80

Image acquisition, 39
  camera
    aperture, 39–40
    image capture, 44–5
    lens, 40–4
  digitization, 48–9
  image formation, 40
  improving image, 46–7
  magnification, 42
  saving image, 45–6
  system checks, 44
Implicit uniform deformations, 136, 142–3
Implicit uniform terms, *see* Implicit uniform deformations
Induced correlation, 111–12
Induced covariance, 110, 112, 119
Influential observations, 119
Isometry, 240, 331, 337, 338
Isotropic change, 62

Jackknife test, 200–1, 417
JPEG format for image files, 45

Kendall's shape space:
  Three-dimensional landmark configurations, in, 390, 395–6, 399, 402
  Two-dimensional landmark configurations, in, 86–95
"Kluge–Kerfoot" phenomenon, 296

Lagrange multiplier, 165, 176
Landmark configuration, 14–15, 76
Landmarks, 6, 7, 23–50
  Bookstein's typology of, 31
  digitizing, 48–9
  criteria for choosing, 24–31
    adequate coverage of the form, 28–9
    consistency of relative position, 27–8
    coplanarity of, 30–1
    homology, 25–7
    repeatability of, 30
    applied to specific cases
      external body form, piranha, 34–7
      scapula, squirrel, 32–4
      skull, rodent, 37–39
  configuration of, 14–5, 75
  curves without, 395–6
  definition of, 75
  differential weighting of, 399
  image acquisition, manipulation and enhancement, 39–47
  superimposing configurations of, 387–9
  three-dimensional configuration of, 386–7
  software, for digitizing, 48
Latent variables, 262
Least squares, 119, 232–3
Linear regression, 230, 262
Linear combinations, 5, 155–6, 162, 164, 170, 264
Linearity and regression, 236–40

Mahalanobis distance, 177–8
MakeFan program, 406
MANCOVA, 58, 217, 336
MANOVA, 58, 65, 209–28, 210–23
  analytic techniques
    categorical with continuous variable, 217–18
    complex trait with categorical variable, 218–22
    simple trait with more than two groups, 214–15
    simple trait with two groups, 210–14
    two or more categorical variables, 215–17
  software, 223–8
  see also Computer-based statistical methods, *F*-test, Goodall's *F*-test

Means, comparing statistically, *see* ANOVA, MANOVA, computer-based statistical methods, Student's *t*-test, *F*-test, Goodall's *F*-test
  Standard error of, 194, 200
Mean square, 215
Measurement theory, 2–10
Mglinstaller, 20
Monte Carlo simulations, 201–2, 305
Morphological disparity, *see* Disparity
Multiple regression, 262–3, 264–5
Multiple triangles, 65–8
Multivariate analysis of covariance, *see* MANCOVA
Multivariate analysis of variance, *see* MANOVA
Multivariate regression, *see* Regression

Nearest-neighbor analysis, 303–8
Negative allometry, 330
Neoteny, 338, 340
Non-affine transformations, *see* Non-uniform transformations
Non-uniform deformation, *see* Non-uniform transformations
Non-uniform transformations, 143–51, 391–2
  algebraic introduction, 146–50
  intuitive introduction, 143–6
  see also Bending energy, partial warps, uniform transformations
Normal distribution, 191, 194–5, 211, 212, 215, 219, 234
Normalized vector, 251, 256, *see also* Orthonormal basis

Ontogenetic scaling, 336–7
Ontogenetic allometry, 322, 324, 325, 333–4
Ontogenetic trajectory, 341–5
  computing by multivariate regression, *see* Regression
  comparing, 336–47
  see also Ontogenetic allometry
Ontogeny, in relation to phylogeny, 321–62
Ordination methods, *see* Principal components analysis, canonical variates analysis
Orthogonality, 163, 164
Orthonormal basis, 163–4
Outliers, 119
Outlines, *see* Semilandmarks
Over-dispersion, 307

Paedomorphosis, 334, 337, 341
Parallelism, between ontogeny and phylogeny, 321, 322, 334
Parameters, 191
Partial disparity, 302
Partial least squares analysis, 261–90
  comparison with CCA, 264–5
  comparison with PCA, 263–4
  comparison with regression, 262–3
  comparing patterns of covariances across groups, 265, 276–7
  using to test hypotheses of morphological integration, 266–76, 277–9
  software, 285–9
Partial Procrustes distance, 83, 85, 87, 93–6, 99, 222
Partial Procrustes superimposition, 93, 115
Partial warp scores, 129, 143
Partial warps, 143–51, 298
  algebraic introduction, 146–50
  degrees of freedom, 128–9
  intuitive introduction, 143–6
  why not to use as phylogenetic characters, 369–72
  see also Bending energy, non-uniform transformations
PCA, *see* Principal components analysis
PCAGen program, 151, 181–4
Peramorphosis, 334, 337, 341
Permutation test, 199–200
Phylogeny:
  Ontogeny and, 321–61
  Inferring from morphometric data, 363–5, 367–80
Photo-editing software, 47
Pillai's trace, 219
Pinhole camera, 40–4
Pinocchio effect, 119, 120, 132
PLS, *see* Partial least squares analysis
Population, 190
Positive allometry, 330
Power (statistical), 205
Power law, formalism for analysis of allometry, 330–1
Pre-shape space, 79–83
  definition, 79
  fibers in, 81–3
  shape of, 79–80
Principal axes of a transformation, 61
  variables implied by, 62–5
Principal components analysis, 15, 156–70, 328–30
  algebraic description of, 161–5
  comparison with PLS, 263–4
  finding phylogenetic characters by, 372–6
  geometric description of, 156–61
  interpretation of results, 166–70
    relationship to allometric coefficients, 328–30
  principal component scores, 159–61
  relationship to canonical variates analysis, 155, 170–1

scree plots, 168
software, 181–4
testing whether successive eigenvalues are distinct, 169
Principal strains of a transformation, 61
Principal warps, 149, *see also* Partial warps
Probability distribution, 190
Procrustes distance, 82–3, 93, 94, 389–90
full, 85, 115, 390
partial, 83, 85, 87, 93–6, 99, 389
Procrustes methods, 113–21
Procrustes residuals, *see* Shape coordinates, GLS
Procrustes superimposition, 113–19
axes of symmetry, and, 117
degrees of freedom, 97–8
partial, 93, 115
full, 94, 96, 99, 115

Reference configuration, 83
Selecting, 96–7
Regress6 program, 255–6
Regression, 229–59, 423
assumption of linearity, 236–40
bivariate, 232–4
comparison with PLS, 262–3
comparing regressions
directions of change, 249–53
rates of response, 254
correlation coefficient, 235
MANCOVA, 240–2
multivariate, 14, 58, 235–6
regression equations
bivariate, 230, 232
multivariate, 235–6
software, 255–9, 406
standardization by, 242–4
Relative warps, 170, 423, *see also* Principal components analysis
Repeatability, 30
Repeated median as a criterion for superimposition, 119–20
Resampling-based methods, 222–3
bootstrap, 195–9
jackknife, 200–1
Monte Carlo methods, 201–2
permutation tests, 199–200
Rescaling, 52
Residual variance, 263
Resistant-fit superimposition, 119–21
Resistant-fit theta-rho analysis, *see* Resistant-fit superimposition
Resources available electronically, *see* Web resources
Robust methods, 119
Rotation, 52, 74, 389
Rotational effects, 111
Row vector, 76

Saliences, 263
Sample, 190
Sampling with replacement, 196
Scalar, 209
Scale, 55
Scaling, 74, 388–9
Scaling factor, 425
Scatter-plot, 54
Scree plot, 168
SemiLand program, 406
Semilandmarks:
superimposition
differential weighting of landmarks and semilandmarks, 399, 403
equal weighting of landmarks and semilandmarks, 397–9
sliding to minimize bending energy, 399–401
sliding to perpendicular alignment on reference, 401–2
sensitivity analysis, 403–4
statistical analysis, 404–5
software, 406
Shape:
definition of, geometric, 11, 72–3
space, 83–7, *see also* Kendall's shape space, pre-shape space, tangent space
variables, 5, 23–152, *see also* Bookstein shape coordinates, superimposition, partial warps
see also Shape change, shape, theory of
Shape change:
anisotropy of, 62
as a deformation, 131–2
covariance with separately measured factors, *see* Regression
ontogenetic, *see* Ontogenetic trajectory
visualization of, 150–1
Shape coordinates, *see* Bookstein shape coordinates, superimpositions
Shape space, 83–7, *see also* Kendall's shape space, tangent space
Shape, theory of, 72–104
configuration matrix, 76–9
configuration space, 77
position or location, 77–8
size, 78–9
definition of shape, 11, 72–5
dimensions and degrees of freedom of, 97–8
pre-shape space, 79–83
fibers in, 81–3
shape of, 79–80
reference configuration, 96–7
rotation to maximal similarity, 87
shape spaces, 83–7
tangent spaces, 95–6

Shape variable, 5, *see also* Bookstein shape coordinates, superimposition, partial warps
Shear, 137–42
ShuffleAllometry program, 258
Simple3D program, 406
Singular axes, 262
Singular value, 263
Singular value decomposition, 262–3
Singular warps, 262
Size, 11–13, 51–3, 55–6, 230, *see also* Centroid size
Sliding baseline registration, 109–13
  uniform transformations, calculation from, 142–3
Sliding semilandmarks, 399–402
Software, 19–20
  Bookstein shape coordinates, 69
  canonical variates analysis, 184–6
  comparing directions of regression vectors, 257–9
  comparing subspaces of variation, 317
  customizing symbols for scatter-plots, 186–7
  disparity, 314–17
  MANOVA, 223–8
  partial least squares analysis, 285–9
  principal components analysis
    three-dimensional landmarks, 406
    two-dimensional landmarks, 181–4
  regression
    three-dimensional landmarks, 406
    two-dimensional landmarks, 255–9
  semilandmarks, 406
  sources on the web, 19
  standardization by regression, 257
  traditional morphometric variables, computing from landmarks, 358–60
  superimposition
    three-dimensional data, 406
    two-dimensional data, 121, 127–8
  variation, 314–17
  see also individual programs
SpaceAngle program, 317
Standard6 program, 257
Standardization by regression, 242–4
Strain, *see* Principal strain
Strauss–Sadler estimate of range, 307
Student's *t*-test, 58, 197–9, 211–12, 222, *see also* Hotelling's T^2 test
Superimposition, 105–8, 387–390
  methods
    generalized least squares Procrustes, 113–19, 387–390
    resistant-fit, 119–21
    sliding baseline registration, 109–13
    three-dimensional landmarks, 387–390
  resolving apparent inconsistencies among methods, 124–6
  software, 127–8, 406
SuperPoser program, 121, 127
Symmetric tensors, 61

t-test, *see* Student's *t*-test
Tangent space, 95–6
Target form, 137–8, 147
Taxonomic discrimination, 365–7
Taxonomy, 364
TBox program, 226
Thin-plate spline, 129–52
  non-uniform components from three-dimensional landmarks, 392
  non-uniform components from two-dimensional landmarks, 143–50
  physical metaphor, 133–4
  shape change as deformation, 131–2
  software, 151–2
  uniform (affine) components, 134–43
  visualization of shape change, 150–1
ThreeDPCA6 program, 406
ThreeDRegress6 program, 406
TIFF format for image files, 45
tpsDig program, 48–9, 406
tpsrelw program, 406
TradMorphGen program, 359–60
Translation, 51, 52, 57, 74, 77, 80, 84, 387–8, *see also* Centered configuration matrix
Transpose of a matrix, 162
Transpositional allometry, 341
Triangles, shape coordinates, 52–5
Truss measurement scheme, 4–5
Two-point registration, 51–72, *see also* Bookstein shape coordinates
TwoGroup program, 69, 71–2, 224–6
Type 1 landmark, 31–2
Type 2 landmark, 31–2
Type 3 landmark, 31–2
Type I error, 193
Type II error, 193

Uniform (affine) transformations, 134–43
  Conceptual framework for, 137–8
  Mathematical derivation of, 138–42
  shear and compression/dilation terms, 137–42
  six component model, 136
  sliding baseline registration and, 142
  Bookstein shape coordinates and, 142
  RFTRA, and, 143
  component scores, 181, 185

Variables, 190
  selection of, 23–4
Variance, 162, 214
Variance-covariance matrices, 263, 295, 298

Variance transfer, 105–9
Variation, 295–6, *see also* Disparity
  metrics, 296–7
  structure, analysis of
    nearest-neighbor analysis, 303–8
    common subspace analysis, 308–14
    see also Principal components analysis
VecCompare program, 257–8
VecDisplay program, 258–9
Vectors:
  eigenvectors, 163–4, 167
  inner (dot) product of, 251
  see also Linear combinations, partial warps, principal components analysis

Web resources, 19
Wilks' lambda, 58, 178, 219, 236
Windows metafiles, 45