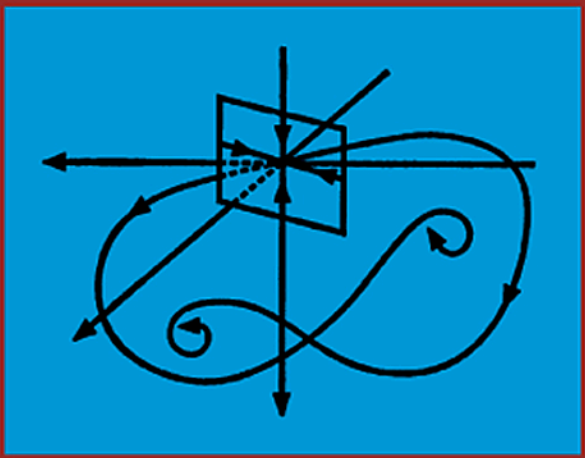
MATHEMATICS IN SCIENCE AND ENGINEERING Volume 212
SERIES EDITOR: C.K. CHUI

Computational Methods *for* Modelling *of* Nonlinear Systems



A. Torokhti and P. Howlett

Computational Methods for Modelling of Nonlinear Systems

This is volume 212 in
MATHEMATICS IN SCIENCE AND ENGINEERING
Edited by C.K. Chui, *Stanford University*

A list of recent titles in this series appears at the end of this volume.

Computational Methods for Modelling of Nonlinear Systems

A. Torokhti
P. Howlett
SCHOOL OF MATHEMATICS AND STATISTICS
UNIVERSITY OF SOUTH AUSTRALIA
MAWSON LAKES, SA
AUSTRALIA

ELSEVIER
Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

First edition 2007

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

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

ISBN-13: 978-0-444-53044-8
ISBN-10: 0-444-53044-4
ISSN: 0076-5392

Printed and bound in The Netherlands

07 08 09 10 11 10 9 8 7 6 5 4 3 2 1

To my family.
Anatoli Torokhti

This page intentionally left blank

Preface

In broad terms, the purpose of this book is twofold. First, we present a theoretical basis for mathematical modelling of nonlinear systems which can be described from observations of their input-output signals only. Second, we give specified computational methods which follow from the general theoretical framework. The term "nonlinear system" means a device which transforms signals. The setting is often rather abstract but is carefully detailed. This is implied by our desire to embrace a wide spectrum of possible types of nonlinear systems and to provide a rigorous analysis of topics we study.

While our work is mainly motivated by research in systems theory, we are very concerned with mathematical framing for problems under consideration. This is because the subjects considered in the book represent an absorbing blend of special areas in approximation theory, numerical methods, mathematical statistics and optimal filtering of stochastic signals. Many of the questions we ask are new (see Overview). In many cases, our view of problems we consider is quite specific, and therefore we assume the reader's willingness to accept new terminology.

The book marks the coming together of two basic interlacing research streams. One stream consists of work on operator approximation with a given accuracy and operator interpolation, and has its origin in the work of P. Prenter, V. Istrăţescu, V. Bruno and I. Daugavet around 1980. These pioneering researchers generally worked on rather general problems. We present new, recently developed methods which have been motivated by their fundamental results and which generalize them. The second stream, which studies the best operator approximation techniques, began with Wiener, Hotelling, Karhunen, Loève, Bode and Shannon around 1940 and 1950. We provide new methods that have been initiated with their pioneering results and which advance them to solution of more general problems. Those methods have been published very recently.

Often a book summarizes knowledge in the field. While our book presents very recent accomplished achievements in the area with their rigorous theoretical justification, it may also be viewed as a report on the work in progress where a number of questions are still open.

The book contains a number of numerical examples. In many cases, we used data obtained at http://sipi.usc.edu/database/. This data is the digital basis for a number of images that we have used and that have been used a number of times in the open literature. We have made a concerted effort to locate the source of the original images but unfortunately have not been completely successful.

We wish to acknowledge many debts. A. Torokhti is grateful to I. Daugavet and V. Malozemov (both are with the St. Petersburg State University, St. Petersburg, Russia) for many discussions and useful comments related to the first stream of the book. Both co-authors wish to thank P. Pudney (the University of South Australia, Adelaide, Australia) for his enormous time spent assisting us with numerical simulations. We are specifically grateful to our colleagues in the School of Mathematics and Statistics at the University of South Australia for supporting us in many aspects of the work in this book.

Finally, we are pleased to thank the Australian Research Council (the Large Research Grant A49943121 for 1999–2002 and the ARC Discovery Grant DP0453236 for 2004–2006) and the University of South Australia (a number of internal research grants in 1996–2005) for supporting the research provided here.

Contents

This page intentionally left blank

Chapter 1

Overview

In this book, we study theoretical and practical aspects of computational methods for mathematical modelling of nonlinear systems. A number of computational techniques are considered, such as

- methods of operator approximation with any given accuracy,
- operator interpolation techniques including a non-Lagrange interpolation,
- methods of system representation subject to constraints associated with concepts of causality, memory and stationarity,
- methods of system representation with an accuracy that is the best within a given class of models,
- methods of covariance matrix estimation,
- methods for low-rank matrix approximations,
- hybrid methods based on a combination of iterative procedures and best operator approximation, and
- methods for information compression and filtering under condition that a filter model should satisfy restrictions associated with causality and different types of memory.

As a result, the book represents a blend of new methods in general computational analysis, and specific, but also generic, techniques for study of systems theory ant its particular branches, such as optimal filtering and information compression.

We are interested in the following basic questions:

1. Suppose a nonlinear system can be described from observations of its input-output signals only. What is a constructively defined mathematical model of such a system?
2. What is the accuracy of such a model?

3. How can we compute this model?

The following example illustrates a motivation for the problems under consideration. Suppose that x, y and u are functions of discrete time so that $x(k), y(k)$ and $u(k)$ are values of x, y and u at t_k with $k = 1, 2, \ldots$. Consider the discrete-time time-invariant system [15]

$$\begin{aligned} u(k+1) &= G[u(k), x(k)], \\ y(k) &= F[u(k), x(k)], \quad u(0) = u_0, \end{aligned} \tag{1.1}$$

where $G, F : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$, $u(k)$ is the state vector, u_0 is the initial state, $x(k)$ and $y(k)$ are the input and output, G is the one-step ahead state transition operator and F is the input-output map. "The function F that describes the input-output of the system is of primary importance in systems theory because this is all an external observer can see" [15].

This book, firstly, brings together and systematizes known results in the branch of general computational analysis associated with modelling of nonlinear systems and, secondly, presents a number of new results which are natural but very specific extensions of known techniques.

In practice, one has to be content with an approximate solution to the problem. As a rule, an exact solution to the problem can only be obtained when the problem is completely described by a finite number of input and output data. While *"The best material model of a cat is another, or preferably the same, cat"* (N. Wiener, Philosophy of Science, 1945), the difficulty is that we may not have "a cat" in hand and even a complete description is often not available. In practice, we wish to model a system which is known only partially and indeed is often known only from the observed input-output characteristics. Such observations are typically incomplete. In reality, the observations are stochastic and are disrupted by noise. Moreover, inputs are often unknown and the model may use only sampled observation of the output and some available (*a priori*) knowledge of the system. The models in this book are developed both for the case of "deterministic" spaces and for the case of probability spaces under the practical assumption that description of the system is only realized via observed "external" information.

In broad terms, we set two aims for ourselves. Firstly, we wish to develop a system of general conceptions which allows us to formulate and solve problems related to a representation of models which should have properties associated with properties of real world phenomenon such as causality, memory, stationarity, etc. Secondly, we wish to elaborate the general conceptions to specific techniques which can be applied to practical problems.

In our view, it is possible to develop a unified way to the solution of many problems in modelling of nonlinear systems. In this book we present a number of techniques, many of which are united by one basic idea: to increase degrees of freedom via different types of nonlinearity.

In Chapter 2, we study models of nonlinear systems formed by operator polynomials. More precisely, a nonlinear system is treated as a nonlinear operator, and we study its approximation by operator polynomials. It should be noted that nonlinear operator approximation is an intensively developed branch of general computational analysis. In recent decades, there have been a number of significant achievements in this area spread through diverse publications. From a theoretical point of view, methods in operator approximation theory [45]–[156] are important since they generalize classical results in function approximation theory. The practical importance of results in nonlinear operator approximation arises directly from a demand in the modelling of nonlinear systems [98], [106]–[151]. We consider both existence theorems for approximating operators of preassigned accuracy and numerical schemes for their practical realization.

Chapter 2 is organized as follows. In Section 2.2, the general formulation of the problem is presented. In Sections 2.3 and 2.4, we describe methods of nonlinear operator approximation in the space $C([0,1])$ and in Banach spaces as elaborated by Gallman and Narendra [45], Daugavet and Lanne [23] and Daugavet [24]–[26]. These methods are not widely known and the results presented in Sections 2.3 and 2.4 draw attention to these effective procedures.

A generic approach to operator approximation in topological vector spaces is considered in Section 2.5 and is elaborated in Sections 2.5.5–2.7. In Section 2.5.5, we give theorems on the existence of an operator approximating an operator defined on a compact set in a topological vector space. In Section 2.6 the technique of Section 2.5.5 is modified to establish some theorems on constructive approximation on noncompact sets. In Section 2.7, the results of Sections 2.5.5 and 2.6 are specified in terms of moduli of continuity.

In Chapter 3, we consider some fundamental principles of the general theory for nonlinear operator interpolation. Interpolating operators are naturally connected to modelling of nonlinear systems. If a system is given by finite sets of input-output signals, then an interpolating operator provides a model of such a system. We give both an extension of the Lagrange procedure to the case of interpolation in Hilbert space, and consider a specific interpolation method for the case when the Lagrange interpolation idea does not work.

So-called realistic operators are the subject of study in Chapter 4. The class of operators described in this chapter generalizes causal operators, operators with finite memory, stationary operators, etc. The generalization is given in the form of special operators which we call $\mathcal{R}$-continuous operators and approximately $\mathcal{R}$-continuous operators. This study is motivated by the necessity of formalizing and modeling of physical properties of realistic nonlinear systems.

All the next chapters are devoted to the study of different forms for best computational methods, that is the methods with a highest possible accuracy. In Chapter 5, the methods of best approximation for the system acting in "deterministic" and probability spaces are considered. We are specifically concerned with different types on nonlinear models and give a rigorous analysis of their accuracy. In particular, we study the so called Hadamard-quadratic model, the r-degree model, the best causal model, and the so-called hybrid method for finding models which combine advantages of best approximation techniques and iterative procedures.

Computational methods for optimal filtering stochastic signals are given in Chapter 6. Here, we consider generalizations of the Wiener filter to the optimal linear filters in Hilbert spaces, and linear and nonlinear optimal filters with different types of memory. This includes optimal causal linear filtering with piecewise constant memory, optimal causal polynomial filtering with arbitrarily variable memory and optimal unconstrained nonlinear filtering based on an extension of the hybrid method studied in Chapter 5. Our methodology is based on the Wiener-Kolmogorov approach. In this regard, Sorenson ([142], p. 14) points out that the "pioneering work of Wiener and Kolmogorov" enabled us to "bridge the gap between communication engineers and statisticians" and observes that the introduction of "communication engineering and mathematical concepts" assisted "the synthesis of ideas from both fields in order to obtain a more powerful technique". Our extensions of the Wiener-Kolmogorov ideas exploit an underlying Volterra functional form, which has been studied extensively by many authors. See for example [16, 92, 95, 112, 136, 159, 175, 182].

At the same time, our treatment of techniques associated with the Volterra series differs essentially from the known approaches. Firstly, our estimator has the specific structure to accommodate both causality and finite memory. Secondly, we establish an equivalence between the original problem (6.66) formulated in terms of the vectors $\mathbf{x}$ and $\mathbf{y}$ and the collection of problems (6.74) formulated in terms of the components of $\mathbf{x}$ and $\mathbf{y}$. This allows us to represent the estimator in a natural coordinate form (6.62). Such a representation implies a significant computational benefit related to the small sizes of matrices to be computed in 6.77).

The results described in Chapter 6 are new and are based on a substantial extension of our earlier results [153]–[160].

Many known results assume the relevant covariance matrices are invertible. Such an assumption is quite restrictive from a computational point of view and is often associated with numerical instability. We avoid such a drawback by using generalized inverse operators [6] which always exist.

In Chapter 7, computational methods for optimal compression and reconstruction of random data are studied. We begin with the standard Principal Component Analysis (PCA) and Karhunen-Loéve transform (KLT), and give a generalized PCA–KLT. Then these techniques are extended to an optimal hybrid transform based on Hadamard-quadratic approximation, an optimal transform formed by a combination of nonlinear operators and an optimal generalized hybrid transform.

To be more specific, we list the following questions to be studied. These questions are completely or partially answered in the book.

1. What is a computationally realizable nonlinear model of a nonlinear system in spaces which are algebras? We say that the space is an algebra if the operation of multiplication can be defined in the space.
2. Is there a way to provide a similar model in a space which is not an algebra?
3. What kind of restrictions should be imposed on the spaces above?
4. Can we find a nonlinear model that approximates the system with any given accuracy?
5. Can we provide a nonlinear model that approximates a system which transforms signals belonging to a non-compact set?
6. If the answer to the above question is "yes", what type of a space should be chosen? For instance, should it be a Banach space or a Hilbert space? What kind of topology should be chosen there?
7. Can the Lagrange's idea of real function interpolation be extended to the case of nonlinear operator interpolation? What kind of spaces should we use?
8. What are the limitations for such an extension?
9. How can we overcome those limitations?
10. What is a unified definition of causality, memory and stationarity in operator terms?
11. Can we determine a system model which is "equipped" with the properties mentioned in the preceding question, so that this model approximates the system with any given accuracy?

12. What kind of spaces should be chosen to determine such a system? Can we use Banach spaces or even linear topological spaces? What kind of limitations should be imposed on associated topologies?
13. What kind of nonlinearity should be chosen for the model which realizes an approximation of the system with a highest possible accuracy? Such a model is called the best model.
14. Suppose the nonlinear model is chosen in the form of an operator polynomial of an arbitrary degree. Then the degree of the operator polynomial is the only degree of freedom in the model performance. Can we choose a different way in a model determination so that the model has more than one degree of freedom?
15. Is the best model unique? If not, what are conditions for its uniqueness?
16. What are different types of memory in realistic systems? How can we formalize them?
17. How and what particularities in the structure of the model should be chosen to satisfy different types of memory?
18. Can we find a best model which is "equipped" with specific type of memory?
19. What is a generic representation for the linear transform of stochastic vectors subject to constraint on its rank?
20. Is such a representation unique? If not, what is an analytical form for a family of such transforms?
21. Can we find nonlinear transforms with a better compression ratio and associated accuracy than those for linear rank-constrained transforms?

We believe that this book could give an opportunity to compare different methods, including a number of new ones, to choose the best suitable algorithm for applications, and will initialize a future theoretical development of the presented results.

Finally, the way for referring to material within the book is as follows. Theorems, lemmas, definitions, equations, et. are separately numbered for each chapter. A reference to material within the same chapter does not name the chapter. A reference to material in a different chapter names the chapter.

Part I

Methods of Operator Approximation in System Modelling

This page intentionally left blank

Chapter 2

Nonlinear Operator Approximation with Preassigned Accuracy

2.1 Introduction

Nonlinear systems theory is a direct area of applications for the methods of nonlinear operator approximation. Gallman and Narendra [45] appear to have been the first to propose an application of the Stone–Weierstrass theorem generalization to the modelling of nonlinear systems. Further development in nonlinear operator approximation techniques and their application to nonlinear systems has been made by a number of authors. We list references [12], [15]–[17], [22]–[26], [57, 60, 65, 98, 99, 104], [106]–[108], [117]–[131], [148, 153, 154, 157] as examples of such methods.

In this chapter, we consider both existence theorems for approximating operators of preassigned accuracy and numerical schemes for their practical realization.

The chapter is organized as follows. A general formulation of the problem is given in Section 2.2. In Sections 2.3 and 2.4, we describe methods of nonlinear operator approximation in the space $C([0,1])$ and in Banach spaces as elaborated by Gallman and Narendra [45], Daugavet and Lanne [23], and Daugavet [24]–[26]. These methods are not widely known and the results presented in Sections 2.3 and 2.4 draw attention to these effective procedures.

A generic approach to operator approximation in topological vector spaces is considered in Section 2.5 and is further elaborated in Section 2.6. In Section 2.5.5, we give theorems on the existence of an operator approximating an operator defined on a compact set in a topological vector space. In Section 2.6, the technique of Section 2.5 is modified to establish some theorems on constructive approximation on noncompact sets. In Section 2.7, the results of Sections 2.5 and 2.6 are specified in terms of modulus of continuity.

2.2 Generic Formulation of the Problem.

We begin with some remarks on a general formulation of the problem for nonlinear operator approximation.

Let X and Y be locally convex topological vector spaces, with $K \subseteq X$ a subset and $F : K \to Y$ a continuous map.

The problem is to prove that for a given neighbourhood of zero $\tau \subseteq Y$, there exists a constructively–defined operator $S : X \to Y$ and a neighbourhood of zero $\varepsilon \subseteq X$ such that, for all $x \in K$ and all $x' \in X$ with

$$x' - x \in \tau,$$

we have

$$F(x) - S(x') \in \tau.$$

This general statement contains a few issues which must be clarified. Firstly, it is necessary to establish some restrictions on the subset K. *Should K be a compact set?* If so, *what kind of restrictions should then be imposed on X and Y?* Next, suppose K is bounded but not necessarily compact. *What kind of topology should then be used for X and Y?*

Secondly, a locally convex topological vector space is not an algebra, as the ordinary multiplication operation cannot be defined for this space. This causes corresponding difficulties for the structure of an approximating operator S.

In the following sections, we give and discuss a variety of possible solutions to this problem depending on the specific type of spaces X, Y, set K and approximating operator S.

2.3 Operator Approximation in Space $\mathbf{C}([0,1])$.

The case where X and Y are the space $C = C([0,1])$ of continuous functions on the interval $[0,1] \subset \mathbb{R}$ and $K_c \subset C$ a compact subset, allows a specific structure for the operator S. The method has been presented in [45, 23].

Let $F : K_c \to C([0,1])$ be continuous. It is assumed that $F(\mathbb{O}) = \mathbb{O}$ if $\mathbb{O} \in K_c$, where $\mathbb{O}$ is the zero element in $C([0,1])$.

In the next theorem we shall need some notation. We denote by $\mathcal{Z}_+$ the set of positive integers and by C_m an $m-$dimensional subspace of C, with $P_m \in \mathcal{L}(C, C_m)$ given by $P_m(x) = x_m$. The functions $g_j : C_m \to C_m$ $(j = 1, \ldots, N)$ are assumed continuous and $S_{m,N} \in \mathcal{L}(C_m, C_m^N)$ is given by

$$S_{m,N}(x_m) = (g_1(x_m), \ldots, g_N(x_m)).$$

The operator $Q : C_m^N \to C$ is continuous and such that

$$Q(g_1(x_m), \ldots, g_N(x_m)) = \sum a_{j_1,\ldots,j_N} [g_1(x_m)]^{j_1} \ldots [g_N(x_m)]^{j_N},$$

with

$$a_{j_1,\ldots,j_N} \in \mathbb{R}, \quad j_k \geq 0 \quad \text{and} \quad \sum j_k \geq 0.$$

Theorem 1. [23] *For any $\varepsilon > 0$, there exist $\delta > 0$, $N \in \mathcal{Z}_+$ and operators P_m, $S_{m,N}$ and Q such that, for all $x \in K_c$ and $u \in C$ satisfying the condition*

$$\|x - u\| \leq \delta,$$

we have the inequality

$$\|F(x) - QS_{m,N}P_m(u)\| < \varepsilon.$$

The proof in [23] depends greatly on the structure of the operator $S_{m,N}$. To be specific, the operator $S_{m,N}$ must be such that

1. if $u_m \neq \mathbb{O}$, then
$$S_{m,N}(u_m)(t) \neq \mathbb{O}^N$$
for all $t \in [0,1]$, where $\mathbb{O}^N$ is the zero element in C^N;

2. if
$$S_{m,N}(u_m^{(1)})(t_1) = S_{m,N}(u_m^{(2)})(t_2),$$

then either

$$u_m^{(1)} = u_m^{(2)} = \mathbb{O} \quad \text{or} \quad u_m^{(1)} = u_m^{(2)}$$

and $t_1 = t_2$.

A special study of operators $S_{m,N}$ with the above properties is provided in [24]. In particular, it is shown that there always exists an $S_{m,N}$ satisfying these conditions.

Note that Theorem 1 is the generalization of a similar theorem by Gallman and Narendra [45].

The space C is an algebra, since multiplication of elements takes place in C. This is essentially exploited for the structure of the approximating operator in Theorem 1; namely, the operator Q is constructed from the product of functions $g_k(x_m)^{j_k}$.

Nevertheless the structure of the approximating operator above is quite complicated. A desire to study the possibility of simpler structures for an approximating operator has motivated Theorems 2 and 3 [25] below.

Let $C = C(T)$, where T is metric and compact. Set

$$A(u) = L_0 + L_1 u + L_2 u^2 + \ldots + L_n u^n, \tag{2.1}$$

where $L_0 \in C$, $L_k \in \mathcal{L}(C, C)$ is a linear operator and $u^k = [u(t)]^k$. The class of operators A is denoted by $\mathcal{A}$.

An important feature in the structure of the operator A is that L_k is linear but not a $k-$linear operator as commonly supposed. See the following sections in this connection.

We write $\mathbb{N}$ for the set of natural numbers, $\mathbb{R}^+$ for the set of real positive numbers and $\mathbb{Z}$ for the set of integers.

Theorem 2. [25] *Let C_m be a $m-$dimensional subspace of $C(T)$ with basis $\varphi_1(t), \ldots, \varphi_m(t)$ such that, for any $N \in \mathbb{N}$, all functions $\prod_{j=1}^{m} \varphi_j^{k_j}(t)$ are linearly independent, where $k_i \in \mathbb{N}$ for $i = 1, \ldots, m$ and $k_1 + \ldots + k_m = N$. Let K_m be a compact subset in C_m. Then any continuous mapping $F : K_m \to C$ can be approximated by an operator of class $\mathcal{A}$ with any prescribed accuracy.*

Note that the assumption of the theorem is not very restrictive. The following example [25] describes a subspace of $C(T)$ satisfying this condition.

Example 1. *Let $\alpha_1, \ldots, \alpha_m \in \mathbb{R}^+$ be such that*

$$q_1\alpha_1 + \ldots + q_m\alpha_m = 0$$

implies

$$q_1 = \ldots = q_m = 0,$$

where $q_1, \ldots, q_m \in \mathbb{Z}$. *Then all numbers* $k_1\alpha_1 + \ldots + k_m\alpha_m$ *are different if* $k_1, \ldots, k_m \in \mathbb{N}$ *and* $k_1 + \ldots + k_m = N$. *Therefore the functions*

$$\varphi_1(t) = t^{\alpha_1}, \quad \ldots, \quad \varphi_m(t) = t^{\alpha_m}$$

satisfy the assumption in Theorem 2 because of linear independence.

Theorem 2 is recast in [25] in terms of the modulus of continuity. We give the definition first.

Definition 1. *Let* K_c *be a compact subset of* $C(T)$ *and* $F : K_c \to C(T)$ *a continuous mapping. The function*

$$\omega(F;\delta) = \sup_{x_1,x_2\in K:\ \|x_1-x_2\|\le\delta} \|F(x_1) - F(x_2)\|$$

is called the modulus of continuity for F.

Theorem 3. [25] *Let* $K_c \subset C([0,1])$ *be a compact set. For any* $\varepsilon > 0$, *there exists a continuous operator* $Q_\varepsilon : K_c \to C([0,1])$ *such that, for all* $x \in K_c$,

$$\|Q_\varepsilon(x) - x\| < \varepsilon,$$

and then for any continuous $F : K_c \to C([0,1])$ *there exists* $A \in \mathcal{A}$ *such that*

$$\|F(x) - A(Q_\varepsilon(x))\| < \frac{1}{2}\omega(F;4\varepsilon) + \varepsilon.$$

Remark 1. [25] *It might be taken from Theorems 2 and 3 that class* $\mathcal{A}$ *is a reasonable approximation tool. Unfortunately this is not so, since an operator* $A \in \mathcal{A}$ *can be unstable to small disturbances of its argument.*

2.4 Operator Approximation in Banach Spaces by Operator Polynomials

The structures of approximating operators considered in the preceding section cannot directly be extended to the approximation of operators acting on abstract Banach spaces. In this and subsequent sections we present some possible forms of approximating operators acting on spaces more general than C.

We begin with the following definition.

Definition 2. *Let X and Y be Banach spaces and $\mathcal{B}(X^k, Y)$ a set of continuous $k-$linear operators. Let $L_0 \in Y$, $L_k \in \mathcal{B}(X^k, Y)$ and $L_k(x^k) = L_k(\underbrace{x, \ldots, x}_{k})$. The operator P_n defined by*

$$P_n(x) = L_0 + L_1 x + L_2 x^2 + \ldots + L_n x^n \tag{2.2}$$

is called an operator polynomial of degree n.

The structure of the operator P_n is quite general. A popular choice of $k-$linear operators L_k are multiple Volterra integrals [117]. The $k-$linear operators in (2.2) can also be designed from superpositions of sigmoidal functions ([22], [4]), radial functions [98] or wavelets.

Prenter [104] first used an operator polynomial to prove the Stone–Weierstrass theorem for operators on Hilbert space.

Let H be a real, separable Hilbert space, K a compact subset of H, and let $\mathcal{H}(K, H)$ denote the family of continuous operators from K into H together with the uniform norm topology

$$\|F - G\| = \max_{x \in K} \|F(x) - G(x)\|$$

where $F, G \in \mathcal{H}(K)$. Prenter proved the following theorem.

Theorem 4. (Prenter, [104]) *The the family of continuous finite rank polynomial operators $\{P_n\}$ on H, restricted to K, is dense in $\mathcal{H}(K, H)$ restricted to K.*

Istrăţescu [65] and Bruno [12] extended Prenter's result respectively to operators on Banach spaces and topological vector spaces.

Theorem 5. (Istrăţescu, [65]) *Let K be a compact set in a Banach space X. Given $\varepsilon > 0$, there exists an operator polynomial P_n such that, for all $x \in K$, the inequality*

$$\|F(x) - P_n(x)\| < \varepsilon$$

holds.

The extension [12] of the Stone–Weierstrass theorem to topological vector spaces is based on the following concepts.

Definition 3. *Let X be a topological vector space and let $\mathcal{L}(X, X_m)$ be the set of all continuous linear mappings from X into X_m, where $X_m \subseteq X$ is a subspace of dimension m. Let $\{G_m\}_{m=1,2,\ldots}$ be a sequence of continuous linear operators $G_m \in \mathcal{L}(X, X_m)$. The sequence $\{G_m\}_{m=1,2,\ldots}$ is said to be* equicontinuous on compacta*, if for any given compact set $K \subseteq X$ and*

any given neighbourhood of zero $\mu \subseteq X$ we can find a neighbourhood of zero $\sigma = \sigma(\mu) \subseteq X$ such that $G_m(x_1) - G_m(x_2) \in \mu$ for all $m = 1, 2, \ldots$ whenever $x_1, x_2 \in K$ and $x_1 - x_2 \in \sigma$.

Definition 4. [133] *We say that X possesses the Grothendieck property of approximation if there exists a sequence $\{G_m\}_{m=1,2,\ldots}$ of continuous linear operators $G_m \in \mathcal{L}(X, X_m)$ such that the operators G_m are equicontinuous on compacta and uniformly convergent to unit operators on the same compacta*[1].

Remark 2. *The conditions in Definition 4 are related. The condition that the sequence of operators is equicontinuous on a compact set implies that a uniformly convergent subsequence can be found. On the other hand, if the sequence of operators converges uniformly to the unit operator on a compact set, then the sequence is equicontinuous.*

See also reference [108] in this connection.

Let X and Y be topological vector spaces. We denote by $C(X, Y)$ the family of continuous operators from X into Y in compact open topology.

Theorem 6. (Bruno, [12]) *Let X and Y be real Hausdorff topological vector spaces with the Grothendieck property of approximation. Then operator polynomials $P_n : X \to Y$ of finite rank are dense on $C(X, Y)$.*

A related fundamental result is due to Fréchet [86].

Fréchet Theorem [86] *Any functional defined and continuous on a space of functions is representable as a limit of a sequence of polynomial integral functions on any bounded subset of this space.*

In [26], the Fréchet Theorem is generalized to continuous operator approximation on an arbitrary set in a separable Banach space.

For an *open* subset of a separable Banach space X, the generalization of the Fréchet Theorem is as follows.

Theorem 7. [26] *Suppose $\Omega \subseteq X$ is an open subset of a separable Banach space $\mathcal{X}$ and $F : \Omega \to Y$ a continuous operator. Then there exists a sequence of operator polynomials $P_n : X \to Y$ such that*

$$P_n \to F$$

for all $x \in \Omega$.

[1] *The sequence $\{G_m\}$ is said to* converge uniformly *on the compact set $K \subseteq X$ to the unit operator on K if for any given neighbourhood of zero $\mu \subseteq X$ we can find $M > 0$ such that $G_m(x) - x \in \mu$ whenever $x \in K$ and $m > M$.*

The proof is mainly based on the following theorem by Istrăţescu [65].

The class of sets Ω in Theorem 7 can be extended so that continuous operators defined on them can be approximated, in the sense of pointwise convergence, by operator polynomials. In particular, if the set $\Omega \subseteq X$ is such that any continuous operator defined on Ω can be extended continuously to the complete space $\mathcal{X}$, or at least to a neighbourhood Ω, then for such a set Ω the statement of Theorem 7 holds. For example, convex closed bodies in $\mathcal{X}$ belong to the above–mentioned class of sets Ω.

An extension of Theorem 7 to the case where Ω is an *arbitrary* set in $\mathcal{X}$ requires an essential constraint for F as follows.

Theorem 8. [26] *Let Ω be an arbitrary set in $\mathcal{X}$ and $F : \Omega \to Y$ a uniformly continuous operator. Then there exists a sequence $\{P_n\}$ of operator polynomials such that $P_n \to F$ for all $x \in \Omega$.*

Without loss of generality, it can be assumed that Ω is closed, since any operator uniformly continuous on Ω can be extended with respect to continuity onto its closure.

Further, not every operator defined and uniformly continuous on a closed bounded set in a Banach space can be arbitrarily uniformly approximated by generalized polynomials. An example of such a functional on the unit ball of a Hilbert space is given in [94]. Another, perhaps simpler, example, is given in [26].

Example 2. [26] *Let X be the space of sequences*

$$x = (\xi_1, \xi_2, \ldots)$$

with $\xi_n \to 0$ and set

$$\|x\| = \max |\xi_k|.$$

Suppose that $\Omega \subset X$ consists of all sequences of the form

$$y = (\sigma_1, \sigma_2, \ldots, \sigma_k, 0, 0, \ldots),$$

where σ_i is $+1$ or -1. The set Ω belongs to the unit ball of the space X and is closed.

We define a functional f on Ω by

$$f(y) = \sigma_1 \sigma_2 \ldots \sigma_k.$$

Since we have

$$\|y' - y''\| \geq 1$$

for distinct points $y', y'' \in \Omega$, the functional f is uniformly continuous. It is easily extended to the unit ball in X with preservation of uniform continuity.

Let

$$X_m = \{x = \xi_1, \ldots, \xi_m, 0, 0, \ldots)\}$$

be a subspace of X. We make no distinction between this space and the space l_m^∞. The restriction of a polynomial functional p_n of degree n on X_m is a polynomial in m variables of degree not greater than n. Therefore if

$$\Omega_m \subset \Omega \cap X_m$$

is the set of vertices of the $M-$dimensional cube $[-1,1]^m$ and if f_m is the restriction of the functional f to Ω_m, then for any polynomial p_n of degree n and for any m, the inequality

$$\sup\{|f(x) - p_n(x)| \ \textit{for} \ x \in \Omega\} \geq E_n(f_m)$$

holds, where $E_n(f_m)$ is the best uniform approximation of f_m by polynomials in m variables of degree n.

The function f_m is odd with respect to each argument, and in the set of all best–approximation polynomials of this function, there also exists an odd polynomial with respect to each argument. In particular, for $m > n$ the zero polynomial is the best approximation polynomial, since this is the unique polynomial of degree n of m variables which is odd with respect to all arguments, and therefore

$$E_n(f_m) = 1$$

for $m > n$.

Thus for any polynomial function p_n, we have

$$\sup\{|f(x) - p_n(x)| \ \textit{for} \ x \in \Omega\} \geq 1.$$

2.5 Approximation on Compact Sets in Topological Vector Spaces

In this section, we consider a systematic theoretical procedure for the constructive approximation of non-linear operators and show how this procedure can be applied to the modelling of dynamical systems. There are several properties which we have sought to preserve in the modelling process. In many cases the only given information about such a system is information pertaining to an abstract operator F. We wish to construct an approximating operator S which can be realized in physical terms, will approximate F with a given accuracy and must be stable to *small* disturbances. The operator S defines our model of the real system and will

be constructed from an algebra of elementary continuous functions by a process of finite arithmetic. For this reason we regard S as computer-processable.

A number of specific examples are presented for the particular purpose of illustrating the theoretical results. Although the examples have been simplified for computational convenience and are somewhat artificial they are none-the-less representative of real situations. In these examples we have used an underlying polynomial algebra but we note that this is simply a matter of theoretical convenience. A suitable wavelet algebra could be used instead. Another currently popular alternative is an algebra generated by superpositions of a sigmoidal function. Such algebras are discussed in detail by Cybenko [22] and Barron [4]. In general we require only that the underlying algebra satisfy the conditions of Stone's Algebra [113]. For the purposes of this Section we have assumed that the elementary continuous functions which generate the algebra can be evaluated by a process of finite arithmetic. While this assumption may not be strictly correct the errors involved are limited only by machine accuracy and in principle do not disrupt our analysis.

The study in this Section has been motivated by a desire to understand the nature of the modelling process for simulation of a real dynamical system. A dynamical system is defined by a mapping that transforms a set of input signals to a corresponding set of output signals. A signal is normally defined by a set of real number parameters. In practice these parameter sets could be uncountably infinite. For a computer simulation of the system each signal must be represented by a finite set of real number parameters and the mapping must be represented by a finite arithmetical process. We must nevertheless show that the simulated system is a good approximation to the real system.

To justify the approximation process we impose a basic topological structure and use the consequent notions of continuity to establish theorems of Weierstrass type. In the case of a general continuous map $F : X \to Y$ where X, Y are locally convex topological vector spaces we will show that the approximation procedure can be used on any given compact subset $K \subseteq X$. Indeed if we assume that F is known only on K then for some suitable neighbourhood ϵ of zero in X the construction of an extended operator $S : K + \epsilon \subseteq X \to Y$ is an important ingredient in our approximation procedure. The extension of the domain allows us to consider the effect of a small disturbance in the input signal. Such disturbances are unavoidable in the modelling process.

This section describes a generic approach and is concerned with applicable conditions that will allow the simulated system to represent the real system to within an arbitrarily prescribed accuracy. The problem of

relating the various error bounds to the dimensions of the model is not considered and may be more effectively resolved in a specific context. One could certainly consider these questions at the level of our particular examples.

There are other aspects of the approximation process which we do not consider here. A real system is normally causal and may also be stationary or have finite memory. These systems are studied in Section 3.

2.5.1 Relation to previous studies

The extension of the classical Stone-Weierstrass theorem to the approximation of continuous mappings on topological vector spaces by polynomial mappings has been known for some time and was developed by Prenter [104], Istratescu [65], Prolla and Machado [108] and Bruno [12]. In these papers the approximation procedure relies directly on the classical theorem via an underlying algebra of real valued polynomials.

Our procedure is essentially an elaboration of the procedure used by Bruno but is more explicitly constructive and we believe more directly related to the representation of real dynamical systems. In particular we show that the model is stable to small disturbances in the input signal. We have also considered the role of parameters in the representation process and have adapted our methods accordingly. Our procedure is not limited to polynomial approximation. On the other hand our analysis is restricted to locally convex topological vector spaces. The present work is developed from an approach used by Torokhti [151, 152, 153, 154].

2.5.2 A remark on the compactness condition

The assumption of compactness for the set K on which the operator F is to be approximated is an important part of the modelling process and cannot be totally removed. For a continuous real valued function on the real line it is well known that uniform approximation by a polynomial can be guaranteed only on a compact subset.

We believe that the compactness assumption is reasonable in practice. Suppose the dynamical system is defined by an operator $F : X \to Y$ where X and Y are topological vector spaces. Some knowledge of the operator is necessary if we wish to simulate the given system. It may happen that F is known only on the basis of a finite subset

$$\{(x_n, y_n) \mid x_n \in X \text{ and } y_n = F(x_n) \in Y \text{ for } n = 1, \ldots, N\} \subseteq X \times Y \quad (2.3)$$

or alternatively on a set

$$\{(x_\gamma, y_\gamma) \mid x_\gamma \in X \text{ and } y_\gamma = F(x_\gamma) \in Y \text{ for } \gamma \in \Gamma \subseteq \mathbb{R}^n\} \subseteq X \times Y \quad (2.4)$$

where Γ is compact. Such knowledge may be empirical or based on a restricted analysis of the system concerned. Of course there may be some situations where the compactness assumption is not reasonable. If the set on which the approximation is required is not compact then a stronger continuity condition is needed. In a subsequent section we will use stronger topological assumptions to consider this more difficult problem.

2.5.3 Generic approximant representation in topological vector spaces

The constructive approximation of nonlinear operators in topological vector spaces encounters some special difficulties. First, a topological vector space is not an algebra, as the ordinary multiplication operation is not defined. This necessitates using the structure (2.2) for the approximating operator. Secondly, the structure (2.2) is quite general. It is not clear what specific kind of $k-$linear operator is preferable in (2.2). In the studies [151]–[57] by Torokhti and Howlett, these difficulties have been overcome by the further elaboration of the ideas of Prenter [104] and Bruno [12].

One of the main aims in [151]–[57] is the constructive definition of an operator S to approximate the operator $F : K \to Y$ given on the the compact set K of the topological vector space X possessing the Grothendieck property of approximation (see Definition 4 above), with values in the topological vector space Y. Furthermore there are certain properties that must be satisfied by S if we wish to construct a useful model of a real nonlinear system.

The generic structure of the approximating operator S in [151]–[57] is as follows.

Let X and Y be topological vector spaces with the Grothendieck property of approximation and with approximating sequences $\{G_m\}_{m=1,2,\ldots}$ and $\{H_n\}_{n=1,2,\ldots}$ of continuous linear operators

$$G_m \in \mathcal{L}(X, X_m) \quad \text{and} \quad H_n \in \mathcal{L}(Y, Y_n),$$

where $X_m \subseteq X, Y_n \subseteq Y$ are subspaces of dimension m, n as described in Definition 14. Write

$$X_m = \left\{ x_m \in X \mid x_m = \sum_{j=1}^{m} a_j u_j \right\}$$

and

$$Y_n = \left\{ y_n \in Y \mid y_n = \sum_{k=1}^{n} b_k v_k \right\},$$

where

$$a = (a_1, a_2, \ldots, a_m) \in \mathbb{R}^m, \quad b = (b_1, b_2, \ldots, b_n) \in \mathbb{R}^n$$

and

$$\{u_j\}_{j=1}^m, \quad \{v_k\}_{k=1}^n$$

are bases in X_m, Y_n respectively. Let $\{g\} = \mathcal{G}$ be an algebra of continuous functions $g : \mathbb{R}^m \to \mathbb{R}$ that satisfies the conditions of Stone's Algebra [113]. Define the operators

$$Q \in \mathcal{L}(X_m, \mathbb{R}^m), \quad Z : \mathbb{R}^m \to \mathbb{R}^n \quad \text{and} \quad W \in \mathcal{L}(\mathbb{R}^n, Y_n)$$

by the formulæ

$$Q(x_m) = a, \; Z(a) = (g_1(a), g_2(a), \ldots, g_n(a)) \text{ and } W(z) = \sum_{k=1}^{n} z_k v_k, \quad (2.5)$$

where each $g_k \in \mathcal{G}$ and $z_k = g_k(a)$.

Then $S : X \to Y_n$ is defined by the composition

$$S = WZQG_m. \quad (2.6)$$

In the following sections, it will be shown that subject to an appropriate choice of the functions $\{g_k\} \in \mathcal{G}$, the operator S supplies an approximation to F with any preassigned accuracy on both a compact set and noncompact set in the corresponding topological vector spaces. Moreover [57] the generic structure (2.6) of operator S provides so–called weak interpolation to a nonlinear mapping in the space $\mathcal{C}([0, 1])$.

A diagram for the realization of the approximating operator S is given in Section 1.8.

2.5.4 Preliminaries

We begin with some preliminary results.

Definition 5. *Let X, Y be real Hausdorff topological vector spaces and let A be a subset of X. The map $F : A \to Y$ is uniformly continuous on A if for each open neighbourhood of zero*[2] *$\tau \subseteq Y$ we can find a neighbourhood of zero $\sigma \subseteq X$ such that*

$$F\left[(x + \sigma) \cap A\right] \subseteq F(x) + \tau \quad (2.7)$$

for all $x \in A$.

[2] *In a topological vector space a set τ with $0 \in int(\tau)$ will be called a* neighbourhood of zero.

Lemma 1. *Let X, Y be real Hausdorff topological vector spaces and let K be a compact subset of X. If $F : K \to Y$ is continuous on K then it is uniformly continuous on K.*

Proof. Let τ be a neighbourhood of zero in Y. Choose a neighbourhood of zero $\nu \subseteq Y$ with $\nu - \nu \subseteq \tau$. For each $x \in K$ we choose a neighbourhood of zero $\mu(x) \subseteq X$ such that

$$F\left[(x + \mu(x)) \cap K\right] \subseteq F(x) + \nu. \tag{2.8}$$

Now choose a neighbourhood of zero $\sigma(x) \subseteq X$ such that $\sigma(x) + \sigma(x) \subseteq \mu(x)$. We write

$$\Omega(x) = x + \sigma(x). \tag{2.9}$$

Since

$$K \subseteq \bigcup_{x \in K} \Omega(x) \tag{2.10}$$

and since K is compact we can find a finite subcollection $\Omega_1, \Omega_2, \ldots, \Omega_r$ (where we write $\Omega_i = x_i + \sigma_i, \sigma_i = \sigma(x_i)$ and $\mu_i = \mu(x_i)$) such that

$$K \subseteq \bigcup_{i=1}^{r} \Omega_i. \tag{2.11}$$

Define

$$\sigma = \bigcap_{i=1}^{r} \sigma_i. \tag{2.12}$$

It is clear that σ is an open neighbourhood of zero in X. If we choose any $x \in K$ then we can find k such that $x \in \Omega_k$. Thus

$$F(x) \in F(x_k) + \nu. \tag{2.13}$$

Since

$$\begin{aligned} x + \sigma &\subseteq \Omega_k + \sigma \\ &\subseteq (x_k + \sigma_k) + \sigma_k \\ &\subseteq x_k + \mu_k \end{aligned}$$

it follows that

$$F\left[(x + \sigma) \cap K\right] \subseteq F(x_k) + \nu \tag{2.14}$$

and hence

$$\begin{aligned} F\left[(x+\sigma)\cap K\right]-F(x) &= \{F\left[(x+\sigma)\cap K\right]-F(x_k)\} \\ &\qquad -\{F(x)-F(x_k)\} \\ &\subseteq \nu-\nu \\ &\subseteq \tau. \end{aligned} \tag{2.15}$$

Therefore

$$F\left[(x+\sigma)\cap K\right]\subseteq F(x)+\tau. \tag{2.16}$$

The lemma is proven. □

Lemma 2. *Let X be a normal*[3] *topological vector space and let Y be a locally convex topological vector space. Let K be a compact subset of X and $F : K \to Y$ a continuous map. For each convex neighbourhood of zero $\tau \subseteq Y$ there exists a neighbourhood of zero $\sigma \subseteq X$ and a continuous map $\mathcal{F}_\sigma : K+\sigma \to Y$ in the form*

$$\mathcal{F}_\sigma(u)=\sum_{i=1}^{r}\kappa_i(u)F(x_i) \tag{2.17}$$

where $x_i \in K$ for each $i = 1, 2, \ldots, r$ and where $\kappa_i : K+\sigma \to \mathbb{R}$ is continuous with

(i) $\kappa_i(u) \in [0,1]$, *and*

(ii) $\displaystyle\sum_{i=1}^{r}\kappa_i(u)=1,$

such that

$$F(x)-\mathcal{F}_\sigma(u)\in\tau \tag{2.18}$$

whenever $x \in K$ and $x-u \in \sigma$.

Proof. . Choose a neighbourhood of zero $\mu \subseteq X$ so that for all $x \in K$

$$F[(x+\mu)\cap K]\subseteq F(x)+\tau \tag{2.19}$$

and choose a neighbourhood of zero $\sigma \subseteq X$ with $\sigma+\sigma \subseteq \mu$. Let $\Omega_x = x+\sigma$. Since

$$K\subseteq\bigcup_{x\in K}\Omega_x \tag{2.20}$$

[3] *A topological vector space is said to be* normal *if for each pair of disjoint closed sets $A, B \subseteq X$ there exists a pair of disjoint open sets $U, V \subseteq X$ with $A \subseteq U$ and $B \subseteq V$.*

we can find a finite subcollection $\Omega_1, \Omega_2, \ldots, \Omega_r$ such that

$$K \subseteq \bigcup_{i=1}^{r} \Omega_i. \tag{2.21}$$

Since X is a normal topological vector space we can construct a collection of continuous functions $\kappa_i : K + \sigma \to \mathbb{R}$ for each $i = 1, 2, \ldots, r$ with the properties that

1. $\kappa_i(u) \in [0, 1]$,
2. $\sum_{i=1}^{r} \kappa_i(u) = 1$, and
3. $\kappa_i(u) = 0$ for $u \notin \Omega_i$.

We define a map $\mathcal{F}_\sigma : K + \sigma \to Y$ by the formula

$$\mathcal{F}_\sigma(u) = \sum_{i=1}^{r} \kappa_i(u) F(x_i). \tag{2.22}$$

Now $\kappa_i(u) \neq 0$ implies $u \in \Omega_i$ and if $x - u \in \sigma$ then we have $x \in x_i + \mu$. Hence if $x \in K$ then

$$F(x) - F(x_i) \in \tau \tag{2.23}$$

and so

$$\begin{aligned} F(x) - \mathcal{F}_\sigma(u) &= \sum_{i=1}^{r} \kappa_i(u)[F(x) - F(x_i)] \\ &= \sum_{\{i \mid \kappa_i(u) \neq 0\}} \kappa_i(u)[F(x) - F(x_i)] \\ &\in \tau \end{aligned} \tag{2.24}$$

since the right hand side is a convex combination and τ is a convex set. □

Corollary 1. *If in addition to the conditions of Lemma 6 we have $F(0) = 0$ then we can choose $\mathcal{F}^*_\sigma : K + \sigma \to Y$ such that $\mathcal{F}^*_\sigma$ satisfies the conditions of Lemma 6 and also satisfies $\mathcal{F}^*_\sigma(0) = 0$.*

Proof. Choose a neighbourhood of zero $\sigma_0 \subseteq X$ such that $F(\sigma_0) \subseteq \tau$. Choose another neighbourhood of zero $\sigma \subseteq X$ such that $\sigma + \sigma \subseteq \sigma_0$ and such that

$$F(x) - \mathcal{F}_\sigma(u) \in \tau \tag{2.25}$$

whenever $x \in K$ and $x - u \in \sigma$. In accordance with Urysohn's Lemma [31] there is a continuous function $f : X \to [0, 1]$ such that $f(0) = 0$ and such that $f(u) = 1$ when $u \notin \sigma$. Let

$$\mathcal{F}^*_\sigma(u) = f(u)\mathcal{F}_\sigma(u). \tag{2.26}$$

When $u \notin \sigma$ we have

$$\begin{aligned} F(x) - \mathcal{F}^*_\sigma(u) &= F(x) - \mathcal{F}_\sigma(u) \\ &\in \tau \end{aligned}$$

and when $u \in \sigma$ we have $x \in \sigma_0$ and hence

$$\begin{aligned} F(x) - \mathcal{F}^*_\sigma(u) &= [1 - f(u)]F(x) + f(u)[F(x) - \mathcal{F}_\sigma(u)] \\ &\in \tau, \end{aligned}$$

since

$$F(x) \in \tau \quad \text{and} \quad F(x) - \mathcal{F}_\sigma(u) \in \tau$$

and the right hand side is a convex combination. □

Remark 3. *The condition $F(0) = 0$ in Corollary 1 can be interpreted as follows. If the operator F is the mathematical model of some dynamical system then the output y is related to the input x by $y = F(x)$. Thus the condition $F(0) = 0$ means that a zero input produces a zero output.*

2.5.5 Constructive determination of approximating operator S on compact set of locally convex topological vector space

Recall that our aim has been the constructive determination of an operator S to approximate the given operator F. Furthermore there are certain properties that must be satisfied by S if we wish to construct a useful model of the real system.

Subject to an appropriate choice of the functions $\{g_k\} \in \mathcal{G}$ we now show that S given by (2.6) supplies the required approximation to F.

Theorem 9. *Let X, Y be locally convex topological vector spaces with the Grothendieck property of approximation and let X be normal. Let $K \subseteq X$ be a compact set and $F : K \to Y$ a continuous map. For a given convex neighbourhood of zero $\tau \subseteq Y$ there exists a neighbourhood of zero $\sigma \subseteq X$ with an associated continuous operator $S : X \to Y_n$ in the form*

$$S = WZQG_m$$

and a neighbourhood of zero $\epsilon \subseteq X$ such that for all $x \in K$ and all $x' \in X$ with $x' - x \in \epsilon$ we have

$$F(x) - S(x') \in \tau. \tag{2.27}$$

Remark 4. *This theorem can be regarded as a generalization of the well known Weierstrass approximation theorem.*

To prove the theorem we need to establish that certain sets are compact. Since $G_m \in \mathcal{L}(X, X_m)$ is continuous and since K is compact it follows that $G_m(K)$ is compact. To show that the set $QG_m(K)$ is compact we need to establish that $Q \in \mathcal{L}(X_m, \mathbb{R}^m)$. Since

$$Q(\sum_{j=1}^{m} a_j u_j) = a$$

we need to show that there exists a constant M_Q and a seminorm $\rho : X \to \mathbb{R}$ with

$$\|a\| \leq M_Q \rho(\sum_{j=1}^{m} a_j u_j).$$

We have the following preliminary results.

Lemma 3. *Let X be a locally convex topological vector space and let X_m be the subspace defined above. We can find a sequence $\{\rho_s\}_{s=1,2,\ldots,r}$ of seminorms $\rho_s : X \to \mathbb{R}$ where $r \leq m$ such that the function $\rho : X \to \mathbb{R}$ defined by*

$$\rho(x) = \left[\sum_{s=1}^{r} \{\rho_s(x)\}^2\right]^{\frac{1}{2}} \tag{2.28}$$

is a norm on X_m.

Proof. Let $\{\rho_s\}_{s=1,2,\ldots,r}$ be a sequence of seminorms and let $\rho : X \to \mathbb{R}$ be the function defined above. Let

$$N_s = \{x \mid x \in X_m \text{ and } \rho_s(x) = 0\}$$

for each $s = 1, 2, \ldots, r$ and let $N = \{x \mid x \in X_m \text{ and } \rho(x) = 0\}$. It is easily shown that

1. ρ is a seminorm,
2. N_s is a subspace of X_m for each $s = 1, 2, \ldots, r$, and

3. $N = \bigcap_{s=1}^{r} N_s$ is also a subspace of X_m.

Since X is a locally convex linear topological space we can choose a sequence $\{\rho_s\}_{s=1,2,\ldots,r}$ of seminorms so that $r \leq m$ and $N = \{0\}$. In this case the function $\rho : X \to \mathbb{R}$ defined above is the required norm on X_m. □

Lemma 4. *Let X be a locally convex topological vector space and let X_m be the subspace defined above. If $\rho : X \to \mathbb{R}$ is a norm on X_m then we can find $\alpha > 0$ such that*

$$\rho(\sum_{j=1}^{m} a_j u_j) \geq \alpha \|a\| \tag{2.29}$$

for each $a \in \mathbb{R}^m$.

Proof. It is sufficient to prove that there exists some $\alpha > 0$ with

$$\rho(\sum_{j=1}^{m} a_j u_j) \geq \alpha \tag{2.30}$$

whenever $\|a\| = 1$. If not we can find $\{a^{(p)}\}_{p=1,2,\ldots}$ such that

$$\rho(\sum_{j=1}^{m} a_j^{(p)} u_j) < \frac{1}{p} \tag{2.31}$$

and $\|a^{(p)}\| = 1$. Thus we can find a convergent subsequence (which for convenience we also denote by $\{a^{(p)}\}_{p=1,2,\ldots}$) with $a^{(p)} \to a$ as $p \to \infty$ for some $a \in \mathbb{R}^m$. It now follows that

$$\rho(\sum_{j=1}^{m} a_j u_j) = 0$$

and also that $\|a\| = 1$. But

$$\begin{aligned} \rho(\sum_{j=1}^{m} a_j u_j) = 0 \quad &\Rightarrow \quad \sum_{j=1}^{m} a_j u_j = 0 \\ &\Rightarrow \quad a = 0. \end{aligned}$$

Since $\|a\| = 1$ this is a contradiction. □

We are now able to prove Theorem 11.

Proof. By the approximation property of the space X, for any neighbourhood of zero $\xi \subseteq X$ and for all $x \in K$, we can find $M = M(\xi) > 0$ such that

$$G_m(x) - x \in \xi$$

for $m > M$. By Lemma 6 we can choose σ and a continuous map $\mathcal{F}_\sigma : K + \sigma \to Y$ given by

$$\mathcal{F}_\sigma(u) = \sum_{i=1}^{r} \kappa_i(u) F(x_i) \tag{2.32}$$

with the property that

$$F(x) - \mathcal{F}_\sigma(u) \in \frac{\tau}{4} \tag{2.33}$$

when $x - u \in \sigma$ and hence if we choose $\xi \subseteq \sigma$ and $m > M$ then

$$F(x) - \mathcal{F}_\sigma G_m(x) \in \frac{\tau}{4} \tag{2.34}$$

for each $x \in K$. If we write

$$G_m(x) = \sum_{j=1}^{m} a_j u_j \tag{2.35}$$

then

$$\begin{aligned} \mathcal{F}_\sigma G_m(x) &= \mathcal{F}_\sigma(\sum_{j=1}^{m} a_j u_j) \\ &= \sum_{i=1}^{r} \kappa_i(\sum_{j=1}^{m} a_j u_j) F(x_i) \end{aligned} \tag{2.36}$$

and hence

$$\begin{aligned} H_n \mathcal{F}_\sigma G_m(x) &= \sum_{k=1}^{n} b_k \left[\sum_{i=1}^{r} \kappa_i(\sum_{j=1}^{m} a_j u_j) F(x_i) \right] v_k \\ &= \sum_{k=1}^{n} f_k(a) v_k. \end{aligned} \tag{2.37}$$

We note that $\mathcal{F}_\sigma G_m(K) \subseteq Y$ is a compact subset. By the approximation property of the space Y, for any given neighbourhood of zero $\nu \subseteq Y$, we can choose $N_m > 0$ so that

$$H_n \mathcal{F}_\sigma G_m(x) - \mathcal{F}_\sigma G_m(x) \in \nu \tag{2.38}$$

for all $x \in K$ when $n > N_m$. We also note that

$$\begin{aligned} H_n\mathcal{F}_\sigma G_m(x) - S(x) &= H_n\mathcal{F}_\sigma G_m(x) - WZQG_m(x) \\ &= \sum_{k=1}^{n}[f_k(a) - g_k(a)]v_k. \end{aligned} \tag{2.39}$$

If we suppose that the algebra $\mathcal{G}$ satisfies the conditions of Stone's Algebra then since $a \in QG_m(K)$ and since $QG_m(K)$ is compact it follows that we can choose $\{g_k\}_{k=1,2,\ldots,n} \in \mathcal{G}$ so that

$$H_n\mathcal{F}_\sigma G_m(x) - S(x) \in \nu. \tag{2.40}$$

Thus, if we choose $\nu \subseteq \frac{\tau}{8}$, then

$$\begin{aligned} \mathcal{F}_\sigma G_m(x) - S(x) &\in \frac{\tau}{8} + \frac{\tau}{8} \\ &\subseteq \frac{\tau}{4} \end{aligned}$$

and hence

$$\begin{aligned} F(x) - S(x) &\in \frac{\tau}{4} + \frac{\tau}{4} \\ &\subseteq \frac{\tau}{2}. \end{aligned}$$

Finally we note that

$$\begin{aligned} S(x) - S(x + \Delta x) &= \sum_{k=1}^{n}[g_k(a) - g_k(a + \Delta a)]v_k \\ &\in \frac{\tau}{2} \end{aligned}$$

where $\Delta a \in \mathbb{R}^m$ is defined by

$$G_m(x + \Delta x) = \sum_{j=1}^{m}(a_j + \Delta a_j)u_j, \tag{2.41}$$

provided we choose $\Delta x \in \epsilon$ where ϵ is a sufficiently small neighbourhood of zero in X. Now it follows that

$$\begin{aligned} F(x) - S(x') &\in \frac{\tau}{2} + \frac{\tau}{2} \\ &\subseteq \tau \end{aligned}$$

where $x' = x + \Delta x$. □

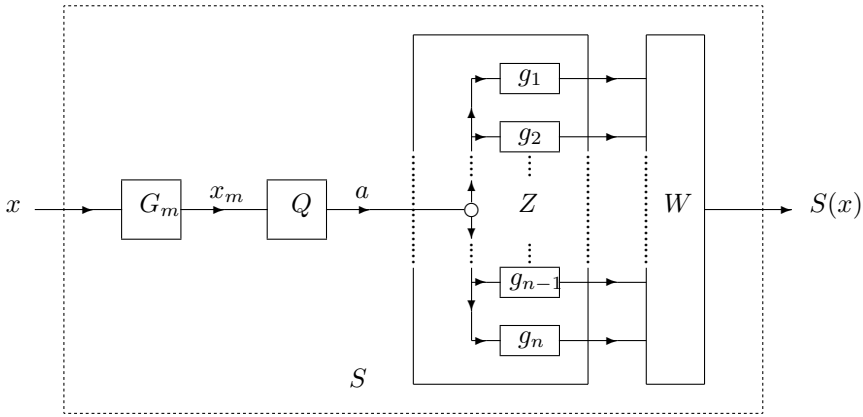


Figure 2.1: Block diagram for realization of S.

Theorem 11 can have the following interpretation. The operator S represents a mathematical model of the real system [109, 152, 56, 57]. A diagram for the realization of the approximating operator S is given Fig. 1.

In this context x is the input signal, $F(x)$ is the output signal from the real system, Δx is the noise that is added to the input signal in practice, and $S(x + \Delta x)$ is the output signal from the constructed system. Thus S is a practical realization of the given abstract operator F. Note that the noise term in the input signal could result from truncation of the parametric description.

We illustrate this theorem with examples.

2.5.6 Examples

Example 3. *Let $X = Y = \mathcal{C}[-1, 1]$ be the Banach space of continuous functions on $[-1, 1]$ with the uniform norm*

$$\|f\| = \sup_{t\in[-1,1]} |f(t)|.$$

For each $\gamma = (\gamma_1, \gamma_2, \gamma_3) \in \mathbb{R}^3$ define

$$x_\gamma(t) = \gamma_1 \cos(\gamma_2 t + \gamma_3)$$

and

$$y_\gamma(t) = (\gamma_1 \cos\gamma_3)^4 \cosh(\gamma_2 t)$$

and let $K \subseteq X$ *be the compact set given by*

$$K = \{x \mid x = x_\gamma \text{ for some } \gamma \in \Gamma = [0,1] \times [\frac{1}{2}, 1] \times [0, 2\pi] \subseteq \mathbb{R}^3\}. \quad (2.42)$$

Let the non-linear operator

$$F : K \to L = F(K) \subseteq Y$$

be defined by the formula

$$F(x_\gamma) = y_\gamma \quad (2.43)$$

and consider the dynamical system described by the mapping $F : K \to L$.

We wish to construct a practical model of the given system. We suppose that the input signal is disturbed by an additional noise term that is essentially unrelated in structure to the true input signal. In this example we choose to approximate the input signal by a polynomial and hence it is convenient for the noise term to be modelled by a polynomial of the same degree. Thus we assume that the actual input signal x' *is given by*

$$x' = x + \Delta x = x_\gamma + \Delta x \quad (2.44)$$

for some $\gamma \in \Gamma$ *where* $\Delta x = h$ *is an appropriate polynomial. We will also approximate the output signal by a polynomial.*

For some given tolerance $\alpha > 0$ *and a corresponding restriction* $h \in \epsilon$ *on the magnitude of the noise term (in this context* ϵ *is some suitable neighbourhood of zero) we wish to find an operator* $S : K + \epsilon \to Y$ *such that*

$$\|F(x) - S(x')\| < \alpha \quad (2.45)$$

for all $x \in K$ *and all* $x' - x \in \epsilon$.

To construct the operator S *it is necessary to extend the given set* K *of input signals to include the additional noise terms. Some initial discussion of calculation procedures is therefore desirable. To this end let* $\mathcal{P}_s$ *denote the space of polynomials with real coefficients and of degree at most* $s-1$. *We define a Chebyshev projection operator* $\Pi_s : \mathcal{C}[-1,1] \to \mathcal{P}_s$ *by the formula*

$$\Pi_s(u) = \sum_{j=1}^{s} c_j(u) T_{j-1} \quad (2.46)$$

where

$$T_{j-1}(\cos\theta) = \cos(j-1)\theta$$

is the Chebyshev polynomial of the first kind and where the coefficients $c_j = c_j(u)$ *are defined by the formulae*

$$c_1 = \frac{1}{\pi}\int_{-1}^{1} \frac{u(t)dt}{\sqrt{(1-t^2)}}$$

and

$$c_j = \frac{2}{\pi}\int_{-1}^{1} \frac{u(t)T_{j-1}(t)dt}{\sqrt{(1-t^2)}}$$

for each $j = 2, 3, \ldots, s$. *In this example we will not use the integral formulae but will calculate approximate Chebyshev coefficients where necessary by using a standard economization procedure [38].*

Let $X_m = \mathcal{P}_m$ *and* $Y_n = \mathcal{P}_n$. *We define linear operators* $G_m : X \to X_m$ *and* $H_n : Y \to Y_n$ *by setting*

$$G_m = \Pi_m \quad \textit{and} \quad H_n = \Pi_n,$$

respectively. For convenience we will use the following approximate calculation procedure to determine

$$X_m(x_\gamma) = x[\gamma, m] \quad \textit{for} \quad x_\gamma \in K$$

and

$$H_n(y_\gamma) = y[\gamma, n] \quad \textit{for} \quad y_\gamma \in L.$$

For any given values of $\mu, \nu > 0$ *we can choose* $m = m(\mu), n = n(\nu)$ *and polynomials* $p_m, q_m \in \mathcal{P}_m$ *and* $r_n \in \mathcal{P}_n$ *with*

$$p_m(\tau) = \sum_{j=1}^{[\frac{m+1}{2}]} p_{2j-1}\tau^{2j-2}, \quad q_m(\tau) = \sum_{j=1}^{[\frac{m+1}{2}]} q_{2j}\tau^{2j-1}$$

and

$$r_n(\tau) = \sum_{j=1}^{[\frac{n+1}{2}]} r_{2j-1}\tau^{2j-2}$$

such that

$$|p_m(\tau) - \cos\tau| + |q_m(\tau) - \sin\tau| < \mu$$

and

$$|r_n(\tau) - \cosh\tau| < \nu$$

for all $\tau \in [-1, 1]$. *Now define* $x[\gamma, m] \in \mathcal{P}_m$ *and* $y[\gamma, n] \in \mathcal{P}_n$ *by*

$$\begin{aligned} x[\gamma, m](t) &= \gamma_1 \left[(\cos\gamma_3)p_m(\gamma_2 t) - (\sin\gamma_3)q_m(\gamma_2 t)\right] \\ &= \sum_{j=1}^{m} a_j[\gamma, m]t^{j-1} \end{aligned}$$

and

$$y[\gamma, n](t) = (\gamma_1 \cos \gamma_3)^4 r_n(\gamma_2 t) = \sum_{j=1}^{n} b_j[\gamma, n] t^{j-1} \tag{2.47}$$

and note that

$$|x_\gamma(t) - x[\gamma, m](t)| < \mu \tag{2.48}$$

and

$$|y_\gamma(t) - y[\gamma, n](t)| < \nu \tag{2.49}$$

for all $t \in [-1, 1]$.

By observing that

$$G_m(p) = p \quad \textit{and} \quad H_n(q) = q$$

when $p \in \mathcal{P}_m$, $q \in \mathcal{P}_n$ *and by using the linearity of the operators* G_m, H_n *we can extend the above calculation procedure to polynomial neighbourhoods of* K, L.

Since the hypothetical input signal $x_\gamma \in K$ *is approximated by a polynomial* $x[\gamma, m] \in \mathcal{P}_m$ *we suppose that the noise term is also modelled by a polynomial* $h \in \mathcal{P}_m$. *Thus we assume that*

$$h(t) = \sum_{j=1}^{m} w_j t^{j-1}. \tag{2.50}$$

where $w = (w_1, w_2, \ldots, w_m) \in \mathbb{R}^m$ *is an unknown constant.*

At this stage we need to point out that we will not follow the specific construction procedure described in our theoretical development. In this example the compact set K *is described by a parameter* $\gamma \in \Gamma$ *and the preceding theory suggests that we should choose an appropriate neighbourhood of zero* $\sigma \subseteq X$ *and construct an operator*

$$\mathcal{F}_\sigma : K + \sigma \to Y$$

by choosing a finite collection $\{\gamma^{(i)}\}_{i=1,2,\ldots,r}$ *of points in* Γ *and an appropriate partition of unity. In practice it is often easier to choose a neighbourhood of zero* $\zeta \subseteq X_m \subseteq X$ *and use direct methods to construct an operator*

$$\begin{aligned} \mathcal{R}_\zeta : G_m(K + \zeta) &= G_m(K) + \zeta \\ &\to Y_n \end{aligned}$$

which effectively replaces the operator $\mathcal{F}_\sigma$ used in the theoretical development by providing an approximate representation of the formal composition $H_n\mathcal{F}_\sigma {G_m}^{-1}$. We will show that the operator

$$R : G_m(K) \to H_n(L)$$

given by

$$R(x[\gamma, m]) = y[\gamma, n] \tag{2.51}$$

can be extended to provide the desired approximation. To define R we note that

$$a_k[\gamma, m] = \begin{cases} {\gamma_1}{\gamma_2}^{k-1}(\cos\gamma_3)p_k & \text{if } k \text{ is odd} \\ -\gamma_1{\gamma_2}^{k-1}(\sin\gamma_3)q_k & \text{if } k \text{ is even} \end{cases} \tag{2.52}$$

and

$$b_k[\gamma, m] = \begin{cases} {\gamma_1}^4{\gamma_2}^{k-1}(\cos\gamma_3)^4 r_k & \text{if } k \text{ is odd} \\ 0 & \text{if } k \text{ is even.} \end{cases} \tag{2.53}$$

In particular we note that

$$b_k[\gamma, m] = \left(\frac{a_1[\gamma, m]}{p_1}\right)^3 \left(\frac{a_k[\gamma, m]}{p_k}\right) r_k \tag{2.54}$$

for each $k = 1, 2, \ldots, n$. Therefore if we define

$$g_k(a) = \left(\frac{a_1}{p_1}\right)^3 \left(\frac{a_k}{p_k}\right) r_k \tag{2.55}$$

for each $k = 1, 2, \ldots, n$ and set

1. $G_m(x_\gamma) = x[\gamma, m]$,
2. $Q(x[\gamma, m]) = (a_1[\gamma, m], a_2[\gamma, m], \ldots, a_m[\gamma, m]) = a[\gamma, m]$,
3. $Z(a) = (g_1(a), g_2(a), \ldots, g_n(a)) = g(a)$, *and*
4. $W(b[\gamma, n]) = y[\gamma, n]$.

then the desired operator

$$R : G_m(K) \to H_n(L)$$

is given by

$$R = WZQ.$$

For any fixed neighbourhood of zero $\zeta \in \mathcal{P}_m$ the extended operator

$$\mathcal{R}_\zeta : G_m(K + \zeta) \to Y_n$$

is simply defined by noting that

1. $G_m(x_\gamma + h) = x[\gamma, m] + h$, *and*
2. $Q(x[\gamma, m] + h) = (a_1[\gamma, m] + w_1, a_2[\gamma, m] + w_2, \ldots, a_m[\gamma, m] + w_m)$.

The operator $S : K + \zeta \to Y_n$ *is now given by the composition*

$$S = WZQG_m.$$

We note that

$$S(x[\gamma, m]) = y[\gamma, n]$$

and that

$$\begin{aligned} S(x[\gamma, m] + h)(t) &= \sum_{k=1}^{n} \left(\frac{(a_1[\gamma, m] + w_1}{p_1} \right)^3 \left(\frac{a_k[\gamma, m] + w_k}{p_k} \right) \\ &\quad \times r_k t^{k-1}. \end{aligned} \tag{2.56}$$

Suppose that the actual level of approximation required is given by

$$\|F(x) - S(x')\| < .01. \tag{2.57}$$

Of course it is necessary to understand that the achievable level of approximation will be limited by the magnitude of h. *By the same token we can only quantify this limitation when we have decided on the precise structure of* S. *To begin the process we let* $m = n = 3$ *and construct the necessary polynomial approximations by applying a standard Chebyshev economization procedure [38] to the appropriate Maclaurin series. We have*

$$\begin{aligned} \cos\tau &\approx 1 - \frac{\tau^2}{2} + \frac{\tau^4}{24} - \frac{\tau^6}{720} \\ &= \frac{1763}{2304}T_0(\tau) - \frac{353}{1536}T_2(\tau) + \frac{19}{3840}T_4(\tau) - \frac{1}{23040}T_6(\tau) \\ &\approx \frac{1763}{2304}T_0(\tau) - \frac{353}{1536}T_2(\tau) \\ &= \frac{4585}{4608} - \frac{353}{768}\tau^2 \\ &= p_3(\tau), \end{aligned} \tag{2.58}$$

$$\begin{aligned} \sin\tau &\approx \tau - \frac{\tau^3}{6} + \frac{\tau^5}{120} \\ &= \frac{169}{192}T_1(\tau) - \frac{5}{128}T_3(\tau) + \frac{1}{1920}T_5(\tau) \\ &\approx \frac{169}{192}T_1(\tau) \\ &= \frac{169}{192}\tau \\ &= q_3(\tau), \end{aligned} \tag{2.59}$$

and

$$\begin{aligned}\cosh\tau &\approx 1+\frac{\tau^2}{2}+\frac{\tau^4}{24}+\frac{\tau^6}{720}\\ &= \frac{2917}{2304}T_0(\tau)+\frac{139}{512}T_2(\tau)+\frac{7}{1280}T_4(\tau)+\frac{1}{23040}T_6(\tau)\\ &\approx \frac{2917}{2304}T_0(\tau)+\frac{139}{512}T_2(\tau)\\ &= \frac{4583}{4608}-\frac{139}{256}\tau^2\\ &= r_3(\tau). \end{aligned} \tag{2.60}$$

With these approximations it can be seen that

$$|p_3(\tau)-\cos\tau|+|q_3(\tau)-\sin\tau|<.05 \tag{2.61}$$

and

$$|r_3(\tau)-\cosh\tau|<.006 \tag{2.62}$$

for all $\tau\in[-1,1]$. *It follows that*

$$\|x_\gamma - x[\gamma,3]\|<.05$$

and

$$\|y_\gamma - y[\gamma,3]\|<.006.$$

Now we have

1. $x[\gamma,3](t)=\frac{4585}{4608}\gamma_1\cos\gamma_3-\frac{169}{192}\gamma_1\gamma_2(\sin\gamma_3)t-\frac{353}{768}\gamma_1{\gamma_2}^2(\cos\gamma_3)t^2$
2. $a[\gamma,3]=(\frac{4585}{4608}\gamma_1\cos\gamma_3,-\frac{169}{192}\gamma_1\gamma_2\sin\gamma_3,-\frac{353}{768}\gamma_1{\gamma_2}^2\cos\gamma_3)$,
3. $g(a)=(\frac{4583\times(4608)^3}{4585}{a_1}^4,0,\frac{139\times768\times(4608)^3}{256\times353\times(4585)^3}{a_1}^3a_3)$, *and*
4. $y[\gamma,3](t)=\frac{4583}{4608}(\gamma_1\cos\gamma_3)^4-\frac{139}{256}(\gamma_1\cos\gamma_3)^4{\gamma_2}^2t^2$.

In this particular example we suppose that the noise term has the form $h(t)=wt^2$ *where* $|w|<.002$. *Therefore*

$$S(x')(t)=y[\gamma,3](t)+\frac{139\times768\times4608}{256\times353\times4585}(\gamma_1\cos\gamma_3)^3wt^2 \tag{2.63}$$

and hence

$$\|S(x')-y[\gamma,3]\|<.003. \tag{2.64}$$

Therefore

$$\begin{aligned} \|F(x) - S(x')\| &\leq \|y_\gamma - y[\gamma, 3]\| + \|y[\gamma, 3] - S(x')\| \\ &< .006 + .003 \\ &< .01. \end{aligned} \tag{2.65}$$

In the above example we note that the input signal depends on a finite number of real parameters. It is natural to investigate what happens when the error in the input signal is caused by an inherent uncertainty in our knowledge of the parameter values. We will motivate further discussion by considering a second example.

Example 4. *We consider the system described in Example 6 and suppose that the error in the input signal is due entirely to an inherent uncertainty $\Delta\gamma$ in our knowledge of the value of γ. For convenience we write*

$$\gamma' = \gamma + \Delta\gamma \quad \textit{and} \quad x' = x_{\gamma'}.$$

For a sufficiently small neighbourhood of zero $\theta \subseteq \mathbb{R}^3$ and with the same definitions as we used in Example 6 we can define an operator

$$S : K_{\Gamma+\theta} \to Y_n$$

such that

$$S(x_{\gamma'}) = y[\gamma', n] \tag{2.66}$$

whenever $\gamma' - \gamma \in \theta$. Since

$$F(x_\gamma) - S(x') = y_\gamma - y[\gamma', n] \tag{2.67}$$

it follows that

$$\|F(x_\gamma) - S(x')\| \leq \|y_\gamma - y_{\gamma'}\| + \|y_{\gamma'} - y[\gamma', n]\|. \tag{2.68}$$

It is now easy to see that the achievable level of approximation is limited by the uncertainty in γ. In particular we note that

$$\begin{aligned} y_{\gamma'}(t) - y_\gamma(t) &\approx \frac{\partial y}{\partial \gamma_1}(t)\Delta\gamma_1 + \frac{\partial y}{\partial \gamma_2}(t)\Delta\gamma_2 + \frac{\partial y}{\partial \gamma_3}(t)\Delta\gamma_3 \\ &= 4(\gamma_1 \cos\gamma_3)^3(\cos\gamma_3)(\cosh\gamma_2 t)\Delta\gamma_1 \\ &\quad +(\gamma_1 \cos\gamma_3)^4(\sinh\gamma_2 t)t\Delta\gamma_2 \\ &\quad\quad -4(\gamma_1 \cos\gamma_3)^3\gamma_1(\sin\gamma_3)(\cosh\gamma_2 t)\Delta\gamma_3 \end{aligned} \tag{2.69}$$

and hence calculate that

$$\|y_{\gamma'} - y_\gamma\| < \sqrt{(32\cosh^2 1 + \sinh^2 1)}\|\Delta\gamma\|.$$

If we suppose that $\|\Delta\gamma\| < .0005$ *we have*

$$\|y_{\gamma'} - y_\gamma\| < .042. \tag{2.70}$$

Suppose the actual level of approximation required is given by

$$\|F(x) - S(x')\| < .05. \tag{2.71}$$

If we let $m = n = 3$ *as we did in the previous example then we again obtain*

$$\|y_\gamma - y[\gamma, 3]\| < .006 \tag{2.72}$$

and hence

$$\begin{aligned} \|F(x) - S(x')\| &< .042 + .006 \\ &< .05. \end{aligned}$$

Remark 5. *The problem of parameter estimation in signal analysis is well known to electrical engineers and has been studied extensively. The determination of a suitable estimate* $\hat{\gamma}$ *for the parameter* $\gamma \in \mathbb{R}^3$ *in the above examples is a classic single-tone estimation problem and is discussed in detail by Rife and Boorstyn [111]. The estimation procedure is based on the following observation. If we define*

$$E(c) = \frac{1}{2}\int_{-1}^{1} |\gamma_1 \exp[i(\gamma_2 t + \gamma_3)] - c_1 \exp[i(c_2 t + c_3)]|^2 dt \tag{2.73}$$

for each $c \in \mathbb{R}^3$ *then*

$$\begin{aligned} E(c) &= {\gamma_1}^2 \\ &\quad -\gamma_1 c_1 \left[\frac{\sin[(c_2 - \gamma_2) + (c_3 - \gamma_3)] - \sin[(c_2 - \gamma_2) - (c_3 - \gamma_3)]}{c_2 - \gamma_2}\right] \\ &\quad +{c_1}^2 \end{aligned}$$

and it is now easy to establish that

$$\begin{aligned} \min_{c_3} E(c) &= E(c_1, c_2, \gamma_3) \\ &= {\gamma_1}^2 - 2\gamma_1 c_1 \left[\frac{\sin(c_2 - \gamma_2)}{c_2 - \gamma_2}\right] + {c_1}^2, \qquad (2.74) \\ \min_{c_2} E(c_1, c_2, \gamma_3) &= E(c_1, \gamma_2, \gamma_3) \\ &= {\gamma_1}^2 - 2\gamma_1 c_1 + {c_1}^2 \qquad (2.75) \end{aligned}$$

and

$$\begin{aligned} \min_{c_1} E(c_1, \gamma_2, \gamma_3) &= E(\gamma) \\ &= 0. \end{aligned}$$

The estimate $\hat{\gamma}$ for γ is found by an elementary search over a suitably chosen finite set $\{c\} \subseteq \mathbb{R}^3$. On the basis of the above analysis the search procedure can be seen to consist of three consecutive one dimensional searches. When the full signal $\gamma_1 \exp[i(\gamma_2 t + \gamma_3)]$ is not known we define

$$E_1(c) = \frac{1}{2}\int_{-1}^{1} [\gamma_1 \cos(\gamma_2 t + \gamma_3) - c_1 \cos(c_2 t + c_3)]^2 dt \tag{2.76}$$

and search over $\{c\} \subseteq \mathbb{R}^3$ to find the minimum value $E_1(\gamma) = 0$. The search is a true three dimensional search because the problem is no longer separable. On the other hand if the signal $\gamma_1 \cos(\gamma_2 t + \gamma_3)$ is observed for all $t \in (-\infty, \infty)$ then we have

$$\gamma_1 \sin(\gamma_2 s + \gamma_3) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\gamma_1 \cos(\gamma_2 t + \gamma_3)}{s - t} dt \tag{2.77}$$

which reconstructs the full signal and allows us to use the original method. Although our explanation does not consider the influence of noise on the estimation process the above procedure is valid in the presence of additive Gaussian noise.

2.5.7 Simplification of the canonical structure in the approximating operator

Consider application of the above approach in the approximation of real non-linear dynamical systems where the system is completely described by a finite number of real parameters.

Let X, Y be locally convex topological vector spaces and let

$$F : K \subseteq X \to Y$$

be a given continuous map. As above we will consider F as an abstract model of some dynamical system where the sets K and $L = F(K) \subseteq Y$ are understood to be the sets of input and output signals respectively. It may be that both sets depend continuously on a finite number of real parameters. In this regard we will therefore assume the existence of closed and bounded intervals $\Gamma \subseteq \mathbb{R}^m$ and $\Delta \subseteq \mathbb{R}^n$ and continuous maps

$$\varphi : \Gamma \to K_\Gamma \quad \text{and} \quad \psi : \Delta \to L_\Delta = L$$

with

$$K_\Gamma = K^4, \quad \varphi(\gamma) = x_\gamma \tag{2.78}$$

and

$$\psi(\delta) = y_\delta \tag{2.79}$$

where

$$F(x_\gamma) = y_\delta. \tag{2.80}$$

If we also assume that ψ is a homeomorphism (thus we assume that ψ is a one to one map and that ψ^{-1} is continuous) then we effectively assume the existence of a continuous map

$$R : \Gamma \to \Delta$$

defined by the composition

$$R = \psi^{-1} F \varphi$$

which describes the continuous dependence $\delta = R(\gamma)$ of the output parameters on the input parameters. The non-linear system described by the continuous map $F : K_\Gamma \to L_\Delta$ where

$$K_\Gamma = \{x \mid x = x_\gamma \quad \text{where} \ \ \gamma \in \Gamma\} \tag{2.81}$$

and

$$L_\Delta = \{y \mid y = y_\delta \quad \text{where} \ \ \delta \in \Delta\} \tag{2.82}$$

can now be represented in alternative form on the compact set $\Gamma \subseteq \mathbb{R}^m$ by the continuous map $R : \Gamma \to \Delta$.

For each neighbourhood of zero $\eta \subseteq \mathbb{R}^n$ we can use Lemma 6 to find a neighbourhood of zero $\zeta \subseteq \mathbb{R}^m$ and a continuous map

$$\mathcal{R}_\zeta : \Gamma + \zeta \to \mathbb{R}^n$$

such that

$$R(\gamma) - \mathcal{R}_\zeta(\hat{\gamma}) \in \eta$$

whenever $\gamma \in \Gamma$ and $\gamma - \hat{\gamma} \in \zeta$. We choose ζ to be closed and bounded and assume that the map $\mathcal{R}_\zeta$ can be represented approximately on the compact set $\Gamma + \zeta$ by a continuous map

$$Z : \Gamma + \zeta \to \mathbb{R}^n$$

with the property that

$$Z(\hat{\gamma}) - \mathcal{R}_\zeta(\hat{\gamma}) \in \eta$$

whenever $\hat{\gamma} \in \Gamma + \zeta$. This representation is normally constructed from a given algebra $\{g\} = \mathcal{G}$ of continuous functions $g : \mathbb{R}^m \to \mathbb{R}^n$ that satisfies the conditions of Stone's Algebra [113].

In practice our knowledge of the parameter values is subject to systematic and pseudo-random instrumental errors. Thus the assumed parameter value is given by $\gamma' = \gamma + \Delta\gamma$ where γ is the true value and where the error $\Delta\gamma$ is bounded by $\Delta\gamma \in \theta$ for some known neighbourhood of zero $\theta \subseteq \mathbb{R}^m$. We will assume that $\theta + \theta \subseteq \zeta$.

In addition to the problem of instrumental errors it may be necessary to use some prescribed method of approximate calculation to determine the parameter values from measurements of the input signal. To this end we assume that for each θ we can choose a neighbourhood of zero $\xi \subseteq \theta$ and a continuous operator

$$\hat{V}_\xi : K_{\Gamma+\theta} \to \Gamma + \zeta$$

which is used to calculate the parameter value γ and for which the calculated value $\hat{\gamma'} = \hat{V}_\xi(x_{\gamma'})$ satisfies the constraint

$$\hat{\gamma'} - \gamma' \in \xi$$

for all $\gamma' \in \Gamma + \theta$. Thus $\hat{\gamma'} \in \Gamma + \zeta$.

To describe the system we define an operator

$$\hat{S} : K_{\Gamma+\theta} \to L_\Delta + \tau$$

in the form of a composition

$$\hat{S} = WZ\hat{V}_\xi. \tag{2.83}$$

We can now state the following theorem.

Theorem 10. *Let X, Y be locally convex linear topological spaces and let*

$$F : K_\Gamma \subseteq X \to L_\Delta \subseteq Y$$

be a continuous map as described above. Then, for each given neighbourhood of zero $\tau \subseteq Y$, we can find neighbourhoods of zero $\xi \subseteq \theta \subseteq \mathbb{R}^m$ and an operator

$$\hat{S} : K_{\Gamma+\theta} \to L_\Delta + \tau$$

in the form of a composition

$$\hat{S} = WZ\hat{V}_\xi$$

such that

$$F(x_\gamma) - \hat{S}(x_{\gamma'}) \in \tau \tag{2.84}$$

whenever $\gamma \in \Gamma$ and $\gamma' - \gamma \in \theta$.

Proof. In terms of the notation introduced above we define $W = \psi$ and let

$$\eta = \psi^{-1}(\nu)$$

where the neighbourhood of zero $\nu \subseteq Y$ is chosen so that

$$\nu + \nu \subseteq \tau.$$

We also suppose that the neighbourhoods of zero $\xi \subseteq \theta \subseteq X$ are chosen in the manner suggested in the above preamble.

Now we can write

$$\begin{aligned} F(x_\gamma) - W\mathcal{R}_\zeta \hat{V}_\xi(x_{\gamma'}) &= \psi[\psi^{-1}F\varphi](\gamma) - \psi\mathcal{R}_\zeta \hat{V}_\xi(x_{\gamma'}) \\ &= \psi[R(\gamma) - \mathcal{R}_\zeta(\hat{\gamma'})]. \end{aligned} \tag{2.85}$$

We choose ζ so that

$$R(\gamma) - \mathcal{R}_\zeta(\hat{\gamma'}) \in \eta \tag{2.86}$$

whenever $\gamma \in \Gamma$ and $\hat{\gamma'} - \gamma \in \zeta$. Hence

$$\psi[R(\gamma) - \mathcal{R}_\zeta(\hat{\gamma'})] \in \nu. \tag{2.87}$$

Since $\hat{\gamma'} \in \Gamma + \zeta$ it follows that

$$\mathcal{R}_\zeta(\hat{\gamma'}) - Z(\hat{\gamma'}) \in \eta \tag{2.88}$$

and hence

$$\begin{aligned} F(x_\gamma) - \hat{S}(x_{\gamma'}) &= [F(x_\gamma) - W\mathcal{R}_\zeta \hat{V}_\xi(x_{\gamma'})] \\ &\qquad + [W\mathcal{R}_\zeta \hat{V}_\xi(x_{\gamma'}) - \hat{S}(x_{\gamma'})] \\ &= \psi[R(\gamma) - \mathcal{R}_\zeta(\hat{\gamma'})] + \psi[\mathcal{R}_\zeta(\hat{\gamma'}) - Z(\hat{\gamma'})] \\ &\in \nu + \nu \\ &\in \tau. \end{aligned}$$

This completes the proof. □

Remark 6. *Practical considerations allow us, as a rule, to determine R on a set $\Gamma + \zeta$ for some neighbourhood of zero $\zeta \subseteq \mathbb{R}^m$ and hence we can set $\mathcal{R}_\zeta = R$.*

Below, a specification of the above technique will be given for the case of Hilbert space. Such a specialization is motivated by a number of practical modelling problems for real dynamical systems where the input-output mapping is known only on some bounded subset of the input space.

Our main result can be regarded as an extension of the classical Stone-Weierstrass Theorem.

2.6 Approximation on Noncompact Sets in Hilbert Spaces

In this section, we consider the constructive approximation of a non-linear operator that is known on a bounded but not necessarily compact set. To justify our proposed construction we introduce an appropriate topology for each of the various vector spaces and assume that the given mapping is uniformly continuous in the introduced topology.

We will assume that the dynamical system is defined by an abstract non-linear operator

$$F : B \subseteq X \to Y$$

where X and Y are suitable vector spaces and B is a bounded subset of X. In such cases it is desirable to construct a model of the real system with a complete input-output map that preserves, in some approximate sense, the known mapping. The model is normally constructed from an algebra of elementary continuous functions. In particular we wish to construct an operator

$$S : X \to Y$$

which will approximate F with a given accuracy on B and will be stable to *small* disturbances. The operator S defines our model of the real system.

In the preceding sections we considered the approximation of non-linear operator $F : K \subseteq X \to Y$ where X and Y are locally convex linear topological spaces and K is a compact subset of X.

In this section we use stronger topological assumptions to solve an analogous approximation problem for operators defined on bounded but non-compact sets. To obtain the necessary topological structure and the consequent notions of continuity required to prove a theorem in the above form we believe it is necessary to consider an operator $F : B \subseteq H \to Y$ where B is a bounded subset of a separable Hilbert space H and Y is a locally convex linear topological space. By introducing a special class

$$\mathcal{A} = \mathcal{A}(H)$$

of linear operators on H we define a collection of semi-norms $\{\rho_A\}_{A\in\mathcal{A}(H)}$ and an associated weak topology for H. This topology is called the $\mathcal{A}$-weak topology for H and is due to Sazonov [132]. When the operator F is uniformly $\mathcal{A}$-weak continuous on B we will show that it is possible to construct an approximating operator $S : H \to Y$ with the desired properties. This is our main result and is stated in Theorem 11. The nature of our approximation is elaborated in more detail when the output space Y is a Banach space. This result is given in Theorem 26.

2.6.1 Preliminaries

Let H be a separable Hilbert space. Consider the class $\mathcal{A} = \mathcal{A}(H)$ of linear operators $A \in \mathcal{L}(H, H)$ defined by

$$\mathcal{A} = \{A \mid A = T^*T\} \tag{2.89}$$

where $T \in \mathcal{L}(H, H)$ and $T^* \in \mathcal{L}(H, H)$ is the adjoint operator and where

$$\sum_j \|T(u_j)\|^2 < \infty \tag{2.90}$$

for each complete orthonormal set

$$\{u_j\}_{j=1,2,\ldots} \subseteq H.$$

Operators of this type are discussed in [32]. For each $A \in \mathcal{A}$ define a semi-norm

$$\rho_A : H \to \mathbb{R}$$

by the formula

$$\rho_A(x) = \langle A(x), x\rangle^{\frac{1}{2}} = \|T(x)\| \tag{2.91}$$

for all $x \in H$ where $\langle\cdot,\cdot\rangle$ denotes the inner product on H. We use the following convenient notation. Let $\mathcal{Z}$ be the set of integers and let $\mathcal{Z}_+$ denote the set of positive integers. When $s \in \mathcal{Z}_+$ we write

$$\mathcal{A}^s = \{A \mid A = (A_1, A_2, \ldots, A_s) \tag{2.92}$$

where $A_k \in \mathcal{A} \ \forall \ k = 1, 2, \ldots, s\}$.

When $A \in \mathcal{A}^s$ we write

$$\rho_k(x) = \langle A_k(x), x\rangle^{\frac{1}{2}} = \|T_k(x)\|$$

for each $k = 1, 2, \ldots, s$ and each $x \in H$.

We use the following notation.

Let $s \in \mathcal{Z}_+$ and let

$$A = (A_1, A_2, \ldots, A_s) \in \mathcal{A}^s.$$

For each $k = 1, 2, \ldots, s$ let $\rho_k : H \to \mathbb{R}$ be the semi-norms defined above. The function $\rho_A : H \to \mathbb{R}^s$ is defined by

$$\rho_A(x) = (\rho_1(x), \rho_2(x), \ldots, \rho_s(x)) \tag{2.93}$$

for each $A \in \mathcal{A}^s$ and each $x \in H$.

For each finite sequence $\{\alpha_k\}_{k=1,2,\ldots,s}$ of real numbers we write

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_s) \in \mathbb{R}^s.$$

Let $\alpha, \beta \in \mathbb{R}^s$. We write $\alpha > \beta$ if and only if $\alpha_j > \beta_j$ for all $j = 1, 2, \ldots, s$ and we will use the notation

$$\mathbb{R}^s_+ = \{\alpha \mid \alpha \in \mathbb{R}^s \text{ with } \quad \alpha > 0\}. \tag{2.94}$$

We make the following definitions.

Definition 6. *Let $s \in \mathcal{Z}_+$, $A \in \mathcal{A}^s$ and $\alpha \in \mathbb{R}^s$ with $\alpha > 0$. A set $\sigma = \sigma(A, \alpha) \subseteq H$ in the form*

$$\sigma(A, \alpha) = \{u \mid u \in H \quad \textit{and} \quad \rho_A(u) < \alpha\} \tag{2.95}$$

will be called an $\mathcal{A}$-weak neighbourhood of zero.

Note that for each $p \in \mathbb{R}_+$ we write

$$\begin{aligned} p\sigma(A, \alpha) &= \{pu \mid u \in H \text{ and } \quad \rho_A(u) < \alpha\} \\ &= \{u \mid u \in H \text{ and } \quad \rho_A(\frac{u}{p}) < \alpha\} \\ &= \{u \mid u \in H \text{ and } \quad \rho_A(u) < p\alpha\} \\ &= \sigma(A, p\alpha). \end{aligned} \tag{2.96}$$

Definition 7. *A set $U \in H$ is said to be an $\mathcal{A}$-weak open set if for each $u \in U$ there exists an $\mathcal{A}$-weak neighbourhood of zero*

$$\sigma = \sigma(u) = \sigma(A(u), \alpha(u)) \subseteq H$$

such that

$$u + \sigma \subseteq U.$$

It can be shown that the collection $\mathcal{U}$ of all $\mathcal{A}$-weak open sets $U \subseteq H$ forms a topology. Since any positive operator $A \in \mathcal{L}(H, H)$ can be written in the form

$$A = T^*T$$

this topology is identical to the topology used by Sazonov [132] and will be known as the $\mathcal{A}$-weak topology for H. With this topology it can be shown that H is a locally convex linear topological vector space. Henceforth we refer to the original Hilbert space topology as the strong topology. In general we can say that the points of H are more clearly distinguished by the strong topology.

Definition 8. *Let $\sigma = \sigma(A, \alpha)$ be an $\mathcal{A}$-weak neighbourhood of zero. For each set*

$$U \subseteq H$$

a set

$$D = D_U(\sigma) \subseteq H$$

is said to be an $\mathcal{A}$-weak σ-net for the set U and the neighbourhood σ, if for each $x \in U$, there exists $d \in D$ with

$$\rho_A(x - d) < \alpha.$$

We can now prove a basic preliminary result.

Lemma 5. *Let $B \subseteq H$ be a bounded subset. For each $\mathcal{A}$-weak neighbourhood of zero $\sigma = \sigma(A, \alpha) \subseteq H$ there exists a finite $\mathcal{A}$-weak σ-net for the set B.*

Proof. We suppose that

$$B \subseteq \{x \mid x \in H \text{ and } \|x\| \leq b\}$$

where $b > 0$ is a known bound for the set B.

Let

$$\{u_j\}_{j=1,2,\ldots} \subseteq H$$

be a complete orthonormal set. For each $x \in H$ we note that

$$x = \sum_{j=1}^{\infty} \langle x, u_j \rangle u_j. \tag{2.97}$$

Choose an integer m such that

$$\left(\sum_{j=m+1}^{\infty} \|T_k(u_j)\|^2 \right)^{\frac{1}{2}} < \frac{\alpha_k}{2b} \tag{2.98}$$

and choose $\alpha_0 > 0$ with

$$\alpha_0 < \frac{\alpha_k}{2m\|T_k\|} \tag{2.99}$$

for each $k = 1, 2, \ldots, s$. Now choose $q \in \mathcal{Z}_+$ so that

$$q\alpha_0 \leq b < (q+1)\alpha_0$$

and define the finite set

$$\begin{aligned} D &= \{d \mid d = \alpha_0 z \quad \text{for all} \quad z \in H \quad \text{with} \quad z = \sum_{j=1}^{m} p_j u_j \\ &\qquad \text{and} \quad p_j \in [-q, q] \cap \mathcal{Z} \quad \text{for each} \quad j = 1, 2, \ldots, m\}. \end{aligned}$$

The set D is the desired $\mathcal{A}$-weak σ-net for B. We confirm this by noting that for any $x \in B$ we have

$$|\langle x, u_j \rangle| < b$$

and hence we can choose $d = \alpha_0 z \in D$ such that

$$|\langle x, u_j \rangle - \alpha_0 p_j| < \alpha_0$$

for each $j = 1, 2, \ldots, m$.

Therefore

$$\begin{aligned} \rho_k(x - d) &= \|T_k(x - \alpha_0 z)\| \\ &\leq \|\sum_{j=1}^{m} (\langle x, u_j \rangle - \alpha_0 p_j) T_k(u_j)\| + \|\sum_{j=m+1}^{\infty} \langle x, u_j \rangle T_k(u_j)\| \\ &\leq \sum_{j=1}^{m} |\langle x, u_j \rangle - \alpha_0 p_j| \|T_k(u_j)\| \\ &\quad + \left(\sum_{j=m+1}^{\infty} |\langle x, u_j \rangle|^2 \right)^{\frac{1}{2}} \left(\sum_{j=m+1}^{\infty} \|T_k(u_j)\|^2 \right)^{\frac{1}{2}} \\ &< \alpha_k \end{aligned}$$

for each $k = 1, 2, \ldots, s$. This completes the proof. □

Definition 9. *Let Y be a topological vector space. A map $F : B \to Y$ is called uniformly $\mathcal{A}$-weak continuous on $B \subseteq H$ if, for each open neighbourhood of zero $\tau \subseteq Y$, there exists a corresponding $\mathcal{A}$-weak neighbourhood of zero*

$$\sigma = \sigma(A, \alpha) \subseteq H$$

such that

$$F[(x+\sigma)\cap B]\subseteq F(x)+\tau \tag{2.100}$$

for all $x\in B$.

Example 5. *In this example we exhibit a non-linear uniformly* $\mathcal{A}$*-weak continuous map on a set which is closed and bounded but non-compact. We write*

$$\mathcal{S}=L_2([0,1]) \tag{2.101}$$

and define

$$\mathcal{T}=\{y\mid y\in L_2([0,1])\ \textit{and}\ \int_0^1 y(t)dt=0\}. \tag{2.102}$$

For each $x\in\mathcal{S}$ *we define an associated function* $\hat{x}:\mathbb{R}\to\mathbb{R}$ *in the following way. Let* $e_j:\mathbb{R}\to\mathbb{R}$ *be defined by*

$$e_j(s)=\sqrt{2}\sin j\pi s \tag{2.103}$$

for each $s\in\mathbb{R}$ *and each* $j=1,2,\ldots$. *For each* $x\in\mathcal{S}$ *define a corresponding sequence of real numbers* $\{x_j\}_{j=1,2,\ldots}$ *by setting*

$$x_j=\int_0^1 x(s)e_j(s)ds \tag{2.104}$$

and let $\hat{x}:\mathbb{R}\to\mathbb{R}$ *be the associated function defined by*

$$\hat{x}=\sum_{j=1}^{\infty}x_je_j. \tag{2.105}$$

It is easily seen that

$$\sum_{j=1}^{\infty}{x_j}^2<\infty \tag{2.106}$$

and that

$$\hat{x}(-s)=-\hat{x}(s) \tag{2.107}$$

and

$$\hat{x}(s+1)=\hat{x}(s-1) \tag{2.108}$$

for each $s\in\mathbb{R}$. *Furthermore it is well known that*

$$\hat{x}(s)=x(s) \tag{2.109}$$

for almost all $s \in [0,1]$. *We say that the function* $\hat{x}$ *is the Fourier sine series representation for* x *and note that* $\hat{x}$ *is the odd periodic extension of period two for the function* x.

For each $y \in \mathcal{T}$ *we define an associated function* $\check{y} : \mathbb{R} \to \mathbb{R}$ *in the following way. Let* $f_j : \mathbb{R} \to \mathbb{R}$ *be defined by*

$$f_j(t) = \sqrt{2}\cos j\pi t \tag{2.110}$$

for each $t \in \mathbb{R}$ *and each* $j = 1, 2, \ldots$. *For each* $y \in \mathcal{T}$ *define a corresponding sequence of real numbers* $\{y_j\}_{j=1,2,\ldots}$ *by setting*

$$y_j = \int_0^1 y(t) f_j(t) dt \tag{2.111}$$

and let $\check{y} : \mathbb{R} \to \mathbb{R}$ *be the associated function defined by*

$$\check{y} = \sum_{j=1}^{\infty} y_j f_j. \tag{2.112}$$

It is easily seen that

$$\sum_{j=1}^{\infty} {y_j}^2 < \infty \tag{2.113}$$

and that

$$\check{y}(-t) = \check{y}(t) \tag{2.114}$$

and

$$\check{y}(t+1) = \check{y}(t-1) \tag{2.115}$$

for each $t \in \mathbb{R}$. *Furthermore it is well known that*

$$\check{y}(t) = y(t) \tag{2.116}$$

for almost all $t \in [0,1]$. *We say that the function* $\check{y}$ *is the Fourier cosine representation for* y *and note that* $\check{y}$ *is the even periodic extension of period two for the function* y.

We define a non-linear operator $F : \mathcal{S} \to \mathcal{S}$ *in the following way. First we define a linear operator* $A : \mathcal{S} \to \mathcal{T}$ *by setting*

$$\begin{aligned} A[x](t) &= \int_0^1 [u(s-t) - s] x(s) ds \\ &= \bar{X} - X(t) \end{aligned} \tag{2.117}$$

where $u : \mathbb{R} \to \mathbb{R}$ *is the unit step function defined by*

$$u(s) = \begin{cases} 0 & \text{if } s < 0 \\ 1 & \text{if } s > 0, \end{cases} \tag{2.118}$$

$X : [0,1] \to \mathbb{R}$ *is the function defined by*

$$X(t) = \int_0^t x(s)ds \tag{2.119}$$

for each $t \in [0,1]$ *and* $\bar{X}$ *is the average value of* X *given by*

$$\bar{X} = \int_0^1 X(t)dt. \tag{2.120}$$

Note that if $y = A[x]$ *where* $x \in \mathcal{S}$ *then* $y \in \mathcal{T}$. *Now for each* $x \in \mathcal{S}$ *we can define* $F[x] \in \mathcal{S}$ *by the formula*

$$F[x](s) = \frac{1}{2}\int_0^1 [\hat{x}(s-t) + \hat{x}(s+t)]X(t)dt. \tag{2.121}$$

The function $F[x]$ *is defined by a convolution integral and can be interpreted as the symmetric correlation of* x *and* $A[x]$. *Such operators are used frequently in the representation and analysis of non-linear systems. In Fourier series form we have*

$$A[\sum_{j=1}^{\infty} x_j e_j] = \sum_{j=1}^{\infty} \frac{x_j}{\pi j} f_j \tag{2.122}$$

and

$$F[\sum_{j=1}^{\infty} x_j e_j] = \sum_{j=1}^{\infty} \frac{{x_j}^2}{\sqrt{2}\pi j} e_j. \tag{2.123}$$

We will show that F *is uniformly* $\mathcal{A}$*-weak continuous on the unit sphere*

$$\begin{aligned} S(0;1) &= \{x \mid x \in \mathcal{S} \text{ with } \|x\| \leq 1\} \\ &\subseteq \mathcal{S}. \end{aligned}$$

Note that the set $S(0;1)$ *is bounded and closed but is not compact. Let* $\epsilon > 0$ *be an arbitrary positive number. Choose* N *such that*

$$\sum_{j=N+1}^{\infty} \frac{1}{j^2} < \epsilon\pi^2 \tag{2.124}$$

and define operators $T_k : \mathcal{S} \to \mathcal{S}$ *for each* $k = 1, 2, \ldots, N$ *by the formula*

$$T_k(\sum_{j=1}^{\infty} x_j e_j) = x_k e_k \tag{2.125}$$

with associated semi-norms $\rho_k : \mathcal{S} \to \mathbb{R}$ *given by*

$$\rho_k(x) = |x_k| \tag{2.126}$$

and consider the $\mathcal{A}$*-weak neighbourhood of zero* $\sigma \subseteq \mathcal{S}$ *defined by*

$$\sigma = \{h \mid \rho_k(h) < \sqrt{\frac{3\epsilon}{2}} \text{ for each } k = 1, 2, \ldots, N\}. \tag{2.127}$$

Now for

$$x + h, x \in S(0; 1) \quad \text{and} \quad h \in \sigma$$

it follows that

$$\begin{aligned} \|F[x+h] - F[x]\| &= \|\sum_{j=1}^{\infty} \frac{(x_j + h_j)^2 - {x_j}^2}{\sqrt{2}\pi j} e_j\| \\ &= \sum_{j=1}^{\infty} \frac{(2x_j + h_j)^2 {h_j}^2}{2\pi^2 j^2} \\ &\leq \sum_{j=1}^{N} \frac{2{h_j}^2}{\pi^2 j^2} + \sum_{j=N+1}^{\infty} \frac{1}{2\pi^2 j^2} \\ &\leq \epsilon. \end{aligned} \tag{2.128}$$

Thus the uniform $\mathcal{A}$*-weak continuity of* F *on the unit sphere* $S(0; 1)$ *is established.*

We now consider the construction of an auxiliary operator that is defined on the entire space of input signals and which approximates the known operator $F : B \subseteq H \to Y$ on the given set B in a well defined way. This operator will be used in the proof of the main result. We suppose that B is a bounded set. The set of uniformly $\mathcal{A}$-weak continuous maps $F : B \to Y$ will be denoted by $\mathcal{C}_{\mathcal{A}}(B, Y)$.

Lemma 6. *Let* Y *be a locally convex topological vector space and let* $B \subseteq H$ *be a bounded set. Let* $F \in \mathcal{C}_{\mathcal{A}}(B, Y)$*. For each convex neighbourhood of zero* $\tau \subseteq Y$ *there exists a corresponding* $\mathcal{A}$*-weak neighbourhood of zero*

$$\sigma = \sigma(A, \alpha) \subseteq H$$

and an associated continuous map $\mathcal{F}_\sigma : H \to Y$ *in the form*

$$\mathcal{F}_\sigma(u) = \sum_{i=1}^{r} \kappa_i(u) F(x_i) \tag{2.129}$$

where $x_i \in B$ *for each* $i = 1, 2, \ldots, r$ *and where* $\kappa_i : H \to \mathbb{R}$ *is continuous with*

1. $\kappa_i(u) \in [0, 1]$, *and*
2. $\sum_{i=1}^{r} \kappa_i(u) = 1$,

such that

$$F(x) - \mathcal{F}_\sigma(v) \in \tau \tag{2.130}$$

whenever $x \in B$ *and* $x - v \in \sigma$.

Proof. Let $A \in \mathcal{A}^s$ where $s \in \mathcal{Z}_+$ with $s > 0$. Choose $\alpha > 0$ and an associated $\mathcal{A}$-weak neighbourhood of zero $\mu = 2\sigma(A, \alpha) \subseteq H$ so that

$$F[(x + \mu) \cap B] \subseteq F(x) + \tau \tag{2.131}$$

for all $x \in B$. If we also define the $\mathcal{A}$-weak neighbourhood of zero $\sigma = \sigma(A, \alpha) \subseteq H$ then

$$\sigma + \sigma \subseteq \mu.$$

Let $D = D(\sigma) = \{x_i\}_{i=1,2,\ldots,r}$ denote an $\mathcal{A}$-weak σ-net for the set B and let

$$\Omega_i = x_i + \sigma.$$

Therefore

$$B \subseteq \bigcup_{i=1}^{r} \Omega_i. \tag{2.132}$$

Define continuous functions

$$\chi : \mathbb{R} \to \mathbb{R} \quad \text{and} \quad \pi_k : H \to \mathbb{R}$$

for each $k = 1, 2, \ldots, s$ by setting

$$\chi(t) = \max\{1 - |t|, 0\}$$

and

$$\pi_k(u) = \min_{i=1,2,\ldots,r} \rho_k(u - x_i)$$

and construct a collection of continuous functions $\lambda_i : H \to \mathbb{R}$ for each $i = 1, 2, \ldots, r$ by setting

$$\lambda_i(u) = \prod_{k=1}^{s} \chi\left(\frac{2\rho_k(u - x_i)}{\pi_k(u) + \alpha_k}\right). \tag{2.133}$$

Now define $\lambda : H \to \mathbb{R}$ by setting

$$\lambda(u) = \sum_{i=1}^{r} \lambda_i(u)$$

and finally define a collection of continuous functions $\kappa_i : H \to \mathbb{R}$ for each $i = 1, 2, \ldots, r$ given by

$$\kappa_i(u) = \frac{\lambda_i(u)}{\lambda(u)} \tag{2.134}$$

and with the properties that

1. $\kappa_i(u) \in [0, 1]$,
2. $\sum_{i=1}^{r} \kappa_i(u) = 1$, and
3. $\kappa_i(u) = 0$ for $u \notin \Omega_i$.

We define a map $\mathcal{F}_\sigma : H \to Y$ by the formula

$$\mathcal{F}_\sigma(u) = \sum_{i=1}^{r} \kappa_i(u) F(x_i). \tag{2.135}$$

Now $\kappa_i(u) \neq 0$ implies $u \in \Omega_i$ and if $x - u \in \sigma$ then we have

$$x \in x_i + \mu.$$

Hence if $x \in B$ then

$$F(x) - F(x_i) \in \tau \tag{2.136}$$

and so

$$\begin{aligned} F(x) - \mathcal{F}_\sigma(u) &= \sum_{i=1}^{r} \kappa_i(u)[F(x) - F(x_i)] \\ &= \sum_{\{i \mid \kappa_i(u) \neq 0\}} \kappa_i(u)[F(x) - F(x_i)] \\ &\in \tau \end{aligned} \tag{2.137}$$

since the right hand side is a convex combination and τ is a convex set. □

It is important to note that although the operator $\mathcal{F}_\sigma : H \to Y$ is defined on the entire space of input signals it is not necessarily an operator that can be realised in a real system.

2.6.2 Constructive determination of approximating operator S on noncompact set of separable Hilbert space

In this section we establish our main theorem. To explain the statement of the theorem it is convenient to review some standard terminology and introduce some additional notation.

It is easy to see that a separable Hilbert space possesses the Grothendieck property of approximation (see Definition 4 in Section 1.5).

Let $\{u_j\}_{j=1,2,\ldots} \subseteq H$ be a complete orthonormal set. For each $m = 1, 2, \ldots$ let

$$X_m = \{x_m \mid x_m \in H \quad \text{and} \quad \langle x_m, u_j \rangle = 0 \ \ \forall \ \ j > m\} \tag{2.138}$$

and define a sequence $\{U_m\}_{m=1,2,\ldots}$ of continuous linear operators $U_m \in \mathcal{L}(H, X_m)$ given by

$$U_m(x) = \sum_{j=1}^{m} \langle x, u_j \rangle u_j \tag{2.139}$$

for each $x \in H$. For convenience write $a_j = \langle x, u_j \rangle$ for each $j = 1, 2, \ldots, m$.

Let Y be a topological vector space with the Grothendieck property of approximation and with approximating sequence $\{V_n\}_{n=1,2,\ldots}$ of continuous linear operators $V_n \in \mathcal{L}(Y, Y_n)$ where $Y_n \subseteq Y$ is a subspace of dimension n as described in Definition 11. Write

$$Y_n = \{y_n \mid y_n \in Y \text{ and } y_n = \sum_{k=1}^{n} b_k v_k\} \tag{2.140}$$

where $b = (b_1, b_2, \ldots, b_n) \in \mathbb{R}^n$ and $\{v_k\}_{k=1,2,\ldots,n}$ is a basis in Y_n. Let $\{g\} = \mathcal{G}$ be an algebra of continuous functions $g : \mathbb{R}^m \to \mathbb{R}$ that satisfies the conditions of Stone's Algebra [113]. Define the operators

$$Q \in \mathcal{L}(X_m, \mathbb{R}^m)^5, \quad Z : \mathbb{R}^m \to \mathbb{R}^n \quad \text{and} \quad W \in \mathcal{L}(\mathbb{R}^n, Y_n)$$

by the formulae

$$Q(x_m) = a, \ Z(a) = (g_1(a), g_2(a), \ldots, g_n(a)), \text{ and } W(z) = \sum_{k=1}^{n} z_k v_k$$

[5] It is necessary to justify the assertion that $Q \in \mathcal{L}(X_m, \mathbb{R}^m)$. An elementary proof of this assertion was given in Section 1.5.

where each $g_k \in \mathcal{G}$ and $z_k = g_k(a)$ and let $S : H \to Y_n$ be defined by the composition

$$S = WZQU_m. \tag{2.141}$$

The corresponding process of numerical realisation of S is discussed in detail in Section 1.5.

Theorem 11. *Let H be a Hilbert space and let $B \subseteq H$ be a bounded subset. Let Y be a locally convex topological vector space with the Grothendieck property of approximation*[6] *and let $F \in \mathcal{C}_{\mathcal{A}}(B, Y)$ be a uniformly $\mathcal{A}$-weak continuous map. For a given convex neighbourhood of zero $\tau \subseteq Y$ there exists a corresponding $\mathcal{A}$-weak neighbourhood of zero $\sigma = \sigma(A, \alpha) \subseteq H$ with an associated continuous operator $S : H \to Y_n$ in the form*

$$S = WZQU_m$$

and a strong closed neighbourhood of zero $\epsilon \subseteq H$ such that for all $x \in B$ and all $x' \in H$ with

$$x' - x \in \epsilon$$

we have

$$F(x) - S(x') \in \tau. \tag{2.142}$$

Proof. First we show that for any $\mathcal{A}$-weak neighbourhood of zero $\sigma = \sigma(A, \alpha)$ and each $p \in \mathbb{R}_+$ and all $x \in B$ we can find $M = M(\sigma, p) > 0$ such that

$$U_m(x) - x \in p\sigma$$

when $m > M$. Since the map $F : B \to Y$ can be extended by continuity to a map $\bar{F} : \bar{B} \to Y$, where $\bar{B}$ denotes the closure of B, we can suppose without loss of generality that B is closed. Let

$$B \subseteq \{x \mid x \in H \text{ and } \|x\| < b\}.$$

For any given $\mathcal{A}$-weak neighbourhood of zero

$$\sigma = \sigma(A, \alpha) \subseteq H$$

where $A \in \mathcal{A}^s$ for some $s \in \mathcal{Z}_+$ with $s > 0$ and fixed $p \in \mathbb{R}_+$ we can find $M = M(\sigma, p) > 0$ such that

$$\left(\sum_{j=M+1}^{\infty} \|T_k(u_j)\|^2 \right)^{\frac{1}{2}} < \frac{p\alpha_k}{b} \tag{2.143}$$

[6] *That is we suppose the existence of a sequence $\{V_n\}_{n=1,2,\ldots}$ of continuous operators $V_n \in \mathcal{L}(Y, Y_n)$ as described above.*

for each $k = 1, 2, \ldots, s$. Now for each $x \in B$ and $m > M$ it is clear that

$$\rho_A(U_m(x) - x) < r\alpha.$$

Thus

$$U_m(x) - x \in p\sigma.$$

The proof now follows the proof of the corresponding theorem given in Section 1.5. By Lemma 6 we can choose an $\mathcal{A}$-weak neighbourhood of zero σ and a continuous map

$$\mathcal{F}_\sigma : B + \sigma \to Y$$

given by

$$\mathcal{F}_\sigma(u) = \sum_{i=1}^{r} \kappa_i(u) F(x_i) \tag{2.144}$$

with the property that

$$F(x) - \mathcal{F}_\sigma(u) \in \frac{\tau}{4} \tag{2.145}$$

when $x - u \in \sigma$ and hence if we choose $p < 1$ so that

$$p\sigma \subseteq \sigma \quad \text{and} \quad m > M(\sigma, p)$$

then

$$F(x) - \mathcal{F}_\sigma U_m(x) \in \frac{\tau}{4} \tag{2.146}$$

for each $x \in B$. If we write

$$U_m(x) = \sum_{j=1}^{m} a_j u_j \tag{2.147}$$

then

$$\begin{aligned} \mathcal{F}_\sigma U_m(x) &= \mathcal{F}_\sigma\Big(\sum_{j=1}^{m} a_j u_j\Big) \\ &= \sum_{i=1}^{r} \kappa_i\Big(\sum_{j=1}^{m} a_j u_j\Big) F(x_i) \end{aligned} \tag{2.148}$$

and hence

$$\begin{aligned} V_n\mathcal{F}_\sigma U_m(x) &= \sum_{k=1}^{n} b_k \left[\sum_{i=1}^{r} \kappa_i(\sum_{j=1}^{m} a_j u_j)F(x_i)\right] v_k \\ &= \sum_{k=1}^{n} f_k(a)v_k. \end{aligned} \tag{2.149}$$

Note that the set B is bounded and closed and so the set $U_m(B) \subseteq X_m$ is also bounded and closed. Since X_m is finite dimensional it follows that $U_m(B)$ is compact. Thus the set $\mathcal{F}_\sigma U_m(B)$ is also compact. By the approximation property of the space Y, for any given neighbourhood of zero $\nu \subseteq Y$, we can choose $N_m > 0$ so that

$$V_n\mathcal{F}_\sigma U_m(x) - \mathcal{F}_\sigma U_m(x) \in \nu \tag{2.150}$$

for all $x \in B$ when $n > N_m$. We also note that

$$\begin{aligned} V_n\mathcal{F}_\sigma U_m(x) - S(x) &= V_n\mathcal{F}_\sigma U_m(x) - WZQU_m(x) \\ &= \sum_{k=1}^{n}[f_k(a) - g_k(a)]v_k. \end{aligned} \tag{2.151}$$

If we suppose that the algebra $\mathcal{G}$ satisfies the conditions of Stone's Algebra then since $a \in QU_m(B)$ and since $QU_m(B)$ is compact it follows that we can choose $\{g_k\}_{k=1,2,\ldots,n} \in \mathcal{G}$ so that

$$V_n\mathcal{F}_\sigma U_m(x) - S(x) \in \nu. \tag{2.152}$$

Thus, if we choose $\nu \subseteq \frac{\tau}{8}$, then

$$\begin{aligned} \mathcal{F}_\sigma U_m(x) - S(x) &\in \frac{\tau}{8} + \frac{\tau}{8} \\ &\subseteq \frac{\tau}{4} \end{aligned}$$

and hence

$$\begin{aligned} F(x) - S(x) &\in \frac{\tau}{4} + \frac{\tau}{4} \\ &\subseteq \frac{\tau}{2}. \end{aligned}$$

Finally we define $\Delta a \in \mathbb{R}^m$ by setting

$$U_m(x + \Delta x) = \sum_{j=1}^{m}(a_j + \Delta a_j)u_j, \tag{2.153}$$

and we note that

$$\begin{aligned} S(x) - S(x + \Delta x) &= \sum_{k=1}^{n} [g_k(a) - g_k(a + \Delta a)] v_k \\ &\in \frac{\tau}{2} \end{aligned}$$

provided we choose $\Delta x \in \epsilon$ where ϵ is a sufficiently small strong closed neighbourhood of zero in X. In this context we note that the set $U_m(B+\epsilon)$ is compact and that $\|\Delta a\| \leq \|\Delta x\|$. Now it follows that

$$\begin{aligned} F(x) - S(x') &\in \frac{\tau}{2} + \frac{\tau}{2} \\ &\subseteq \tau \end{aligned}$$

where $x' = x + \Delta x$. □

Remark 7. *We would like to emphasize the fact that the above theorem shows that the operator S is stable to small disturbances $\Delta x = x' - x$.*

Remark 8. *The condition that the map $F : B \to Y$ be uniformly $\mathcal{A}$-weak continuous is a relatively strong condition. For example this condition implies that $F : B \to Y$ is uniformly continuous in the strong topology. Note that the latter condition is not sufficient to establish the finite covering required in Lemma* 8.

We can consider the main result of the present paper from a different viewpoint. Let $A \in \mathcal{A}^s$. If the set B is bounded in the strong topology the argument of Lemma 8 *can be used to show that the set $T_k(B) \subseteq H$ is totally bounded. If B is closed then $T_k(B)$ is also closed. Since H is complete it follows that $T_k(B)$ is compact in the strong topology. If we can write the map $F : B \to Y$ as a sum of compositions in the form*

$$F = \sum_{k=1}^{s} E_k T_k$$

where $E_k : T_k(B) \to Y$ then $\mathcal{A}$-weak continuity for F is implied by continuity for each E_k in the strong topology. Since $T_k(B)$ is compact the latter condition implies uniform continuity for $E_k : T_k(B) \to Y$ and this in turn implies uniform $\mathcal{A}$-weak continuity for $F : B \to Y$.

Hence in the case where

$$F = \sum_{k=1}^{s} E_k T_k$$

for some $E_k : T_k(B) \to Y$ our approximation procedure for F is equivalent to the original approximation procedure of Section 1.5 applied simultaneously to each E_k.

2.7 Special Results for Maps into Banach Spaces

We wish to elaborate the results of the previous section in the case where Y is a Banach space. Our specific purpose is to show how the results can be quantified by measuring the modulus of continuity for the operator F.

Let Σ_0 denote the collection of all $\mathcal{A}$-weak neighbourhoods of zero in H. Let $B \subseteq H$ be a bounded set and let $F \in \mathcal{C}_{\mathcal{A}}(B, Y)$ be an $\mathcal{A}$-uniformly continuous map.

Definition 10. *The function $\omega : \Sigma_0 \to \mathbb{R}$ given by*

$$\omega(\sigma) = \sup\{\|F(x) - F(u)\| \mid x \in B, u \in B \quad \textit{and} \quad x - u \in \sigma\} \tag{2.154}$$

is called the $\mathcal{A}$-weak modulus of continuity for $F : B \to Y$.

It is often useful to consider the behaviour of the modulus of continuity when $\sigma = \sigma(A, \alpha)$ for a fixed $A \in \mathcal{A}^s$ where $s \in \mathcal{Z}_+$.

Definition 11. *Let $s \in \mathcal{Z}_+$ and $A \in \mathcal{A}^s$. The function $\omega_A : \mathbb{R}_+^s \to \mathbb{R}$ given by*

$$\omega_A(\alpha) = \omega[\sigma(A, \alpha)] \tag{2.155}$$

is called the A-modulus of continuity for $F : B \to Y$.

We can now restate the assertions of Lemma 6 and Theorem 11 in a more specific form.

Lemma 7. *Let Y be a Banach space and let $B \subseteq H$ be a bounded set. Let $F \in \mathcal{C}_{\mathcal{A}}(B, Y)$. For each real number $\beta > 0$ and each $p \in (0, \frac{1}{2}] \subseteq \mathbb{R}$ there exists a corresponding*

$$\alpha = \alpha(\beta, p) > 0$$

and associated $\mathcal{A}$-weak neighbourhoods of zero

$$\begin{aligned} \sigma &= p\mu \\ &= \sigma(A, p\alpha) \subseteq \mu \\ &= \sigma(A, \alpha) \subseteq H \end{aligned}$$

such that

$$\omega_A(\alpha) \leq \beta$$

and such that the continuous map $\mathcal{F}_\sigma : H \to Y$ satisfies the inequality

$$\|F(x) - \mathcal{F}_\sigma(u)\| \leq \omega_A[\rho_A(u - x) + p\alpha] \tag{2.156}$$

for all $x \in B$.

Proof. Let

$$\tau = \{y \mid y \in Y \quad \text{and} \quad \|y\| < \beta\}. \tag{2.157}$$

Choose an $\mathcal{A}$-weak neighbourhood of zero

$$\mu = \sigma(A, \alpha) \subseteq H$$

such that

$$F[(x + \mu) \cap B] \subseteq F(x) + \tau \tag{2.158}$$

for all $x \in B$. It follows that

$$\omega_A(\alpha) \leq \beta.$$

We now consider the $\mathcal{A}$-weak neighbourhood of zero

$$\sigma = p\mu = \sigma(A, p\alpha)$$

and the associated continuous map $\mathcal{F}_\sigma : H \to Y$ defined in Lemma 6. We observe that

$$\kappa_i(u) \neq 0 \quad \text{implies} \quad \rho_A(u - x_i) < p\alpha.$$

Since

$$\begin{aligned} \rho_A(x - x_i) &\leq \rho_A(x - u) + \rho_A(u - x_i) \\ &\leq \rho_A(x - u) + p\alpha \end{aligned} \tag{2.159}$$

it follows that

$$x - x_i \in \sigma(A, \rho_A(x - u) + p\alpha)$$

and hence

$$\|F(x) - F(x_i)\| \leq \omega_A[\rho_A(x - u) + p\alpha]. \tag{2.160}$$

Since $\kappa_i(u) \in [0, 1]$ we have

$$\begin{aligned} \Big\| \sum_{\{i|\kappa_i(u)\neq 0\}} \kappa_i(u)[F(x) - F(x_i)] \Big\| &\leq \sum_{\{i|\kappa_i(u)\neq 0\}} \kappa_i(u)\|F(x) - F(x_i)\| \\ &\leq \sum_{\{i|\kappa_i(u)\neq 0\}} \kappa_i(u)\omega_A[\rho_A(x - u) + p\alpha] \\ &= \omega_A[\rho_A(x - u) + p\alpha]. \end{aligned}$$

The lemma is proven. □

Theorem 12. *Let H be a Hilbert space and let $B \subseteq H$ be a bounded subset. Let Y be a Banach space with the Grothendieck property of approximation*[7] *and let $F \in \mathcal{C}_{\mathcal{A}}(B, Y)$ be a uniformly $\mathcal{A}$-weak continuous map. For a given real number $\beta > 0$ there exists $\alpha = \alpha(\beta) > 0$ and an $\mathcal{A}$-weak neighbourhood of zero*

$$\sigma = \sigma(A, \frac{\alpha}{4}) \subseteq H$$

such that

$$\omega_A(\alpha) \leq \frac{\beta}{2}$$

with an associated continuous operator $S : H \to Y_n$ in the form

$$S = WZQU_m$$

and a strong neighbourhood of zero $\epsilon \subseteq H$ such that for all $x \in B$ and all $x' \in H$ with $x' - x \in \epsilon$ we have

$$\|F(x) - S(x')\| \leq \omega_A \left(\rho_A[U_m(x - x')] + \frac{\alpha}{2}\right) + \frac{\beta}{2}. \tag{2.161}$$

Proof. Let

$$\tau = \{y \mid y \in Y \quad \text{and} \quad \|y\| < \beta\}. \tag{2.162}$$

From Lemma 7 it follows that we can choose $\alpha > 0$ and an associated $\mathcal{A}$-weak neighbourhood of zero

$$\sigma = \sigma(A, \frac{\alpha}{4})$$

such that

$$\omega_A(\alpha) < \frac{\beta}{2}$$

and

$$\|F(x) - \mathcal{F}_\sigma(u)\| \leq \omega_A \left(\rho_A(x - u) + \frac{\alpha}{4}\right) \tag{2.163}$$

when $x \in B$.

Consider the $\mathcal{A}$-weak neighbourhood of zero

$$\sigma = \sigma(A, \frac{\alpha}{4}).$$

[7] *Once again we suppose the existence of a sequence $\{V_n\}_{n=1,2,\ldots}$ of continuous operators $V_n \in \mathcal{L}(Y, Y_n)$ with the appropriate properties.*

As in the proof of Theorem 11 we can choose $M = M(\sigma)$ such that

$$U_m(x) - x \in \sigma$$

for each $x \in B$ when $m > M$. Now we know that

$$\begin{aligned} \rho_A[x - U_m(x')] &\leq \rho_A[x - U_m(x)] + \rho_A[U_m(x - x')] \\ &\leq \rho_A[U_m(x - x')] + \frac{\alpha}{4} \end{aligned} \tag{2.164}$$

and hence, from the definition of ω_A, it follows that

$$\begin{aligned} \|F(x) - \mathcal{F}_\sigma U_m(x')\| &\leq \omega_A\left(\{\rho_A[U_m(x - x')] + \frac{\alpha}{4}\} + \frac{\alpha}{4}\right) \\ &= \omega_A\left(\rho_A[U_m(x - x')] + \frac{\alpha}{2}\right). \end{aligned} \tag{2.165}$$

Since $U_m(B + \epsilon) \subseteq \mathbb{R}^m$ is compact we can choose N_m such that

$$\|\mathcal{F}_\sigma U_m(x') - V_n \mathcal{F}_\sigma U_m(x')\| \leq \frac{\beta}{4} \tag{2.166}$$

for all

$$x' \in B + \epsilon$$

and

$$n > N_m.$$

We can also choose functions from the algebra $\mathcal{G}$ as described earlier in Theorem 11 so that

$$\|V_n \mathcal{F}_\sigma U_m(x') - S(x')\| \leq \frac{\beta}{4} \tag{2.167}$$

for all

$$x' \in B + \epsilon.$$

The required result follows from the previous three inequalities. □

2.8 Concluding Remarks

In this chapter, we have presented several techniques for nonlinear operator approximation with any pre-assigned accuracy. The special attention has been given to applications of these methods to modelling of nonlinear systems.

We have shown that realistic models for non-linear dynamical systems can be constructed in such a way that the model provides an accurate representation of the *input-output* behavior of the given system and is stable to small disturbances.

We have presented the constructive operator approximation methodology that can be used to provide a useful numerical model of a non-linear system in a realistic situation where there is limited initial information about the system. In particular we have shown that a system defined by an abstract operator known only on some bounded set of input signals can nevertheless be realized by a satisfactory numerical model provided that the operator satisfies certain reasonable continuity requirements.

This page intentionally left blank

Chapter 3

Interpolation of Nonlinear Operators

3.1 Introduction

In this chapter, we consider some fundamental principles of the general theory for nonlinear operator interpolation. Interpolating operators are naturally connected to modelling of nonlinear systems. If a system is given by finite sets of input-output signals, then interpolating operator provides a model of such a system.

The most widely known formula for interpolation is the formula due to Lagrange for real valued functions on the real line. The Lagrange formula has since been extended to the interpolation of mappings on more general vector spaces. Let H be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and let $F : H \to H$ be a continuous mapping. Let

$$\{(x_r, y_r)\}_{r=1,2,\ldots,p} \subseteq H \times H,$$

where $y_r = F(x_r)$ and $x_r \neq x_s$ for $r \neq s$ be a known finite collection of data points. In this case, Prenter [105] has proposed an extended Lagrange interpolation formula for each $x \in H$. A similar formula can also be used on Banach space where the inner products are replaced by suitable linear mappings. The formula was developed in association with a systematic study of multi-linear mappings that formed the basis of a generalized Weierstrass approximation theorem in Banach space.

In Section 3.2, we consider a detailed presentation of the Prenter's results. In Section 3.3, we define and justify a non-Lagrangean procedure for the *weak* interpolation of non-linear operators on $\mathcal{C}([0,1])$.

3.2 Lagrange Interpolation in Banach Spaces

Let X and Y be Banach spaces and let F be an operator mapping X into Y. Let $c_1, \ldots, c_n$ be points of X. *The interpolation problem* is that of funding, for each sequence $\{x_1, \ldots, x_n\}$ of distinct points of X, a polynomial operator P which interpolates $\{c_1, \ldots, c_n\}$ at $\{x_1, \ldots, x_n\}$, so that

$$P(x_i) = c_i$$

for all $i = 1, \ldots, n$. We shall show that there always exists a polynomial of degree $(n-1)$ which solves the interpolation problem.

3.2.1 Fréchet derivatives of operators

If L is n-linear $(n > 1)$, we shall let $\partial_i L$ denote the $n > 1$-linear operator on X into $\mathcal{L}_1[X, Y]$ defined by

$$\partial_i L(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = L(x_1, \ldots, x_{i-1}, \cdot, x_{i+1}, \ldots, x_n),$$

where

$$(L(x_1, \ldots, x_{i-1}, \cdot, x_{i+1}, \ldots, x_n))(x) = L(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n).$$

Definition 12. *Let F be an operator mapping an open subset V of a Banach space X into a Banach space Y. Let $x_0 \in V$. If there exists a linear operator $U \in \mathcal{L}_1[X, Y]$ such that*

$$\|F(x_0 + \Delta x) - F(x_0) - U(\Delta x)\| = o(\|\Delta x\|),$$

then $U = F'(x_0)$ is called the Fréchet derivative of F at x_0. Equivalently,

$$U(x) = \lim_{t \to 0} \frac{F(x_0 + tx) - F(x_0)}{t},$$

where the convergence is uniform on the sphere $\{x \mid \|x\| = 1\}$.

It follows from the definition that if L is bounded, n-linear operator on X, and

$$F(x) = L(x^n),$$

where $x^n = (\underbrace{x, \ldots, x}_{n})$, then

$$F'(x) = \sum_{i=1}^{n} \partial_i L(x^{n-1}).$$

In particular, if L is bilinear and

$$F(x) = L(x^2), \quad \text{then } F'(x) = L(x, \cdot) + L(\cdot, x).$$

If L is symmetric, then, clearly, $F(x) = nL(x^{n-1})$.

We shall need the derivative of W. Let L be n-linear and let $x_1, \ldots, x_n$ be points of X. We let $\partial_i W$ or $W/(x - x_i)$ denote the operator on X into $\mathcal{L}_i[X, Y]$ defined by

$$\partial_i W(z) = L(z - x_1, \ldots, z - x_{i-1}, \cdot, z - x_{i+1}, \ldots, z - x_n).$$

We set

$$\begin{aligned} \partial_i W(z) &= (W/(x - x_i))(z) \\ &= W(z)/(x - x_i). \end{aligned}$$

It should be noted that the operator $W/(x - x_i)$ is completely independent of the x in the denominator; the denominator $(x - x_i)$ is purely symbolic.

Theorem 13. *Let L be bounded, n-linear operator. Let $x_1, \ldots, x_n \in X$, and set*

$$W(x) = L(x - x_1, \ldots, x - x_n).$$

Then

$$W'(x_0) = \sum_{i=1}^{n} W(x_0)/(x - x_i)$$

and, in particular,

$$\begin{aligned} W'(x_i) &= W(x_i)/(x - x_i) \\ &= \partial_i W(x_i). \end{aligned}$$

Proof. Let x_0 be a fixed point of X. Then, using the multilinearity and boundedness of L,

$$\begin{aligned}
&\left\| W(x_0+\Delta x) - W(x_0) - \sum_{i=1}^{n} \frac{W(x_0)}{x-x_i}(\Delta x) \right\| \\
&= \|L(x_0 - x_1 + \Delta x, \ldots, x_0 - x_n + \Delta x) \\
&\quad - L(x_0 - x_1, \ldots, x_0 - x_n \\
&\qquad - \sum_{i=1}^{n} L(x_0 - x_1, \ldots, x_0 - x_{i-1}, \Delta x, x_0 - x_{i+1}, \ldots, x_0 - x_n) \Big\| \\
&\qquad \leq \sum_{k=2}^{n} M_k \|\Delta x\|^k = o(\|\Delta x\|),
\end{aligned}$$

where each M_k is a positive constant arising from $\|L\|$ and from the norms $\|x_0 - x_i\|$, with $i = 1, \ldots, n$. □

3.2.2 The interpolation problem - solution

Let $\mathcal{L}_n[X, Y]$, $n = 0, 1, 2, \ldots$, denote the set of n-linear operators on X into Y. If $X = Y$, we write $\mathcal{L}_n[X]$ and we shall identify $\mathcal{L}_0[X]$ with X. Let L be a bounded n-linear operator in $\mathcal{L}_n[X]$; let $x_1, \ldots, x_n$ be distinct points of X and let

$$W(x) = L(x - x_1, \ldots, x - x_n).$$

Then W is a polynomial of degree n mapping X into X of the form

$$L_n x^n + L_{n-1} x^{n-1} + \ldots + L_1 x + L_0,$$

where $L_n = L$, and $L_0 = (-1)^n L(x_1, \ldots, x_n)$. For example, if L is bilinear,

$$L(x - x_1, x - x_2) = L(x^2) - L(x_1, x) - L(x, x_1) + L(x_1, x_2).$$

Thus, $L_2 = L$, $L_0 = L(x_1, x_2)$, and $L_1 = -L(x_1, \cdot) - L(\cdot, x_2)$.

Also,

$$\frac{W(x)}{(x - x_i)} = \partial_i W(x) = L(x - x_1, \ldots, x - x_{i-1}, \cdot, x - x_{i+1}, \ldots, x - x_n)$$

is a polynomial of degree $(n-1)$ which maps X into $L_1[X]$. We have shown that

$$W'(x) = \sum_{i=1}^{n} W(x)/(x - x_i)$$

so that

$$W'(x_i) = W(x_i)/(x - x_i) = \partial_i W(x_i)$$

is a linear operator. Thus, should $W'(x_i)$ be nonsingular for $i = 1, \ldots, n$, then since

$$l_i(x) = [W'(x_i)]^{-1} W(x)/(x - x_i),$$

l_i would be a linear operator-valued function having the property

$$l_i(x_j) = \delta_{ij} I.$$

Furthermore, for each $x_0 \in X$, it is easily seen that

$$[l_i(x_j)](x_0) = l_i(x)x_0$$

is a polynomial operator of degree $(n-1)$. That is, we have proved

Theorem 14. *If there exists an n-linear operator L such that $[W'(x_i)]^{-1}$ exists for each $i = 1, \ldots, n$, where*

$$W(x) = L(x - x_1, \ldots, x - x_n),$$

then the Lagrange polynomial $y(x)$ of degree $(n-1)$ given by

$$y(x) = \sum_{i=1}^{n} l_i(x) c_i \left(= \sum_{i=1}^{n} l_i(x) F(x_i) \right),$$

where

$$l_i(x) = [W'(x_i)]^{-1} \frac{W(x)}{(x - x_i)} = [W'(x_i)]^{-1} \partial_i W(x_i),$$

solves the interpolation problem (interpolates the operator F at the n distinct points $x_1, \ldots, x_n$ of X).

Thus, to solve the interpolation problem, it is enough to prove that such an n-linear operator exists. It would actually suffice to prove the existence of a family $\{L_1, \ldots, L_n\}$ of n-linear operators having the property that $[W_i'(x_i)]^{-1}$ exists for $i = 1, \ldots, n$, where

$$W_i(x) = L_i(x - x_1, \ldots, x - x_n).$$

If this were the case, we could take

$$y(x) = \sum_{i=1}^{n} [W_i'(x_i)]^{-1} \frac{W_i(x)}{(x - x_i)}(c_i)$$

as our interpolating polynomial. we shall prove the existence of such a family of L_i's.

Theorem 15. *Let $x_1, \ldots, x_n$ be distinct points of a Banach space X. Then for each $i = 1, \ldots, n$ there exists an n-linear operator L_i for which $[W_i'(x_i)]^{-1}$ exists, where*

$$W_i(x) = L_i(x - x_1, \ldots, x - x_n).$$

Furthermore, the L_i's can be chosen so that

$$W_i'(x_i) = I,$$

where I is the identity operator in $\mathcal{L}_1[X]$.

Proof. We start with $i = 1$. We must produce an n-linear operator L_1 for which $W_i'(x_i)$ exists and nonsingular, where

$$W_1(x) = L_1(x - x_1, \ldots, x - x_n).$$

Recall that if such an L_1 exists, then

$$\begin{aligned} W_1'(x_1) &= \frac{W_1(x_1)}{(x - x_1)} = \partial_1 W_1(x_1) \\ &= L_1(\cdot, x_1 - x_2, \ldots, x_1 - x_n), \end{aligned}$$

which belongs to $\mathcal{L}_1[X]$. Also, $\mathcal{L}_1 : X^{n-1} \to \mathcal{L}_1[X]$. With this in mind, let

$$X_{ij} = \operatorname{span}\{x_1 - x_j\}.$$

Since each X_{1j} $(j = 2, \ldots, n)$ is one-dimensional, there exist continuous projections P_{1j} of X onto X_{1j}. Define

$$\tilde{T}_1 : X_{12} \times X_{13} \times \ldots \times X_{1n} \to \mathcal{L}_1[X]$$

by linearity, through the equation

$$\tilde{T}_1(x_1 - x_2, \ldots, x_1 - x_n) = I.$$

Then $\tilde{T}_1$ is a bounded (continuous), $(n-1)$-linear operator in

$$\mathcal{L}_1[X_{12} \times X_{13} \times \ldots \times X_{1n}, Y].$$

That is,

$$\begin{aligned} &\|\tilde{T}_1(a_2(x_1 - x_2), \ldots, a_n(x_1 - x_n))\| \\ &= \|a_2 \ldots a_n \tilde{T}(x_1 - x_2, \ldots, x_1 - x_n)\| \\ &= |a_2 \ldots a_n| \cdot \|I\| \\ &= \frac{1}{\|x_1 - x_2\| \ldots \|x_1 - x_n\|} \|a_1(x_1 - x_2)\| \ldots \|a_n(x_1 - x_n)\|, \end{aligned}$$

so that

$$\|\tilde{T_1}\| = \frac{1}{\|x_1 - x_2\| \dots \|x_1 - x_n\|}.$$

We extend $\tilde{T_1}$ to a continuous, $(n-1)$-linear operator $T_1 : X^{n-1} \to \mathcal{L}_1[X]$ through the projections P_{ij}. That is, we define

$$T_1(y_1, \dots, y_{n-1}) = \tilde{T_1}(P_{12}y_1, \dots, P_{1n}y_{n-1}).$$

Since the operators P_{1j} are linear and continuous, it follows that T_1 is $(n-1)$-linear and continuous. In particular, the map P,

$$P : X^{n-1} \to X_{12} \times X_{13} \times \dots \times X_{1n}$$

given by

$$P(y_2, \dots, y_n) = (P_{12}y_2, \dots, P_{1n}y_n)$$

is continuous, so that the composition

$$\tilde{T_1} \circ P = T_1$$

is continuous.

Now define the n-linear operator L_1 by

$$L_1(y_1, \dots, y_n) = [T_1(y_1, \dots, y_n)](y_1).$$

The n-linearity of L_1 follows directly from the $(n-1)$-linearity of T_1 and the fact that T_1 is linear and operator-valued. The boundedness of T_1 is also apparent. If

$$P_{1k}y_k = a_k \frac{x_1 - x_k}{\|x_1 - x_k\|},$$

then

$$\|P_{1k}y_k\| = |a_k|.$$

Thus,

$$\begin{aligned} L_1(y_1, y_2 \dots, y_{n-1}, y_n) &= [T_1(y_2, \dots, y_n)](y_1) \\ &= (-\tilde{T_1}[P_{12}y_2, \dots, P_{1n}y_n)](y_1) \\ &= \frac{a_2 \dots a_n}{\|x_1 - x_2\| \dots \|x_1 - x_n\|} \\ &\quad \times [\tilde{T_1}(x_1 - x_2, \dots, x_1 - x_n)](y_1) \\ &= \frac{a_2 \dots a_n}{\|x_1 - x_2\| \dots \|x_1 - x_n\|} y_1. \end{aligned}$$

Therefore, if

$$K = \frac{1}{\|x_1 - x_2\| \dots \|x_1 - x_n\|},$$

then

$$\begin{aligned}\|L_1(y_1,\ldots,y_n)\| &= K|a_1|\ldots|a_n|\|y_1\| \\ &= K\|P_{12}y_2\|\ldots\|P_{1n}y_n\|\|y_1\| \\ &\leq \bar{K}\|y_1\|\|y_2\|\ldots\|y_n\|,\end{aligned}$$

since each P_{1k} is a projection and

$$\|P_{1k}y\| = \|P_{1k}\|\|y\|.$$

Now let

$$W_1(x) = L_1(x-x_1,\ldots,x-x_n).$$

Since L_1 is a bounded, n-linear operator, $W_1(x)$ is differentiable and

$$\begin{aligned}W_1'(x_1) &= \frac{W_1(x_1)}{(x-x_1)} \\ &= L_1(\cdot, x_1-x_2,\ldots,x_1-x_n) \\ &= \tilde{T}_1(x_1-x_2,\ldots,x_1-x_n) \\ &= I.\end{aligned}$$

Thus, $W_1'(x_1)$ is a non-singular, linear operator.

A similar line of arguments proves the existence, for each $i=1,\ldots,n$, of an n-linear operator L_i for which $W_i'(x_i)=I$, where

$$W_i(x_i) = L_i(x-x_1,\ldots,x-x_n).$$

This completes the proof of the theorem. □

As a direct result of Theorem 15 we have

Theorem 16. *The interpolation problem can always be solved by a polynomial $y(x)$ of degree $(n-1)$ having a Lagrange representation*

$$y(x) = \sum_{i=1}^{n} l_i(x)c_i,$$

where

$$l_i(x) = [W'(x_i)]^{-1}\frac{W(x)}{(x-x_i)} = [W'(x_i)]^{-1}\partial_i W_i(x)$$

and

$$W_i(x) = L_i(x-x_1,\ldots,x-x_n)$$

for appropriately chosen n-linear operators $L_1, \ldots, L_n$.

In the event X is a Hilbert space with inner product $\langle x, y\rangle$, Theorem 15 also yields a representation theorem. Consider the projection P_{1j} of X onto X given in the proof of Theorem 15. If X is a Hilbert space, then

$$P_{1j}y_j = \left\langle y_j, \frac{x_1 - x_j}{\|x_1 - x_j\|} \right\rangle \frac{x_1 - x_j}{\|x_1 - x_j\|}.$$

Thus,

$$L_1(y_1, \ldots, y_n) = \frac{\langle y_2, x_1 - x_2\rangle\langle y_3, x_1 - x_3\rangle \ldots \langle y_n, x_1 - x_n\rangle}{\|x_1 - x_2\|^2\|x_1 - x_3\|^2 \ldots \|x_1 - x_n\|^2} I(y_1).$$

In particular, since $W_{,1}(x_1) = I$,

$$\begin{aligned} l_1(x) &= I \circ \frac{W_1(x)}{(x - x_1} \\ &= L_1(\cdot, x - x_2, \ldots, x - x_n) \\ &= \frac{\langle x - x_2, x_1 - x_2\rangle\langle x - x_3, x_1 - x_3\rangle \ldots \langle x - x_n, x_1 - x_n\rangle}{\|x_1 - x_2\|^2\|x_1 - x_3\|^2 \ldots \|x_1 - x_n\|^2} I. \end{aligned}$$

Analogously, one can prove that

$$l_j(x) = \left[\prod_{\substack{k=1 \\ k\neq j}}^{n} \langle x - x_k, x_j - x_k\rangle\right] \left[\prod_{\substack{k=1 \\ k\neq j}}^{n} \|x_j - x_k\|\right]^{-1} I.$$

Thus, we arrive at

Theorem 17. *Let X be a Hilbert space with inner product $\langle x, y\rangle$ and let $c_1, \ldots, c_n$ be points of X. Then, for any distinct points $x_1, \ldots, x_n$ of X, the polynomial $y(x)$ of degree $(n-1)$, given by*

$$y(x) = \sum_{i=1}^{n} \frac{\pi_i(x)}{\pi_i(x_i)} c_i,$$

where

$$\pi_i(x) = \prod_{\substack{k=1 \\ k\neq j}}^{n} \langle x - x_k, x_j - x_k\rangle$$

satisfies

$$y(x_i 0 = c_i$$

for $i = 1, \ldots, n$.

Proof. The theorem is evident by inspection; however, it is interesting to note how it followed naturally from the theory of Theorems 15 and 16. □

3.3 Weak Interpolation of Nonlinear Operators

Suppose that the mapping $F : X \to Y$ is defined by an empirical data set

$$D_p = \{(x_r, y_r) \mid y_r = F[x_r] \text{ for all } r = 1, 2, \ldots, p\} \subseteq X \times Y.$$

The natural assumption in general is that the data points (x_r, y_r) are all known elements of $X \times Y$. However when each data point

$$(x_r, y_r) \in \mathcal{C}([0,1]) \times \mathcal{C}([0,1])$$

is an ordered pair of functions and if each function pair is known only by an evaluation vector pair

$$(\xi_r, \eta_r) \in \mathbb{R}^m \times \mathbb{R}^n,$$

where

$$\xi_r = (x_r(s_i)) \quad \text{and} \quad \eta_r = (y_r(t_k))$$

then *the extended Lagrange formula cannot be applied.*

In this section, we consider the so called *weak* interpolation procedure [**?**] which will avoid this difficulty.

It will be shown that the *weak* interpolation can become a *strong* interpolation when the image space is a finite dimensional Chebyshev subspace of $\mathcal{C}([0,1])$ and we also show that for each $\epsilon > 0$ there exists $\delta = \delta(\epsilon) > 0$ and an output evaluation set $\mathcal{N} = \mathcal{N}(\epsilon)$ so that the corresponding *weak* interpolation $S[x]$ provides an approximation in $\mathcal{C}([0,1])$ with

$$\|S[x] - y_r\| < \epsilon$$

when

$$\|x - x_r\| < \delta$$

for each $r = 1, 2, \ldots, p$.

The general form

$$S = WKQG_m$$

of the *weak* interpolation operator is motivated by the structure of the approximating operator considered in the preceding chapter. The desire to develop a *weak* interpolation procedure is motivated by the consideration that in many practical problems the input-output pairs (x_r, y_r) are likely to be known only by an evaluation of each function on some finite subset of $[0,1]$.

The *weak* interpolation procedure will be illustrated by an elementary example.

In the case of the weak interpolation of nonlinear operators, we suppose that the mapping

$$F : \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$$

is defined by an empirical data set

$$D_p = \{(x_r, y_r) \mid y_r = F[x_r] \text{ for each } r = 1, 2, \ldots, p\} \subseteq \mathcal{C}([0,1]) \times \mathcal{C}([0,1])$$

and that the data point (x_r, y_r) is known only by the evaluation vectors $(x_r(s_i)) \in \mathbb{R}^m$ and $(y_r(t_k)) \in \mathbb{R}^n$ on some finite collections

$$\mathcal{M} = \{s_i\} \subseteq [0,1] \quad \text{and} \quad \mathcal{N} = \{t_k\} \subseteq [0,1]$$

of fixed points. The desired interpolation is defined by constructing a mapping

$$S : \mathcal{C}[0,1] \to \mathcal{C}[0,1]$$

with

$$S[u] = S[x]$$

when $u(s_i) = x(s_i)$ for each $i = 1, 2, \ldots, m$ and with

$$S[x_r](t_k) = y_r(t_k)$$

for each $r = 1, 2, \ldots, p$ and each $k = 1, 2, \ldots, n$.

3.3.1 Weak interpolation

We write

$$S[x] \simeq F[x]$$

for all $x \in \mathcal{C}([0,1])$.

Let $\mathcal{M} = \{s_i\}_{i=1,2,\ldots,m}$ where

$$0 = s_1 < s_2 < \ldots < s_{m-1} < s_m = 1$$

and $\mathcal{N} = \{t_k\}_{k=1,2,\ldots,n}$ where

$$0 = t_1 < t_2 < \ldots < t_{n-1} < t_n = 1$$

be ordered collections of fixed points in the interval $[0,1]$ and let

$$E_{\mathcal{M}} : \mathcal{C}([0,1]) \to \mathbb{R}^m$$

and

$$E_{\mathcal{N}} : \mathcal{C}([0,1]) \to \mathbb{R}^n$$

be the linear mappings defined by

$$E_{\mathcal{M}}[x] = \xi, \qquad E_{\mathcal{N}}[y] = \eta \tag{3.1}$$

where

$$\xi = (x(s_i)) \in \mathbb{R}^m \quad \text{and} \quad \eta = (y(t_k)) \in \mathbb{R}^n$$

for each $x, y \in \mathcal{C}([0,1])$. When $E_{\mathcal{M}}[x]$ and $E_{\mathcal{N}}[y]$ are known we say that x is evaluated on $\mathcal{M}$ and y is evaluated on $\mathcal{N}$.

We note that in many situations experimental data will be determined in this way.

Consider a mapping

$$F : \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$$

defined by

$$y_r = F[x_r]$$

for each $r = 1, 2, \ldots, p$, where x_r is evaluated on $\mathcal{M}$ and y_r is evaluated on $\mathcal{N}$.

Definition 13. *We will say that*

$$S : \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$$

is an $(\mathcal{M}, \mathcal{N})$ *weak interpolation of*

$$F : \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$$

if

1. $S[u] = S[x]$ *whenever* $E_{\mathcal{M}}[u] = E_{\mathcal{M}}[x]$, *and*
2. $E_{\mathcal{N}} S[x_r] = E_{\mathcal{N}} F[x_r]$ *for each* $r = 1, 2, \ldots, p$.

Let $\sigma, \tau : \mathbb{R} \to \mathbb{R}$ be continuous and non-decreasing with

$$\sigma(s), \tau(t) \downarrow 0 \quad \text{as} \quad s, t \downarrow -\infty$$

and

$$\sigma(s), \tau(t) \uparrow 1 \quad \text{as} \quad s, t \uparrow \infty.$$

We will use these *sigmoidal* functions [22] to construct an operator S which provides an $(\mathcal{M}, \mathcal{N})$ weak interpolation of F.

We need the following preliminary result.

Lemma 8. *Let*

$$f : \mathbb{R}^{m\times m} \to \mathbb{R}, \quad g : \mathbb{R}^{n\times n} \to \mathbb{R}$$

be defined by

$$f(\Phi) = \det \Sigma, \qquad g(\Psi) = \det T \tag{3.2}$$

where

$$\Phi = (\phi_{ij}) \in \mathbb{R}^{m\times m}, \quad \Psi = (\psi_{kl}) \in \mathbb{R}^{n\times n}$$

and where

$$\Sigma = (\sigma_{ij}) \in \mathbb{R}^{m\times m}, \quad T = (\tau_{kl}) \in \mathbb{R}^{n\times n}$$

are given by

$$\sigma_{ij} = \begin{cases} \phi_{ij} & (i < j) \\ 1 - \phi_{ij} & (i \geq j) \end{cases} \quad \text{and} \quad \tau_{kl} = \begin{cases} \psi_{kl} & (k < l) \\ 1 - \psi_{kl} & (k \geq l). \end{cases} \tag{3.3}$$

Then we can find $\phi, \psi > 0$ such that

$$f(\Phi), g(\Psi) \neq 0$$

when

$$|\phi_{ij}| \leq \phi \quad \text{and} \quad |\psi_{kl}| \leq \psi \quad \forall \quad i, j, k, l.$$

Proof. The result follows by observing that $f(\Phi)$, $g(\Psi)$ are polynomials with

$$f(0) = g(0) = 1.$$

□

For each $j = 1, 2, \ldots, m$ and $l = 1, 2, \ldots, n$ define $\alpha_j, \beta_j, \gamma_l, \delta_l \in \mathbb{R}$ with $\alpha_j, \gamma_l > 0$ such that

$$\sigma(\alpha_1 s_1 + \beta_1) = 1 - \phi, \qquad \tau(\gamma_1 t_1 + \delta_1) = 1 - \psi \tag{3.4}$$

and such that

$$\sigma(\alpha_j s_{j-1} + \beta_j) = \phi, \qquad \tau(\gamma_l t_{l-1} + \delta_l) = \psi \tag{3.5}$$

and

$$\sigma(\alpha_j s_j + \beta_j) = 1 - \phi, \qquad \tau(\gamma_l t_l + \delta_l) = 1 - \psi \tag{3.6}$$

for $j, l > 1$ where $\phi, \psi > 0$ are defined in Lemma 8. We also define $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{m\times m}$, $T = (\tau_{kl}) \in \mathbb{R}^{n\times n}$ by setting

$$\sigma_{ij} = \sigma(\alpha_j s_i + \beta_j), \qquad \tau_{kl} = \tau(\gamma_l t_k + \delta_l) \tag{3.7}$$

for each i, j, k, l. Let

$$\xi_r = (x_{ir}) = (x_r(s_i)) = E_{\mathcal{M}}[x_r] \in \mathbb{R}^m$$

and

$$\eta_r = (y_{kr}) = (y_r(t_k)) = E_{\mathcal{N}}[y_r] \in \mathbb{R}^n$$

and define $X \in \mathbb{R}^{m\times p}$, $Y \in \mathbb{R}^{n\times p}$ by writing

$$X = (\xi_1, \ldots, \xi_p), \qquad Y = (\eta_1, \ldots, \eta_p). \tag{3.8}$$

Since Σ, T are non-singular we can find $\mu_r = (\mu_{jr}) \in \mathbb{R}^m$, $\nu_r = (\nu_{lr}) \in \mathbb{R}^n$ by solving the equations

$$\xi_r = \Sigma\mu_r, \qquad \eta_r = T\nu_r \tag{3.9}$$

for each $r = 1, 2, \ldots, p$. The equations (3.9) can be written in the more explicit form

$$x_r(s_i) = \sum_{j=1}^{m} \mu_{jr}\sigma(\alpha_j s_i + \beta_j), \qquad y_r(t_k) = \sum_{l=1}^{n} \nu_{lr}\tau(\gamma_l t_k + \delta_l) \tag{3.10}$$

for each $i = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, n$ and each $r = 1, 2, \ldots, p$. On the other hand, if we define $M \in \mathbb{R}^{m\times p}$, $N \in \mathbb{R}^{n\times p}$ by writing

$$M = (\mu_1, \ldots, \mu_p), \qquad N = (\nu_1, \ldots, \nu_p) \tag{3.11}$$

then the equations (3.9) can be written collectively in the form

$$X = \Sigma M, \qquad Y = TN. \tag{3.12}$$

In general, for each $x \in \mathcal{C}([0, 1])$, we define

$$\xi = (x(s_i)) = E_{\mathcal{M}}[x] \in \mathbb{R}^m$$

and calculate $\theta = (\theta_j) \in \mathbb{R}^m$ by solving the equation

$$\xi = \Sigma\theta. \tag{3.13}$$

This equation can be written as a system of equations in the more explicit form

$$x(s_i) = \sum_{j=1}^{m} \theta_j\sigma(\alpha_j s_i + \beta_j) \tag{3.14}$$

for each $i = 1, 2, \ldots, m$. Define a mapping $G_m : \mathcal{C}([0, 1]) \to \mathcal{X}_m \subseteq \mathcal{C}([0, 1])$ by the formula

$$G_m[x](s) = \sum_{j=1}^{m} \theta_j\sigma(\alpha_j s + \beta_j) \tag{3.15}$$

and a mapping $Q : \mathcal{X}_m \to \mathbb{R}^m$ by setting

$$Q(G_m[x]) = \theta. \tag{3.16}$$

If we assume that $\mu_1, \ldots, \mu_p$ are linearly independent in $\mathbb{R}^m$ then we can define a mapping $K : \mathbb{R}^m \to \mathbb{R}^n$ by the composition

$$K = N(M^T M)^{-1} M^T. \tag{3.17}$$

Indeed we can find $\lambda = (\lambda_r) \in \mathbb{R}^p$ so that

$$\|M\lambda - \theta\| \tag{3.18}$$

is minimized by solving the equation

$$M^T M\lambda = M^T \theta \tag{3.19}$$

then it follows that $\kappa = K\theta \in \mathbb{R}^n$ can be rewritten in the form

$$\kappa = N\lambda. \tag{3.20}$$

Next we define the mapping $W : \mathbb{R}^n \to \mathcal{Y}_n \subseteq \mathcal{C}([0,1])$ by the formula

$$W[\kappa](t) = \sum_{l=1}^{n} \kappa_l \tau(\gamma_l t + \delta_l) \tag{3.21}$$

and finally $S : \mathcal{C}([0,1]) \to \mathcal{Y}_n$ by the composition

$$S = WKQG_m. \tag{3.22}$$

We have the following basic result.

Theorem 18. *Let*

$$S : \mathcal{C}([0,1]) \to \mathcal{Y}_n \subseteq \mathcal{C}([0,1])$$

be the operator defined above. Then S is an $(\mathcal{M}, \mathcal{N})$ weak interpolation of F.

Proof. Let $u, x \in \mathcal{C}([0,1])$ and suppose that

$$E_{\mathcal{M}}[u] = E_{\mathcal{M}}[x].$$

We must show that $S[u] = S[x]$. If we write

$$E_{\mathcal{M}}[u] = \omega \quad \text{and} \quad E_{\mathcal{M}}[x] = \xi$$

then it follows that $\omega = \xi$. Now we note that

$$G_m[u](s) = \sum_{j=1}^{m} \rho_j \sigma(\alpha_j s + \beta_j) \tag{3.23}$$

and

$$G_m[x](s) = \sum_{j=1}^{m} \theta_j \sigma(\alpha_j s + \beta_j) \tag{3.24}$$

where

$$\rho = (\rho_j) \in \mathbb{R}^m \quad \text{and} \quad \theta = (\theta_j) \in \mathbb{R}^m$$

are determined by solving the equations

$$\omega = \Sigma\rho \quad \text{and} \quad \xi = \Sigma\theta. \tag{3.25}$$

Since

$$\rho = \Sigma^{-1}\omega = \Sigma^{-1}\xi = \theta$$

it follows that $G_m[u] = G_m[x]$ and hence that

$$\begin{aligned} S[u] &= WKQG_m[u] \\ &= WKQG_m[x] \\ &= S[x]. \end{aligned}$$

We must also show that

$$E_{\mathcal{N}} S[x_r] = E_{\mathcal{N}} F[x_r]$$

for all $r = 1, 2, \ldots, p$. Since

$$\xi_r = \Sigma\mu_r \tag{3.26}$$

it follows that

$$G_m[x_r] = \sum_{j=1}^{m} \mu_{jr} \sigma(\alpha_j s + \beta_j) \tag{3.27}$$

and hence that

$$QG_m[x_r] = \mu_r. \tag{3.28}$$

Because

$$N(M^T M)^{-1} M^T \mu_r = \nu_r \tag{3.29}$$

it follows that

$$KQG_m[x_r] = \nu_r \tag{3.30}$$

and hence

$$S[x_r] = W[\nu_r]. \tag{3.31}$$

Consequently

$$S[x_r](t) = \sum_{l=1}^{n} \nu_{lr}\tau(\gamma_l t + \delta_l). \tag{3.32}$$

If we use the notation $z_r = S[x_r]$ and if we define $\zeta_r = (z_{kr}) = (z_r(t_k)) \in \mathbb{R}^n$ then the equations

$$z_r(t_k) = \sum_{l=1}^{n} \nu_{lr}\tau(\gamma_l t_k + \delta_l) \tag{3.33}$$

for each $k = 1, 2, \ldots, n$ can be rewritten in the form

$$\zeta_r = T\nu_r \tag{3.34}$$

and since $\zeta_r = T\nu_r = \eta_r$ it follows that

$$\begin{aligned} E_{\mathcal{N}}S[x_r] &= E_{\mathcal{N}}[z_r] \\ &= E_{\mathcal{N}}[y_r] \\ &= E_{\mathcal{N}}F[x_r] \end{aligned}$$

for each $r = 1, 2, \ldots, p$. □

Example 6. *Let*

$$\mathcal{C}_0 = \{x \mid x \in \mathcal{C}([0,1]) \text{ with } x(0) = x(1) = 0\}. \tag{3.35}$$

Define a linear operator $A : \mathcal{C}_0 \to \mathcal{C}[0,1]$ *by setting*

$$A[x](t) = \int_0^1 [u(s-t) - s]x(s)ds \tag{3.36}$$

where $u : \mathbb{R} \to \mathbb{R}$ *is the unit step function given by*

$$u(s) = \begin{cases} 0 & \text{if} \quad s < 0 \\ 1 & \text{if} \quad s > 0 \end{cases} \tag{3.37}$$

and for each $x \in \mathcal{C}_0$ *define an associated function* $\hat{x} : \mathbb{R} \to \mathbb{R}$ *by taking the odd periodic extension of period two for* x. *This extension can be easily constructed using the Fourier sine series representation for* x. *Let* $F : \mathcal{C}_0 \to \mathcal{C}_0$ *be the mapping defined by the formula*

$$F[x](s) = \frac{1}{2}\int_0^1 [\hat{x}(s-t) + \hat{x}(s+t)]A[x](t)dt \tag{3.38}$$

for each s with $0 < s < 1$. The function $F[x]$ is defined by a convolution integral and can be interpreted as the symmetric correlation of x and $A[x]$. Such operators are used frequently in the representation and analysis of non-linear systems. It is useful to observe that if we represent x as a Fourier sine series

$$x(s) = \sum_{j=1}^{\infty} \xi_j \sin j\pi s \tag{3.39}$$

then

$$F[x](s) = \sum_{j=1}^{\infty} \frac{\xi_j^2}{2j\pi} \sin j\pi s \tag{3.40}$$

for each $s \in [0,1]$. If

$$x_1(s) = \begin{cases} 2s & \text{if} \quad 0 < s < \frac{1}{2} \\ 2 - 2s & \text{if} \quad \frac{1}{2} < s < 1 \end{cases} \tag{3.41}$$

and

$$x_2(s) = \begin{cases} 4s & \text{if} \quad 0 \le s \le \frac{1}{4} \\ 2 - 4s & \text{if} \quad \frac{1}{4} \le s \le \frac{3}{4} \\ 4s - 4 & \text{if} \quad \frac{3}{4} \le s \le 1 \end{cases} \tag{3.42}$$

then elementary calculations show that

$$A[x_1](t) = \begin{cases} \frac{1}{4} - t^2 & \text{if} \quad 0 < t < \frac{1}{2} \\ -\frac{1}{4} + (1-t)^2 & \text{if} \quad \frac{1}{2} < t < 1 \end{cases} \tag{3.43}$$

and

$$A[x_2](t) = \begin{cases} \frac{1}{8} - 2t^2 & \text{if} \quad 0 < t < \frac{1}{4} \\ -\frac{1}{8} + 2(t - \frac{1}{2})^2 & \text{if} \quad \frac{1}{4} < t < \frac{3}{4} \\ \frac{1}{8} - 2(1-t)^2 & \text{if} \quad \frac{3}{4} < t < 1. \end{cases} \tag{3.44}$$

Further calculations give

$$F[x_1](s) = \frac{1}{3}[\frac{1}{4} - \sigma^2][\frac{3}{4} - \sigma^2]$$

and

$$F[x_2](s) = \begin{cases} \frac{8}{3}[\frac{1}{16} - (\sigma - \frac{1}{4})^2][\frac{5}{16} - (\sigma - \frac{1}{4})^2] & \text{if} \quad 0 \le s \le \frac{1}{2} \\ -\frac{8}{3}[\frac{1}{16} - (\sigma - \frac{1}{4})^2][\frac{5}{16} - (\sigma - \frac{1}{4})^2] & \text{if} \quad \frac{1}{2} \le s \le 1 \end{cases}$$

where we have used the notation $\sigma = |s - \frac{1}{2}|$ *for convenience. We suppose that the mapping is not known but that for some* $\mathcal{M}$ *and* $\mathcal{N}$ *the data vectors* $E_{\mathcal{M}}(x_1)$, $E_{\mathcal{M}}(x_2)$, $E_{\mathcal{N}}(F[x_1])$ *and* $E_{\mathcal{N}}(F[x_2])$ *are given. Take*

$$\sigma(s) = \begin{cases} 0 & \text{if} \quad s \le 0 \\ s & \text{if} \quad 0 \le s \le 1 \\ 1 & \text{if} \quad s \ge 1 \end{cases} \tag{3.45}$$

and

$$\tau(t) = \begin{cases} 0 & \text{if} \quad t \le 0 \\ \frac{1}{2}[1 - \cos \pi t] & \text{if} \quad 0 \le t \le 1 \\ 1 & \text{if} \quad t \ge 1. \end{cases} \tag{3.46}$$

Choose

$$\{s_i\} = \{t_k\} = \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\} \tag{3.47}$$

and set $\phi = \psi = 0$. *Choose* $\alpha_j = \gamma_l = 4$ *for each* j, l *and* $\{\beta_j\} = \{\delta_l\} = \{1, 0, -1, -2, -3\}$ *so that*

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} = T. \tag{3.48}$$

The vectors $\xi_r = (x_r(s_i)) \in \mathbb{R}^5$ *and* $\mu_r = \Sigma^{-1}\xi_r \in \mathbb{R}^5$ *for each* $r = 1, 2$ *are given by*

$$\xi_1 = \begin{pmatrix} 0 \\ \frac{1}{2} \\ 1 \\ \frac{1}{2} \\ 0 \end{pmatrix}, \xi_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{pmatrix}, \mu_1 = \begin{pmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}, \mu_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \tag{3.49}$$

and the vectors $\eta_r = (y_r(t_k)) \in \mathbb{R}^5$ *and* $\nu_r = T^{-1}\eta_r \in \mathbb{R}^5$ *for each* $r = 1, 2$ *are given by*

$$\eta_1 = \begin{pmatrix} 0 \\ \frac{11}{256} \\ \frac{1}{16} \\ \frac{11}{256} \\ 0 \end{pmatrix}, \eta_2 = \begin{pmatrix} 0 \\ \frac{5}{96} \\ 0 \\ -\frac{5}{96} \\ 0 \end{pmatrix}, \nu_1 = \begin{pmatrix} 0 \\ \frac{11}{256} \\ \frac{5}{256} \\ -\frac{5}{256} \\ -\frac{11}{256} \end{pmatrix}, \nu_2 = \begin{pmatrix} 0 \\ \frac{5}{96} \\ -\frac{5}{96} \\ -\frac{5}{96} \\ \frac{5}{96} \end{pmatrix}.$$

Therefore M *and* N *are given by*

$$M = \begin{pmatrix} 0 & 0 \\ \frac{1}{2} & 1 \\ \frac{1}{2} & -1 \\ -\frac{1}{2} & -1 \\ -\frac{1}{2} & 1 \end{pmatrix}, N = \begin{pmatrix} 0 & 0 \\ \frac{11}{256} & \frac{5}{96} \\ \frac{5}{256} & -\frac{5}{96} \\ -\frac{5}{256} & -\frac{5}{96} \\ -\frac{11}{256} & \frac{5}{96} \end{pmatrix} \tag{3.50}$$

from which it follows that $M^T M = I$ *and*

$$N(M^T M)^{-1} M^T = \frac{1}{1536} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 113 & -47 & -113 & 47 \\ 0 & -65 & 95 & 65 & -95 \\ 0 & -95 & 65 & 95 & -65 \\ 0 & 47 & -113 & -47 & 113 \end{pmatrix}. \tag{3.51}$$

Now for a general $x \in \mathcal{C}([0,1])$ *we have*

$$\xi = \begin{pmatrix} x(0) \\ x(\frac{1}{4}) \\ x(\frac{1}{2}) \\ x(\frac{3}{4}) \\ x(1) \end{pmatrix} \tag{3.52}$$

and hence $\theta = \Sigma^{-1}\xi$ *is given by*

$$\theta = \begin{pmatrix} x(0) \\ x(\frac{1}{4}) - x(0) \\ x(\frac{1}{2}) - x(\frac{1}{4}) \\ x(\frac{3}{4}) - x(\frac{1}{2}) \\ x(1) - x(\frac{3}{4}) \end{pmatrix} \tag{3.53}$$

and $\kappa = N(M^T M)^{-1} M^T \theta$ *by*

$$\kappa = \frac{1}{1536} \begin{pmatrix} 0 \\ -113x(0) + 160x(\frac{1}{4}) + 66x(\frac{1}{2}) - 160x(\frac{3}{4}) + 47x(1) \\ 65x(0) - 160x(\frac{1}{4}) + 30x(\frac{1}{2}) + 160x(\frac{3}{4}) - 95x(1) \\ 95x(0) - 160x(\frac{1}{4}) - 30x(\frac{1}{2}) + 160x(\frac{3}{4}) - 65x(1) \\ -47x(0) + 160x(\frac{1}{4}) - 66x(\frac{1}{2}) - 160x(\frac{3}{4}) + 113x(1) \end{pmatrix}.$$

For $t \in [0,1]$ we have

$$\tau(4t) = [1 - \cos 4\pi t]/2$$

for all t,

$$\tau(4t-1) = [1 + \cos 4\pi t]/2$$

for $t \geq 1/4$,

$$\tau(4t-2) = [1 - \cos 4\pi t]/2$$

for $t \geq 1/2$ and

$$\tau(4t-3) = [1 + \cos 4\pi t]/2$$

for $t \geq 3/4$ with all functions equal to zero to the left of the specified points. It can now be seen that

$$S[x](t) = \begin{cases} \frac{1}{3072}(-113x(0) + 160x(\frac{1}{4}) + 66x(\frac{1}{2}) - 160x(\frac{3}{4}) + 47x(1)) \\ \qquad \times(1 - \cos 4\pi t) \\ \qquad \text{if} \quad 0 \leq t \leq \frac{1}{4} \\ \frac{1}{3072}[(-48x(0) + 96x(\frac{1}{2}) - 48x(1)) \\ \qquad +(-178x(0) - 320x(\frac{1}{4}) - 36x(\frac{1}{2}) \\ \qquad +320x(\frac{3}{4}) - 142x(1))\cos 4\pi t] \\ \qquad \text{if} \quad \frac{1}{4} \leq t \leq \frac{1}{2} \\ \frac{1}{3072}[(47x(0) - 160x(\frac{1}{4}) + 66x(\frac{1}{2}) + 160x(\frac{3}{4}) - 113x(1)) \\ \qquad +(-143x(0) - 160x(\frac{1}{4}) - 6x(\frac{1}{2}) \\ \qquad +160x(\frac{3}{4}) - 77x(1))\cos 4\pi t] \\ \qquad \text{if} \quad \frac{1}{2} \leq t \leq \frac{3}{4} \\ \frac{1}{3072}[(36x(0) - 72x(\frac{1}{2}) + 36x(1))\cos 4\pi t] \\ \qquad \text{if} \quad \frac{3}{4} \leq t \leq 1. \end{cases}$$

3.4 Strong interpolation

We begin with a simple definition.

Definition 14. *We will say that $S : \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$ is a strong interpolation of $F : \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$ if $S[x_r] = F[x_r]$ for each $r = 1, 2, \ldots, p$.*

To construct a strong interpolation of the operator $F : \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$ we need to make some more assumptions. To this end we have the following definition.

Definition 15. *Let $\mathcal{Y}_n \subseteq \mathcal{C}([0,1])$ be an n-dimensional subspace of $\mathcal{C}([0,1])$ with basis functions $\{\omega_l\}_{l=1,2,\ldots,n}$. We define*

$$T = (\tau_{kl}) = (\omega_l(t_k))$$

and say that $\mathcal{Y}_n$ is a Chebyshev subspace if $\det T \neq 0$ for each collection $\mathcal{N} = \{t_k\}_{k=1,2,\ldots,n}$ of fixed points with $0 = t_1 < t_2 < \ldots < t_{n-1} < t_n = 1$.

Note that for each $y \in \mathcal{Y}_n$ we can find $\nu = (\nu_l) \in \mathbb{R}^n$ such that

$$y = \sum_{l=1}^{n} \nu_l \omega_l \tag{3.54}$$

and note also that ν is uniquely defined and can be calculated by solving the equations

$$y(t_k) = \sum_{l=1}^{n} \nu_l \omega_l(t_k) \tag{3.55}$$

for each $k = 1, 2, \ldots, n$. If we define $\eta = (y(t_k)) \in \mathbb{R}^n$ then the equations (3.55) can be written in the form

$$\eta = T\nu. \tag{3.56}$$

These observations can be used to construct a strong interpolation when we have a mapping $F : \mathcal{C}([0,1]) \to \mathcal{Y}_n \subseteq \mathcal{C}([0,1])$. We use the same basic idea as we used in Section 3.3.1 but use the new matrix T to find $\nu_r \in \mathbb{R}^n$ such that

$$\eta_r = T\nu_r \tag{3.57}$$

for each $r = 1, 2, \ldots, p$. The mapping $W : \mathbb{R}^n \to \mathcal{Y}_n \subseteq \mathcal{C}([0,1])$ is now defined by the formula

$$W[\kappa](t) = \sum_{l=1}^{n} \kappa_l \omega_l(t) \tag{3.58}$$

but all other definitions remain formally the same and in particular the operator $S : \mathcal{C}([0,1]) \to \mathcal{Y}_n$ is defined by the same formal composition

$$S = WKQG_m. \tag{3.59}$$

We have the following theorem.

Theorem 19. *Let $S : \mathcal{C}([0,1]) \to \mathcal{Y}_n \subseteq \mathcal{C}([0,1])$ be the operator defined above. If $\mathcal{Y}_n$ is a Chebyshev subspace then S is a strong interpolation of the operator $F : \mathcal{C}([0,1]) \to \mathcal{Y}_n \subseteq \mathcal{C}([0,1])$.*

Proof. Let $z_r = S[x_r]$ for each $r = 1, 2, \ldots, p$. By following the proof of Theorem 18 we can show that

$$z_r(t) = \sum_{l=1}^{n} \nu_{lr} \omega_l(t). \tag{3.60}$$

On the other hand we know that $y_r \in \mathcal{Y}_n$ implies that we can find $\pi_r = (\pi_{lr}) \in \mathbb{R}^n$ such that

$$y_r(t_k) = \sum_{l=1}^{n} \pi_{lr} \omega_l(t_k) \tag{3.61}$$

for each $k = 1, 2, \ldots, n$. Thus we have

$$\eta_r = T\pi_r \tag{3.62}$$

and since T is non-singular it follows from equation (3.57) that $\pi_r = \nu_r$ and hence that $z_r = y_r$ for all $r = 1, 2, \ldots, p$. □

3.5 Interpolation and approximation

Under certain circumstances we can show that the $(\mathcal{M}, \mathcal{N})$ weak interpolation of Section 3.3.1 also provides an approximation to the mapping $F : \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$. Although the principles of the construction are the same as they were in Section 3.3.1 we will need to be more careful in the way we choose the various parameters that define the interpolation operator. We use essentially the same notation as we used in Section 3.3.1 and suppose that $F : \mathcal{C}([0,1]) \to \mathcal{C}_0 \subseteq \mathcal{C}([0,1])$ where

$$\mathcal{C}_0 = \{y | y \in \mathcal{C}([0,1]) \text{ with } y(0) = y(1) = 0\}. \tag{3.63}$$

Suppose that $y_r = F[x_r] \in \mathcal{C}_0$ is known for each $r = 1, 2, \ldots, p$. Let $\epsilon > 0$ and choose $\delta_1 = \delta_1(\epsilon) > 0$ so that

$$|y_r(t) - y_r(t^*)| < \frac{\epsilon}{4} \tag{3.64}$$

when $|t - t^*| < \delta_1$. Choose $\mathcal{N}$ so that

$$|t_{k+1} - t_k| < \delta_1 \tag{3.65}$$

for each $k = 1, 2, \ldots, n-1$. Choose $\phi, \psi^* > 0$ in the way that $\phi, \psi > 0$ were chosen in Lemma 8. For each $\psi \in (0, \psi^*)$ choose $\gamma_l(\psi), \delta_l(\psi) \in \mathbb{R}$ with $\gamma_l(\psi) > 0$ for $l = 1, 2, \ldots, n$ such that

$$\tau(\gamma_1(\psi)t_1 + \delta_1(\psi)) = 1 - \psi \tag{3.66}$$

and such that

$$\tau(\gamma_l(\psi)t_{l-1} + \delta_l(\psi)) = \psi \tag{3.67}$$

and

$$\tau(\gamma_l(\psi)t_l + \delta_l(\psi)) = 1 - \psi \tag{3.68}$$

for $l > 1$. Define $T_\psi = (\tau_{kl}(\psi)) \in \mathbb{R}^{n\times n}$ by setting

$$\tau_{kl}(\psi) = \tau(\gamma_l(\psi)t_k + \delta_l(\psi)) \tag{3.69}$$

for each k, l. Since T_ψ is non-singular we can find $\nu_r(\psi) = (\nu_{lr}(\psi)) \in \mathbb{R}^n$ by solving the equations

$$\eta_r = T_\psi \nu_r(\psi) \tag{3.70}$$

for each $r = 1, 2, \ldots, p$. The equations (3.70) can be written in the more explicit form

$$y_r(t_k) = \sum_{l=1}^{n} \nu_{lr}(\psi)\tau(\gamma_l(\psi)t_k + \delta_l(\psi)) \tag{3.71}$$

for each $k = 1, 2, \ldots, n$ and each $r = 1, 2, \ldots, p$. On the other hand, if $N_\psi \in \mathbb{R}^{n\times p}$ is defined by writing

$$N_\psi = (\nu_1(\psi), \ldots, \nu_p(\psi)) \tag{3.72}$$

then the equations (3.70) can be written collectively in the form

$$Y = T_\psi N_\psi. \tag{3.73}$$

Define $T_0 = (\tau_{kl}(0)) \in \mathbb{R}^{n\times n}$ by setting

$$\tau_{kl}(0) = \begin{cases} 0 & (k < l) \\ 1 & (k \geq l) \end{cases} \tag{3.74}$$

and note that $T_\psi \to T_0$ as $\psi \to 0$. Since T_0 is non-singular we can define $\nu_r(0) = (\nu_{lr}(0)) \in \mathbb{R}^n$ by solving the equation

$$\eta_r = T_0 \nu_r(0) \tag{3.75}$$

and because $T_\psi \to T_0$ then $\nu_r(\psi) \to \nu_r(0)$ as $\psi \to 0$. Equation (3.75) can be rewritten in the form

$$y_r(t_k) = \sum_{l=1}^{k} \nu_{lr}(0) \tag{3.76}$$

and hence

$$\nu_{lr}(0) = \begin{cases} y_r(t_1) & (l = 1) \\ y_r(t_l) - y_r(t_{l-1}) & (l > 1). \end{cases} \tag{3.77}$$

Define the mapping $S_\psi : \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$ by the composition

$$S_\psi = W K_\psi Q G_m \tag{3.78}$$

where $K_\psi : \mathbb{R}^m \to \mathbb{R}^n$ is defined by

$$K_\psi = N_\psi (M^T M)^{-1} M^T \tag{3.79}$$

and all other operators are defined in the same way as they were in Section 3.3.1. Therefore

$$S_\psi[x_r](t) = \sum_{l=1}^{n} \nu_{lr}(\psi) \tau(\gamma_l(\psi) t + \delta_l(\psi)) \tag{3.80}$$

and from this equation and equation (3.71) we note that

$$S_\psi[x_r](t_k) = y_r(t_k)$$

and in particular that

$$S_\psi[x_r](0) = S_\psi[x_r](1) = 0.$$

Now since

$$\begin{aligned} \kappa(\psi) &= K_\psi \theta \\ &= N_\psi \lambda \\ &= \sum_{r=1}^{p} \lambda_r \nu_r(\psi) \end{aligned} \tag{3.81}$$

it follows that

$$\kappa_l(\psi) = \sum_{r=1}^{p} \lambda_r \nu_{lr}(\psi) \tag{3.82}$$

and hence that

$$\begin{aligned} W[\kappa(\psi)](t_k) &= \sum_{l=1}^{n}\left[\sum_{r=1}^{p}\lambda_r\nu_{lr}(\psi)\right]\tau(\gamma_l(\psi)t_k+\delta_l(\psi)) \\ &= \sum_{r=1}^{p}\lambda_r\left[\sum_{l=1}^{n}\nu_{lr}(\psi)\tau(\gamma_l(\psi)t_k+\delta_l(\psi))\right] \\ &= \sum_{r=1}^{p}\lambda_r S_\psi[x_r](t_k). \end{aligned} \tag{3.83}$$

It is now obvious that

$$W[\kappa(\psi)](0) = W[\kappa(\psi)](1) = 0$$

and hence $S_\psi[x] \in \mathcal{C}_0$ for all $x \in \mathcal{C}([0,1])$. In other words $S_\psi : \mathcal{C}([0,1]) \to \mathcal{C}_0 \subseteq \mathcal{C}([0,1])$.

To show that $\|S_\psi[x] - F[x_r]\|$ can be made arbitrarily small when ψ and $\|x - x_r\|$ are sufficiently small and $\mathcal{N}$ is sufficiently fine we recall that

$$y_r = F[x_r] \quad \text{and} \quad S_\psi[x_r](t_k) = y_r(t_k)$$

and consider the inequality

$$\begin{aligned} |S_\psi[x](t) - y_r(t)| \leq\ & |S_\psi[x](t) - S_\psi[x_r](t)| + |S_\psi[x_r](t) - y_r(t_k)| \\ & + |y_r(t_k) - y_r(t)| \end{aligned} \tag{3.84}$$

where $t \in [0,1]$ and k is chosen so that $t \in [t_{k-1}, t_k]$. For $t \in [t_{k-1}, t_k]$ we note that

$$\begin{aligned} y_r(t_k) - S_\psi[x_r](t) = & \sum_{l=1}^{n}\nu_{lr}(\psi)\left[\tau(\gamma_l(\psi)t_k+\delta_l(\psi))\right. \\ & \left. -\tau(\gamma_l(\psi)t+\delta_l(\psi))\right] \end{aligned}$$

and since

$$\begin{aligned} 1-\psi &\leq \tau(\gamma_l(\psi)t+\delta_l(\psi)) \\ &\leq \tau(\gamma_l(\psi)t_k+\delta_l(\psi)) \\ &\leq 1 \end{aligned}$$

when $l < k$ and

$$\begin{aligned} 0 &\leq \tau(\gamma_l(\psi)t+\delta_l(\psi)) \\ &\leq \tau(\gamma_l(\psi)t_k+\delta_l(\psi)) \\ &\leq \psi \end{aligned}$$

when $l > k$ then

$$|y_r(t_k) - S_\psi[x_r](t)| \leq \sum_{l \neq k} |\nu_{lr}(\psi)|\psi + |\nu_{kr}(\psi)|. \tag{3.85}$$

Since

$$|\nu_{1r}(\psi)| \to 0$$

and

$$|\nu_{lr}(\psi)| \to |y_r(t_l) - y_r(t_{l-1})| < \epsilon/4$$

for $l > 1$ as $\psi \to 0$ it follows we can find ψ so that

$$|\nu_{lr}(\psi)| < \epsilon/4$$

for all $l = 1, 2, \ldots, n$ and if ψ is chosen so that we also have

$$(n-1)\psi < 1$$

then

$$\sum_{l \neq k} |\nu_{lr}(\psi)|\psi + |\nu_{kr}(\psi)| < \frac{\epsilon}{2} \tag{3.86}$$

for all $k = 1, 2, \ldots, n$. It follows that

$$|y_r(t_k) - S_\psi[x_r](t)| < \frac{\epsilon}{2} \tag{3.87}$$

for all $t \in [t_{k-1}, t_k]$. We now consider the value of ψ to be fixed. Incidentally we note that our earlier choice of $\mathcal{N}$ also implies that

$$|y_r(t_k) - y_r(t)| < \frac{\epsilon}{4} \tag{3.88}$$

for $t \in [t_{k-1}, t_k]$. We note that

$$|x(s_i) - x_r(s_i)| < \|x - x_r\|$$

for each $i = 1, 2, \ldots, m$ and hence

$$\|\xi - \xi_r\| < \|x - x_r\|\sqrt{m}. \tag{3.89}$$

Thus

$$\|\theta - \mu_r\| < \|\Sigma^{-1}\| \, \|x - x_r\|\sqrt{m} \tag{3.90}$$

and therefore

$$\|\kappa(\psi) - \nu_r\| < \|K_\psi\| \, \|\Sigma^{-1}\| \, \|x - x_r\|\sqrt{m}. \tag{3.91}$$

It follows that

$$\begin{aligned} |S_\psi[x](t) - S_\psi[x_r](t)| &= |W[\kappa(\psi)](t) - W[\nu_r](t)| \\ &= |\sum_{l=1}^{n} [\kappa_l(\psi) - \nu_{lr}]\tau(\gamma_l(\psi)t + \delta_l(\psi))| \\ &\leq \sum_{l=1}^{n} |\kappa_l(\psi) - \nu_{lr}| \\ &\leq \|\kappa(\psi) - \nu_r\|\sqrt{n} \\ &< \|K_\psi\| \, \|\Sigma^{-1}\| \, \|x - x_r\|\sqrt{mn} \end{aligned} \tag{3.92}$$

for all $t \in [0,1]$ and hence if $\|x - x_r\| < \delta_2$ and we choose δ_2 sufficiently small then

$$\|S_\psi[x] - S_\psi[x_r]\| < \frac{\epsilon}{4}. \tag{3.93}$$

Since $y_r = F[x_r]$, we observe that the inequalities (3.84,3.87,3.88,3.93) imply

$$|S_\psi[x](t) - F[x_r](t)| < \epsilon \tag{3.94}$$

for all $t \in [0,1]$ and hence

$$\|S_\psi[x] - F[x_r]\| < \epsilon. \tag{3.95}$$

We can summarize the preceding discussion in the following way.

Theorem 20. *Let*

$$S_\psi = WK_\psi QG_m : \mathcal{C}([0,1]) \to \mathcal{C}_0 \subseteq \mathcal{C}([0,1])$$

be the operator defined above. For each $\epsilon > 0$ we can choose $\psi = \psi(\epsilon) > 0$ sufficiently small and $\mathcal{N} = \mathcal{N}(\epsilon)$ sufficiently fine and find $\delta_2 = \delta_2(\epsilon) > 0$ such that

$$\|S_\psi[x] - F[x_r]\| < \epsilon \tag{3.96}$$

whenever $\|x - x_r\| < \delta_2$ for each $r = 1, 2, \ldots, p$.

3.5.1 An idle comparison

Although it is inappropriate to compare the *weak* interpolation procedure in this section with the Lagrangean interpolation proposed by Prenter [105] it is nevertheless of some interest to apply the latter interpolation to our earlier example. Of course it is now necessary to assume that the data set is completely known in $\mathcal{C}[0,1]$.

Example 7. *We use the mapping of Example 6 and essentially the same data set. We suppose once again that the mapping is not known but must now assume that the data set is completely known. Since*

$$\begin{aligned} \max|(x_1 - x_2)(s)| &= (x_1 - x_2)(\frac{3}{4}) \\ &= \frac{3}{2} \end{aligned}$$

we can define an associated function $v_{12} : [0,1] \to [0,1] \in NBV([0,1])$ *of normalised bounded total variation by the formula*

$$v_{12}(s) = \begin{cases} 0 & \text{if} \quad 0 < s < \frac{3}{4} \\ 1 & \text{if} \quad \frac{3}{4} < s < 1 \end{cases} \tag{3.97}$$

and a corresponding linear functional $f_{12} \in \mathcal{C}([0,1])^*$ *by setting*

$$\begin{aligned} f_{12}(x) &= \int_{[0,1]} x(s)v(ds) \\ &= x(\frac{3}{4}). \end{aligned}$$

It is clear that

$$\begin{aligned} f_{12}(x_1 - x_2) &= (x_1 - x_2)(\frac{3}{4}) \\ &= \|x_1 - x_2\| \\ &= \frac{3}{2} \end{aligned}$$

Define $f_{21} \in \mathcal{C}([0,1])^*$ *by setting* $f_{21} = -f_{12}$ *and apply the extended Lagrange formula to obtain*

$$\begin{aligned} R[x] &= f_{12}[\frac{x - x_2}{\|x_1 - x_2\|}]F[x_1] + f_{21}[\frac{x - x_1}{\|x_2 - x_1\|}]F[x_2] \\ &= \frac{x(\frac{3}{4}) - x_2(\frac{3}{4})}{\frac{3}{2}}F[x_1] + \frac{x_1(\frac{3}{4}) - x(\frac{3}{4})}{\frac{3}{2}}F[x_2] \end{aligned} \tag{3.98}$$

from which it follows that

$$R[x](t) = \begin{cases} \frac{2}{9}[x(\frac{3}{4})+1][\frac{1}{4}-\tau^2][\frac{3}{4}-\tau^2] \\ \qquad -\frac{16}{9}[x(\frac{3}{4})-\frac{1}{2}][\frac{1}{16}-(\tau-\frac{1}{4})^2][\frac{5}{16}-(\tau-\frac{1}{4})^2] \\ \\ \frac{2}{9}[x(\frac{3}{4})+1][\frac{1}{4}-\tau^2][\frac{3}{4}-\tau^2] \\ \qquad +\frac{16}{9}[x(\frac{3}{4})-\frac{1}{2}][\frac{1}{16}-(\tau-\frac{1}{4})^2][\frac{5}{16}-(\tau-\frac{1}{4})^2] \end{cases}$$

in the intervals

$$0 \le t \le \frac{1}{2} \quad \textit{and} \quad \frac{1}{2} \le t \le 1$$

respectively. The notation $\tau = |t - \frac{1}{2}|$ *has been used for convenience. For a given* $x \in \mathcal{C}([0,1])$ *we could compare the Lagrange interpolation* $R[x]$ *of this example with the interpolation* $S[x]$ *of the previous example by evaluating each function at selected points.*

3.6 Some Related Results

W. Porter [106, 107], extended the result by Prenter [105]] to the case of causal operators. With H a Hilbert space and $\{(x_i, y_i) : i = 1, \cdots, m\} \subset H \times H$ a basic problem in [106] is to determine the existence and uniqueness of causal operators, F, on H satisfying $y_i = F(x_i)$ $i = 1, \cdots, m$. In [106], classes of polynomial functions are considered which minimize an operator norm. The results include explicit necessary and sufficient conditions and an explicit synthesis procedure for realizing the resultant polynomial operators.

A. Torokhti [151, 152] considered synthesis of set-valued interpolation systems on the basis of a special application of some interpolation polynomial operators considered in this chapter.

V. Khlobystov [74] generalized the result by W. Porter [106, 107] of finding an interpolation polynomial in $\mathcal{L}_2(a, b)$ with minimal norm to the case of an abstract Hilbert space with a measure. V. Khlobystov also obtained a solution to the extremal problem which is a generalization of the theorem of M. Golomb and H. Weinberger [49] for a bounded set of operator interpolants.

In the book by V. L. Makarov and V. V. Khlobystov [91], the theory of polynomial operators in Hilbert space due to Hermite and Hermite-Birkhoff is studied. The necessary and sufficient conditions for the solvability of different operator interpolation problems are given. The error analysis and study of convergence of associated interpolation techniques are provided.

3.7 Concluding Remarks

We have developed a non-Lagrangean procedure to construct a *weak* interpolation of a non-linear mapping on $\mathcal{C}[0,1]$ defined by a finite number of observed input-output pairs. We have shown that the *weak* interpolation can become a *strong* interpolation in the case of a finite dimensional range space and that when the parameters are chosen appropriately and the output evaluation set is sufficiently fine it can also provide an approximation to the original mapping in terms of the uniform norm on $\mathcal{C}([0,1])$. In this context we claim that a non-linear system described by an empirical data set can be synthesized in the form $S = WKQG_m$. We have also provided an elementary example to illustrate the *weak* interpolation procedure.

This page intentionally left blank

Chapter 4

Realistic Operators and their Approximation

4.1 Introduction

In the real world each object can be defined by a legend of historical information. The legend represents the complete history of the object and specifies the state of the object at all times. The current legend specifies the current state. The systematic evolution of state for a collection of real world objects is called a dynamical system. The system is specified by specifying each pair of initial and final states. Any such collection of input–output pairs defines a realistic operator. There are many real world objects that we may wish to model and some may be *non-deterministic.*

Example 8. *Experiments at Harvard University reported by Prof. Susan Greenfield of Oxford University* (ABC TV: *Brain Story; Compass* 19/2/01) *have supported the contention that when processing visual images the human brain records only key parts of the external image and uses memory of known and apparently similar images to reconstruct appropriate background information. One might regard the input-output relationship in this instance as only partially deterministic.*

For many centuries the modelling of real-world objects has been a primary interest for both natural science and philosophy. The notion of *cause* was considered by Aristotle who presented a basic metaphysical view of the material object, the essential object, the object as a body of work and the object as a realization of purpose. The English philosopher David Hume rejected this notion and argued that *causality* is a condition of constant conjunction, proximity in space and time and succession [69]. Immanuel Kant proposed that every event has a deterministic cause while John Stuart Mill related causality to the *natural laws* of physics [69] and argued that it could be analysed by experimental methods. Bertrand Russell also considered the notion of cause [114].

A qualitative theory of causality in Suppes [146] extends the notion to probabilistic systems. On the other hand the more traditional idea of *determinism* proposed by Laplace [69] suggests that knowledge of the positions of all physical bodies and the forces acting upon them at any instant would be sufficient to predict all future and past positions. Even with the Laplacian view we argue that imperfect knowledge of the state will allow or even *require* a probabilistic interpretation. The role of probabilistic scenarios is central.

Brinksma *et al* [11], Eerola [35] and Petrović [103] used stochastic models to extend the approach by Suppes [146].

The implementation of a different approach to the representation of a causal object has been developed, in particular, by De Santis [116] and Porter [106].

Porter [106] applied Prenter's theorems [104] to causal systems. Bertuzzi, Gandolfi and Germani [8] further extended Prenter's theorem [104] to the causal approximation of input-output maps in Hilbert space.

A significant development in applications of approximation theory to modelling nonlinear systems has been made by Sandberg [117]–[131]. In particular, it has been shown in [117], [118], [119] that a causal nonlinear input–output map can be approximated arbitrarily well in a meaningful sense by a finite Volterra series, even though it may not have a Volterra series expansion. Park and Sandberg [98] proved that radial–basis–function networks are capable of universal approximation. Sandberg [120, 121] showed that causal time–invariant maps satisfying certain continuity and approximately finite–memory conditions can be uniformly approximated arbitrarily well by finite sums formed from some simple linear operators.

A complete characterization of the input–output maps of causal time–invariant systems that can be uniformly approximated by the maps of certain simple structures is given in [122]. Reference [99] concerns conditions for the approximation of discrete time–invariant nonlinear systems that act between bounded real sequences.

The main theorem in [124] gives, in a certain setting, a necessary and sufficient condition under which multidimensional shift–invariant input–output maps with vector–valued inputs drawn from a certain large set can be uniformly approximated arbitrarily well using a structure consisting of a linear preprocessing stage followed by a memoryless nonlinear network. Further extensions of these results for the approximation of input–output maps of some special nonlinear systems are given in [125], [126].

Sandberg [117, 118, 119] generalized earlier theories of *causality* and *memory* for approximation of nonlinear systems. Bode and Shannon in the work [9] initiated an implementation of the causality principle into optimal linear filters. The causal models, developed in [116] - [121], approximate the input-output map with any given accuracy.

Daugavet [23] introduced the concept of a general mathematical formalism to describe a class of *realistic* properties such as causality, memory, and stationarity. The extension of this methodology to encompass the notion of a complete history or legend associated with each real world object and the development and interpretation of the idea, described later in the proposal, is due to Howlett, Torokhti and Pearce [58, 59]. The evolution of state in dynamical systems whereby one thing affects another is expressed through the agency of *operators*. Our representation of realistic operators is inextricably linked to optimal approximation.

4.2 Formalization of Concepts Related to Description of Real-World Objects

4.2.1 Causal operators, operators with finite memory, and stationary operators with finite memory

Here, we give Definitions and examples of causal operators, operators with finite memory, and stationary operators. These specific operators are motivated by the following observations.

Suppose an operator A is a mathematical model of a nonlinear system.

By the heuristic definition of causality, the present value of the output of a system is not affected by future values of the input [116]. To determine the output signal at time t_0, *the causal system* should "remember" the input signal up to time t_0.

A system with *finite memory* Δ is "able" to determine the output signal at time t_0 from a fragment of the input signal on the segment $[t_0 - \Delta,\ \ t_0]$ only. In other words, the system with finite memory Δ should "remember" the input signal on the segment of the length Δ.

A stationary system with finite memory is invariant with respect to time. In other words, if any fragments of the input signal are the same over different segments of the same length then the corresponding outputs are the same as well.

The formalizations of the above concepts are given in Definitions 1–3 below.

Let X and Y be functional Banach spaces such that

$$X = \{x(t) \mid t \in [a\ \ b]\} \quad \text{and} \quad Y = \{y(t) \mid t \in [c\ \ d]\}$$

where $x, y : \mathbb{R} \to \mathbb{R}$, $[a\ \ b] \subset \mathbb{R}$ and $[c\ \ d] \subset \mathbb{R}$.

Definition 16. *Let $[a\ \ b] = [c\ \ d]$, $t_0 \in [a\ \ b]$ and $K \subset X$. An operator $A : K \to Y$ is called causal if for any $x_1, x_2 \in K$,*

$$x_1(t) = x_2(t) \quad \forall \quad t \in [a\ \ t_0]$$

implies

$$[A(x_1)](t) = [A(x_2)](t) \quad \forall \quad t \in [a\ \ t_0].$$

Example 9. *Let $y = A(x)$ so that*

$$y(t_0) = \int_a^{t_0} x(t)dt.$$

Then A is the causal operator.

The operator A given by

$$y(t_0) = \int_a^b B(t_0, t)x(t)dt,$$

where $B : [a\ \ b] \times [a\ \ b] \to [a\ \ b]$, is not causal.

Let $C([a\ \ b])$ be the space of continuous functions on segment $[a\ \ b]$. For the sake of clarity, we set $X = C([a\ \ b])$ and $Y = C([a+\Delta\ \ b])$ where $\Delta \geq 0$.

Definition 17. *Operator $A : X \to Y$ is said to be the operator with finite memory Δ if for any $x_1, x_2 \in K \subset X$,*

$$x_1(t_0 - \Delta + s) = x_2(t_0 - \Delta + s) \quad \forall \quad s \in [0\ \ \Delta]$$

implies

$$[A(x_1)](t_0) = [A(x_2)](t_0) \quad \forall \quad t_0 \in [a + \Delta,\ \ b].$$

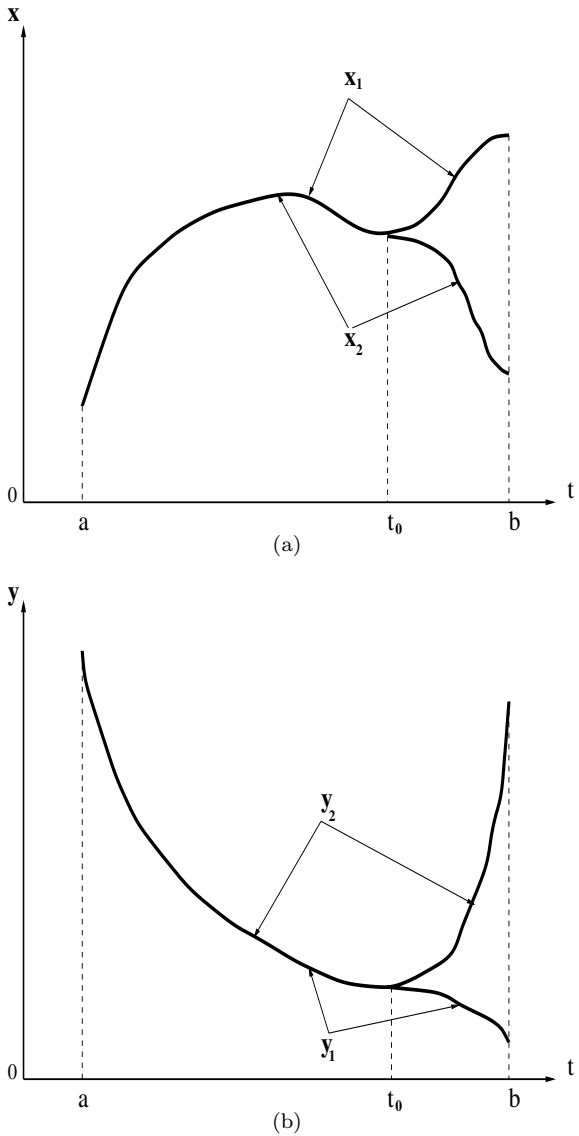


Figure 4.1: Illustration to the definition of the causal operator. Here, $y_1 = A(x_1)$ and $y_2 = A(x_2)$.

Example 10. *Let $y = A(x)$ so that*

$$y(t_0) = \int_0^{\Delta} x(t_0 - \Delta + s)ds.$$

Then A is the operator with finite memory Δ.

Definition 18. *Operator $A : X \to Y$ is called the stationary operator with finite memory Δ if for any $x_1, x_2 \in K \subset X$,*

$$x_1(t_1 - \Delta + s) = x_2(t_2 - \Delta + s) \quad \forall \quad s \in [0 \ \ \Delta] \quad \text{and} \quad t_1, t_2 \in [a + \Delta \ \ b]$$

implies

$$[A(x_1)](t_1) = [A(x_2)](t_2).$$

4.2.2 $\mathcal{R}-$continuous operators

Preliminaries

To begin we make the following definition.

Definition 19. *Let X and Y be separable Banach spaces. Let $K \subseteq X$ be a compact set and let $F : K \to Y$ be a continuous map. The modulus of continuity $\omega = \omega[F] : \mathbb{R}_+ \to \mathbb{R}_+$ is given by the formula*

$$\omega(\delta) = \sup_{x_1, x_2 \in K, \ \|x_1 - x_2\| \leq \delta} \|F(x_1) - F(x_2)\|.$$

It is easy to see that $\omega(0) = 0$ and that $\omega(\delta) \leq \omega(\delta')$ whenever $\delta \leq \delta'$. We will show that ω is also a uniformly continuous function.

Lemma 9. *Let X and Y be separable Banach spaces. Let $K \subseteq X$ be a compact set and $F : K \to Y$ a continuous map. Let $\omega = \omega[F] : \mathbb{R}_+ \to \mathbb{R}_+$ be the corresponding modulus of continuity. Then for each $\tau > 0$ we can find $\sigma = \sigma(\tau) > 0$ such that*

$$0 \leq \omega(\delta') - \omega(\delta) \leq \tau$$

whenever $0 \leq \delta' - \delta \leq \sigma$.

Proof. Define $\Delta F : K \times K \to Y$ by setting

$$\Delta F(x_1, x_2) = F(x_2) - F(x_1)$$

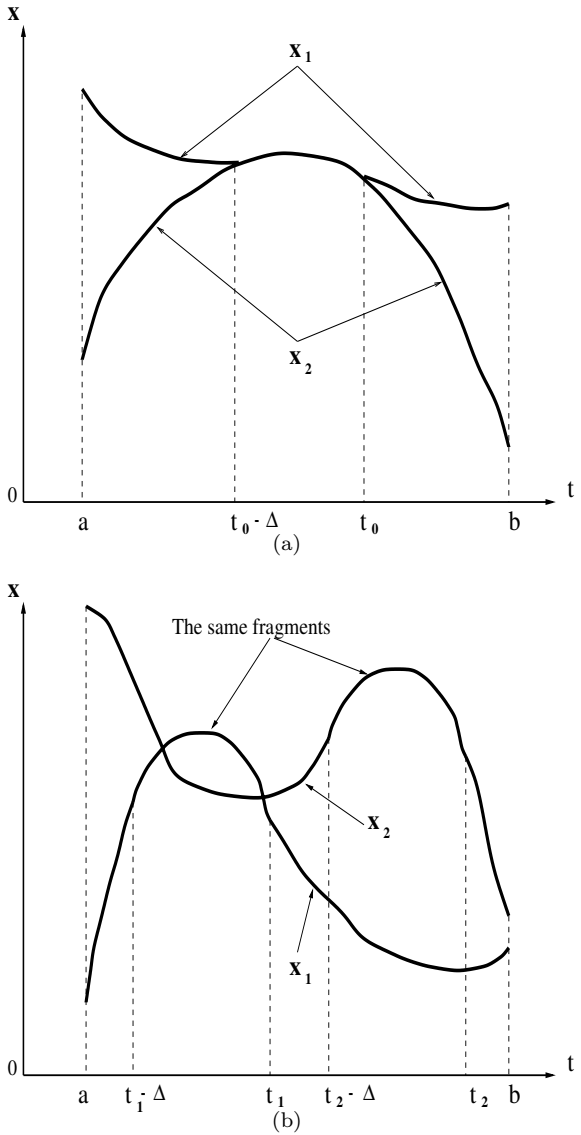


Figure 4.2: (a) Illustration to the definition of the operator with finite memory. (b) Illustration to the definition of the stationary operator.

for each $x_1, x_2 \in K$. Clearly ΔF is continuous with respect to the norm

$$\|(x_1, x_2)\|_{K\times K} = \|x_1\| + \|x_2\|$$

and hence, since $K \times K$ is compact, ΔF is uniformly continuous. If we define

$$D_\delta = \{(x_1, x_2) \mid \|x_2 - x_1\| \leq \delta\}$$

then $D_\delta \subseteq K \times K$ is compact and

$$\omega(\delta) = \sup_{(x_1,x_2)\in D_\delta} \|\Delta F(x_1, x_2)\|$$

for each $\delta \geq 0$. Fix $\tau > 0$ and choose $\sigma = \sigma(\tau) > 0$ such that

$$\|\Delta F(x_1', x_2') - \Delta F(x_1, x_2)\| < \tau$$

whenever

$$\|(x_1', x_2') - (x_1, x_2)\|_{K\times K} < \sigma.$$

Now suppose that

$$0 \leq \delta' - \delta \leq \sigma.$$

Find $(x_1', x_2') \in D_{\delta'}$ with

$$\omega(\delta') = \|\Delta F(x_1', x_2')\|$$

and define $\theta \in [0, 1]$ so that

$$\theta\|x_2' - x_1'\| = \delta.$$

Let

$$(x_1, x_2) = \theta(x_1', x_2') + (1 - \theta)(\frac{x_1' + x_2'}{2}, \frac{x_1' + x_2'}{2}).$$

It is easy to see that

$$\|x_2 - x_1\| = \delta$$

and that

$$\|(x_1', x_2') - (x_1, x_2)\|_{K\times K} \leq \sigma.$$

It follows that

$$\begin{aligned} \omega(\delta') &= \|\Delta F(x_1', x_2')\| \leq \|\Delta F(x_1, x_2)\| + \tau \\ &\leq \omega(\delta) + \tau. \end{aligned}$$

Thus

$$0 \leq \omega(\delta') - \omega(\delta) \leq \tau$$

whenever $0 \leq \delta' - \delta \leq \sigma$. Hence ω is uniformly continuous on $\mathbb{R}_+$. □

Theorem 21. *Let X and Y be separable Banach spaces. Let $K \subseteq X$ be a compact set and $F : K \to Y$ a continuous map. For any given numbers $\delta > 0$ and $\tau > 0$ and for all $x \in K$ and all $x' \in X$ with*

$$\|x' - x\| \leq \delta$$

there exists an operator

$$S = WZQG_m : X \to Y$$

defined by finite arithmetic such that

$$||F(x) - S(x')|| \leq \frac{1}{2}\omega[F](2\delta) + \tau.$$

Remark 9. *This theorem is important for synthesis of non-linear systems because the error $||F(x) - S(x')||$ in the output for a given level δ of the noise $x' - x$ is not dependent on the arbitrarily chosen positive real number τ.*

Proof. This proof follows the methods of Daugavet [23].

It is well known that any separable Banach space is isometric and isomorphic to a subspace of the space $C([0,1])$ of continuous functions on the interval $[0,1]$. Thus without loss of generality we assume $X = Y = C([0,1])$.

Define

$$\varphi : K \times [0,1] \to \mathbb{R}$$

by setting

$$\varphi(x,t) = F[x](t)$$

for all $t \in [0,1]$. Fix $\delta > 0$ and $t \in [0,1]$. For each

$$u \in K_\delta = \{u \mid \|u - x\| \leq \delta \text{ for some } x \in K\}$$

choose

$$x^+[u] = x^+_{\delta,t}[u], \quad x^-[u] = x^-_{\delta,t}[u] \in K$$

so that

$$\varphi^+_\delta(u,t) = \varphi(x^+[u],t) = \max_{x \in K, \|x-u\| \leq \delta} \varphi(x,t)$$

and

$$\varphi^-_\delta(u,t) = \varphi(x^-[u],t) = \min_{x \in K, \|x-u\| \leq \delta} \varphi(x,t)$$

and set

$$\varphi_\delta(u,t) = \frac{1}{2}[\varphi^+_\delta(u,t) + \varphi^-_\delta(u,t)].$$

Define

$$F_\delta : K_\delta \to C([0,1])$$

by setting

$$F_\delta[u](t) = \varphi_\delta(u,t)$$

for all $\delta > 0$ and each $t \in [0,1]$. If $u \in K_\delta$ and $x \in K$ with

$$\|u - x\| \leq \delta$$

then

$$|\varphi(x,t) - \varphi_\delta(u,t)| \leq \omega(2\delta)/2$$

for all $t \in [0,1]$ and hence it follows that

$$\|F(x) - F_\delta(u)\| \leq \frac{1}{2}\omega(2\delta).$$

However F_δ may not be continuous. Therefore for fixed $t \in [0,1]$ and each pair of positive real numbers λ and μ we define

$$\varphi_{\lambda,\mu}(u,t) = \frac{1}{2\mu}\int_{[\lambda,\lambda+\mu]}[\varphi_\xi^+(u,t) + \varphi_\xi^-(u,t)]d\xi$$

and

$$F_{\lambda,\mu} : K_\lambda \to C([0,1])$$

by setting

$$F_{\lambda,\mu}[u](t) = \varphi_{\lambda,\mu}(u,t)$$

for all $t \in [0,1]$. If

$$\|u - v\| < \rho$$

then it can be shown that

$$\|F_{\lambda,\mu}[u] - F_{\lambda,\mu}[v]\| \leq \frac{2F_K\rho}{\mu}$$

where

$$F_K = \max_{x\in K}\|F(x)\|.$$

This shows that the operator $F_{\lambda,\mu}$ is continuous. If $x \in K$ and

$$\|x - u\| < \lambda$$

then it follows that

$$\|F(x) - F_{\lambda,\mu}(u)\| \leq \frac{1}{2}\omega(2\nu)$$

where $\nu = \lambda + \mu$. To prove the desired result we take $\tau > 0$ and choose $\epsilon > 0$ so that

$$\omega(2\delta + \epsilon) \leq \omega(2\delta) + \tau$$

for all $\delta > 0$. Now we set $\lambda = \delta + \epsilon/2$ and $\mu = \epsilon/2$ and note that if

$$\|x - u\| \leq \lambda$$

then

$$\|F(x) - F_{\lambda,\mu}(u)\| \leq \frac{1}{2}\omega(2\delta) + \frac{\tau}{2}.$$

Let $0 = t_0 < \cdots < t_N = 1$ be a partition of the interval $[0, 1]$ and define the operator

$$P_N \in \mathcal{L}(C([0, 1]), PL([0, 1])),$$

where

$$PL([0, 1]) \subseteq C([0, 1])$$

is the subspace of piecewise linear functions, by setting

$$P_N[x](t_k) = x(t_k)$$

for each $k = 0, \ldots, N$ with the partition sufficiently fine to ensure that

$$\|x - P_N(x)\| \leq \epsilon/4$$

for all $x \in K$. Let L_δ denote the closure of the set $P_N(K_\delta)$. Since L_δ lies in an $N + 1$ dimensional subspace and is bounded and closed it follows that L_δ is compact. It can be shown that $L_\delta \subseteq K_\lambda$ and hence $F_{\lambda,\mu}$ is well defined on L_δ. By Theorem 9 in Section 5.5 of Chapter 1, for all $v \in L_\delta$ there exists an operator $S_{\lambda,\mu} : X \to C(T)$ in the form

$$S_{\lambda,\mu} = WZQG_m^\star$$

such that

$$\|F_{\lambda,\mu}(v) - S_{\lambda,\mu}(v)\| \leq \frac{\tau}{2}.$$

We can now define the operator $S : X \to C(T)$ in the form

$$S = WZQG_m,$$

where $G_m = G_m^\star P_N$, by the equality

$$S(u) = S_{\lambda,\mu}(P_N[u])$$

for each $u \in K_\delta$. □

4.2.3 Main definitions and auxiliary results

We use the results discussed above to establish a systematic procedure for the constructive approximation of *realistic* operators.

The notions of *causality*, *finite memory* and *stationarity* have been used for many years in the engineering literature and are properties that one may associate with *realistic* dynamical systems. The object of this discussion is to consider the possibility of a generic description of a *realistic* system that allows us to establish general procedures with which we can effectively simulate such systems. To include the above *realistic* properties we construct special spaces with additional topological structure. The fundamental idea is that each element $x \in X$ must contain a corresponding continuum of historical information. In fact we will assume that each element is uniquely defined by this corresponding history. The definition of an $\mathcal{R}$-space follows Daugavet [23].

Definition 20. *Let X and A be Banach spaces and let $\mathcal{L}(X, A)$ be the set of continuous linear operators from X into A. Let $T = (T, \rho)$ be a compact metric space and let $\mathcal{M} = \{M_t\}_{t \in T}$ be a family of operators $M_t \in \mathcal{L}(X, A)$ with norm*

$$||M_t|| \leq 1 \quad \text{for each } t \in T$$

and such that

$$M_s[u] \to M_t[u] \quad \text{as} \quad \rho(s,t) \to 0 \quad \text{for each} \quad u \in X.$$

The space X equipped with the family of operators $\mathcal{M}$ is called an $\mathcal{R}$-space and is denoted by

$$X_{\mathcal{R}} = (X, A, T, \mathcal{M}).$$

For each $x \in X$ the collection of elements

$$\mathcal{M}[x] = \{M_t[x] \mid t \in T\} \subseteq A$$

specifies the complete history of the element x. We assume that if $\mathcal{M}[x]$ is specified then $x \in X$ is uniquely defined. In other words each element $x \in X$ is defined by specifying the complete history of the element[1]. We will write

$$x = \mathcal{H}(\mathcal{M}[x])$$

where

$$\mathcal{H} : \mathcal{M}[X] \to X$$

[1]This idea is an adaption of the idea that a function is defined by specifying the complete set of function values.

is the appropriate archival function. We wish to define a special class of *realistic* operators. An $\mathcal{R}$-continuous operator is an operator from one $\mathcal{R}$-space to another such that the history of the range space is continuous with respect to the history of the domain space.

Definition 21. *Let*

$$X_{\mathcal{R}} = (X, A, T, \mathcal{M}) \quad \textit{and} \quad Y_{\mathcal{R}} = (Y, B, T, \mathcal{N})$$

be $\mathcal{R}$-spaces and let the closed set

$$E \subseteq T \times T$$

be an equivalence relation. Let $K \subseteq X$ be a compact set and let $v \in K$ and $t \in T$. The operator $F : K \to Y$ is $\mathcal{R}$-continuous at $M_s[x] \in A$ if, for each open neighbourhood of zero $H \subseteq B$, we can find a corresponding open neighbourhood of zero

$$G = G(x, s, H) \subseteq A$$

such that

$$M_r[u] - M_s[x] \in G$$

implies

$$N_r[F(u)] - N_s[F(x)] \in H$$

whenever $(r, s) \in E$ and $u \in K$.

If $F : K \to Y$ is $\mathcal{R}$-continuous at $M_s[x] \in A$ for all $x \in K$ and $s \in T$ then we say that $F : K \to Y$ is an $\mathcal{R}$-continuous operator.

For each $t \in T$ we observe that the set

$$M_t[K] = \{M_t[x] \mid x \in K\} \subseteq A$$

is compact. Indeed, if

$$M_t[K] \subseteq \bigcup_{\gamma \in \Gamma} G_\gamma$$

where each G_γ is open then

$$K \subseteq \bigcup_{\gamma \in \Gamma} U_\gamma$$

where each $U_\gamma = M_t^{-1}[G_\gamma]$ is also open. Since K is compact we can find a finite subcollection $U_{\gamma_1}, \ldots, U_{\gamma_r}$ such that

$$K \subseteq \bigcup_{i=1}^{r} U_{\gamma_i}$$

and hence

$$M_t[K] \subseteq \bigcup_{i=1}^{r} G_{\gamma_i}.$$

Hence $M_t[K]$ is covered by a finite subcollection. It follows that $M_t[K]$ is compact. Let

$$E_t = \{s \mid (s,t) \in E\} \subseteq T$$

and note that E_t is compact. We wish to show that the set

$$\mathcal{M}_t[K] = \{M_s[K] \mid s \in E_t\}$$

is also compact. We need the following result.

Lemma 10. *Let $s \in T$. If $M_s[K] \subseteq G$ where G is an open set then we can find $\delta = \delta(s,G) > 0$ such that*

$$M_r[K] \subseteq G$$

when

$$\rho(r,s) < \delta.$$

Proof. Suppose the result is not true. Then we can find a sequence $\{r_i\} \subseteq T$ with $r_i \to s$ as $i \to \infty$ and a sequence $\{x_i\} \subseteq K$ such that $M_{r_i}[x_i] \notin G$ for each i. Because K is compact we can assume without loss of generality that $x_i \to x$ for some $x \in K$ as $i \to \infty$. Choose a neighbourhood of zero

$$G_\alpha = \{a \mid \|a\| < \alpha\} \subseteq A$$

so that

$$M_s[x] + G_\alpha \subseteq G.$$

Since $\|M_r\| \leq 1$ for all $r \in T$ it follows that

$$M_r[u] \in G_\alpha/2 \quad \text{whenever} \quad u \in U_\alpha/2$$

where

$$U_\alpha = \{u \mid \|u\| < \alpha\} \subseteq X.$$

If we choose i so large that

$$x_i - x \in U_\alpha/2$$

and

$$M_{r_i} \in M_s[x] + G_\alpha/2$$

then

$$\begin{aligned} M_{r_i}[x_i] &= M_s[x] + (M_{r_i}[x] - M_s[x]) + M_{r_i}[x_i - x] \\ &\in M_s[x] + G_\alpha \subseteq G \end{aligned}$$

which is a contradiction. □

Lemma 11. *For each $t \in T$ the set $\mathcal{M}_t[K]$ is a compact subset of A.*

Proof. Suppose that

$$\mathcal{M}_t[K] \subseteq \bigcup_{\gamma\in\Gamma} G_\gamma$$

where each G_γ is an open set. For each $s \in E_t$ we know that $M_s[K]$ is compact and that

$$M_s[K] \subseteq \mathcal{M}_t[K].$$

Therefore we can find a finite subset $\Gamma(s) \subseteq \Gamma$ and a corresponding finite sub-collection $\{G_\gamma\}_{\gamma\in\Gamma(s)}$ such that

$$\begin{aligned} M_s[K] &\subseteq \bigcup_{\gamma\in\Gamma(s)} G_\gamma \\ &= G(s). \end{aligned}$$

Choose $\delta(s) > 0$ such that $M_r[K] \subseteq G(s)$ whenever $\rho(r,s) < \delta(s)$ and define the open sets

$$\begin{aligned} R(s) &= \{r \mid \rho(r,s) < \delta(s)\} \\ &\subseteq T \end{aligned}$$

for each $s \in T$. Since

$$E_t \subseteq \bigcup_{s\in E_t} R(s)$$

and since E_t is compact we can find a finite sub-collection $\{R(s_j)\}_{j=1,2,\ldots,q}$ such that

$$E_t \subseteq \bigcup_{j=1}^{q} R(s_j).$$

Therefore

$$\bigcup_{r\in R(s_j)} M_r[K] \subseteq G(s_j)$$

and hence

$$\begin{aligned} \mathcal{M}_t[K] &= \bigcup_{r\in E_t} M_r[K] \\ &= \bigcup_{j=1}^{q} \left[\bigcup_{r\in R(s_j)} M_r[K] \right] \subseteq \bigcup_{j=1}^{q} G(s_j) \\ &= \bigcup_{j=1}^{q} \left[\bigcup_{\gamma\in\Gamma(s_j)} G_\gamma \right]. \end{aligned} \tag{4.1}$$

Since this is a finite sub-covering it follows that $\mathcal{M}_t[K]$ is compact. □ □

Definition 22. *Let $F : K \to Y$. If for each neighbourhood of zero, $H \subseteq B$ there exists a neighbourhood of zero $G = G(H)$ such that*

$$M_r[u] - M_s[v] \in G$$

implies

$$N_r[Fu] - N_s[Fv] \in H$$

whenever $(r, s) \in E$ and $u, v \in K$, then F is called the uniformly $\mathcal{R}$-continuous operator.

A link between continuous, $\mathcal{R}$-continuous and uniformly $\mathcal{R}$-continuous operators is shown in the Lemma below.

Lemma 12. *Let $F : K \to Y$ be continuous and $\mathcal{R}$-continuous. Then F is uniformly $\mathcal{R}$-continuous.*

Proof. Suppose the result is not true. Then for some $\beta > 0$ we can find neighbourhoods of zero

$$\begin{aligned} H_\beta &= \{b \mid \|b\| < \beta\} \\ &\subseteq B \end{aligned}$$

and

$$\begin{aligned} G_{1/n} &= \{a \mid \|a\| < 1/n\} \\ &\subseteq A \end{aligned}$$

for each $n = 1, 2, \ldots$ and points

$$u_n, v_n \in K \quad \text{and} \quad r(n), s(n), t(n) \in T$$

with $r(n), s(n) \in E_{t(n)}$ for each $n = 1, 2, \ldots$ such that

$$M_{r(n)}[u_n] - M_{s(n)}[v_n] \in G_{1/n}$$

and

$$N_{r(n)}[Fu_n] - N_{s(n)}[Fv_n] \notin H_\beta.$$

Since K is compact we can suppose without loss of generality that there exist $u, v \in K$ with

$$u_n \to u \quad \text{and} \quad v_n \to v \quad \text{as} \quad n \to \infty.$$

Since T is compact we can suppose that there exist points $r, s, t \in T$ with

$$r(n) \to r, \quad s(n) \to s \quad \text{and} \quad t(n) \to t$$

as $n \to \infty$. Note that since $(r(n), t(n)) \in E$ and $(s(n), t(n)) \in E$ and since E is closed it follows that

$$(r, t) \in E \quad \text{and} \quad (s, t) \in E.$$

Hence $r, s \in E_t$.

Now choose $\alpha > 0$ and define the neighborhood of zero

$$\begin{aligned} G_\alpha &= \{a \mid \|a\| < \alpha\} \\ &\subseteq A. \end{aligned}$$

Since $\|M_r\| \leq 1$ for all $r \in T$ we have

$$M_r[x] \in G_\alpha/5$$

whenever $x \in U_\alpha/5$ where

$$U_\alpha = \{x \mid \|x\| < \alpha\} \subseteq X.$$

If we take n so large that

$$u - u_n \in U_\alpha/5,$$

$$v - v_n \in U_\alpha/5,$$

$$M_{r(n)}[u] - M_r[u] \in G_\alpha/5,$$

$$M_{s(n)}[v] - M_s[v] \in G_\alpha/5$$

and

$$G_{1/n} \subseteq G_\alpha/5$$

then

$$\begin{aligned} M_r[u] - M_s[v] &= \big[M_r[u] - M_{r(n)}[u]\big] + \big[M_{r(n)}[u] - M_{r(n)}[u_n]\big] \\ &\quad + \big[M_{r(n)}[u_n] - M_{s(n)}[v_n]\big] + \big[M_{s(n)}[v_n] - M_{s(n)}[v]\big] \\ &\quad + \big[M_{s(n)}[v] - M_s[v]\big] \\ &\in G_\alpha. \end{aligned}$$

Since α is arbitrary it follows that

$$M_r[u] - M_s[v] = 0$$

and since $r, s \in E_t$ the $\mathcal{R}$-continuity of F implies that

$$N_r[Fu] - N_s[Fv] = 0.$$

Define the neighborhood

$$\begin{aligned} V_\beta &= \{y \mid \|y\| < \beta\} \\ &\subseteq Y. \end{aligned}$$

Since

$$\|N_r\| \leq 1 \quad \text{for all} \quad r \in T$$

we have

$$N_r[y] \in H_\beta/4$$

whenever $y \in V_\beta/4$.

Choose n so large that

$$Fu_n - Fu \in V_\beta/4,$$

$$Fv_n - Fv \in V_\beta/4$$

and

$$N_{r(n)}[Fu] - N_r[Fu], N_{s(n)}[Fv] - N_s[Fv] \in H_\beta/4.$$

Now it follows that

$$\begin{aligned} N_{r(n)}[Fu_n] - N_{s(n)}[Fv_n] &= \big[N_{r(n)}[Fu_n] - N_{r(n)}[Fu]\big] \\ &\quad + \big[N_{r(n)}[Fu] - N_r[Fu]\big] \\ &\quad + [N_r[Fu] - N_s[Fv]] \\ &\quad + \big[N_s[Fv] - N_{s(n)}[Fv]\big] \\ &\quad + \big[N_{s(n)}[Fv] - N_{s(n)}[Fv_n]\big] \\ &\in H_\beta \end{aligned}$$

which is a contradiction. □

4.3 Approximation of $\mathcal{R}-$continuous Operators

4.3.1 The collection of auxiliary mappings

In order to establish a constructive $\mathcal{R}$-continuous approximation to the $\mathcal{R}$-continuous mapping $F : K \to Y$ it is convenient to define a collection of auxilliary mappings. For each $t \in T$ we define $f_t : \mathcal{M}_t[K] \to B$ by setting

$$f_t(M_s[x]) = N_s[F(x)]$$

for each $s \in E_t$ and $x \in K$. This is a good definition because

$$M_r[u] = M_s[x]$$

implies

$$N_r[F(u)] = N_s[F(x)]$$

for each $r, s \in E_t$ and each $u \in K$. The mapping $f_t : \mathcal{M}_t[K] \to B$ is continuous at each point $M_s[x] \in \mathcal{M}_t[K]$ because, for each open neighbourhood of zero $H \subseteq B$, we can find a corresponding open neighbourhood of zero $G = G_t(x, s, H) \subseteq A$ such that

$$M_r[u] - M_s[x] \in G$$

implies

$$\begin{aligned} f_t(M_r[u]) - f_t(M_s[x]) &= N_r[F(u)] - N_s[F(x)] \\ &\in H \end{aligned}$$

whenever $r, s \in E_t$ and $u \in K$.

Because $\mathcal{M}_t[K]$ is compact it follows that the mapping $f_t : \mathcal{M}_t[K] \to B$ is uniformly continuous. In other words, for each neighbourhood of zero $H \subseteq B$, we can find a neighbourhood of zero $G = G_t(H) \subseteq A$ such that

$$M_r[u] - M_s[v] \in G$$

implies

$$f_t(M_r[u]) - f_t(M_s[v]) \in H$$

whenever $r, s \in E_t$ and $u, v \in K$. In view of Lemma 12 we know that when $F : K \to Y$ is continuous the collection of mappings $\{f_t\}_{t \in T}$ is uniformly equi-continuous. That is for each open neighbourhood of zero $H \subseteq B$ we can find a neighbourhood of zero $G = G(H) \subseteq A$ such that for all $t \in T$ we have

$$M_r[u] - M_s[v] \in G$$

implies

$$f_t(M_r[u]) - f_t(M_s[v]) \in H$$

whenever $r, s \in E_t$ and $u, v \in K$.

Remark 10. *The continuous operator $F : K \to Y$ is an $\mathcal{R}$-operator in the sense of Daugavet* [23] *if*

$$M_s[u] - M_t[v] = 0$$

implies

$$N_s[F(u)] - N_t[F(v)] = 0$$

for all $u, v \in K$ and all $(s, t) \in E$.

If F is continuous and $\mathcal{R}$-continuous operator then F is an $\mathcal{R}$-operator in the sense of Daugavet.

4.3.2 The special $\mathcal{R}$-modulus of continuity

The $\mathcal{R}$-modulus of continuity will be used to characterize our constructive approximation theorems for realistic operators.

Definition 23. *Let $X_{\mathcal{R}} = \{X, A, T, \mathcal{M}\}$ and $Y_{\mathcal{R}} = \{Y, B, T, \mathcal{N}\}$ be $\mathcal{R}$-spaces and let $E \subseteq T \times T$ be the given equivalence relation. Let $K \subseteq X$ be a compact set and suppose that the map $F : K \to Y$ is an $\mathcal{R}$-continuous operator. The function*

$$\omega_{\mathcal{R}} = \omega_{\mathcal{R}}[F] : \mathbb{R}_+ \to \mathbb{R}_+$$

defined by

$$\omega_{\mathcal{R}}(\delta) = \sup_{\substack{u,v \in K;\ (r,s) \in E: \\ \|M_r[u] - M_s[v]\| \leq \delta}} \|N_r[F(u)] - N_s[F(v)]\|$$

is called the $\mathcal{R}$-modulus of continuity of the operator F.

Definition 24. *We say that $(X_{\mathcal{R}}, Y_{\mathcal{R}})$ is a complete $\mathcal{R}$-pair if*

$$E = T \times T$$

and an incomplete $\mathcal{R}$-pair if

$$E \neq T \times T.$$

In the case where $E = \{(t,t)\}_{t \in T}$ we say that $(X_{\mathcal{R}}, Y_{\mathcal{R}})$ is a simple $\mathcal{R}$-pair.

We make the following elementary observations about the $\mathcal{R}$-modulus of continuity.

Lemma 13. *Let $(X_{\mathcal{R}}, Y_{\mathcal{R}})$ is a complete $\mathcal{R}$-pair and suppose that $F : K \to Y$ is an $\mathcal{R}$-continuous operator. Then the $\mathcal{R}$-modulus of continuity*

$$\omega_{\mathcal{R}} = \omega_{\mathcal{R}}[F] : \mathbb{R}_+ \to \mathbb{R}_+$$

is uniformly continuous with $\omega_{\mathcal{R}}(0) = 0$.

Proof. Since $E_t = T$ for all $t \in T$ it follows that

$$\mathcal{M}[K] = \mathcal{M}_t[K] = \{M_s[x] \mid x \in K \text{ and } s \in T\} \subseteq A$$

for all $t \in T$ and we can define an auxiliary mapping $f : \mathcal{M}[K] \to B$ by setting

$$f(M_t[x]) = N_t[Fx]$$

for each $x \in K$ and $t \in T$. We recall from our earlier remarks about auxiliary mappings that not only is this a good definition but also that the mapping $f : \mathcal{M}[K] \to B$ is uniformly continuous. The function $\omega_f : \mathbb{R}_+ \to \mathbb{R}_+$ is the modulus of continuity of f. Lemma 9 shows us that ω_f is also uniformly continuous. Since

$$\begin{aligned}
\omega_f(\delta) &= \sup_{\substack{p,q \in \mathcal{M}[K]: \\ \|p-q\| \leq \delta}} \|f(p) - f(q)\| \\
&= \sup_{\substack{u,v \in K;\ (r,s) \in E: \\ \|M_r[u] - M_s[v]\| \leq \delta}} \|N_r[F(u)] - N_s[F(v)]\| \\
&= \omega_{\mathcal{R}}(\delta)
\end{aligned}$$

we obtain the desired result. □ □

Lemma 14. *Let $(X_{\mathcal{R}}, Y_{\mathcal{R}})$ be an incomplete $\mathcal{R}$-pair and suppose that $F : K \to Y$ is both a continuous operator and an $\mathcal{R}$-continuous operator. Then the $\mathcal{R}$-modulus of continuity $\omega_{\mathcal{R}} = \omega_{\mathcal{R}}[F] : \mathbb{R}_+ \to \mathbb{R}_+$ is uniformly continuous with $\omega_{\mathcal{R}}(0) = 0$.*

Proof. Since $(X_{\mathcal{R}}, Y_{\mathcal{R}})$ is an incomplete $\mathcal{R}$-pair we consider the various equivalence classes E_t for each $t \in T$. We have seen earlier that for each $t \in T$ there is a auxiliary mapping $f_t : M_t[K] \to B$ defined by setting

$$f_t(M_t[x]) = N_t[F(x)]$$

for all $x \in K$. Let $\omega[f_t] : \mathbb{R}_+ \to \mathbb{R}_+$ be the modulus of continuity for the map f_t and consider the argument used in Lemma 9. Define

$$\Delta f_t : M_t[K] \times M_t[K]$$

by the formula

$$\Delta f_t(p, q) = \|f_t(p) - f_t(q)\|$$

for each

$$(p, q) \in M_t[K] \times M_t[K].$$

Choose $\tau > 0$. From our earlier remarks about the uniform equi-continuity of the family of auxiliary mappings $\{f_t\}_{t \in T}$ we can choose $\sigma = \sigma(\tau) > 0$ such that for all $t \in T$ we have

$$\|\Delta f_t(p', q') - \Delta f_t(p, q)\| < \tau$$

whenever

$$\|(p', q') - (p, q)\| < \sigma.$$

Now it is clear from Lemma 9 that for all $t \in T$ we have

$$0 \leq \omega[f_t](\delta') - \omega[f_t](\delta) \leq \tau$$

whenever

$$0 \leq \delta' - \delta \leq \sigma.$$

Thus the family $\{\omega[f_t]\}_{t \in T}$ is also uniformly equi-continuous. Since

$$\omega_{\mathcal{R}}(\delta) = \sup_{t \in T} \omega[f_t](\delta)$$

it follows that

$$0 \leq \omega_{\mathcal{R}}(\delta') - \omega_{\mathcal{R}}(\delta) \leq \tau$$

whenever $0 \leq \delta' - \delta \leq \sigma$. □

4.3.3 Approximately $\mathcal{R}$-continuous operators

In practice we may have to consider an approximately $\mathcal{R}$-continuous operator which will preserve an approximately continuous sense of history.

Definition 25. *Let $X_{\mathcal{R}} = (X, A, T_X, \mathcal{M})$ and $Y_{\mathcal{R}} = (Y, B, T_Y, \mathcal{N})$ be $\mathcal{R}$-spaces with $T_X = T_Y = T$. Let the closed set*

$$E \subseteq T \times T$$

be an equivalence relation and suppose that $G \subseteq A$ and $H \subseteq B$ are open neighbourhoods of 0. Let $K \subseteq X$ be a compact set and let $v \in K$ and $t \in T$. The operator $F : K \to Y$ is approximately $\mathcal{R}$-continuous with tolerance (G, H) at $M_t[v] \in A$ if

$$M_s[u] - M_t[v] \in G$$

implies

$$N_s[F(u)] - N_t[F(v)] \in H$$

whenever $u \in K$ and $(s, t) \in E$.

If $F : K \to Y$ is approximately $\mathcal{R}$-continuous with tolerance (G, H) at $M_t[v] \in A$ for all $v \in K$ and $t \in T$ then we say that $F : K \to Y$ is an approximately $\mathcal{R}$-continuous operator with tolerance (G, H).

Some elementary examples

Before we begin our discussion of the approximation procedure we will consider some familiar examples of *realistic* operators from the viewpoint of our new definitions. To prepare for the examples we note the following theorem of M. Riesz regarding compactness criteria for a subset $K \subseteq L^p([0,1])$. For each $x \in K$ we use the notation

$$T_h x(r) = x(r+h)$$

for each

$$r, r+h \in [0,1].$$

Theorem 22. *The set $K \subseteq L^p([0,1])$ is compact if and only if we can find $M > 0$ with $\|x\|_p \leq M$ and $\delta = \delta(\epsilon)$ such that*

$$\|T_h x - x\|_p < \epsilon$$

whenever

$$|h| < \delta \quad \text{for all} \quad x \in K.$$

In the case $K \subseteq L^p(\mathbb{R})$ the above conditions and the additional condition, that for each $x \in K$ we have $x(t) = 0$ for $t \notin C_K$ for some fixed compact set $C_K \subseteq \mathbb{R}$, are sufficient for K to be compact.

Example 11. *Let $X = L^1([0,1])$ and $Y = C([0,1])$ and let*

$$K = \{x \mid |x(s) - x(t)| \leq |s-t| \ \forall \ s,t \in [0,1]\}.$$

Define $F : K \to Y$ by setting

$$Fx(t) = e^{-t} \int_{[0,t]} e^s x(s) ds$$

for each $t \in [0,1]$.

Set

$$T_X = T_Y = [0,1] \quad \text{and} \quad A = B = C([0,1])$$

and define $\mathcal{M} = \{M_t\}_{t\in[0,1]}$ and $\mathcal{N} = \{N_t\}_{t\in[0,1]}$ by setting

$$M_t[x](s) = \begin{cases} \displaystyle\int_{[0,s]} x(r)dr & \text{if } s \leq t, \\ \displaystyle\int_{[0,t]} x(r)dr & \text{otherwise,} \end{cases}$$

and

$$N_t[y](s) = \begin{cases} y(s) & \text{if } s \leq t, \\ y(t) & \text{otherwise.} \end{cases}$$

Now $X_{\mathcal{R}}$ and $Y_{\mathcal{R}}$ are $\mathcal{R}$-spaces. Put $E = \{(t,t) \mid t \in T\}$. Therefore $E_t = \{t\}$ for all $t \in T$. We will say that the operator $F : K \to Y$ is a uniformly $\mathcal{R}$-continuous causal operator if for each

$$u, v \in K \quad \text{and} \quad t \in T \quad \text{and for each} \quad \beta > 0$$

we can find

$$\alpha = \alpha(\beta) > 0$$

such that

$$\|M_t[u] - M_t[v]\| < \alpha$$

implies

$$|N_t[F(u)] - N_t[F(v)]\| < \beta.$$

Note that we can use integration by parts to show that

$$\begin{aligned} Fu(\tau) - Fv(\tau) &= e^{-\tau} \int_{[0,\tau]} e^s [u(s) - v(s)] ds \\ &= \int_{[0,\tau]} [u(r) - v(r)] dr \\ &\quad -e^{-\tau} \int_{[0,\tau]} e^s \left[\int_{[0,s]} [u(r) - v(r)] dr \right] ds \end{aligned}$$

for each $\tau \in [0,t]$ and since

$$\|M_t[u] - M_t[v]\| < \alpha$$

implies

$$|\int_{[0,s]} [u(r) - v(r)] dr| < \alpha$$

for all $s \in [0,t]$ it follows that

$$|Fu(\tau) - Fv(\tau)| < (\tau + 1)\alpha.$$

If we set $\alpha = \beta/2$ then

$$\|M_t[u] - M_t[v]\| < \alpha$$

implies

$$\|N_t[Fu] - N_t[Fv]\| < \beta$$

for all $t \in [0,1]$. Therefore F is indeed a uniformly $\mathcal{R}$-continuous causal operator.

Example 12. *Consider the $\mathcal{R}$-continuous causal operator F of Example* 11 *and let $\alpha, \beta \in \mathbb{R}$ be given positive numbers. Define open neighbourhoods of zero $G_\alpha \subseteq A$ and $H_\beta \subseteq B$ by setting*

$$G_\alpha = \{a \mid \|a\| < \alpha\} \quad \textit{and} \quad H_\beta = \{b \mid \|b\| < \beta\}.$$

Choose $\beta > 0$. In Example 11 *we showed that*

$$M_t[u] - M_t[v] \in G_{\beta/2}$$

implies

$$N_t[F(u)] - N_t[F(v)] \in H_\beta$$

for each $u, v \in K$ and each $t \in T$. Hence, for all $\beta > 0$, we can say that F is an approximately $\mathcal{R}$-continuous operator with tolerance $(G_{\beta/2}, H_\beta)$.

Remark 11. *It is clear from the previous example that for any uniformly $\mathcal{R}$-continuous operator F and any $\beta > 0$ we can choose $\alpha > 0$ such that F is an approximately $\mathcal{R}$-continuous operator with tolerance (G_α, H_β).*

The next examples show that concepts such as *finite memory* and *stationarity* can also be formulated in terms of the proposed general framework.

Example 13. *Let*

$$X = L^\infty(\mathbb{R}), \quad A = L^\infty([0, \Delta]), \quad Y = C(\mathbb{R}), \quad B = C([0, 1 + \Delta])$$

and

$$T_X = T_Y = [0, 1 + \Delta]$$

where $\Delta > 0$ is a fixed positive real number. Define

$$K = \{x \mid x(t) = 0 \textit{ for } t \notin [0, 1] \textit{ and } |x(s) - x(t)| \leq |s - t| \ \forall \ s, t \in \mathbb{R}\}$$

and consider the mapping $F = F_\Delta : K \to Y$ given by the formula

$$[Fu](t) = \frac{1}{\Delta} \int_{[t-\Delta, t]} u(r) dr$$

for each $u \in X$ and $t \in \mathbb{R}$. Define $M_t : X \to A$ by

$$M_t[x](r) = x(r + t - \Delta)$$

for each $r \in [0, \Delta]$ and each $t \in T$ and $N_t : Y \to C([0, 1 + \Delta])$ by

$$N_t[y](\tau) = y(t)$$

for all $\tau \in [0, 1+\Delta]$ *and each* $t \in T$.

Let $E = T \times T$. *Since*

$$\begin{aligned} N_t[Fx](\tau) &= [Fx](t) \\ &= \frac{1}{\Delta}\int_{[t-\Delta,t]} x(r)dr \\ &= \frac{1}{\Delta}\int_{[0,\Delta]} M_t[x](r)dr \end{aligned}$$

it is clear that

$$N_s[Fu](\tau) - N_t[Fv](\tau) = \frac{1}{\Delta}\int_{[0,\Delta]} [M_s[u](r) - M_t[v](r)]\, dr.$$

If

$$\|M_s[u] - M_t[v]\| < \beta$$

then

$$\begin{aligned} |N_s[Fu](\tau) - N_t[Fv](\tau)| &\leq \frac{1}{\Delta}\int_{[0,\Delta]} |M_s[u](r) - M_t[v](r)|dr \\ &\leq \frac{1}{\Delta}\int_{[0,\Delta]} \|M_s[u] - M_t[v]\|dr \\ &< \beta \end{aligned}$$

and hence

$$\|M_s[u] - M_t[v]\| < \beta$$

implies

$$\|N_s[Fu] - N_t[Fv]\| < \beta$$

for all $(s,t) \in E$ *and all* $\beta > 0$. *Thus, for all* $\beta > 0$, *we can say that* F *is an approximately* $\mathcal{R}$*-continuous stationary operator of finite memory* Δ *with tolerance* (G_β, H_β).

Of course we have also shown that F *is a uniformly* $\mathcal{R}$*-continuous stationary operator with finite memory* Δ.

Note that the equivalence relation $E = T \times T$ *allows us to consider time pairs in the form* (s,t) *where* $s \neq t$. *This is an essential ingredient in our description of a stationary operator.*

4.3.4 A model for constructive approximation in the class of $\mathcal{R}$-continuous operators

When F is an $\mathcal{R}$-continuous operator we prove the existence of an approximating $\mathcal{R}$-continuous operator S which is *stable* to small disturbances. The

operator S defines a model of the real system and is constructed from an algebra of elementary continuous functions by a process of finite arithmetic.

Definition 26. *We will say that the family $\mathcal{N}$ of operators $\{N_t\}_{t\in T}$ is pointwise normally extreme on Y if, for each $y \in Y$, we can find $t = t_y \in T$ such that*

$$\|N_t[y]\| = \|y\|.$$

Theorem 23. *Let A and B be Banach spaces with the Grothendieck property of approximation and let*

$$X_{\mathcal{R}} = (X, A, T, \mathcal{M}) \quad \textit{and} \quad Y_{\mathcal{R}} = (Y, B, T, \mathcal{N})$$

be $\mathcal{R}$-spaces. Suppose that $(X_{\mathcal{R}}, Y_{\mathcal{R}})$ is a complete $\mathcal{R}$-pair and that $\mathcal{N}$ is pointwise normally extreme on Y. Let $K \subseteq X$ be a compact set and let the map $F : K \to Y$ be an $\mathcal{R}$-continuous operator. Then for any fixed real numbers

$$\delta > 0 \quad \textit{and} \quad \tau > 0$$

there exists an associated $\mathcal{R}$-continuous operator S defined by finite arithmetic in the form

$$S = WZQG : X \to Y$$

such that for all $x \in K$ and $x' \in X$ with $\|x - x'\| \leq \delta$ we have

$$\|F(x) - S(x')\| \leq \frac{1}{2}\omega_{\mathcal{R}}(2\delta) + \tau.$$

Proof. Since $E_t = T$ for all $t \in T$ we can define the auxiliary mapping $f : \mathcal{M}[K] \to B$ by setting

$$f(M_t[x]) = N_t[Fx]$$

for each $x \in K$ and $t \in T$. We recall that not only is this a good definition but also from Lemma 13 that the mapping $f : \mathcal{M}[K] \to B$ is uniformly continuous. Let $A_m \subseteq A$ be a subspace of dimension m and $B_n \subseteq B$ be a subspace of dimension n. We will construct a mapping $\sigma : A \to B$ in the form

$$\sigma = \pi\nu\lambda\theta$$

where $\theta \in \mathcal{L}(A, A_m)$ and $\lambda \in \mathcal{L}(A_m, \mathbb{R}^m)$ are given by

$$\theta(w) = \sum_{i=1}^{m} \alpha_i c_i$$

and

$$\lambda(\sum_{i=1}^{m} \alpha_i c_i) = (\alpha_1, \ldots, \alpha_m)$$

for some suitable basis $c_1, \ldots, c_m$ in A_m, where $\nu : \mathbb{R}^m \to \mathbb{R}^n$ is continuous and where $\pi \in \mathcal{L}(\mathbb{R}^n, B_n)$ is given by

$$\pi(\nu) = \sum_{j=1}^{n} \nu_j d_j$$

for some suitable basis $d_1, \ldots, d_n$ in B_n. By applying Theorem 21 there exists a continuous mapping $\sigma : A \to B$ in the above form such that for all

$$w \in \mathcal{M}[K] \quad \text{and all} \quad w'$$

with

$$\|w - w'\| < \delta$$

we have

$$\|f(w) - \sigma(w')\| \leq \frac{1}{2}\omega_f(2\delta) + \tau$$

where ω_f is the modulus of continuity of f. Since $\omega_f(\alpha) = \omega_{\mathcal{R}}(\alpha)$ for all $\alpha \in \mathbb{R}_+$ we see that

$$\|f(w) - \sigma(w')\| \leq \frac{1}{2}\omega_{\mathcal{R}}(2\delta) + \tau$$

when

$$\|w - w'\| < \delta.$$

Now define $S : X \to Y$ by setting

$$N_t[Sx] = \sigma(M_t[x])$$

for each $x \in X$ and each $t \in T$.

Our indirect definition assumes that if $N_t[y] \in B$ is known for all $t \in T$ then $y \in Y$ is also known. We will follow our earlier notation and write

$$y = \mathcal{K}(\mathcal{N}[y])$$

where $\mathcal{K} : \mathcal{N}[Y] \to Y$ is the appropriate archival function. The mapping $\sigma : A \to B$ is continuous and hence, for each $\beta > 0$, we can find $\alpha > 0$ so that

$$\|M_s[u] - M_t[v]\| < \alpha$$

implies

$$\|\sigma(M_s[u]) - \sigma(M_t[v])\| < \beta$$

and

$$\|N_s[Su] - N_t[Sv]\| < \beta$$

and hence $S : X \to Y$ is an $\mathcal{R}$-continuous operator.

Since

$$\|M_t[x - x']\| \leq \|x - x'\|$$

it follows that

$$\begin{aligned} \|N_t[Fx - Sx']\| &= \|f(M_t[x]) - \sigma(M_t[x'])\| \\ &< \frac{1}{2}\omega_{\mathcal{R}}(2\delta) + \tau \end{aligned}$$

for all $t \in T$ whenever $x \in K$ and

$$\|x - x'\| < \delta.$$

Because we can choose $t \in T$ such that

$$\|N_t[Fx - Sx']\| = \|F(x) - S(x')\|$$

we must have

$$\|F(x) - S(x')\| < \frac{1}{2}\omega_{\mathcal{R}}(2\delta) + \tau$$

whenever $x \in K$ and

$$\|x - x'\| < \delta.$$

The action of the operator S on an element $x \in X$ is defined by the collection of ordered pairs

$$\{(M_t[x], \pi\nu\lambda\theta M_t[x]) \mid t \in T\} = \{(\mathcal{M}[x], \pi\nu\lambda\theta\mathcal{M}[x])\}.$$

Since we defined

$$N_t[Sx] = \pi\nu\lambda\theta M_t[x]$$

we can now write

$$\mathcal{N}[Sx] = \pi\nu\lambda\theta\mathcal{M}[x]$$

or equivalently

$$S(x) = \mathcal{K}\pi\nu\lambda\theta\mathcal{H}^{-1}x$$

for each $x \in X$. If we define

$$G = \theta\mathcal{H}^{-1}, \quad Q = \lambda, \quad Z = \nu \quad \text{and} \quad W = \mathcal{K}\pi$$

then we can see that S has the desired form. We assume that G and W can be defined by finite arithmetic or replaced by suitable approximations. $\square$

To establish our next result on *stable* approximation in the class of $\mathcal{R}$-continuous operators we need the following elementary lemma.

Lemma 15. *Let $K \subseteq X$ be a compact set. Then for each $\epsilon > 0$ we can find $\gamma > 0$ such that*

$$\|M_s[x] - M_t[x]\| < \epsilon$$

for all $x \in K$ whenever $s, t \in T$ and $\rho(s,t) < \gamma$.

Proof. For each $x \in K$ let

$$U(x; \epsilon/3) = \{u \mid u \in X \text{ and } \|u - x\| < \epsilon/3\}.$$

Clearly

$$K \subseteq \bigcup_{x \in K} U(x; \epsilon/3)$$

and because K is compact there is a finite subcovering

$$U(x_1; \epsilon/3), \ldots, U(x_p; \epsilon/3).$$

Since $M_t[x_j]$ is uniformly continuous in $t \in T$ for each $j = 1, 2, \ldots, p$ we can find $\gamma > 0$ so that

$$\rho(s,t) < \gamma$$

implies

$$\|M_s[x_j] - M_t[x_j]\| < \epsilon/3$$

whenever $s, t \in T$. For each $x \in K$ we can find x_j such that $x \in U(x_j; \epsilon/3)$ and hence

$$\begin{aligned} \|M_s[x] - M_t[x]\| &< \|M_s[x - x_j]\| + \|M_s[x_j] - M_t[x_j]\| \\ &\qquad + \|M_t[x_j - x]\| \\ &\leq 2\|u - x\| + \|M_s[x_j] - M_t[x_j]\| \\ &< \epsilon \end{aligned}$$

when $s, t \in T$ and $\rho(s,t) < \gamma$. □

Theorem 24. *Let A and B be Banach spaces with the Grothendieck property of approximation. Let*

$$X_{\mathcal{R}} = (X, A, T, \mathcal{M}) \quad \textit{and} \quad Y_{\mathcal{R}} = (Y, B, T, \mathcal{N})$$

be $\mathcal{R}$-spaces and suppose that $(X_{\mathcal{R}}, Y_{\mathcal{R}})$ is an incomplete $\mathcal{R}$-pair and that $\mathcal{N}$ is pointwise normally extreme on Y. Let $K \subseteq X$ be a compact set and

let the map $F : K \to Y$ be continuous and $\mathcal{R}$-continuous. Then for any fixed real numbers

$$\delta > 0 \quad \textit{and} \quad \tau > 0$$

there exists an associated operator $S : X \to Y$ defined by

$$N_t[Su] = \sum_{j=1}^{N} \psi_j(t) N_t[S_j u]$$

where $\psi_j : T \to \mathbb{R}$ for each $j = 1, 2, \ldots, N$ and $\{\psi_1, \ldots, \psi_N\}$ is a partition of unity and where

$$S_j = W_j Z_j Q_j G_j : X \to Y$$

for each $j = 1, 2, \ldots, N$ and each $u \in K$ and $t \in T$. The mapping S is continuous and $\mathcal{R}$-continuous and is defined by a process of finite arithmetic in such a way that for all $x \in K$ and $x' \in X$ with

$$\|x - x'\| \leq \delta$$

we have

$$\|F(x) - S(x')\| \leq \frac{1}{2}\omega_{\mathcal{R}}(2\delta) + \tau.$$

Proof. Let $t \in T$. The auxilliary mapping $f_t : M_t[K] \to B$ is defined by setting

$$f_t(M_s[x]) = N_s[F(x)]$$

when $s \in E_t$ and $x \in K$. Let

$$\omega[f_t] : \mathbb{R}_+ \to \mathbb{R}_+$$

be the associated modulus of continuity. We recall from Lemmas 12 and 14 that the families $\{f_t\}_{t\in T}$ and $\{\omega[f_t]\}_{t\in T}$ are each uniformly equi-continuous. Hence, for the given $\tau > 0$, it is possible to choose $\epsilon = \epsilon(\tau) > 0$ so that

$$\lambda \leq \delta + \epsilon$$

implies

$$\omega[f_t](2\lambda) \leq \omega[f_t](2\delta) + \tau$$

for all $t \in T$. It now follows that

$$\lambda \leq \delta + \epsilon$$

implies

$$\omega_{\mathcal{R}}(2\lambda) \leq \omega_{\mathcal{R}}(2\delta) + \tau.$$

Our choice of ϵ can also be sufficiently to also ensure that

$$\|M_r[u] - M_s[v]\| < \epsilon$$

implies

$$\|N_r[Fu] - N_s[Fv]\| < \tau/12$$

whenever $(r, s) \in E$ and $u, v \in K$. By Lemma 15 we can find $\gamma > 0$ to ensure that

$$\|M_s[x] - M_t[x]\| < \epsilon$$

for all $x \in K$ when $\rho(s, t) < \gamma$. Using a similar argument and by decreasing γ if necessary we can also ensure that

$$\|N_s[Fx] - N_t[Fx]\| < \tau/4$$

for all $x \in K$ when $\rho(s, t) < \gamma$. Choose a γ-net

$$\{t_1, \ldots, t_N\} \subseteq T$$

such that whenever $t \in T$ we can always find some $j = j(t)$ with

$$\|t - t_j\| < \gamma$$

and let

$$\{\psi_1(t), \ldots, \psi_N(t)\}$$

where $\psi_j : T \to \mathbb{R}$ for each $j = 1, 2, \ldots, N$ be a partition of unity on T such that

- $\psi_1, \ldots, \psi_N \in C(T)$,
- $\psi_j(t) \geq 0$ for all $t \in T$,
- $\sum_{j=1}^{N} \psi_j(t) = 1$ for all $t \in T$, and
- $\psi_j(t) = 0$ if $\rho(t, t_j) \geq \gamma$.

Let $x \in K$ and choose $u \in X$ with

$$\|u - x\| \leq \delta.$$

If $\rho(t, t_j) < \gamma$ then

$$\begin{aligned} \|M_t[u] - M_{t_j}[x]\| &\leq \|M_t[u] - M_t[x]\| + \|M_t[x] - M_{t_j}[x]\| \\ &\leq \|u - x\| + \epsilon \\ &\leq \delta + \epsilon. \end{aligned} \tag{4.2}$$

By applying Theorem 21 we can define a function

$$\sigma_j : A \to B$$

in the form

$$\sigma_j = \pi_j \nu_j \lambda_j \theta_j$$

such that for all $w \in M_{t_j}[K]$ and w' with

$$\|w' - w\| < \lambda$$

we have

$$\begin{aligned} \|f_j(w) - \sigma_j(w')\| &< \frac{1}{2}\omega[f_j](2\lambda) + \frac{\tau}{4} \\ &< \frac{1}{2}\omega_{\mathcal{R}}(2\lambda) + \frac{\tau}{4}. \end{aligned} \tag{4.3}$$

Define $S_j : X \to Y$ by setting

$$N_t[S_j u] = \sigma_j(M_t[u])$$

and $S : X \to Y$ by the formula

$$N_t[Su] = \sum_{j=1}^{N} \psi_j(t)\sigma_j(M_t[u])$$

for all $u \in X$ and $t \in T$.

Now for $x \in K, \quad u \in X$ with

$$\|u - x\| < \delta$$

and all $t \in T$, we have

$$\|N_t[Fx] - N_t[Su]\| = \| \sum_{\rho(t,t_j)<\gamma} \psi_j(t) \left[N_t[Fx] - \sigma_j(M_t[u])\right] \|.$$

We make two observations. Firstly, for $\rho(t, t_j) < \gamma$ we have

$$\begin{aligned} \|N_t[Fx] - \sigma_j(M_t[u])\| &\leq \|N_t[Fx] - N_{t_j}[Fx]\| \\ &\quad + \|N_{t_j}[Fx] - \sigma_j(M_t[u])\| \\ &\leq \|f_j(M_{t_j}[x]) - \sigma_j(M_t u)\| + \frac{\tau}{4} \end{aligned}$$

and secondly, since

$$\|M_{t_j}[x] - M_t[u]\| \leq \|u - x\| + \epsilon$$

it follows that

$$\begin{aligned}\|f_j(M_{t_j}[x]) - \sigma_j(M_t[u])\| &\leq \frac{1}{2}\omega_{\mathcal{R}}(2[\|u-x\|+\epsilon]) + \frac{\tau}{4} \\ &\leq \frac{1}{2}\omega_{\mathcal{R}}(2\delta) + \frac{\tau}{4}.\end{aligned}$$

Hence for all $t \in T$ we have

$$\begin{aligned}\|N_t[Fx] - N_t[Su]\| &\leq [\frac{1}{2}\omega_{\mathcal{R}}(2\delta) + \frac{\tau}{2}] \sum_{\rho(t,t_j)<\gamma} \psi_j(t) \\ &\leq \frac{1}{2}\omega_{\mathcal{R}}(2\delta) + \tau\end{aligned}$$

from which the desired result follows. □

Theorem 25. *Let A and B be Banach spaces with the Grothendieck property of approximation and let*

$$X_{\mathcal{R}} = (X, A, T, \mathcal{M}) \quad \textit{and} \quad Y_{\mathcal{R}} = (Y, B, T, \mathcal{N})$$

be $\mathcal{R}$-spaces and suppose that $(X_{\mathcal{R}}, Y_{\mathcal{R}})$ is a complete $\mathcal{R}$-pair and that $\mathcal{N}$ is pointwise normally extreme on Y. Let $K \subseteq X$ be a compact set and let the continuous map $F : K \to Y$ be a continuous $\mathcal{R}$-operator. Then for any fixed real numbers

$$\delta > 0 \quad \textit{and} \quad \tau > 0$$

there exists an associated approximately $\mathcal{R}$-continuous operator $\hat{S}$ defined by finite arithmetic in the form

$$\hat{S} = WZQG : X \to Y$$

such that for all $x \in K$ and $x' \in X$ with

$$\|x - x'\| \leq \delta$$

we have

$$\|F(x) - \hat{S}(x')\| \leq \frac{1}{2}\omega_{\mathcal{R}}(2\delta) + \tau.$$

Proof. Define the mapping $f : \mathcal{M}[K] \to B$ by setting

$$f(M_t[x]) = N_t[Fx]$$

for each $x \in K$ and $t \in T$. By Theorem 21 we can construct a mapping in the form

$$\sigma = \pi\nu\lambda\theta : A \to B$$

such that for all $w \in \mathcal{M}[K]$ and all w' with

$$\|w - w'\| < \delta$$

we have

$$|f(w) - \sigma(w')| \leq \frac{1}{2}\omega_{\mathcal{R}}(2\delta) + \tau/2$$

where $\omega_{\mathcal{R}}(\cdot)$ is the $\mathcal{R}$-modulus of continuity of F.

At this stage we wish to reformulate our construction and begin by considering an appropriate approximation to the map $\nu : \mathbb{R}^m \to \mathbb{R}^n$.

Let $u \in K$. For each $t \in T$ we have

$$\theta(M_t[u]) = \sum_{i=1}^{m} \alpha_i[u](t)c_i$$

and since

$$M_s[u] \to M_t[u] \quad \text{as} \quad s \to t$$

it follows that

$$\alpha[u] = (\alpha_1[u], \ldots, \alpha_m[u]) \in C(T)^m.$$

Because

$$\alpha[v] \to \alpha[u] \quad \text{as} \quad v \to u$$

and K is compact the set

$$A_K = \{\alpha[u] \mid u \in K\} \subseteq C(T)^m$$

is also compact and we can use the Arzela-Ascoli Theorem to deduce that the functions $\alpha[u] \in A_K$ are uniformly equi-continuous. It follows that we can find a bounded closed interval $I_K \subseteq \mathbb{R}^m$ such that

$$\alpha[u](t) = (\alpha_1[u], \ldots, \alpha_m[u])(t) \in I_K$$

for each $u \in K$ and each $t \in T$. The mapping $\nu : \mathbb{R}^m \to \mathbb{R}^n$ is uniformly continuous on I_K and hence we can find $\gamma > 0$ such that

$$\|\nu(\alpha) - \nu(\alpha')\| < \tau/2n$$

whenever

$$\|\alpha - \alpha'\| < \gamma.$$

As in Lemma 15 choose a finite number of points $\alpha[u_1], \ldots, \alpha[u_q]$ such that

$$A_K \subseteq \bigcup_{l=1}^{q} N(\alpha[u_l]; \gamma)$$

and a collection $\{\psi_1, \ldots, \psi_q\}$ of continuous functions

$$\psi_l : C(T)^m \to \mathbb{R}$$

with

$$\sum_{l=1}^{q} \psi_l(\alpha) = 1$$

for all $\alpha \in A_K$ and such that

$$\psi_l(\alpha) \geq 0, \quad \psi_l(\alpha) > 0$$

when

$$\|\alpha - \alpha[u_l]\| < \gamma/12$$

and

$$\psi_l(\alpha) = 0$$

when

$$\|\alpha - \alpha[u_l]\| \geq \gamma$$

for each $l = 1, 2, \ldots, q$.

It follows that

$$\|\alpha - \sum_{l=1}^{q} \psi_l(\alpha)\alpha[u_l]\| < \gamma$$

for each $\alpha \in A_K$.

To complete the approximation we choose η so that

$$\|\alpha[u](s) - \alpha[u](t)\| < \gamma$$

for all $u \in K$ whenever $\rho(s,t) < \eta$. If $\{t_1, \ldots, t_p\}$ is an η-net for T then we can define continuous functions $\varphi_k : T \to \mathbb{R}$ with

$$\sum_{k=1}^{p} \varphi_k(t) = 1$$

for all $t \in T$ and such that

$$\varphi_k(t) \geq 0, \quad \varphi_k(t) > 0$$

when

$$\rho(t, t_k) < \eta/2$$

and

$$\varphi_k(t) = 0$$

when

$$\rho(t, t_k) \geq \eta$$

for each $k = 1, 2, \ldots, p$.

It can be seen that

$$\begin{aligned}
&\|\alpha[u](t) - \sum_{k=1}^{p} \varphi_k(t) \sum_{l=1}^{q} \psi_l(\alpha[u])\alpha[u_l](t_k)\| \\
&\leq \quad \|\alpha[u](t) - \sum_{k=1}^{p} \varphi_k(t)\alpha[u](t_k)\| \\
&\qquad + \|\sum_{k=1}^{p} \varphi_k(t) \left[\alpha[u](t_k) - \sum_{l=1}^{q} \psi_l(\alpha[u])\alpha[u_l](t_k)\right] \| \\
&\leq \quad \gamma/3
\end{aligned}$$

for each $u \in K$ and all $t \in T$ and hence

$$\|\nu(\alpha[u](t)) - \nu(\sum_{k=1}^{p} \varphi_k(t) \sum_{l=1}^{q} \psi_l(\alpha[u])\alpha[u_l](t_k))\| < \tau/2n.$$

If $\|d_j\| \leq 1$ for all $j = 1, 2, \ldots, n$ then

$$\begin{aligned}
&\|\pi\nu(\alpha[u](t)) - \pi\nu(\sum_{k=1}^{p} \varphi_k(t) \sum_{l=1}^{q} \psi_l(\alpha[u])\alpha[u_l](t_k))\| \\
&= \quad \|\sum_{j=1}^{n} \left[\nu_j(\alpha[u](t)) - \nu_j(\sum_{k=1}^{p} \varphi_k(t) \sum_{l=1}^{q} \psi_l(\alpha[u])\alpha[u_l](t_k))\right] d_j\| \\
&\leq \quad \sum_{j=1}^{n} |\nu_j(\alpha[u](t)) - \nu_j(\sum_{k=1}^{p} \varphi_k(t) \sum_{l=1}^{q} \psi_l(\alpha[u])\alpha[u_l](t_k))| \\
&< \quad \tau/2
\end{aligned}$$

which we can rewrite in the form

$$\|\sigma(M_t[u]) - \pi\nu(\sum_{k=1}^{p} \varphi_k(t) \sum_{l=1}^{q} \psi_l(\alpha[u])\alpha[u_l](t_k))\| < \tau/2.$$

If we define $\hat{S} : X \to Y$ by setting

$$N_t[\hat{S}u] = \pi\nu(\sum_{k=1}^{p} \varphi_k(t) \sum_{l=1}^{q} \psi_l(\alpha[u])\alpha[u_l](t_k))$$

then

$$\|N_t[Su] - N_t[\hat{S}u]\| < \tau/2$$

for all $u \in K$ and all $t \in T$ and hence

$$\|S(u) - \hat{S}(u)\| < \tau/2$$

for all $u \in K$. On the other hand if $\|c_i\| \leq 1$ for all $i = 1, 2, \ldots, m$ then

$$\|M_t[u] - \sum_{i=1}^{m} \left[\sum_{k=1}^{p} \varphi_k(t) \sum_{l=1}^{q} \psi_l(\alpha[u]) \alpha_i[u_l](t_k) \right] c_i\| < \gamma/3$$

for all $u \in K$ and $t \in T$. Therefore

$$\|M_s[u] - M_t[v]\| < \gamma/3$$

implies

$$\|\sum_{i=1}^{m} \left[\sum_{k=1}^{p} \left\{ \varphi_k(s) \sum_{l=1}^{q} \psi_l(\alpha[u]) - \varphi_k(t) \sum_{l=1}^{q} \psi_l(\alpha[v]) \right\} \alpha_i[u_l](t_k) \right] c_i\| < \gamma$$

from which it follows that

$$\|N_s[\hat{S}u] - N_t[\hat{S}v]\| < \tau/2$$

for each $s, t \in T$. In other words

$$M_s[u] - M_t[v] \in U_{\gamma/3}$$

implies

$$N_s[\hat{S}u] - N_t[\hat{S}v] \in V_{\tau/2}$$

and we can say that $\hat{S}$ is an approximately continuous realistic operator with tolerance $(U_{\gamma/3}, V_{\tau/2})$.

We note that if

$$x \in K \quad \text{and} \quad \|x - x'\| < \delta$$

then

$$\begin{aligned} \|N_t[Fx] - N_t[\hat{S}x']\| &\leq \|N_t[Fx] - N_t[Sx']\| + \|N_t[Sx'] - N_t[\hat{S}x']\| \\ &< \frac{1}{2}\omega_{\mathcal{R}}(2\delta) + \tau \end{aligned}$$

for all $t \in T$ and hence

$$\|F(x) - \hat{S}(x')\| < \frac{1}{2}\omega_{\mathcal{R}}(2\delta) + \tau.$$

Finally we address the structure of the operator $\hat{S}$.
The operator $\hat{S} : X \to Y$ is based on a mapping

$$\hat{\sigma} = \pi\hat{\nu}\hat{\lambda}\hat{\theta} : X \to B,$$

where $\hat{\theta} \in \mathcal{L}(X, C(T)^m)$ is defined by

$$\hat{\theta}(M_t u) = \alpha[u](t)$$

for all

$$u \in X \quad \text{and} \quad t \in T,$$

and where $\hat{\lambda} : C(T)^m \to \mathbb{R}^{mpq}$ is given by

$$\hat{\lambda}(\alpha[u]) = \{\alpha_i[u_l](t_k)\}$$

for $i = 1, 2, \ldots, m$, $k = 1, 2, \ldots, p$ and $l = 1, 2, \ldots, q$, and where the mapping $\hat{\nu} : \mathbb{R}^{mpq} \to \mathbb{R}^n$ is specified by

$$\hat{\nu}(\{\alpha_i[u_l](t_k)\}) = \nu(\sum_{k=1}^{p} \varphi_k(t) \sum_{l=1}^{q} \psi_l(\alpha[u])\alpha[u_l(t_k)).$$

Of course with a bit more effort we could do the same sort of approximation with $N_t[y]$. □

4.4 Concluding Remarks

In this chapter, we have presented a unified approach to finding mathematical models of a realistic dynamical system that represent the system with any pre-assigned accuracy. The models are given by constructively defined operators with special properties. "A realistic dynamical system" means an object with real-world properties such as causality, memory, stationarity, etc. We presented a formalization of such properties in the form of the $\mathcal{R}$-continuous operator and the approximately $\mathcal{R}$-continuous operator. The proposed models of the realistic system are $\mathcal{R}$-continuous and approximately $\mathcal{R}$-continuous.

This page intentionally left blank

Chapter 5

Methods of Best Approximation for Nonlinear Operators

5.1 Introduction

The theory of operator approximation has a direct application to the mathematical modelling of nonlinear systems. In recent decades, methods of constructive representation of nonlinear systems have been a topic of intensive research [106]–[155]. In broad terms, the problem is to find a mathematical model of the system which is given by an abstract operator $\mathcal{F}$ representing the input-output relationship. The model must approximate the system in a certain sense subject to some restrictions. Such restrictions follow, in par-

ticular, from initial known information on the system. For example, in the case of the system transforming deterministic signals, this information can be given by equations describing the signals sets. A number of works including, in particular, the fundamental results by Volterra, Wiener, Porter, Sandberg have been devoted to the solution of this problem. The references can be found, in particular, in [106]–[119]. The works [106]–[155] provide models which approximate the system with *any pre-assigned accuracy.*

General theory of the best approximation in normed linear spaces has been developed for many years. A number of deep theoretical results related to the investigation of existence, uniqueness and characterization of elements of the best approximation have been established. See, for example, [29] and the bibliography there. However, theory of the best *constructive* approximation of nonlinear operators is not so well elaborated, and pioneering achievements in this area, such as those obtained in [76, 81, 82], are very recent. The papers [156]–[166] also provide new methods in this area of research.

The methods of Chapter 3 (and those in [12]-[65], [104]-[151]) for constructive approximation of nonlinear operators *with any preassigned accuracy* have mainly been concerned with proving the existence and uniqueness of approximating operators, and with justifying the bounds of errors arising from the approximation methods. The assumptions used are that the spaces of images and pre-images are deterministic and that elements of these spaces can be represented in an analytical form, i.e. by equations.

In many applications, the spaces of images and pre-images are probabilistic and their elements cannot be written analytically. Nevertheless, it is possible to describe the elements of these spaces in terms of their numerical characteristics, such as the estimates of mathematical expectation, of covariance matrices etc. A typical example is image processing where a digitized image, represented by a matrix, is often interpreted as the sample of a stochastic signal.

In this chapter, we provide some new approaches to the best constructive approximation of nonlinear operators in "deterministic" and probability spaces. In the case of approximation in probability spaces, it is assumed that the only available information on the operator is given by certain covariance matrices. The approaches considered in Sections 5.2 and 5.4–5.7 below are based on the approximant produced by a polynomial operator. The approximant minimizes the mean squared error between a desired image and the image of the approximating operator. Standard software packages, such as Matlab, can easily be used to implement the method (see Sections 5.4–5.7 in this regard).

In Section 5.7, we consider a method based on a combination of the special iterative procedure and the best operator approximation used at each iteration.

5.2 Best Approximation of Nonlinear Operators in Banach Spaces: "Deterministic" Case

Let us consider the operator $S = WZQG_m$ considered in preceding chapters.

In this section, we show how the operator $S = WZQG_m$ could be constructed to provide, in some definite sense, the best possible approximation to a given operator F.

Let X, Y be Banach spaces having the Grothendieck property of approximation and consider the following procedure.

Unlike the preceding Chapters, we now suppose that operator Z is given by multidimensional polynomials $g_1(c_k; a)$, ..., $g_n(c_k; a)$ in the form (5.2) below so that $Z = Z_c$ with

$$Z_c(a) = (g_1(c_1; a), g_2(c_2; a), \ldots, g_n(c_n; a)) \tag{5.1}$$

and

$$g_k(c_k; a) = \sum_{s=0}^{p} c_{k,s} r_s(a) \tag{5.2}$$

where $p = (p_1, p_2, \ldots, p_m) \in \mathcal{Z}_+^m$ is given and $\mathcal{Z}_+ = \{0, 1, 2, \ldots\}$ denotes the set of non-negative integers and where

$$c = (c_1, c_2, \ldots, c_n), \quad c_k = \{c_{k,s}\}_{s \in \mathcal{Z}_+^m} \quad \text{and} \quad c_{k,s} \in \mathbb{R}$$

for each $k = 1, 2, \ldots, n$ and each permissible $s \in \mathcal{Z}_+^m$. We assume that each $r_s : \mathbb{R}^m \to \mathbb{R}$ is continuous and that the collection $\{r_s\}_{s \in \mathcal{Z}_+^m}$ generates an algebra that satisfies the conditions of Stone's Algebra [113]. We could for example take

$$r_s(a) = a^s = a_1^{s_1} a_2^{s_2} \ldots a_m^{s_m}. \tag{5.3}$$

Let us assume that the functions $\{r_s\}_{s \in \mathcal{Z}_+^m}$ are linearly independent. Introduce the class $\mathcal{S}$ of operators given by

$$\mathcal{S} = \{S \mid S : X \to Y_n \text{ and } S = S_c = WZ_cQG_m\} \tag{5.4}$$

with fixed operators G_m, Q, H_n and W and with fixed functions $\{r_s\}_{s \in \mathcal{Z}_+^m}$. Thus the operator S_c is completely defined by the coefficients $\{c_{k,s}\}$.

Let $\{g_k(c_k^*; a)\}$ denote the functions which best approximate the given functions $\{f_k(a)\}$ on the set

$$QG_m(K) \subseteq \mathbb{R}^m.$$

We can now state the following theorem.

Theorem 26. *Let X, Y be Banach spaces with the Grothendieck property of approximation, let $K \subseteq X$ be a compact set and $F : K \to Y$ a continuous map. Let the operator $Z^* : \mathbb{R}^m \to \mathbb{R}^n$ be defined by*

$$Z^* = Z_{c^*}. \tag{5.5}$$

Then for some fixed $\epsilon > 0$ and for all $x, x' \in X$ with $\|x' - x\| < \epsilon$ the operator $S^* = S_{c^*} : X \to Y_n$ in the form

$$S^* = WZ^*QG_m \tag{5.6}$$

satisfies the equality

$$\sup_{x \in K} \|F(x) - S^*(x')\| = \inf_{S \in \mathcal{S}} \left\{ \sup_{x \in K} \|F(x) - S(x')\| \right\}. \tag{5.7}$$

Let us now extend the application of the approach presented in the preceding Chapters to the best approximation of non-linear dynamical systems when the system is completely described by a finite number of real parameters.

To this end, let us consider the approximation of operator $F : K_\Gamma \to Y$ in the form

$$\hat{S} = WZ\hat{V}_\xi,$$

where K_Γ and Γ are the same as in Section 1.6.6. (check the section number!!!)

We suppose that X and Y are Banach spaces, and that by analogy with (5.1)–(5.4), $Z = Z_c$ where

$$Z_c(\gamma) = (g_1(c_1; \gamma), g_2(c_2; \gamma), \ldots, g_n(c_n; \gamma)) \tag{5.8}$$

and

$$g_k(c_k; \gamma) = \sum_{s=0}^{p} c_{k,s} r_s(\gamma) \tag{5.9}$$

where $p = (p_1, p_2, \ldots, p_m) \in \mathcal{Z}_+^m$ is fixed and where

$$r_s(\gamma) = \gamma^s = {\gamma_1}^{s_1} {\gamma_2}^{s_2} \ldots {\gamma_m}^{s_m}. \tag{5.10}$$

The neighbourhoods of zero $\xi, \theta, \zeta \subseteq \mathbb{R}^m$ can be chosen to be closed and bounded. Fix ξ, θ, ζ and the method of calculation of the parameters and introduce the class $\hat{\mathcal{S}}$ of operators given by

$$\hat{\mathcal{S}} = \{\hat{S} \mid \hat{S} : K_{\Gamma+\theta} \to Y \text{ and } \hat{S} = \hat{S}_c = WZ_cQ\hat{V}_\xi\}. \tag{5.11}$$

Thus the operator $\hat{S}_c$ is completely defined by the coefficients $\{c_{k,s}\}$. We suppose the map

$$\mathcal{R}_\zeta : \Gamma + \zeta \to \mathbb{R}^n$$

is written in the form

$$\mathcal{R}_\zeta(\gamma) = (f_1(\gamma), f_2(\gamma), \ldots, f_n(\gamma)) \tag{5.12}$$

and let $\{g_k(c_k^*; \gamma)\}$ denote the functions which best approximate the given functions $\{f_k(\gamma)\}$ on the closed and bounded interval $\Gamma + \zeta \subseteq \mathbb{R}^m$.

We have the following theorem.

Theorem 27. *Let X, Y be Banach spaces, let $K \subseteq X$ be a compact set and $F : K \to Y$ a continuous map. Let the operator $Z^* : \mathbb{R}^m \to \mathbb{R}^n$ be defined by*

$$Z^* = Z_{c^*} \tag{5.13}$$

Then for some fixed $\alpha > 0$ and for all $x, x' \in X$ with $\|x' - x\| < \alpha$ the operator $\hat{S}^* = \hat{S}_{c^*} : X \to Y$ in the form

$$\hat{S}^* = WZ^*\hat{V}_\xi \tag{5.14}$$

satisfies the equality

$$\sup_{x \in K} \|F(x) - \hat{S}^*(x')\| = \inf_{\hat{S} \in \hat{\mathcal{S}}} \left\{ \sup_{x \in K} \|F(x) - \hat{S}(x')\| \right\}. \tag{5.15}$$

The scheme of numerical realization of the operator $\hat{S}$ consists of the following steps. Firstly it is necessary to implement a method for approximate determination of the parameter γ. Secondly it is necessary to construct the functions $g_1, g_2, \ldots, g_n$ and consequently the operator Z and thirdly it is necessary to construct an appropriate operator W.

We will illustrate these procedures with an example involving parameter estimation.

Example 14. *Consider the following situation. Let X be a space of measurable functions. We will suppose that a set of incoming signals has the form $K = \{x_\gamma\}_{\gamma \in \Gamma}$ where each signal $x_\gamma \in X$ is completely specified by the*

value of a parameter $\gamma \in \Gamma \subseteq \mathbb{R}^m$. *By observing an individual signal from this set we obtain a measurement*

$$\delta = R(\gamma) \in \Delta \subseteq \mathbb{R}^n$$

from which we wish to estimate the value of the unknown parameter γ.

Therefore the natural estimation procedure can be regarded as a dynamical system represented by a mapping

$$R : \Gamma \to \Delta$$

with input $\gamma \in \Gamma \subseteq \mathbb{R}^m$ *and output* $\delta \in \Delta \subseteq \mathbb{R}^n$. *We wish to construct a best possible approximation to this system in the sense of Theorem 27. Thus we must show that the mapping* $R : \Gamma \to \Delta$ *can be approximated by an operator*

$$\hat{S} : \mathbb{R}^m \to \mathbb{R}^n.$$

Since the output from the system is the parameter δ *itself we have* $y_\delta = \delta$ *and the general output structure is simplified.*

In our example we let $X = L \times L$ *where* L *is the space of all measurable functions* $x : [0, \infty) \to \mathbb{R}$ *such that*

$$\|x\| = \int_0^\infty |x(t)|dt < \infty. \tag{5.16}$$

We assume that the observed signal has the form $x_\gamma = (x_{\gamma,1}, x_{\gamma,2}) \in X$ *where*

$$x_{\gamma,1}(t) = \exp(-t)(\cos \gamma_1 t + \frac{\sin \gamma_2 t}{t}) \tag{5.17}$$

and

$$x_{\gamma,2}(t) = t x_{\gamma,1}(t) \tag{5.18}$$

and where

$$\gamma = (\gamma_1, \gamma_2) \in [-1, 1] \times [-1, 1] = \Gamma \subseteq \mathbb{R}^2$$

is the unknown parameter. To estimate γ *we take a Fourier cosine transform for* x_γ. *In particular the transform is used to determine the DC-component of each signal. Define* $\mathcal{X}_\gamma(\omega) = (\mathcal{X}_{\gamma,1}(\omega), \mathcal{X}_{\gamma,2}(\omega))$. *It is easily shown that*

$$\begin{aligned} \mathcal{X}_{\gamma,1}(\omega) &= \int_0^\infty x_{\gamma,1}(t) \cos \omega t dt \\ &= \frac{1}{2}\left[\frac{1}{1 + (\omega + \gamma_1)^2} + \frac{1}{1 + (\omega - \gamma_1)^2}\right] \\ &\qquad + \frac{1}{2}\left[\arctan(\omega + \gamma_2) - \arctan(\omega - \gamma_2)\right] \end{aligned} \tag{5.19}$$

and

$$\begin{aligned}\mathcal{X}_{\gamma,2}(\omega) &= \int_0^\infty x_{\gamma,2}(t)\cos\omega t dt \\ &= \frac{1}{2}\left[\frac{1-(\omega+\gamma_1)^2}{[1+(\omega+\gamma_1)^2]^2}+\frac{1-(\omega-\gamma_1)^2}{[1+(\omega-\gamma_1)^2]^2}\right] \\ &\quad +\frac{1}{2}\left[\frac{\omega+\gamma_2}{1+(\omega+\gamma_2)^2}-\frac{\omega-\gamma_2}{1+(\omega-\gamma_2)^2}\right]. \quad (5.20)\end{aligned}$$

Thus we calculate

$$\delta_1 = \mathcal{X}_{\gamma,1}(0) = \frac{1}{1+{\gamma_1}^2} + \arctan\gamma_2 \quad (5.21)$$

and

$$\delta_2 = \mathcal{X}_{\gamma,2}(0) = \frac{1-{\gamma_1}^2}{[1+{\gamma_1}^2]^2} + \frac{\gamma_2}{1+{\gamma_2}^2}. \quad (5.22)$$

In effect we have defined a non-linear system which is described by a map $R:\Gamma\to\mathbb{R}^2$ *given by*

$$\delta = R(\gamma) \quad (5.23)$$

where $\delta=(\delta_1,\delta_2)\in\mathbb{R}^2$. *The non-linear system has input* $\gamma\in\Gamma$ *and output* $\delta\in R(\Gamma)=\Delta$. *It is clear that the above formulae for* $\delta=\mathcal{X}_\gamma(0)$ *can be applied to all* $\gamma\in\mathbb{R}^2$ *to define an extended map* $\mathcal{R}:\mathbb{R}^2\to\mathbb{R}^2$.

We seek the best possible approximation to the extended operator $\mathcal{R}$ *in the following sense. Let* H *be the Hilbert space of measurable functions* $f:[-1,1]\to\mathbb{R}$ *such that*

$$\int_{-1}^{1}\frac{[f(s)]^2 ds}{\sqrt{(1-s^2)}} < \infty \quad (5.24)$$

with inner product

$$\langle f,g\rangle = \int_{-1}^{1}\frac{f(s)g(s)ds}{\sqrt{(1-s^2)}}. \quad (5.25)$$

Let $\mathcal{P}_m\subseteq H$ *be the subspace of polynomials of degree at most* $m-1$. *For each* $f\in H$ *there exists a unique polynomial* $p_m=p_m(f)\in\mathcal{P}_m$ *which minimizes the integral*

$$E(f,p) = \|f-p\|^2 = \langle f-p, f-p\rangle \quad (5.26)$$

over all $p\in\mathcal{P}_m$. *It is well known that* $p_m(f)=\Pi_m(f)$ *where* $\Pi_m:H\to\mathcal{P}_m$ *is the Chebyshev projection operator defined in Example 6. Therefore*

$$p_m = \sum_{j=1}^{m} c_j T_{j-1} \quad (5.27)$$

where T_{j-1} is the Chebyshev polynomial of the first kind of degree $j-1$ and where the coefficients $c_j = c_j(f)$ are calculated using the integral formulae given in Example 6. We write $f \sim p_m(f)$.

In this example we will take $m = 6$. Define functions $\{f_{ij}\}_{i,j\in\{1,2\}} \in H$ by the formulae

$$f_{11}(s) = \frac{1}{1+s^2}, \quad f_{12}(s) = \arctan s, \quad f_{21}(s) = \frac{1-s^2}{(1+s^2)^2}$$

and

$$f_{22}(s) = \frac{s}{1+s^2}.$$

The corresponding projections $\{p_{ij}\}_{i,j\in\{1,2\}} \in \mathcal{P}_6$ are given by

$$\begin{aligned} p_{11} &= \frac{\sqrt{2}}{2}T_0 - (3\sqrt{2}-4)T_2 + (17\sqrt{2}-24)T_4 \\ &\approx (.7071)T_0 - (.2426)T_2 + (.0416)T_4, \\ p_{12} &= (2\sqrt{2}-2)T_1 - \frac{(10\sqrt{2}-14)}{3}T_3 + \frac{(58\sqrt{2}-82)}{5}T_5 \\ &\approx (.8284)T_1 - (.0474)T_3 + (.0049)T_5, \\ p_{21} &= \frac{\sqrt{2}}{4}T_0 - \frac{(8-5\sqrt{2})}{2}T_2 + \frac{(112-79\sqrt{2})}{2}T_4 \\ &\approx (.3536)T_0 - (.4645)T_2 + (.1386)T_4, \end{aligned} \tag{5.28}$$

and

$$\begin{aligned} p_{22} &= (2-\sqrt{2})T_1 - (10-7\sqrt{2})T_3 + (58-41\sqrt{2})T_5 \\ &\approx (.5858)T_1 - (.1005)T_3 + (.0172)T_5. \end{aligned} \tag{5.29}$$

The theoretical system is therefore replaced by a more practical system described by a map $Z^* : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$\delta = Z^*(\gamma) \tag{5.30}$$

where

$$\delta_1 = p_{11}(\gamma_1) + p_{12}(\gamma_2) \tag{5.31}$$

and

$$\delta_2 = p_{21}(\gamma_1) + p_{22}(\gamma_2). \tag{5.32}$$

In actual fact the calculations will be based on one further approximation. In practice we choose a large value of T and calculate $\delta_T = \mathcal{R}_T(\gamma)$ using

$$\delta_{T,1} = \int_0^T x_{\gamma,1}(t)dt \tag{5.33}$$

and

$$\delta_{T,2} = \int_0^T x_{\gamma,2}(t)dt. \tag{5.34}$$

Associated with each δ_T there is a uniquely defined (virtual) measurement $\hat{\gamma}$ defined by

$$\hat{\gamma} = \mathcal{R}^{-1}(\delta_T). \tag{5.35}$$

Therefore we have a (virtual) measurement scheme defined by an operator $\hat{V}_T : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$\hat{V}_T = \mathcal{R}^{-1}\mathcal{R}_T \tag{5.36}$$

and written in the form

$$\hat{\gamma} = \hat{V}_T(\gamma). \tag{5.37}$$

The practical measurement system is now described by an operator

$$\hat{S} = Z^*\hat{V}_T = Z^*\mathcal{R}^{-1}\mathcal{R}_T$$

with output given by

$$\begin{aligned} \delta &= \hat{S}(\gamma) \\ &= Z^*\hat{V}_T(\gamma) \\ &= Z^*(\hat{\gamma}). \end{aligned} \tag{5.38}$$

The operator $\hat{S}$ is the best possible approximation in the sense of Theorem 27.

To estimate the parameter γ we take the (observed) value $\delta_T = \mathcal{R}_T(\gamma)$ of δ and compute

$$\gamma_{est} = Z^{*-1}(\delta_T) = Z^{*-1}\mathcal{R}_T(\gamma).$$

For example this could be done by using a Newton iteration to solve the equation

$$Z^*(\gamma_{est}) = \delta_T. \tag{5.39}$$

5.3 Estimation of Mean and Covariance Matrix for Random Vectors

In the next Sections, methods for modelling of nonlinear systems that transform stochastic signals will be considered under an assumption that their mean and covariance matrix either are known or can be estimated. Below, we present some methods for estimating the mean and the covariance matrix.

Throughout the next Sections, (Ω, Σ, μ) signifies a probability space, where Ω is the set of outcomes, Σ a σ–field of measurable subsets $\Delta \subset \Omega$ and $\mu : \Sigma \to [0, 1]$ an associated probability measure on Σ with $\mu(\Omega) = 1$. Each element $\omega \in \Omega$ represents the outcome of an experiment and each subset Δ of Ω is a set of outcomes, called an event. We say that the event Δ has occurred if $\omega \in \Delta$.

5.3.1 Maximum likelihood estimates

Suppose that $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$ and $\mathbf{y} \in L^2(\Omega, \mathbb{R}^n)$ are random vectors such that $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)^T$ and $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$ with $\mathbf{x}_i, \mathbf{y}_k \in L^2(\Omega, \mathbb{R})$ for $i = 1, \ldots, m$ and $k = 1, \ldots, n$, respectively. Let

$$x = \mathbf{x}(\omega) \in \mathbb{R}^m \quad \text{and} \quad y = \mathbf{y}(\omega) \in \mathbb{R}^n$$

be realizations of $\mathbf{x}$ and $\mathbf{y}$ so that $x = (x_1, \ldots, x_m)^T$ and $y = (y_1, \ldots, y_n)^T$ with $x_i, y_k \in \mathbb{R}$ for $i = 1, \ldots, m$ and $k = 1, \ldots, n$.

Random vectors will be denoted by bold letters.

Let us write

$$E[\mathbf{x}_i \mathbf{y}_k] = \int_\Omega \mathbf{x}_i(\omega)\mathbf{y}_k(\omega) d\mu(\omega) < \infty,$$

$$E[\mathbf{x}] = \begin{bmatrix} E[\mathbf{x}_1] \\ \vdots \\ E[\mathbf{x}_m] \end{bmatrix}, \quad E[\mathbf{x}\mathbf{y}^T] = \begin{bmatrix} E[\mathbf{x}_1\mathbf{y}_1] & \ldots & E[\mathbf{x}_1\mathbf{y}_n] \\ \vdots & \ddots & \vdots \\ E[\mathbf{x}_m\mathbf{y}_1] & \ldots & E[\mathbf{x}_m\mathbf{y}_n] \end{bmatrix}$$

and

$$E_{xy} = E[\mathbf{x}\mathbf{y}^T] - E[\mathbf{x}]E[\mathbf{y}^T].$$

Given N independent realizations $x^{(1)}$, ..., $x^{(N)}$ of the random vector $\mathbf{x}$, the maximum likelihood (ML) estimates $\hat{E}[\mathbf{x}]$ and $\hat{E}_{xx}$ for $E[\mathbf{x}]$ and E_{xx} respectively, under the Gaussian assumption, are known to be [1]

$$\hat{E}[\mathbf{x}] = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} \tag{5.40}$$

and

$$\hat{E}_{xx} = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)} - \hat{E}[\mathbf{x}])(x^{(i)} - \hat{E}[\mathbf{x}])^T. \tag{5.41}$$

In many real world situations, such estimates are difficult to use. Below, we consider the estimating methods subject to specific realistic conditions.

5.3.2 Estimates from incomplete realizations

Often, the source of complexity in using formulas (5.40) and (5.41) is that not every component of each realization $x^{(i)}$ is observed, so that (5.40) and (5.41) cannot be used. Ad hoc modifications of these estimators are likely to produce unsatisfactory results. For example, one couuld arbitrary assign the value of zero to all missing components and then directly use (5.40) and (5.41), however severe biases would occer. Another approach would replace the normalizing factor $\frac{1}{N}$ in (5.40) and (5.41) by a factor that would vary from one component to another and which would be chosen to give unbiased estimates. Thus, the ith component of $\hat{E}[\mathbf{x}]$ would be the arithmetic mean of the ith component of $\mathbf{x}$ over all realizations for which it is available. Nevertheless, this procedure could result in a covariance matrix with a negative eigenvalue.

Here, we consider the method [102] which is motivated to give maximum likelihood estimates in an effort to improve upon the ad hoc estimates.

The specific restriction of the method is that it must be possible to order the components of the random vector such that the set of realizations for which the ith component is available is a subset of the set of realizations for which the $(i-1)$th component is available, for $i = 2, \ldots, M$.

This restriction is satisfied in cases where a sequence of limited-resource sensors is used, with each subsequent sensor observing a subset of objects that were observed by the previous sensor.

Let $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$ be a normal random vector with the probability density function given by

$$\begin{aligned} p(x_1, \ldots, x_m) &= \frac{1}{(2\pi)^{m/2}[\det(E_{xx})]^{1/2}} \\ &\quad \times \exp\left[-\frac{1}{2}(x - E[\mathbf{x}])^T E_{xx}^{-1}(x - E[\mathbf{x}])\right], \end{aligned} \tag{5.42}$$

where $\det(E_{xx})$ is the determinant of E_{xx}, $\det(E_{xx}) \neq 0$ and E_{xx} is positive definite.

The problem is to estimate the mean and the covariance matrix of $\mathbf{x}$ under assumption that not all of the m components are necessarily observed.

In what follows the observations are denoted by a set X illustrated in Fig. 1:

$$X = \{\mathbf{x}_{ik} \mid i = 1, \ldots, m, \quad k = 1, \ldots, N_i\},$$

where i indicates the component of the random vector, and k is the index of the realization.

$$X = \begin{bmatrix} \mathbf{x}_{11} & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \mathbf{x}_{1N_1} \\ \mathbf{x}_{21} & \ldots & \ldots & \ldots & \ldots & \ldots & \mathbf{x}_{2N_2} & \circ & \ldots & \circ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}_{m1} & \ldots & \mathbf{x}_{m,N_m} & \circ & \ldots & \circ & \ldots & \circ & \ldots & \circ \end{bmatrix}$$

Fig. 1. Structure of the set comprising independent realizations of the random vector $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$; the symbol $\circ$ represents missing data.

In Fig. 1, the symbol $\circ$ denotes missing data and it is not zero. Thus, $\mathbf{x}_{i_1 k_1}$ and $\mathbf{x}_{i_2 k_2}$ are statistically independent if $k_1 \neq k_2$. The total number of observations of the ith component is N_i. It is assumed that

$$N_m \leq N_{m-1} \leq \ldots \leq N_1.$$

The solution of the problem formulated above is based on a representation of E_{xx} via the LDL^T factorization [50] and a further determination of the probability density function $p(x_1, \ldots, x_i)$.

First, let E_{xx} be nonsingular and let

$$E_{xx} = MDM^T \tag{5.43}$$

be the LDL^T factorization of E_{xx} where M is lower triangular matrix with ones on the main diagonal and D is diagonal matrix. Then

$$E_{xx}^{-1} = M^{-T} D^{-1} M^{-1} = L^{-T} D^{-1} L^{-1},$$

where $L = M^{-1}$.

Each matrix $F : \mathbb{R}^{n \times m}$ defines an associated operator $\mathcal{F} : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^p)$ via the equation

$$[\mathcal{F}(\mathbf{x})](\omega) = F[\mathbf{x}(\omega)] \tag{5.44}$$

for each $\omega \in \Omega$.

Let $\mathcal{L} : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^m)$ be the operator defined similarly to $\mathcal{F}$ via (5.44) and let

$$\mathbf{u} = \mathcal{L}(\mathbf{x}) \quad \text{and} \quad u = Lx$$

where $\mathbf{u} \in L^2(\Omega, \mathbb{R}^m)$ and $u = \mathbf{u}(\omega)$.

Then we have

$$E[\mathbf{u}] = LE[\mathbf{x}]$$

and

$$E[(\mathbf{u} - E[\mathbf{u}])(\mathbf{u} - E[\mathbf{u}])^T] = LE_{xx}L^T = D.$$

Because of the lower triangular structure of L, the mean and the covariance matrix of the first i components of $\mathbf{u}$ are obtainable from the mean and the covariance matrix of the first i components of $\mathbf{x}$ as follows:

$$\begin{bmatrix} E[\mathbf{u}_1] \\ E[\mathbf{u}_2] \\ \vdots \\ E[\mathbf{u}_i] \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ L_{21} & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ L_{i1} & L_{i2} & \dots & L_{i,i-1} & 1 \end{bmatrix} \begin{bmatrix} E[\mathbf{x}_1] \\ E[\mathbf{x}_2] \\ \vdots \\ E[\mathbf{x}_i] \end{bmatrix} \tag{5.45}$$

and

$$\begin{bmatrix} D_1 & 0 & \dots & \dots & 0 \\ 0 & D_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & D_i \end{bmatrix} \tag{5.46}$$

$$= \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ L_{21} & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ L_{i1} & L_{i2} & \dots & L_{i,i-1} & 1 \end{bmatrix} \begin{bmatrix} E_{x_1x_1} & \dots & E_{x_1x_i} \\ E_{x_2x_1} & \dots & E_{x_2x_i} \\ \vdots & \vdots & \vdots \\ E_{x_ix_1} & \dots & E_{x_ix_i} \end{bmatrix}$$

$$\times \begin{bmatrix} 1 & L_{21} & \dots & \dots & L_{i1} \\ 0 & 1 & L_{32} & \dots & L_{i2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{5.47}$$

where D_1, D_2, …, D_i are diagonal entries of D.

From (5.47), it follows that

$$\det \begin{bmatrix} E_{x_1x_1} & E_{x_1x_2} & \dots & E_{x_1x_i} \\ E_{x_2x_1} & E_{x_2x_2} & \dots & E_{x_2x_i} \\ \vdots & \vdots & \vdots & \vdots \\ E_{x_ix_1} & E_{x_ix_2} & \dots & E_{x_ix_i} \end{bmatrix} = \prod_{j=1}^{i} D_j \tag{5.48}$$

and

$$\begin{bmatrix} E_{x_1x_1} & E_{x_1x_2} & \dots & E_{x_1x_i} \\ E_{x_2x_1} & E_{x_2x_2} & \dots & E_{x_2x_i} \\ \vdots & \vdots & \vdots & \vdots \\ E_{x_ix_1} & E_{x_ix_2} & \dots & E_{x_ix_i} \end{bmatrix}^{-1} \tag{5.49}$$

$$= \begin{bmatrix} 1 & L_{21} & \dots & \dots & L_{i1} \\ 0 & 1 & L_{32} & \dots & L_{i2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} D_1 & 0 & \dots & \dots & 0 \\ 0 & D_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & D_i \end{bmatrix}^{-1}$$

$$\times \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ L_{21} & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ L_{i1} & L_{i2} & \dots & L_{i,i-1} & 1 \end{bmatrix} \tag{5.50}$$

On the basis of (5.42), (5.48) and (5.50), the joint log-probability density function[1] for the first i components of $\mathbf{x}$ is

$$\log p(x_1, \dots, x_i) = -\frac{i}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{i}\log D_j$$

$$-\sum_{j=1}^{i}\frac{1}{2D_j}(x_j - E[\mathbf{u}_j] + L_{j1}x_1 + \ldots + L_{j,j-1}x_{j-1})^2. \tag{5.51}$$

Next, an expression for the log-probability density function $\log p(X)$ of the incomplete data set X can now be obtained as follows. Let us represent X in the form

$$X = \bigcup_{i=1}^{m} X_i,$$

where, for $i = 1, \dots, m$,

$$X_i = \{\mathbf{x}_{jk} \mid j = 1, \dots, i, \quad k = N_{i+1} + 1, \dots, N_i\}.$$

Here, $N_{m+1} = 0$ and if $N_{i+1} = N_i$ then X_i is the empty set.

Thus, X_i comprises $N_i - N_{i+1}$ realizations of the first i components of $\mathbf{x}$. The $\mathbf{x}_{jk}$ are uncorrelated over k (representing different realizations), while for a particular value of k, the i correlated components have a log-probability density function of the form (5.51). Noting that the X_i are

[1] Also called the joint log-llikelihood function [**?**], [**?**].

uncorrelated, the log-probability density function for X is

$$\begin{aligned}\log p(X) &= \sum_{i=1}^{m}\sum_{k=N_{i+1}+1}^{N_i}\left[-\frac{i}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{i}\log D_j\right. \\ &- \left.\frac{1}{2}\sum_{j=1}^{i}\frac{1}{D_j}(x_j - E[\mathbf{u}_j] + L_{j1}x_1 + \ldots + L_{j,j-1}x_{j-1})^2\right], \quad (5.52)\end{aligned}$$

which can be simplified to

$$\begin{aligned}\log p(X) &= \sum_{i=1}^{m}\left[-\frac{N_i}{2}\log(2\pi) - \frac{N_i}{2}\log D_i\right. \\ &- \left.\frac{1}{2D_i}\sum_{k=1}^{N_i}(x_{ik} - E[\mathbf{u}_i] + L_{i1}x_{1k} + \ldots + L_{i,i-1}x_{i-1,k})^2\right]. \quad (5.53)\end{aligned}$$

This strikingly simple formula indicates that the computation of the log-probability density function involves applying regression operators of length 1 through m to the largest subsets of X for which the corrresponding components are available. The regression operators, in turn, produce uncorrelated residuals.

The maximum likelihood estimates $\hat{L}$, $\hat{D}$ and $\hat{E}[\mathbf{u}]$ for L, D and $E[\mathbf{u}]$, respectively, follow from choosing these quantities to maximize the log-probability density function (5.53). Estimates of the mean and the covariance matrix are obtained from $\hat{L}$, $\hat{D}$ and $\hat{E}[\mathbf{u}]$.

The structure of $\log p(X)$ by (5.53) implies that the problem of estimating $E[\mathbf{u}_i]$, D_i, and the ith row of L are decoupled for different values of i. Maximizing (5.53) over $E[\mathbf{u}_i]$ and the i row of L involves solving the following least squares problem (for example, using the QR factorization):

$$\min\sum_{k=1}^{N_i}(x_{ik} - E[\mathbf{u}_i] + L_{i1}x_{1k} + \ldots + L_{i,i-1}x_{i-1,k})^2. \quad (5.54)$$

Maximizing (5.53) over D_i is equivalent to choosing D_i to be equal to the minimum of (5.54), divided by N_i,

$$\min\frac{1}{N_i}\sum_{k=1}^{N_i}(x_{ik} - E[\mathbf{u}_i] + L_{i1}x_{1k} + \ldots + L_{i,i-1}x_{i-1,k})^2$$

for $i = 1, \ldots, m$.

If $N_i > i$ for all i, so that

$$N_1 \geq N_2 \geq \ldots \geq N_m \geq m$$

(i.e., every component is observed at least $m+1$ times), and if the true covariance matrix is positive definite, then the D_i are strictly positive (so the estimated covariance matrix is positive definite).[2]

The decoupling of $E[\mathbf{u}_i]$, D_i, and the ith row of L, for different i, facilitates the computation of the Cramer-rao bounds, since it imparts a block-diagonal structure to the Fisher information matrix.

Example 15. *It is instructive to consider in detail the bivariate case (i.e., $m = 2$). We further simplify the problem by assuming that the true mean is zero, $E[\mathbf{x}] = 0$. Here, we present the exact maximum likelihood estimates for thr elements of the covarince matrix.*

For $m = 2$, and with the assumption $E[\mathbf{x}] = 0$, it is straightforward to obtain the maximum likelihood estimates for L and D. Then transforming back to E_{xx} gives the following exact maximum likelihood estimates:

$$\hat{E}_{x_1x_1} = \frac{1}{N_1}\sum_{k=1}^{N_1} x_{1k}^2, \quad \hat{E}_{x_1x_2} = \frac{1}{N_1}\sum_{k=1}^{N_1} x_{1k}x_{2k}\frac{\dfrac{1}{N_1}\displaystyle\sum_{k=1}^{N_1} x_{1k}^2}{\dfrac{1}{N_2}\displaystyle\sum_{k=1}^{N_2} x_{2k}^2}$$

and

$$\hat{E}_{x_2x_2} = \frac{1}{N_2}\sum_{k=1}^{N_2} x_{2k}^2 - \frac{\left[\dfrac{1}{N_2}\displaystyle\sum_{k=1}^{N_2} x_{1k}x_{2k}\right]^2}{\dfrac{1}{N_2}\displaystyle\sum_{k=1}^{N_2} x_{2k}^2}\left[\frac{1}{N_2}\sum_{k=1}^{N_2} x_{2k}^2 - \frac{1}{N_1}\sum_{k=1}^{N_1} x_{1k}^2\right].$$

In summary, the above method provides the maximum likelihood estimates for the mean and the covariance matrix of a random vector $\mathbf{x}$ with the normal distribution. The method works under assumption that the components of the random vector can be ordered such that the set of realizations for which the ith component is available is a subset of realizations for which the $(i-1)$th component is available, for $i = 2, \ldots, m$. When the matrix dimension p is large than the number N of observations available,

5.3.3 A well-conditioned estimator for large-dimensional covariance matrices

In the next Sections, an estimate of a covariance matrix $E_{xx} \in \mathbb{R}^{m\times m}$ and/or its inverse can be required, where m is large compared to the sam-

[2]The covariance matrix (5.43) is positive definite if and only if all the diagonal elements of D are strictly positive.

ple size N. In such situations, the usual estimator – the sample covariance matrix $\hat{E}_{xx}$ by (5.41) – is known to perform poorly. When the matrix dimension m is large than the number N of observations available, the sample covariance matrix $\hat{E}_{xx}$ is not even invertible. When the ratio m/N is less than one but not negligible, $\hat{E}_{xx}$ is invertible but numerically ill-conditioned, which means that inverting it amplifies estimation error dramatically. For large m, it is difficult to find enough observations to make m/N negligible, and therefore, it is important to develop a well-conditioned estimator for large-dimensional covariance matrices.
To the best of our knowledge, no existing estimator is both well-conditioned *and* more accurate than the sample covariance matrix. Here, we consider the method [83] that is both well-conditioned *and* more accurate than the sample covariance matrix asymptotically.

One way to to get a well-conditioned structured estimator is to impose the condition that all variances are the same and all covariances are zero. The estimator which is considered below is *a weighted average of this structured estimator* and *the sample covariance matrix.* The estimator inherits the good conditioning properties of the structured estimator and, by choosing the weight optimally according to a quadratic loss function, it is ensured that the weighted average of the sample covariance matrix and the structured estimator is more accurate than either of them.

The only difficulty is that the true optimal weight depends on the true covariance matrix, which is unobservable. This difficulty is solved by finding a consistent estimator of the optimal weight, and show that replacing the true optimal weight with a consistent estimator makes no difference asymptotically.

Standard asymptotics assume that the number of variables m is finite and fixed, while the number of observations N goes to infinity. Under standard asymptotics, the sample covariance matrix is well-conditioned (in the limit), and has some appealing optimality properties (e.g., it is maximum likelihood estimator for normally distributed data). However, this is a bad approximation of many real-world situations where the number of variables m is of the same order of magnitude as the number of observations N, and possibly large.

In method [83], a different framework is used, called *general* asymptotics, where the number of variables m can go to infinity as well.

The only constraint is that the ratio m/N must remain bounded.

We see standard asymptotics as a special case where it is optimal to put (asymptotically) all the weight on the sample covariance matrix and none on the structured estimator. In the general case, however, the estimator considered below is asypmtotically different from the sample covariance matrix, substantially more accurate, and of course well-conditioned.

Analysis in finite sample

The easiest way to explain the core of the method is to first analyze the finite sample case.

Let $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ be a set of independent and identically distributed (iid) random vectors $\mathbf{x}^{(1)} \in L^2(\Omega, \mathbb{R}^m)$, ..., $\mathbf{x}^{(N)} \in L^2(\Omega, \mathbb{R}^m)$ with mean zero such that $\mathbf{x}^{(k)} = [\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_m^{(k)}]^T$ for $k = 1, \ldots, N$ where $\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_m^{(k)} \in L^2(\Omega, \mathbb{R})$. We call $\mathbf{X}$ the finite sample of random vectors.

Since $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$ are the iid random vectors, we denote

$$E[\mathbf{x}_j] := E[\mathbf{x}_j^{(1)}] = \ldots = E[\mathbf{x}_j^{(N)}] \quad \text{for } j = 1, \ldots, m \tag{5.55}$$

and

$$E[\mathbf{x}_i \mathbf{x}_j] := E[\mathbf{x}_i^{(1)} \mathbf{x}_j^{(1)}] = \ldots = E[\mathbf{x}_i^{(N)} \mathbf{x}_j^{(N)}] \quad \text{for } i, j = 1, \ldots, m. \tag{5.56}$$

Now, we write

$$\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^T,$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_m$ satisfy (5.55) and (5.56). It is clear that $\mathbf{x}$ has mean zero.

Let us also denote

$$\mathbf{S} = \frac{1}{N} \mathbf{X}\mathbf{X}^T$$

and call $\mathbf{S}$ the sample covariance vector. Due to (5.55) and (5.56),

$$E[\mathbf{S}] = E[\mathbf{x}\mathbf{x}^T].$$

Furthermore, we write

$$\mathbf{C} = \rho_1 I + \rho_2 \mathbf{S} \tag{5.57}$$

and

$$J(\mathbf{C}) = E[\|E[\mathbf{x}\mathbf{x}^T] - \mathbf{C}\|^2], \tag{5.58}$$

where $\rho_1, \rho_2 \in \mathbb{R}$, I is the identity matrix and $\|\cdot\|$ is the Frobenius-like norm defined by

$$\|A\| = \left[\frac{\operatorname{tr}(AA^T)}{m}\right]^{1/2}$$

for $A \in \mathbb{R}^{m \times m}$.

The problem is to find ρ_1^0 and ρ_1^0 such that

$$J(\mathbf{C}^0) = \min_{\rho_1, \rho_2} J(\mathbf{C}), \tag{5.59}$$

where

$$\mathbf{C}^0 = \rho_1^0 I + \rho_2^0 \mathbf{S}. \tag{5.60}$$

The solution is given by Theorem 28 below in terms of matrix $E[\mathbf{x}\mathbf{x}^T]$. In this sense, $\mathbf{C}^0$ is not a *bona fide* estimator. In the next subsection, a *bona fide* estimator $\mathbf{S}^*$ is developed with the same properties as $\mathbf{C}^0$ *asymptotically as* N *and* m *go to infinity together.*

For $A_1, A_2 \in \mathbb{R}^{m\times m}$, we write $\langle A_1, A_2\rangle = \mathrm{tr}(A_1 A_2^T)/m$.

Lemma 16. *Let*

$$\kappa = \langle E[\mathbf{x}\mathbf{x}^T], I\rangle, \quad \alpha^2 = \|E[\mathbf{x}\mathbf{x}^T] - \kappa I\|^2,$$

$$\beta^2 = E[\|\mathbf{S} - E[\mathbf{x}\mathbf{x}^T]\|^2] < \infty \quad \textit{and} \quad \delta^2 = E[\|\mathbf{S} - \kappa I\|^2] < \infty.$$

Then

$$\alpha^2 + \beta^2 = \delta^2.$$

Proof. . We have

$$\begin{aligned} E[\|\mathbf{S} - \kappa I\|^2] &= E[\|\mathbf{S} - E[\mathbf{x}\mathbf{x}^T] + E[\mathbf{x}\mathbf{x}^T] - \kappa I\|^2] \\ &= E[\|\mathbf{S} - E[\mathbf{x}\mathbf{x}^T]\|^2] + E[\|E[\mathbf{x}\mathbf{x}^T] - \kappa I\|^2] \\ &\quad +2E[\langle \mathbf{S} - E[\mathbf{x}\mathbf{x}^T], E[\mathbf{x}\mathbf{x}^T] - \kappa I\rangle] \\ &= E[\|\mathbf{S} - E[\mathbf{x}\mathbf{x}^T]\|^2] + E[\|E[\mathbf{x}\mathbf{x}^T] - \kappa I\|^2] \\ &\quad +2\langle E[\mathbf{S} - E[\mathbf{x}\mathbf{x}^T]], E[\mathbf{x}\mathbf{x}^T] - \kappa I\rangle \\ &= E[\|\mathbf{S} - E[\mathbf{x}\mathbf{x}^T]\|^2] + E[\|E[\mathbf{x}\mathbf{x}^T] - \kappa I\|^2] \end{aligned}$$

because $E[\mathbf{S}] = E[\mathbf{x}\mathbf{x}^T]$. □

Theorem 28. *The solution to problem* (5.59) *is given by*

$$\mathbf{C}^0 = \frac{\beta^2}{\delta^2}\kappa I + \frac{\alpha^2}{\delta^2}\mathbf{S}. \tag{5.61}$$

The error associated with $\mathbf{C}^0$ *is given by*

$$E[\|E[\mathbf{x}\mathbf{x}^T] - \mathbf{C}^0\|^2] = \frac{\alpha^2\beta^2}{\delta^2}. \tag{5.62}$$

Proof. By a change of variables, (5.57) and (5.59) can be rewritten as

$$\mathbf{C} = \rho\nu I + (1-\rho)\mathbf{S}$$

and

$$J(\mathbf{C}^0) = \min_{\rho,\nu} J(\mathbf{C}),$$

respectively. With a little algebra, we can rewrite $J(\mathbf{C})$ as

$$\begin{aligned} E[\|E[\mathbf{x}\mathbf{x}^T] - \mathbf{C}\|^2 &= \rho^2 \|E[\mathbf{x}\mathbf{x}^T] - \nu I\|^2 \\ &\quad + (1-\rho)^2 E[\|E[\mathbf{x}\mathbf{x}^T] - \mathbf{S}\|^2. \end{aligned} \tag{5.63}$$

Therefore, the optimal value of ν can be obtained as the solution to a reduced problem that does not depend on ρ:

$$\min_{\nu} \|E[\mathbf{x}\mathbf{x}^T] - \nu I\|^2.$$

The norm of identity is one by convention, so

$$\|E[\mathbf{x}\mathbf{x}^T] - \nu I\|^2 = \|E[\mathbf{x}\mathbf{x}^T]\|^2 - 2\nu\langle E[\mathbf{x}\mathbf{x}^T], I\rangle + \nu^2.$$

The first-order condition is

$$-2\langle E[\mathbf{x}\mathbf{x}^T], I\rangle + 2\nu = 0.$$

The solution is

$$\begin{aligned} \nu &= \langle E[\mathbf{x}\mathbf{x}^T], I\rangle \\ &= \kappa. \end{aligned}$$

replacing ν by its optimal value κ in (5.63), we have

$$E[\|E[\mathbf{x}\mathbf{x}^T] - \mathbf{C}\|^2 = \rho^2\alpha^2 + (1-\rho)^2\beta^2.$$

The desired ρ is

$$\begin{aligned} \rho &= \frac{\beta^2}{\alpha^2 + \beta^2} \\ &= \frac{\beta^2}{\delta^2}. \end{aligned}$$

Note that

$$1 - \rho = \frac{\alpha^2}{\delta^2}.$$

At the optimum, $E[\|E[\mathbf{x}\mathbf{x}^T] - \mathbf{C}^0\|^2]$ becomes

$$\begin{aligned} E[\|E[\mathbf{x}\mathbf{x}^T] - \mathbf{C}^0\|^2] &= \frac{\beta^2}{\delta^2}\alpha^2 + \frac{\alpha^2}{\delta^2}\beta^2 \\ &= \frac{\alpha^2\beta^2}{\delta^2}. \end{aligned}$$

This completes the proof. □

Analysis under general asymptotics

The solution (5.61) does not provide a *bona fide* estimator, since it requires hindsight knowledge of $E[\mathbf{x}\mathbf{x}^T]$. To avoid this difficulty, the consistent estimators for κ, α, β and δ can be obtained in the following way.

First, an appropriate asymptotic framework is chosen. Standard asymptotics consider m fixed while N tends to infinity, implying that the optimal shrinkage intensity vanishes in the limit. This would be reasonable for situations where m is very small in comparison to N. However, in the problems of interest to us m tends to be of the same order as N and can even be larger. Hence, we consider it more appropriate to use a framework that reflects this condition. This is achieving by allowing the number of variables m to go to infinity at the same "speed" as the number of observations N. It is called *general asymptotics*. In this framework, the optimal shrinkage intensity generally does not vanish asymptotically but rather it tends to a limiting constant that it is possible to estimate consistently. the idea then is to use the estimated shrinkage intensity in order to arrive at a *bona fide* estimator.

Let $N = 1, 2, \ldots$ index a sequence of statistical models. For every N, $\mathbf{X}_N = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ is a set of independent and identically distributed random vectors $\mathbf{x}^{(1)} \in L^2(\Omega, \mathbb{R}^{m_N}), \ldots, \mathbf{x}^{(N)} \in L^2(\Omega, \mathbb{R}^{m_N})$ with mean zero such that $\mathbf{x}^{(k)} = [\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_{m_N}^{(k)}]^T$ for $k = 1, \ldots, N$ where $\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_{m_N}^{(k)} \in L^2(\Omega, \mathbb{R})$.

The number m_N can change and even go to infinity with the number N, but nit too fast.

Assumption 1. There exists a constant K_1 independent of N such that $m_N/N \leq K_1$.

Let $E[\mathbf{x}\mathbf{x}^T] = U\sigma U^T$ be the eigenvalue decomposition of $E[\mathbf{x}\mathbf{x}^T]$. Let $Y_N = U^T\mathbf{X}_N$ and $[y_{11}^N, \ldots, y_{m_N 1}^N]^T$ the first column of Y_N.

Assumption 2. There exists a constant K_2 independent of N such that

$$\frac{1}{m_N}\sum_{i=1}^{m_N} E[(y_{i1}^N)^8] \leq K_2.$$

Assumption 3.

$$\lim_{N\to\infty} \frac{m_N^2}{N^2} \frac{\sum_{(i,j,k,l)\in Q_N} (\mathrm{Cov}[y_{i1}^N y_{j1}^N y_{k1}^N y_{l1}^N])^2}{\text{Cardinal of } Q_N}$$

Assumption 2 states that the eight moment is bounded (in average). Assumption 3 states that products of uncorrelated random variables are

themselves uncorrelated (on average, in the limit). In the case where general asymptotics degenerate into standard asymptotics ($\frac{m_N}{N} \to 0$), Assumption 3 is trivially verified as a consequence of Assumption 2. Assumption 3 is verified when random variables are normally or even elliptically distributed, but it is much weaker than that.

For $A \in \mathbb{R}^{m_N \times m_N}$, the Frobenius -like norm is defined by

$$\|A\|_N^2 = \frac{\mathrm{tr}(AA^T)}{m_N}.$$

We follow the notation used in the preceding Section, except that we add the subscript N to signal that all results hold asymptotically. In particular, we now write $\mathbf{C}_N^0$ instead of $\mathbf{C}^0$. Also, let us denote

$$\mathbf{S}_N = \frac{1}{N}\mathbf{X}_N\mathbf{X}_N^T, \quad \kappa_N = \langle E[\mathbf{x}\mathbf{x}^T], I\rangle, \quad \alpha_N^2 = \|E[\mathbf{x}\mathbf{x}^T] - \kappa I\|_N^2,$$

$$\beta_N^2 = E[\|\mathbf{S}_N - E[\mathbf{x}\mathbf{x}^T]\|_N^2] < \infty \quad \text{and} \quad \delta_N^2 = E[\|\mathbf{S}_N - \kappa_N I\|_N^2] < \infty.$$

These four scalars are well behaved asymptotically.

Lemma 17. *κ_N, α_N^2, β_N^2 and δ_N^2 remain bounded as $N \to \infty$.*

We omit proofs of Lemmata and Theorems in this Section. The proofs can be found in [83].

The most basic question is whether $\mathbf{S}_N$ is consistent under general asymptotics. Specifically, we ask whether $\mathbf{S}_N$ converges in quadratic mean to the true covariance matrix, that is, whether β_N^2 vanishes. In general, the answer is no, as shown below. The results stated below are related to special cases of a general result proven by Yin [184]. But presented method works under weaker assumptions than the method [184] does. Also, the goal in [184] is to find the distribution of the eigenvalues of the sample covariance matrix, while this method is to find an improved estimator of the covariance matrix.

Theorem 29. *Let*

$$\theta_N^2 = Var\left[\frac{1}{m_N}\sum_{j=1}^{m_N}(y_{j1}^N)^2\right].$$

The scalar θ_N^2 is bounded as $N \to \infty$, and

$$\lim_{N\to\infty}\{E[\|\mathbf{S}_N - E[\mathbf{x}\mathbf{x}^T]\|_N^2 - \frac{m_N}{N}(\kappa_N^2 - \theta_N^2)\} = 0.$$

For κ_N, a consistent estimator is its sample counterpart as it follows from the next Lemma. We write $\xrightarrow{\text{q.m.}}$ to denote convergence in quadratic mean.

Lemma 18. *Let*

$$m_N = \langle \mathbf{S}_N, I \rangle.$$

Then

$$E[m_N] = \kappa_N,$$

and

$$m_N - \kappa_N \xrightarrow{\text{q.m.}} 0.$$

as $N \to \infty$.

A consistent estimator for δ_N^2 is also its sample counterpart.

Lemma 19. *Let*

$$d_N^2 = \|\mathbf{S}_N - m_N I\|_N^2.$$

Then

$$d_N^2 - \delta_N^2 \xrightarrow{\text{q.m.}} 0.$$

Now, note that $\mathbf{S}_N$ can be represented as

$$\mathbf{S}_N = \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}^{(k)} \mathbf{x}^{(k)T}.$$

Since the matrices $\mathbf{x}^{(k)}\mathbf{x}^{(k)T}$ are iid across k, we can estimate the error $\beta^2 = E[\|\mathbf{S}_N - E[\mathbf{x}\mathbf{x}^T]\|_N^2]$ of their average by seeing how far each one of them deviates from the average.

Lemma 20. *Let*

$$\tilde{b}_N^2 = \frac{1}{N^2} \sum_{k=1}^{N} \|\mathbf{S}_N - \mathbf{x}^{(k)} \mathbf{x}^{(k)T}\|_N^2$$

and

$$b_N^2 = \min\{\tilde{b}_N^2, d_N^2\}.$$

Then

$$\tilde{b}_N^2 - \beta_N^2 \xrightarrow{\text{q.m.}} 0 \quad \textit{and} \quad b_N^2 - \beta_N^2 \xrightarrow{\text{q.m.}} 0.$$

We note that $b_N^2 \leq \delta_N^2$ by Lemma 16. In general, this constraint is rarely binding. But it insures that the following estimator of α_N^2 is nonnegative.

Lemma 21. *Let* $a_N^2 = d_N^2 - b_N^2$. *Then*

$$a_N^2 - \alpha_N^2 \xrightarrow{\text{q.m.}} 0.$$

The next stage of the strategy is to replace the unobservable scalars in the formula defining $\mathbf{C}^0$ with consistent estimators, and to show that the asymptotic properties are unchanged. This yields the *bona fide* estimator of the covariance matrix:

$$\mathbf{S}_N^0 = \frac{b_N^2}{d_N^2} m_N I + \frac{a_N^2}{d_N^2} \mathbf{S}_N. \tag{5.64}$$

The next Theorem shows that $\mathbf{S}_N^0$ has the same asymptotic properties as $\mathbf{C}_N^0$. Thus, we can neglect the error associated with the replacement of the unobservable parameters κ_N, α_N^2, β_N^2 and δ_N^2 by estimators.

Theorem 30. $\mathbf{S}_N^0$ *is a consistent estimator of* $\mathbf{C}_N^0$, *i.e.*

$$\|\mathbf{S}_N^0 - \mathbf{C}_N^0\|_N \xrightarrow{\text{q.m.}} 0.$$

As a consequence, $\mathbf{S}_N^0$ *has the same associated asymptotic error as* $\mathbf{C}_N^0$, *i.e.*

$$E[\|\mathbf{S}_N^0 - \mathbf{C}_N\|_N^2] - E[\|\mathbf{C}_N^0 - \mathbf{C}_N\|_N^2] \xrightarrow{\text{q.m.}} 0.$$

The following result presents the estimate of the associated error of $\mathbf{C}_N^0$ and $\mathbf{S}_N^0$ consistently.

Lemma 22.

$$E\left[\left\| \frac{a_N^2 b_N^2}{d_N^2} - \frac{\alpha_N^2 \beta_N^2}{\delta_N^2} \right\|_N^2\right] \to 0.$$

The final step is to demonstrate that $\mathbf{S}_N^0$, which has been obtained as a consistent estimator for $\mathbf{C}_N^0$, possesses an important optimality property. It follows from Theorem 28 that $\mathbf{C}_N^0$ (hence, $\mathbf{S}_N^0$ in the limit) is optimal among the linear combinations (5.57) with *nonrandom* coefficients. this interesting, but only midly so, because it excludes the other linear shtinkage estimators with random coefficients.

Below, it is shown that $\mathbf{S}_N^0$ is still optimal within a bigger class: the linear combinations like (5.57) but with *random* coefficients. This class includes both the linear combinations that represent *bona fide* estimators, and those with coefficients that require hindsight knowledge of the true (and unobservable) covariance matrix.

Let

$$J(\mathbf{C}_N^*) = \|E[\mathbf{x}\mathbf{x}^T] - \mathbf{C}_N^*\|_N^2$$

and

$$\mathbf{C}_N^* = \boldsymbol{\rho}_1 I + \boldsymbol{\rho}_1 \mathbf{S}_N$$

where unlike ρ_1 and ρ_2 in (5.57), $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_1$ are random variables. Another difference is the norm $\|\cdot\|_N$ instead of the norm $E\|\cdot\|_N$ in (5.58).

Let $\mathbf{C}_N^{*0}$ be such that

$$J(\mathbf{C}_N^{*0}) = \min J(\mathbf{C}_N^*).$$

It turns out that $\mathbf{C}_N^{*0}$ is a function of $E[\mathbf{x}\mathbf{x}^T]$ therefore $\mathbf{C}_N^{*0}$ does not constitute a *bona fide* estimator. By construction, $\mathbf{C}_N^{*0}$ has lower associated error than $\mathbf{C}_N^*$ and $\mathbf{S}_N^0$ almost surely (a.s.), but asymptotically it makes no difference.

Theorem 31. *$\mathbf{S}_N^0$ is a consistent estimator of $\mathbf{C}_N^{*0}$, i.e.*

$$\|\mathbf{S}_N^0 - \mathbf{C}_N^{*0}\|_N \xrightarrow{\text{q.m.}} 0.$$

As a consequence, $\mathbf{S}_N^0$ has the same associated asymptotic error as $\mathbf{C}_N^{*0}$, i.e.

$$E[\|\mathbf{S}_N^0 - \mathbf{C}_N\|_N^2] - E[\|\mathbf{C}_N^{*0} - \mathbf{C}_N\|_N^2] \xrightarrow{\text{q.m.}} 0.$$

Both $\mathbf{C}_N^0$ and $\mathbf{C}_N^{*0}$ have the same asymptotic properties as $\mathbf{S}_N^0$; therefore, they also have the same asymptotic properties as each other.

The most important result of this section is as follows:

The *bona fide* estimator $\mathbf{S}_N^0$ has uniformly minimum associated error asymptotically among all linear combinations of I with $\mathbf{S}_N$, including those that are *bona fide* estimators, and even those that use hindsight knowledge of the true covariance matrix.

Theorem 32. *For any sequence of linear combinations $\hat{\mathbf{C}}_N$ of I and $\mathbf{S}_N$, the estimator $\mathbf{S}_N^0$ defined in (5.64) verifies*

$$\lim_{N_0 \to \infty} \inf_{N \geq N_0} \left\{ E[\|\hat{\mathbf{C}}_N - E[\mathbf{x}\mathbf{x}^T]\|_N^2] - E[\|\mathbf{S}_N^0 - E[\mathbf{x}\mathbf{x}^T]\|_N^2] \right\} \geq 0.$$

In addition, every $\hat{\mathbf{C}}_N$ that performs as well as $\mathbf{S}_N^0$ is identical to $\mathbf{S}_N^0$ in the limit:

$$\lim_{n \to \infty} \left\{ E[\|\hat{\mathbf{C}}_N - E[\mathbf{x}\mathbf{x}^T]\|_N^2] - E[\|\mathbf{S}_N^0 - E[\mathbf{x}\mathbf{x}^T]\|_N^2] \right\} = 0$$

is equivalent to

$$\|\hat{\mathbf{C}}_N - \mathbf{S}_N^0\|_N \xrightarrow{\text{q.m.}} 0.$$

Thus, $\mathbf{S}_N^0$ is an asymptotically optimal linear shrinkage estimator of the covariance matrix with respect to associated error under general asymptotics.

A distinctive feature of this result is that it provides the rigorous justification when the number of variables m_N exceeds the number of observations N. Not only that, but $\mathbf{S}_N^0$ is guaranteed to be always invertible, even in the case $m_N > N$, where rank deficiency makes the sample covariance matrix singular. estimating the inverse covariance matrix when variables outnumber observations is sometimes dismissed as impossible, but the existence of $(\mathbf{S}_N^0)^{-1}$ certainly proves otherwise.

The following theorem shows that $\mathbf{S}_N^0$ is usually well-conditioned.

Theorem 33. *Let the condition number of the true covariance matrix $E[\mathbf{x}\mathbf{x}^T]$ be bounded, and let the normalized variables $y_{i1}/\sqrt{\lambda_i}$ be iid across $i = 1, \ldots, N$. Then the condition number of the estimator $\mathbf{S}_N^0$ is bounded in probability.*

This theorem follows from [2].

If the cross-sectional iid assumption is violated, it does not mean that the condition number goes to infinity, but rather that it is technically too difficult to find out anything about it. Interestingly, there is one case where the estimator $\mathbf{S}_N^0$ is even better-conditioned than the true covariance matrix $E[\mathbf{x}\mathbf{x}^T]$: if the ill-conditioning of $E[\mathbf{x}\mathbf{x}^T]$ comes from eigenvalues close to zero (multi-collinearity in the variables) and the ratio of variables to observations m_N/N is not negligible. In this case, $\mathbf{S}_N^0$ is well-conditioned because the sample observations do not provide enough information to update our prior belief that there is no multi-collinearity.

5.3.4 Some other relevant results on estimates of covariance matrix

The covariance matrix estimation is an area of intensive research. Below, we mention some results which are relevant to the methods discussed above.

Calvin and Dykstra [13] considered the problem of estimating covariance matrix in balanced multivariate variance components models. As with univariate models, it is possible for the traditional estimators, based on differences of the mean square matrices, to produce estimates that are outside the parameter space. In fact, in many cases it is extremely likely that traditional estimates of the covariance matrices will not be non-negative definite. In [13], Calvin and Dykstra developed an iterative procedure, satisfying a least squares criterion, that is guaranteed to produce non-negative definite estimates of covariance matrices and provide an analysis of convergence.

In some applications the covariance matrix of the observations enjoys a particular symmetry: it is not only symmetric with respect to its main diagonal but also with respect to the anti-diagonal. The standard forward-only sample covariance estimate does not impose this extra symmetry. In such cases one often uses the so-called forward-backward sample covariance estimate. Jansson and Stoica [67] performed a direct comparative study of the relative accuracy of the two sample covariance estimates is performed. An explicit expression for the difference between the estimation error covariance matrices of the two sample covariance estimates is given. This expression shows quantitatively the gain of using the forward-backward estimate compared to the forward-only estimate. The results [67] are also useful in the analysis of estimators based on either of the two sample covariances. As an example, in [67], spatial power estimation by means of the Capon method [145] is considered. Using a second-order approximation, it is shown that Capon based on the forward-only sample covariance (F-Capon) underestimates the power spectrum, and also that the bias for Capon based on the forward-backward sample covariance is half that of F-Capon.

Delmash [28] studied estimators, both batch and adaptive, of the eigenvalue decomposition (EVD) of centrosymmetric (CS) covariance matrices. These estimators make use of the property that eigenvectors and eigenvalues of such structured matrices can be estimated via two decoupled eigensystems. As a result, the number of operations is roughly halved, and moreover, the statistical properties of the estimators are improved. In [28], after deriving the asymptotic distribution of the EVD estimators, the closed-form expressions of the asymptotic bias and covariance of the EVD estimators are compared to those obtained when the CS structure is not taken into account. As a by-product, it is shown [28] that the closed-form expressions of the asymptotic bias and covariance of the batch and adaptive EVD estimators are very similar provided that the number of samples is replaced by the inverse of the step size.

Kauermann and Carroll considered the sandwich covariance matrix estimation [72]. The sandwich estimator, also known as robust covariance matrix estimator, heteroscedasticity-consistent covariance matrix estimate, or empirical covariance matrix estimator, has achieved increasing use in the literature as well as with the growing popularity of generalized estimating equations. Its virtue is that it provides consistent estimates of the covariance matrix for parameter estimates even when the fitted parametric model fails to hold or is not even specified. Surprisingly though, there has been little discussion of properties of the sandwich method other than consistency. Kauermann and Carroll investigate the sandwich estimator in quasi-

likelihood models asymptotically, and in the linear case analytically. They show that under certain circumstances when the quasi-likelihood model is correct, the sandwich estimate is often far more variable than the usual parametric variance estimate. The increased variance is a fixed feature of the method and the price that one pays to obtain consistency even when the parametric model fails or when there is heteroscedasticity. It is shown in [72] that the additional variability directly affects the coverage probability of confidence intervals constructed from sandwich variance estimates. In fact, the use of sandwich variance estimates combined with t-distribution quantiles gives confidence intervals with coverage probability falling below the nominal value. Kauermann and Carroll propose an adjustment to compensate for this fact.

Kubokawa and Srivastava [80] considered the problem of estimating the covariance matrix and the generalized variance when the observations follow a nonsingular multivariate normal distribution with unknown mean. They present a new method to obtain a truncated estimator that utilizes the information available in the sample mean matrix and dominates the James-Stein minimax estimator [66]. Several scale equivariant minimax estimators are also given.

This method is then applied to obtain new truncated and improved estimators of the generalized variance; it also provides a new proof to the results of Shorrok and Zidek [138] and Sinha [139].

Champion [14] derived and evaluated an algorithm for estimating normal covariances. A particular concern in [14] is the performance of the estimator when the dimension of the space exceeds the number of observations. The algorithm is simple, tolerably well founded, and seems to be more accurate for its purpose than the alternatives. Other topics discussed in [14] are the joint estimation of variances in one and many dimensions; the loss function appropriate to a variance estimator; and its connection with a certain Bayesian prescription.

Schneider and Willsky [133] proposed a new iterative algorithm for the simultaneous computational approximation to the covariance matrix of a random vector and drawing a sample from that approximation. The algorithm is especially suited to cases for which the elements of the random vector are samples of a stochastic process or random field. The proposed algorithm has close connections to the conjugate gradient method for solving linear systems of equations.

A comparison has been made between the algorithm's structure and complexity and other methods for simulation and covariance matrix approximation, including those based on FFTs and Lanczos methods. The

convergence of the proposed iterative algorithm is analyzed, and a preconditioning technique for accelerating convergence is explored.

5.4 Best Hadamard-quadratic Approximation

In the next sections, we consider the best constructive approximation of the input-output map of the system in a general stochastic setting when the input-output map is an arbitrary nonlinear continuous operator, the inputs and outputs are stochastic signals and the only information on a system is given by certain covariance matrices formed from the input-output signals.

It is known that a nonlinear system provides more flexibility in its performance than that by a linear system. Needless to say that an approximator with a nonlinear structure is a natural tool in nonlinear system modelling, because such an approximator provides, in particular, a higher accuracy than a linear model. The question is what kind of nonlinearity should be used for an effective approximation. The answer depends on criterion which we aim to achieve in the system modelling. In the following sections, we consider different types of nonlinear approximators.

We begin with the so called Hadamard-quadratic approximation.

5.4.1 Statement of the problem

The following preliminaries are necessary to pose the problem properly.

We interpret $\mathbf{x}$ as a given "idealized" input signal (without any distortion) of a nonlinear system, and $\mathbf{y}$ as its observed input signal. In particular, $\mathbf{y}$ can be interpreted as $\mathbf{x}$ contaminated with noise so that no specific relationships between signal and noise are assumed to be known. For instance, noise can be additive or multiplicative or their combination.

Let $\mathcal{F} : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^p)$ be the input-output map of a nonlinear system.

The terms "system" and "input-output map" will be identified.

We consider the class $\mathbb{A}$ of models $\mathcal{A} : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^p)$ of a nonlinear system with $\mathcal{A}$ given by the equation

$$\mathcal{A}(y) = \mathcal{A}_0 + \mathcal{A}_1(\mathbf{y}) + \mathcal{A}_2(\mathbf{y}^2), \tag{5.65}$$

where $\mathcal{A}_0 \in L^2(\Omega, \mathbb{R}^p)$, $\mathcal{A}_1, \mathcal{A}_2 : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^p)$ and $\mathbf{y}^2$ is defined by the Hadamard product [50] so that

$$y^2 = (y_1^2, \dots, y_n^2)^T.$$

Operators $\mathcal{A}_1$ and $\mathcal{A}_2$ are linear and are defined by matrices $A_1 \in \mathbb{R}^{p \times n}$ and $A_2 \in \mathbb{R}^{p \times n}$ so that

$$[\mathcal{A}_1(\mathbf{x})](\omega) = A_1[\mathbf{x}(\omega)] \quad \text{and} \quad [\mathcal{A}_2(\mathbf{x})](\omega) = A_2[\mathbf{x}(\omega)]. \tag{5.66}$$

For any random vector $\mathbf{z} \in L^2(\Omega, \mathbb{R}^m)$, we write

$$E[\|\mathbf{z}\|^2] = \int_\Omega \|\mathbf{z}(\omega)\|^2 d\mu(\omega), \tag{5.67}$$

where $\|\mathbf{z}(\omega)\|$ is the Euclidean norm of $\mathbf{z}(\omega)$.

Then in accordance with (5.67),

$$E[\|\mathcal{F}(\mathbf{x}) - \mathcal{A}(\mathbf{y})\|^2] = \int_\Omega \|[\mathcal{F}(\mathbf{x})](\omega) - [\mathcal{A}(\mathbf{y})](\omega)\|^2 d\mu(\omega).$$

Hereinafter, calligraphic letters will be used to signify operators defined similarly to $\mathcal{F}$, $\mathcal{A}_1$ and $\mathcal{A}_2$.

Let us denote

$$J(A_0, A_1, A_2) = E[\|\mathcal{F}(\mathbf{x}) - \mathcal{A}(\mathbf{y})\|^2]. \tag{5.68}$$

We wish to find A_0^0, A_1^0, A_2^0 so that

$$J(A_0^0, A_1^0, A_2^0) = \min_{A_0, A_1, A_2} J(A_0, A_1, A_2) \tag{5.69}$$

for all A_j with $j = 0, 1, 2$.

In other words, we wish to find the model $\mathcal{A}^0$ of the system $\mathcal{F}$ which is the best in the class $\mathbb{A}$ in the sense (5.69).

It is natural to call $\mathcal{A}^0$ the best Hadamard-quadratic approximation to $\mathcal{F}$.

Generalizations of the Hadamard-quadratic model $\mathcal{A}$ are discussed in the next Sections.

5.4.2 Auxiliary results

We denote by $\mathcal{N}(M)$ and $\mathcal{R}(M)$ the null space and range space respectively of a matrix M, and by $M^\dagger$ the Moore-Penrose pseudo-inverse of M.

Note that

$$\mathcal{N}(M) = \mathcal{R}(M^T)^\perp \quad \text{and} \quad \mathcal{N}(M^T) = \mathcal{R}(M)^\perp$$

where $\mathcal{R}(M^T)^\perp$ and $\mathcal{R}(M)^\perp$ are the orthogonal compliments of $\mathcal{R}(M^T)$ and $\mathcal{R}(M)$, respectively (see, for example, [89], p. 155).

Proposition 1. *For any random vector* $\mathbf{z} \in L^2(\Omega, \mathbb{R}^m)$,

$$E[\|\mathbf{z}\|^2] = \operatorname{tr} E[\mathbf{z}\mathbf{z}^T].$$

Proof. We have

$$\begin{aligned} E[\|\mathbf{z}\|^2] &= \int_\Omega \|\mathbf{z}(\omega)\|^2 d\mu(\omega) \\ &= \int_\Omega \mathrm{tr}\,\{\mathbf{z}(\omega)\mathbf{z}^T(\omega)\} d\mu(\omega) \\ &= \sum_{j=1}^{m} \int_\Omega [\mathbf{z}_j(\omega)]^2 d\mu(\omega) \\ &= \mathrm{tr}\,\{E_{zz}\}, \end{aligned}$$

where $\mathbf{z}^T(\omega) = [\mathbf{z}(\omega)]^T$. □

Lemma 23. Let $P \in \mathbb{R}^{n\times m}$ and $Q \in \mathbb{R}^{n\times m}$. Then

$$\mathcal{N}(P) \subseteq \mathcal{N}(Q) \tag{5.70}$$

implies

$$QP^\dagger P = Q.$$

Proof. If $q_\mathcal{N} \in \mathcal{N}(P)$, then

$$QP^\dagger P q_\mathcal{N} = 0$$

and since equation (5.70) shows that $q_\mathcal{N} \in \mathcal{N}(Q)$, we also have $Qq_\mathcal{N} = 0$. Hence

$$Q(I - P^\dagger P)q_\mathcal{N} = 0,$$

where I is the identity matrix.

On the other hand, if $q_\mathcal{R} \in \mathcal{N}(P)^\perp = \mathcal{R}(P^T)$, then there exists $p \in \mathbb{R}^n$ such that $q_\mathcal{R} = P^\dagger p$ [89]. Hence

$$QP^\dagger P q_\mathcal{R} = QP^\dagger P P^\dagger p = Qq_\mathcal{R}$$

and therefore

$$Q(I - P^\dagger P)q_\mathcal{R} = 0.$$

The desired result follows from the unique representation of any vector $q \in \mathbb{R}^n$ in the form $q = q_\mathcal{N} + q_\mathcal{R}$. □

We write

$$\mathbf{s} = \mathcal{F}(\mathbf{x}) \quad \text{and} \quad s = \mathbf{s}(\omega),$$

and

$$\mathbf{z} = \mathbf{y}^2 \quad \text{and} \quad z = \mathbf{z}(\omega).$$

Lemma 24. *The following equations hold:*

$$E_{sy}E_{yy}^{\dagger}E_{yy} = E_{sy}, \quad E_{zy}E_{yy}^{\dagger}E_{yy} = E_{zy} \tag{5.71}$$

and

$$E_{sz}E_{zz}^{\dagger}E_{zz} = E_{sz}. \tag{5.72}$$

Proof. If $u \in \mathcal{N}(E_{yy})$ then $u^T E_{yy} u = 0$ and hence

$$E[(\mathbf{y}^T u)^2] = 0.$$

But for each $w \in \mathbb{R}^m$ we have

$$\begin{aligned} |w^T E_{sy} u| &= |E[(w^T \mathbf{s})(\mathbf{y}^T u)]| \\ &\leq (E[(w^T \mathbf{s})^2])^{1/2} (E[(\mathbf{y}^T u)^2])^{1/2} \\ &= 0. \end{aligned}$$

Therefore $E_{sy}u = 0$ and hence $u \in \mathcal{N}(E_{sy})$. This means that

$$\mathcal{N}(E_{yy}) \subseteq \mathcal{N}(E_{sy}) \tag{5.73}$$

and then the first equation in (5.71) follows from (5.73) on the basis of Lemma 23. Other equations in (5.71) and (5.72) are proved similarly. □

Lemma 25. *Let*

$$B = E_{zz} - E_{zy}E_{yy}^{\dagger}E_{yz} \quad \textit{and} \quad G = E_{sz} - E_{sy}E_{yy}^{\dagger}E_{yz}.$$

Then

$$GB^{\dagger}B = G. \tag{5.74}$$

Proof. Let $u \in \mathcal{N}(B)$. Then $u^T B u = 0$ and

$$u^T E[(\mathbf{z} - E_{zy}E_{yy}^{\dagger}\mathbf{y})(\mathbf{z}^T - \mathbf{y}^T E_{yy}^{\dagger}E_{yz})]u = 0.$$

Therefore

$$E[\{(\mathbf{z} - E_{zy}E_{yy}^{\dagger}\mathbf{y})^T u\}^2] = 0.$$

Next, for all $v \in \mathbb{R}^m$ we have

$$\begin{aligned} (v^T G u)^2 &= (E[v^T(\mathbf{s} - E_{sy}E_{yy}^{\dagger}\mathbf{y})(\mathbf{z}^T - \mathbf{y}^T E_{yy}^{\dagger}E_{yz})u])^2 \\ &\leq E[\{v^T(\mathbf{s} - E_{sy}E_{yy}^{\dagger}\mathbf{y})\}^2] E[\{(\mathbf{z} - E_{zy}E_{yy}^{\dagger}\mathbf{y})^T u\}^2] \\ &= 0 \end{aligned}$$

i.e. $Gu = 0$ or $u \in \mathcal{N}(G)$. Hence

$$\mathcal{N}(B) \subseteq \mathcal{N}(G)$$

and then (5.74) follows from Lemma 23. □

Note, that B, $G \neq \mathbb{O}$, in general. Here, $\mathbb{O}$ is the zero matrix. The following elementary example illustrates this fact.

Example 16. *Let* $\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}$ *and* $\mathbf{x} = \begin{pmatrix} \mathbf{y}_1\mathbf{y}_2 \\ 0 \end{pmatrix}$, *where* $\mathbf{y}_1$, $\mathbf{y}_2$ *are independent random variables taking values 1 and* -1. *Then we have*

$$E[\mathbf{y}_i = -1] = 1/2, \quad E[\mathbf{y}_i = 1] = 1/2, \quad \textit{for each} \quad i = 1, 2,$$

$$E_{xz} = \begin{bmatrix} E[\mathbf{y}_1^3\mathbf{y}_2] & E[\mathbf{y}_1^2\mathbf{y}_2^2] \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

$$E_{xy} = \begin{bmatrix} E[\mathbf{y}_1^2\mathbf{y}_2] & E[\mathbf{y}_1\mathbf{y}_2^2] \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$E_{zy} = \begin{bmatrix} E[\mathbf{y}_1^3] & E[\mathbf{y}_1^2\mathbf{y}_2] \\ E[\mathbf{y}_1^2\mathbf{y}_2] & E[\mathbf{y}_1\mathbf{y}_2] \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$E_{zz} = \begin{bmatrix} E[\mathbf{y}_1^4] & E[\mathbf{y}_1^3\mathbf{y}_2] \\ E[\mathbf{y}_1^3\mathbf{y}_2] & E[\mathbf{y}_1^2\mathbf{y}_2^2] \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Therefore

$$B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \textit{and} \quad G = I.$$

Next, it is well known (see, for example, [10], p. 8) that for any matrix M,

$$M^\dagger = M^T(MM^T)^\dagger. \tag{5.75}$$

We denote by $M^{1/2}$ a matrix such that $M^{1/2}M^{1/2} = M$.

If M is a symmetric non-negative definite matrix then we can write

$$M = V\Sigma V^T$$

where V is orthogonal and Σ is a non-negative diagonal matrix. We note that

$$M^\dagger = V\Sigma^\dagger V^T \quad \text{and} \quad M^{1/2} = V\Sigma^{1/2}V^T$$

and that consequently

$$(M^{1/2})^\dagger = (M^\dagger)^{1/2}.$$

Let us denote $M^{\dagger 1/2} = (M^\dagger)^{1/2}$.

We will use (5.75) in the next Lemma.

Lemma 26. *The equations*

$$A_1 E_{yy}^{1/2} = (E_{sy} - A_2 E_{zy}) E_{yy}^{\dagger} E_{yy}^{1/2}, \tag{5.76}$$

$$A_2 B^{1/2} = G B^{\dagger} B^{1/2} \tag{5.77}$$

are respectively equivalent to

$$A_1 E_{yy} = E_{sy} - A_2 E_{zy}, \tag{5.78}$$

$$A_2 B = G. \tag{5.79}$$

Proof. Let us suppose that (5.77) is true. Multiplying on the right by $B^{1/2}$ gives

$$A_2 B - G B^{\dagger} B = 0.$$

Then $A_2 B - G = 0$ follows on the basis of (5.74).

On the other hand, if $A_2 B - G = 0$ then multiplying on the right by $B^{\dagger}$ gives

$$A_2 B B^{\dagger} - G B^{\dagger} = 0. \tag{5.80}$$

If we set $A = B^{1/2}$ then (5.75) implies

$$B^{1/2\dagger} = B^{1/2} B^{\dagger}$$

from which it follows that

$$B^{1/2} B^{1/2\dagger} = B B^{\dagger}.$$

Hence, equation (5.80) can be rewritten as

$$A_2 B^{1/2} B^{1/2\dagger} - G B^{\dagger} = 0.$$

Multiplying on the right by $B^{1/2}$ gives the required result

$$(A_2 - G B^{\dagger}) B^{1/2} = 0.$$

The equivalence of (5.76) and (5.78) is proved similarly. Namely, if (5.76) is true then multiplying on the right by $E_{yy}^{1/2}$ gives (5.78). If (5.78) is true then multiplying on the right by $E_{yy}^{\dagger}$ and applying (5.75) gives (5.76). □

5.4.3 The best Hadamard-quadratic model

Theorem 34. *The solution to problem* (5.69) *is given by*

$$A_0^0 = E[\mathbf{s}] - A_1^0 E[\mathbf{y}] - A_2^0 E[\mathbf{z}], \tag{5.81}$$

$$A_1^0 = (E_{sy} - A_2^0 E_{zy})E_{yy}^{\dagger} + M_1(I - E_{yy}E_{yy}^{\dagger}) \tag{5.82}$$

and

$$A_2^0 = GB^{\dagger} + M_2(I - BB^{\dagger}), \tag{5.83}$$

where $M_1 \in \mathbb{R}^{p\times n}$, $M_2 \in \mathbb{R}^{p\times n}$ *are arbitrary matrices and* I *is the identity matrix.*

Proof. The functional $J(A_0, A_1, A_2)$ can be written as

$$J(A_0, A_1, A_2) = J_0 + J_1(A_0, A_1, A_2) + J_2(A_1, A_2) + J_3(A_2), \tag{5.84}$$

where

$$\begin{aligned} J_0 &= \mathrm{tr}\{E_{ss} - E_{sy}E_{yy}^{\dagger}E_{ys}\} - \Delta, && (5.85)\\ \Delta &= \|G(B^{\dagger})^{\frac{1}{2}}\|^2, && (5.86)\\ J_1(A_0, A_1, A_2) &= \|A_0 - (E[\mathbf{s}] - A_1E[\mathbf{y}] - A_2E[\mathbf{z}])\|^2 && (5.87)\\ J_2(A_1, A_2) &= \|[A_1 - (E_{sy} - A_2E_{zy})E_{yy}^{\dagger}]E_{yy}^{1/2}\|^2 && (5.88) \end{aligned}$$

and

$$J_3(A_2) = \|(A_2 - GB^{\dagger})B^{1/2}\|^2. \tag{5.89}$$

Equations (5.84)–(5.89) are deduced from the representation of the corresponding functionals as follows:

$$\begin{aligned} J_1(A_0, A_1, A_2) &= \mathrm{tr}\{(A_0 - E[\mathbf{s}] + A_1E[\mathbf{y}] + A_2E[\mathbf{z}])\\ &\times (A_0 - E[\mathbf{s}] + A_1E[\mathbf{y}] + A_2E[\mathbf{z}])^T\}, && (5.90) \end{aligned}$$

$$\begin{aligned} J_2(A_1, A_2) &= \mathrm{tr}\{[A_1 - (E_{sy} - A_2E_{zy})E_{yy}^{\dagger}]E_{yy}\\ &\times [A_1 - (E_{sy} - A_2E_{zy})E_{yy}^{\dagger}]^T\} && (5.91) \end{aligned}$$

and

$$\begin{aligned} J_3(A_2) &= \mathrm{tr}\{[A_2 - (E_{sz} - E_{sy}E_{yy}^{\dagger}E_{yz})B^{\dagger}]B\\ &\times [A_2 - (E_{sz} - E_{sy}E_{yy}^{\dagger}E_{yz})B^{\dagger}]^T\}. && (5.92) \end{aligned}$$

Then on the basis of Lemma 24, we obtain

$$\begin{aligned}
&J_0 + J_1(A_0, A_1, A_2) + J_2(A_1, A_2) + J_3(A_2) \\
&= J_0 + J_1(A_0, A_1, A_2) + \mathrm{tr}\{D_1 + D_2 + A_1 E_{yy} A_1^T \\
&-A_1 E_{ys} + A_1 E_{yz} A_2^T - E_{sy} A_1^T + A_2 E_{zy} A_1^T \\
&+(E_{sy} E_{yy}^\dagger E_{yz} B^\dagger B - E_{sz} B^\dagger B - E_{sy} E_{yy}^\dagger E_{yz}) A_2^T \\
&+A_2 (E_{sy} E_{yy}^\dagger E_{yz} B^\dagger B E_{sz} B^\dagger B - E_{sy} E_{yy}^\dagger E_{yz})^T \\
&+A_2 (B + E_{zy} E_{yy}^\dagger E_{yz}) A_2^T\},
\end{aligned}$$

where

$$D_1 = E_{sy} E_{yy}^\dagger E_{ys} \quad \text{and} \quad D_2 = G B^\dagger G^T.$$

Next, taking into account Lemma 25,

$$\begin{aligned}
&J_0 + J_1(A_0, A_1, A_2) + J_2(A_1, A_2) + J_3(A_2) \\
&= J_0 + J_1(A_0, A_1, A_2) + \mathrm{tr}\{D_1 + D_2 + A_1 E_{yy} A_1^T - A_1 E_{ys} \\
&\quad -E_{sy} A_1^T - E_{sz} A_2^T - A_2 E_{zs} + A_2 E_{zz} A_2^T + A_1 E_{yz} A_2^T \\
&\quad +A_2 E_{zy} A_1^T\} \\
&= \mathrm{tr}\{E[(s - A_0 - A_1 y - A_2 z)(s - A_0 - A_1 y - A_2 z)^T]\} \\
&= J(A_0, A_1, A_2).
\end{aligned}$$

It follows from (5.84) – (5.89) that $J(A_0, A_1, A_2)$ is minimized when

$$A_0 = E[\mathbf{s}] - A_1 E[\mathbf{y}] - A_2 E[\mathbf{z}], \tag{5.93}$$

$$A_1 E_{yy}^{1/2} = (E_{sy} - A_2 E_{zy}) E_{yy}^\dagger E_{yy}^{1/2}, \tag{5.94}$$

$$A_2 B^{1/2} = G B^\dagger B^{1/2} \tag{5.95}$$

since

$$\begin{aligned}
E_{yy}^\dagger E_{yy}^{1/2} &= ([E_{yy}^{1/2}]^T E_{yy}^{1/2})^\dagger [E_{yy}^{1/2}]^T \\
&= (E_{yy}^{1/2})^\dagger \\
&= (E_{yy}^\dagger)^{1/2}
\end{aligned}$$

and also

$$B^\dagger B^{1/2} = (B^{1/2})^\dagger$$

for the same reason.

By Lemma 26, the equation (5.95) is equivalent to the equation (5.79). The necessary and sufficient conditions [6] for the equations (5.79) to have the solution is

$$G = G B^\dagger B$$

which is true by Lemma 25. Then the solution is given [6] by (5.83).

Next, also by Lemma 26, the equation (5.94) is equivalent to the equation (5.78) where we now set $A_2 = A_2^0$. The necessary and sufficient conditions [6] for the equations (5.78) to have the solution is

$$(E_{sy} - A_2^0 E_{zy}) E_{yy}^{\dagger} E_{yy} = E_{sy} - A_2^0 E_{zy}$$

which is satisfied by Lemma 24. Therefore, the solution [6] is provided by (5.82). □

Remark 12. *Matrices E_{yy} and B are positive semidefinite and their eigenvalues are nonnegative* ([143], p. 309). *Hence $E_{yy}^{1/2}$ and $B^{1/2}$ have real entries.*

Remark 13. *The solution of each matrix equation* (5.79) *and* (5.78) *is not unique* [6] *and this fact is reflected by the arbitrary matrices M_1 and M_2 in* (5.82), (5.83)*: for any M_1 and M_2, the matrices A_1^0 and A_2^0 are the solutions to the corresponding equations* (5.79) *and* (5.78).

In this connection we note that a possible and natural choice for M_1 and M_2 in equations (5.79) and (5.78) is $M_1 = \mathbb{O}$ and $M_2 = \mathbb{O}$ where $\mathbb{O}$ is the zero matrix, and then equations (5.79) and (5.78) are simplified.

Also note that the best operator $\mathcal{A}^0$ defined by the equations (5.81), (5.82) and (5.83) requires knowledge of the matrices

$$E[\mathbf{s}], \quad E[\mathbf{y}], \quad E[\mathbf{z}], \quad E_{sy}, \quad E_{sz}, \quad E_{yy}, \quad E_{yz} \quad \text{and} \quad E_{zz}. \tag{5.96}$$

The methods for their estimation have been considered in Section 4.3.

Corollary 2. *The best approximation $\tilde{\mathbf{s}}$ of $\mathbf{s} = \mathcal{F}(\mathbf{x})$ in the class $\mathbb{A}$ of the models (5.65) is*

$$\tilde{\mathbf{s}} = \mathcal{A}^0(\mathbf{y}). \tag{5.97}$$

Theorem 35. *The error of approximation of the system $\mathcal{F}$ by the best Hadamard-quadratic model $\mathcal{A}^0$ is*

$$E[\|\mathcal{F}(\mathbf{x}) - \mathcal{A}^0(\mathbf{y})\|^2] = \mathrm{tr}\{E_{ss}\} - \|E_{sy}(E_{yy}^{\dagger})^{1/2}\|^2 - \|G(B^{\dagger})^{1/2}\|^2. \tag{5.98}$$

Proof. The proof follows directly from equations (5.68) and (5.81)–(5.89). □

A particular case of the model $\mathcal{A}$ is the first-degree model $\mathcal{A}_{(1)}$ given by

$$\mathcal{A}_{(1)}(\mathbf{y}) = \mathcal{A}_0 + \mathcal{A}_1(\mathbf{y}). \tag{5.99}$$

Corollary 3. *The best first degree model* $\mathcal{A}^0_{(1)}$, *minimizing the functional* (5.68) *with* $A_2 = 0$, *is defined by*

$$A^0_0 = E[\mathbf{s}] - A^0_1 E[\mathbf{y}] \tag{5.100}$$

and

$$A^0_1 = E_{sy} E^{\dagger}_{yy} + K_1[I - E^{1/2}_{yy}(E^{1/2}_{yy})^{\dagger}], \tag{5.101}$$

where $K_1 \in \mathbb{R}^{p\times n}$ *is an arbitrary matrix.*

Corollary 4. *The error of approximation of the system* $\mathcal{F}$ *by the best first degree model* $\mathcal{A}^0_{(1)}$ *is*

$$E[\|\mathcal{F}(\mathbf{x}) - \mathcal{A}^0_{(1)}(\mathbf{y})\|^2] = \mathrm{tr}\{E_{ss}\} - \|E_{sy}(E^{\dagger}_{yy})^{1/2}\|^2. \tag{5.102}$$

Proof. The equation (5.102) follows from (5.100) and (5.101) on the basis of Lemma 1. □

Remark 14. *A comparison of the error equations* (5.98) *and* (5.102) *shows that the error associated with the best Hadamard-quadratic model is less for* $\|G(B^{\dagger})^{1/2}\|^2$ *than the error associated with the best first degree model.*

Remark 15. *Knowledge of matrices* $E[\mathbf{s}]$, $E[\mathbf{y}]$ *and* $E[\mathbf{z}]$ *allows us to simplify procedures* (5.81), (5.82), (5.83) *and* (5.100), (5.101) *above by the replacement of* $\mathbf{s}$, $\mathbf{y}$ *and* $\mathbf{z}$ *with* $\hat{\mathbf{s}} = \mathbf{s} - E[\mathbf{s}]$, $\hat{\mathbf{y}} = \mathbf{y} - E[\mathbf{y}]$ *and* $\hat{\mathbf{z}} = \mathbf{z} - E[\mathbf{z}]$ *respectively. Then*

$$E[\hat{\mathbf{s}}] = 0, \quad E[\hat{\mathbf{y}}] = 0 \quad \textit{and} \quad E[\hat{\mathbf{z}}] = 0$$

and therefore equations (5.81), (5.82), (5.83) *and* (5.100), (5.101) *with* $\hat{\mathbf{s}}$, $\hat{\mathbf{y}}$ *and* $\hat{\mathbf{z}}$ *instead of* $\mathbf{s}$, $\mathbf{y}$ *and* $\mathbf{z}$ *correspondingly, are reduced to simpler forms with*

$$E_{ss} = E[\mathbf{s}\mathbf{s}^T], \quad E_{sy} = E[\mathbf{s}\mathbf{y}^T], \quad E_{yy} = E[\mathbf{y}\mathbf{y}^T],$$
$$E_{zy} = E[\mathbf{z}\mathbf{y}^T], \quad E_{sz} = E[\mathbf{s}\mathbf{z}^T] \quad \textit{and} \quad E_{zz} = E[\mathbf{z}\mathbf{z}^T].$$

5.4.4 Simulations

In our simulations, the performance of the proposed approach is illustrated with an example of its application to image processing.

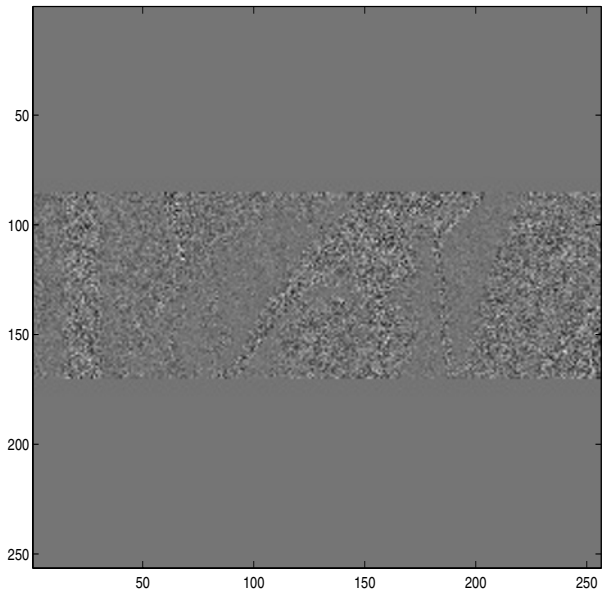
We suppose that the signals to be transformed by a system are given by digitized images presented by matrices. A column of the matrix can be considered as a realization of a stochastic signal.

The known digitized images (see sub-figures (a) and (b) in Fig. 5.1) have been chosen to represent the sets of input and output signals. We denote by

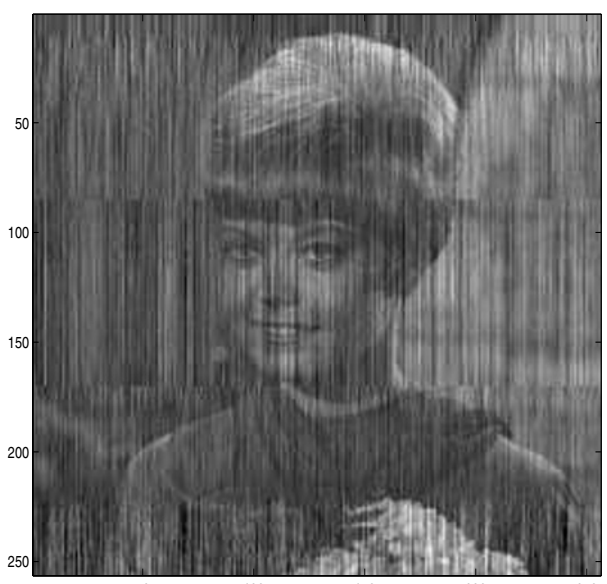
(a) Given "idealized" input.

(b) Desired output.

Figure 5.1: Illustration to the performance of the proposed method. These digitized images have been taken from http://sipi.usc.edu/database/.



(a) Observed input.



(b) Output of the model A^0 constructed from (5.65), (5.81)–(5.83.

Figure 5.2: Illustration to the performance of the proposed method.

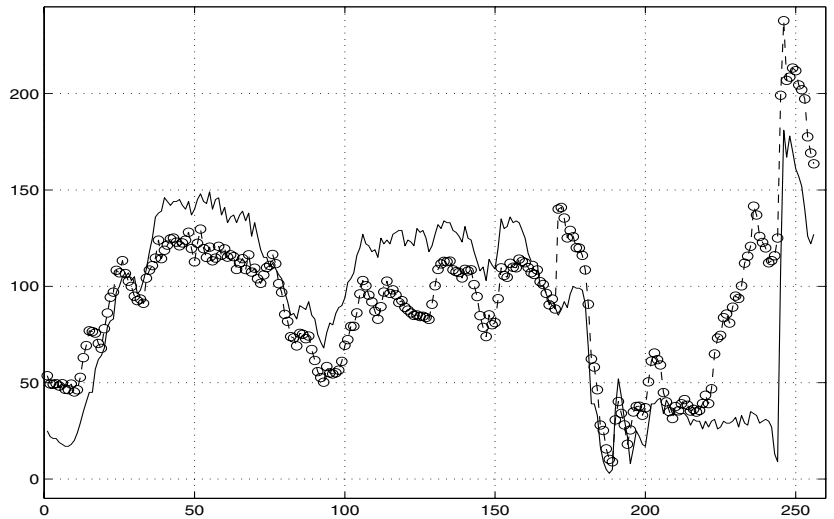


Figure 5.3: Approximation (dashed line with circles) of the 205-th column (solid line) in matrix V by (5.65), (5.81)–(5.83).

$U \in \mathbb{R}^{256\times 256}$ and $V \in \mathbb{R}^{256\times 256}$ matrices which are numerical counterparts of the images in the sub-figures (a) and (b) in Fig. 5.1, respectively.

To illustrate the performance of the presented technique for different types of signals, the matrices U and V have been partitioned into sub-matrices

$$U^{(1)} = U(1:85,:), \quad U^{(2)} = U(86:170,:), \quad U^{(3)} = U(171:256,:)$$

and

$$V^{(1)} = V(1:85,:), \quad V^{(2)} = V(86:170,:), \quad V^{(3)} = V(171:256,:),$$

where $U(n_1 : n_2,:)$ is a matrix formed by $n_2 - n_1 + 1$ consequent rows of U beginning with the n_1-th row. Then each sub-matrix $U^{(i)}$ has been distorted to the matrix $W^{(i)}$ so that

$$W^{(1)} = R_1 .* \cos(U^{(i)}), \quad W^{(2)} = 100R_2 .* U^{(2)} \quad \text{and} \quad W^{(3)} = \sin(U^{(3)}),$$

where R_1 is a matrix with uniformly distributed entries in the interval $(0, 1)$, R_2 is a matrix with normally distributed entries with mean 0 and

variance 1, $\cos(U^{(i)}) = \cos(u_{kj}^{(i)})$, $\sin(U^{(i)}) = \sin(u_{kj}^{(i)})$, $u_{kj}^{(i)}$ is the entry of $U^{(i)}$, and the symbol $.*$ denotes the Hadamard product of matrices.

The proposed method has been applied to each pair $W^{(i)}$, $V^{(i)}$ separately to find the best approximation A^0 to the operator $F = F_i$ where $F_i : U^{(i)} \to V^{(i)}$ for each $i = 1, 2, 3$. The approximator A^0 has been constructed from (5.65), (5.81)–(5.83), (5.97). The input of the system A^0 is a column of the matrix $U^{(i)}$ and the input of the approximating system is a column of the matrix $W^{(i)}$. Covariance matrices have been estimated by the known sample estimates formed by the matrices $W^{(i)}$, $V^{(i)}$ and $Z^{(i)} = V^{(i)}. * V^{(i)}$ for each $i = 1, 2, 3$. For example, for $i = 1$ we calculated E_{sy} as $V^{(1)}W^{(1)T}/85$ etc. The estimation techniques have been discussed in Section 4.3. The simple estimation method used in our simulations has been chosen for the purpose of illustration only.

Sub-figures (b) in Fig. 5.1 and (a) in Fig. 5.2, respectively, have been created from the matrices $[V^{(1)^T}\ V^{(2)^T}\ V^{(3)^T}]^T$ and $[W^{(1)^T}\ W^{(2)^T}\ W^{(3)^T}]^T$ correspondingly.

Sub-figure (b) in Fig. 5.2 is a digitized image created from matrix $[V_1^{(1)^T}\ V_1^{(2)^T}\ V_1^{(3)^T}]^T$ obtained by the proposed method. To illustrate the same performance of the method in a more conspicuous way, we represent in Fig. 5.3 a plot of the 205-th column in the matrix V and plots of the 205-th column in matrix $[V_1^{(1)^T}\ V_1^{(2)^T}\ V_1^{(3)^T}]^T$.

5.5 Best r-Degree Polynomial Approximation

While an advantage of the method considered in Section 5.4 is its computational simplicity, the method proposed here provides a significantly better performance than the technique considered in Section 5.4 and is based on a broad generalization of the procedure presented in Section 5.4. The results of this section supplement the general system theory which has been developed in fundamental works by Volterra, Wiener, Sandberg, and in works by many other authors. The relevant bibliography can be found, for example in [117]–[131].

The proposed approach is based on the approximant produced by a polynomial operator of the rth degree for some natural number $r \in \mathbb{N}$. The approximant minimizes the mean squared error between a desired image and the image of the approximating operator. Standard software packages can easily be used to implement the method.

The statement of the problem is given in Section 5.5.1. Some auxiliary results are presented in Section 5.5.3. In Section 5.5.4 we provide a solution to the problem and prove a theorem on the error associated with

the solution. Some methods for matrix equations solution are considered in Sections 5.5.5 and 5.5.5. Numerical simulations with digitized images in Section 5.5.6 demonstrate the clear advantages of the presented method over the method used in Section 5.4.

5.5.1 Problem formulation

Let $r \in \mathbb{N}$ and let $P_r : \mathbb{R}^n \to \mathbb{R}^m$ be an operator with associated input-output map $\mathcal{P}_r : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^m)$ such that

$$[\mathcal{P}_r(\mathbf{y})](\omega) = P_r[\mathbf{y}(\omega)],$$

and let P_r be given by

$$P_r(y) = A^{(0)} + \sum_{q=1}^{r} A^{(q)}(y^q). \tag{5.103}$$

Here $A^{(0)} \in \mathbb{R}^m$, $r \leq n$, $A^{(q)} : (\mathbb{R}^n)^q \to \mathbb{R}^m$ is a $q-$linear operator, $y^q = (\underbrace{y, \ldots, y}_{q}) \in (\mathbb{R}^n)^q$ and $(\mathbb{R}^n)^q$ is the qth degree of $\mathbb{R}^n$.

The operator P_r is completely defined by $A^{(q)}$ with $q = 0, 1, \ldots, r$. We call P_r a generalized polynomial operator of the rth degree.

Let $\mathcal{F} : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^m)$. The problem is to find $\tilde{A}^{(q)}$ with $q = 0, 1, \ldots, r$ such that

$$J(\tilde{A}^{(0)}, \tilde{A}^{(1)}, \ldots, \tilde{A}^{(r)}) = \min_{A^{(0)}, A^{(1)}, \ldots, A^{(r)}} J(A^{(0)}, A^{(1)}, \ldots, A^{(r)}), \tag{5.104}$$

where

$$J(A^{(0)}, A^{(1)}, \ldots, A^{(r)}) = E[\|\mathcal{F}(\mathbf{x}) - \mathcal{P}_r(\mathbf{y})\|^2]. \tag{5.105}$$

In other words, we wish to find the best mathematical model $\tilde{P}_r$ (defined by $\tilde{A}^{(0)}$, $\tilde{A}^{(1)}$, $\ldots$, $\tilde{A}^{(r)}$) of the system F in the class of generalized polynomial operators (5.103).

The problem considered in Section 5.4 is a particular case of (5.103), (5.104) if $r = 2$,

$$P_2(y) = A^{(0)} + A^{(1)}y + A^{(2)}\tilde{y}$$

and $\tilde{y}$ defined by the Hadamard product so that $\tilde{y} = (y_1^2, \ldots, y_n^2)^T$ where $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$.

5.5.2 Approximation with any given accuracy

The problem above motivates the following question: Is there exists an operator $\mathcal{P}_r$ which approximates $\mathcal{F}$ with any given accuracy? A natural conjecture is that the positive answer can follow from the results presented in Chapter 1. Below, we show that this is indeed true.

Theorem 36. *Let K_Y be a compact set of signals in the space $L^2(\Omega, \mathbb{R}^n)$, $K_X = \mathcal{F}(K_Y)$ and $\mathcal{F}$ defined as above. Then for any $\mathbf{x} \in K_X$ and $\varepsilon > 0$ there exists $\mathcal{P}_r$ such that*

$$E[\|\mathbf{x} - \mathcal{P}_r(\mathbf{y})\|^2] \leq \varepsilon \tag{5.106}$$

for all $\mathbf{x} \in K_X$ and $\mathbf{y} \in K_Y$.

Proof. For any $\mathbf{g} \in L^2(\Omega, \mathbb{R}^m)$, $(E[\|\mathbf{g}\|^2])^{1/2}$ defined by (5.67) is the norm $\|\mathbf{g}\|_{L^2}$ in $L^2(\Omega, \mathbb{R}^m)$ so that

$$(E[\|\mathbf{g}\|^2])^{1/2} = \|\mathbf{g}\|_{L^2}.$$

Then

$$\begin{aligned} E[\|\mathbf{x} - \mathcal{P}_r(\mathbf{y})\|^2] &= \|\mathbf{x} - \mathcal{P}_r(\mathbf{y})\|_{L^2}^2 \\ &= \|\mathcal{F}(\mathbf{y}) - \mathcal{P}_r(\mathbf{y})\|_{L^2}^2 \end{aligned}$$

and the statement of the theorem follows directly from the references [104] or [65]. □

We note that compactness of the sets K_Y and K_X is an essential condition of this result. Theorem 36 is important because it constitutes the existence of $\mathcal{P}_r$ which estimates $\mathbf{x}$ with any desirable accuracy.

Let us now show that $\mathcal{P}_r$ can be determined in the form which guarantees a smallest associated error among all $\mathcal{P}_r$ of the same degree r.

5.5.3 Reduction of $P_r(y)$ to a matrix form representation

For our purposes, it is convenient to use a representation of the operator P_r in matrix terms. The following lemma establishes this representation.

Lemma 27. *There exist matrices $A^{(q)}_{j_1,\ldots,j_{q-1}} \in \mathbb{R}^{m\times n}$ such that*

$$P_r(y) = A^{(0)} + A^{(1)}y + \sum_{q=2}^{r}\sum_{j_1=1}^{n}\cdots\sum_{j_{q-1}=1}^{n} y_{j_1}\ldots y_{j_{q-1}}A^{(q)}_{j_1,\ldots,j_{q-1}}y \tag{5.107}$$

$$= A^{(0)} + A^{(1)}y + \sum_{q=2}^{r}\sum_{q-1\leq\sigma_{q-1}\leq(q-1)n} y_{j_1}\ldots y_{j_{q-1}}A^{(q)}_{j_1,\ldots,j_{q-1}}y \tag{5.108}$$

where $\sigma_{q-1} = j_1 + \ldots + j_{q-1}$ *and the inner sum is extended for all summands with subscripts satisfying the inequality* $q - 1 \le \sigma_{q-1} \le (q-1)n$.

Proof. Let $\{e_1, \ldots, e_n\}$ be the standard basis in $\mathbb{R}^n$. Then

$$\begin{aligned} A^{(q)}(y^q) &= A^{(q)}(\sum_{j_1=1}^{n} y_{j_1} e_{j_1}, \ldots, \sum_{j_q=1}^{n} y_{j_q} e_{j_q}) \\ &= \sum_{j_1=1}^{n} \ldots \sum_{j_q=1}^{n} A^{(q)}(e_{j_1}, \ldots, e_{j_q}) y_{j_1} \ldots, y_{j_q}. \end{aligned}$$

If we write

$$A^{(q)}(e_{j_1}, \ldots, e_{j_q}) = a^{(q)}_{j_1,\ldots,j_q} \in \mathbb{R}^m$$

then we can define the matrix $A^{(q)}_{j_1,\ldots,j_{q-1}} \in \mathbb{R}^{m\times n}$ by the formula

$$A^{(q)}_{j_1,\ldots,j_{q-1}} v = \sum_{j_q=1}^{n} a^{(q)}_{j_1,\ldots,j_q} v_{j_q},$$

where $v = (v_1, \ldots v_n)^T \in \mathbb{R}^n$. Therefore

$$\begin{aligned} A^{(q)}(y^q) &= \sum_{j_1=1}^{n} \ldots \sum_{j_{q-1}=1}^{n} y_{j_1} \ldots y_{j_{q-1}} (\sum_{j_q=1}^{n} a^{(q)}_{j_1,\ldots,j_q} y_{j_q}) \\ &= \sum_{q-1\le\sigma_{q-1}\le(q-1)n} y_{j_1} \ldots y_{j_{q-1}} A^{(q)}_{j_1,\ldots,j_{q-1}} y \end{aligned} \tag{5.109}$$

and the lemma is proved. □

Example 17. *For* $r = 2$, *the formula* (5.107) *takes the form*

$$P_2(y) = A^{(0)} + A^{(1)} y + \sum_{j_1=1}^{n} y_{j_1} A^{(2)}_{j_1} y, \tag{5.110}$$

where

$$A^{(2)}_1 = \{a^{(2)}_{i,1,j_2}\}^{i=m,j_2=n}_{i,j_2=1} \in \mathbb{R}^{m\times n}, \ldots, A^{(2)}_n = \{a^{(2)}_{i,n,j_2}\}^{i=m,j_2=n}_{i,j_2=1} \in \mathbb{R}^{m\times n}$$

or

$$A^{(2)}_1 = \{a^{(2)}_{i,j_1,1}\}^{i=m,j_1=n}_{i,j_1=1} \in \mathbb{R}^{m\times n}, \ldots, A^{(2)}_n = \{a^{(2)}_{i,j_1,n}\}^{i=m,j_1=n}_{i,j_1=1} \in \mathbb{R}^{m\times n}$$

and $a^{(2)}_{i,j_1,j_2}$ *are entries of the* $m \times n \times n$ *tensor* $A^{(2)}$.

Let us now reduce the expression (5.107) to a more compact form which is similar to (5.110) but will be given for $r > 2$.

First, we observe that the products $y_{j_1} \dots y_{j_{q-1}}$ are not ordered in (5.107). Therefore, the model (5.107) contains the same factors $y_{j_1} \dots y_{j_{q-1}}$ for different matrices $A^{(q)}_{j_1,\dots,j_{q-1}}$. For example, for $r = 3$, $P_r(y)$ contains products $y_1 y_2$ and $y_2 y_1$ which are the same. We call this issue *a symmetry effect*.

This circumstance may lead to an unnecessary increase in the computation load when determining the optimal estimator. To avoid this inconvenience, we collect together all terms with the same factors $y_{j_1} \dots y_{j_{q-1}}$ and write (5.107) in the form

$$P_r(y) = A^{(0)} + A^{(1)}y + \sum_{q=2}^{r} \sum_{\substack{i_1+\dots+i_n=q-1 \\ i_1,\dots,i_n=0,1,\dots,q-1}} y_1^{i_1} \dots y_n^{i_n} A_{(q),i_1,\dots,i_n} y \quad (5.111)$$

$$= A^{(0)} + A^{(1)}y + \sum_{\substack{i_1+\dots+i_n\le r-1 \\ i_1,\dots,i_n=0,1,\dots,r-1}} y_1^{i_1} \dots y_n^{i_n} A_{i_1,\dots,i_n} y. \quad (5.112)$$

The inner sum in (5.111) and the sum in (5.112) are extended for terms such that $i_1 + \dots + i_n = q - 1$ and $i_1 + \dots + i_n \le r - 1$, respectively, where $i_1, \dots, i_n = 0, 1, \dots, q - 1$ and at least one i_j is not zero for $j = 1, \dots, n$. Each product $y_1^{i_1} \dots y_n^{i_n}$ is written in ascending order with respect to subscripts. Also, $A_{(q),j_1,\dots,j_{q-1}} \in \mathbb{R}^{n\times n}$ follows from collecting similar terms in (5.107) and $A_{j_1,\dots,j_{q-1}}$ represent $A_{(q),j_1,\dots,j_{q-1}}$ which are rewritten with different subscripts after representing the double sum (5.111) in the form given by (5.112).

Thus, (5.111)–(5.112) has no similar terms $y_1^{i_1} \dots y_n^{i_n}$. This circumstance allows us to avoid the symmetry effect mentioned above. Such an rearrangement leads to a smaller matrix $\mathcal{D}$ in Theorem 37 below.

The expression $P_r(y)$ given by (5.107) contains

$$1 + n + n^2 + \dots + n^r = \frac{n^{r+1} - 1}{n - 1}$$

terms. The number of terms in $P_r(y)$ by (5.112) is reduced to

$$1 + \left[\binom{n-1}{0} + \binom{n}{1} + \dots + \binom{n+r-2}{r-1} \right] \cdot n \quad (5.113)$$

$$= 1 + \binom{n+r-1}{r-1} \cdot n. \quad (5.114)$$

As a result, computational work needed for computation of $\mathcal{D}^\dagger$ in Theorem 37 is smaller than that for the matrix $\mathcal{D}^\dagger$ without suppressing the symmetry effect.

By (5.113), the expression (5.112) contains

$$N = \left(\begin{array}{c} n+r-1 \\ r-1 \end{array} \right) \cdot n \tag{5.115}$$

matrices $A_{i_1,\ldots,i_n}$. If we denote them by $C_1, \ldots, C_N$ with the corresponding operands denoted by $u_1, \ldots, u_N \in \mathbb{R}^n$ then we can write

$$\sum_{\substack{i_1+\ldots+i_n \leq r-1 \\ i_1,\ldots,i_n=0,1,\ldots,r-1}} y_1^{i_1} \ldots y_n^{i_n} A_{i_1,\ldots,i_n} y = \sum_{j=1}^{N} C_j u_j. \tag{5.116}$$

Therefore the following is true.

Corollary 5. *The polynomial operators P_r and $\mathcal{P}_r$ can respectively be written as*

$$P_r(\mathbf{y}) = A^{(0)} + A^{(1)} y + \sum_{j=1}^{N} C_j u_j \tag{5.117}$$

and

$$\mathcal{P}_r(\mathbf{y}) = \mathcal{A}^{(0)} + \mathcal{A}^{(1)}(\mathbf{y}) + \sum_{j=1}^{N} \mathcal{C}_j(\mathbf{u}_j), \tag{5.118}$$

where $\mathcal{C}_j : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^m)$ are defined by matrices C_j in the manner of (5.44).

The operator P_r is represented diagrammatically in Fig. 5.4.

For any $\mathbf{x} \in L^2(\Omega, \mathbb{R}^n)$ and $\mathbf{v} \in L^2(\Omega, \mathbb{R}^m)$, we denote

$$\mathbf{v} = \mathcal{F}(\mathbf{x}).$$

The results in the next Lemmas will be used in the proof of Theorems 37 and 38 in the next Section.

Lemma 28. *The following equations hold:*

$$E_{vy} E_{yy}^{\dagger} E_{yy} = E_{vy}, \; E_{u_j y} E_{yy}^{\dagger} E_{yy} = E_{u_j y} \quad \textit{and} \quad E_{v u_j} E_{u_j u_j}^{\dagger} E_{u_j u_j} = E_{v u_j}.$$

Proof. The proof follows from the proof of Lemma 24 above. □

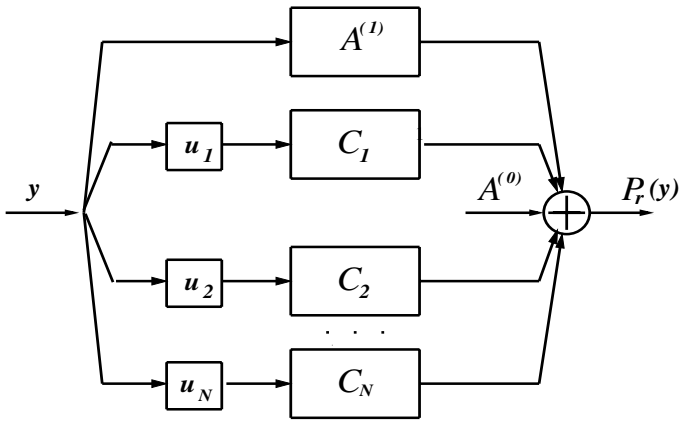


Figure 5.4: Representation of the operator P_r.

Lemma 29. *Let*

$$D_{ij} = E_{u_i u_j} - E_{u_i y} E_{yy}^{\dagger} E_{y u_j}$$

and

$$G_j = E_{v u_j} - E_{v y} E_{yy}^{\dagger} E_{y u_j},$$

so that $D_{ij} \in \mathbb{R}^{n \times n}$ *and* $G_j \in \mathbb{R}^{m \times n}$. *Then*

$$G_j D_{ij}^{\dagger} D_{ij} = G_j \tag{5.119}$$

for all $i, j = 1, 2, \ldots N$.

Proof. The proof follows from the proof of Lemma 25 in Section 4.4.1(b). □

Lemma 30. *Let* $D = \begin{bmatrix} D_{11} & \ldots & D_{1N} \\ D_{21} & \ldots & D_{2N} \\ \ldots & \ldots & \ldots \\ D_{N1} & \ldots & D_{NN} \end{bmatrix}$ *and* $G = [G_1\ G_2\ \ldots\ G_N]$.

Then

$$G D^{\dagger} D = G. \tag{5.120}$$

Proof. We observe that

$$D = E_{uu} - E_{uy} E_{yy}^{\dagger} E_{yu}$$

and

$$G = E_{vu} - E_{vy}E_{yy}^{\dagger}E_{yu},$$

where $\mathbf{u} = (\mathbf{u}_1^T, \dots \mathbf{u}_N^T)^T$. Then the proof follows from Lemma 25 in Section 4.4.1(b). □

5.5.4 Determination of $\tilde{A}^{(0)}$, $\tilde{A}^{(1)}$ and $\tilde{C}_1$, $\tilde{C}_2$, ..., $\tilde{C}_N$

In accordance with Corollary 5, the problem formulated in Section 4.4.2(a) is reduced to finding $\tilde{A}^{(0)} \in \mathbb{R}^m$, $\tilde{A}^{(1)} \in \mathbb{R}^{m\times n}$ and $\tilde{C}_j \in \mathbb{R}^{m\times n}$ with $j = 1, 2, \dots, N$ which minimize the functional (5.105), where $\mathcal{P}_r(\mathbf{y})$ is in the form (5.118) and where

$$J(\tilde{A}^{(0)}, \tilde{A}^{(1)}, \tilde{C}_1, \dots, \tilde{C}_N) = \min_{A^{(0)}, A^{(1)}, C_1, \dots, C_N} J(A^{(0)}, A^{(1)}, C_1, \dots, C_N) \quad (5.121)$$

with

$$J(A^{(0)}, A^{(1)}, C_1, \dots, C_N) = E[\|\mathcal{F}(\mathbf{x}) - (\mathcal{A}^{(0)} + \mathcal{A}^{(1)}(\mathbf{y}) + \sum_{j=1}^{N} \mathcal{C}_j(\mathbf{u}_j)\|^2]. \quad (5.122)$$

Theorem 37. *The solution to problem specified by* (5.121) *and* (5.122) *is given by*

$$\tilde{A}^{(0)} = E[\mathbf{v}] - \tilde{A}^{(1)}E[\mathbf{y}] - \sum_{j=1}^{N} \tilde{C}_j E[\mathbf{u}_j], \quad (5.123)$$

$$\tilde{A}^{(1)} = (E_{vy} - \sum_{j=1}^{N} \tilde{C}_j E_{u_j y})E_{yy}^{\dagger} + K_1(I - E_{yy}E_{yy}^{\dagger}) \quad (5.124)$$

and

$$[\tilde{C}_1\ \tilde{C}_2\ \dots\ \tilde{C}_N] = GD^{\dagger} + K_2(I - DD^{\dagger}), \quad (5.125)$$

where K_1 and K_2 are arbitrary matrices.

Proof. Note that each matrix $\tilde{C}_j \in \mathbb{R}^{m\times n}$ in (5.123) - (5.125) is defined as a corresponding $m \times n$ submatrix of matrix $GD^{\dagger} + K_2(I - DD^{\dagger})$.

We write

$$D^{\dagger} = \begin{bmatrix} Q_{11} & \dots & Q_{1N} \\ Q_{21} & \dots & Q_{2N} \\ \dots & \dots & \dots \\ Q_{N1} & \dots & Q_{NN} \end{bmatrix},$$

where $Q_{ij} \in \mathbb{R}^{n\times n}$ for $i,j = 1,\dots,N$.

Let us show that $J(A^{(0)}, A^{(1)}, C_1, \dots, C_N)$, given by (5.122), can be represented as

$$\begin{aligned} J(A^{(0)}, A^{(1)}, C_1, \dots, C_N) &= J_0 + J_1(A^{(0)}, A^{(1)}, C_1, \dots, C_N) \\ &\quad + J_2(A^{(1)}, C_1, \dots, C_N) + J_3(C_1, \dots, C_N), \end{aligned} \tag{5.126}$$

where

$$\begin{aligned} J_0 &= \operatorname{tr}\{E_{vv}\} - \|E_{vy}(E_{yy}^{\dagger})^{1/2}\|^2 - \sum_{i=1}^{N} \|G_i Q_{ii}^{1/2}\|^2 \\ &\quad - \sum_{j,k=1,\dots,N,\ j\neq k} \operatorname{tr}\{G_j Q_{jk} G_k^T\}, \end{aligned}$$

$$\begin{aligned} &J_1(A^{(0)}, A^{(1)}, C_1, \dots, C_N) \\ &\quad = \|A^{(0)} - E[\mathbf{v}] + A^{(1)} E[\mathbf{y}] + \sum_{j=1}^{N} C_j E[\mathbf{u}_j]\|^2, \end{aligned} \tag{5.127}$$

$$J_2(A^{(1)}, C_1, \dots, C_N) = \|[A^{(1)} - (E_{vy} - \sum_{j=1}^{N} C_j E_{u_j y}) E_{yy}^{\dagger}] E_{yy}^{1/2}\|^2 \tag{5.128}$$

and

$$J_3(C_1, \dots, C_N) = \|([C_1, \dots, C_N] - GD^{\dagger}) D^{1/2}\|^2. \tag{5.129}$$

We have

$$J_0 = \operatorname{tr}\{E_{vv} - E_{vy} E_{yy}^{\dagger} E_{yv} - \sum_{j,k=1,\dots,N} G_j Q_{jk} G_k^T\} \tag{5.130}$$

and

$$\begin{aligned} &J_1(A^{(0)}, A^{(1)}, C_1, \dots, C_N) \\ &\quad = \operatorname{tr}\{A^{(0)} A^{(0)T} - A^{(0)} E[\mathbf{v}^T] + A^{(0)} E[\mathbf{y}^T] A^{(1)} + A^{(0)} \sum_{k=1}^{N} E[\mathbf{u}_k^T] C_k^T \\ &\quad - E[\mathbf{v}] A^{(0)T} + E[v] E[\mathbf{v}^T] - E[v] E[\mathbf{y}^T] A^{(1)T} + (A^{(1)} E[\mathbf{y}] - E[\mathbf{v}]) \\ &\quad \times \sum_{k=1}^{N} E[\mathbf{u}_k^T] C_k^T + A^{(1)} E[\mathbf{y}] A^{(0)T} - A^{(1)} E[\mathbf{y}] E[\mathbf{v}^T] \\ &\quad + A^{(1)} E[\mathbf{y}] E[\mathbf{y}^T] A^{(1)T} + \sum_{k=1}^{N} C_k E[\mathbf{u}_k] (A^{(0)T} - E[\mathbf{v}^T] \end{aligned} \tag{5.131}$$

$$+E[y^T]A^{(1)T}) + \sum_{j,k=1,\ldots,N} C_j E[\mathbf{u}_j]E[\mathbf{u}_k^T]C_k^T\},$$

and

$$\begin{aligned} &J_2(A^{(1)}, C_1, \ldots, C_N) + J_3(C_1, \ldots, C_N) \\ &\quad = \mathrm{tr}\{A^{(1)}E_{yy}A^{(1)T} - A^{(1)}(E_{yv} - \sum_{j=1}^{N} E_{yu_j}C_j^T) - E_{vy}A^{(1)T} \\ &\qquad + \sum_{j=1}^{N} C_j E_{u_j y}A^{(1)T} + E_{vy}E_{yy}^{\dagger}E_{yv} + \sum_{j,k=1,\ldots,N} C_j E_{u_j u_k}C_k^T \\ &\qquad - \sum_{j=1}^{N} C_j E_{u_j v} - \sum_{j=1}^{N} E_{vu_j}C_j^T + \sum_{j,k=1,\ldots,N} G_j Q_{jk} G_k^T\}. \end{aligned} \tag{5.132}$$

Then, on the basis of Lemmata 28 and 29, and combining (5.130) - (5.132), we obtain

$$\begin{aligned} &J_0 + J_1(A^{(0)}, A^{(1)}, C_1, \ldots, C_N) + J_2(A^{(1)}, C_1, \ldots, C_N) + J_3(C_1, \ldots, C_N) \\ &\quad = \mathrm{tr}\{E[\mathbf{v}\mathbf{v}^T] - E[\mathbf{v}]A^{(0)T} - E[\mathbf{v}\mathbf{y}^T]A^{(1)T} - \sum_{k=1}^{N} E[\mathbf{v}\mathbf{u}_k^T]C_k^T \\ &\qquad - A^{(0)}E[\mathbf{v}^T] + A^{(0)}A^{(0)T} + A^{(0)}E[\mathbf{y}^T]A^{(1)T} \\ &\qquad + A^{(0)}\sum_{k=1}^{N} E[\mathbf{u}_k^T]C_k^T - A^{(1)}E[\mathbf{y}\mathbf{v}^T] + A^{(1)}E[\mathbf{y}]A^{(0)T} \\ &\qquad + A^{(1)}E[\mathbf{y}\mathbf{y}^T]A^{(1)T} + A^{(1)}\sum_{k=1}^{N} E[\mathbf{y}\mathbf{u}_k^T]C_k^T - \sum_{k=1}^{N} C_k E[\mathbf{u}_k\mathbf{v}^T] \\ &\qquad + \sum_{k=1}^{N} C_k E[\mathbf{u}_k]A^{(0)T} + \sum_{k=1}^{N} C_k E[\mathbf{u}_k\mathbf{y}^T]A^{(1)T} \\ &\qquad + \sum_{j,k=1,\ldots,N} C_j E[\mathbf{u}_j\mathbf{u}_k^T]C_k^T\} \\ &\qquad = J(A^{(0)}, A^{(1)}, C_1, \ldots, C_N). \end{aligned}$$

Thus (5.126) is true.

It follows from (5.126) - (5.129) that $J(A^{(0)}, A^{(1)}, C_1, \ldots, C_N)$ is minimized when

$$A^{(0)} = \tilde{A}^{(0)},$$

$$A^{(1)}E_{yy}^{1/2} = (E_{vy} - \sum_{j=1}^{N} C_j E_{u_j y})E_{yy}^{\dagger}E_{yy}^{1/2}, \tag{5.133}$$

$$[C_1 C_2 \dots C_n]D^{1/2} = GD^{\dagger}D^{1/2}. \tag{5.134}$$

Similarly to Lemma 26, it can be shown that the equations (5.133) and (5.134) are respectively equivalent to the equations

$$A^{(1)}E_{yy} = E_{vy} - \sum_{j=1}^{N} C_j E_{u_j y} \tag{5.135}$$

and

$$[C_1 C_2 \dots C_n]D = GD. \tag{5.136}$$

The necessary and sufficient conditions [6] for the equations (5.135) and (5.136) to have solutions are

$$E_{vy} - \sum_{j=1}^{N} C_j E_{u_j y} = (E_{vy} - \sum_{j=1}^{N} C_j E_{u_j y})E_{yy}^{\dagger}E_{yy}$$

and

$$G = GD^{\dagger}D,$$

respectively. They are satisfied on the basis of Lemmata 28 and 29. Therefore, as it follows from [6], the solutions to the equations (5.135) and (5.136) are given by (5.124) and (5.125), respectively. □

Note that a possible and natural choice for K_1 and K_2 in the expressions (5.124) and (5.125) is $K_1 = \mathbb{O}$ and $K_2 = \mathbb{O}$ where $\mathbb{O}$ is the zero matrix.

Also note that the best polynomial operator $\tilde{P}_r$ defined by the equations (5.123), (5.124) and (5.125) requires knowledge of the matrices $E[\mathbf{v}]$, $E[\mathbf{y}]$, $E[\mathbf{u}_k]$, E_{vy}, E_{vu_k}, E_{yy}, E_{yu_k} and $E_{u_j u_k}$. Methods for estimation of these matrices have been considered in Section 4.3.

Theorem 38. *The error of approximation by the best polynomial operator* $\tilde{\mathcal{P}}_r$ *defined by* (5.123)–(5.125) *is*

$$\begin{aligned} E[\|\mathcal{F}(\mathbf{x}) - \tilde{\mathcal{P}}_r(\mathbf{y})\|^2] \\ = \operatorname{tr}\{E_{vv}\} - \|E_{vy}(E_{yy}^{\dagger})^{1/2}\|^2 \\ - \sum_{i=1}^{N} \|G_i Q_{ii}^{1/2}\|^2 - \sum_{j,k=1,\dots,N,\ j\neq k} \operatorname{tr}\{G_j Q_{jk} G_k^T\}. \end{aligned} \tag{5.137}$$

Proof. The proof follows directly from equations (5.123) - (5.129). □

Comparison of (5.137) with the error associated with the method of Section 5.4.3 (see Theorem 35 in Section 5.4.3) demonstrates the clear advantage of the proposed method over the procedure considered in Section 4.4.1.

5.5.5 Towards methods for matrix equation solution

In the proof of Theorem 37, we show that $\tilde{A}^{(0)}$, $\tilde{A}^{(1)}$, $\tilde{C}_1$, $\tilde{C}_2$, ,..., $\tilde{C}_N$ can be found by solving the equation (5.133) and (5.134). Their solutions are based on a solution of the matrix equation

$$X\mathbb{E}_{ff} = \mathbb{E}_{gf} \tag{5.138}$$

where $\mathbf{f} : \Omega \to \mathbb{R}^n$ and $\mathbf{g} : \Omega \to \mathbb{R}^k$ are random vectors and $X \in \mathbb{R}^{k\times n}$. Since the null space of $\mathbb{E}_{ff}$ is a subset of the null space of $\mathbb{E}_{gf}$ it follows from [6] that

$$X = \mathbb{E}_{gf}\mathbb{E}_{ff}{}^{\dagger} + F(I - \mathbb{E}_{ff}\mathbb{E}_{ff}{}^{\dagger}) \tag{5.139}$$

where $F \in \mathbb{R}^{k\times n}$ is arbitrary.

In practice k and n may be large and consequently (5.138) becomes a large system. For large systems, the generalized inverse may be difficult to compute. One might reasonably expect to facilitate the solution using the idea of Gaussian elimination [50] with full pivoting or some suitable variant.

Here, we exploit the special structure of the original system (5.138) to propose a new conceptual block elimination procedure that separates the original system into two independent smaller subsystems each with the same general form. This provides the basis for an efficient solution algorithm that will be described in the next subsection.

Let

$$f = \begin{bmatrix} p \\ q \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} Y & Z \end{bmatrix}$$

where $\mathbf{p} : \Omega \to \mathbb{R}^{n_1}$ and $\mathbf{q} : \Omega \to \mathbb{R}^{n_2}$ with $n_1 + n_2 = n$ are random vectors and where $Y \in \mathbb{R}^{k\times n_1}$ and $Z \in \mathbb{R}^{k\times n_2}$. Write $\mathbb{E}_{ff}$ and $\mathbb{E}_{gf}$ in partitioned form as

$$\mathbb{E}_{ff} = \begin{bmatrix} \mathbb{E}_{pp} & \mathbb{E}_{pq} \\ \mathbb{E}_{qp} & \mathbb{E}_{qq} \end{bmatrix} \quad \text{and} \quad \mathbb{E}_{gf} = \begin{bmatrix} \mathbb{E}_{gp} & \mathbb{E}_{gq} \end{bmatrix}$$

and hence rewrite the original equation (5.138) as

$$\begin{bmatrix} Y & Z \end{bmatrix} \begin{bmatrix} \mathbb{E}_{pp} & \mathbb{E}_{pq} \\ \mathbb{E}_{qp} & \mathbb{E}_{qq} \end{bmatrix} = \begin{bmatrix} \mathbb{E}_{gp} & \mathbb{E}_{gq} \end{bmatrix}.$$

The following lemma and theorem are the key to the new elimination procedure.

Lemma 31. *Let $n \in \mathbb{N}$ and $\mathbf{p},\ \mathbf{q} \in L^2(\Omega, \mathbb{R}^n)$. Define $\mathcal{R} = \mathbf{q} - \mathbb{E}_{qp}\mathbb{E}_{pp}^{\dagger}\mathbf{p}$. Then*

$$\mathbb{E}_{rr} = \mathbb{E}_{qq} - \mathbb{E}_{qp}\mathbb{E}_{pp}^{\dagger}\mathbb{E}_{pq}.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}_{rr} &= E[(\hat{\mathbf{q}} - \mathbb{E}_{qp}\mathbb{E}_{pp}^{\dagger}\hat{\mathbf{p}})(\hat{\mathbf{q}} - \mathbb{E}_{qp}\mathbb{E}_{pp}^{\dagger}\hat{\mathbf{p}})^T] \\
&= E[\hat{\mathbf{q}}\hat{\mathbf{q}}^T] - \mathbb{E}_{qp}\mathbb{E}_{pp}^{\dagger}E[\hat{\mathbf{p}}\hat{\mathbf{q}}^T] - E[\hat{\mathbf{q}}\hat{\mathbf{p}}^T]\mathbb{E}_{pp}^{\dagger}\mathbb{E}_{pq} \\
&\qquad + \mathbb{E}_{qp}\mathbb{E}_{pp}^{\dagger}E[\hat{\mathbf{p}}\hat{\mathbf{p}}^T]\mathbb{E}_{pp}^{\dagger}\mathbb{E}_{pq} \\
&= \mathbb{E}_{qq} - \mathbb{E}_{qp}\mathbb{E}_{pp}^{\dagger}\mathbb{E}_{pq}
\end{aligned}$$

as required. □

Theorem 39. *Let $\mathbf{p} : \Omega \to \mathbb{R}^{n_1}$, $\mathbf{q} : \Omega \to \mathbb{R}^{n_2}$ and $\mathbf{g} : \Omega \to \mathbb{R}^k$ be random vectors and consider the system of equations*

$$\begin{bmatrix} Y & Z \end{bmatrix} \begin{bmatrix} \mathbb{E}_{pp} & \mathbb{E}_{pq} \\ \mathbb{E}_{qp} & \mathbb{E}_{qq} \end{bmatrix} = \begin{bmatrix} \mathbb{E}_{gp} & \mathbb{E}_{gq} \end{bmatrix}. \tag{5.140}$$

If we define $r = q - \mathbb{E}_{qp}\mathbb{E}_{pp}{}^{\dagger}p$ and $Y^* = Y + Z\mathbb{E}_{qp}\mathbb{E}_{pp}{}^{\dagger}$ then the original system of equations can be rewritten equivalently as two separate systems

$$Y^*\mathbb{E}_{pp} = \mathbb{E}_{gp} \quad \textit{and} \quad Z\mathbb{E}_{rr} = \mathbb{E}_{gr}. \tag{5.141}$$

The solutions to the separated systems are given by

$$Y^* = \mathbb{E}_{gp}\mathbb{E}_{pp}{}^{\dagger} + P(I - \mathbb{E}_{pp}\mathbb{E}_{pp}{}^{\dagger}) \tag{5.142}$$

and

$$Z = \mathbb{E}_{gr}\mathbb{E}_{rr}{}^{\dagger} + R(I - \mathbb{E}_{rr}\mathbb{E}_{rr}{}^{\dagger}). \tag{5.143}$$

Proof. If we make the transformation

$$[Y^*, Z^*] = [Y, Z] \begin{bmatrix} I & \mathbb{O} \\ \mathbb{E}_{qp}\mathbb{E}_{pp}{}^{\dagger} & I \end{bmatrix} \quad \Leftrightarrow \quad [Y, Z] = \begin{bmatrix} I & \mathbb{O} \\ -\mathbb{E}_{qp}\mathbb{E}_{pp}{}^{\dagger} & I \end{bmatrix}$$

where $\mathbb{O}$ is the zero block, then

$$\begin{aligned}
&[Y^*, Z^*] \begin{bmatrix} I & \mathbb{O} \\ -\mathbb{E}_{qp}\mathbb{E}_{pp}{}^{\dagger} & I \end{bmatrix} \begin{bmatrix} \mathbb{E}_{pp} & \mathbb{E}_{pq} \\ \mathbb{E}_{qp} & \mathbb{E}_{qq} \end{bmatrix} \begin{bmatrix} I & -\mathbb{E}_{pp}{}^{\dagger}\mathbb{E}_{pq} \\ \mathbb{O} & I \end{bmatrix} \\
&\qquad = [\mathbb{E}_{gp}, \mathbb{E}_{gq}] \begin{bmatrix} I & -\mathbb{E}_{pp}{}^{\dagger}\mathbb{E}_{pq} \\ \mathbb{O} & I \end{bmatrix}
\end{aligned}$$

and if we define $r = q - \mathbb{E}_{qp}\mathbb{E}_{pp}{}^{\dagger}p$ and use Lemmata A1.2 and A2.1 then the original system reduces to

$$[Y^*, Z^*] \begin{bmatrix} \mathbb{E}_{pp} & \mathbb{O} \\ \mathbb{O} & \mathbb{E}_{rr} \end{bmatrix} = [\mathbb{E}_{gp}, \mathbb{E}_{gq}] .$$

The result is now easily established. □

It is clear that the separated systems each have the same form as the original system. In practice the separation is designed to remove a small system for which the solution can be easily calculated. The remaining system takes exactly the same form as the original and so the separation procedure can be repeated.

A recursive algorithm for solution of the matrix equation

Here, we present a new algorithm that reduces the original system of equations (5.138) to a collection of independent smaller subsystems each with the same general form. Solution of the collection of smaller subsystems requires significantly less computational work and enables greater precision in the calculations. Hence the new algorithm is highly efficient.

Let $\mathbf{p}_i : \Omega \to \mathbb{R}^{n_i}$ be a random vector for each $i = 1, 2, \ldots, r$ and let $n = \sum_{i=1}^{r} n_i$. If we define $\mathbb{E}_{i,j} = E[\mathbf{p}_i\mathbf{p}_j^T] - E[\mathbf{p}_i]E[\mathbf{p}_j^T]$ and $\mathbb{E}_j = E[\mathbf{g}\mathbf{p}_j^T] - E[\mathbf{g}]E[\mathbf{p}_j^T]$ and let $\mathbf{f}^T = [\mathbf{p}_1^T, \mathbf{p}_2^T, \ldots, \mathbf{p}_r^T]$ then the equation $X\mathbb{E}_{ff} = \mathbb{E}_{gf}$ can be rewritten in the partitioned form

$$[X_1, X_2, \ldots, X_r] \begin{bmatrix} \mathbb{E}_{1,1} & \mathbb{E}_{1,2} & \cdots & \mathbb{E}_{1,r} \\ \mathbb{E}_{2,1} & \mathbb{E}_{2,2} & \cdots & \mathbb{E}_{2,r} \\ \vdots & \vdots & & \vdots \\ \mathbb{E}_{r,1} & \mathbb{E}_{r,2} & \cdots & \mathbb{E}_{r,r} \end{bmatrix} = [\mathbb{E}_1, \mathbb{E}_2, \ldots, \mathbb{E}_r] .$$

We use the following algorithm to calculate the solution.

Solution algorithm

1. Set $\ell := 1$.

2. Set $\mathbb{E}_\ell := E[\mathbf{g}\mathbf{p}_\ell^T] - E[\mathbf{g}]E[\mathbf{p}_\ell^T]$

3. For all (i, j) with $\ell \le i \le j \le r$ set $\mathbb{E}_{i,j} := E[\mathbf{p}_i\mathbf{p}_j^T] - E[\mathbf{p}_i]E[\mathbf{p}_j^T]$.

4. For all j with $\ell + 1 \le j \le r$ set $\mathbf{p}_j := \mathbf{p}_j - \mathbb{E}_{j,\ell}\mathbb{E}_{\ell,\ell}{}^{\dagger}p_\ell$.

5. If $\ell < r$ set $\ell := \ell + 1$ and go to Step 2.

6. If $\ell = r$ go to Step 7.

7. Set $X_\ell := \mathbb{E}_\ell \mathbb{E}_{\ell,\ell}^\dagger + P_\ell(I - \mathbb{E}_{\ell,\ell}\mathbb{E}_{\ell,\ell}^\dagger)$ where $P_\ell \in \mathbb{R}^{k\times m_\ell}$ is arbitrary.

8. Set $\ell := \ell - 1$.

9. Set $X_\ell := (\mathbb{E}_\ell - \sum_{j=\ell+1}^r \mathbb{E}_j \mathbb{E}_{j,\ell})\mathbb{E}_{\ell,\ell}^\dagger + P_\ell(I - \mathbb{E}_{\ell,\ell}\mathbb{E}_{\ell,\ell}^\dagger)$ where $P_\ell \in \mathbb{R}^{k\times m_\ell}$ is arbitrary.

10. If $\ell > 1$ go to Step 8.

11. End.

The algorithm essentially exploits the idea of Gauss–Jordan elimination [50] and is based on the new block-elimination procedure described in the previous subsection. The first stage reduces the system to block lower triangular form. The algorithm moves a pointer through the matrix $\mathbb{E}_{ff}$ along the leading diagonal from the $(1,1)$ position to the (r,r) position. We consider what happens during stage ℓ when the pointer is in the (ℓ,ℓ) position. The *current* equations

$$[\mathbf{p}_\ell, \mathbf{p}_{\ell+1}, \ldots, \mathbf{p}_r] \begin{bmatrix} \mathbb{E}_{\ell,\ell} & \mathbb{E}_{\ell,\ell+1} & \cdots & \mathbb{E}_{\ell,r} \\ \mathbb{E}_{\ell+1,\ell} & \mathbb{E}_{\ell+1,\ell+1} & \cdots & \mathbb{E}_{\ell+1,r} \\ \vdots & \vdots & & \vdots \\ \mathbb{E}_{r,\ell} & \mathbb{E}_{r,\ell+1} & \cdots & \mathbb{E}_{r,r} \end{bmatrix} = [\mathbb{E}_\ell, \mathbb{E}_{\ell+1}, \ldots, \mathbb{E}_r]$$

are defined in terms of the *current* vectors $\mathbf{p}_\ell, \mathbf{p}_{\ell+1}, \ldots, \mathbf{p}_r$ by the formulae

$$\mathbb{E}_{ij} = E[\mathbf{p}_i \mathbf{p}_j^T] - E[\mathbf{p}_i]E[\mathbf{p}_j] \quad \text{and} \quad \mathbb{E}_j = E[\mathbf{g}\mathbf{p}_j^T] - E[\mathbf{g}]E[\mathbf{p}_j^T].$$

The current *pivoting* coefficient $\mathbb{E}_{\ell,\ell} = E[\mathbf{p}_\ell \mathbf{p}_\ell^T] - E[\mathbf{p}_\ell]E[\mathbf{p}_\ell^T]$ is used to update the *remaining* vectors $\mathbf{p}_{\ell+1}, \mathbf{p}_{\ell+2}, \ldots, \mathbf{p}_r$ and all elements in the *remaining* equations

$$[\mathbf{p}_{\ell+1}, \mathbf{p}_{\ell+2}, \ldots, \mathbf{p}_r] \begin{bmatrix} \mathbb{E}_{\ell+1,\ell+1} & \mathbb{E}_{\ell+1,\ell+2} & \cdots & \mathbb{E}_{\ell+1,r} \\ \mathbb{E}_{\ell+2,\ell+1} & \mathbb{E}_{\ell+2,\ell+2} & \cdots & \mathbb{E}_{\ell+2,r} \\ \vdots & \vdots & & \vdots \\ \mathbb{E}_{r,\ell+1} & \mathbb{E}_{r,\ell+2} & \cdots & \mathbb{E}_{r,r} \end{bmatrix} = [\mathbb{E}_{\ell+1}, \mathbb{E}_{\ell+2}, \ldots, \mathbb{E}_r]$$

according to the formulae

$$\mathbf{p}_j := \mathbf{p}_j - \mathbb{E}_{j,\ell}\mathbb{E}_{\ell,\ell}^\dagger \mathbf{p}_\ell, \quad \mathbb{E}_{i,j} := \mathbb{E}_{i,j} - \mathbb{E}_{i,\ell}\mathbb{E}_{\ell,\ell}^\dagger \mathbb{E}_{\ell,j}$$

and

$$\mathbb{E}_j := \mathbb{E}_j - \mathbb{E}_\ell \mathbb{E}_{\ell,\ell}^\dagger \mathbb{E}_{\ell,j}.$$

Note that the *remaining* elements $\mathbb{E}_{\ell,\ell+1}, \mathbb{E}_{\ell,\ell+2}, \ldots, F_{\ell,r}$ in row ℓ of the coefficient matrix are eliminated at this stage because Lemma A1.2 (Appendix 1 and [154]) shows us that

$$\mathbb{E}_{\ell,j} - \mathbb{E}_{\ell,\ell} \mathbb{E}_{\ell,\ell}^\dagger \mathbb{E}_{\ell,j} = \mathbb{O},$$

where $\mathbb{O}$ is the zero matrix.

The pointer now moves forward one place to the $(\ell+1, \ell+1)$ position. Our earlier results show us that the updated system at stage $\ell+1$ retains the same relative structure as the original system at stage ℓ. In particular we preserve the relationships $\mathbb{E}_{i,j} = E[\mathbf{p}_i \mathbf{p}_j^T] - E[\mathbf{p}_i] E[\mathbf{p}_j^T]$ and $\mathbb{E}_j = E[\mathbf{g}\mathbf{p}_j^T] - E[\mathbf{g}] E[\mathbf{p}_j^T]$.

The second stage of the algorithm is a block back substitution. In general we can see from the lower triangular form of the system at stage q that

$$X_q = \left[\mathbb{E}_q - \sum_{j=q+1}^{r} X_j \mathbb{E}_{j,q} \right] \mathbb{E}_{q,q}^\dagger + P_q (I - \mathbb{E}_{q,q} \mathbb{E}_{q,q}^\dagger)$$

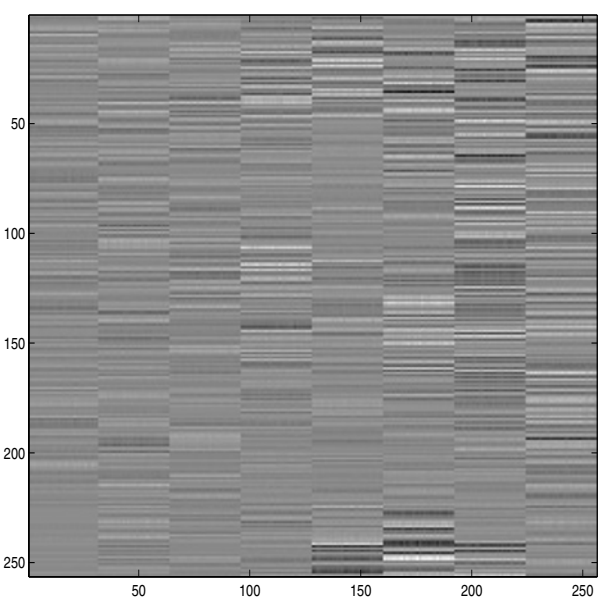
where $P_q \in \mathbb{R}^{k \times m_q}$ is arbitrary.

5.5.6 Simulations

We wish to demonstrate the advantages of the methods considered in Sections 5.5.4 and 5.5.5 with the simulation of systems transforming digitized images.

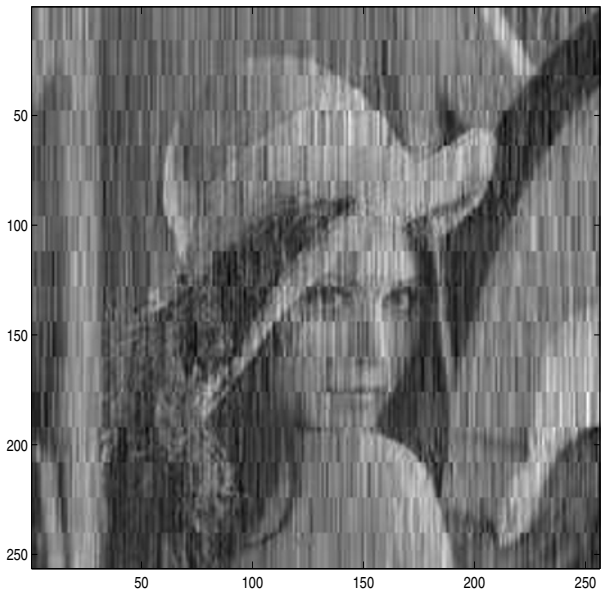
Let matrices $X \in \mathbb{R}^{256 \times 256}$ and $V \in \mathbb{R}^{256 \times 256}$ be counterparts of the image presented in Fig. 1 (a) and the known image "Lenna," respectively. We partition X and V into 128 submatrices $X_{ij}, V_{ij} \in \mathbb{R}^{16 \times 32}$ with $i = 1, \ldots, 16$ and $j = 1, \ldots, 8$ so that $X = \{X_{ij}\}$ and $V = \{V_{ij}\}$. Let $F_{ij} : X_{ij} \to V_{ij}$. Each submatrix X_{ij} has been treated as a set of 32 realizations of a random signal with columns representing realizations. The operator $F_{ij} : X_{ij} \to V_{ij}$ is interpreted as the mathematical model of a nonlinear system, where a column of X_{ij} is the input signal and the corresponding column of V_{ij} is the output signal.

(a) Given "idealized" inputs X_{ij}. This digitized image has been taken from http://sipi.usc.edu/database/.



(b) Observed inputs $Y_{ij}^{(1)}$.

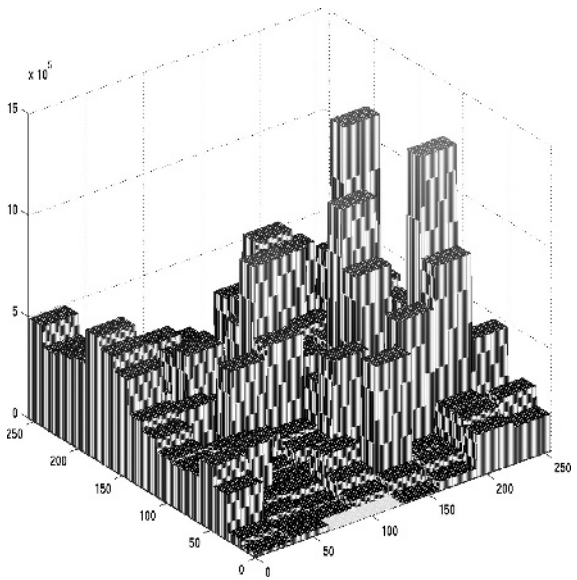
Figure 5.5: Illustration to the performance of the considered method.



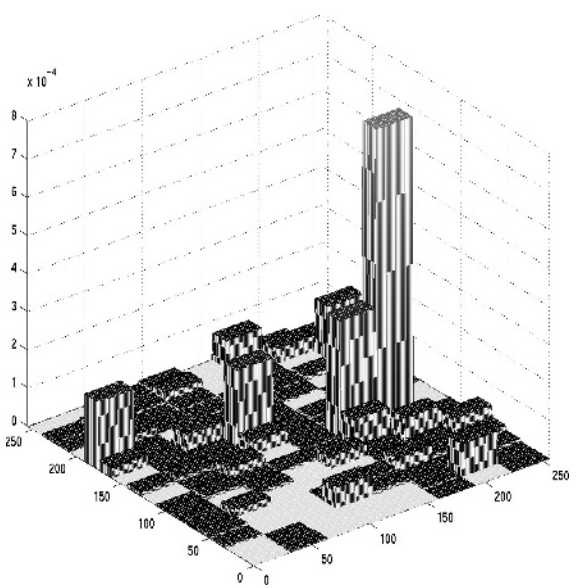
(a) Outputs of $\tilde{H}_{ij}^{(k)}$ by [151].

(b) Outputs of $\tilde{P}_{2,ij}^{(1)}$.

Figure 5.6: Illustration to the performance of the considered method. The digitized image has been taken from http://sipi.usc.edu/database/.



(a) Errors associated with method [151].



(b) Errors associated with $\tilde{P}^{(1)}_{2,ij}(Y^{(1)}_{ij})$.

Figure 5.7: Illustration to errors associated with the considered method.

Table 1.

k	Observed inputs	$\max_{ij} \Delta^{(k)}_{H,ij}$	$\max_{ij} \Delta^{(k)}_{2,ij}$
1	$Y^{(1)}_{ij} = 20\mathcal{R}^{(1)}_{ij} X_{ij} \mathcal{Q}^{(1)}_{ij} + 50\mathcal{Q}^{(2)}_{ij}$	1.4074×10^{6}	7.7375×10^{-4}
2	$Y^{(2)}_{ij} = X_{ij}\mathcal{R}^{(2)}_{ij} + 20\mathcal{Q}^{(3)}_{ij}$	6.0448×10^{4}	2.4749×10^{-15}
3	$Y^{(3)}_{ij} = \mathcal{Q}^{(4)}_{ij} X_{ij} + X_{ij}\mathcal{R}^{(3)}_{ij}$	2.4671×10^{5}	1.2915×10^{-11}

In practice, input signals are contaminated with noise. We simulated three different types of noisy input signal in the form $Y^{(k)}_{ij}$ with $k = 1, 2, 3$ presented in Table 1, where $\mathcal{R}^{(k)}_{ij}$ is a matrix with normally distributed entries with mean 0 and variance 1 and $\mathcal{Q}^{(k)}_{ij}$ is a matrix with uniformly distributed entries in the interval $(0, 1)$. In Table 1,

$$\Delta^{(k)}_{H,ij} = \|F_{ij}(X_{ij}) - \tilde{H}^{(k)}_{ij}(Y^{(k)}_{ij})\|^2 \quad \text{and} \quad \Delta^{(k)}_{2,ij} = \|F_{ij}(X_{ij}) - \tilde{P}^{(k)}_{2,ij}(Y^{(k)}_{ij})\|^2,$$

where $\tilde{H}^{(k)}_{ij}$ is the best approximation for F_{ij} by the method [151] and $\tilde{P}^{(k)}_{2,ij}$ is the best approximation of the second degree for F_{ij} constructed from (5.110), (5.118), (5.123)-(5.125) with $r = 2$.

$\tilde{H}^{(k)}_{ij}(Y^{(k)}_{ij})$ and $\tilde{P}^{(k)}_{2,ij}(Y^{(k)}_{ij})$ have been calculated with Matlab for each i, j and k (i.e. the method has been tested 384 times). We put $K_1 = K_2 = \mathbb{O}$ in (5.124), (5.125).

The figures illustrate the performance of the methods for $k = 1$, i.e. for $Y^{(1)}_{ij}$ in Table 1. The matrices for the digitized images in Figs. (b)-(d) have been composed from sub-matrices $Y^{(1)}_{ij}$, $\tilde{H}^{(1)}_{ij}(Y^{(1)}_{ij})$ and $\tilde{P}^{(1)}_{2,ij}(Y^{(1)}_{ij})$ correspondingly. The expectations and covariance matrices in (5.123)-(5.125) have been estimated from known simple equations [174]. For instance, for each i, j and k we estimated E_{vz_q} as $V_{ij}(Z^{(k)}_{ij})^T/32 - \hat{V}_{ij}(\hat{Z}^{(k)}_{ij})^T$, where $Z^{(k)}_{ij} = Y^{(k)}_{ij}(\mathrm{diag}Y^{(k)}_{ij}(q,:)), \mathrm{diag}Y^{(k)}_{ij}(q,:))$ is a diagonal matrix with the elements from the qth row of $Y^{(k)}_{ij}$ on the diagonal, and $\hat{M}$ means a vector formed from means of the rows of a matrix M. These simple estimates have been chosen to illustrate the performance of the considered method. Special estimates have been considered in Section 4.3.

Fig. 5.7 represents the matrices $\{\Delta^{(k)}_{H,ij}\}$ and $\{\Delta^{(k)}_{2,ij}\}$ of errors associated with the best approximations $\tilde{H}^{(k)}_{ij}(Y_{ij})$ by [151] and the best approximations $\tilde{P}^{(1)}_{2,ij}(Y_{ij})$ by (5.118), (5.123)- (5.125) with $r = 2$, respectively.

We see from Table 1 and the figures that the best approximations $\tilde{P}^{(k)}_{2,ij}$ give significant improvements in the accuracy of approximation to F_{ij} compared with the best approximations by [151]: the error associated with the

considered method with $r = 2$ is at least 10^{10} times less than the error associated with the method [151]. In fact, this is an illustration of Theorem 38.

In summary, the approach presented in this Section, is based on the solution of the best approximation problem for the input-output map. It is supposed that the observed input is distorted by an unknown effect. The approximant is given by the special polynomial operator of the rth degree with $r = 1, 2, \ldots$ and minimizes the mean squared error between the desired output and the output of the approximating system.

5.6 Best Causal Approximation

In Chapter 3, we have considered a new concept for the representation of realistic systems with any *pre-assigned accuracy*. In this section, we provide a new technique for the *best causal* representation of nonlinear systems in the sense of minimizing an associated error.

The approach presented here develops some ideas from [9] - [159] and is based on the best constrained approximation of mapping $\mathcal{F}$ in probability spaces by polynomial operator $\mathcal{P}_r$ of degree r. The operator $\mathcal{P}_r$ is designed from matrices of a special form. This allows us to solve the best approximation problem with the constraints on the matrix structures. The special matrix structures imply the incorporation of the causality concept into the models. As a result, the approximant preserves the causality principle and minimizes the mean square difference between a desired output $\mathcal{F}(\mathbf{x})$ and the output $\mathcal{P}_r(\mathbf{y})$ of the approximating model $\mathcal{P}_r$. It is supposed that the observable input $\mathbf{y}$ represents an idealized input $\mathbf{x}$ contaminated with noise. Unlike the known approaches to the modelling of nonlinear systems, it is not assumed here that $\mathbf{x}$ and $\mathbf{y}$ can be presented as analytical expressions. The inputs and outputs of the system under consideration are elements of the probability spaces and therefore relationships between them are assumed to be given by some covariance matrices only. Another difference is that we consider the *best* causal model of nonlinear systems. In other words, the model that we provides guarantees the smallest associated error in the entire class of models under consideration.

In Section 5.6.1, we present the model of the nonlinear system, reformulate and extend the heuristic definition of causality and show how the model is adjusted to the causality concept. In particular, we define so called (δ, ε)-causality which is closer to realistic conditions than the earlier notion of 'idealized' causality. To satisfy the causality concept, the model is reduced to a representation by matrices of special form.

The rigorous statement of the problem is given in Section 5.6.2.

In Section 5.6.3, we provide a solution to the problem, i.e. we obtain the equations for the matrices which determine the optimal model $\mathcal{P}_r^0$. We also establish the error equation associated with $\mathcal{P}_r^0$. It is shown that the model has a degree of freedom, the degree r of the operator $\mathcal{P}_r$. In particular, we establish that the error is decreased if the degree r of $\mathcal{P}_r^0$ is increased. This fact gives us the degree of freedom in manipulating with the model adjustment.

Simulations are described in Section 5.6.4.

5.6.1 Causality

Suppose that $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$, $\mathbf{y} \in L^2(\Omega, \mathbb{R}^m)$ and $\mathbf{u} \in L^2(\Omega, \mathbb{R}^m)$ are random vectors with realizations $\mathbf{x}(\omega) \in \mathbb{R}^m$, $\mathbf{y}(\omega) \in \mathbb{R}^m$ and $\mathbf{u}(\omega) \in \mathbb{R}^m$, respectively. As before, we denote $x = \mathbf{x}(\omega)$, $y = \mathbf{y}(\omega)$ and $u = \mathbf{u}(\omega)$.

Let $P_r : \mathbb{R}^m \to \mathbb{R}^m$ be given by (5.107), i.e. by

$$P_r(y) = A^{(0)} + A^{(1)}y + \sum_{q=2}^{r} \left[\sum_{q-1 \le \sigma_{q-1} \le (q-1)n} \tilde{A}_{j_1,\ldots,j_{q-1}} z_{j_1,\ldots,j_{q-1}} \right], \quad (5.144)$$

where $A^{(0)} \in \mathbb{R}^m$, $A^{(1)} \in \mathbb{R}^{m\times m}$, $\tilde{A}_{j_1,\ldots,j_{q-1}} \in \mathbb{R}^{m\times m}$, $z_{j_1,\ldots,j_{q-1}} = y_{j_1} \ldots y_{j_{q-1}} y$, with $j_1, \ldots, j_{q-1} = 1, \ldots, m$ and $q = 2, \ldots, r$, and where $\sigma_{q-1} = j_1 + \ldots + j_{q-1}$ and the inner sum is extended for all summands with subscripts satisfying the inequality $q - 1 \le \sigma_{q-1} \le (q-1)m$.

Let $\mathcal{P}_r : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^m)$ be given by

$$[\mathcal{P}_r(\mathbf{y})](\omega) = P_r[\mathbf{y}(\omega)].$$

We write

$$\mathcal{P}_r(\mathbf{y}) = \begin{bmatrix} \mathbf{p}_1(\mathbf{y}_1, \ ..., \ \mathbf{y}_m) \\ \vdots \\ \mathbf{p}_m(\mathbf{y}_1, \ ..., \ \mathbf{y}_m) \end{bmatrix},$$

where $\mathbf{p}_i : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R})$.

Next, let us denote $\mathbf{h} = \mathcal{P}_r(\mathbf{y})$ where $\mathbf{h} = (\mathbf{h}_1 \ ... \ \mathbf{h}_m)^T \in L^2(\Omega, \mathbb{R}^m)$. This equation can be rewritten as the set of equations

$$\mathbf{h}_i = \mathbf{p}_i(\mathbf{y}_1, \ ..., \ \mathbf{y}_m)$$

for $i = 1, ..., m$. Note that each component $\mathbf{h}_i$ (or $\mathbf{y}_i$) can be interpreted as a value of $\mathbf{h}$ (or $\mathbf{y}$, respectively) at time t_i.

We recall that by the heuristic definition of causality, the present value of the output of a physical system is not affected by future values of the input [116]. Since the operator $\mathcal{P}_r$ is treated as a model of the system, we formalize the causality concept in terms of the operator $\mathcal{P}_r$.

Definition 27. *We call the operator $\mathcal{P}_r$ causal if $\mathbf{h}_i$ is determined from components $\mathbf{y}_1, ..., \mathbf{y}_k$ with $k \leq i$, i.e. if for $i = 1, ..., m$,*

$$\mathbf{h}_i = \mathbf{p}_i(\mathbf{y}_1, \ ..., \ \mathbf{y}_k) \quad with \quad k \leq i.$$

An alternative definition is as follows.

Definition 28. *The operator $\mathcal{P}_r$ is called causal if for any $\mathbf{v} = (\mathbf{v}_1 \ ... \ \mathbf{v}_m)^T \in L^2(\Omega, \mathbb{R}^m)$ and $\mathbf{w} = (\mathbf{w}_1 \ ... \ \mathbf{w}_m)^T \in L^2(\Omega, \mathbb{R}^m)$,*

$$(\mathbf{v}_1 \ ... \ \mathbf{v}_k)^T = (\mathbf{w}_1 \ ... \ \mathbf{w}_k)^T$$

implies

$$\mathbf{p}_i(\mathbf{v}_1 \ ... \ \mathbf{v}_k) = \mathbf{p}_i(\mathbf{w}_1 \ ... \ \mathbf{w}_k),$$

where $k \leq i$.

In other words, $\mathcal{P}_r$ is causal if matrix $\tilde{A}_{j_1 ... j_{q-1}} = \{a_{kl}\}$ is such that $a_{kl} = 0$ for all $l = 1, ..., m$ if $k = 1, ..., j-1$ where $j = \max\{j_1, ..., j_{q-1}\}$, and also $a_{kl} = 0$ if $k = j, j+1, ..., m$ and $k < l$. Such a matrix is called *j-lower trapezoidal.*

An example of 4×4 3-lower trapezoidal matrix is as follows:

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}.$$

In particular, if $\mathcal{P}_r$ is linear, i.e. P_r is a matrix A, then $\mathcal{P}_r$ is causal if A is lower triangular.

The class of j-lower trapezoidal matrices is denoted by $\mathcal{T}_j$. Note that the 1-lower trapezoidal matrix is lower triangular.

The above implies the following definition.

Definition 29. *The operator $\mathcal{P}_r$ is called causal if matrix $\tilde{A}_{j_1 ... j_{q-1}}$ is j-lower trapezoidal where $j = \max\{j_1, ..., j_{q-1}\}$.*

In the real world, information is often obtained with some error, caused in particular, by the influence of external factors, data and instrument inexactness, etc. In this sense, the definition above is rather idealistic. A more realistic definition of causality for the operator $\mathcal{P}_r$ is as follows.

Definition 30. *The operator $\mathcal{P}_r$ is called (δ, ε)-causal if for any $\delta \geq 0$ there exists $\varepsilon \geq 0$ such that for arbitrary $v = (v_1 \ ... \ v_m)^T \in \mathbb{R}^m$ and $w = (w_1 \ ... \ w_m)^T \in \mathbb{R}^m$,*

$$\|(v_1 \ ... \ v_k)^T - (w_1 \ ... \ w_k)^T\|^2 \leq \delta$$

implies

$$|p_i(v_1 \ldots v_k) - p_i(w_1 \ldots w_k)| \le \varepsilon,$$

where $k \le i$.

It is clear that the (0, 0)-causal operator is causal in the sense of Definition 1.

Proposition 2. *If matrix* $\tilde{A}_{j_1 \ldots j_{q-1}}$ *is* j*-lower trapezoidal for each* j *then* $\mathcal{P}_r$ *is the* (δ, ε)*-causal operator.*

Proof. The proof follows directly from the above definitions. □

Next, similarly to (5.111)–(5.118), it is shown that

$$\sum_{j_1 + \ldots j_{q-1} \le (q-1)m} \tilde{A}_{j_1, \ldots, j_{q-1}} z_{j_1, \ldots, j_{q-1}} = \sum_{i=1}^{N} A_{i+1} u_i,$$

$$P_r(y) = A_0 + A_1 y + \sum_{i=1}^{N} A_{i+1} u_i \tag{5.145}$$

and

$$\mathcal{P}_r(\mathbf{y}) = \mathcal{A}_0 + \mathcal{A}_1(\mathbf{y}) + \sum_{i=1}^{N} \mathcal{A}_{i+1}(\mathbf{u}_i), \tag{5.146}$$

where N is defined similarly to that in (5.115), and matrices $A_{i+1} \in \mathbb{R}^{m \times m}$, vectors $u_i \in \mathbb{R}^m$ and $\mathbf{u}_i \in L^2(\Omega, \mathbb{R}^m)$, and operators $\mathcal{A}_{i+1} : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^m)$ are defined in the manner of matrices C_j, vectors u_j and $\mathbf{u}_j$, and operators $\mathcal{C}_j$ in (5.117) and (5.118), respectively.

5.6.2 Statement of the problem

We identify a nonlinear system with the continuous operator $\mathcal{F} : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^m)$. It is supposed that a structure of $\mathcal{F}$ is either unknown or is difficult to compute. It is also assumed that a relationship between the input $\mathbf{x}$ of the system $\mathcal{F}$ and an observable input $\mathbf{y}$ is not known.

We wish to find a causal optimal model of the system $\mathcal{F}$ which minimizes the associated error.

Let

$$J(A_0, A_1, \ldots, A_{N+1}) = E[\|\mathcal{F}(\mathbf{x}) - \mathcal{P}_r(\mathbf{y})\|^2],$$

where $A_j \in \mathcal{T}_j$ for $j = 1, \ldots, N+1$.

The problem is to find A_0^0, A_1^0, ... A_{N+1}^0 such that

$$J(A_0^0, A_1^0, ..., A_{N+1}^0) = \min_{A_0, A_1, ..., A_{N+1}} J(A_0, A_1, ..., A_{N+1}) \tag{5.147}$$

subject to

$$A_j^0 \in \mathcal{T}_j \quad \text{for each} \quad j = 1, ..., N+1. \tag{5.148}$$

Let us denote $P_r^0 = P_r^0(A_0^0, A_1^0, ..., A_{N+1}^0)$. By Proposition 2, the condition $A_j^0 \in \mathcal{T}_j$ for $j = 1, ..., N+1$ implies (δ, ε)-causality of $\mathcal{P}_r^0$.

We note that the only restriction imposed on $\mathcal{F}$ is its continuity. It is not supposed that some properties of $\mathcal{F}$, such as causality, memory, etc., are known but we wish that $\mathcal{P}_r^0$ is (δ, ε)-causal and constructively defined.

5.6.3 Best causal polynomial model of the system

In this section, we provide a solution to the problem posed above and also present an error analysis associated with the solution.

Let

$$\mathbf{s} = \mathcal{F}(\mathbf{x}), \ \ s = F(x), \ \ u = (u_1 \ ... \ u_N)^T, \ \ G = E_{su} - E_{sy} E_{yy}^{\dagger} E_{yu}, \tag{5.149}$$

$$D = E_{uu} - E_{uy} E_{yy}^{\dagger} E_{yu}, \quad H = G D^{\dagger} \in \mathbb{R}^{m \times N}$$

and

$$D^{1/2} = \begin{bmatrix} Q_{11} & ... & Q_{1N} \\ ... & ... & ... \\ Q_{N1} & .... & Q_{NN} \end{bmatrix} \quad \text{with} \quad Q_{ij} \in \mathbb{R}^{m \times m},$$

where $E_{yy}^{\dagger}$ and $D^{\dagger}$ are the pseudo-inverses of E_{yy} and D respectively, and where sub-matrices Q_{ij} are assumed to be nonsingular for all $i, j = 1, ..., N$.

Let matrices $Q_i = Q_{ii} Q_{ii}^T$ be positive definite for all $i = 1, ..., N$ so that there exists the Cholesky factorization [50] for Q_i,

$$Q_i = L_i L_i^T, \tag{5.150}$$

where L_i is lower triangular. We write $[H_1 \ ... \ H_N] = H$ where $H_i \in \mathbb{R}^{m \times m}$, and

$$(H_i - \sum_{k=1, k \neq i}^{N} (A_{k+1}^0 - H_k) Q_{ki} Q_{ii}^{-1}) L_i = K_{1i} + K_{2i} + K_{3i}, \tag{5.151}$$

where A_{k+1}^0 is defined by the following Theorem 40, K_{2i} is i-lower trapezoidal, K_{3i} is strictly upper triangular (i.e. with the zero entries on the

main diagonal) and K_{1i} is a matrix which supplements K_{2i} to lower triangular matrix.

For example, matrix $\begin{bmatrix} a_{11} & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ supplements 4×4 3-lower trapezoidal matrix to the lower triangular matrix.

We also suppose that E_{yy} is positive definite, therefore $E_{yy}^{\dagger} = E_{yy}^{-1}$, and there exists the Cholesky factorization for E_{yy},

$$E_{yy} = RR^T, \tag{5.152}$$

where R is lower triangular.

Each matrix can be presented as a sum of lower triangular and strictly upper triangular matrices. We write

$$(E_{sy} - \sum_{k=1}^{N} A_{k+1}^0 E_{u_k y})R^{-T} = M_1 + M_2, \tag{5.153}$$

where M_1 is lower triangular and M_2 is strictly upper triangular.

Theorem 40. *Under the assumptions above, the best (δ, ε)-causal model $\mathcal{P}_r^0$ of the system $\mathcal{F}$ is given by*

$$P_r^0(y) = A_0^0 + A_1^0 y + \sum_{i=1}^{N} A_{i+1}^0 u_i, \tag{5.154}$$

where

$$A_0^0 = E[\mathbf{s}] - A_1^0 E[\mathbf{y}] - \sum_{k=1}^{N} A_{k+1} E[\mathbf{u}_k], \tag{5.155}$$

$$A_1^0 = M_1 R^{-1} \tag{5.156}$$

and for each $i = 1, ..., N$,

$$A_{i+1}^0 = K_{2i} L_i^{-1}. \tag{5.157}$$

Proof. It follows from Proposition 6.9 and Section ... that

$$J(A_0, A_1, ..., A_{N+1}) = J_0 + J_1 + J_2 + J_3,$$

where

$$J_0 = \|E_{ss}^{1/2}\|^2 - \|E_{sy}(E_{yy}^{-1})^{1/2}\|^2 - \|G(D^{\dagger})^{1/2}\|^2,$$

$$J_1 = \left\| A_0 - E[\mathbf{s}] + A_1 E[\mathbf{y}] + \sum_{k=1}^{N} A_{k+1} E[\mathbf{u}_k] \right\|^2,$$

$$J_2 = \left\| \left[A_1 - (E_{sy} - \sum_{k=1}^{N} A_{k+1} E_{u_k y}) E_{yy}^{-1} \right] E_{yy}^{1/2} \right\|^2$$

and

$$J_3 = \left\| ([A_2 ... A_{N+1}] - H)\, D^{1/2} \right\|^2.$$

We have

$$\begin{aligned}
J_2 &= \operatorname{tr}\left\{ \left[A_1 - \left(E_{sy} - \sum_{k=1}^{N} A_{k+1} E_{u_k y} \right) E_{yy}^{-1} \right] E_{yy} \right. \\
&\qquad \left. \times \left[A_1 - \left(E_{sy} - \sum_{k=1}^{N} A_{k+1} E_{u_k y} \right) E_{yy}^{-1} \right]^T \right\} \\
&= \operatorname{tr}\{[A_1 R - M_1 - M_2][R^T A_1^T - M_1^T - M_2^T]\} \\
&= \operatorname{tr}\{[A_1 R - M_1][R^T A_1^T - M_1^T]\} - \operatorname{tr}\{A_1 R M_2^T \\
&\quad + M_2 R^T A_1^T\} \\
&\quad + \operatorname{tr}\{M_2 M_1^T + M_1 M_2^T + M_2 M_2^T\} \\
&= \operatorname{tr}\{[A_1 R - M_1][R^T A_1^T - M_1^T]\} \\
&= \|A_1 R - M_1\|^2
\end{aligned}$$

where

$$\operatorname{tr}\{A_1 R M_2^T + M_2 R^T A_1^T\} = 0$$

and

$$\operatorname{tr}\{M_2 M_1^T + M_1 M_2^T + M_2 M_2^T\} = 0$$

since A_1 is lower triangular.

Hence (5.156) is true. Next,

$$\begin{aligned}
J_3 &= \left\| \left[\sum_{k=1}^{N} (A_{k+1} - H_k) Q_{k1},\ \ldots,\ \sum_{k=1}^{N} (A_{k+1} - H_k) Q_{kN} \right] \right\|^2 \\
&= \left\| \sum_{k=1}^{N} (A_{k+1} - H_k) Q_{k1} \right\|^2 + \ldots + \left\| \sum_{k=1}^{N} (A_{k+1} - H_k) Q_{kN} \right\|^2 \\
&= \sum_{j=1}^{N} J(A_{j+1}),
\end{aligned}$$

where

$$\begin{aligned} J(A_{j+1}) &= \left\| \left[A_{j+1} - \left(H_j - \sum_{\substack{k=2 \\ k\neq j}}^{N} [A_{k+1} - H_k] Q_{kj} Q_{jj}^{-1} \right) \right] Q_{jj} \right\|^2 \\ &= \mathrm{tr}[A_{j+1}L_j - (K_{1j} + K_{2j} + K_{3j})][L_j^T A_{j+1}^T \\ &\quad -(K_{1j}^T + K_{2j}^T + K_{3j}^T)] \\ &= \mathrm{tr}\{[(A_{j+1}L_j - K_{2j}) - (K_{1j} + K_{3j})]\,[(L_j^T A_{j+1}^T - K_{2j}^T) \\ &\quad -(K_{1j}^T + K_{3j}^T)]\} \\ &= \mathrm{tr}\{(A_{j+1}L_j - K_{2j})(L_j^T A_{j+1}^T - K_{2j}^T)\} \\ &= \|A_{j+1}L_j - K_{2j}\|^2 \end{aligned}$$

since

$$\mathrm{tr}\{(A_{j+1}L_j - K_{2j})(K_{1j}^T + K_{3j}^T)\} = 0,$$

$$\mathrm{tr}\{(K_{1j} + K_{3j})(L_j^T A_{j+1}^T - K_{2j}^T)\} = 0$$

and

$$\mathrm{tr}\{(K_{1j} + K_{3j})(K_{1j}^T + K_{3j}^T)\} = 0.$$

Thus matrices A_{j+1}^0 with $j = 1, ..., N$ minimize J_3, and therefore (5.157) is true.

Matrices A_1^0 and A_{i+1}^0 defined by (5.156) and (5.157) are lower triangular and i-lower trapezoidal, respectively. It implies the (δ, ε)-causality of the operator $\mathcal{P}_r^0$.

The theorem is proven. □

We note that the operator $\mathcal{P}_r^0$ defined by Theorem 40 is constructive, i.e. $\mathcal{P}_r^0$ is numerically realizable with the standard software packages. Matrices A_{i+1}^0 for $i = 1, ..., N$ are determined form the set of equations (5.157).

Theorem 41. *The error associated with the best (δ, ε)-causal model $\mathcal{P}_r^0$ presented by Theorem* 40, *is*

$$\begin{aligned} E[\|\mathcal{F}(\mathbf{x}) - \mathcal{P}_r^0(\mathbf{y})\|^2] &= \|E_{ss}^{1/2}\|^2 - \|E_{sy}(E_{yy}^{-1})^{1/2}\|^2 \\ &\quad -\|G(D^\dagger)^{1/2}\|^2. \end{aligned} \tag{5.158}$$

Proof. The proof follows directly from the above. □

Corollary 6. *The error $E[\|\mathcal{F}(\mathbf{x}) - \mathcal{P}_r^0(\mathbf{y})\|^2]$ associated with the best representation $\mathcal{P}_r^0$ of the system $\mathcal{F}$ is decreased if the degree r of $\mathcal{P}_r^0$ is increased.*

Proof. It follows from Theorem 1 that A_{i+1}^0 turns to zero the functional $J(A_{i+1})$ for each $i = 1, ..., N$. But

$$J_3 = \sum_{i=1}^{N} J(A_{i+1})$$

therefore $A_2^0, ..., A_{N+1}^0$ turn to zero J_3. This means that

$$[A_2^0 \; ... \; A_{N+1}^0] = GD^\dagger, \tag{5.159}$$

where G and D are subject to conditions (5.150)–(5.153). Next, it follows from (5.159) that $\|GD^\dagger\|$ is increased with an increase of N. Since $N = m + m^2 + ...m^{r-1}$, then the right hand side of (5.158) is decreased if degree r of $\mathcal{P}_r^0$ is increased. □

Thus the degree r of the model $\mathcal{P}_r^0$ is a degree of freedom which allow us to adjust the model to an accuracy which is defined by conditions of a particular modelling problem.

5.6.4 Simulations

To illustrate the performance of the considered method, we simulate the inputs and outputs of the system $\mathcal{F}$ in the form of matrices $X = \{x_{ij}\}$ and $S = \{s_{ij}\}$ respectively, where

$$x_{ij} = (\exp(t_i) + a_j \exp(-t_i)) \sin(10t_i)$$

and

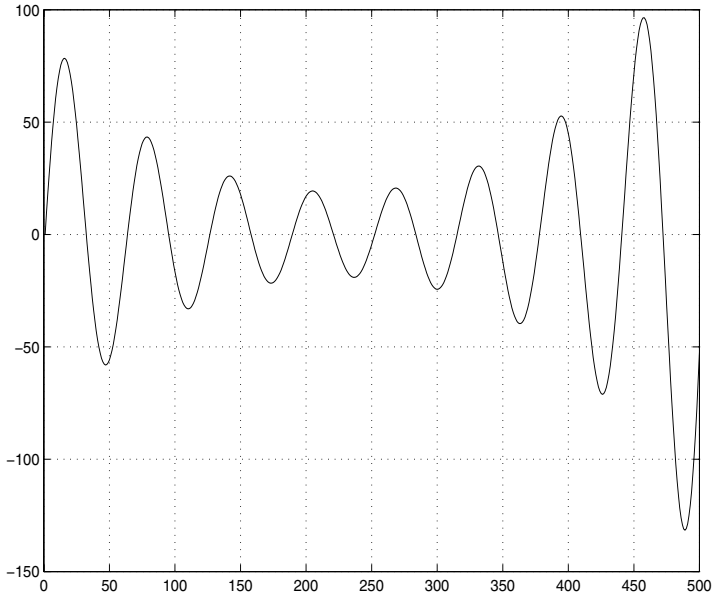
$$s_{ij} = b_j \sin(10t_i) \sin(t_i)$$

with $t_i = t_{i-1} + 0.01$, $a_j = a_{j-1} + 0.5$, $b_j = b_{j-1} + 0.5$, $i = 1, \ldots, 500$, $j = 1, \ldots, 100$, $t_0 = 0$, $a_0 = 0$ and $b_0 = -30$. We note that 'trigonometric' signals are often exploited in real-world data processing.

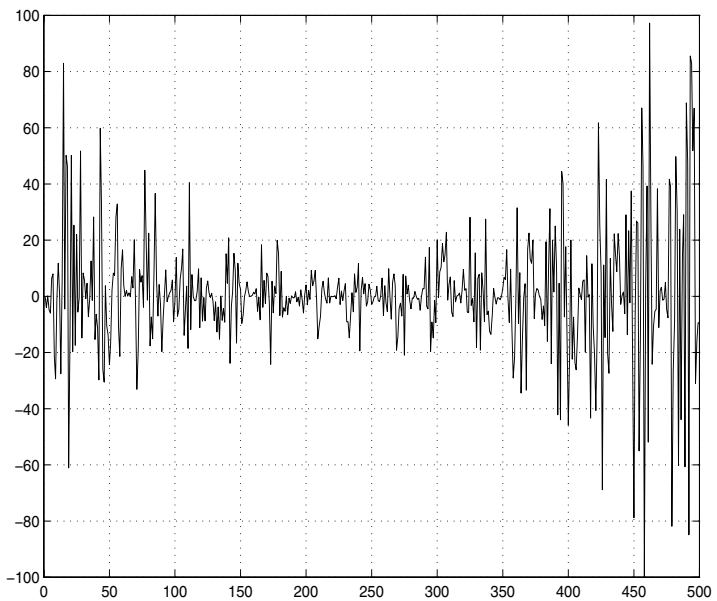
Columns of X and S represent realizations of the random vectors $\mathbf{x}$ and $\mathbf{s} = \mathcal{F}(\mathbf{x})$ correspondingly. The observed inputs of the system have been simulated in the form of the matrix $Y = M. * X\,C$ where C is a diagonal matrix with the nonzero entries $c_j = c_{j-1} + 0.5$, $j = 1, \ldots, 100$, $c_0 = -20$, $M = \{\mu_{ij}\}$ is a matrix with normally distributed entries with mean 0 and variance 1, and the symbol $.*$ means the Hadamard product.

We model the system $\mathcal{F}$ in the form of operator P_r^0, given by Theorem 40, for $r = 1$ and $r = 2$ so that

$$P_1^0(y) = A_0^0 + A_1^0 y \tag{5.160}$$

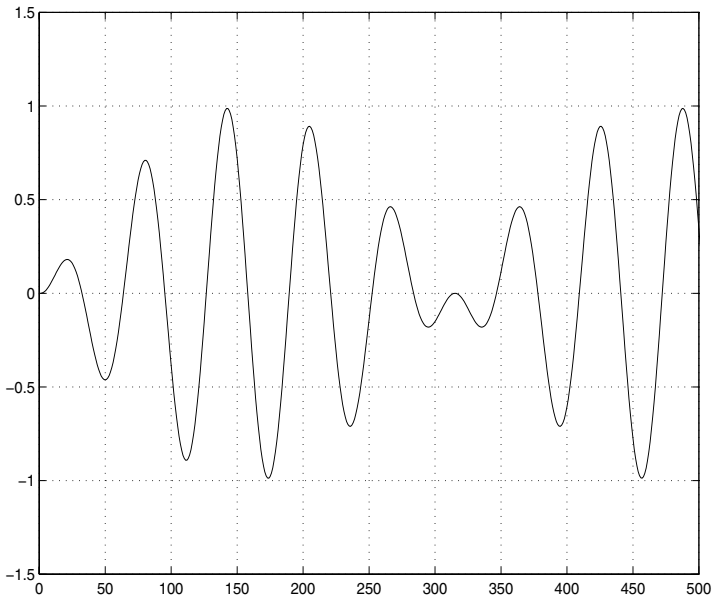


(a) Reference input x.

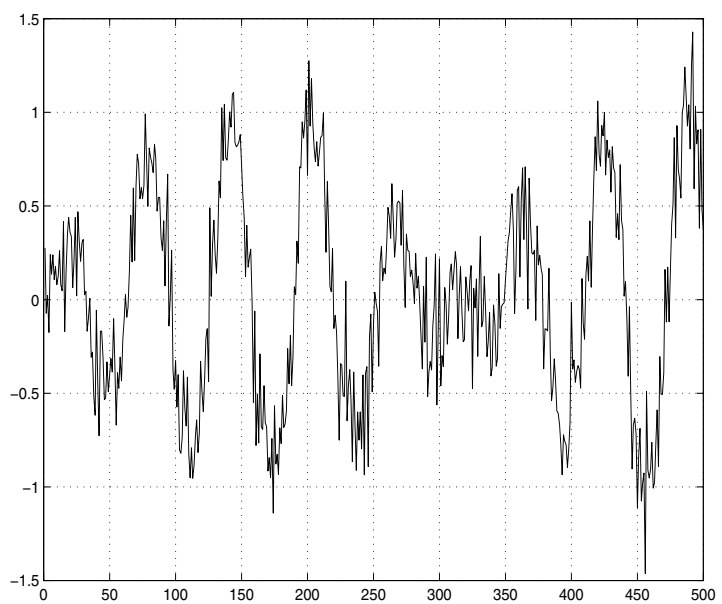


(b) Observed input y.

Figure 5.8: Illustration to performance of considered method.

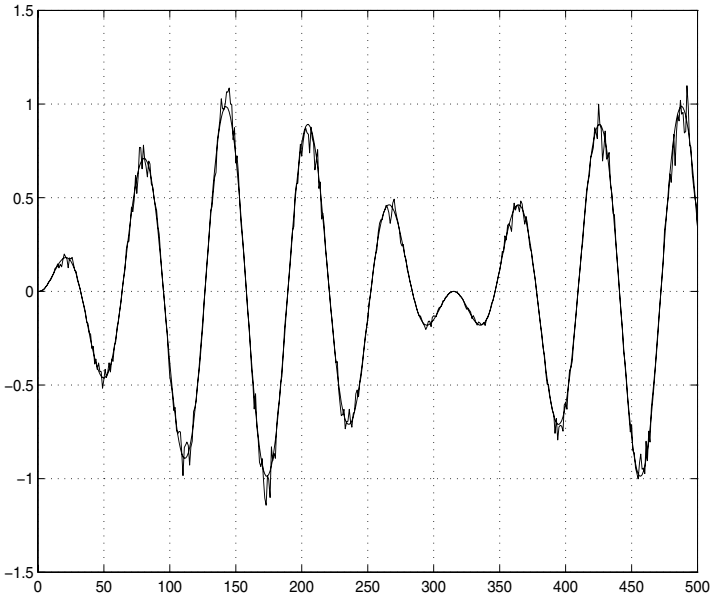


(a) Reference output s.

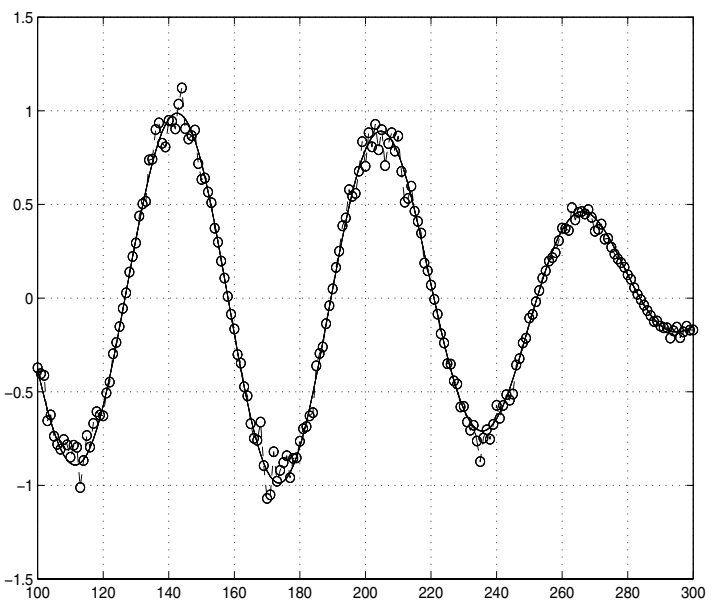


(b) Output of the model P_1^0.

Figure 5.9: Illustration to performance of considered method.



(a) Output p_2 of the model P_2^0.



(b) Fragments of outputs s (solid line) and p_2 (line with circles).

Figure 5.10: Illustration to performance of considered method.

and

$$P_2^0(y) = A_0^0 + A_1^0 y + \sum_{j_1=1}^{m} \tilde{A}_{j_1}^0 z_{j_1} = A_0^0 + A_1^0 y + \sum_{i=1}^{m} A_{i+1}^0 u_i \qquad (5.161)$$

with $z_{j_1} = y_{j_1} y = u_i$ for each $j_1, i = 1, \ldots, m$ (see (5.146) and (5.145)) where A_0^0, A_1^0, ..., A_{m+1}^0 have been determined from (5.155)–(5.157).

In these simulations, the expectations and covariance matrices used in (5.149) – (5.157), have been estimated from the entire samples X, S and Y of $\mathbf{x}$, $\mathbf{s}$ and $\mathbf{y}$, respectively.[3]

The figures 5.8– 5.10 present results of the simulations for $a_j = 90$, $b_j = 1$ and $c_j = 0.5$.

The figures demonstrate the significant improvement in quality of modelling when the degree r of the model P_r^0 is increased.

5.7 Best Hybrid Approximations

In this section, we consider a new approach to constructive representation of nonlinear systems which is based on a combination of ideas of the best approximation and iterative procedures.

The motivation for such an approach is as follows. Best–approximation methods have the aim of obtaining the best solution within a certain class, and therefore the solution cannot be improved by these techniques in cases when the approximation is not satisfactory. In contrast, iterative methods are normally convergent, but the error associated with each iteration of the particular method is not the smallest possible. As a result, convergence can be quite slow for a rather wide variety of problems. Moreover, in practice only a finite number of iterations can be carried out, and therefore the final approximate solution is often unsatisfactorily inaccurate.

A natural idea is to combine the above techniques to exploit their advantageous features. We present a method which realizes this idea. First, a special iterative procedure is considered with the aim of improving the accuracy of $\mathcal{F}$ approximation with each consequent iteration. Second, on each iteration, the best approximation problem is solved providing the smallest associated error within the chosen class of approximations for each iteration.

We show that the combination of these techniques allows us to build a computationally efficient and flexible method which has three degrees of freedom. See Remarks 19 and 20 in Section 5.7.7 in this connection. In

[3]The special methods of the estimation and related references are given in Section 4.3.

particular, we prove that the error in approximating $\mathcal{F}$ by the considered method decreases with an increase in the number of iterations.

This section delivers a substantially more effective methodology compared to the primary methods considered in Sections 5.4 and 5.5.

5.7.1 Some preliminaries

Before a formal statement of the problem, we describe the motivating idea.

Let $\mathcal{F}$ be the input-output map of a nonlinear system, and $\mathbf{x}$ and $\mathbf{s}$ the stochastic input and stochastic output of $\mathcal{F}$, respectively. Let $\mathbf{y}$ be the noise-corrupted version of $\mathbf{x}$. It is supposed that the input of the system, which approximates $\mathcal{F}$, is $\mathbf{y}$ and that information on $\mathcal{F}$ is given in terms of the statistical characteristics of $\mathbf{s}$ and $\mathbf{y}$ such as the mean, covariance matrices etc.

To find a system which approximates $\mathcal{F}$, the idea of a concatenation of approximating subsystems can be exploited in the following way. Let $\mathcal{B}_0$ approximate $\mathcal{F}$ in a certain sense. We call $\mathcal{B}_0$ a subsystem. The output of $\mathcal{B}_0$ is used as the input of the subsequent approximating subsystem $\mathcal{B}_1$ that has to be determined, and then the procedure is repeated. As a result, the link between $\mathbf{y}$ and $\mathbf{s}$ is modelled from the concatenation

$$\mathcal{P}_k = \mathcal{B}_k \circ \mathcal{B}_{k-1} \circ \ldots \circ \mathcal{B}_0$$

with $k = 0, 1, \ldots$. This device initiates the problem as follows. Find a constructive approximation $\mathcal{P}_k$ for $\mathcal{F}$ such that each $\mathcal{B}_k$ approximates $\mathcal{F}$ with a minimal possible error for every $k = 0, 1, \ldots$, and further, the error is decreased when k is increased. Since $\mathcal{P}_k$ is determined by $\mathcal{B}_k, \mathcal{B}_{k-1}, \ldots, \mathcal{B}_0$, the more precise formulation in terms of the approximating subsystem $\mathcal{B}_k$ is given in the Section 5.7.2.

We note that while the system concatenation is a natural idea, the methodology of an optimal determination of the parameters of each subsystem is not obvious. In particular, we point out that the nonlinearity of each approximating subsystem $\mathcal{B}_i$ is essential. No improvement in the accuracy can be achieved by the following subsystem $\mathcal{B}_{i+1}$ if $\mathcal{B}_{i+1}$ is linear. This observation is justified in Remark 21 of Section 5.7.7.

Suppose that $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$, $\mathbf{y} \in L^2(\Omega, \mathbb{R}^n)$ are random vectors with realizations $\mathbf{x}(\omega) \in \mathbb{R}^m$ and $\mathbf{y}(\omega) \in \mathbb{R}^n$.

Let the input-output map $\mathcal{F} : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^p)$ be such that $[\mathcal{F}(\mathbf{x})](\omega) = F[x]$ for each $\omega \in \Omega$ so that

$$\mathbf{s} = \mathcal{F}(\mathbf{x}) \quad \text{and} \quad s = F(x)$$

where $F \in L^2(\mathbb{R}^m, \mathbb{R}^p)$.

The approach described in Section 2.1, is implemented through the following device.

Let $\mathbf{s}_k \in L^2(\Omega, \mathbb{R}^p)$, $s_k = \mathbf{s}_k(\omega)$ and let us suppose that s_k is known for $k = 0, 1, 2, \ldots$ where $s_0 = y$. We set

$$s_{k+1} = P_k(y), \tag{5.162}$$

where $k = 0, 1, ...,$ and the nonlinear operator $P_k \in L^2(\mathbb{R}^n, \mathbb{R}^p)$ is determined by

$$P_k(y) = \tilde{B}_k(s_k), \tag{5.163}$$

where

$$\tilde{B}_k(s_k) = \tilde{A}_k^{(0)} + \sum_{q=1}^{r} \tilde{A}_k^{(q)}(s_k^q) \tag{5.164}$$

and where $\tilde{A}_k^{(0)} \in \mathbb{R}^p$, $\tilde{A}_k^{(q)} : (\mathbb{R}^\nu)^q \to \mathbb{R}^p$ is the q-linear operator with $\nu = n$ for $k = 0$ and $\nu = p$ for $k = 1, 2, \ldots$.

Let $\mathcal{B}_k : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^p)$ be defined by $[\mathcal{B}_k(\mathbf{s}_k)](\omega) = B_k(s_k)$ with

$$\mathcal{B}_k(\mathbf{s}_k) = A_k^{(0)} + \sum_{q=1}^{r} \mathcal{A}_k^{(q)}(\mathbf{s}_k^q) \tag{5.165}$$

and

$$B_k(s_k) = A_k^{(0)} + \sum_{q=1}^{r} A_k^{(q)}(s_k^q) \tag{5.166}$$

where

$$A_k^{(q)}(s_k^q) = [\mathcal{A}_k^{(q)}(\mathbf{s}_k^q)](\omega), \quad s_k^q = \mathbf{s}_k^q(\omega), \quad \mathcal{A}_k^{(q)} : L^2(\Omega, (\mathbb{R}^\nu)^q) \to L^2(\Omega, \mathbb{R}^p)$$

and ν is as above.

5.7.2 Statement of the problem

It follows from the above that the subsystem $\tilde{B}_k$ is defined by $\tilde{A}_k^{(0)}$, $\tilde{\mathcal{A}}_k^{(1)}$, ..., $\tilde{\mathcal{A}}_k^{(r)}$. Therefore it is natural to state the problem in the following form.

Let

$$J(A_k^{(0)}, \mathcal{A}_k^{(1)}, \ldots, \mathcal{A}_k^{(r)}) = E[\|\mathcal{F}(\mathbf{x}) - \mathcal{B}_k(\mathbf{s}_k)\|^2]. \tag{5.167}$$

For each $k = 0, 1, ...$, find the vector $\tilde{A}_k^{(0)}$ and operators $\tilde{\mathcal{A}}_k^{(1)}, \ldots, \tilde{\mathcal{A}}_k^{(r)}$ such that

$$J(\tilde{A}_k^{(0)}, \tilde{\mathcal{A}}_k^{(1)}, \ldots, \tilde{\mathcal{A}}_k^{(r)}) = \min_{A_k^{(0)}, \mathcal{A}_k^{(1)}, \ldots, \mathcal{A}_k^{(r)}} J(A_k^{(0)}, \mathcal{A}_k^{(1)}, \ldots, \mathcal{A}_k^{(r)}). \quad (5.168)$$

Thus, the solution will completely define $\tilde{B}_k$.

5.7.3 Method for the solution of problem (5.168)

In this section, we present the general structure of the considered method and its particularities, that is, a solution of the best approximation problem and the algorithm for a numerical realization.

The considered approach (5.162)–(5.168) implies the solutions of the sequence of the problems (5.167), (5.168) for each $k = 0, 1, ...$ with $s_0 = y$ and s_{k+1} in the form (5.162)–(5.164). For each kth iteration of the procedure (5.162)–(5.167), the operator $\tilde{B}_k$ represents the best polynomial approximant of the rth degree for F. Note that (5.163) can be written as

$$P_k(y) = \tilde{B}_k \circ \tilde{B}_{k-1} \circ \ldots \circ \tilde{B}_0(s_0). \quad (5.169)$$

We begin with a representation of B_k and (5.168) in a different form. For $q = 1, \ldots, r$, let us represent $A_k^{(q)}$ in the form

$$A_k^{(q)} = R_k^{(q)} \circ T_k^{(q)}$$

where $T_k^{(q)} : (\mathbb{R}^\nu)^q \to \mathbb{R}^l$ is the q-linear operator with ν defined in Section 2.2, and $R_k^{(q)} : \mathbb{R}^l \to \mathbb{R}^p$. In particular, $l = p$.

The reason for such a representation is twofold. First, we wish to determine $T_k^{(q)}$ in a way which simplifies the computational procedure for $\tilde{B}_k$. Second, we wish to optimize this simplified procedure by determining $R_k^{(q)}$ so that the associated error is minimized.

Next, $T_k^{(1)}$, ..., $T_k^{(r)}$ are multi-linear operators, i.e. the tensors. For our purposes, it is convenient to use a representation of the operator B_k in terms of linear operators, i.e. in matrix terms.

We proceed with this device in the following way. Lemma 32 below gives a matrix representation of B_k. In fact, Lemma 32 is a reformulation of Lemma 27 of Section 4.4.2(b). Section 4.4.2(d) provides procedures for determining the operator $T_k^{(q)}$ in terms of operator $\mathcal{D}_k^{(j+1)}$ which is introduced by (5.171)–(5.173) below. As a result, we reformulate the problem (5.168) in special matrix terms (5.186).

We write $s_k = (s_{k,1}, \ldots, s_{k,\nu})^T \in \mathbb{R}^\nu$, $z_{k,j_1,\ldots,j_{q-1}} = s_{kj_1} \ldots s_{kj_{q-1}} s_k$, where $q = 2, \ldots, r$ and $j_i = 1, \ldots, \nu$ for $i = 1, \ldots, q-1$.

Lemma 32. *There exist matrices* $T^{(q)}_{k,j_1,\ldots,j_{q-1}} \in \mathbb{R}^{\nu\times l}$ *such that*

$$B_k(s_k) = A_k^{(0)} + R_k^{(1)}T_k^{(1)}s_k + \sum_{q=2}^{r} R_k^{(q)}\left[\sum_{\sigma_{q-1}\le(q-1)\nu} T^{(q)}_{k,j_1,\ldots,j_{q-1}} z_{k,j_1,\ldots,j_{q-1}}\right], \quad (5.170)$$

where $\sigma_{q-1} = j_1+\ldots+j_{q-1}$ *and the inner sum is extended for all summands with subscripts satisfying the inequality* $\sigma_{q-1} \le (q-1)\nu$.

Next, let us write $Q^{(q)}_{k,j_1,\ldots,j_{q-1}} = R_k^{(q)}T^{(q)}_{k,j_1,\ldots,j_{q-1}}$. Similarly to (5.111)–(5.118), we can reduce (5.170) to a representation with a lesser number of terms. Namely, we set

$$\sum_{q=2}^{r}\sum_{\sigma_{q-1}\le(q-1)\nu} Q^{(q)}_{k,j_1,\ldots,j_{q-1}} z_{k,j_1,\ldots,j_{q-1}} = \sum_{j=2}^{N+1} C_k^{(j)}D_k^{(j)}u_{k,j-1}, \quad (5.171)$$

where matrices $C_k^{(j)}D_k^{(j)}$ and operands $u_{k,j-1} \in \mathbb{R}^{\nu}$ denote matrices $R_k^{(q)} \times T^{(q)}_{k,j_1,\ldots,j_{q-1}}$ and vectors $z_{k,j_1,\ldots,j_{q-1}}$, respectively, in the manner of that used in (5.116). The number N is defined similarly to (5.115).

We also set $C_k^{(1)}D_k^{(1)} = R_k^{(1)}T_k^{(1)}$ and $u_{k0} = s_k$. Then

$$B_k(s_k) = A_k^{(0)} + \sum_{j=0}^{N} C_k^{(j+1)}D_k^{(j+1)}u_{kj} \quad (5.172)$$

and

$$\mathcal{B}_k(\mathbf{s}_k) = A_k^{(0)} + \sum_{j=0}^{N} \mathcal{C}_k^{(j+1)}[\mathcal{D}_k^{(j+1)}(\mathbf{u}_{kj})], \quad (5.173)$$

where $[\mathcal{D}_k^{(j+1)}(\mathbf{u}_{kj})](\omega) = D_k^{(j+1)}u_{kj}$, and $\mathcal{C}_k^{(j+1)}$ is defined similarly.

5.7.4 Orthogonality of random vectors

We recall that for any $\mathbf{x} \in L^2(\Omega,\mathbb{R}^m)$ and $\mathbf{y} \in L^2(\Omega,\mathbb{R}^n)$, we denote

$$E_{xy} = E[\mathbf{x}\mathbf{y}^T] = \left\{E[\mathbf{x}_i\mathbf{y}_j]\right\}_{i,j=1}^{n} \quad \text{and} \quad \mathbb{E}_{xy} = E[\mathbf{x}\mathbf{y}^T] - E[\mathbf{x}]E[\mathbf{y}^T],$$

where

$$E[\mathbf{x}_i\mathbf{y}_j] \stackrel{\text{def}}{=} \int_{\Omega} \mathbf{x}_i(\omega)\mathbf{y}_j(\omega)d\mu(\omega).$$

Definition 31. *Let* $\mathbf{u}_k \in L^2(\Omega, \mathbb{R}^n)$ *and* $\mathbf{v}_k = \mathcal{Z}_k(\mathbf{u}_k)$. *The operators* $\mathcal{Z}_1, \ldots, \mathcal{Z}_p$ *are called pairwise orthonormal if*

$$\mathbb{E}_{v_i v_j} = \begin{cases} \mathbb{O}, & i \neq j, \\ I, & i = j \end{cases} \qquad \text{for any } i, j = 1, \ldots, p.$$

Here, $\mathbb{O}$ *and* I *are the zero matrix and identity matrix, respectively.*

If

$$\mathbb{E}_{v_i v_j} = \mathbb{O} \quad \text{for} \quad i \neq j, \quad i, j = 1, \ldots, p,$$

and if $\mathbb{E}_{v_i v_j}$ *is not necessarily equal to* I *for* $i = j$ *then* $\mathcal{Z}_1, \ldots, \mathcal{Z}_p$ *are called pairwise orthogonal.*

The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$ *will also be called orthonormal and orthogonal, respectively.*

If M is a square matrix then we write $M^{1/2}$ for a matrix such that $M^{1/2} M^{1/2} = M$.

For the case when matrix $\mathbb{E}_{v_k v_k}$ is invertible for any $k = 1, \ldots, p$, the orthonormalization procedure is as follows. For $\mathbf{u}_k \in L^2(\Omega, \mathbb{R}^n)$, we write

$$[\mathcal{Z}_k(\mathbf{u}_k)](\omega) = Z_k \mathbf{u}_k(\omega), \tag{5.174}$$

where $Z_k \in \mathbb{R}^{n \times n}$. For $\mathbf{u}_k, \mathbf{v}_j, \mathbf{w}_j \in L^2(\Omega, \mathbb{R}^n)$, we also define operators $\mathcal{E}_{u_k v_j}, \mathcal{E}^{-1}_{v_j v_j} : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^n)$ by the equations

$$[\mathcal{E}_{u_k v_j}(\mathbf{w}_j)](\omega) = \mathbb{E}_{u_k v_j} \mathbf{w}_j(\omega) \tag{5.175}$$

and

$$[\mathcal{E}^{-1}_{v_j v_j}(\mathbf{w}_j)](\omega) = \mathbb{E}^{-1}_{v_j v_j} \mathbf{w}_j(\omega), \tag{5.176}$$

respectively.

Lemma 33. *Let*

$$\mathbf{w}_1 = \mathbf{u}_1 \quad \text{and} \quad \mathbf{w}_i = \mathbf{u}_i - \sum_{k=1}^{i-1} \mathcal{E}_{u_i w_k} \mathcal{E}^{-1}_{w_k w_k}(\mathbf{w}_k) \tag{5.177}$$

where $i = 1, \ldots, p$ *and* $\mathcal{E}^{-1}_{w_k w_k}$ *exists. Then*

(i) the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_p$ *are pairwise orthogonal, and*

(ii) the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$, *defined by*

$$\mathbf{v}_i = \mathcal{Z}_i(\mathbf{u}_i) \tag{5.178}$$

with

$$\mathcal{Z}_i(\mathbf{u}_i) = (\mathcal{E}^{1/2}_{w_i w_i})^{-1}(\mathbf{w}_i) \tag{5.179}$$

for $i = 1, \ldots, p$, *are pairwise orthonormal.*

Proof. Let us write

$$\mathbf{w}_1 = \mathbf{u}_1 \quad \text{and} \quad \mathbf{w}_i = \mathbf{u}_i - \sum_{k=1}^{i-1} \mathcal{U}_{ik}(\mathbf{w}_k) \quad \text{for} \quad i = 1, \ldots, p,$$

with $\mathcal{U}_{ik} : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^n)$ chosen so that, for $k = 1, \ldots, i-1$,

$$\mathbb{E}_{w_i w_k} = \mathbb{O} \quad \text{if} \quad i \neq k. \tag{5.180}$$

We wish (5.180) is true for any k, i.e.

$$\begin{aligned}
\mathbb{E}_{w_i w_k} &= E_{w_i w_k} - E[\mathbf{w}_i]E[\mathbf{w}_k^T] \\
&= E[(\mathbf{u}_i - \sum_{l=1}^{i-1} \mathcal{U}_{il}(\mathbf{w}_l))\mathbf{w}_k^T] - E[(\mathbf{u}_i - \sum_{l=1}^{i-1} \mathcal{U}_{il}(\mathbf{w}_l))]E[\mathbf{w}_k^T] \\
&= E_{u_i w_k} - U_{ik} E_{w_k w_k} - E[\mathbf{u}_i]E[\mathbf{w}_k^T] + E[\mathbf{w}_k]E[\mathbf{w}_k^T] \\
&= \mathbb{E}_{u_i w_k} - U_{ik} \mathbb{E}_{w_k w_k} \\
&= \mathbb{O}.
\end{aligned}$$

Thus, $U_{ik} = \mathbb{E}_{u_i w_k} \mathbb{E}_{w_k w_k}^{-1}$, and the statement (i) is true.

It is clear that vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$, defined by (5.178), are orthogonal. For $\mathcal{Z}_k$, defined by (5.179), we have $Z_k = (\mathbb{E}_{w_k w_k}^{1/2})^{-1}$ and

$$\begin{aligned}
\mathbb{E}_{v_k v_k} &= E[(\mathbb{E}_{w_k w_k}^{1/2})^{-1} \mathbf{w}_k \mathbf{w}_k^T (\mathbb{E}_{w_k w_k}^{1/2})^{-1}] \\
&\qquad - E[(\mathbb{E}_{w_k w_k}^{1/2})^{-1} \mathbf{w}_k] E[\mathbf{w}_k^T (\mathbb{E}_{w_k w_k}^{1/2})^{-1}] \\
&= (\mathbb{E}_{w_k w_k}^{1/2})^{-1} \mathbb{E}_{w_k w_k} (\mathbb{E}_{w_k w_k}^{1/2})^{-1} \\
&= I.
\end{aligned}$$

Hence, $\mathbf{v}_1, \ldots, \mathbf{v}_p$, defined by (5.178), are orthonormal. □

For the case when matrix $\mathbb{E}_{v_k v_k}$ is singular for $k = 1, \ldots, p$, the orthogonalizing operators $\mathcal{Z}_1, \ldots, \mathcal{Z}_p$ are determined by Lemma 34 below. Another difference from Lemma 33 is that the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$ in Lemma 34 are pairwise orthogonal but not orthonormal.

Lemma 34. *Let*

$$\mathbf{v}_i = \mathcal{Z}_i(\mathbf{u}_i)$$

for $i = 1, \ldots, p$, where $\mathcal{Z}_1, \ldots, \mathcal{Z}_p$ are such that

$$\mathcal{Z}_1(\mathbf{u}_1) = \mathbf{u}_1 \quad \textit{and} \quad \mathcal{Z}_i(\mathbf{u}_i) = \mathbf{u}_i - \sum_{k=1}^{i-1} \mathcal{Z}_{ik}(\mathbf{v}_k) \tag{5.181}$$

for $i = 2, \ldots, p$ *with* $\mathcal{Z}_{ik} : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^n)$ *defined by*

$$Z_{ik} = \mathbb{E}_{u_i v_k} \mathbb{E}^{\dagger}_{v_k v_k} + A_{ik}(I - \mathbb{E}_{v_k v_k} \mathbb{E}^{\dagger}_{v_k v_k}) \tag{5.182}$$

with $A_{ik} \in \mathbb{R}^{n \times n}$ *arbitrary. Then the vectors* $\mathbf{v}_1, \ldots, \mathbf{v}_p$ *are pairwise orthogonal.*

Proof. We wish that $\mathbb{E}_{v_i v_k} = \mathbb{O}$ for $i \neq k$. If Z_{ik} has been chosen so that this condition is true for all $k = 1, \ldots, i-1$ then we have

$$\begin{aligned} E[(\mathbf{u}_i - \sum_{l=1}^{i-1} \mathcal{Z}_{il}(\mathbf{v}_l))\mathbf{v}_k^T] &= \mathbb{E}_{u_i v_k} - \sum_{l=1}^{i-1} Z_{il} \mathbb{E}_{v_l v_k} \\ &= \mathbb{E}_{u_i v_k} - Z_{ik} \mathbb{E}_{v_k v_k} = \mathbb{O}. \end{aligned}$$

Thus,

$$Z_{ik} \mathbb{E}_{v_k v_k} = \mathbb{E}_{u_i v_k}. \tag{5.183}$$

The necessary and sufficient condition [6] for the solution of the matrix equation (5.183) is given by

$$\mathbb{E}_{u_i v_k} \mathbb{E}^{\dagger}_{v_k v_k} \mathbb{E}_{v_k v_k} = \mathbb{E}_{u_i v_k}. \tag{5.184}$$

By Lemma 24, (5.184) is true. Then, on the basis of [6], the general solution to (5.183) is given by (5.182). □

We note that Lemma 34 does not require invertibility of matrix $\mathbb{E}_{v_k v_k}$. At the same time, if $\mathbb{E}^{-1}_{v_k v_k}$ exists, then vectors $\mathbf{w}_1, \ldots, \mathbf{w}_p$ and $\mathbf{v}_1, \ldots, \mathbf{v}_p$ defined by (5.177) and Lemma 34 respectively, coincide.

Remark 16. *Orthogonalization of random vectors is not, of course, a new idea. In particular, generalizations of the Gram-Schmidt orthogonalization procedure have been considered in* [92, 49]. *The considered orthogonalization prcedures in Lemmata* 33 *and* 34 *are different from those in* [92, 49].

5.7.5 Reformulation of the problem

It follows from the above Section that random vectors $\mathbf{u}_{k0}, \ldots, \mathbf{u}_{kN} \in L^2(\Omega, \mathbb{R}^\nu)$ in (5.173) are always reduced to the pairwise orthogonal vectors $\mathbf{v}_{k0}, \ldots, \mathbf{v}_{kN}$ so that for $k = 0, 1, \ldots,$ and $j = 0, \ldots, N$,

$$\mathbf{v}_{kj} = \mathcal{D}_k^{(j+1)}(\mathbf{u}_{kj}), \tag{5.185}$$

where operators $\mathcal{D}_k^{(j+1)}$ are constructed from operators $\mathcal{Z}_{kj}$ given by Lemmata 33 and 34.

Hereinafter, $\mathcal{D}_k^{(j+1)}$ means the operator which converts $\mathbf{u}_{k0}, \ldots, \mathbf{u}_{kN}$ from (5.172) into the pairwise orthogonal vectors $\mathbf{v}_{k0}, \ldots, \mathbf{v}_{kN}$.

Without loss of generality we now also assume that all random vectors have zero mean.

As a result, the problem given by (5.167) and (5.168) is now presented in the following form.

Let $N = 1, \ldots, N$. For each $k = 0, 1, \ldots$, find $\tilde{A}^{(0)}$ and $\tilde{\mathcal{C}}_k^{(1)}, \ldots, \tilde{\mathcal{C}}_k^{(N+1)}$ such that

$$J(\tilde{A}_k^{(0)}, \tilde{\mathcal{C}}_k^{(1)}, \ldots, \tilde{\mathcal{C}}_k^{(N+1)}) = \min_{A_k^{(0)}, \mathcal{C}_k^{(1)}, \ldots, \mathcal{C}_k^{(N+1)}} J(A_k^{(0)}, \mathcal{C}_k^{(1)}, \ldots, \mathcal{C}_k^{(N+1)}), \tag{5.186}$$

where

$$J(A_k^{(0)}, \mathcal{C}_k^{(1)}, \ldots, \mathcal{C}_k^{(N+1)}) = E[\|\mathcal{F}(\mathbf{x}) - (A_k^{(0)} + \sum_{j=0}^{N} \mathcal{C}_k^{(j+1)}(\mathbf{v}_{kj}))\|^2] \tag{5.187}$$

and $\mathbf{v}_{k0}, \ldots, \mathbf{v}_{kN}$ are pairwise orthogonal vectors constructed from Lemma 34 and (5.185).

5.7.6 Solution of problem (5.186)

First, by Lemma 24 of Section 4.4.1, we have

$$E_{sv_{kj}} E^{\dagger}_{v_{kj}v_{kj}} E_{v_{kj}v_{kj}} = E_{sv_{kj}}. \tag{5.188}$$

Now we are in a position to solve the problem (5.186).

Theorem 42. *Let* $\mathbf{s} = \mathcal{F}(\mathbf{x})$. *The minimum in* (5.186) *is achieved if*

$$\begin{aligned} \tilde{A}_k^{(0)} &= \mathbb{O}_{p\times 1} \quad \textit{and} \\ \tilde{C}_k^{(j+1)} &= E_{sv_{kj}} E^{\dagger}_{v_{kj}v_{kj}} + M_{kj}(I - E_{v_{kj}v_{kj}} E^{\dagger}_{v_{kj}v_{kj}}) \end{aligned} \tag{5.189}$$

with M_{kj} *arbitrary,* $k = 0, 1, \ldots$ *and* $j = 0, \ldots, N$.

Proof. Let us denote $J = J(A_k^{(0)}, \mathcal{C}_k^{(1)}, \ldots, \mathcal{C}_k^{(N+1)})$. We have

$$\begin{aligned} J = \operatorname{tr}\{E_{ss} + A_k^{(0)} A_k^{(0)T} - \sum_{j=0}^{N} E_{sv_{kj}} E^{\dagger}_{v_{kj}v_{kj}} E_{v_{kj}s}\} \\ + \sum_{j=0}^{N} \operatorname{tr}\{(C_k^{(j+1)} - E_{sv_{kj}} E^{\dagger}_{v_{kj}v_{kj}}) E_{v_{kj}v_{kj}} \\ \times (C_k^{(j+1)T} - E^{\dagger}_{v_{kj}v_{kj}} E_{v_{kj}s})\} \end{aligned}$$

which is true on the basis of (5.188). Then

$$J = \|E_{ss}^{1/2}\|^2 + \|A_k^{(0)1/2}\|^2 - \sum_{j=0}^{N} \|E_{sv_{kj}}(E_{v_{kj}v_{kj}}^{\dagger})^{1/2}\|^2$$

$$+ \sum_{j=0}^{N} \|(C_k^{(j+1)} - E_{sv_{kj}}E_{v_{kj}v_{kj}}^{\dagger})E_{v_{kj}v_{kj}}^{1/2}\|^2. \tag{5.190}$$

Hence, J is minimized when $A_k^{(0)} = \mathbb{O}_{p\times 1}$ and

$$C_k^{(j+1)}E_{v_{kj}v_{kj}}^{1/2} - E_{sv_{kj}}E_{v_{kj}v_{kj}}^{\dagger}E_{v_{kj}v_{kj}}^{1/2} = \mathbb{O}_{p\times\nu}.$$

The solution to the latter equation is given in (5.189). □

Note that a possible and natural choice for M_{kj} in (5.189) is $M_{kj} = \mathbb{O}_{p\times\nu}$.

Remark 17. *The attractive feature of the solution presented in Theorem* 42 *is that* $\tilde{\mathcal{C}}_k^{(1)}, \ldots, \tilde{\mathcal{C}}_k^{(N+1)}$ *are subsequently determined from the sequence of the simple single independent equations* (5.189). *This is achieved by exploiting the orthogonalization procedure by Lemma 34. Otherwise matrices* $\tilde{C}_k^{(1)}, \ldots, \tilde{C}_k^{(N+1)}$ *which minimize* (5.168) *should be determined from a system of matrix equations. Such a solution would require substantially more computational work.*

Description of the algorithm. It follows from the above that the numerical realization of the operator $\mathcal{P}_k$ consists of the iteration procedure (5.162)–(5.164), (5.172) with the vector orthogonalization given by Lemma 34, and computation of the matrices $\tilde{C}_k^{(1)}, \ldots, \tilde{C}_k^{(N+1)}$ by (5.189) on each stage of the procedure (5.162)– (5.164), (5.172).

The device of numerical realization for the operator $\mathcal{P}_k$ is summarized as follows.

Initial parameters: $\mathbf{s} \in L^2(\Omega, \mathbb{R}^p)$, $\mathbf{y} \in L^2(\Omega, \mathbb{R}^n)$, $N, q \in \mathbb{N}$.

Final parameter: $P_q(y)$.

Algorithm:

- $s_0 := y$;
- for $k := 0$ to q do
  begin
  $v_{k0} := u_{k0}$;
  (here and below, u_{kj} is defined in accordance with (5.171))
  ◇ for $j := 0$ to N do

begin

$Z_{kjl} := E_{u_{kj}v_{kl}} E^{\dagger}_{v_{kl}v_{kl}} + K_{jl}(I - E_{v_{kl}v_{kl}} E^{\dagger}_{v_{kl}v_{kl}})$;

$v_{kj} := u_{kj} - \sum_{l=0}^{j-1} Z_{kjl}(v_{kl})$;

end;

$\tilde{C}_k^{(j+1)} := E_{sv_{kj}} E^{\dagger}_{v_{kj}v_{kj}} + M_{kj}(I - E_{v_{kj}v_{kj}} E^{\dagger}_{v_{kj}v_{kj}})$;

$\tilde{B}_k(s_k) := \sum_{j=0}^{N} \tilde{\mathcal{C}}_k^{(j+1)}(\mathbf{v}_{kj})$;

$s_k := \tilde{B}_k(s_k)$;

end;

- $P_q(y) := B_q(s_q)$.

The number of iterations, q can be determined by the stopping criterion: if $\|s_{k+1} - s_k\|^2$ is not less than $\|s_k - s_{k-1}\|^2$ then the algorithm should be stopped.

5.7.7 Error analysis associated with the operator $\mathcal{P}_k$

Theorem 43. *The error of $\mathcal{F}$ approximation by the operator $\mathcal{P}_k$ is*

$$E[\|\mathcal{F}(\mathbf{x}) - \mathcal{P}_k(\mathbf{y})\|^2] = \|E_{ss}^{1/2}\|^2 - \|E_{sy}E_{yy}^{\dagger 1/2}\|^2 - \sum_{j=0}^{k}\sum_{l=1}^{N} \|E_{sv_{jl}} E^{\dagger 1/2}_{v_{jl}v_{jl}}\|^2, \tag{5.191}$$

where $k = 0, 1, \ldots$.

Proof. We write

$$E[\|\mathbf{s} - \mathbf{s}_{k+1}\|^2] = J(\tilde{A}_k^{(0)}, \tilde{\mathcal{C}}_k^{(1)}, \ldots, \tilde{\mathcal{C}}_k^{(N+1)}).$$

It follows from (5.190) that

$$E[\|\mathbf{s} - \mathbf{s}_{k+1}\|^2] = \|E_{ss}^{1/2}\|^2 - \|E_{ss_k} E^{\dagger 1/2}_{s_k s_k}\|^2 - \Delta(s_k), \tag{5.192}$$

where

$$\Delta(s_k) = \sum_{j=1}^{N} \|E_{sv_{kj}} E^{\dagger 1/2}_{v_{kj}v_{kj}}\|^2.$$

For $k = 0$, the theorem follows from (5.192) directly. To prove that (5.191) is true for any $k = 1, 2, \ldots$, we denote

$$W_k = [C_k^{(1)} \ldots C_k^{(N+1)}]$$

and

$$w_k = [v_{k0}^T \ldots v_{kN}^T]^T$$

so that

$$W_k w_k = [\mathcal{W}_k(\mathbf{w}_k)](\omega)$$

where

$$\mathcal{W}_k : L^2(\Omega, \mathbb{R}^{\nu(N+1)}) \to L^2(\Omega, \mathbb{R}^p).$$

Then the minimum of the functional $E[\|\mathbf{s} - \mathcal{W}_k(\mathbf{w}_k)\|^2]$ is achieved for

$$\tilde{\mathcal{W}}_k = E_{sw_k} E^{\dagger}_{w_k w_k} + K_k(I - E_{w_k w_k} E^{\dagger}_{w_k w_k}) \tag{5.193}$$

with K_k arbitrary. We note that (5.193) is true since

$$E_{sw_k} = E_{sw_k} E^{\dagger}_{w_k w_k} E_{w_k w_k}$$

by Lemma 24. The associated error is

$$E[\|\mathbf{s} - \tilde{\mathcal{W}}_k(\mathbf{w}_k)\|^2] = \|E_{ss}^{1/2}\|^2 - \|E_{sw_k} E^{\dagger 1/2}_{w_k w_k}\|^2.$$

Therefore

$$\begin{aligned} E[\|\mathbf{s} - \mathbf{s}_{k+1}\|^2] &= E[\|\mathbf{s} - \tilde{\mathcal{W}}_k(\mathbf{w}_k)\|^2] \\ &= \|E_{ss}^{1/2}\|^2 - \|E_{sw_k} E^{\dagger 1/2}_{w_k w_k}\|^2 \\ &= \|E_{ss}^{1/2}\|^2 - \|E_{ss_k} E^{\dagger 1/2}_{s_k s_k}\|^2 - \Delta(s_k). \end{aligned} \tag{5.194}$$

Let us suppose that the theorem is true for $k = i - 1$. Then

$$\begin{aligned} E[\|\mathbf{s} - \mathbf{s}_i\|^2] &= \|E_{ss}^{1/2}\|^2 - \|E_{sw_{i-1}} E^{\dagger 1/2}_{w_{i-1} w_{i-1}}\|^2 \\ &= \|E_{ss}^{1/2}\|^2 - \|E_{sy} E^{\dagger 1/2}_{yy}\|^2 - \sum_{j=0}^{i-1} \Delta(s_j). \end{aligned} \tag{5.195}$$

It is easy to show that

$$\|E_{ss_i} E^{\dagger 1/2}_{s_i s_i}\|^2 = \|E_{sw_{i-1}} E^{\dagger 1/2}_{w_{i-1} w_{i-1}}\|^2.$$

Thus on the basis of (5.195),

$$E[\|\mathbf{s} - \mathbf{s}_{i+1}\|^2] = \|E_{ss}^{1/2}\|^2 - \|E_{ss_i} E^{\dagger 1/2}_{s_i s_i}\|^2 - \Delta(s_i) \tag{5.196}$$

$$= \|E_{ss}^{1/2}\|^2 - \|E_{sw_{i-1}} E^{\dagger 1/2}_{w_{i-1} w_{i-1}}\|^2 - \Delta(s_i) \tag{5.197}$$

$$= \|E_{ss}^{1/2}\|^2 - \|E_{sy} E^{\dagger 1/2}_{yy}\|^2 - \sum_{j=0}^{i-1} \Delta(s_j) - \Delta(s_i). \tag{5.198}$$

Hence the theorem is proven. □

Remark 18. *It follows from* (5.191) *that the error decreases if both the number of iterations* k *and the number* N *of the coefficient matrices* $\tilde{C}_k^{(1)}$, $\ldots$, $\tilde{C}_k^{(N+1)}$ *increases.*

Remark 19. *It follows from* (5.191) *that the accuracy of approximation associated with the considered method can be adjusted by a variation of the two degrees of freedom, namely the degree* r *of the operator* $\mathcal{B}_k$ *and the number of iterations* k.

Remark 20. *Another degree of freedom is a form of the polynomial* $\mathcal{B}_k(\mathbf{s}_k)$ *for each* k *in* (5.165). *For example,* $\mathcal{B}_k(\mathbf{s}_k)$ *can be chosen in the form of the Hadamard-quadratic polynomial considered in Section* 4.4.1.

Remark 21. *It follows from* (5.191) *that if* $C_k^{(j+1)} = \mathbb{O}$ *for all* $j = 1, \ldots, N$ *then for any* $k = 1, 2, \ldots,$

$$E[\|\mathcal{F}(\mathbf{x}) - \mathcal{P}_k(\mathbf{y})\|^2] = \|E_{ss}^{1/2}\|^2 - \|E_{sy}E_{yy}^{\dagger 1/2}\|^2. \tag{5.199}$$

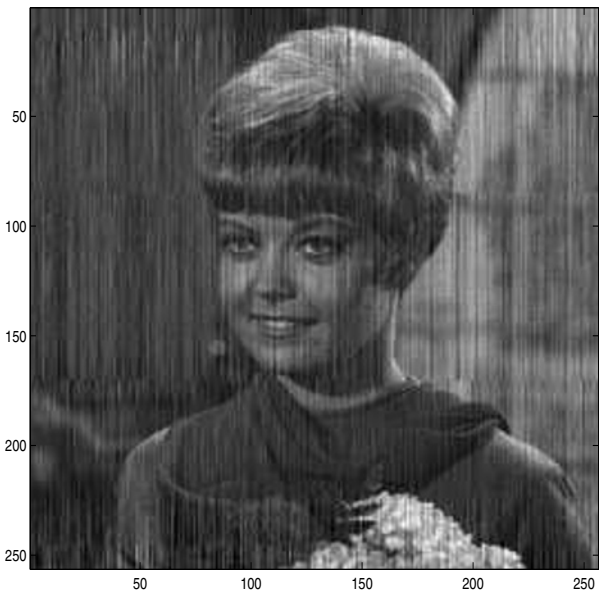
Since the right hand side in (5.199) *does not depend on* k, *this means that the error* $E[\|\mathcal{F}(\mathbf{x}) - \mathcal{P}_k(\mathbf{y})\|^2]$ *remains the same for any* k *if* $C_k^{(j+1)} = \mathbb{O}$ *for all* $j = 1, \ldots, N$. *In other words, nonlinearity of* $\mathcal{P}_k$, *which is implied by* $C_k^{(j+1)}$ *is an essential ingredient of the considered procedure. No improvement in accuracy of* $\mathcal{P}_{k+1}$ *over* $\mathcal{P}_k$ *can be reached if* $\mathcal{P}_{k+1}$ *is linear.*

Remark 22. *The idea of this method has been outlined but not justified in the reference* [156]. *The above method presented by* (5.162)–(5.164), (5.172), (5.189) *provides a broad generalization and substantial improvement both in the technique* [156] *and its modifications considered in* [156].

5.7.8 Simulations

To illustrate the performance of the considered approach, we use data from Section 4.4.1 (d).

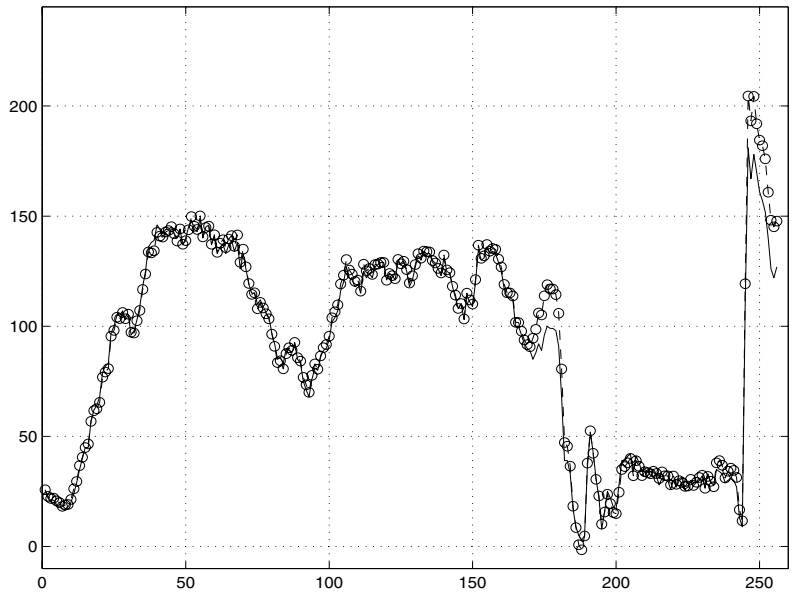
The considered method has been applied to each pair of matrices $W^{(i)}$, $V^{(i)}$ (see Section 4.4.1 (d)) separately to find the best approximation in the form (5.169), (5.172), (5.173), (5.181), (5.185) to the operator $F = F_i$ where $F_i : U^{(i)} \to V^{(i)}$.) To compare the method considered in this Section with the method of Sections 4.4.1(a)-(d) from which the data is used, B_k has been constructed from the Hadamard-quadratic polynomial (5.65), (5.81)–(5.83) for each k, in accordance with Remark 20. The input of the system



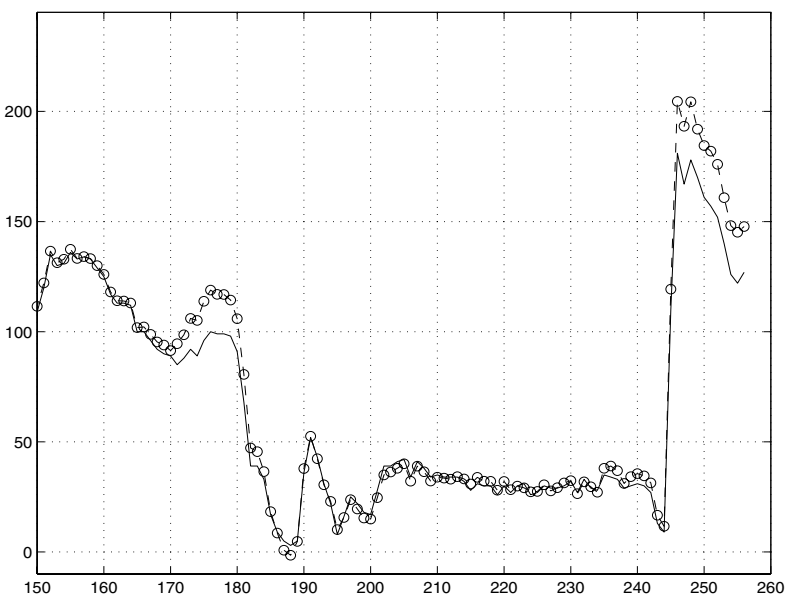
(a) Output of the model P_1.

(b) Output of the model P_{49}.

Figure 5.11: Illustration to the performance of the considered method. The digitized image has been taken from http://sipi.usc.edu/database/.



(a) Approximation (dashed line with circles) of the 205-th column (solid line) in matrix V by the procedure (5.169), (5.172), (5.173), (5.181), (5.185) with $k = 49$.



(b) A bigger scale of the fragment of sub-figure (a) above.

Figure 5.12: Illustration to the performance of the considered method.

B_k is a column of the matrix $U^{(i)}$ and the input of the approximating system is a column of the matrix $W^{(i)}$. Covariance matrices have been estimated by the known sample estimates formed by the matrices $W^{(i)}$, $V^{(i)}$ and $Z^{(i)} = V^{(i)} .* V^{(i)}$ for each $i = 1, 2, 3$.

We denote by $V_{k+1}^{(i)}$ the best approximation of $V^{(i)}$ (in the sense (5.186)) obtained in the $k-$th iteration of the procedure (5.169) with s and s_k substituted by $V^{(i)}$ and $V_k^{(i)}$ respectively.

Subfigures (a) and (b) in Fig. 5.11 are digitized images created from matrices $[V_2^{(1)^T}\ V_2^{(2)^T}\ V_2^{(3)^T}]^T$ and $[V_{50}^{(1)^T}\ V_{50}^{(2)^T}\ V_{50}^{(3)^T}]^T$ respectively.

In Fig. 5.12, we represent a plot of the 205-th column in the matrix V and plots of the 205-th column in matrices $[V_1^{(1)^T}\ V_1^{(2)^T}\ V_1^{(3)^T}]^T$ and $[V_{50}^{(1)^T}\ V_{50}^{(2)^T}\ V_{50}^{(3)^T}]^T$.

Fig. 5.12 (b) represents the part of Fig. 5.12 (a) in a bigger scale.

A comparison with Fig. 5.2 and Fig. 5.3 of Section 4.4.1(d) clearly demonstrates the efficiency of the method presented above.

5.8 Concluding Remarks

In this chapter, we have presented different approaches to the best constructive approximation of nonlinear operators and have given rigorous analysis of their properties. The major part of the chapter is devoted to approximating methods in probability spaces but in Section 5.2, we have also considered the best operator approximation technique for the so called "deterministic" case. In Sections 5.4–5.7, the specific methods for nonlinear operator approximation have been given. It is assumed that covariance matrices associated with those methods or their estimates are known. Therefore, in Section 5.3, some methods for a covariance matrix estimation have been considered. Applications to modelling of nonlinear systems have been discussed.

This page intentionally left blank

Part II

Optimal Estimation of Random Vectors

This page intentionally left blank

Chapter 6

Computational Methods for Optimal Filtering of Stochastic Signals

6.1 Introduction

In this chapter, we consider different approaches and computational methods for constructing mathematical models for optimal filtering of stochastic signals. In Sections 6.2–6.4, we give wide generalizations of the known Wiener filter to the cases when an associated *linear* operator is not invertible, noise is arbitrary, and the filter should satisfy conditions of causality and different types of memory. In Sections 6.5–6.6, we provide further generalizations of those approaches to the case of *nonlinear* models.

Throughout this chapter, we use the same notation for a probability space (Ω, Σ, μ) as in Chapter 4: Ω is the set of outcomes, Σ a σ–field of measurable subsets $\Delta \subset \Omega$ and $\mu : \Sigma \to [0,1]$ an associated probability measure on Σ with $\mu(\Omega) = 1$. Each element $\omega \in \Omega$ represents the outcome of an experiment and each subset Δ of Ω is a set of outcomes, called an event. We say that the event Δ has occurred if $\omega \in \Delta$.

6.2 Optimal Linear Filtering in Finite Dimensional Vector Spaces

The Wiener filtering problem has received a great deal of attention since the time when Wiener published his pioneering work [179]. In the next Sections 6.2–6.4, we present a wide generalization of the original Wiener problem and provide its solution.

6.2.1 Statement of the problem

Let $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$ and $\mathbf{y} \in L^2(\Omega, \mathbb{R}^n)$ be a reference stochastic signal and an observable data, respectively.

Similar to preceding chapters, for a matrix $A \in \mathbb{R}^{m\times n}$, we define a linear operator $\mathcal{A} : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^n)$ by the formula

$$[\mathcal{A}(\mathbf{y})](\omega) = A[\mathbf{y}(\omega)] \tag{6.1}$$

for all $\mathbf{y} \in L^2(\Omega, \mathbb{R}^n)$ and $\omega \in \Omega$, so that

$$\tilde{\mathbf{x}} = \mathcal{A}(\mathbf{y})$$

is an estimate of $\mathbf{x}$.

For any $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$, $\mathbf{y} \in L^2(\Omega, \mathbb{R}^n)$ and a continuous $\mathcal{A}$, let

$$J(A) = E\left[\|\mathbf{x} - \mathcal{A}(\mathbf{y})\|^2\right], \tag{6.2}$$

where

$$E\left[\|\mathbf{x} - \mathcal{A}(\mathbf{y})\|^2\right] = \int_\Omega \|\mathbf{x}(\omega) - [\mathcal{A}(\mathbf{y})](\omega)\|^2 d\mu(\omega) \tag{6.3}$$

with $\|\cdot\|$ the Euclidean norm.

The problem is to find a linear continuous operator $\mathcal{A}^0$ such that

$$J(A^0) = \min_{A\in\mathbb{R}^{m\times n}} J(A). \tag{6.4}$$

Here, $[\mathcal{A}^0(\mathbf{y})](\omega) = A^0[\mathbf{y}(\omega)]$.

The problem (6.4) is a generalization of the known Wiener filtering problem [179]. Unlike [179] it is not assumed that the operator $\mathcal{A}$ is invertible, and that $\mathbf{y} = \mathbf{x} + \xi$ where ξ is noise. Instead, it is assumed that

(i) $\mathbf{x}$ is unknown and no relationship between $\mathbf{x}$ and $\mathbf{y}$ is known except covariance matrices or their estimates formed from subvectors of $\mathbf{y}$ and $\mathbf{x}$, and

(ii) the operator $\mathcal{A}$ can be singular.

We note that that the assumption concerning covariance matrices is conventional for the known methods [9, 37, 60, 115, 142, 152, 153, 154, 155, 156, 157, 179] of the best operator approximation. The methods of a covariance matrix estimation can be found in Section 4.3.

6.2.2 Solution of the problem: optimal linear filter

Below, we give the solution to the problem (6.4) in terms of the pseudoinverse matrix $E_{yy}E_{yy}^{\dagger}$. This means that the solution always exists.

Theorem 44. *The operator $\mathcal{A}^0$ which satisfy* (6.4) *is determined by*

$$A^0 = E_{xy}E_{yy}^{\dagger} + M(I - E_{yy}E_{yy}^{\dagger}), \tag{6.5}$$

where $M \in \mathbb{R}^{m\times n}$ is an arbitrary matrix.

Proof. If we choose $A_0 = \mathbb{O}$ and $A_2 = \mathbb{O}$ in Theorem 34 of Section 5.4 then the proof follows directly from the proof of Theorem 34. □

We would like to point out that the model $\mathcal{A}^0$ is not unique due to an arbitrary M in (6.5). A natural practical choice for M is $M = \mathbb{O}$.

Numerical simulations associated with the model (6.5) are given in Section 6.6.

6.3 Optimal Linear Filtering in Hilbert Spaces

Let $\mathbf{u}$ be a random signal with realizations $\mathbf{u}(\omega) = x$ in an *infinite* dimensional vector space X for each outcome ω from the set of all possible outcomes. We seek an estimate of the signal $\mathbf{u}$ by observing an associated random signal $\mathbf{v}$ and we suppose that the outcome of the observed data signal $\mathbf{v}(\omega) = y$ is realized as an element of some *finite* dimensional subspace $Y \subseteq X$. Our goal is to find the best possible estimate $\hat{\mathbf{u}}$ of $\mathbf{u}$ using a linear filter on $\mathbf{v}$.

6.3.1 A generic example

For the beginning, we show that an elementary random signal is equivalent to a random vector with realizations in an infinite dimensional vector space. This generic example will be used later in the paper to illustrate the proposed optimal filter. It is well–known that a signal $x : [0, 1] \to \mathbb{R}$ for which

$$\int_0^1 [x(t)]^2 dt < \infty$$

can be represented by a Fourier sine series

$$x(t) = \sum_{k=1}^{\infty} x_k \sqrt{2} \sin k\pi t$$

or equivalently by an infinite dimensional vector

$$x = (x_1, x_2, x_3, \cdots)^T$$

where

$$\sum_{k=1}^{\infty} |x_k|^2 < \infty.$$

In this case we say that the vector $x \in X = l^2$. For the purpose of practical calculations with these signals it is necessary to use a suitable finite dimensional approximation. Thus we write

$$x \approx (x_1, x_2, \cdots, x_n, 0, 0, \cdots)^T$$

for some fixed value of n.

We can generate random vectors with realizations in an infinite dimensional Hilbert space by thinking of each coefficient x_k in the Fourier sine series as the realization of a real valued random variable. If Ω is the set of all possible outcomes and $\mathbf{u}_k : \Omega \to \mathbb{R}$ is a real valued random variable then $\mathbf{u}_k(\omega) = x_k \in \mathbb{R}$ and we obtain a realization

$$\mathbf{u}(\omega, t) = \sum_{k=1}^{\infty} u_k(\omega) \sqrt{2} \sin k\pi t$$

of the random signal $\mathbf{u}(\cdot, t)$, or equivalently a realization

$$\mathbf{u}(\omega) = (\mathbf{u}_1(\omega), \mathbf{u}_2(\omega), \mathbf{u}_3(\omega), \cdots)^T$$

of the infinite dimensional random vector $\mathbf{u}(\cdot)$ for each outcome $\omega \in \Omega$. For the above realizations to be meaningful it is of course necessary that

$$\sum_{k=1}^{\infty} |\mathbf{u}_k(\omega)|^2 < \infty$$

for almost all $\omega \in \Omega$. That is for all ω except possibly a set of measure zero. This statement is understood in terms of the associated probability measure $\mu : \Omega \to [0,1]$ where $\mu(\Delta) = P\{\omega \in \Delta\} \in [0,1]$ is well defined for each suitable event $\Delta \subseteq \Omega$.

6.3.2 Random vectors in Banach space

In this section, we outline a theoretical basis for the description of random vectors with realizations in Banach space. We follow the methods of Halmos [51], Dunford and Schwartz [31] and Yosida [185]. Although many of the results are natural extensions of the results for real–valued random variables, the extensions are nontrivial. This material is essential for a proper understanding of the expectation operator.

The Bochner integral

Suppose $E_j \in \Sigma$ $(j = 1, \ldots, n)$ are mutually disjoint sets and $\xi_j \in X$ (a Banach space) for $j = 1, 2, \ldots, n$. We may define a finitely–valued function $\mathbf{u} : \Omega \to X$ by

$$\mathbf{u}(\omega) = \sum_{j=1}^{n} \chi_j(\omega)\xi_j, \tag{6.6}$$

where $\chi_j : \Omega \to \{0,1\}$, the characteristic function of the set E_j, is given by

$$\chi_j(\omega) = \begin{cases} 1 & \omega \in E_j \\ 0 & \omega \notin E_j. \end{cases}$$

A function $\mathbf{u} : \Omega \to X$ is said to be strongly Σ–measurable if there exists a sequence $\{\mathbf{u}_n\}_{n\geq 1}$ of finitely–valued functions $\mathbf{u}_n : \Omega \to X$ such that

$$\|\mathbf{u}(\omega) - \mathbf{u}_n(\omega)\| \to 0 \quad \text{as} \quad n \to \infty$$

for almost all $\omega \in \Omega$. The value $\mathbf{u}(\omega)$ of a strongly Σ–measurable function $\mathbf{u}$ depends on the outcome $\omega \in \Omega$ of the experiment and we refer to u

as a random vector. When u is finitely–valued, the Bochner μ–integral $\mathcal{I}(\mathbf{u}) \in X$ is prescribed by

$$\mathcal{I}\left(\sum_{j=1}^{n} \chi_j \xi_j\right) = \sum_{j=1}^{n} \mu(E_j)\xi_j.$$

When $\mathbf{u}$ is strongly Σ–measurable, we say that $\mathbf{u}$ is Bochner μ–integrable if there exists a sequence $\{\mathbf{u}_n\}_{n\geq 1}$ of finitely–valued functions $\mathbf{u}_n : \Omega \to X$ with

$$\|\mathbf{u}_n(\omega) - \mathbf{u}(\omega)\| \to 0$$

for μ–almost all $\omega \in \Omega$ in such a way that

$$\int_\Omega \|\mathbf{u}_n(\omega) - \mathbf{u}(\omega)\|\mu(d\omega) \to 0$$

as $n \to \infty$. In this case the Bochner μ–integral is defined by

$$\mathcal{I}(\mathbf{u}) = \int_\Omega \mathbf{u}(\omega)\mu(d\omega),$$

where $\mathcal{I}(\mathbf{u}) \in X$ is the unique element with

$$\|\mathcal{I}(\mathbf{u}) - \mathcal{I}(\mathbf{u}_n)\| \to 0$$

as $n \to \infty$. In general, for each set $E \in \Sigma$, we define

$$\int_E \mathbf{u}(\omega)\mu(d\omega) = \int_\Omega \chi_E(\omega)\mathbf{u}(\omega)\mu(d\omega),$$

where χ is the characteristic function of the set E. The following general results can be found in Yosida [185].

Theorem 45. *A strongly Σ–measurable function $\mathbf{u}$ is Bochner μ–integrable if and only if $\|\mathbf{u}\|$ is μ–integrable.*

Corollary 7. *If $\|\mathbf{u}\|$ is μ–integrable, then*

$$\left\|\int_\Omega \mathbf{u}(\omega)\mu(d\omega)\right\| \leq \int_\Omega \|\mathbf{u}(\omega)\|\mu(d\omega).$$

Corollary 8. *Let X and Y be Banach spaces and $A \in \mathcal{L}(X, Y)$ a bounded linear map. If $\mathbf{u} : \Omega \to X$ is Bochner μ–integrable in X and if $\mathbf{v} = A[\mathbf{u}]$, then $\mathbf{v} : \Omega \to Y$ is Bochner μ–integrable on Y and*

$$\int_\Omega \mathbf{v}(\omega)\mu(d\omega) = A\left[\int_\Omega \mathbf{u}(\omega)\mu(d\omega)\right].$$

Definition 32. *Suppose X and Y are Banach spaces. Let $\mathbf{u} : \Omega \to X$ be a Bochner μ–integrable random vector in X. The expected value of $\mathbf{u}$ is defined by*

$$\mathcal{E}[\mathbf{u}] = \int_\Omega \mathbf{u}(\omega)\mu(d\omega).$$

We note from Corollary 7 that

$$\|\mathcal{E}[\mathbf{u}]\| \leq \mathcal{E}[\|\mathbf{u}\|].$$

When $A \in \mathcal{L}(X, Y)$ is a bounded linear map, it follows from Corollary 8 that

$$\mathcal{E}[A(\mathbf{u})] = A(\mathcal{E}[\mathbf{u}]).$$

6.3.3 Random vectors in Hilbert space

The theory of random vectors in Hilbert space is an extension of the theory of random vectors in Banach space. Of particular interest are properties relating to the scalar product. These properties are used directly in defining the special operators for the optimal filter.

Suppose X is a Hilbert space with scalar product $\langle \cdot, \cdot \rangle$. Let χ_j, E_j $(j = 1, \ldots, n)$ be as in Section 6.3.2 and let $\mathbf{u}$ be the finitely–valued random vector given by (6.6). Since

$$\|\mathbf{u}(\omega)\|^2 = \sum_{j=1}^{n} \chi_j(\omega)\|\xi_j\|^2,$$

it follows that if $A \in \mathcal{L}(X, X)$ is a bounded linear map, then

$$\begin{aligned}
\left\langle \int_\Omega \mathbf{u}(\omega)\mu(d\omega), \int_\Omega A[\mathbf{u}(\omega)]\mu(d\omega) \right\rangle &= \sum_{j=1}^{n}\sum_{k=1}^{n} \mu(E_j)\mu(E_k)\langle \xi_j, A[\xi_k]\rangle \\
&= \|A\| \sum_{j=1}^{n}\sum_{k=1}^{n} \mu(E_j)\mu(E_k)\|\xi_j\| \\
&\quad \times \|\xi_k\| \\
&\leq \|A\| \sum_{j=1}^{n} [\mu(E_j)]^2 \|\xi_j\|^2 \\
&\leq \|A\| \sum_{j=1}^{n} \mu(E_j)\|\xi_j\|^2 \\
&= \|A\| \int_\Omega \|\mathbf{u}(\omega)\|^2 \mu(d\omega).
\end{aligned}$$

By taking appropriate limits, we can easily extend the above argument to establish the following general results. These results are used to justify the construction of the optimal filter.

Theorem 46. *If* $\mathbf{u} : \Omega \to X$ *is strongly* Σ*–measurable and* $\|\mathbf{u}\|^2$ *is* μ*–integrable, then* $\mathbf{u}$ *is Bochner* μ*–integrable and for each bounded linear map* $A \in \mathcal{L}(X, X)$ *we have*

$$\left\langle \int_\Omega \mathbf{u}(\omega)\mu(d\omega), \int_\Omega A[\mathbf{u}(\omega)]\mu(d\omega) \right\rangle \leq \|A\| \int_\Omega \|\mathbf{u}(\omega)\|^2 \mu(d\omega).$$

Corollary 9. *If* $\mathbf{u} : \Omega \to X$ *is strongly* Σ*–measurable and* $\|\mathbf{u}\|^2$ *is* μ*–integrable, then*

$$\left\| \int_\Omega \mathbf{u}(\omega)\mu(d\omega) \right\|^2 \leq \int_\Omega \|\mathbf{u}(\omega)\|^2 \mu(d\omega).$$

The results of this subsection can be rewritten in terms of expected values. Let $A \in \mathcal{L}(X, X)$ and let $\mathbf{u} : \Omega \to X$ be a random vector. If $\|\mathbf{u}\|^2$ is μ–integrable, then

$$\langle \mathcal{E}[\mathbf{u}], \mathcal{E}[A(\mathbf{u})] \rangle \leq \|A\| \mathcal{E}[\|\mathbf{u}\|^2]$$

and in particular

$$\|\mathcal{E}[\mathbf{u}]\|^2 \leq \mathcal{E}[\|\mathbf{u}\|^2].$$

We write $L^2(\Omega, X)$ for the set of all strongly Σ–measurable functions $\mathbf{u} : \Omega \to X$ with $\mathcal{E}[\|\mathbf{u}\|^2] < \infty$.

6.3.4 Finite–dimensional maps on Hilbert space

In this section we review some basic structural results for bounded linear maps with finite–dimensional ranges on Hilbert space. These results are used directly in our construction of the optimal estimates. We assume that X is a separable Hilbert space and $Y \subseteq X$ is a finite–dimensional subspace, with dimension n, say. The material on Hilbert–Schmidt operators follows Balakrishnan [3].

We consider a bounded linear map $A \in \mathcal{L}(X, Y)$.

Let $\mathcal{R}(A) \subseteq Y$ denote the range space of A and suppose $\mathcal{R}(A)$ has dimension $r \leq n$. Let $\mathcal{N}(A) \subseteq X$ denote the null space of A. The bounded linear map $A^T : Y \to X$ is defined uniquely by the equation

$$\langle A^T(y), x \rangle = \langle y, A(x) \rangle$$

for each $y \in Y$.

We write $\mathcal{R}(A^T) \subseteq X$ for the range space of A^T. Since $\mathcal{R}(A)$ has dimension $r \leq n$, it follows that $\mathcal{R}(A^T)$ also has dimension r.

Let $\mathcal{N}(A^T) \subseteq Y$ denote the null space of A^T. Since $\mathcal{R}(A)$ is finite–dimensional and therefore closed, it follows that

$$Y = \mathcal{R}(A) \oplus \mathcal{N}(A^T)$$

and that each $y \in Y$ can be written uniquely in the form

$$y = y_{\mathcal{R}} + y_{\mathcal{N}},$$

where $y_{\mathcal{R}} \in \mathcal{R}(A)$ and $y_{\mathcal{N}} \in \mathcal{N}(A^T)$ and where $\langle y_{\mathcal{R}}, y_{\mathcal{N}} \rangle = 0$. In a similar fashion

$$X = \mathcal{R}(A^T) \oplus \mathcal{N}(A)$$

and each $x \in X$ can be written uniquely in the form

$$x = x_{\mathcal{R}} + x_{\mathcal{N}},$$

where $x_{\mathcal{R}} \in \mathcal{R}(A^T)$, $x_{\mathcal{N}} \in \mathcal{N}(A)$ and $\langle x_{\mathcal{R}}, x_{\mathcal{N}} \rangle = 0$.

The generalized inverse $A^\dagger \in \mathcal{L}(Y, X)$ of A is a bounded linear map defined as follows. Let $y \in Y$, put

$$y = y_{\mathcal{R}} + y_{\mathcal{N}}$$

and choose $x \in X$ such that $A(x) = y_{\mathcal{R}}$. Write

$$x = x_{\mathcal{R}} + x_{\mathcal{N}}$$

and define

$$A^\dagger(y) = x_{\mathcal{R}}.$$

The bounded linear operators

$$A^T A \in \mathcal{L}(X, X) \quad \text{and} \quad AA^T \in \mathcal{L}(Y, Y)$$

are positive–definite and self–adjoint. Since

$$A^T A : \mathcal{R}(A^T) \to \mathcal{R}(A^T) \quad \text{and} \quad AA^T : \mathcal{R}(A) \to \mathcal{R}(A),$$

we can find orthonormal vectors $\{e_i\}_{i=1}^r$ forming a basis for $\mathcal{R}(A^T)$ and $\{f_i\}_{i=1}^r$ forming a basis for $\mathcal{R}(A)$ which satisfy

$$A^T A e_i = {s_i}^2 e_i \qquad \text{and} \qquad AA^T f_i = {s_i}^2 f_i$$

for each $i = 1, 2, \ldots, r$. Here $s_1 > s_2 > \cdots > s_r > 0$ are real numbers and

$$f_i = \frac{1}{s_i} A e_i \qquad \text{and} \qquad e_i = \frac{1}{s_i} A^T f_i$$

for each $i = 1, 2, \ldots, r$. Because X is separable, the orthonormal sets

$$\{e_i\}_{i=1}^r \quad \text{and} \quad \{f_i\}_{i=1}^r$$

can be extended to form complete orthonormal sets

$$\{e_i\}_{i=1}^\infty \quad \text{and} \quad \{f_i\}_{i=1}^\infty$$

in X and the operators A and A^T are Hilbert–Schmidt operators because

$$\begin{aligned} \|A\|_{HS}^2 &= \sum_{i=1}^\infty \|Ae_i\|^2 \\ &= \sum_{i=1}^r {s_i}^2 \\ &< \infty \end{aligned}$$

and

$$\begin{aligned} \|A^T\|_{HS}^2 &= \sum_{i=1}^\infty \|A^T f_i\|^2 \\ &= \sum_{i=1}^r {s_i}^2 \\ &< \infty. \end{aligned}$$

It follows that the operators $A^T A$ and AA^T are nuclear operators with finite traces given by

$$\begin{aligned} \operatorname{tr}(A^T A) &= \sum_{i=1}^\infty \langle A^T A e_i, e_i \rangle \\ &= \sum_{i=1}^r {s_i}^2 \\ &< \infty \end{aligned}$$

and

$$\begin{aligned} \operatorname{tr}(AA^T) &= \sum_{i=1}^\infty \langle AA^T f_i, f_i \rangle \\ &= \sum_{i=1}^r {s_i}^2 \\ &< \infty. \end{aligned}$$

6.3.5 The correlation and covariance operators

Let X be a separable Hilbert space and suppose that Y is a finite–dimensional subspace of X.

To define the correlation and covariance operators we begin by considering an auxiliary mapping. For each $x \in X$, define a bounded linear map $J_x \in \mathcal{L}(\mathbb{R}, X)$ by

$$J_x(\alpha) = \alpha x.$$

The adjoint mapping ${J_x}^T \in \mathcal{L}(X, \mathbb{R})$ is given by

$${J_x}^T(y) = \langle x, y \rangle.$$

Now ${J_x}^T J_x \in \mathcal{L}(\mathbb{R}, \mathbb{R})$ satisfies

$$\begin{aligned} {J_x}^T J_x(\alpha) &= {J_x}^T(\alpha x) \\ &= \langle x, \alpha x \rangle \\ &= \alpha \|x\|^2 \end{aligned}$$

and clearly

$$\|{J_x}^T J_x\| = \|x\|^2.$$

On the other hand, $J_x {J_x}^T \in \mathcal{L}(X, X)$ is prescribed by

$$\begin{aligned} J_x {J_x}^T(y) &= J_x(\langle x, y \rangle) \\ &= \langle x, y \rangle x \end{aligned}$$

and hence, once again,

$$\|J_x {J_x}^T\| = \|x\|^2.$$

Let $\{e_i\}$ be a complete orthonormal set in X. We have

$$\begin{aligned} \sum_{i=1}^{\infty} \langle J_x {J_x}^T(e_i), e_i \rangle &= \sum_{i=1}^{\infty} \langle x, e_i \rangle^2 \\ &= \sum_{i=1}^{\infty} x_i^2 \\ &= |x|^2 \end{aligned}$$

and

$${J_x}^T J_x(1) = \|x\|^2.$$

Hence we see that ${J_x}^T J_x$ and $J_x {J_x}^T$ are both nuclear operators with finite trace given by

$$\begin{aligned} \mathrm{tr}({J_x}^T J_x) &= \mathrm{tr}(J_x {J_x}^T) \\ &= \|x\|^2. \end{aligned}$$

If $A \in \mathcal{L}(X,Y)$ and $B \in \mathcal{L}(Y,X)$ then

$$J_{Ax} = AJ_x \quad \text{and} \quad J_{By} = BJ_y$$

for all $x \in X$ and all $y \in Y$.

Let $\mathbf{u} : \Omega \to X$ and $\mathbf{v} : \Omega \to Y$ be random vectors with

$$\mathcal{E}[\|\mathbf{u}\|^2] < \infty \quad \text{and} \quad \mathcal{E}[\|\mathbf{v}\|^2] < \infty.$$

Lemma 35. *Suppose $q \in Y$ is a fixed vector. Then the vectors*

$$J_u {J_v}^T q : \Omega \to X \quad \textit{and} \quad J_v {J_v}^T q : \Omega \to Y$$

defined by

$$[J_u {J_v}^T q](\omega) = \langle \mathbf{v}(\omega), q\rangle \mathbf{u}(\omega) \quad \textit{and} \quad [J_v {J_v}^T q](\omega) = \langle \mathbf{v}(\omega), q\rangle \mathbf{v}(\omega)$$

for each $\omega \in \Omega$ are strongly Σ–measurable with

$$\mathcal{E}[\|J_u {J_v}^T q\|] < \infty \quad \textit{and} \quad \mathcal{E}[\|J_v {J_v}^T q\|] < \infty.$$

Proof. Let $\{\mathbf{u}_n\}$ and $\{\mathbf{v}_n\}$ be sequences of finitely–valued random vectors with

$$\|\mathbf{u}_n(\omega) - \mathbf{u}(\omega)\| \to 0 \quad \text{and} \quad \|\mathbf{v}_n(\omega) - \mathbf{v}(\omega)\| \to 0$$

as $n \to \infty$ for almost all $\omega \in \Omega$. Then $\{\langle \mathbf{v}_n(\omega), q\rangle \mathbf{u}_n(\omega)\}$ is a sequence of finitely–valued random vectors with

$$\begin{aligned}
&\|\langle \mathbf{v}_n(\omega), q\rangle \mathbf{u}_n(\omega) - \langle \mathbf{v}(\omega), q\rangle \mathbf{u}(\omega)\| \\
&\quad \le \|\langle \mathbf{v}_n(\omega) - \mathbf{v}(\omega), q\rangle \mathbf{u}_n(\omega)\| + \|\langle \mathbf{v}(\omega), q\rangle [\mathbf{u}_n(\omega) - \mathbf{u}(\omega)]\| \\
&\quad \le \|\mathbf{v}_n(\omega) - \mathbf{v}(\omega)\| \cdot \|q\| \cdot \|\mathbf{u}(\omega)\| + \|\mathbf{v}(\omega)\| \cdot \|q\| \cdot \|\mathbf{u}_n(\omega) - \mathbf{u}(\omega)\| \\
&\quad \to 0
\end{aligned}$$

as $n \to \infty$ for almost all $\omega \in \Omega$.

From the definition of strong measurability ([11], page 130) we see that $J_u {J_v}^T q$ is strongly Σ–measurable. Similarly $J_v {J_v}^T q$ is strongly Σ–measurable.

It follows that

$$\begin{aligned}
\|\mathcal{E}[J_u {J_v}^T q]\|^2 &\le (\mathcal{E}[\|\langle \mathbf{v}, q\rangle \mathbf{u}\|])^2 \\
&\le \|q\|^2 (\mathcal{E}[\|\mathbf{u}\| \cdot \|\mathbf{v}\|])^2 \\
&\le \|q\|^2 \mathcal{E}[\|\mathbf{u}\|^2] \mathcal{E}[\|\mathbf{v}\|^2] \\
&< \infty
\end{aligned}$$

and likewise that

$$\|\mathcal{E}[J_v {J_v}^T q]\|^2 < \infty.$$

This completes the proof. □

Definition 33. *The correlation operators*

$$\mathcal{E}[J_v {J_u}^T] \in \mathcal{L}(X,Y) \quad \textit{and} \quad \mathcal{E}[J_u {J_v}^T] \in \mathcal{L}(Y,X)$$

are defined by setting

$$\mathcal{E}[J_v {J_u}^T]p = \mathcal{E}[\langle \mathbf{u}, p\rangle \mathbf{v}] \quad \textit{and} \quad \mathcal{E}[J_u {J_v}^T]q = \mathcal{E}[\langle \mathbf{v}, q\rangle \mathbf{u}]$$

for each $p \in X$ and $q \in Y$.

It follows that

$$\begin{aligned} \langle p, \mathcal{E}[J_u {J_v}^T]q\rangle &= \langle p, \mathcal{E}[\langle \mathbf{v}, q\rangle \mathbf{u}]\rangle \\ &= \mathcal{E}[\langle p, \mathbf{u}\rangle \langle \mathbf{v}, q\rangle] \\ &= \langle \mathcal{E}[\langle p, \mathbf{u}\rangle \mathbf{v}], q\rangle \\ &= \langle \mathcal{E}[J_v {J_u}^T]p, q\rangle \end{aligned}$$

and hence

$$\mathcal{E}[J_u {J_v}^T]^T = \mathcal{E}[J_v {J_u}^T].$$

Definition 34. *The self–adjoint covariance operator $\mathcal{E}[J_v {J_v}^T] \in \mathcal{L}(Y,Y)$ is defined by setting*

$$\mathcal{E}[J_v {J_v}^T]q = \mathcal{E}[\langle \mathbf{v}, q\rangle \mathbf{v}]$$

for each $q \in Y$.

Note that since

$$\sum_{i=1}^{\infty} \langle \mathcal{E}[J_u {J_v}^T]e_i, e_i\rangle = \mathcal{E}[\langle u, v\rangle]$$

and

$$\sum_{i=1}^{\infty} \langle \mathcal{E}[J_v {J_v}^T]e_i, e_i\rangle = \mathcal{E}[\|\mathbf{v}\|^2]$$

it follows from Appendix C that $\mathcal{E}[J_u {J_v}^T]$ and $\mathcal{E}[J_v {J_v}^T]$ are both nuclear operators.

6.3.6 Statement of the problem

Suppose $\mathbf{u} \in L^2(\Omega, X)$ and $\mathbf{v} \in L^2(\Omega, Y)$. For each $F \in \mathcal{L}(Y,X)$, the linear transformation $\mathcal{M}_F \in \mathcal{L}(L^2(\Omega,Y), L^2(\Omega,X))$ is defined by

$$[\mathcal{M}_F \mathbf{v}](\omega) = F[\mathbf{v}(\omega)]$$

for each $\omega \in \Omega$. Once again it is customary to write Fv rather than $\mathcal{M}_F v$ since we then have $[F\mathbf{v}](\omega) = F[\mathbf{v}(\omega)] = F\mathbf{v}(\omega)$ for each $\omega \in \Omega$.

We wish to solve the following problem.

Let $\mathbf{u} \in L^2(\Omega, X)$ be a random vector and $\mathbf{v} \in L^2(\Omega, Y)$ an observable random vector. Suppose that $\mathcal{E}[J_u J_v{}^T]$ and $\mathcal{E}[J_v J_v{}^T]$ are known. Let $Q : \mathcal{L}(Y, X) \to \mathbb{R}$ be defined by

$$Q(F) = \mathcal{E}[\|\mathbf{u} - F\mathbf{v}\|^2] \tag{6.7}$$

for each $F \in \mathcal{L}(Y, X)$. We wish to find $\hat{F} \in \mathcal{L}(Y, X)$ such that

$$Q(\hat{F}) \leq Q(F) \tag{6.8}$$

for all F.

6.3.7 Solution to the problem (6.7)–(6.8)

Lemma 36. *The null space $\mathcal{N}(\mathcal{E}[J_v J_v{}^T])$ of the operator*

$$\mathcal{E}[J_v J_v{}^T] \in \mathcal{L}(Y, X)$$

is a subspace of the null space $\mathcal{N}(\mathcal{E}[J_u J_v{}^T])$ of the operator

$$\mathcal{E}[J_u J_v{}^T] \in \mathcal{L}(Y, X).$$

Proof. Suppose $q_{\mathcal{N}} \in \mathcal{N}(\mathcal{E}[J_v J_v{}^T])$. Then

$$\langle q_{\mathcal{N}}, \mathcal{E}[\langle \mathbf{v}, q_{\mathcal{N}} \rangle \mathbf{v}] \rangle = 0$$

and hence

$$\mathcal{E}[\langle \mathbf{v}, q_{\mathcal{N}} \rangle^2] = 0.$$

But for each $p \in X$ we have

$$\begin{aligned} |\langle p, \mathcal{E}[J_u J_v{}^T] q_{\mathcal{N}} \rangle| &= |\mathcal{E}[\langle p, \mathbf{u} \rangle \langle \mathbf{v}, q_{\mathcal{N}} \rangle]| \\ &\leq \left(\mathcal{E}[(\langle p, \mathbf{u} \rangle)^2]\right)^{1/2} \left(\mathcal{E}[(\langle \mathbf{v}, q_{\mathcal{N}} \rangle)^2]\right)^{1/2} \\ &= 0. \end{aligned}$$

Therefore $\mathcal{E}[J_u J_v{}^T] q_{\mathcal{N}} = 0$ and hence $q_{\mathcal{N}} \in \mathcal{N}(\mathcal{E}[J_u J_v{}^T])$. □

Corollary 10.

$$\mathcal{E}[J_u J_v{}^T] \mathcal{E}[J_v J_v{}^T]^\dagger \mathcal{E}[J_v J_v{}^T] = \mathcal{E}[J_u J_v{}^T].$$

Proof. If $q_{\mathcal{N}} \in \mathcal{N}(\mathcal{E}[J_v {J_v}^T])$, then

$$\mathcal{E}[J_u {J_v}^T]\mathcal{E}[J_v {J_v}^T]^\dagger \mathcal{E}[J_v {J_v}^T] q_{\mathcal{N}} = 0$$

and since the previous lemma shows that $q_{\mathcal{N}} \in \mathcal{N}(\mathcal{E}[J_u {J_v}^T])$, we have also

$$\mathcal{E}[J_u {J_v}^T] q_{\mathcal{N}} = 0.$$

On the other hand, if

$$q_{\mathcal{R}} \in \mathcal{R}(\mathcal{E}[J_v {J_v}^T]) = \mathcal{N}(\mathcal{E}[J_v {J_v}^T)^\perp,$$

then there exists $k \in Y$ such that

$$q_{\mathcal{R}} = \mathcal{E}[J_v {J_v}^T]^\dagger k.$$

Hence

$$\begin{aligned}
&\mathcal{E}[J_u {J_v}^T]\mathcal{E}[J_v {J_v}^T]^\dagger \mathcal{E}[J_v {J_v}^T] q_{\mathcal{R}} \\
&= \mathcal{E}[J_u {J_v}^T]\mathcal{E}[J_v {J_v}^T]^\dagger \mathcal{E}[J_v {J_v}^T]\mathcal{E}[J_v {J_v}^T]^\dagger k \\
&= \mathcal{E}[J_u {J_v}^T] q_{\mathcal{R}}.
\end{aligned}$$

The desired result follows from the fact that any element of Y can be written in the form $q = q_{\mathcal{N}} + q_{\mathcal{R}}$. □

Remark 23. *If $\mathbf{w} \in L^2(\Omega, X)$ and $z = \mathbf{w}(\omega)$ for some $\omega \in \Omega$ then the operators $J_z {J_z}^T$ and ${J_z}^T J_z$ are each nuclear operators and the trace is well defined. The trace is used in establishing the next identity and the subsequent theorem. It is therefore necessary to show that the operators concerned are nuclear operators. Nuclear operators are discussed by Dunford and Schwartz* [31], *and Yosida* [185].

Theorem 47. *The solution to problem (6.7)–(6.8) is given by $\hat{F} = F_K$ where*

$$F_K = F_0 + K[I - (\mathcal{E}[J_v {J_v}^T])^{1/2}(\mathcal{E}[J_v {J_v}^T]^\dagger)^{1/2}]$$

and

$$F_0 = \mathcal{E}[J_u {J_v}^T]\mathcal{E}[J_v {J_v}^T]^\dagger$$

and where $K \in \mathcal{L}(Y, X)$ is an arbitrary bounded linear operator.

The corresponding uniquely defined minimum value of $Q(F)$ is

$$\begin{aligned}
Q(F_K) &= Q(F_0) \\
&= \operatorname{tr}\{\mathcal{E}[J_u {J_u}^T] - \mathcal{E}[J_u {J_v}^T]\mathcal{E}[J_v {J_v}^T]^\dagger \mathcal{E}[J_v {J_u}^T]\}.
\end{aligned}$$

Proof. For each $F \in \mathcal{L}(Y,X)$ we know that $J_{u-Fv}{J_{u-Fv}}^T$ is a nuclear operator. It follows that

$$\begin{aligned} Q(F) &= \mathcal{E}[\|\mathbf{u}-F\mathbf{v}\|^2] \\ &= \mathcal{E}[\mathrm{tr}(J_{u-Fv}{J_{u-Fv}}^T)] \\ &= \mathrm{tr}(\mathcal{E}[(J_u-FJ_v)(J_u-FJ_v)^T]) \end{aligned}$$

and if we define $\Delta Q(F) = Q(F) - Q(F_0)$ then

$$\begin{aligned} \Delta Q(F) &= \mathrm{tr}\{F\mathcal{E}[J_v{J_v}^T]F^T - \mathcal{E}[J_u{J_v}^T]F^T - F\mathcal{E}[J_v{J_u}^T] \\ &\quad +\mathcal{E}[J_u{J_v}^T]\mathcal{E}[J_v{J_v}^T]^\dagger\mathcal{E}[J_v{J_u}^T]\} \\ &= \mathrm{tr}\{(F-\mathcal{E}[J_u{J_v}^T]\mathcal{E}[J_v{J_v}^T]^\dagger)\mathcal{E}[J_v{J_v}^T] \\ &\quad \times(F-\mathcal{E}[J_u{J_v}^T]\mathcal{E}[J_v{J_v}^T]^\dagger)^T\} \\ &= \|(F-\mathcal{E}[J_u{J_v}^T]\mathcal{E}[J_v{J_v}^T]^\dagger)(\mathcal{E}[J_v{J_v}^T])^{1/2}\|^2_{HS} \end{aligned}$$

where the norm is the Hilbert–Schmidt norm.

Hence

$$Q(F) - Q(F_0) \geq 0.$$

The minimum value is achieved if and only if

$$F = F_K$$

for some $K \in \mathcal{L}(X,Y)$. □

Corollary 11. *The best estimate $\hat{u}$ of $\mathbf{u}$ using a bounded linear operator on $\mathbf{v}$ is given by*

$$\hat{\mathbf{u}} = F_0\mathbf{v} + K[I - (\mathcal{E}[J_v{J_v}^T])^{1/2}(\mathcal{E}[J_v{J_v}^T]^\dagger)^{1/2}]\mathbf{v}$$

where $K \in \mathcal{L}(Y,X)$ is arbitrary. The minimum norm estimate is given by

$$\hat{\mathbf{u}} = F_0\mathbf{v}.$$

Example 18. *The generic example of Subsection* 6.3.1 *will be used to demonstrate construction of an optimal filter. In this example random signals are represented by infinite dimensional random vectors. We will show that the optimal filter can be represented by infinite dimensional matrices with suitable limits on the size of the matrix coefficients. Let $X = l^2$. Suppose that we wish to estimate the random signal*

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \end{pmatrix} : \Omega \to X$$

on the basis of an observed signal

$$\mathbf{v}(\omega) = \begin{pmatrix} \mathbf{v}_1(\omega) \\ \mathbf{v}_2(\omega) \\ \mathbf{v}_3(\omega) \\ \mathbf{v}_4(\omega) \\ 0 \\ \vdots \end{pmatrix}$$

with realizations in a four dimensional subspace $Y \subseteq X$. *We assume that* $\mathbf{v} = A\mathbf{u}$ *where*

$$A = \begin{pmatrix} A_{11} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{pmatrix} \quad \textit{and} \quad A_{11} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

and where $\mathbb{O}$ *is an infinite dimensional zero submatrix. Therefore* $\mathbf{v}_1 = \mathbf{u}_1 + \mathbf{u}_2$, $\mathbf{v}_2 = \mathbf{u}_2 + \mathbf{u}_3$, $\mathbf{v}_3 = \mathbf{u}_3 + \mathbf{u}_4$, $\mathbf{v}_4 = \mathbf{u}_4 + \mathbf{u}_1$ *and* $\mathbf{v}_k = 0$ *for all* $k \geq 5$.

To find the best estimate $\hat{u}$ *of* $\mathbf{u}$ *using a linear filter on* $\mathbf{v}$ *we need to define some special operators. For each*

$$\mathbf{u} : \Omega \to l^2 \quad \textit{and} \quad \mathbf{v} : \Omega \to l^2 \quad \textit{and each} \quad y \in Y,$$

the functions

$$J_u {J_v}^T y : \Omega \to X \quad \textit{and} \quad J_v {J_v}^T y : \Omega \to Y$$

are defined by

$$J_u {J_v}^T y(\omega) = \langle \mathbf{v}(\omega), y \rangle \mathbf{u}(\omega) \quad \textit{and} \quad J_v {J_v}^T y(\omega) = \langle \mathbf{v}(\omega), y \rangle \mathbf{v}(\omega)$$

for each $\omega \in \Omega$. *We suppose that the random variables* $\mathbf{u}_k$ *are are pairwise independent with*

$$\mathcal{E}[\mathbf{u}_k] = \rho_k \quad \textit{and} \quad \mathcal{E}[(\mathbf{u}_k - \rho_k)^2] = {\sigma_k}^2.$$

In practice this could occur as a deterministic signal with coefficients $\{\rho_k\}$ *and an additive noise term* $\xi_k = x_k - \rho_k$. *We also suppose that*

$$|\rho_k| \leq Rk^{-1}$$

for some fixed constant $R > 0$. We can now calculate

$$\mathcal{E}[J_u {J_v}^T y] = \begin{pmatrix} {\sigma_1}^2(y_1+y_4) \\ {\sigma_2}^2(y_1+y_2) \\ {\sigma_3}^2(y_2+y_3) \\ {\sigma_4}^2(y_3+y_4) \\ 0 \\ \vdots \end{pmatrix} + \begin{pmatrix} \rho_1[(\rho_1+\rho_2)y_1+(\rho_2+\rho_3)y_2+(\rho_3+\rho_4)y_3+(\rho_4+\rho_1)y_4] \\ \rho_2[(\rho_1+\rho_2)y_1+(\rho_2+\rho_3)y_2+(\rho_3+\rho_4)y_3+(\rho_4+\rho_1)y_4] \\ \rho_3[(\rho_1+\rho_2)y_1+(\rho_2+\rho_3)y_2+(\rho_3+\rho_4)y_3+(\rho_4+\rho_1)y_4] \\ \rho_4[(\rho_1+\rho_2)y_1+(\rho_2+\rho_3)y_2+(\rho_3+\rho_4)y_3+(\rho_4+\rho_1)y_4] \\ \rho_5[(\rho_1+\rho_2)y_1+(\rho_2+\rho_3)y_2+(\rho_3+\rho_4)y_3+(\rho_4+\rho_1)y_4] \\ \vdots \end{pmatrix}$$

and

$$\mathcal{E}[J_v {J_v}^T y] = \begin{pmatrix} ({\sigma_1}^2+{\sigma_2}^2)y_1+{\sigma_2}^2 y_2+{\sigma_1}^2 y_4 \\ {\sigma_2}^2 y_1+({\sigma_2}^2+{\sigma_3}^2)y_2+{\sigma_3}^2 y_3 \\ {\sigma_3}^2 y_2+({\sigma_3}^2+{\sigma_4}^2)y_3+{\sigma_4}^2 y_4 \\ {\sigma_1}^2 y_1+{\sigma_4}^2 y_3+({\sigma_4}^2+{\sigma_1}^2)y_4 \\ 0 \\ \vdots \end{pmatrix} + \begin{pmatrix} (\rho_1+\rho_2)[(\rho_1+\rho_2)y_1+(\rho_2+\rho_3)y_2+(\rho_3+\rho_4)y_3+(\rho_4+\rho_1)y_4] \\ (\rho_2+\rho_3)[(\rho_1+\rho_2)y_1+(\rho_2+\rho_3)y_2+(\rho_3+\rho_4)y_3+(\rho_4+\rho_1)y_4] \\ (\rho_3+\rho_4)[(\rho_1+\rho_2)y_1+(\rho_2+\rho_3)y_2+(\rho_3+\rho_4)y_3+(\rho_4+\rho_1)y_4] \\ (\rho_4+\rho_1)[(\rho_1+\rho_2)y_1+(\rho_2+\rho_3)y_2+(\rho_3+\rho_4)y_3+(\rho_4+\rho_1)y_4] \\ 0 \\ \vdots \end{pmatrix}$$

for all $y \in Y$. We are now able to write down a matrix representation for each of these operators. Note that these representations are essentially infinite matrices with some limit on the size of the matrix coefficients. In this case the size of the coefficients is limited by the inequality

$$\sum_{i,j=1}^{\infty} (\rho_i \rho_j)^2 < \frac{R\pi^4}{36}.$$

If we define

$$S = \begin{pmatrix} S_{11} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{pmatrix} \quad \textit{where} \quad S_{11} = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 \\ 0 & 0 & 0 & \sigma_4 \end{pmatrix} \quad \textit{and} \quad \rho = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \end{pmatrix}$$

then

$$\mathcal{E}[J_u {J_v}^T] = [SS^T + \rho\rho^T]A^T$$

and

$$\mathcal{E}[J_v {J_v}^T] = A[SS^T + \rho\rho^T]A^T.$$

We now show that the operator $\mathcal{E}[J_v {J_v}^T]$ is not invertible and calculate the generalized inverse. Define an orthogonal transformation

$$U_T = \begin{pmatrix} {U_{11}}^T & \mathbb{O} \\ \mathbb{O} & I \end{pmatrix} \quad \textit{where} \quad {U_{11}}^T = \begin{pmatrix} 1/2 & 1/2 & -1/2 & 1/2 \\ -1/2 & 1/2 & -1/2 & -1/2 \\ 1/2 & 1/2 & 1/2 & -1/2 \\ -1/2 & 1/2 & 1/2 & 1/2 \end{pmatrix}$$

and observe that

$$\begin{aligned} U\mathcal{E}[J_v {J_v}^T]U^T \quad = \quad & \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & \sum_{i=1}^4 {\sigma_i}^2 & {\sigma_4}^2 - {\sigma_2}^2 & {\sigma_1}^2 - {\sigma_3}^2 & 0 & \cdots \\ 0 & {\sigma_4}^2 - {\sigma_2}^2 & {\sigma_2}^2 + {\sigma_4}^2 & 0 & 0 & \cdots \\ 0 & {\sigma_1}^2 - {\sigma_3}^2 & 0 & {\sigma_1}^2 + {\sigma_3}^2 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \\ & + \begin{pmatrix} 0 \\ \sum_{i=1}^4 \rho_i \\ \rho_4 - \rho_2 \\ \rho_1 - \rho_3 \\ 0 \\ \vdots \end{pmatrix} \begin{pmatrix} 0 & \sum_{i=1}^4 \rho_i & \rho_4 - \rho_2 & \rho_1 - \rho_3 & 0 & \cdots \end{pmatrix}. \end{aligned}$$

Using an appropriate partition we can therefore write

$$U\mathcal{E}[J_v {J_v}^T]U^T = \begin{pmatrix} \mathbb{O} & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & P & \mathbb{O} \\ \mathbb{O} & \mathbb{O} & \mathbb{O} \end{pmatrix}$$

where

$$P = \begin{pmatrix} \sum_{i=1}^{4} {\sigma_i}^2 & {\sigma_4}^2 - {\sigma_2}^2 & {\sigma_1}^2 - {\sigma_3}^2 \\ {\sigma_4}^2 - {\sigma_2}^2 & {\sigma_2}^2 + {\sigma_4}^2 & 0 \\ {\sigma_1}^2 - {\sigma_3}^2 & 0 & {\sigma_1}^2 + {\sigma_3}^2 \end{pmatrix} + \begin{pmatrix} \sum_{i=1}^{4} \rho_i \\ \rho_4 - \rho_2 \\ \rho_1 - \rho_3 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{4} \rho_i & \rho_4 - \rho_2 & \rho_1 - \rho_3 \end{pmatrix}.$$

Since

$$\begin{vmatrix} \sum_{i=1}^{4} {\sigma_i}^2 & {\sigma_4}^2 - {\sigma_2}^2 & {\sigma_1}^2 - {\sigma_3}^2 \\ {\sigma_4}^2 - {\sigma_2}^2 & {\sigma_2}^2 + {\sigma_4}^2 & 0 \\ {\sigma_1}^2 - {\sigma_3}^2 & 0 & {\sigma_1}^2 + {\sigma_3}^2 \end{vmatrix} = \sum_{1 \le i < j < k \le 4} 4{\sigma_i}^2 {\sigma_j}^2 {\sigma_k}^2 > 0$$

it follows that P^{-1} *exists and*

$$\mathcal{E}[J_v {J_v}^T]^\dagger = U \begin{pmatrix} \mathbb{O} & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & P^{-1} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} & \mathbb{O} \end{pmatrix} U^T.$$

It has been shown that the best estimate $\hat{u}$ *of* $\mathbf{u}$ *using a linear filter on the observed signal* $\mathbf{v}$ *is given by*

$$\hat{u} = \mathcal{E}[J_u {J_v}^T] \mathcal{E}[J_v {J_v}^T]^\dagger v.$$

In this example we have seen that this filter can be easily computed. Since the filter involves an infinite dimensional matrix our implementation must necessarily be a truncation of the true optimal filter. This approximation can be made as accurate as we please.

6.4 Optimal Causal Linear Filtering with Piecewise Constant Memory

This section concerns the best constructive approximation of random vectors subject to a specialized minimization criterion criterion associated with the notion of piecewise-constant finite memory. The problem stems from an observation considered in Sections 4.1 and 4.2. A formulation of the problem is given in Section 6.4.4. The solution is provided in Section 6.4.7.

First, we need the following preliminary notation.

6.4.1 Preliminary notation

Let $\boldsymbol{\vartheta}, \boldsymbol{\varrho} : T \times \Omega \to \mathbb{R}$ where $T = \{t_k \mid k = 1, \ldots, n,\ t_1 \le \ldots \le t_n\} \subset \mathbb{R}$ and Ω is the set of outcomes in a probability space (Ω, Σ, μ).[1] The random

[1] The finite set T can be interpreted as a collection of time instants.

variables $\mathbf{x}_k : \Omega \to \mathbb{R}$ and $\mathbf{y}_k : \Omega \to \mathbb{R}$ are defined by

$$\mathbf{x}_k(\omega) = \boldsymbol{\vartheta}(t_k, \omega) \quad \text{and} \quad \mathbf{y}_k(\omega) = \boldsymbol{\varrho}(t_k, \omega)$$

for each $\omega \in \Omega$ where $\mathbf{x}_k$ and $\mathbf{y}_k$ are measurable functions on Ω for each $k - 1, 2, \ldots, n$. If $\mathbf{x}_k$ and $\mathbf{y}_k$ are square integrable for each $k = 1, 2, \ldots, n$ then the square integrable random vectors $\mathbf{x} \in L^2(\Omega, \mathbb{R}^n)$ and $\mathbf{y} \in L^2(\Omega, \mathbb{R}^n)$ are denoted by $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ and $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T$.

For each given $\omega \in \Omega$ we suppress the dependence on ω and write

$$x_k = \mathbf{x}_k(\omega), \quad y_k = \mathbf{y}_k(\omega), \quad x = \mathbf{x}(\omega), \quad y = \mathbf{y}(\omega), \tag{6.9}$$

$$x = [x_1, \ldots, x_n]^T \quad \text{and} \quad y = [y_1, \ldots, y_n]^T. \tag{6.10}$$

We write

$$\tilde{\mathbf{x}} = \mathcal{A}(\mathbf{y}),$$

where $\tilde{\mathbf{x}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n]^T$.

Next, let us partition $\tilde{\mathbf{x}}$ in such a way that

$$\tilde{\mathbf{x}} = [\tilde{\mathbf{u}}_1^T, \tilde{\mathbf{u}}_2^T, \ldots, \tilde{\mathbf{u}}_l^T]^T, \tag{6.11}$$

where

$$\tilde{\mathbf{u}}_i = [\tilde{\mathbf{x}}_{p_1+\ldots+p_{i-1}+1}, \ldots, \tilde{\mathbf{x}}_{p_1+\ldots+p_i}]^T, \quad i = 1, \ldots, l, \quad p_0 = 0,$$
$$\tilde{\mathbf{u}}_i \in L^2(\Omega, \mathbb{R}^{p_i}) \quad \text{and} \quad p_1 + \ldots + p_l = n.$$

6.4.2 The underlying problem

In many applications, similar to those presented in [9, 15, 37, 179], random vectors $\mathbf{y}$ and $\mathbf{x}$ are interpreted as observable data and reference vector, respectively. It is assumed that $\mathbf{y}$ depends on $\mathbf{x}$ and is contaminated with a random noise, and it is required to find A so that $\mathcal{A}(\mathbf{y})$ approximates $\mathbf{x}$ in the best possible way (usually, in terms of minimizing the mean square error). Moreover, to determine a best $\tilde{\mathbf{u}}_i$ in (6.11), the operator $\mathcal{A}$ may transform no more than m_i components $\mathbf{y}_{s_i}, \ldots, \mathbf{y}_{p_1+\ldots+p_i}$ of $\mathbf{y}$, where

$$m_i = (p_1 + \ldots + p_i) - s_i + 1, \quad s_i = q_i, q_i + 1, \ldots, (p_1 + \ldots + p_i), \tag{6.12}$$
$$q_i = 1, 2, \ldots, (p_1 + \ldots + p_i) \quad \text{and} \quad i = 1, \ldots, l.$$

Such an operator $\mathcal{A}$ is called the operator with piecewise-constant memory $\{m_1, \ldots, m_l\}$ (see Fig. 6.2 as an example).

The above constraint implies that the operator $\mathcal{A}$ and consequently the matrix A, must have a compatible structure. Essential conditions are that the components $\tilde{\mathbf{x}}_{p_1+\ldots+p_i}$ and $\mathbf{y}_{p_1+\ldots+p_i}$ have the same subscript and that s_i in (6.12) is different for each i, i.e., for each $\tilde{\mathbf{u}}_i$ in (6.11). This respectively

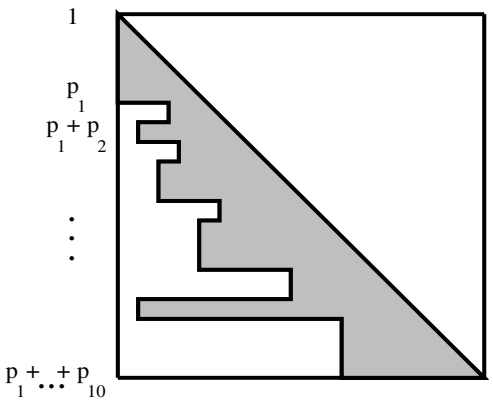


Figure 6.1: A lower stepped matrix A.

means that all entries above the diagonal of the matrix A are zeros and second, that for each i, there can be a zero-rectangular block in A from the left hand side of the diagonal.

An example of such a matrix A is given in Fig. 6.4 for $l = 10$ where the shaded part designates non-zero entries and non-shaded parts denote zero entries of A. The numbers in Fig. 6.4 denote numbers of rows of matrix A. For example, $p_1 + p_2$ denotes a $(p_1 + p_2)$-th row.

For lack of a better name, we will refer to A similar to that in Fig. 6.4 as the lower stepped matrix. We say that non-zero entries of the matrix A form a lower stepped part of A.

Such an unusual structure of the operator $\mathcal{A}$ makes the problem of finding the best $\mathcal{A}$ quite specific. This subject has a long history [9], but to the best of our knowledge, even for a much simpler structure of the operator $\mathcal{A}$ when $\mathcal{A}$ is defined by a lower triangular matrix, the problem of determining the best $\mathcal{A}$ has only been solved under the hard assumption of positive definiteness of an associated covariance matrix (see [9, 37, 115]). We avoid such an assumption and solve the problem in the general case of the lower stepped matrix (Theorem 1). The proposed technique is substantially different from those considered in [9, 37, 115].

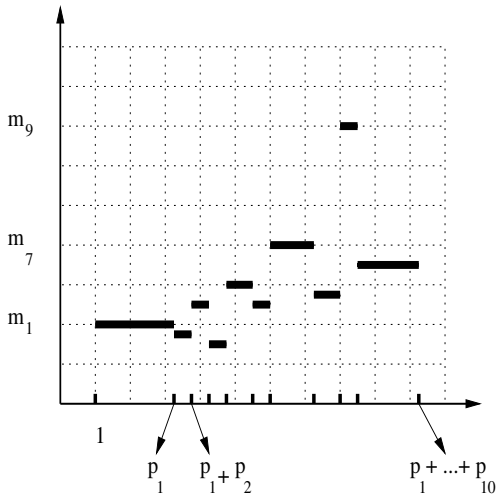


Figure 6.2: Illustration to piecewise memory associated with matrix in Fig. 1.

6.4.3 Linear causal operators with piecewise-constant memory

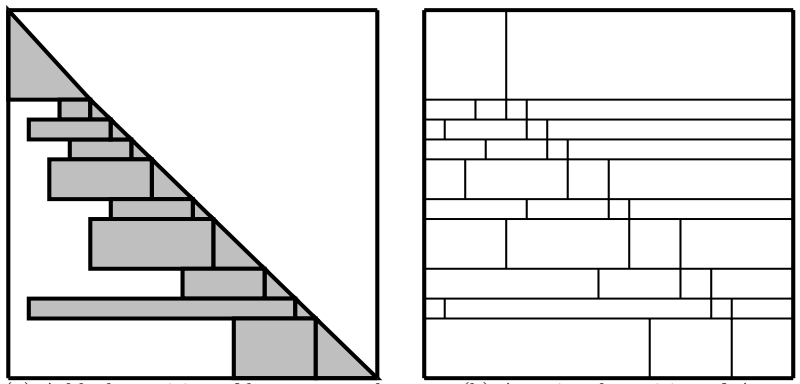
To define a linear causal[2] operators with piecewise-constant memory, we first need to formally define a lower stepped matrix. It is done below with a special partition of A in such a way that its lower stepped part consists from rectangular and lower triangular blocks as it is illustrated in Fig. 6.3(a). As before, the shaded parts in Fig. 6.3(a) designate non-zero entries and non-shaded parts denote zero entries of such a matrix. To realize such a representation, we need to choose a non-uniform partition of A in a form similar to that in Fig. 6.3(b), where a partition associated with the representation of the lower stepped part in the form of Fig. 6.3(a) is given. In Fig. 6.3(b), non-zero entries and zero entries are not allocated.

The block-matrix representation for $\mathcal{A}$ is as follows.

Let

$$A = \{A_{ij} \mid A_{ij} \in \mathbb{R}^{p_i \times q_{ij}}, i = 1, \ldots, l, j = 1, \ldots, 4\}, \tag{6.13}$$

[2]By the heuristic definition of causality, the present value of the reference vector estimate is not affected by future values of observable data containing the reference vector [117].



(a) A block-partition of lower stepped part of A.

(b) Associated partition of A.

Figure 6.3: Illustration to block-partition of lower stepped matrix A.

where

$$p_1 + \ldots + p_l = n \quad \text{and} \quad q_{i1} + \ldots + q_{i4} = n.$$

Let $\emptyset$, $\mathbb{O}_{ij} \in \mathbb{R}^{p_i \times q_{ij}}$, $L_{ij} \in \mathbb{R}^{p_i \times q_{ij}}$ and $R_{ij} \in \mathbb{R}^{p_i \times q_{ij}}$ be the empty block, zero block, lower triangular block and rectangular block, respectively.

We write $A = \begin{bmatrix} A_1 \\ \vdots \\ A_l \end{bmatrix}$, where $A_i = [A_{i1}, \ldots, A_{i4}]$ for each $i = 1, \ldots, l$. Here, A_i is called the block-row.

Now, let

$$A_1 = [\emptyset, \emptyset, L_{13}, \mathbb{O}_{14}], \quad A_i = [\mathbb{O}_{i1}, R_{i2}, L_{i3}, \mathbb{O}_{i4}] \text{ and } A_{l1} = [\mathbb{O}_{l1}, R_{l2}, L_{l3}, \emptyset],$$

where $i = 2, \ldots, l-1$.

For $i = 1, \ldots, l-1$, we also set

$$m(1) = q_{13}, \quad q_{i3} = p_i, \quad m(i+1) = q_{i+1,2} + p_{i+1} \tag{6.14}$$

$$\text{and} \quad q_{i+1,1} + q_{i+1,2} = q_{i,1} + m_i, \tag{6.15}$$

where $q_{11} = 0$. Then the matrix A is represented as follows:

$$A = \begin{bmatrix} L_{13} & & & \mathbb{O}_{14} & \\ \mathbb{O}_{21} & R_{22} & L_{23} & & \mathbb{O}_{24} \\ \vdots & \ddots & \ddots & & \vdots \\ \mathbb{O}_{l-1,1} & & R_{l-1,2} & L_{l-1,3} & \mathbb{O}_{l-1,4} \\ & \mathbb{O}_{l1} & & R_{l2} & L_{l3} \end{bmatrix} \tag{6.16}$$

Definition 35. *The matrix A given by* (6.16) *is called a lower stepped matrix. The set of lower stepped matrices is denoted by* $\mathbb{L}_m^n$.

Definition 36. *The linear operator* $\mathcal{A} : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^n)$ *is called a causal operator with piecewise-constant memory* $m = \{m(1), \ldots, m(l)\}$ *where*

$$m_i = \begin{cases} q_{13} & \text{if} \quad i = 1, \\ q_{i2} + q_{i3} & \text{if} \quad i = 2, \ldots, l, \end{cases} \tag{6.17}$$

if $\mathcal{A}$ *is defined by the lower stepped matrix* $A \in \mathbb{R}^{n \times n}$ *given by* (6.16). *The set of such operators is denoted by* $\mathbb{A}_m^n$.

6.4.4 Statement of the problem

As before, we write

$$J(A) = E[\|\mathbf{x} - \mathcal{A}(\mathbf{y})\|^2] \tag{6.18}$$

with $E[\|\mathbf{x} - \mathcal{A}(\mathbf{y})\|^2]$ defined by (6.3).

The problem is to find $\mathcal{A}^0 \in \mathbb{A}_m^n$ such that

$$J(A^0) = \min_{A \in \mathbb{L}_m^n} J(A) \tag{6.19}$$

for any $\mathbf{x}, \mathbf{y} \in L^2(\Omega, \mathbb{R}^n)$.

Here, $[\mathcal{A}^0(\mathbf{y})](\omega) = A^0[\mathbf{y}(\omega)]$ and $A \in \mathbb{L}_m^n$.

It is assumed that $\mathbf{x}$ is unknown and no relationship between $\mathbf{x}$ and $\mathbf{y}$ is known except covariance matrices or their estimates formed from sub-vectors of $\mathbf{y}$ and $\mathbf{x}$.

We note that the problem (6.19) is, in fact, the problem of finding the best approximation $\mathcal{A}^0$ to the identity mapping subject to $\mathcal{A} \in \mathbb{A}_m^n$.

6.4.5 Partition of x and y, and compatible representation of $\mathcal{A}(\mathbf{y})$

If $\mathbf{x} = \mathbf{y}$ than the solution is trivial: $\mathcal{A}^0$ is the identity mapping.

In general case, the solution of the problem (6.19) given below, consists of the following steps. First, vector $\mathbf{y}$ is partitioned in sub-vectors $\mathbf{v}_{13}, \mathbf{v}_{22}, \mathbf{v}_{23}, \ldots, \mathbf{v}_{l2}, \mathbf{v}_{l3}$ in a way which is compatible with the partition of matrix A in (6.16). Then the original problem can be represented as l independent problems (6.34)–(6.35). Second, to solve the problems (6.34)–(6.35), orthogonalization of sub-vectors $\mathbf{v}_{13}, \mathbf{v}_{22}, \mathbf{v}_{23}, \ldots, \mathbf{v}_{l2}, \mathbf{v}_{l3}$ is used. Finally, in Theorem 1, the solution of the original problem is derived in terms of matrices formed from orthogonalized sub-vectors.

We begin with partitions of $\mathbf{x}$ and $\mathbf{y}$.

Partitions of $\mathbf{x}$ and $\mathbf{y}$ which are compatible with the partition of matrix A above are as follows.

We write

$$x = [u_1^T, u_2^T, \dots, u_l^T]^T \quad \text{and} \quad \mathbf{x} = [\mathbf{u}_1^T, \mathbf{u}_2^T, \dots, \mathbf{u}_l^T]^T \tag{6.20}$$

where $u_1 \in \mathbb{R}^{p_1}$, $u_2 \in \mathbb{R}^{p_2}$, ..., $u_l \in \mathbb{R}^{p_l}$ are such that

$$u_1 = [x_1, \dots, x_{p_1}]^T, \quad u_2 = [x_{p_1+1}, \dots, x_{p_1+p_2}]^T, \quad \dots, \tag{6.21}$$

$$u_l = [x_{p_1+\dots+p_{l-1}+1}, \dots, x_{p_1+\dots+p_l}]^T, \tag{6.22}$$

and $\mathbf{u}_1 \in L^2(\Omega, \mathbb{R}^{p_1})$, $\mathbf{u}_2 \in L^2(\Omega, \mathbb{R}^{p_2})$, ..., $\mathbf{u}_l \in L^2(\Omega, \mathbb{R}^{p_l})$ are defined via $u_1, u_2, \dots, u_l$ similarly to (6.9).

Next, let

$$v_{11} = \emptyset, \quad v_{12} = \emptyset, \quad v_{13} = [y_1, \dots, y_{q_{13}}]^T \quad \text{and} \quad v_{14} = \emptyset.$$

For $i = 2, \dots, l-1$, we set

$$v_{i1} = [y_1, \dots, y_{q_{i1}}]^T, \quad v_{i2} = [y_{q_{i1}+1}, \dots, y_{q_{i1}+q_{i2}}]^T,$$
$$v_{i3} = [y_{q_{i1}+q_{i2}+1}, \dots, y_{q_{i1}+q_{i2}+q_{i3}}]^T, \quad v_{i4} = [y_{q_{i1}+q_{i2}+q_{i3}+1}, \dots, y_n]^T.$$

If $i = l$, then

$$v_{l1} = [y_1, \dots, y_{q_{l1}}]^T, \quad v_{l2} = [y_{q_{l1}+1}, \dots, y_{q_{l1}+q_{l2}}]^T,$$
$$v_{l3} = [y_{q_{l1}+q_{l2}+1}, \dots, y_n]^T, \quad v_{l4} = \emptyset.$$

Therefore

$$Ay = \begin{bmatrix} L_{13}v_{13} \\ R_{22}v_{22} + L_{23}v_{23} \\ \vdots \\ R_{l2}v_{l2} + L_{l3}v_{l3} \end{bmatrix} \text{ and } \mathcal{A}(\mathbf{y}) = \begin{bmatrix} \mathcal{L}_{13}(\mathbf{v}_{13}) \\ \mathcal{R}_{22}(\mathbf{v}_{22}) + \mathcal{L}_{23}(\mathbf{v}_{23}) \\ \vdots \\ \mathcal{R}_{l2}(\mathbf{v}_{l2}) + \mathcal{L}_{l3}(\mathbf{v}_{l3}) \end{bmatrix} \tag{6.23}$$

where $\mathcal{L}_{ij}$ and $\mathcal{R}_{ij}$ are defined via L_{ij} and R_{ij} respectively, in the manner of $\mathcal{A}$ defined via A by (6.1). The vector $\mathbf{v}_{ij} \in L^2(\Omega, \mathbb{R}^{q_{ij}})$ are defined similarly to those in (6.9).

Now, we can represent $J(A)$ given by (6.18), in the form

$$J(A) = J_1(L_{13}) + \sum_{i=2}^{l} J_i(R_{i2}, L_{i3}) \tag{6.24}$$

where

$$J_1(L_{13}) = E\left[\|\mathbf{u}_1 - \mathcal{L}_{13}(\mathbf{v}_{13})\|^2\right]$$

and

$$J_i(R_{i2}, L_{i3}) = E\left[\|\mathbf{u}_i - [\mathcal{R}_{i2}(\mathbf{v}_{i2}) + \mathcal{L}_{i3}(\mathbf{v}_{i3})]\|^2\right]. \tag{6.25}$$

We note that matrix A can be represented so that

$$Ay = BPy,$$

where $B \in \mathbb{R}^{n\times q}$ and $P \in \mathbb{R}^{q\times n}$ with $q = q_{13} + \sum_{i=1}^{l}(q_{i2} + q_{i3})$ are such that

$$B = \begin{bmatrix} L_{13} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & R_{22} & L_{23} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} & \mathbb{O} & R_{32} & L_{33} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & & \vdots & \mathbb{O} \\ \mathbb{O} & \dots & \dots & \dots & \mathbb{O} & R_{l-1,2} & L_{l-1,3} & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & \dots & \dots & \dots & \mathbb{O} & \mathbb{O} & \mathbb{O} & R_{l2} & L_{l3} \end{bmatrix} \tag{6.26}$$

and $Py = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_{l-1} \\ v_l \end{bmatrix}$. Here, $\mathbb{O}$ is the zero block, $v_1 = v_{13}$ and $v_i = \begin{bmatrix} v_{i2} \\ v_{i3} \end{bmatrix}$ for $i = 2, \dots, l-1$. The size of each zero block is such that BPy is represented in the form (6.23). The matrix B consists of $l \times (2l-1)$ blocks. The vector $v = Py$ consists of $2l-1$ sub-vectors v_{13}, v_{22}, v_{23}, …, v_{l2}, v_{l3}.

The operator $\mathcal{A}$ can be written as

$$\mathcal{A}(\mathbf{y}) = \mathcal{B}\mathcal{P}(\mathbf{y})$$

where

$$[\mathcal{B}(\mathbf{v})](\omega) = B[(\mathbf{v})(\omega)], \quad \mathbf{v} = \mathcal{P}(\mathbf{y}) \text{ and } [\mathcal{P}(\mathbf{y})](\omega) = P[(\mathbf{y})(\omega)]. \tag{6.27}$$

6.4.6 A representation of the approximator

We recall that for any $\mathbf{x}, \mathbf{y} \in L^2(\Omega, \mathbb{R}^n)$, we denote

$$E_{xy} = E[\mathbf{x}\mathbf{y}^T] = \left\{E[\mathbf{x}_i\mathbf{y}_j]\right\}_{i,j=1}^{n} \text{ where } E[\mathbf{x}_i\mathbf{y}_j] \stackrel{\text{def}}{=} \int_\Omega \mathbf{x}_i(\omega)\mathbf{y}_j(\omega)d\mu(\omega).$$

In (6.24), the terms $J_1(L_{13})$ and $J_i(R_{i2}, L_{i3})$ is defined by the operators $\mathcal{L}_{13}$, $\mathcal{R}_{i2}$ and $\mathcal{L}_{i3}$ and their action on the random block-vectors $\mathbf{v}_{13}$, $\mathbf{v}_{i2}$ and

$\mathbf{v}_{i3}$ respectively. The corresponding mutually orthogonal random vectors are

$$\mathbf{w}_{13} = \mathbf{v}_{13}, \quad \mathbf{w}_{i2} = \mathbf{v}_{i2} \quad \text{and} \quad \mathbf{w}_{i3} = \mathbf{v}_{i3} - \mathcal{Z}_i(\mathbf{v}_{i2}) \quad i = 2, \ldots, l, \tag{6.28}$$

where the operator $\mathcal{Z}_i : L^2(\Omega, \mathbb{R}^{q_{i2}}) \to L^2(\Omega, \mathbb{R}^{q_{i3}})$ is defined by the matrix

$$Z_i = E_{v_{i3}v_{i2}} E^{\dagger}_{v_{i2}v_{i2}} + M_i(I - E_{v_{i2}v_{i2}} E^{\dagger}_{v_{i2}v_{i2}}) \tag{6.29}$$

with $M_i \in \mathbb{R}^{q_{i3} \times q_{i2}}$ arbitrary (See Section 4.4.4(d)) .

We write $\mathbf{w}(\omega) = \begin{bmatrix} \mathbf{w}_{13}(\omega) \\ \mathbf{w}_{22}(\omega) \\ \mathbf{w}_{23}(\omega) \\ \mathbf{w}_{32}(\omega) \\ \mathbf{w}_{33}(\omega) \\ \vdots \\ \mathbf{w}_{l2}(\omega) \\ \mathbf{w}_{l3}(\omega) \end{bmatrix}$ and

$$Z = \begin{bmatrix} I_{13} & \mathbb{0} & \mathbb{0} & \mathbb{0} & \mathbb{0} & \mathbb{0} & \ldots & \mathbb{0} \\ \mathbb{0} & I_{22} & \mathbb{0} & \mathbb{0} & \mathbb{0} & \mathbb{0} & \ldots & \mathbb{0} \\ \mathbb{0} & -Z_2 & I_{23} & \mathbb{0} & \mathbb{0} & \mathbb{0} & \ldots & \mathbb{0} \\ \mathbb{0} & \mathbb{0} & \mathbb{0} & I_{32} & \mathbb{0} & \mathbb{0} & \ldots & \mathbb{0} \\ \mathbb{0} & \mathbb{0} & \mathbb{0} & -Z_3 & I_{33} & \mathbb{0} & \ldots & \mathbb{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & & \vdots \\ \mathbb{0} & \ldots & \ldots & \ldots & \mathbb{0} & \mathbb{0} & I_{l2} & \mathbb{0} \\ \mathbb{0} & \ldots & \ldots & \ldots & \mathbb{0} & \mathbb{0} & -Z_l & I_{l3} \end{bmatrix}$$

where I_{ij} is $q_{ij} \times q_{ij}$ identity matrix for $i = 1, \ldots, l$ and $j = 2, 3$, and Z_i is defined by (6.29) for $i = 2, \ldots, l$. The matrix Z consists of $(2l-1) \times (2l-1)$ blocks.

Then (6.28) can be written in the matrix form as

$$\mathbf{w}(\omega) = Z\mathbf{v}(\omega)$$

with $\mathbf{v}$ given by (6.27). Matrix Z implies the operator $\mathcal{Z} : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R})$ defined in the manner of (6.1).

Since Z is invertible, we can represent $\mathcal{A}$ as follows:

$$\mathcal{A}(\mathbf{y}) = \mathcal{K}[\mathcal{Z}(\mathcal{P}(\mathbf{y}))] \quad \text{where} \quad \mathcal{K} = \mathcal{B}\mathcal{Z}^{-1}. \tag{6.30}$$

A matrix representation of $\mathcal{K}$ is

$$K = \begin{bmatrix} L_{13} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & T_2 & L_{23} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} & \mathbb{O} & T_3 & L_{33} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & & \vdots & \vdots \\ \mathbb{O} & \dots & \dots & \dots & \mathbb{O} & T_{l-1} & L_{l-1,3} & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & \dots & \dots & \dots & \mathbb{O} & \mathbb{O} & \mathbb{O} & T_l & L_{l3} \end{bmatrix}$$

where

$$T_i = R_{i2} + L_{i3} Z_i \tag{6.31}$$

for $i = 2, \dots, l$. We note that K consists of $l \times (2l-1)$ blocks.

As a result, in (6.25),

$$\begin{aligned} R_{i2}\mathbf{v}_{i2}(\omega) + L_{i3}\mathbf{v}_{i3}(\omega) &= R_{i2}\mathbf{w}_{i2}(\omega) + L_{i3}[\mathbf{w}_{i3}(\omega) + Z_i\mathbf{w}_{i2}(\omega)] \\ &= T_i\mathbf{w}_{i2}(\omega) + L_{i3}\mathbf{w}_{i3}(\omega) \end{aligned}$$

and hence

$$J(A) = J_1(L_{13}) + \sum_{i=2}^{l} \mathcal{J}_i(T_i, L_{i3}), \tag{6.32}$$

where

$$\mathcal{J}_i(T_i, L_{i3}) = E[\|\mathbf{u}_i - [\mathcal{T}_i\mathbf{w}_{i2}(\omega) + \mathcal{L}_{i3}\mathbf{w}_{i3}]\|^2] \tag{6.33}$$

with $\mathcal{T}_i$ defined by

$$[\mathcal{T}_i\mathbf{w}_{i2}](\omega) = T_i[\mathbf{w}_{i2}(\omega)]$$

for all $i = 2, \dots, l$.

6.4.7 Solution of the problem (6.19)

Lemma 37. *For $A \in \mathbb{L}_m^n$, the following is true:*

$$\min_{A \in \mathbb{L}_m^n} J(A) = \min_{L_{13}} J_1(L_{13}) + \sum_{i=2}^{l} \min_{T_i, L_{i3}} \mathcal{J}_i(T_i, L_{i3}) \tag{6.34}$$

$$= \min_{L_{13}} J_1(L_{13}) + \sum_{i=2}^{l} \min_{R_{i2}, L_{i3}} J_i(R_{i2}, L_{i3}). \tag{6.35}$$

Proof. Let A^0 be a solution to the problem (6.19). Let us partition matrix A^0 similarly to (6.16) so that

$$A^0 = \begin{bmatrix} L^0_{13} & & & \mathbb{O}_{14} & \\ \mathbb{O}_{21} & R^0_{22} & L^0_{23} & & \mathbb{O}_{24} \\ \vdots & \ddots & \ddots & & \vdots \\ \mathbb{O}_{l-1,1} & & R^0_{l-1,2} & L^0_{l-1,3} & \mathbb{O}_{l-1,4} \\ & \mathbb{O}_{l1} & & R^0_{l2} & L^0_{l3} \end{bmatrix} \tag{6.36}$$

and set

$$K^0 = \begin{bmatrix} L^0_{13} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & T^0_2 & L^0_{23} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} & \mathbb{O} & T^0_3 & L^0_{33} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & & \vdots & \vdots \\ \mathbb{O} & \dots & \dots & \dots & \mathbb{O} & T^0_{l-1} & {}^0L_{l-1,3} & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & \dots & \dots & \dots & \mathbb{O} & \mathbb{O} & \mathbb{O} & T^0_l & L^0_{l3} \end{bmatrix},$$

where $T^0_i = R^0_{i2} + L^0_{i3} Z_i$ for $i = 2, \dots, l$.

Then

$$\begin{aligned} \min_{A \in \mathbb{L}^n_m} J(A) &= J_1(L^0_{13}) + \sum_{i=2}^{l} \mathcal{J}_i(T^0_i, L^0_{i3}) \\ &\geq \min_{L_{13}} J_1(L_{13}) + \sum_{i=2}^{l} \min_{T_i, L_{i3}} \mathcal{J}_i(T_i, L_{i3}) \end{aligned} \tag{6.37}$$

because L^0_{13}, T^0_i and L^0_{i3} are fixed.

Next, let L^*_{13}, T^*_i and L^*_{i3} be such that

$$J_1(L^*_{13}) = \min_{L_{13}} J_1(L_{13}) \quad \text{and} \quad \mathcal{J}_i(T^*_i, L^*_{i3}) = \min_{T_i, L_{i3}} \mathcal{J}_i(T_i, L_{i3}).$$

Then

$$\begin{aligned} &\min_{L_{13}} J_1(L_{13}) + \sum_{i=2}^{l} \min_{T_i, L_{i3}} \mathcal{J}_i(T_i, L_{i3}) \\ &= J_1(L^*_{13}) + \sum_{i=2}^{l} \mathcal{J}_i(T^*_i, L^*_{i3}) \\ &= J(A^*) \\ &\geq \min_{A \in \mathbb{L}^n_m} J(A), \end{aligned} \tag{6.38}$$

where $A^* = \begin{bmatrix} L^0_{13} & & & \mathbb{O}_{14} & \\ \mathbb{O}_{21} & R^*_{22} & L^*_{23} & & \mathbb{O}_{24} \\ \vdots & \ddots & \ddots & & \vdots \\ \mathbb{O}_{l-1,1} & & R^*_{l-1,2} & L^*_{l-1,3} & \mathbb{O}_{l-1,4} \\ & \mathbb{O}_{l1} & & R^*_{l2} & L^*_{l3} \end{bmatrix}$. The inequalities (6.37) and (6.38) imply the equality (6.34). The statement (6.35) follows from (6.32) and (6.24) □

Let us denote the Frobenius norm by $\|\cdot\|_F$.

Lemma 38. *If $A \in \mathbb{R}^{s\times s}$ and $A = B + C$ where $b_{ij}c_{ij} = 0$ for all i, j then*

$$\|A\|^2_F = \|B\|^2_F + \|C\|^2_F.$$

Proof. The proof is obvious. □

Now, we are in the position to prove the main result given in Theorem 1 below. To this end, we use the following notation.

For $i = 1, \ldots, l$, let λ_i be the rank of the matrix $E_{w_{i3}w_{i3}} \in \mathbb{R}^{p_i\times p_i}$ and let[3]

$$E^{1/2}_{w_{i3}w_{i3}} = Q_iU_i$$

be the QR-decomposition for $E^{1/2}_{w_{i3}w_{i3}}$ where $Q_i \in \mathbb{R}^{p_i\times\lambda_i}$ and $Q^T_iQ_i = I$ and $U_i \in \mathbb{R}^{\lambda_i\times p_i}$ is upper trapezoidal with rank λ_i. We write $G_i = U^T_i$ and use the notation

$$G_i = [g_{i1}, \ldots, g_{i\lambda_i}] \in \mathbb{R}^{p_i\times\lambda_i}$$

where $g_{ij} \in \mathbb{R}^{p_i}$ denotes the j-th column of G_i. We also write

$$G_{i,s} = [g_{i1}, \ldots, g_{is}] \in \mathbb{R}^{p_i\times s}$$

for $s \le \lambda_i$ to denote the matrix consisting of the first s columns of the matrix G_i.

The s-th row of the unit matrix $I \in \mathbb{R}^{p_i\times p_i}$ is denoted by $e^T_s \in \mathbb{R}^{1\times p_i}$.

For a square matrix $M = \{m_{ij}\}^n_{i,j=1}$, we also write

$$M = M_\nabla + M_\triangle$$

where

$$M_\nabla = \{m_{ij} \mid m_{ij} = 0 \quad \text{if} \quad i < j\}$$

and

$$M_\triangle = \{m_{ij} \mid m_{ij} = 0 \quad \text{if} \quad i \ge j\},$$

i.e. M_∇ is lower triangular and $M_\triangle$ is strictly upper triangular.

[3]We recall that by (6.14), $q_{i3} = p_i$.

Theorem 48. *The solution to the problem* (6.19) *is given by the operator* $\mathcal{A}^0 \in \mathbb{A}_m^n$ *defined by the lower stepped matrix* $A^0 \in \mathbb{L}_m^n$ *in the form* (6.36) *where*

$$L_{i3}^0 = \begin{bmatrix} \ell_{i,1}^0 \\ \vdots \\ \ell_{i,p_i}^0 \end{bmatrix} \quad \text{and} \quad R_{i2}^0 = T_{i2}^0 - L_{i3}^0 Z_i \quad \text{for} \quad i = 1, \ldots, l. \tag{6.39}$$

In (6.39), *for each* $i = 1, 2, \ldots, l$ *and* $s = 1, 2, \ldots, p_i$, *the* s*-th row* $\ell_{i,s}^0$ *is defined by*

$$\ell_{i,s}^0 = e_s^T E_{u_i w_{i3}} E_{w_{i3} w_{i3}}^\dagger G_{i,s} G_{i,s}^\dagger + b_i^T (I - G_{i,s} G_{i,s}^\dagger) \tag{6.40}$$

where $b_i^T \in \mathbb{R}^{1 \times p_i}$ *is arbitrary; the matrix* T_{i2}^0 *is such that*

$$T_{i2}^0 = E_{u_i w_{i2}} E_{w_{i2} w_{i2}}^\dagger + F_i (I - E_{w_{i2} w_{i2}} E_{w_{i2} w_{i2}}^\dagger) \tag{6.41}$$

with $F_i \in \mathbb{R}^{p_i \times q_{i2}}$ *arbitrary and* I *the* $q_{i2} \times q_{i2}$ *identity matrix.*

The error associated with the operator $\mathcal{A}^0$ *is given by*

$$\begin{aligned} E[\|\mathbf{x} - \mathcal{A}^0(\mathbf{y})\|^2] \;=\; & \sum_{i=1}^{l} \left[\sum_{s=1}^{\lambda_i} \sum_{j=s+1}^{p_i} E\left[|e_s^T E_{u_i w_{i3}} E_{w_{i3} w_{ki3}}^\dagger g_{i,j}|^2 \right] \right. \\ & \left. + \|E_{u_i u_i}^{1/2}\|_F^2 - \|E_{u_i w_{i2}} E_{w_{i2} w_{i2}}^{\dagger 1/2}\|^2 - \|E_{u_i w_{i3}} E_{w_{i3} w_{i3}}^{\dagger 1/2}\|_F^2 \right]. \end{aligned} \tag{6.42}$$

Proof. Since $E_{w_{i3} w_{i3}} = G_i G_i^T$, we have

$$\begin{aligned} J_1(L_{13}) &= E\left[\|\mathbf{u}_1 - \mathcal{L}_{13}(\mathbf{v}_{13})\|^2\right] \\ &= \operatorname{tr}\left\{E_{u_1 u_1} - E_{v_{13} v_{13}} L_{13}^T - L_{13} E_{v_{13} u_1} + L_{13} E_{v_{13} v_{13}} L_{13}^T\right\} \\ &= \operatorname{tr}\left\{(L_{13} - E_{u_1 v_{13}} E_{v_{13} v_{13}}^\dagger) E_{v_{13} v_{13}} (L_{13}^T - E_{v_{13} v_{13}}^\dagger E_{v_{13} u_1})\right\} \\ &= \operatorname{tr}\left\{(L_{13} - E_{u_1 v_{13}} E_{v_{13} v_{13}}^\dagger) G_1 G_1^T (L_{13}^T - E_{v_{13} v_{13}}^\dagger E_{v_{13} u_1})\right\} \\ &= E\left[\|(L_{13} - E_{u_1 v_{13}} E_{v_{13} v_{13}}^\dagger) G_1\|_F^2\right] \end{aligned} \tag{6.43}$$

and in the similar manner, for $i = 2, \ldots, l$,

$$\begin{aligned} \mathcal{J}_i(T_{i2}, L_{i3}) &= E\left[\|\mathbf{u}_i - [\mathcal{T}_{i2}(\mathbf{w}_{i2}) + \mathcal{L}_{i3}(\mathbf{w}_{i3})]\|^2\right] \\ &= \operatorname{tr}\left\{E_{u_i u_i} - E_{u_i w_{i2}} T_{i2}^T - E_{u_i w_i} L_{i3}^T - T_{i2} E_{w_{i2} u_i} \right. \\ &\qquad + T_{i2} E_{w_{i2} w_{i2}} T_{i2}^T + T_{i2} E_{w_{i2} w_{i3}} L_{i3}^T - L_{i3} E_{w_{i3} u_i} \\ &\qquad \left. + L_{i3} E_{w_{i3} w_{i2}} T_{i2}^T + L_{i3} E_{w_{i3} w_{i3}} L_{i3}^T \right\} \end{aligned}$$

$$
\begin{aligned}
&= \operatorname{tr}\big\{E_{u_iu_i} - E_{u_iw_{i2}}T_{i2}^T - E_{u_iw_{i3}}L_{i3}^T - T_{i2}E_{w_{i2}u_i} \\
&\qquad + T_{i2}E_{w_{i2}w_{i2}}T_{i2}^T - L_{i3}E_{w_{i3}u_i} + L_{i3}E_{w_{i3}w_{i3}}L_{i3}^T\big\} \\
&= \operatorname{tr}\big\{(T_{i2} - E_{u_iw_{i2}}E_{w_{i2}w_{i2}}^{\dagger})E_{w_{i2}w_{i2}}(T_{i2}^T - E_{w_{i2}w_{i2}}^{\dagger}E_{w_{i2}u_i}) \\
&\quad + (L_{i3} - E_{u_iw_{i3}}E_{w_{i3}w_{i3}}^{\dagger})G_iG_i^T(L_{i3}^T - E_{w_{i3}w_{i3}}^{\dagger}E_{w_{i3}u_i}) \\
&\quad + E_{u_iu_i} - E_{u_iw_{i2}}E_{w_{i2}w_{i2}}^{\dagger}E_{w_{i2}u_i} - E_{u_iw_{i3}}E_{w_{i3}w_{i3}}^{\dagger}E_{w_{i3}u_i}\big\} \\
&= E\left[\|(T_{i2} - E_{u_iw_{i2}}E_{w_{i2}w_{i2}}^{\dagger})E_{w_{i2}w_{i2}}^{1/2}\|_F^2\right] \qquad (6.44) \\
&\quad + E\left[\|(L_{i3} - E_{u_iw_{i3}}E_{w_{i3}w_{i3}}^{\dagger})G_i\|_F^2\right] \qquad (6.45) \\
&\quad + \|E_{u_iu_i}^{1/2}\|_F^2 - \|E_{u_iw_{i2}}E_{w_{i2}w_{i2}}^{\dagger 1/2}\|^2 - \|E_{u_iw_{i3}}E_{w_{i3}w_{i3}}^{\dagger 1/2}\|_F^2. \qquad (6.46)
\end{aligned}
$$

For symmetry, we set $\mathbf{w}_{13} = \mathbf{v}_{13}$. Then on the basis of Lemma 38 and the fact that the matrix $L_{i3}G_i$ is lower triangular, (6.43) and (6.45) can be written collectively for $i = 1, \ldots, l$ as follows:

$$
\begin{aligned}
&E\left[\|(L_{i3} - E_{u_iw_{i3}}E_{w_{i3}w_{i3}}^{\dagger})G_i\|_F^2\right] \\
&\quad = E\left[\|(L_{i3}G_i - E_{u_iw_{i3}}E_{w_{i3}w_{i3}}^{\dagger}G_i)_{\nabla} \right. \\
&\qquad\qquad \left. +(L_{i3}G_i - E_{u_iw_{i3}}E_{w_{i3}w_{i3}}^{\dagger}G_i)_{\triangle}\|_F^2\right] \\
&= \sum_{s=1}^{\lambda_i}\sum_{j=1}^{s} E\left[|(\ell_{i,s}g_{i,j} - e_s^TE_{u_iw_{i3}}E_{w_{i3}w_{i3}}^{\dagger}g_{i,j})|^2\right] \\
&\quad + \sum_{s=1}^{\lambda_i}\sum_{j=s+1}^{p_i} E\left[|e_s^TE_{u_iw_{i3}}E_{w_{i3}w_{i3}}^{\dagger}g_{i,j}|^2\right] \\
&\qquad = \sum_{s=1}^{p_i} E\left[|(\ell_{i,s}g_{i,s} - e_s^TE_{u_iw_{i3}}E_{w_{i3}w_{i3}}^{\dagger}g_{i,s})|^2\right] \\
&\qquad\quad + \sum_{s=1}^{\lambda_i}\sum_{j=s+1}^{p_i} E\left[|e_s^TE_{u_iw_{i3}}E_{w_{i3}w_{i3}}^{\dagger}g_{i,j}|^2\right].
\end{aligned}
$$

The first sum calculates the contribution from all elements with $j \le s$ that are on or below the leading diagonal of the matrix $(L_{i3} - E_{u_iw_{i3}}E_{w_{i3}w_{i3}}^{\dagger}) \times G_i$ and the second sum calculates the contribution from all elements with $j > s$ that are strictly above the leading diagonal. To minimize the overall expression it would be sufficient to set the terms on or below the leading diagonal to zero. Thus we wish to solve the matrix equation

$$\ell_{i,s}G_{i,s} - e_s^TE_{u_iw_{i3}}E_{w_{i3}w_{i3}}^{\dagger}G_{i,s} = 0$$

for each $i = 1, 2, \ldots, l$. This equation is a system of $(2p_i - \lambda_i + 1)\lambda/2$ equations in $(p_i + 1)p_i/2$ unknowns. Hence there is always at least one

solution. Indeed by applying similar arguments to those used earlier in Lemma 26 of Section 5.4.2, it can be seen that the general solution is given by $\ell_{i,s} = \ell^0_{i,s}$ for each $i = 1, 2, \ldots, l$.

Next, it follows from (6.44) that minimum of

$$E\left[\|(T_{i2} - E_{u_i w_{i2}} E^{\dagger}_{w_{i2} w_{i2}}) E^{1/2}_{w_{i2} w_{i2}}\|^2_F\right]$$

is attained if

$$(T_{i2} - E_{u_i w_{i2}} E^{\dagger}_{w_{i2} w_{i2}}) E^{1/2}_{w_{i2} w_{i2}} = 0 \tag{6.47}$$

is true. By Lemma 26 (Section 5.4.2), this equation is equivalent to the equation

$$T_{i2} E_{w_{i2} w_{i2}} - E_{u_i w_{i2}} = 0. \tag{6.48}$$

The general solution [6] to (6.48) is given by (6.44). Therefore, R^0_{i2} by (6.39) is true on the basis of (6.31).

The error representation (6.42) follows from (6.44)–(6.46). □

Remark 24. *The matrix $G_i \in \mathbb{R}^{p_i \times r}$ has rank λ_i and hence has λ_i independent columns. It follows that $G_{i,s} \in \mathbb{R}^{p_i \times s}$ also has independent columns and therefore also has rank s. Thus $G^T_{i,s} G_{i,s} \in \mathbb{R}^{\lambda_i \times \lambda_i}$ is non-singular and so $G^{\dagger}_{i,s} = (G^T_{i,s} G_{i,s})^{-1} G^T_{i,s}$. Hence*

$$\ell^0_{i,s} = e^T_s E_{u_i w_{i3}} E^{\dagger}_{w_{i3} w_{i3}} G_{i,s} (G^T_{i,s} G_{i,s})^{-1} G^T_{i,s} + b^T_i [I - G_{i,s} (G^T_{i,s} G_{i,s})^{-1} G^T_{i,s}]$$

for all $i = 1, 2, \ldots, l$.

Remark 25. *The results by Bode and Shannon* [9]*, Fomin and Ruzhansky* [37]*, Ruzhansky and Fomin* [115]*, and Wiener* [179]*, which concern a linear operator approximation, are particular cases of Theorem* 48 *above.*

6.4.8 Simulations

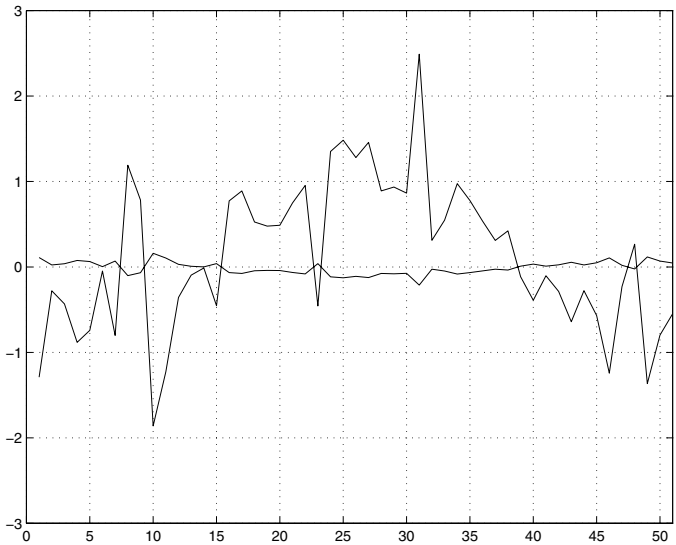
To illustrate the proposed method, we consider the best approximator $\mathcal{A}^0 \in \mathbb{A}^n_m$ with $n = 51$ and memory $m = \{m(1), \ldots, m(5)\}$, where $m(1) = 20$, $m(2) = 25$, $m(3) = 15$, $m(4) = 35$ and $m(5) = 25$.

Then the blocks of the matrix A^0 are

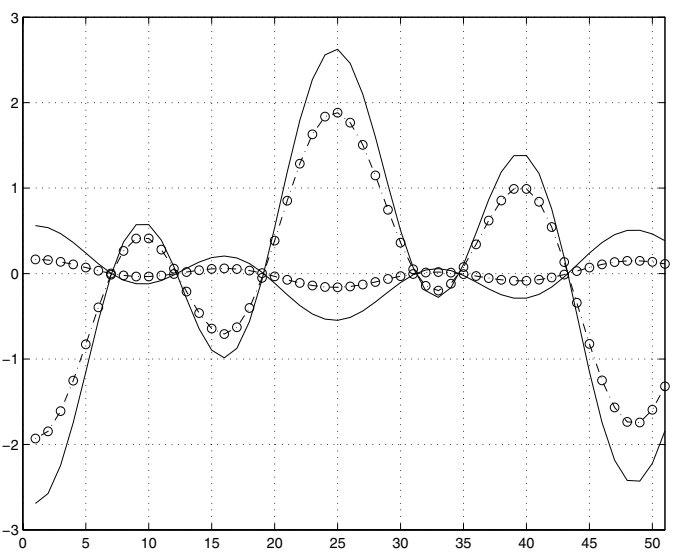
$$L^0_{13} \in \mathbb{R}^{20 \times 20}, \quad R^0_{22} \in \mathbb{R}^{10 \times 15}, \quad L^0_{23} \in \mathbb{R}^{10 \times 10}, \tag{6.49}$$

$$R^0_{32} \in \mathbb{R}^{5 \times 10}, \quad L^0_{33} \in \mathbb{R}^{5 \times 5}, \quad R^0_{42} \in \mathbb{R}^{10 \times 25}, \quad L^0_{43} \in \mathbb{R}^{10 \times 10}. \tag{6.50}$$

$$R^0_{52} \in \mathbb{R}^{5 \times 20} \quad \text{and} \quad L^0_{53} \in \mathbb{R}^{5 \times 5}. \tag{6.51}$$



(a) Two typical realizations of noisy observed data $\mathbf{y}$.



(b) Two related realizations of reference vector $\mathbf{x}$ (solid line) and their estimates (dashed line with circles) by the proposed approximator.

Figure 6.4: Illustration to the performance of the causal filter with piecewise-constant memory.

We apply $\mathcal{A}^0 \in \mathbb{A}_m^{51}$ to the random vector $\mathbf{y}$ under conditions as follows. In accordance with the assumption made above, we suppose that a reference random vector $\mathbf{x} \in L^2(\Omega, \mathbb{R}^{51})$ is unknown and that noisy observed data $\mathbf{y} \in L^2(\Omega, \mathbb{R}^{51})$ is given by q realizations of $\mathbf{y}$ in the form of a matrix $Y \in \mathbb{R}^{n \times q}$ with $q = 101$. Matrices

$$E_{u_1 v_{13}}, \quad E_{v_{13} v_{13}} \tag{6.52}$$

and matrices

$$E_{u_i v_{i2}}, \quad E_{u_i v_{i3}}, \quad E_{v_{i2} v_{i2}} \quad \text{and} \quad E_{v_{i3} v_{i3}} \tag{6.53}$$

for $i = 2, \ldots, 5$, or their estimates are assumed to be known.

In practice, these matrices or their estimates are given numerically, not analytically. Similarly to our methods presented in [61] and [156]–[162], the proposed method works, of course, under this condition. In this example, we model the matrices (6.52)–(6.53) and Y with analytical expressions in the following way. First, we set $X \in \mathbb{R}^{n \times q}$ and $Y \in \mathbb{R}^{n \times q}$ by

$$X = [\cos(\alpha) + \cos(0.3\alpha)]^T [\cos(0.5\beta) + \sin(5\beta)]$$

and

$$Y = [\cos(\alpha) \bullet r_1 + \cos(0.3\alpha)]^T [\cos(0.5\beta) + \sin(5\beta) \bullet r_2],$$

where

$$\alpha = [\alpha_0, \alpha_1, \ldots, \alpha_{n-1}], \quad \alpha_{k+1} = \alpha_k + 0.4, \quad k = 0, 1, \ldots, n-1, \quad \alpha_0 = 0,$$

$$\beta = [\beta_0, \beta_1, \ldots, \beta_{q-1}], \quad \beta_{j+1} = \beta_j + 0.4, \quad j = 0, 1, \ldots, q-1, \quad \beta_0 = 0,$$

$$\cos(\alpha) = [\cos(\alpha_0), \ldots, \cos(\alpha_n)], \quad \sin(\beta) = [\sin(\beta_0), \ldots, \sin(\beta_{q-1})],$$

the symbol $\bullet$ means the Hadamard product, r_1 is a $1 \times n$ normally distributed random vector and r_2 is a $1 \times q$ uniformly distributed random vector. Here, r_1 and r_2 simulate noise.[4]

Each column of Y is a particular realization of $\mathbf{y}$.

By the procedure described in Section 5.2.2(e), we partition each column of X and Y in sub-vectors $u_1, \ldots, u_5$ and v_{13}, v_{22}, v_{23}, …, v_{52}, v_{53}, respectively.

Furthermore, v_{13}, v_{22}, v_{23}, v_{32}, v_{33} and v_{34} have been orthogonalized to w_{11}, w_{22}, w_{23}, w_{32}, w_{33} and w_{34}. Matrices (6.49)–(6.51) have then been evaluated by (6.39)–(6.44) from $u_1, \ldots, u_3$, and w_{11}, w_{22}, w_{23}, w_{32}, w_{33} and w_{34}.

[4]The matrix X can be interpreted as a sample of $\mathbf{x}$. By the assumptions of the proposed method, it is not necessary to know X. We use matrix X for illustration purposes only.

As a result, the estimate $\hat{\mathbf{x}}^0$ has been evaluated in the form $\hat{x}^0$ such that

$$\hat{x}^0 = \begin{bmatrix} L^0_{13} w_{13} \\ R^0_{22} w_{22} + L^0_{23} w_{23} \\ \vdots \\ R^0_{52} w_{52} + L^0_{53} w_{53} \end{bmatrix}. \tag{6.54}$$

On Fig. 6.4(a), the plots of columns 51 and 52 of the matrix Y are presented. They are typical representatives of the noisy data under consideration. On Fig. 6.4(b), the plots of columns 51 and 52 of the matrix X (solid line) and their estimates (dashed line with circles) by our approximator are given.

6.5 Optimal Causal Polynomial Filtering with Arbitrarily Variable Memory

As before, we represent the raw data by a random vector $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)^T \in L^2(\Omega, \mathbb{R}^m)$. The reference random vector to be estimated from $\mathbf{y}$ is denoted by $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)^T \in L^2(\Omega, \mathbb{R}^m)$.[5]

Unlike the method considered in the preceding section, we now consider a case when memory may vary arbitrarily. This means that to estimate a component $\mathbf{x}_k$ of the reference vector, an filter $\mathcal{A}$ uses no more than the $p_k = 1, \ldots, v_k$ most recent components $\mathbf{y}_{s_k}, \ldots, \mathbf{y}_{v_k}$ from the measurement vector, where s_k and v_k are respectively defined by

$$s_k = v_k - p_k + 1 \quad \text{and} \quad v_k = 1, \ldots, k. \tag{6.55}$$

We say that such an filter $\mathcal{A}$ has arbitrarily variable memory $p = \{p_1, \ldots, p_m\}$.

In addition to motivations considered in the previous section, we also motivated by the fact that a non-linear approximator has more degrees of freedom than the linear approximator considered above and it is natural to expect that an optimal *non-linear* filter will improve the accuracy of the optimal estimate. It is often possible to regulate the accuracy by changing the free parameters associated with a non-linear filter. If the filter is a polynomial operator of degree r then it may be possible to improve the accuracy of the approximation by increasing the degree.

As a result, another major difference from he preceding section is that here, we propose a generalized *polynomial* filter of degree r to reduce the inherent noise. The accuracy of the estimation will be regulated by the

[5]The index $k \in \{1, 2, \ldots, m\}$ may specify the time $t_k \in T = \{t_k \mid t_1 < \cdots < t_m\}$ at which the measurement is taken.

degree of the operator and the estimation procedure will be defined in terms of the *generalized inverse* of an observed covariance matrix. An optimal filter from this class will *always* exists but will be unique only if the covariance matrix is non-singular.

To satisfy conditions of causality and finite memory p, we construct the filter in terms of lower *variable-band* matrices. The optimal filter is not easy to determine because the natural minimization procedure for the expected square norm of the error does not preserve the embedded lower variable-band structure. We avoid these difficulties by reformulating the problem as a collection of linear problems on a set of new multinomial variables $y_s^{i_s} \cdots y_v^{i_v}$, where $s = s_k$ and $v = v_k$, using observations at times $t = s, \ldots, k$ to obtain an optimal estimate of the reference vector component $\mathbf{x}_k$ at time $t = k$. Hence the variable memory and causality restrictions are imposed indirectly. It is a remarkable fact that the minimum value for the sum of the square errors over all components is equal to the sum of the individual minimum square errors. We establish the reformulation in Proposition 3 by showing that the original problem can be reduced to m independent problems defined by estimation of the separate components $\mathbf{x}_1, \ldots, \mathbf{x}_m$ of the reference vector. The construction procedures are described in Sections 6.5.1–6.5.6.

While the problem under consideration is presented as a problem of random vector estimation it could also be seen as a generalized form of non-linear regression. Nevertheless *our statement* of the problem in Sections 6.5.4 and 6.5.5, and *our solution methodology* (presented in Proposition 3, Theorems 49 and 50, and in their proofs) differ significantly from those given in the literature. We cite [1, 5, 30, 33, 110, 137, 176, 178, 181] for work on non-linear regression and [16], [56]–[62], [92, 95, 103], [118], [136], [153]–[160], [175, 182] for random vector estimation.

A distinguishing feature of the presented method is that the filter should be non-linear and causal with finite memory. The simplest particular case of the desired filter is an optimal *linear* filter defined by a lower p-band matrix (see Example 19 in Section 6.5.1).

In Sections 6.5.5 and 6.5.6, we describe a new algorithm to perform the necessary calculations for the proposed filter model. In addition, the reduction procedure (Section 6.5.5) means that our optimal filters are defined by matrices of reduced size. Consequently the computational load should compare favorably with known methods [9, 37, 71, 88, 115, 142, 154, 161]. On the other hand we use non-linear filters to provide improved accuracy and it is natural to expect additional computation in problems where increased accuracy is desired.

6.5.1 Non-linear causal filters with arbitrarily variable memory

We use the same notation for $\mathbf{x}$ and $\mathbf{y}$ as in Section 6.4. Using an heuristic idea of causality we expect that the present value of the estimate is not affected by future values of the data [103]–[112]. Since the filters under consideration must have variable memory, the estimate of $\mathbf{x}_k$ must be obtained from the data components $\mathbf{y}_{s_k}, \ldots, \mathbf{y}_{v_k}$ with s_k and v_k given by (6.55). In the following definition we combine the concepts of causality and variable memory.

Definition 37. *Let s_k and v_k be defined by* (6.55). *For each $k = 1, 2, \ldots, m$ and $\mathbf{y} \in L^2(\Omega, \mathbb{R}^m)$ define $\boldsymbol{\tau}_{s_k v_k} : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^{p_k})$ by*

$$\boldsymbol{\tau}_{s_k v_k}(\mathbf{y}) = (\mathbf{y}_{s_k}, \mathbf{y}_{s_k+1}, \ldots, \mathbf{y}_{v_k}).$$

Let $\mathcal{A} : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^m)$. If $\mathcal{A}_k : L^2(\Omega, \mathbb{R}^{p_k}) \to L^2(\Omega, \mathbb{R})$ for each $k = 1, 2, \ldots, m$ is such that

$$\mathbf{u} = \mathcal{A}(\mathbf{y}) = \begin{bmatrix} \mathcal{A}_1(\boldsymbol{\tau}_{s_1 v_1}(\mathbf{y})) \\ \vdots \\ \mathcal{A}_m(\boldsymbol{\tau}_{s_m v_m}(\mathbf{y})) \end{bmatrix} \tag{6.56}$$

for each $\mathbf{y} \in L^2(\Omega, \mathbb{R}^m$, then $\mathcal{A}$ is called a causal *filter with* arbitrarily variable finite memory $p = \{p_1, \ldots, p_m\}$. *If $p_k < k$ for some $k = 1, \ldots, m$, the memory is called* incomplete *and if $p_k = k$ and $v_k = k$ for each $k = 1, \ldots, m$, the memory is said to be* complete.

The relation (6.56) can be represented in the more explicit form

$$\begin{aligned} \mathbf{u}_1 &= \mathcal{A}_1(\mathbf{y}_1), \\ \mathbf{u}_2 &= \mathcal{A}_p(\mathbf{y}_{s_2}, \ldots, \mathbf{y}_{v_2}) \\ &\vdots \\ \mathbf{u}_m &= \mathcal{A}_p(\mathbf{y}_{s_m}, \ldots, \mathbf{y}_{v_m}). \end{aligned}$$

For an appropriate choice of $\mathcal{A}$, the vector $\mathbf{u}$ can be interpreted as an estimate of $\mathbf{x}$ from $\mathbf{y}$. We illustrate Definition 37 for the case in which $\mathcal{A}(\mathbf{y})$ is given by a first-order polynomial.

Example 19. *Suppose, $\mathcal{A}$ is defined by*

$$\mathcal{A}(\mathbf{y}) = a + \mathcal{B}_1(\mathbf{y}),$$

where $a \in \mathbb{R}^m$ *and* $\mathcal{B}_1 : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^m)$ *is defined by*

$$[\mathcal{B}_1(\mathbf{y})](\omega) = B_1\mathbf{y}(\omega) \tag{6.57}$$

with $B_1 \in \mathbb{R}^{m\times m}$. *By Definition* 37, *the operator* $\mathcal{A}$ *is causal with arbitrarily variable memory* p *if the matrix* $B_1 = \{b_{kj}\}$ *is such that*

$$b_{kj} = 0 \quad \textit{for } j = \{1, \ldots, s_k - 1\} \bigcup \{v_k + 1, \ldots, m\}$$

We call B_1 *a* lower p-variable-band matrix. *The set of lower* p*-variable-band matrices in* $\mathbb{R}^m$ *is denoted by* $\mathcal{R}^{m\times m}_{p_1,\ldots,p_m}$.

For instance, if

$$\begin{aligned} &m = 4, \quad p_1 = 1, \quad p_2 = 2, \quad p_3 = 2, \quad p_4 = 3,\\ &v_1 = 1, \quad v_2 = 2, \quad v_3 = 3, \quad v_4 = 3, \end{aligned}$$

then $B_1 \in \mathcal{R}^{4\times 4}_{1223}$ *is given by*

$$B_1 = \begin{bmatrix} \bullet & 0 & 0 & 0 \\ \bullet & \bullet & 0 & 0 \\ 0 & \bullet & \bullet & 0 \\ \bullet & \bullet & \bullet & 0 \end{bmatrix}$$

where $\bullet$ *denotes an entry which is not necessarily equal to zero.*

In the next section, we present a model of a causal filter with with arbitrarily variable memory $p = \{p_1, \ldots, p_m\}$ in the form of an r-degree operator $\mathcal{T}_r$.

The problem is to find an filter $\mathcal{T}_r^0$ which minimizes the associated mean square error on the class of all causal filters of degree r with arbitrarily variable memory p. In Section 6.5.2 we construct a general r-degree filter $\mathcal{T}_r$ and in Section 6.5.3 we restrict this representation to include only causal operators with arbitrarily variable memory p. A rigorous statement of the problem is given in Sections 6.5.4 and 6.5.5. The *optimal* filter $\mathcal{T}_r^0$ is given in Section 6.5.6. We shall show that it is possible to reduce the error associated with the optimal filter by increasing its degree.

6.5.2 Model of a general r-degree filter

We follow the procedure used in [159]. For $r \in \mathbb{N}$, let $\mathcal{P}_r : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^m)$ be given by

$$\mathcal{P}_r(\mathbf{y}) = a + \mathcal{B}_1(\mathbf{y}) + \sum_{q=2}^{r} \mathcal{B}_q(\mathbf{y}^q),$$

where $a = a(r) \in \mathbb{R}^m$ and $\mathcal{B}_q = \mathcal{B}_q(r) : L^2(\Omega, (\mathbb{R}^m)^q) \to L^2(\Omega, \mathbb{R}^m)$ for $q \geq 1$ is a q-linear operator[6] such that

$$[\mathcal{B}_q(\mathbf{y}^q)](\omega) = B_q[y^q] \tag{6.58}$$

where $B_q : (\mathbb{R}^m)^q \to \mathbb{R}^m$ is a tensor (i.e. a '$(q+1)$-dimensional' matrix), $y^q = \mathbf{y}(\omega)^q = (\mathbf{y}(\omega), \ldots, \mathbf{y}(\omega)) = (y, \ldots, y) \in (\mathbb{R}^m)^q$ and $(\mathbb{R}^m)^q$ denotes the q-th power of $\mathbb{R}^m$.

We consider $\mathcal{P}_r$ as an estimator of $\mathbf{x}$ from $\mathbf{y}$. The motivation for using $\mathcal{P}_r$ as an estimator follows from known results [56, 57, 159] where strong estimating properties of $\mathcal{P}_r$ have been justified and demonstrated. Here, and in the rest of the section, operators acting on spaces of random vectors and defined similarly to those in (6.57) and (6.58), are denoted by calligraphic characters. We call $\mathcal{P}_r$ an r-degree operator. For $q \geq 2$ note that

$$\begin{aligned} B_q[y^q] &= B_q\left[\sum_{j_1=1}^m y_{j_1} e_{j_1}, \ldots, \sum_{j_{q-1}=1}^m y_{j_{q-1}} e_{j_{q-1}}, y\right] \\ &= \sum_{j_1=1}^m \cdots \sum_{j_{q-1}=1}^m y_{j_1} \cdots y_{j_{q-1}} B_q[e_{j_1}, \ldots, e_{j_{q-1}}, y] \\ &= \sum_{j_1=1}^m \cdots \sum_{j_{q-1}=1}^m y_{j_1} \cdots y_{j_{q-1}} B_{q,j_1 \cdots j_{q-1}} y \end{aligned}$$

where $B_{q,j_1 \cdots j_{q-1}} \in \mathbb{R}^{m \times m}$. Thus, in matrix terminology, we write $\mathcal{P}_r(\mathbf{y})(\omega) = P_r(y)$ in the form

$$P_r(y) = a + B_1(y) + \sum_{q=2}^r \sum_{j_1=1}^m \cdots \sum_{j_{q-1}=1}^m y_{j_1} \cdots y_{j_{q-1}} B_{q,j_1 \cdots j_{q-1}} y. \tag{6.59}$$

For each $k = 1, 2, \ldots, m$ the k-th element of $P_r(y)$ is denoted by $P_{r,k}(y)$ and is given by

$$\begin{aligned} P_{r,k}(y) &= a_k + B_{1(k)} y + \sum_{q=2}^r \sum_{j_1=1}^m \cdots \sum_{j_{q-1}=1}^m y_{j_1} \ldots y_{j_{q-1}} B_{q,j_1 \cdots j_{q-1}(k)} y \\ &= a_k + \sum_{l=1}^m b_{1(kl)} y_l \\ &\quad + \sum_{q=2}^r \sum_{j_1=1}^m \cdots \sum_{j_{q-1}=1}^m y_{j_1} \ldots y_{j_{q-1}} \sum_{l=1}^m b_{q,j_1 \ldots j_{q-1}(kl)} y_l, \end{aligned} \tag{6.60}$$

[6]The operator $\mathcal{B}_q$ is called q-linear if it is linear in each of its arguments.

where $B_{1(k)} \in \mathbb{R}^{1\times m}$ and $B_{q,j_1\cdots j_{q-1}(k)} \in \mathbb{R}^{1\times m}$ denote the k-th rows of the matrices B_1 and $B_{q,j_1\cdots j_{q-1}}$ respectively and $b_{1(kl)}$ and $b_{q,j_1\ldots j_{q-1}(kl)}$ denote the elements in the kl position in the respective matrices. The expression $P_{r,k}(y)$ contains

$$1 + m + m^2 + \cdots + m^r = \frac{m^{r+1} - 1}{m - 1}$$

scalar terms.

To avoid the *symmetry effect* discussed in Section 5.5.3, we use the same device as in Section 5.5.3. Namely, we collect together all terms with the same factors $y_1^{i_1} \cdots y_m^{i_m}$ into a sub-class $\mathcal{S}(i_1, \ldots, i_m)$ for each combination of non-negative integers $(i_1, \ldots, i_m)$ in the class $\mathcal{S}_{m,r}$ with $i_1 + \cdots + i_m \le r$ and write (6.60) in the form

$$P_{r,k}(y) \quad = \quad a_k \quad + \sum_{(i_1,\ldots,i_m) \,\in\, \mathcal{S}_{m,r}} y_1^{i_1} \cdots y_m^{i_m} \quad B_{i_1\cdots i_m(k)} y$$

where $B_{0\cdots 0} = B_1$ and

$$B_{i_1\cdots i_m} = \sum_{y_{j_1}\cdots y_{j_{q-1}} \,\in\, \mathcal{S}(i_1,\ldots,i_m)} B_{q,j_1\cdots j_{q-1}}$$

for $(i_1, \ldots, i_m) \neq (0, \ldots, 0)$ and where $B_{0\cdots 0(k)}$ and $B_{i_1\cdots i_m(k)}$ denote the k-th rows of the respective matrices. This reduces the number of terms in $P_{r,k}(y)$ to

$$\begin{aligned} &1 + \left[\begin{pmatrix} m-1 \\ 0 \end{pmatrix} + \begin{pmatrix} m \\ 1 \end{pmatrix} + \cdots + \begin{pmatrix} m+r-2 \\ r-1 \end{pmatrix} \right] \cdot m \\ &\quad = 1 + \begin{pmatrix} m+r-1 \\ r-1 \end{pmatrix} \cdot m \end{aligned}$$

and allows us to avoid the symmetry effect. In operator form we have

$$\mathcal{P}_{r,k}(\mathbf{y}) = a_k \quad + \sum_{(i_1,\ldots,i_m) \,\in\, \mathcal{S}_{m,r}} \mathbf{y}_1^{i_1} \ldots \mathbf{y}_m^{i_m} \quad \mathcal{B}_{i_1\ldots i_m(k)}(\mathbf{y}) \qquad (6.61)$$

with $\mathcal{B}_{i_1\ldots i_m(k)} : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R})$ defined by $B_{i_1\ldots i_m(k)}$ for each $(i_1, \ldots, i_m) \in \mathcal{S}_{m,r}$.

6.5.3 Model of a causal r-degree filter with arbitrarily variable memory

A causal r-degree filter $\mathcal{T}_r : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^m)$ with arbitrarily variable memory $p = \{p_1, \ldots, p_m\}$ is constructed from (6.61) in the following way. We set

$$\mathcal{T}_r(\mathbf{y}) = \begin{bmatrix} \mathcal{T}_{r,1}(\tau_{s_1 v_1}(\mathbf{y})) \\ \mathcal{T}_{r,2}(\tau_{s_2 v_2}(\mathbf{y})) \\ \vdots \\ \mathcal{T}_{r,m}(\tau_{s_m v_m}(\mathbf{y})) \end{bmatrix} = \begin{bmatrix} \mathcal{T}_{r,1}(\mathbf{y}_1) \\ \mathcal{T}_{r,2}(\mathbf{y}_{s_2}, \mathbf{y}_{v_2}) \\ \vdots \\ \mathcal{T}_{r,m}(\mathbf{y}_{s_m}, \ldots, \mathbf{y}_{v_m}) \end{bmatrix}, \quad (6.62)$$

where for each $k = 1, \ldots, m$ we have

$$\mathcal{T}_{r,k}(\boldsymbol{\tau}_{s_k v_k}(\mathbf{y})) = a_k + \sum_{(0,\ldots,0,i_{s_k},\ldots,i_{v_k},0,\ldots,0) \in \mathcal{S}_{m,r}} \mathbf{y}_{s_k}^{i_{s_k}} \ldots \mathbf{y}_{v_k}^{i_{v_k}} \quad \boldsymbol{\beta}_{0\cdots 0 i_{s_k} \cdots i_{v_k} 0 \cdots 0(k)}(\boldsymbol{\tau}_{s_k v_k}(\mathbf{y})), \quad (6.63)$$

where $\boldsymbol{\beta}_{0\cdots 0 i_{s_k} \cdots i_{v_k} 0 \cdots 0(k)} : L^2(\Omega, \mathbb{R}^{p_k}) \to L^2(\Omega, \mathbb{R})$ is an appropriate restriction of $\mathcal{B}_{i_1 \ldots i_m(k)}$. Thus $\mathcal{T}_{r,k}$ is constructed from $\mathcal{P}_{r,k}$ when the general terms $\mathbf{y}_1^{i_1} \cdots \mathbf{y}_m^{i_m} \; \mathcal{B}_{i_1 \ldots i_m(k)}(\mathbf{y})$ in (6.61) are restricted to terms of the form $\mathbf{y}_{s_k}^{i_{s_k}} \cdots \mathbf{y}_{v_k}^{i_{v_k}} \; \boldsymbol{\beta}_{0\cdots 0 i_{s_k} \cdots i_{v_k} 0 \cdots 0(k)}(\boldsymbol{\tau}_{s_k v_k}(\mathbf{y}))$. In the usual way we write

$$\mathcal{T}_{r,k}(\boldsymbol{\tau}_{s_k v_k}(\mathbf{y}))(\omega) = \mathcal{T}_{r,k}(y_{s_k}, \ldots, y_{v_k})$$

for a particular observation. The number of scalar terms in this expression is

$$1 + \left[\begin{pmatrix} v_k - s_k \\ 0 \end{pmatrix} + \begin{pmatrix} p_k \\ 1 \end{pmatrix} + \cdots + \begin{pmatrix} v_k - s_k + r - 1 \\ r - 1 \end{pmatrix} \right] \cdot p_k$$
$$= 1 + \begin{pmatrix} v_k - s_k + r \\ r - 1 \end{pmatrix} \cdot p_k.$$

Once again we avoid repetition of terms. According to Definition 37 the operator $\mathcal{T}_{r,k}$ is causal with arbitrarily variable memory $\{p_1, \ldots, p_m\}$. Note that it is possible to have different degree operators for each component. In such cases we simply replace r by r_k in (6.63).

Example 20. *We illustrate the structure of $\mathcal{P}_{r,k}$ and $\mathcal{T}_{r,k}$. If $m = 3$, $r = 4$, $v_k = 3$ and $p_k = 2$ for all $k = 1, \ldots, 4$ then*

$$P_{4,3}(y_1, y_2, y_3) = a_3 + \Big[B_{000(3)} + (y_1 B_{100(3)} + y_2 B_{010(3)} + y_3 B_{001(3)})$$

$$+(y_1^2 B_{200(3)} + y_1 y_2 B_{110(3)} + y_1 y_3 B_{101(3)} + y_2^2 B_{020(3)} + y_2 y_3 B_{011(3)} + y_3^2 B_{002(3)})$$

$$+(y_1^3 B_{300(3)} + y_1^2 y_2 B_{210(3)} + y_1^2 y_3 B_{201(3)} + y_1 y_2^2 B_{120(3)} + y_1 y_2 y_3 B_{111(3)} + y_1 y_3^2 B_{102(3)} + y_2^3 B_{030(3)} + y_2^2 y_3 B_{021(3)} + y_2 y_3^2 B_{012(3)} + y_3^3 B_{003(3)})\Bigg] \times \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

and $[\mathcal{T}_{4,3}(\mathbf{y}_2, \mathbf{y}_3)](\omega) = T_{4,3}(y_2, y_3)$ *is given by*

$$T_{4,3}(y_2, y_3) = a_3 + \Bigg[\beta_{000(3)} + (y_2 \beta_{010(3)} + y_3 \beta_{001(3)}) + (y_2^2 \beta_{020(3)} + y_2 y_3 \beta_{011(3)} + y_3^2 \beta_{002(3)}) + y_2^3 \beta_{030(3)} + y_2^2 y_3 \beta_{021(3)} + y_2 y_3^2 \beta_{012(3)} + y_3^3 \beta_{003(3)})\Bigg] \begin{bmatrix} y_2 \\ y_3 \end{bmatrix}$$

where each operator $\boldsymbol{\beta}_{0i_1i_2(3)} : L^2(\Omega, \mathbb{R}^2) \to L^2(\Omega, \mathbb{R})$ *is represented by a vector* $\beta_{0i_1i_2(3)}$ *with* $\beta^T_{0i_1i_2(3)} \in \mathbb{R}^2$. *We observe that the original expression for* $P_{4,3}(y_1, y_2, y_3)$ *in (6.60) contains* 121 *scalar terms and requires* $\mathcal{O}(350)$ *flops. When the symmetry effect is removed* $P_{4,3}(y_1, y_2, y_3)$ *contains only* 61 *scalar terms and this is further reduced in* $T_{4,3}(y_2, y_3)$ *to* 21 *scalar terms. As a result, computation of* $P_{4,3}(y_1, y_2, y_3)$ *requires* $\mathcal{O}(150)$ *flops while computation of* $T_{4,3}(y_2, y_3)$ *requires* $\mathcal{O}(50)$ *flops.*

Although $\mathcal{T}_{r,k}(\mathbf{y}_{s_k}, \ldots, \mathbf{y}_{v_k})$ is a multi-linear operator on the original variables $\mathbf{y}_1, \ldots, \mathbf{y}_m$ the dependence on the key product terms $\mathbf{y}_{s_k}^{i_{s_k}} \ldots \mathbf{y}_{v_k}^{i_{v_k}}$ $\boldsymbol{\tau}_{s_k v_k}(\mathbf{y})$ is linear. There are $N_k = N_k(r)$ such terms where

$$N_k = \begin{pmatrix} p_k \\ 0 \end{pmatrix} + \cdots + \begin{pmatrix} v_k - s_k + r - 1 \\ r - 1 \end{pmatrix} = \begin{pmatrix} v_k - s_k + r \\ r - 1 \end{pmatrix}. \quad (6.64)$$

We denote the terms by $\mathbf{h}_{js_kv_k} = \mathbf{h}_{js_kv_k}(r)$ and the corresponding linear operators by $\boldsymbol{\eta}^T_{js_kv_k} = \boldsymbol{\eta}^T_{js_kv_k}(r)$ for each $j = 1, 2, \ldots, N_k$. The precise ordering is not important. Thus we write

$$\mathcal{T}_{r,k}(\mathbf{y}_{s_k}, \ldots, \mathbf{y}_{v_k}) = a_k + \sum_{j=1}^{N_k} \boldsymbol{\eta}^T_{js_kv_k}(\mathbf{h}_{js_kv_k}) = a_k + \boldsymbol{\eta}^T_{s_kv_k}(\mathbf{h}_{s_kv_k}) \quad (6.65)$$

where $\boldsymbol{\eta}_{s_k v_k}^T = (\boldsymbol{\eta}_{1sk}^T, \ldots, \boldsymbol{\eta}_{N_k sk}^T)$ and $\mathbf{h}_{s_k v_k}^T = (\mathbf{h}_{1sk}^T, \ldots, \mathbf{h}_{N_k sk}^T)$ for each $k = 1, 2, \ldots, m$.

Example 21. *Let us consider $T_{4,3}(y_2, y_3)$ from Example 20. We define the variables*

$$h_{123} = \begin{bmatrix} y_2 \\ y_3 \end{bmatrix},\ h_{223} = y_2 \begin{bmatrix} y_2 \\ y_3 \end{bmatrix},\ h_{323} = y_3 \begin{bmatrix} y_2 \\ y_3 \end{bmatrix},\ h_{423} = y_2^2 \begin{bmatrix} y_2 \\ y_3 \end{bmatrix},$$
$$h_{523} = y_2 y_3 \begin{bmatrix} y_2 \\ y_3 \end{bmatrix},\ h_{623} = y_3^2 \begin{bmatrix} y_2 \\ y_3 \end{bmatrix},\ h_{723} = y_2^3 \begin{bmatrix} y_2 \\ y_3 \end{bmatrix},$$
$$h_{823} = y_2^2 y_3 \begin{bmatrix} y_2 \\ y_3 \end{bmatrix},\ h_{923} = y_2 y_3^2 \begin{bmatrix} y_2 \\ y_3 \end{bmatrix} \text{ and } h_{10,23} = y_3^3 \begin{bmatrix} y_2 \\ y_3 \end{bmatrix}$$

and the corresponding vector coefficients

$$\eta_{123}^T = \beta_{000(3)}, \quad \eta_{223}^T = \beta_{010(3)}, \quad \eta_{323}^T = \beta_{001(3)}, \quad \eta_{423}^T = \beta_{020(3)},$$
$$\eta_{523}^T = \beta_{011(3)}, \quad \eta_{623}^T = \beta_{002(3)}, \quad \eta_{723}^T = \beta_{030(3)}, \quad \eta_{823}^T = \beta_{021(3)},$$
$$\eta_{923}^T = \beta_{012(3)} \quad \text{and} \quad \eta_{10,23}^T = \beta_{003(3)}.$$

Remark 26. *Note that $T_{r,k}(y_{s_k}, \ldots, y_{v_k})$ does not contain repeated terms and depends only on $y_{s_k}, \ldots, y_{v_k}$. The number of terms N_k in $T_{r,k}(y_{s_k}, \ldots, y_{v_k})$ is much less than the number of terms in a general r-degree polynomial. Consequently a model using $T_{r,k}(y_{s_k}, \ldots, y_{v_k})$ requires much less computational work than a model using $P_{r,k}(y)$.*

6.5.4 Formulation of the problem

Let $n \in \mathbb{N}$. We note that for any vector $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_n)^T \in L^2(\Omega, \mathbb{R}^n)$,

$$|E[\mathbf{p}_i]|^2 \leq E[\|\mathbf{p}\|^2] < \infty$$

for all $i = 1, 2, \ldots, n$. Let

$$J(\mathcal{T}_r) = E[\|\mathbf{x} - \mathcal{T}_r(\mathbf{y})\|^2],$$

where $\mathcal{T}_r$ is defined by (6.62) and (6.65). The problem is to find $\mathcal{T}_r^0$ such that

$$J(\mathcal{T}_r^0) = \min_{\mathcal{T}_r} J(\mathcal{T}_r). \tag{6.66}$$

An optimal filter $\mathcal{T}_r^0$ in the class of causal r-degree filter with arbitrarily variable memory p takes the general form

$$\mathcal{T}_r^0(\mathbf{y}) = \begin{bmatrix} \mathcal{T}_{r,1}^0(\mathbf{y}_{s_k}, \ldots, \mathbf{y}_{v_k}) \\ \vdots \\ \mathcal{T}_{r,m}^0(\mathbf{y}_{s_k}, \ldots, \mathbf{y}_{v_k}) \end{bmatrix}, \tag{6.67}$$

where the component $\mathcal{T}^0_{r,k}$ is given by

$$\mathcal{T}^0_{r,k}(\mathbf{y}_{s_k},\ldots,\mathbf{y}_{v_k}) = a^0_k + \sum_{j=1}^{N_k} \boldsymbol{\eta}^{0\ T}_{js_kv_k}(\mathbf{h}_{js_kv_k}) = a^0_k + \boldsymbol{\eta}^{0\ T}_{s_kv_k}(\mathbf{h}_{s_kv_k}) \quad (6.68)$$

for each $k = 1,\ldots,m$. Finding an optimal representative $\mathcal{T}^0_r$ is therefore a matter of finding optimal values a^0_k and $\boldsymbol{\eta}^{0\ T}_{s_kv_k}$ for the constants a_k and operators $\boldsymbol{\eta}^{\ T}_{s_kv_k}$.

6.5.5 Reduction of the problem (6.66) to m independent problems

The special structure of the operators makes direct solution of (6.66) a difficult problem. Suffice it to say that a solution is known only for the special case where $\mathcal{T}_r$ is linear and has complete memory [37, 115]. Moreover the solution in [37, 115] has been obtained with a quite restrictive assumption that the covariance matrix $E[\mathbf{y}\mathbf{y}^T]$ is non-singular. Indeed we observe that direct determination of $\mathcal{T}^0_r$ from (6.66) is not straightforward because of difficulties imposed by the embedded lower p-variable-band structure of the matrices. To avoid these difficulties we show that the problem (6.66) can be reduced to m independent problems. Define

$$J_k(\mathcal{T}_{r,k}) = E[|\mathbf{x}_k - \mathcal{T}_{r,k}(\mathbf{y}_{s_k},\ldots,\mathbf{y}_{v_k})|^2] \quad (6.69)$$

for each $k = 1,\ldots,m$ where $\mathcal{T}_{r,k}$ is defined by (6.65). We have the following Proposition.

Proposition 3. *Let $\mathcal{T}_r$ and $\mathcal{T}_{r,k}$ be given by* (6.62) *and* (6.65) *respectively. Then*

$$\min_{\mathcal{T}_r} J(\mathcal{T}_r) = \sum_{k=1}^{m} \min_{\mathcal{T}_{r,k}} J_k(\mathcal{T}_{r,k}) \quad (6.70)$$

Proof. The proof is, in fact, a consequence of the proof of Lemma 37 in Section 6.4.7. By elementary algebra

$$J(\mathcal{T}_r) = \sum_{k=1}^{m} J_k(\mathcal{T}_{r,k}). \quad (6.71)$$

Let $\mathcal{T}^0_r$ be a solution to the problem (6.66). Then from (6.71),

$$\min_{\mathcal{T}_r} J(\mathcal{T}_r) \quad = \quad J(\mathcal{T}^0_r)$$

$$\begin{aligned} &= \sum_{k=1}^{m} J_k(\mathcal{T}^0_{r,k}) \qquad (6.72)\\ &\geq \sum_{k=1}^{m} \min_{\mathcal{T}_{r,k}} J_k(\mathcal{T}_{r,k}). \end{aligned}$$

In (6.72) the operator $\mathcal{T}^0_{r,k}$ is the k-th component of the optimal operator $\mathcal{T}^0_r$.

Let now $\mathcal{T}^*_{r,k}$ be a solution to the problem (6.69) for each $k = 1, 2, \ldots, m$. Then

$$\begin{aligned} \sum_{k=1}^{m} \min_{\mathcal{T}_{r,k}} J_k(\mathcal{T}_{r,k}) &= \sum_{k=1}^{m} J_k(\mathcal{T}^*_{r,k}) \\ &= J(\mathcal{T}^*_r) \qquad (6.73)\\ &\geq \min_{\mathcal{T}_r} J(\mathcal{T}_r), \end{aligned}$$

where

$$\mathcal{T}^*_{r_k}(\mathbf{y}_{s_k}, \ldots, \mathbf{y}_{v_k})^T = \left[\mathcal{T}^*_{r_1}(\mathbf{y}_{s_k}, \ldots, \mathbf{y}_{v_k}), \ldots, \mathcal{T}^*_{r_m}(\mathbf{y}_{s_k}, \ldots, \mathbf{y}_{v_k})\right].$$

The inequalities (6.72) and (6.73) establish the desired result. □

Expression (6.70) allows us to reformulate problem (6.66) in an equivalent form as follows.

For each $k = 1, \ldots, m$ find $\mathcal{T}^0_{r,k}$ such that

$$J_k(\mathcal{T}^0_{r,k}) = \min_{\mathcal{T}_{r,k}} J_k(\mathcal{T}_{r,k}). \qquad (6.74)$$

Any optimal operator $\mathcal{T}^0_{r,k}$ is the k-th component of an optimal filter $\mathcal{T}^0_r$. Hence an optimal filter $\mathcal{T}^0_r$ can be constructed from any solutions $\mathcal{T}^0_{r,1}, \ldots, \mathcal{T}^0_{r,m}$ to the m independent problems (6.74). This construction satisfies the structural requirements of Definition 1. An additional bonus is that each individual problem (6.74) can be solved by extending results obtained in Section 4.4.2. In this context solution of the m problems (6.74) is more tractable than a direct solution of the original problem (6.66).

6.5.6 Determination of the optimal causal r-degree filter with arbitrarily variable memory p

Let $n \in \mathbb{N}$ and let $\mathbf{p} \in L^2(\Omega, \mathbb{R}^n)$ and $\mathbf{q} \in L^2(\Omega, \mathbb{R}^n)$ be random vectors. In general we write $\hat{\mathbf{p}} = \mathbf{p} - E[\mathbf{p}]$ and $\hat{\mathbf{q}} = \mathbf{q} - E[\mathbf{q}]$ and define

$$\mathbb{E}_{fg} = E[\hat{\mathbf{p}}\hat{\mathbf{q}}^T] \quad = \quad E[(\mathbf{p} - E[\mathbf{p}])(\mathbf{q} - E[\mathbf{q}])^T]$$

$$= E[\mathbf{p}\mathbf{q}^T] - E[\mathbf{p}]E[\mathbf{q}^T].$$

It is convenient to use a special notation in two particular cases.

We define $\mathbb{H}_{s_k v_k} = \mathbb{H}_{s_k v_k}(r)$ and $\mathbb{Q}_{s_k v_k} = \mathbb{Q}_{s_k v_k}(r)$ by the formulae

$$\mathbb{H}_{s_k v_k} = E[\mathbf{h}_{s_k v_k} \mathbf{h}_{s_k v_k}^T] - E[\mathbf{h}_{s_k v_k}]E[\mathbf{h}_{s_k v_k}^T] \tag{6.75}$$

and

$$\mathbb{Q}_{s_k v_k} = E[\mathbf{x}_k \mathbf{h}_{s_k v_k}^T] - E[\mathbf{x}_k]E[\mathbf{h}_{s_k v_k}^T]. \tag{6.76}$$

The following theorem provides the solution to problem (6.74) for each $k = 1, 2, \ldots, m$.

Theorem 49. *For each $k = 1, 2, \ldots, m$ an optimal causal r-degree filter $\mathcal{T}_{r,k}^0(y_{s_k}, \ldots, y_{v_k})$ in (6.68) with arbitrarily variable memory p is defined by*

$$\eta_{s_k v_k}^{0\ T} = \mathbb{Q}_{s_k v_k} \mathbb{H}_{s_k v_k}^\dagger + M_k[I_k - \mathbb{H}_{s_k v_k} \mathbb{H}_{s_k v_k}^\dagger], \tag{6.77}$$

where $M_k \in \mathbb{R}^{1 \times N_k}$ is arbitrary and $I_k \in \mathbb{R}^{N_k \times N_k}$ is the identity matrix, and

$$a_k^0 = E[\mathbf{x}_k] - \eta_{s_k v_k}^{0\ T} E[\mathbf{h}_{s_k v_k}]. \tag{6.78}$$

Proof. First, we note that

$$\mathbb{Q}_{s_k v_k} \mathbb{H}_{s_k v_k}^\dagger \mathbb{H}_{s_k v_k} = \mathbb{Q}_{s_k v_k}. \tag{6.79}$$

Indeed, since

$$\mathbb{Q}_{s_k v_k} = E[\mathbf{x}_k \mathbf{h}_{s_k v_k}^T] - E[\mathbf{x}_k]E[\mathbf{h}_{s_k v_k}^T]$$

and

$$\mathbb{H}_{s_k v_k} = E[\mathbf{h}_{s_k v_k} \mathbf{h}_{s_k v_k}^T] - E[\mathbf{h}_{s_k v_k}]E[\mathbf{h}_{s_k v_k}^T]$$

then (6.79) follows from Lemma 23 of Section 5.4.2.

Next, we have

$$\begin{aligned} J_k(\mathcal{T}_{r,k}) &= E\left[\|\mathbf{x}_k - a_k - \boldsymbol{\eta}_{s_k v_k}^T \mathbf{h}_{s_k v_k}\|^2\right] \\ &= E\left[\|(\mathbf{x}_k - E[\mathbf{x}_k]) + (E[\mathbf{x}_k] - a_k - \boldsymbol{\eta}_{s_k v_k}^T E[\mathbf{h}_{s_k v_k}]) \right. \\ &\qquad \left. - \boldsymbol{\eta}_{s_k v_k}^T (\mathbf{h}_{s_k v_k} - E[\mathbf{h}_{s_k v_k}])\|^2\right] \\ &= E\left[\|\hat{\mathbf{x}}_k + (\alpha_k - a_k) - \boldsymbol{\eta}_{s_k v_k}^T \hat{\mathbf{h}}_{s_k v_k}\|^2\right] \end{aligned}$$

where we use the standard notation $\hat{\mathbf{x}}_k = \mathbf{x}_k - E[\mathbf{x}_k]$ and $\hat{\mathbf{h}}_{s_k v_k} = \mathbf{h}_{s_k v_k} - E[\mathbf{h}_{s_k v_k}]$ and define $\alpha_k = E[\mathbf{x}_k] - \boldsymbol{\eta}^T_{s_k v_k} E[\mathbf{h}_{s_k v_k}]$. Hence

$$\begin{aligned}
J_k(\mathcal{T}_{r,k}) &= \mathrm{tr}E\Big[(\hat{\mathbf{x}}_k + (\alpha_k - a_k) - \boldsymbol{\eta}^T_{s_k v_k}\hat{\mathbf{h}}_{s_k v_k})(\hat{\mathbf{x}}_k + (\alpha_k - a_k) \\
&\qquad - \hat{\mathbf{h}}^T_{s_k v_k}\boldsymbol{\eta}_{s_k v_k})\Big] \\
&= \mathrm{tr}E\Big[\hat{\mathbf{x}}_k^2 - \hat{\mathbf{x}}_k\hat{\mathbf{h}}^T_{s_k v_k}\boldsymbol{\eta}_{s_k v_k} + (\alpha_k - a_k)^2 - \boldsymbol{\eta}^T_{s_k v_k}\hat{\mathbf{h}}_{s_k v_k}\hat{\mathbf{x}}_k^T \\
&\qquad + \boldsymbol{\eta}^T_{s_k v_k}\hat{\mathbf{h}}_{s_k v_k}\hat{\mathbf{h}}^T_{s_k v_k}\boldsymbol{\eta}_{s_k v_k}\Big] \\
&= E[\hat{\mathbf{x}}_k^2] + (\alpha_k - a_k)^2 + \mathrm{tr}\left[\eta^T_{s_k v_k}\mathbb{H}_{s_k v_k}\eta_{s_k v_k} - \mathbb{Q}_{s_k v_k}\eta_{s_k v_k}\right. \\
&\qquad \left. - \eta^T_{s_k v_k}\mathbb{Q}_{s_k v_k}\right] \\
&= \mathbb{E}_{x_k x_k} + (\alpha_k - a_k)^2 + \mathrm{tr}\left[(\eta^T_{s_k v_k}\mathbb{H}_{s_k v_k} - \mathbb{Q}_{s_k v_k})\right. \\
&\qquad \left. \times \mathbb{H}^\dagger_{s_k v_k}(\eta^T_{s_k v_k}\mathbb{H}_{s_k v_k} - \mathbb{Q}_{s_k v_k})^T - \mathbb{Q}_{s_k v_k}\mathbb{H}^\dagger_{s_k v_k}\mathbb{Q}^T_{s_k v_k}\right] \\
&= \mathbb{E}_{x_k x_k} + (\alpha_k - a_k)^2 + \|(\eta^T_{s_k v_k}\mathbb{H}_{s_k v_k} - \mathbb{Q}_{s_k v_k})\mathbb{H}^{1/2\,\dagger}_{s_k v_k}\|^2 \\
&\qquad - \mathbb{Q}_{s_k v_k}\mathbb{H}^\dagger_{s_k v_k}\mathbb{Q}^T_{s_k v_k}.
\end{aligned}$$

Therefore, $J_k(\mathcal{T}_{r,k})$ has a minimum possible value if

$$a_k^0 = \alpha_k^0 = E[\mathbf{x}_k] - \eta^{0\,T}_{s_k v_k} E[\mathbf{h}_{s_k v_k}]$$

and

$$(\eta^T_{s_k v_k}\mathbb{H}_{s_k v_k} - \mathbb{Q}_{s_k v_k})\mathbb{H}^{1/2\,\dagger}_{s_k v_k} = \mathbb{O}.$$

Similarly to Lemma 26 of Section 5.4.2, it can be shown that the latter equation is equivalent to the equation

$$\eta^T_{s_k v_k}\mathbb{H}_{s_k v_k} - \mathbb{Q}_{s_k v_k} = \mathbb{O},$$

which has the solution [6]

$$\eta^{0\,T}_{s_k v_k} = \mathbb{Q}_{s_k v_k}\mathbb{H}^\dagger_{s_k v_k} + M_k(I - \mathbb{H}_{s_k v_k}\mathbb{H}^\dagger_{s_k v_k}).$$

As a result,

$$J_k(\mathcal{T}_{r,k})(a_k^0, \eta^0_{s_k v_k}) = \mathbb{E}_{x_k x_k} - \mathbb{Q}_{s_k v_k}\mathbb{H}^\dagger_{s_k v_k}\mathbb{Q}^T_{s_k v_k} \tag{6.80}$$

is clearly the minimum possible value for $J_k(\mathcal{T}_{r,k})$. □

Remark 27. *The covariances used in* (6.77) *and* (6.78) *and in similar relationships below, are assumed to be known or to be estimated by the known methods considered, in particular, in Section* 4.3.

Theorem 50. *The error* $E[\|\mathbf{x}-\mathcal{T}_r^0(\mathbf{y})\|^2]$ *for any optimal filter* $\mathcal{T}_r^0$ *defined by* (6.68)*,* (6.77) *and* (6.78) *is*

$$\begin{aligned} E[\|\mathbf{x}-\mathcal{T}_r^0(\mathbf{y})\|^2] &= J(\mathcal{T}_r^0) \\ &= \sum_{k=1}^{m} J_k(\mathcal{T}_{r,k}^0) \\ &= \sum_{k=1}^{m} \left[\mathbb{E}_{x_k x_k} - \mathbb{Q}_{s_k v_k}\mathbb{H}_{s_k v_k}^{\dagger}\mathbb{Q}_{s_k v_k}^{T}\right]. \end{aligned} \tag{6.81}$$

Proof. The result follows from (6.80) in the previous proof. □

Corollary 12. *For each* $k = 1, \ldots, m$*, the error* $E[|\mathbf{x}_k-\mathcal{T}_{r,k}^0(\mathbf{y}_{s_k}, \ldots, \mathbf{y}_{v_k})|^2]$ *decreases as the degree* r *of* $\mathcal{T}_{r,k}^0(\mathbf{y}_{s_k}, \ldots, \mathbf{y}_{v_k})$ *increases.*

Proof. The proof follows directly from the proof of Theorems 49 and 50. □

6.5.7 Particular cases

The proposed approach generalizes the methods considered in the preceding Sections 5.4–5.6 and 6.4.3 as follows.

If $v_k = p_k = m$ in (6.55) for all $k = 1, \ldots, m$ then the solution to the problem (5.104) in Section 5.5.1 can be given in terms of (6.77)–(6.78). The solution to the problem (5.69) in Section 5.4.1 is a particular case of the model obtained in Section 5.5 and therefore, it can also been constructed in terms of (6.77)–(6.78).

If $v_k = p_k = k$ in (6.55) for all $k = 1, \ldots, m$ then the model (6.67)–(6.68) is causal with complete memory (see Definition 37 in Section 6.5.1) and hence, the causal model of Section 5.6 can be constructed from Theorem 49.

If for all $k = 1, \ldots, m$ in (6.55), $v_k = k$ and s_k is defined by (6.12) (with $i = k$ in (6.12)) then the linear filter with piecewise-constant memory of Section 6.4.3 also follows from Theorem 49 for $r = 1$ and $a_k = 0$.

The above demonstrates the flexibility of the proposed method and shows that the choice of s_k and v_k in (6.55) provides an additional degree of freedom for the method. At the same time, the present method may require more computational work than those in the preceding Sections.

6.5.8 Simulations

To illustrate the performance of the method, we apply the proposed filters to the problem of extracting information about images of the surface of the earth obtained by air observations. The reference matrix $\mathcal{X} \in \mathbb{R}^{256\times 256}$ to be estimated is a numerical representation of the image of a chemical plant[7]. We consider two different cases. In the first case the data is disturbed by additive noise and in the second by multiplicative noise. In each case the raw data set is represented by a matrix $\mathcal{Y} \in \mathbb{R}^{256\times 256}$. In the first case we set $\mathcal{Y} = \mathcal{Y}^{(1)}$ and in the second we set $\mathcal{Y} = \mathcal{Y}^{(2)}$ where

$$\mathcal{Y}^{(1)} = \mathcal{X} + 150\ R_1 \quad \text{and} \quad \mathcal{Y}^{(2)} = \mathcal{X} * R_2,$$

and where $R_1 \in \mathbb{R}^{256\times 256}$ and $R_2 \in \mathbb{R}^{256\times 256}$ are matrices of randomly generated numbers from independent uniform distributions on the interval $(0, 1)$. The symbol "$*$" denotes the Hadamard product[8]

Because the procedure is formally the same in each case we will give a generic description with $\mathcal{X}$ denoting the reference matrix that we wish to estimate and $\mathcal{Y}$ denoting the observed data. In each case we begin the analysis by partitioning $\mathcal{X}$ and $\mathcal{Y}$ into smaller blocks and we consider two different schemes. In the first instance we use 64 separate blocks with sub-matrices

$$\{X_{ij}\}_{i,j=1,\dots,32} \in \mathbb{R}^{32\times 32} \quad \text{and} \quad \{Y_{ij}\}_{i,j=1,\dots,32} \in \mathbb{R}^{32\times 32}$$

and in the second we use a more refined partition

$$\{X_{ij}\}_{i,j=1,\dots,16} \in \mathbb{R}^{16\times 16} \quad \text{and} \quad \{Y_{ij}\}_{i,j=1,\dots,16} \in \mathbb{R}^{16\times 16}$$

with 256 separate blocks. Since the procedure is essentially the same whichever scheme is used our subsequent description will not distinguish between the two.

To apply the estimation procedure to each fixed block (i, j) we set $X = X_{ij}$ and $Y = Y_{ij}$. The ℓ-th columns

$$x^{(\ell)} = x_{ij}^{(\ell)} = \mathbf{x}_{ij}(\omega_\ell) = \mathbf{x}(\omega_\ell)$$

and

$$y^{(\ell)} = y_{ij}^{(\ell)} = \mathbf{y}_{ij}(\omega_\ell) = \mathbf{y}(\omega_\ell)$$

of X and Y respectively are regarded as the ℓ-th realizations of the random vectors $\mathbf{x}$ and $\mathbf{y}$. To model the arbitrarily variable memory requirement we

[7]The data can be found in http://sipi.usc.edu/services/database/Database.html .
[8]If $A = \{a_{ij}\} \in \mathbb{R}^{n\times n}$ and $B = \{b_{ij}\} \in \mathbb{R}^{n\times n}$ then $A * B = \{a_{ij}b_{ij}\} \in \mathbb{R}^{n\times n}$.

assume that at each time k we can observe at most the seven most recent rows of data. Thus our estimate for the k-th element $\mathbf{x}_k$ can use only the observed elements $y_{s_k}, \ldots, y_{v_k}$ with $p = 7$. We have applied standard Matlab routines to compute these estimates using our proposed optimal causal filters of degrees one and two. For each $k = 1, \ldots, m$ and each $r = 1, 2$ the optimal filters are denoted by $T^0_{r,k}$ and are given by (6.68), (6.77) and (6.78) with $M_k = 0$ in the form

$$\begin{aligned} T^0_{r,k}(y_{s_k}, \ldots, y_{v_k}) &= a^0_k + \sum_{j=1}^{N_k} \boldsymbol{\eta}^{0\ T}_{js_kv_k}(\mathbf{h}_{js_kv_k}) \\ &= a^0_k + \boldsymbol{\eta}^{0\ T}_{s_kv_k}(\mathbf{h}_{s_kv_k}). \end{aligned}$$

Here, by (6.64), $N_k = N_k(r)$.

The covariances have been estimated from the samples using an elementary method [44]. We have used this method for illustrative purposes only. The results of the simulations are presented in Figures 6.5–6.7 and Table 1. For each case in Table 1 we write

$$\Delta_{r,ij} = \|X_{ij} - T^0_r(Y_{ij})\|^2$$

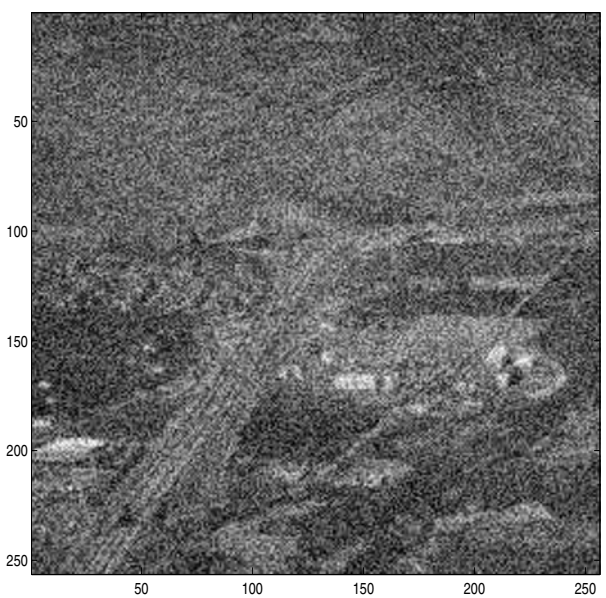
for $r = 1, 2$. The results are consistent with the theoretical analysis. Table 1 shows that the error associated with the second degree filter T^0_2 is less than that for the first degree filter T^0_1.

Table 1. Maximum errors for the proposed filters

	16×16 sub-matrices		32×32 sub-matrices	
	Errors by T^0_1 and T^0_2		Errors by T^0_1 and T^0_2	
Case	$\max_{ij} \Delta_{1,ij}$	$\max_{ij} \Delta_{2,ij}$	$\max_{ij} \Delta_{1,ij}$	$\max_{ij} \Delta_{2,ij}$
1	1.16×10^5	0.02×10^5	5.32×10^5	0.71×10^5
2	2.85×10^5	0.54×10^5	1.05×10^6	0.29×10^6

The proposed method has also been tested with other simulations including EEG data similar to that presented in [46]. Those tests were also consistent with the theoretical results obtained above. It is inappropriate to compare causal filters with arbitrarily variable memory to filters that are not restricted in this way. One would naturally expect unrestricted filters to exhibit superior performance but there are many realistic applications where such filters cannot be used.

(a) Reference signals.



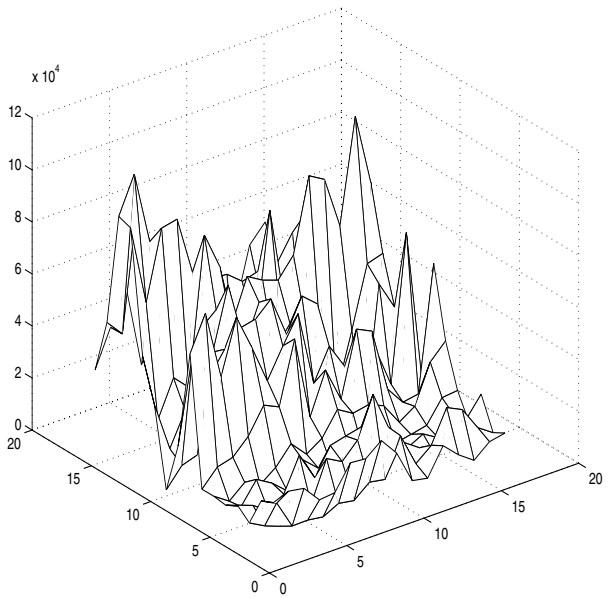
(b) Observed data.

Figure 6.5: Illustration to performance of method of Section 6.5. This digitized image has been taken from http://sipi.usc.edu/database/.

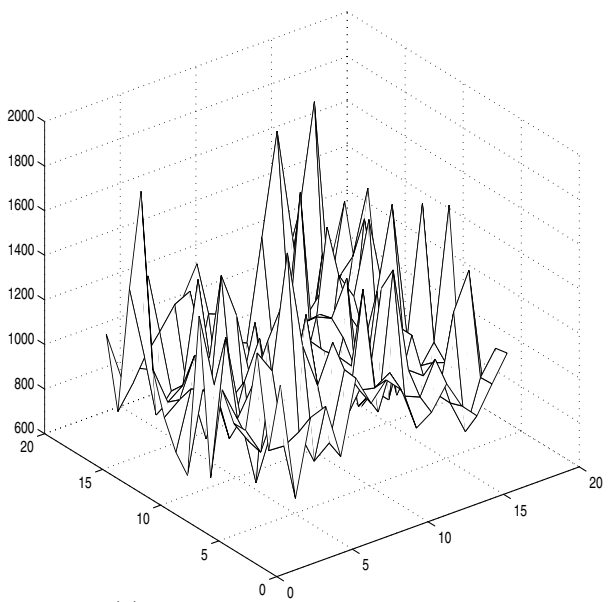
(a) 1st degree estimates.

(b) 2nd degree estimates.

Figure 6.6: Illustration to performance of method of Section 6.5. This digitized image has been taken from http://sipi.usc.edu/database/.



(a) Errors of the 1st degree estimates.



(b) Errors of the 2nd degree estimates.

Figure 6.7: Illustration to performance of method of Section 6.5.

6.6 Optimal Nonlinear Filtering with no Memory Constraint

The methods considered in the preceding chapters concern various types of approximation for operator $\mathcal{F}$ and its image $\mathcal{F}(\mathbf{x})$. If $\mathcal{F}$ is the identity operator then such methods can be interpreted as methods for optimal filtering. Below, we consider special types of optimal filters without restrictions associated with notions of causality and memory. We call them unconstrained filters.

6.6.1 Unconstrained polynomial filter

If $\mathcal{F}(\mathbf{x}) = \mathbf{x}$, then the polynomial operator $\tilde{\mathcal{P}}_r$ constructed in Section 5.5 of Chapter 4 is a model of the optimal filter. We illustrate the efficiency of such a filter by its applications to processing of data given by digitized color images.

A colour image is numerically represented by three matrices $X^{(1)}$, $X^{(2)}$, $X^{(3)} \in \mathbb{R}^{M\times N}$, where the elements in $X^{(1)}$ are interpreted as red intensities, in $X^{(2)}$ as green intensities, and in $X^{(3)}$ as blue intensities. The $M \times N \times 3$ tensor, composed from these matrices, is a numerical counterpart of the colour image. We denote such a tensor by $\mathcal{T} = \mathcal{T}(X^{(1)}, X^{(2)}, X^{(3)})$.

The known image 'Sailboat on lake' [9] has numerically been represented by the tensor $\mathcal{T}(X^{(1)}, X^{(2)}, X^{(3)})$ with $M = N = 256$. For each $k = 1, 2, 3$, matrix $X^{(k)}$ has been partitioned into 2,048 sub-matrices $X_{ij}^{(k)} \in \mathbb{R}^{4\times 8}$ with $i = 1, \ldots, 64$ and $j = 1, \ldots, 32$ so that $X^{(k)} = \{X_{ij}^{(k)}\}$. Each sub-matrix $X_{ij}^{(k)}$ has been interpreted as a set of eight realizations of a random vector with columns representing realizations.

We simulated observed data in the form $Y_{ij}^{(k)}$ presented in Table 2, where $\mathcal{R}_{ij}^{(k)}$ is a matrix with normally distributed entries with mean 0 and variance 1, and $\mathcal{Q}_{ij}^{(k)}$ and $\mathcal{Q}_{ij}^{(kk)}$ are matrices with uniformly distributed entries in the interval $(0, 1)$. In Table 2,

$$\Delta_{1,ij}^{(k)} = \|X_{ij}^{(k)} - L_{Q,ij}^{0(k)}(Y_{ij}^{(k)})\|^2, \quad \Delta_{2,ij}^{(k)} = \|X_{ij}^{(k)} - T_{2,ij}^{0(k)}(Y_{ij}^{(k)})\|^2,$$

and

$$\Delta_{ij}^{(k)} = \|X_{ij}^{(k)} - T_{ij}^{0(k)}(Y_{ij}^{(k)})\|^2,$$

where $L_{Q,ij}^{0(k)}(Y_{ij}^{(k)})$, $T_{2,ij}^{0(k)}(Y_{ij}^{(k)})$ and $T_{ij}^{0(k)}(Y_{ij}^{(k)})$ are the best first-degree, Hadamard-quadratic and multiquadratic estimates for $X_{ij}^{(k)}$ respectively.

[9] The database can be found in http://sipi.usc.edu/services/database/Database.html.

Estimates $L^{0(k)}_{Q,ij}(Y^{(k)}_{ij})$, $T^{0(k)}_{2,ij}(Y^{(k)}_{ij})$ and $T^{0(k)}_{ij}(Y^{(k)}_{ij})$ have been calculated from (5.100)-(5.101), (5.81)-(5.83) and (5.123)-(5.125), correspondingly, with Matlab for each i, j and k (i.e. the method has been applied 6,144 times). We put $M_1 = M_2 = K = \mathbb{O}$.

The expectations and covariance matrices in (5.100)-(5.101), (5.81)-(5.83) and (5.123)-(5.125) have been estimated from the maximum likelihood estimates considered in Section 5.3 of Chapter 4. For instance, for each i, j and k we estimated E_{xz_q} as $X^{(k)}_{ij}(Z^{(k)}_{ij})^T/8 - \hat{X}^{(k)}_{ij}(\hat{Z}^{(k)}_{ij})^T$, where $Z^{(k)}_{ij} = Y^{(k)}_{ij}(\text{diag}Y^{(k)}_{ij}(q,:))$, and $\text{diag}Y^{(k)}_{ij}(q,:))$ is a diagonal matrix with the elements from the qth row of $Y^{(k)}_{ij}$ on its diagonal, and $\hat{M}$ means a vector formed from means of the rows of the matrix M. These simple estimates have only been chosen to illustrate the performance of the proposed method.

Fig. 6.9 illustrates the performance of the method. The tensors for the digitized images in Fig. 6.9 have been composed from sub-matrices $Y^{(k)}_{ij}$, $L^{0(k)}_{Q,ij}(Y^{(k)}_{ij})$ and $T^{0(k)}_{ij}(Y^{(k)}_{ij})$ respectively.

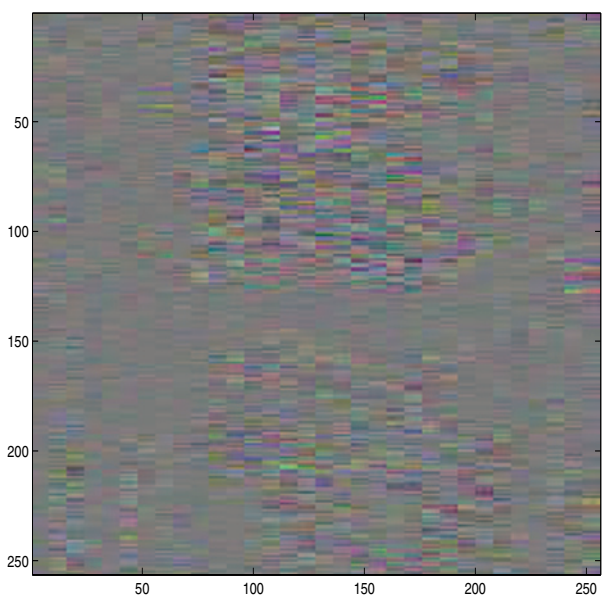
Diagrams in Fig. 6.9 (a) and (b) represent the matrices $\Delta^{(2)}_1 = \{\Delta^{(2)}_{1,ij}\}$ and $\Delta^{(2)} = \{\Delta^{(2)}_{ij}\}$ of errors associated with the best first-degree and multiquadratic estimates for the matrices $X^{(2)}_{ij}$, whose entries are interpreted as green intensities.

The estimates by Sorenson [142] cannot be applied here since the estimates of E_{yy} and $E[yy^T]$ are very close to singular and Matlab warned that calculations may be inaccurate

In Table 3 and Fig. 6.10 we present the results of similar simulations with the well known image 'Lenna' [2] given by a tensor $\mathcal{T}(X^{(1)}, X^{(2)}, X^{(3)})$ with $M = N = 256$ (i.e. the method has been applied 6,144 times again). The notation in Table 3 is the same as in Table 2. In these simulations, estimates by Sorenson [142] can be applied and they coincide with the best first-degree estimates $L^{0(k)}_{Q,ij}(Y^{(k)}_{ij})$.

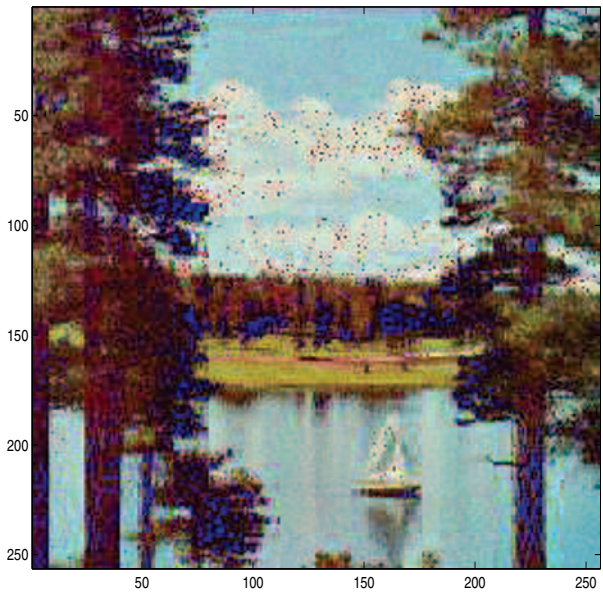
In all 12,288 applications, the best multiquadratic estimates give significant improvements in the accuracy of $X^{(k)}_{ij}$ estimation compared to the best first-degree estimates.

(a) Reference signals. This digitized image has been taken from http://sipi.usc.edu/database/.



(b) Observed data.

Figure 6.8: Illustration to unconstrained polynomial filter.



(a) The best first-degree estimate $L_Q^0(y)$.

(b) The best multiquadratic estimate $T^0(y)$.

Figure 6.9: Illustration to unconstrained polynomial filter.

(a) The best estimate by Sorenson [142].

(b) The best multiquadratic estimate $T^0(y)$.

Figure 6.10: Illustration to unconstrained polynomial filter. This digitized image has been taken from http://sipi.usc.edu/database/.

Table 2.

k	Observed data	$\max_{ij} \Delta_{1,ij}^{(k)}$	$\max_{ij} \Delta_{2,ij}^{(k)}$	$\max_{ij} \Delta_{ij}^{(k)}$
1	$Y_{ij}^{(1)} = 0.1\,\mathcal{R}_{ij}^{(1)} X_{ij}^{(1)} \mathcal{Q}_{ij}^{(1)} \mathcal{Q}_{ij}^{(11)}$	0.7005×10^5	0.00921×10^5	39.6747
2	$Y_{ij}^{(2)} = \mathcal{R}_{ij}^{(2)} X_{ij}^{(2)} \mathcal{Q}_{ij}^{(2)} \mathcal{Q}_{ij}^{(22)}$	2.0059×10^5	0.0452×10^5	12.1697
3	$Y_{ij}^{(3)} = 10\,\mathcal{R}_{ij}^{(3)} X_{ij}^{(3)} \mathcal{Q}_{ij}^{(3)} \mathcal{Q}_{ij}^{(33)}$	1.8218×10^5	0.0020×10^5	21.2261

Table 3.

k	Observed data	$\max_{ij} \Delta_{1,ij}^{(k)}$	$\max_{ij} \Delta_{2,ij}^{(k)}$	$\max_{ij} \Delta_{ij}^{(k)}$
1	$Y_{ij}^{(1)} = 20\,\mathcal{R}_{ij}^{(1)} X_{ij}^{(1)} \mathcal{Q}_{ij}^{(1)} + 50\,\mathcal{Q}_{ij}^{(11)}$	2.5324	0.1713×10^{-4}	0.0212×10^{-4}
2	$Y_{ij}^{(2)} = 10\,\mathcal{R}_{ij}^{(2)} X_{ij}^{(2)} \mathcal{Q}_{ij}^{(2)} \mathcal{Q}_{ij}^{(22)}$	3.1533	0.0034	0.0001
3	$Y_{ij}^{(3)} = \mathcal{Q}_{ij}^{(3)} X_{ij}^{(3)} \mathcal{R}_{ij}^{(3)}$	1.1997	0.0036	4.9951×10^{-7}

6.6.2 Unconstrained hybrid filter

The method of the best hybrid approximations considered in Section 5.7 of Chapter 4, can be extended to the case which we call the optimal hybrid filtering. The solution to the problem (5.186) – (5.187) in Section 5.7 of Chapter 4 for $\mathcal{F} = I$, i.e. the operator $\mathcal{P}_k$ presented by (5.169), (5.172), (5.189) with $\mathcal{F} = I$, represents a model of the filter. This filter is constructed from the consequent 'blocks' $\tilde{B}_0, \ldots, \tilde{B}_k$ in accordance with (5.169).

It is clear from Theorem 43 that the filter, defined by (5.169), (5.172), (5.189) with $\mathcal{F} = I$, possesses the obvious advantages over conventional filters based on a least-square approximation [63], [157]. In particular, even for $k = 0$, this filter produces a smaller associated error than that for filters [63], [157]. This is due to the higher degree r of approximation compared with the case of the approximation in [63], [157]. For the number of iteration k greater than zero, this error is further decreased.

At the same time, such a filtering can be considered as a special case of the method which we develop in Section 7.7 "Optimal generalized hybrid transform" of the next Chapter 7. Therefore, we refer to that section for more details.

6.7 Concluding Remarks

In this chapter, we have presented computational methods for optimal filtering of stochastic signals. The wide generalizations of the Wiener approach to linear filtering have been considered in both finite dimensional vector spaces and the Hilbert space. For different types of memory such as piecewise-constant memory and arbitrarily variable memory, models of optimal filters have been provided. Methods for optimal nonlinear filtering with no memory constraint have also been given. A rigorous theoretical analysis of the presented methods have been presented. Algorithms for numerical computation of the considered filters have been provided.

Chapter 7

Computational Methods for Optimal Compression and Reconstruction of Random Data

7.1 Introduction

In this chapter, we consider computational methods for simultaneous data dimensionality reduction and filtering, and subsequent data reconstruction

with the highest possible accuracy.

In signal processing, data dimensionality reduction (often called compression) is motivated by a necessity to diminish expenditures for transmission, processing and storage of large signal arrays. The known associated methods have also been applied successfully to the solution of problems related to clustering, feature selection and forecasting.

In statistics, data dimensionality reduction is often identified with a procedure for finding the so called principal components of a large random vector, i.e. of components of a smaller vector which preserves principal features of the original vector. In particular, this means that the original vector can be reconstructed from the smaller one with the least possible error.

Observed data is normally corrupted with random noise. Therefore, any procedure of data compression (or finding principal components) should be accompanied by filtering. We note that filtering and data compression could be separated. Nevertheless, simultaneous filtering and compression is more effective in the sense of minimizing the associated error (see [182], for example).

The known methods for filtering and data compression can be applied in either a probabilistic setting (as in [53]–[55], [63, 68, 79, 88, 92, 96, 133, 149, 150, 158, 166, 170, 182, 183]) or a deterministic setting (as in [21, 147]). The associated techniques are often based on the use of reduced-rank operators.

In this chapter, a further advance in the development of reduced-rank transforms is presented. We study a new approach to data dimensionality reduction in a *probabilistic* setting based on the development of ideas presented in [63, 133, 158, 166, 170, 183].

Computational methods considered below are based on solution of best approximation problems, special iterative procedures and their combination.

In Section 7.2, we present the standard Principal Component Analysis and Karhunen-Loève transform (PCA–KLT). In Section 7.4, this method is extended to more general cases. In Sections 7.5–7.7, more advanced techniques are described.

In Section 7.3, methods of rank-constrained matrix approximations are considered.

A so-called generic PCA–KLT is given in Section 7.4 and its generalizations are studied in Sections 7.5–7.7. The methods considered in Sections 7.5–7.7 are motivated by the following observation. In general, the reduced-rank transforms for random data (such as those in [63, 133, 158, 166, 170, 183]) consist of three companion operations – filtering, compression and reconstruction. Filtering and compression are performed simultaneously to estimate a reference signal $\mathbf{x}$ with m components from noisy observed

data $\mathbf{y}$ and to filter and reduce the data to a vector $\hat{\mathbf{x}}$ with η components, $\eta < m$. The components of $\hat{\mathbf{x}}$ are often called the principal components. The quotient η/m is called the compression ratio. Reconstruction returns a vector $\tilde{\mathbf{x}}$ with m components so that $\tilde{\mathbf{x}}$ should be close to the original $\mathbf{x}$. It is natural to perform these three operations so that the reconstruction error and the related computational burden are minimal. As a result, the performance of the reduced-rank transform is characterized by three issues which are (i) associated accuracy, (ii) compression ratio, and (iii) computational work. The methods presented in Sections 7.5–7.7 improve those issues compared to the techniques given in Sections 7.2 and 7.4.

7.2 Standard Principal Component Analysis and Karhunen-Loève Transform (PCA–KLT)

Jolliffe [68] writes: *'Principal component analysis is probably the oldest and best known of the techniques of multivariate analysis.'* Principal component analysis (PCA) was discovered by Pearson [100] in 1901 and then independently developed by Hotelling [55] in 1933, by Karhunen [71] in 1947 and by Loève [88] in 1948. Owing to its versatility in applications, PCA has been extended in many directions (see, in particular, [63], [96], [133], [182] and the corresponding bibliographies). In engineering literature, PCA is normally called the Karhunen-Loève transform (KLT). We use the abbreviation 'PCA–KLT' for this technique.

Note that PCA–KLT can be reformulated as a technique which provides the best linear estimator of given rank for a random vector (see [63], [134]). The error associated with the estimators [44], [63], [68], [96] based on PCA–KLT idea is the smallest in the corresponding class of linear estimators with the same rank. Nevertheless, the performance of these *linear* estimators may not be as good as required. See Sections 7.5–7.7 for more details.

PCA–KLT can be represented in the following way. Let

$$\mathbf{x} \in L^2(\Omega, \mathbb{R}^m), \quad E_{xx} = E[\mathbf{x}\mathbf{x}^T]$$

and let the spectral decomposition of E_{xx} be given by

$$E_{xx} = \sum_{j=1}^{m} \lambda_j u_j u_j^T,$$

where u_j and λ_j are corresponding eigenvectors and eigenvalues of E_{xx}, and E is the expectation operator.

Given $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$, PCA–KLT produces a linear operator $\mathcal{P}_0 : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^m)$ of maximum possible rank $r(\leq m)$ that mini-

mizes

$$\mathcal{J}(P) = E[\|\mathbf{x} - \mathcal{P}(\mathbf{x})\|^2]$$

over all linear operators $\mathcal{P} : L^2(\Omega, \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^m)$ of the same rank r.

Here, (see, for example, [78])

$$\text{rank}(\mathcal{P}) = \dim \mathcal{P}(L^2(\Omega, \mathbb{R}^m)).$$

The matrix P_0, associated with operator $\mathcal{P}_0$, is given by

$$P_0 = U_r U_r^T,$$

where $U_r = [u_1, u_2, \ldots, u_r]$.

Thus, U_r^T performs compression of $\mathbf{x}$ to a shorter vector in $L^2(\Omega, \mathbb{R}^r)$ and U_r performs a resonstruction of the compressed vector to $\hat{\mathbf{x}}$ so that

$$\hat{\mathbf{x}} = \mathcal{P}_0(\mathbf{x}).$$

Components of the compressed vector are called the principal components.

The compression ratio is given by

$$c = \frac{r}{m}, \tag{7.1}$$

where r is the number of principal components of vector $\mathbf{x}$.

7.3 Rank-constrained Matrix Approximations

7.3.1 Classical rank-constrained matrix approximation

We start with the classical result [19, 34] concerning determination of the matrix $X \in \mathbb{R}^{m\times n}$ of rank $= r$ that is nearest to matrix $A \in \mathbb{R}^{m\times n}$ in the Frobenius norm $\|\cdot\|$. The result presented in Theorem 51 below is known as the Eckart-Young theorem [34]. We note that the work [34] involves a number of extensions. We cite [43, 50, 76, 90] as some recent references.

Let the SVD of A be

$$U\Sigma V^T = A, \tag{7.2}$$

where $U = [u_1, \ldots, u_m] \in \mathbb{R}^{m\times m}$ and $V = [v_1, \ldots, v_n] \in \mathbb{R}^{n\times n}$ are orthogonal and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n}$ is diagonal where $p = \min\{m, n\}$, and $\sigma_1 \geq \ldots \geq \sigma_p \geq 0$.

Let

$$A_r = U_r \Sigma_r V_r^T, \tag{7.3}$$

where $U = [u_1, \ldots, u_r] \in \mathbb{R}^{m\times r}$, $V = [v_1, \ldots, v_r] \in \mathbb{R}^{n\times r}$ and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$.

Lemma 39. *Let $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ be orthogonal. Then for any matrix $M \in \mathbb{R}^{m\times n}$,*

$$\|U^T MV\|^2 = \|M\|^2.$$

Proof. Let

$$U = \{u_{ij}\}_{i,j=1}^{m}, \quad V = \{v_{ij}\}_{l,k=1}^{n} \quad \text{and} \quad M = \{m_{jk}\}_{j,k=1}^{m,n}.$$

Let us denote $Z = U^T MV$ so that $Z = \{z_{il}\}_{i,l=1}^{m}$. Then

$$z_{il} = \sum_{j=1}^{m}\sum_{k=1}^{n} u_{ij} m_{jk} v_{lk}$$

and

$$\begin{aligned}
\|U^T MV\|^2 &= \sum_{i=1}^{m}\sum_{l=1}^{m} z_{il}^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{q=1}^{m}\sum_{s=1}^{n}\sum_{l=1}^{n}(u_{ij}m_{jk}v_{lk})(u_{iq}m_{qs}v_{ls}) \\
&= \sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{q=1}^{m}\sum_{s=1}^{n} m_{jk}m_{qs}\sum_{i=1}^{m} u_{ij}u_{iq}\sum_{l=1}^{n} v_{lk}v_{ls} \\
&= \sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{q=1}^{m}\sum_{s=1}^{n} m_{jk}m_{qs}\delta_{jq}\delta_{ks} \\
&= \sum_{j=1}^{m}\sum_{k=1}^{n} m_{jk}^2 \\
&= \|M\|^2
\end{aligned}$$

as required. □

Lemma 40. *If $P \in \mathbb{R}^{m\times m}$ and $Q \in \mathbb{R}^{n\times n}$ are non-singular, then for any $X \in \mathbb{R}^{m\times n}$,*

$$\operatorname{rank}(PX) = \operatorname{rank}(XQ) = \operatorname{rank}(X).$$

Proof. All rows of PX are linear combinations of rows of X, therefore, the number of linearly independent rows in PX is not greater than the number of linearly independent rows in X, i.e.

$$\operatorname{rank}(PX) \le \operatorname{rank}(X).$$

Next, since P is non-singular, $X = P^{-1}(PX)$ implies $\operatorname{rank}(X) \leq \operatorname{rank}(PX)$. Hence,

$$\operatorname{rank}(PX) = \operatorname{rank}(X).$$

The proof of the equality

$$\operatorname{rank}(XQ) = \operatorname{rank}(X)$$

is similar. □

Lemma 41. *For any $W \in \mathbb{R}^{m\times n}$, and $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ orthogonal, there exists $M \in \mathbb{R}^{m\times n}$ such that*

$$U^T W V = M \Longleftrightarrow U M V^T = W.$$

Proof. The proof is obvious. □

Theorem 51. *Let A and A_r be as those in* (7.2) *and* (7.3)*, respectively. Then for any $X \in \mathbb{R}^{m\times n}$,*

$$\|A - A_r\|^2 = \min_{X:\ \operatorname{rank}(X)=r} \|A - X\|^2. \tag{7.4}$$

Proof. On the basis of Lemma 39,

$$\|U^T(A - X)V\|^2 = \|A - X\|^2.$$

We write $Y = U^T X V$. It follows from Lemma 40 that

$$\operatorname{rank}(X) = \operatorname{rank}(Y).$$

Therefore, the problem

$$\|U^T X V - \Sigma\|^2 \to \min_{X:\ \operatorname{rank}(X)=r}$$

is equivalent to the problem

$$\|Y - \Sigma\|^2 \to \min_{Y:\ \operatorname{rank}(Y)=r}. \tag{7.5}$$

This is true because, on the basis of Lemma 41, for every X there exists Y defined as above, and for every Y there exists X such that $X = UYV^T$.

The solution to (7.5) is $Y^0 = \Sigma_r$. Then (7.4) follows. □

7.3.2 Generalized rank-constrained matrix approximations

Justification of the techniques presented in Sections 7.5–7.7 below is based on the solution of generalized forms of the problem considered in Section 7.3.1 above. First, we study the following generalization.

Let $\mathbb{C}^{m\times n}$ be the set of $m \times n$ complex valued matrices and let

$$J(X) = \|M_1 - XM_2\|^2, \tag{7.6}$$

where $M_1 \in \mathbb{C}^{m\times n}$, $X \in \mathbb{C}^{m\times n}$ and $M_2 \in \mathbb{C}^{n\times n}$. Given M_1 and M_2, find X such that

$$J(X^0) = \min J(X) \tag{7.7}$$

subject to

$$\text{rank } X = r \leq \min\{m, n\}. \tag{7.8}$$

An elegant solution to this problem has been obtained by S. Friedland [43]. We present Friedland's result[1] in Theorem 52 below.

Theorem 52. *Let the SVD of M_2 be*

$$M_2 = U\Sigma V^T,$$

where $U \in \mathbb{C}^{n\times n}$ and $V \in \mathbb{C}^{n\times n}$ are orthogonal, and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_p, 0 \ldots, 0) \in \mathbb{C}^{n\times n}$. Let

$$M = M_1V \quad \textit{and} \quad M = [F \;\; G]$$

where $F \in \mathbb{C}^{n\times p}$ and $G \in \mathbb{C}^{n\times(n-p)}$. Let $F_{(r)} \in \mathbb{C}^{n\times p}$ be the best rank r approximation of F. Then a solution to the problem (7.7)–(7.8) is given by a family $\{X^0\}$ of matrices

$$X^0 = [F_{(r)}\Sigma_p^{-1} \;\; L]U^T, \tag{7.9}$$

where $\Sigma_p = \text{diag}(\sigma_1, \ldots, \sigma_p)$ and $L \in \mathbb{C}^{n\times(n-p)}$ with its columns in the column space of $F_{(r)}\Sigma_p^{-1}$ in order to satisfy the rank restriction.

Proof. We write

$$Y = XU \quad \text{and} \quad Y = [K \;\; L],$$

[1] The authors are grateful to S. Friedland for his generous consent to present the solution [43] here.

where $K \in \mathbb{C}^{n\times r}$ and $L \in \mathbb{C}^{n\times(n-r)}$. Then

$$\begin{aligned} J(X) &= \|M_1 - XM_2\|^2 \\ &= \|M_1V - XU\Sigma V^TV\|^2 \\ &= \|M - Y\Sigma\|^2 \\ &= \|M - [K\ \ L]\begin{bmatrix} \Sigma_p & \mathbb{O}_{12} \\ \mathbb{O}_{21} & \mathbb{O}_{22} \end{bmatrix}\|^2 \\ &= \|[F\ \ G] - [K\Sigma_p\ \ \mathbb{O}]\|^2 \\ &= \|[F - K\Sigma_p]\|^2 + \|G\|^2. \end{aligned} \tag{7.10}$$

where

$$\mathbb{O}_{12} \in \mathbb{C}^{p\times(n-p)}, \quad \mathbb{O}_{21} \in \mathbb{C}^{(n-p)\times p}, \quad \mathbb{O}_{22} \in \mathbb{C}^{(n-p)\times(n-p)}, \quad \mathbb{O} \in \mathbb{C}^{n\times(n-p)}$$

are the zero matrices. The term $\|G\|^2$ does not depend on X. The minimum of the term $\|[K\Sigma_p - F]\|^2$ subject to $\operatorname{rank}(K\Sigma_p) = r$ is attained when

$$K = K^0 \quad \text{where} \quad K^0 = F_{(r)}\Sigma_p^{-1}.$$

Since $X = YU^T$, we have

$$\begin{aligned} X^0 &= [K^0\ \ L]U^T \\ &= [F_{(r)}\Sigma_p^{-1}\ \ L]U^T. \end{aligned}$$

Here, $L = K^0S$ for some $S \in \mathbb{C}^{p\times(n-p)}$ so that[2]

$$\operatorname{rank}[K^0\ \ L] = \operatorname{rank}[K^0\ \ K^0S] = \operatorname{rank} K^0.$$

□

The above approach has been further developed by Friedland and Torokhti in [42] as follows.

Let $\mathbb{C}^{m\times n}$ be set of $m \times n$ complex valued matrices, and denote by $\mathcal{R}(m,n,k) \subseteq \mathbb{C}^{m\times n}$ the variety of all $m \times n$ matrices of rank k at most. Fix $A = [a_{ij}]_{i,j=1}^{m,n} \in \mathbb{C}^{m\times n}$. Then $A^* \in \mathbb{C}^{n\times m}$ is the conjugate transpose of A. Let the SVD of A be given by

$$A = U_A\Sigma_AV_A^*,$$

where $U_A \in \mathbb{C}^{m\times m}$ and $V_A \in \mathbb{C}^{n\times n}$ are unitary matrices and

$$\Sigma_A := \operatorname{diag}(\sigma_1(A), \ldots, \sigma_{\min(m,n)}(A)) \in \mathbb{C}^{m\times n}$$

[2]The matrix L must not increase the rank X above r. Hence, the columns of L must be linear combinations of the columns of K^0, i.e. $L = K^0S$ for some $S \in \mathbb{C}^{p\times(n-p)}$.

is a generalized diagonal matrix, with the singular values $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq 0$ on the main diagonal.

Let $U_A = [u_1\ u_2\ \ldots u_m]$ and $V_A = [v_1\ v_2\ \ldots v_n]$ be the representations of U and V in terms of their m and n columns, respectively, and let

$$P_{A,L} := \sum_{i=1}^{\text{rank } A} u_i u_i^* \in \mathbb{C}^{m\times m} \quad \text{and} \quad P_{A,R} := \sum_{i=1}^{\text{rank } A} v_i v_i^* \in \mathbb{C}^{n\times n}, \tag{7.11}$$

be the corresponding orthogonal projections onto the ranges of A and A^*. Define

$$A_k := (A)_k := \sum_{i=1}^{k} \sigma_i(A) u_i v_i^* = U_{Ak}\Sigma_{Ak}V_{Ak}^* \in \mathbb{C}^{m\times n} \tag{7.12}$$

for $k = 1, \ldots, \text{rank } A$, where

$$U_{Ak} = [u_1\ u_2\ \ldots u_k], \quad \Sigma_{Ak} = \text{diag}(\sigma_1(A), \ldots, \sigma_k(A)) \tag{7.13}$$

and

$$V_{Ak} = [v_1\ v_2\ \ldots v_k]. \tag{7.14}$$

For $k > \text{rank } A$, we write $A_k := A\ (= A_{\text{rank } A})$. For $1 \leq k < \text{rank } A$, the matrix A_k is uniquely defined if and only if $\sigma_k(A) > \sigma_{k+1}(A)$.

Below, we provide generalizations of both the classical minimal problem given in (7.4) and the problem given in (7.7)–(7.8). First, we present the result obtained in [42].

Theorem 53. [42] *Let matrices $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{m\times p}$ and $C \in \mathbb{C}^{q\times n}$ be given. Then*

$$X = B^\dagger (P_{B,L} A P_{C,R})_k C^\dagger \tag{7.15}$$

is a solution to the minimization problem

$$\min_{X\in\mathcal{R}(p,q,k)} ||A - BXC|| \tag{7.16}$$

with minimal norm $||X||$. This solution is unique if and only if either

$$k \geq \text{rank } (P_{B,L} A P_{C,R})$$

or

$$1 \leq k < \text{rank } (P_{B,L} A P_{C,R}) \quad \textit{and} \quad \sigma_k(P_{B,L} A P_{C,R}) > \sigma_{k+1}(P_{B,L} A P_{C,R}).$$

Proof. Recall that the Frobenius norm is invariant under multiplication from the left and the right by compatible unitary matrices. Hence

$$||A - BXC|| = ||\widetilde{A} - \Sigma_B \widetilde{X} \Sigma_C||,$$

where

$$\widetilde{A} := U_B^* A V_C \quad \text{and} \quad \widetilde{X} := V_B^* X U_C.$$

Clearly, X and $\widetilde{X}$ have the same rank and the same Frobenius norm. Thus, it is enough to consider the minimal problem

$$\min_{X \in \mathcal{R}(p,q,k)} ||\widetilde{A} - \Sigma_B \widetilde{X} \Sigma_C||.$$

Let $s = \text{rank } B$ and $t = \text{rank } C$. Clearly if B or C is a zero matrix, then $X = \mathbb{O}$ is the solution to the minimal problem (7.16). In this case either $P_{B,L}$ or $P_{C,R}$ are zero matrices, and the theorem holds trivially in this case.

Let us consider the case $1 \leq s, 1 \leq t$. Define

$$B_1 := \text{diag}(\sigma_1(B), \ldots, \sigma_s(B)) \in \mathbb{C}^{s \times s}$$

and

$$C_1 := \text{diag}(\sigma_1(C), \ldots, \sigma_t(C)) \in \mathbb{C}^{t \times t}.$$

Partition $\widetilde{A}$ and $\widetilde{X}$ into four block matrices A_{ij} and X_{ij} with $i, j = 1, 2$ so that $\widetilde{A} = [A_{ij}]_{i,j=1}^2$ and $\widetilde{X} = [X_{ij}]_{i,j=1}^2$, where $A_{11}, X_{11} \in \mathbb{C}^{s \times t}$. (For certain values of s and t, we may have to partition $\widetilde{A}$ or $\widetilde{X}$ to less than four block matrices.) Next, observe that

$$Z := \Sigma_B \widetilde{X} \Sigma_C = [Z_{ij}]_{i,j=1}^2,$$

where $Z_{11} = B_1 X_{11} C_1$ and all other blocks Z_{ij} are zero matrices. Since B_1 and C_1 are invertible we deduce

$$\text{rank } Z = \text{rank } Z_{11} = \text{rank } X_{11} \leq \text{rank } \widetilde{X} \leq k.$$

The approximation property of $(A_{11})_k$ yields the inequality

$$||A_{11} - Z_{11}|| \geq ||A_{11} - (A_{11})_k|| \tag{7.17}$$

for any Z_{11} of rank k at most. Hence, for any Z of the above form,

$$||\widetilde{A} - Z||^2 = ||A_{11} - Z_{11}||^2 + \sum_{2<i+j\leq 4} ||A_{ij}||^2 \geq ||A_{11} - (A_{11})_k||^2 + \sum_{2<i+j\leq 4} ||A_{ij}||^2.$$

Thus,

$$\widehat{X} = [X_{ij}]_{i,j=1}, \tag{7.18}$$

where

$$X_{11} = B_1^{-1}(A_{11})_k C_1^{-1} \quad \text{and} \quad X_{ij} = \mathbb{O} \quad \text{for all } (i,j) \neq (1,1), \tag{7.19}$$

is a solution to the problem

$$\min_{\widetilde{X} \in \mathcal{R}(p,q,k)} ||\widetilde{A} - \Sigma_B \widetilde{X} \Sigma_C|| \tag{7.20}$$

with the minimal Frobenius form. This solution is unique if and only if the solution

$$Z_{11} = (A_{11})_k$$

is the unique solution to the problem

$$\min_{Z_{11} \in \mathcal{R}(s,t,k)} ||A_{11} - Z_{11}||.$$

For $k \geq 1$, this happens if and only if $\sigma_k(A_{11}) > \sigma_{k+1}(A_{11})$.

Let us now show that

$$\widehat{X} = \Sigma_B^{\dagger}(P_{\Sigma_B,L}\widetilde{A}P_{\Sigma_C,R})_k\Sigma_C^{\dagger}. \tag{7.21}$$

On the basis of (7.18)-(7.19), we write

$$\begin{aligned}\widehat{X} &= \begin{bmatrix} B_1^{-1}(A_{11})_k C_1^{-1} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} \\ &= \begin{bmatrix} B_1^{-1} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} \begin{bmatrix} (A_{11})_k & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} \begin{bmatrix} C_1^{-1} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} \\ &= \Sigma_B^{\dagger} \begin{bmatrix} (A_{11})_k & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} \Sigma_C^{\dagger}. \end{aligned} \tag{7.22}$$

Here,

$$\begin{bmatrix} (A_{11})_k & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} = \left(\begin{bmatrix} A_{11} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} \right)_k. \tag{7.23}$$

To see this is true, we write the SVD of A_{11} as

$$A_{11} = U\Sigma_{st}V^*.$$

Then

$$\begin{bmatrix} A_{11} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} = \begin{bmatrix} U & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} \begin{bmatrix} \Sigma_{st} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} \begin{bmatrix} V^* & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}.$$

Therefore, we have

$$\begin{aligned}\left(\begin{bmatrix} A_{11} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}\right)_k &= \left(\begin{bmatrix} U & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}\begin{bmatrix} \Sigma_{st} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}\begin{bmatrix} V^* & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}\right)_k \\ &= \begin{bmatrix} U_k & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}\begin{bmatrix} \Sigma_{st,k} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}\begin{bmatrix} V_k^* & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} \\ &= \begin{bmatrix} (A_{11})_k & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix},\end{aligned}$$

where U_k, $\Sigma_{st,k}$ and V_k are truncated versions of U, Σ_{st} and V, respectively, defined similarly to U_{Ak}, Σ_{Ak} and V_{Ak} in (7.13). Thus, (7.23) is true.

Next, it follows from (7.19) and (7.22) that

$$\widehat{X} = \Sigma_B^\dagger \begin{bmatrix} A_{11} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}_k \Sigma_C^\dagger.$$

Let us write

$$\widehat{X} = \Sigma_B^\dagger \left(\begin{bmatrix} I_{ss} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} I_{tt} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}\right)_k \Sigma_C^\dagger$$

where I_{ss} is the $s \times s$ identity matrix. The SVD for $\Sigma_B = \begin{bmatrix} B_1 & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}$ is given by

$$\Sigma_B = U_{\Sigma_B} \begin{bmatrix} B_1 & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} V_{\Sigma_B}^*.$$

Here $U_{\Sigma_B} = [e_1, \ldots, e_m]$ and $V_{\Sigma_B} = [e_1, \ldots, e_p]$ where $e_j = [0, \ldots, 0, 1, 0, \ldots, 0]^T$ with 1 on the jth position. Therefore,

$$P_{\Sigma_B L} = \sum_{i=1}^{s} e_i e_i^T = \begin{bmatrix} I_{ss} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}$$

and by analogy,

$$P_{\Sigma_C R} = \sum_{i=1}^{t} e_i e_i^T = \begin{bmatrix} I_{tt} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}.$$

Thus,

$$\begin{bmatrix} A_{11} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} = P_{\Sigma_B L} \widetilde{A} P_{\Sigma_C R} \tag{7.24}$$

and (7.21) is true.

Next,

$$\begin{aligned} X &= V_B \widehat{X} U_C^* \\ &= V_B \Sigma_B^\dagger (P_{\Sigma_B,L} \widetilde{A} P_{\Sigma_C,R})_k \Sigma_C^\dagger U_C^* \\ &= B^\dagger U_B (P_{\Sigma_B L} U_B^* A V_C P_{\Sigma_C R})_k V_C^* C^\dagger \\ &= B^\dagger U_B (U_B^* Q V_C)_k V_C^* C^\dagger \end{aligned} \tag{7.25}$$

where

$$Q = P_{B,L} A P_{C,R}.$$

Let the SVD of Q be $Q = W \Sigma_Q Z^*$. Then we have

$$\begin{aligned} & U_B (U_B^* Q V_C)_k V_C^* = U_B (U_B^* W \Sigma_Q Z^* V_C)_k V_C^* \\ &= [u_1 \ \dots \ u_m] \left(\begin{bmatrix} u_1^* \\ \vdots \\ u_m^* \end{bmatrix} [w_1 \ \dots \ w_m] \Sigma_Q \begin{bmatrix} z_1^* \\ \vdots \\ z_n^* \end{bmatrix} [v_1 \ \dots \ v_n] \right)_k \begin{bmatrix} v_1^* \\ \vdots \\ v_n^* \end{bmatrix} \\ &= [u_1 \ \dots \ u_m] \begin{bmatrix} u_1^* w_1 & \dots & u_1^* w_k \\ \vdots & \vdots & \vdots \\ u_m^* w_1 & \dots & u_m^* w_k \end{bmatrix} \Sigma_{Qk} \begin{bmatrix} z_1^* v_1 & \dots & z_1^* v_n \\ \vdots & \vdots & \vdots \\ z_k^* v_1 & \dots & z_k^* v_n \end{bmatrix} \begin{bmatrix} v_1^* \\ \vdots \\ v_n^* \end{bmatrix} \end{aligned}$$

where Σ_{Qk} is a truncated version of Σ_Q constructed similar to Σ_{Ak} in (7.13). Therefore,

$$\begin{aligned} U_B (U_B^* Q V_C)_k V_C^* &= \left[(\sum_{i=1}^m u_i u_i^*) w_i \dots (\sum_{i=1}^m u_i u_i^*) w_k \right] \Sigma_{Qk} \\ &\qquad \times \begin{bmatrix} z_1^* \sum_{j=1}^n v_j v_j^* \\ \vdots \\ z_k^* \sum_{j=1}^n v_j v_j^* \end{bmatrix} \qquad (7.26) \\ &= [w_i \dots w_k] \Sigma_{Qk} \begin{bmatrix} z_1^* \\ \vdots \\ z_k^* \end{bmatrix} \\ &= (W \Sigma_Q Z^*)_k \\ &= Q_k. \qquad (7.27) \end{aligned}$$

As a result, it follows from (7.25), (7.26) and (7.27) that a solution of (7.16) with the minimal Frobenius norm is given by (7.15).

This solution is unique if and only if either $k \geq \text{rank } P_{B,L} A P_{C,R}$ or $1 \leq k < \text{rank } P_{B,L} A P_{C,R}$ and $\sigma_k(P_{B,L} A P_{C,R}) > \sigma_{k+1}(P_{B,L} A P_{C,R})$. □

Remark 28. *First observe that the classical approximation problem given by* (7.4) *is equivalent to the case* $m = p, n = q, B = I_{mm}, C = I_{nn}$. *(Here,* I_{mm} *is the* $m \times m$ *identity matrix.) Clearly* $P_{I_{mm},L} = I_{mm}$, $P_{I_{nn},R} = I_{nn}$. *In this case we obtain the classical solution* $B^{\dagger}(P_{B,L}AP_{C,R})_k C^{\dagger} = A_k$. *Second, if* $p = m$, $q = n$ *and* B, C *are non-singular, then* rank $(BXC) =$ rank X. *In this case,* $P_{B,L} = I_{mm}$ *and* $P_{C,R} = I_{nn}$, *and the solution to* (7.16) *is given by* $X = B^{-1}A_k C^{-1}$.

Next, the above Theorem 53 can be extended as follows.

It follows from (7.17)–(7.18) that a family of solutions to the minimization problem (7.20) with no requirement of a minimal Frobenius norm is given by

$$\widehat{X} = \begin{bmatrix} B_1^{-1}(A_{11})_k C_1^{-1} & X_{12} \\ X_{21} & X_{22} \end{bmatrix},$$

where X_{12}, X_{21} and X_{22} should be chosen in such a way that $\widehat{X} \in \mathcal{R}(p, q, k)$. In Theorem 54 below, we show how to choose X_{12}, X_{21} and X_{22} (see (7.30)) to satisfy the condition $\widehat{X} \in \mathcal{R}(p, q, k)$.

Theorem 54. *Let* $P \in \mathbb{C}^{(p-s)\times s}$, $Q \in \mathbb{C}^{t\times(q-t)}$ *be arbitrary matrices,* $V_B = [V_1\ V_2] \in \mathbb{C}^{p\times p}$, $U_C = [U_1\ U_2] \in \mathbb{C}^{q\times q}$ *where* $V_1 \in \mathbb{C}^{p\times s}$ *and* $U_1 \in \mathbb{C}^{q\times t}$, *and*

$$K = [B^{\dagger}B,\ \ I - B^{\dagger}B] \begin{bmatrix} \mathbb{O} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} CC^{\dagger} \\ I - CC^{\dagger} \end{bmatrix}, \tag{7.28}$$

where

$$K_{12} = V_1 X_{12} U_2^*, \quad K_{21} = V_2 X_{21} U_1^*, \quad K_{22} = V_2 X_{22} U_2^*, \tag{7.29}$$

$$X_{12} = X_{11}Q, \quad X_{21} = PX_{11}, \quad X_{22} = PX_{11}Q \tag{7.30}$$

and [3]

$$X_{11} = B_1^{-1}(A_{11})_k C_1^{-1}. \tag{7.31}$$

If the constraint of the minimal $\|X\|$ *is omitted in the problem* (7.16) *then its solution is not unique. A family of solutions to the problem* (7.16) *without this constraint is given by*

$$X = B^{\dagger}(P_{B,L}AP_{C,R})_k C^{\dagger} + K \tag{7.32}$$

with K *defined by* (7.28)–(7.31).

[3] *We note that matrices* K_{12}, K_{21} *and* K_{22} *depend on arbitrary matrices* P *and* Q.

Proof. To preserve rank $\widehat{X} \leq k$, we should choose X_{12}, X_{21} and X_{22} in a compatible form. In particular, the columns of X_{12} and the rows of X_{21} must be linear combinations of the columns and the rows of X_{11}, respectively, while for X_{22}, both the rows and columns must be linear combinations of the rows and columns of X_{11}. To this end, we need to show that there exist matrices $P \in \mathbb{C}^{(p-s)\times s}$ and $Q \in \mathbb{C}^{t\times(q-t)}$ such that, for $(i,j) \neq (1,1)$, X_{ij} can be written in the form (7.30). The existence follows from the next observation. By Gaussian elimination, there are matrices $P \in \mathbb{C}^{(p-s)\times s}$ and $Q \in \mathbb{C}^{t\times(q-t)}$ such that

$$\begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} I & -Q \\ \mathbb{O} & I \end{bmatrix} = \begin{bmatrix} X_{11} & \mathbb{O} \\ X_{21} & X_{22} - X_{21}Q \end{bmatrix}$$

and

$$\begin{bmatrix} I & \mathbb{O} \\ -P & I \end{bmatrix} \begin{bmatrix} X_{11} & \mathbb{O} \\ X_{21} & X_{22} - X_{21}Q \end{bmatrix} = \begin{bmatrix} X_{11} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}$$

which is true if

$$X_{12} - X_{11}Q = \mathbb{O}, \quad X_{21} - PX_{11} = \mathbb{O} \quad \text{and} \quad X_{22} - X_{21}Q = \mathbb{O}. \tag{7.33}$$

The last condition in (7.33) follows from the observation that the matrices $\begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$ and $\begin{bmatrix} X_{11} & \mathbb{O} \\ \mathbb{O} & X_{22} - X_{21}Q \end{bmatrix}$ have the same rank. Therefore, to ensure rank $\widehat{X} \leq k$, we must choose $X_{22} - X_{21}Q = \mathbb{O}$.[4] Thus, (7.30) follows.

Next, let us write

$$\begin{aligned} \widehat{X} &= \begin{bmatrix} B_1^{-1}(A_{11})_k C_1^{-1} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} + \begin{bmatrix} \mathbb{O} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \\ &= \Sigma_B^{\dagger}(P_{\Sigma_B,L}\widetilde{A}P_{\Sigma_C,R})_k\Sigma_C^{\dagger} + \begin{bmatrix} \mathbb{O} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}. \end{aligned}$$

This is true because of (7.21)–(7.24). Therefore,

$$\begin{aligned} X &= V_B\widehat{X}U_C^* \\ &= B^{\dagger}(P_{B,L}AP_{C,R})_kC^{\dagger} + [V_1\ V_2]\begin{bmatrix} \mathbb{O} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}\begin{bmatrix} U_1^* \\ U_2^* \end{bmatrix} \\ &= B^{\dagger}(P_{B,L}AP_{C,R})_kC^{\dagger} + V_2X_{21}U_1^* + V_1X_{12}U_2^* + V_2X_{22}U_2^*. \end{aligned}$$

[4]Otherwise, the columns of $\begin{bmatrix} \mathbb{O} \\ X_{22} - X_{21}Q \end{bmatrix}$ and the rows of $[\mathbb{O}\ \ X_{22} - X_{21}Q]$ are linear combinations of the columns $\begin{bmatrix} X_{11} \\ \mathbb{O} \end{bmatrix}$ and the rows of $[X_{11}\ \ \mathbb{O}]$, respectively, and then rank $\widehat{X}$ could be grater than k.

We note that

$$V_1^*V_1 = I_{ss}, \quad V_2^*V_2 = I_{p-s,p-s}, \quad U_1^*U_1 = I_{tt}, \quad U_2^*U_2 = I_{q-t,q-t},$$

$$V_1V_1^* + V_2V_2^* = I_{pp}, \quad U_1U_1^* + U_2U_2^* = I_{qq}$$

and

$$B^\dagger B = V_1V_1^* \quad \text{and} \quad CC^\dagger = U_1U_1^*.$$

Thus,

$$\begin{aligned} X &= B^\dagger(P_{B,L}AP_{C,R})_kC^\dagger + (I - V_1V_1^*)K_{21}U_1U_1^* \\ &+ V_1V_1^*K_{12}(I - U_1U_1^*) + (I - V_1V_1^*)K_{22}(I - U_1U_1^*) \\ &= B^\dagger(P_{B,L}AP_{C,R})_kC^\dagger + (I - B^\dagger B)K_{21}CC^\dagger \\ &+ +B^\dagger BK_{12}(I - CC^\dagger) + (I - B^\dagger B)K_{22}(I - CC^\dagger) \\ &= B^\dagger(P_{B,L}AP_{C,R})_kC^\dagger + K. \end{aligned} \tag{7.34}$$

Thus, (7.32) is true. □

Remark 29. *Let us set $p = m$, $q = n$, $A = M_1 \in \mathbb{C}^{m\times n}$, $B = I_{mm}$, $C = M_2 \in \mathbb{C}^{n\times n}$ and $\widetilde{A} = AV_C = [\widetilde{A}_1\ \widetilde{A}_2]$ where $\widetilde{A}_1 \in \mathbb{C}^{m\times t}$. Let $Q \in \mathbb{C}^{t\times(n-t)}$ be arbitrary. Then Theorem* 52 *follows as a special case of Theorem* 54 *in that the solution to the problem*

$$\min_{X\in\mathcal{R}(m,n,k)} ||A - XC|| \tag{7.35}$$

is given by

$$X = (AP_{C,R})_kV_C^*C^\dagger + (\widetilde{A}_1)_kC_1^{-1}QU_2^*(I - CC^\dagger). \tag{7.36}$$

Indeed, a solution to the problem

$$\min_{\widetilde{X}\in\mathcal{R}(m,n,k)} ||\widetilde{A} - \widetilde{X}\Sigma_C|| \tag{7.37}$$

where $\widetilde{X} = [X_1\ X_2]$, is $\widehat{X} = [\widehat{X}_1\ \widehat{X}_2]$ with $\widehat{X}_1 = (\widetilde{A}_1)_kC_1^{-1}$ and $\widehat{X}_2 = \widehat{X}_1Q$. Then $X = \widehat{X}U_C^*$ is the solution to the problem (7.35). *We have $\widehat{X} = [\widehat{X}_1\ \mathbb{O}] + [\mathbb{O}\ \widehat{X}_2]$ where*

$$\widehat{X}_1 = [(\widetilde{A}_1)_k\ \mathbb{O}]\begin{bmatrix} C_1^{-1} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} = [(\widetilde{A}_1)_k\ \mathbb{O}]\Sigma_C^\dagger = ([\widetilde{A}_1\ \mathbb{O}])_k\Sigma_C^\dagger$$

with $[\widetilde{A}_1\ \mathbb{O}] = [\widetilde{A}_1\ \widetilde{A}_2]\begin{bmatrix} I_{tt} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} = \widetilde{A}P_{\Sigma_CR}$. Therefore,

$$X = (\widetilde{A}P_{\Sigma_CR})_k\Sigma_C^\dagger U_C^* + [\mathbb{O},\ \widehat{X}_1Q]U_C^*.$$

Here,

$$(\widetilde{A}P_{\Sigma_C R})_k\Sigma_C^\dagger U_C^* = (\widetilde{A}P_{\Sigma_C R})_k V_C^* C^\dagger \quad \textit{and} \quad [\mathbb{O},\ \widehat{X}_1 Q]U_C^* = \widehat{X}_1 Q U_2^*.$$

Next, let $K_2 = X_2 U_2^*$. *Then*

$$X_2 U_2^* = K_2 U_2 U_2^* = K_2(I - U_1 U_1^*) = K_2(I - CC^\dagger).$$

Thus,

$$X = (\widetilde{A}P_{\Sigma_C R})_k\Sigma_C^\dagger U_C^* + \widehat{X}_1 Q U_2^*(I - CC^\dagger)$$

and (7.36) *follows.*

Remark 30. *The problems in the next Sections are reduced to problems which are similar to the one considered in Theorem* 54 *with* $p = m$, $q = n$, $A = E_{xy}E_{yy}^{1/2\dagger} \in \mathbb{C}^{m\times n}$, $B = I_{mm}$ *and* $C = E_{yy}^{1/2} \in \mathbb{C}^{n\times n}$ *where we write* $E_{yy}^{1/2\dagger} = (E_{yy}^{1/2})^\dagger$.

Let the SVD of $E_{yy}^{1/2}$ *be given by* $E_{yy}^{1/2} = U_C \Sigma U_C^*$ *and let*

$$\operatorname{rank} E_{yy}^{1/2} = r.$$

We write $U_C = [U_1\ U_2]$ *where* $U_1 \in \mathbb{C}^{n\times r}$, *and* $E_{xy}E_{yy}^{1/2\dagger}U_C = [\widetilde{A}_1\ \widetilde{A}_2]$ *where* $\widetilde{A}_1 \in \mathbb{C}^{m\times r}$.

Let $Q \in \mathbb{C}^{r\times(n-r)}$ *be arbitrary. By Remark* 29, *the solution to this particular case of the problem* (7.16), *without the constraint for the minimal Frobenius norm, is given by*

$$X = (E_{xy}E_{yy}^{1/2\dagger}U_1U_1^*)_k E_{yy}^{1/2\dagger} + K,$$

where

$$K = (\widetilde{A}_1)_k C_1^{-1} Q U_2^*(I - E_{yy}^{1/2}E_{yy}^{1/2\dagger}),\ \ C_1 = \operatorname{diag}(\sigma_1(C), \ldots, \sigma_r(C)) \in \mathbb{C}^{r\times r}$$

and

$$U_2 \in \mathbb{C}^{n\times(n-r)}.$$

Here, $(\widetilde{A}_1)_k$ *is defined similarly to* $(A_{11})_k$ *in the proof of Theorem* 53, *i.e. via a truncated SVD for* A_1 *defined by* (7.12). *Moreover,* $E_{yy}^{1/2\dagger}U_1U_1^* = E_{yy}^{1/2\dagger}$ *and by Lemma* 43 *below,* $E_{yy}^{1/2}E_{yy}^{1/2\dagger} = E_{yy}E_{yy}^\dagger$. *Therefore,*

$$X = (E_{xy}E_{yy}^{1/2\dagger})_k E_{yy}^{1/2\dagger} + K_1, \tag{7.38}$$

where

$$K_1 = (\widetilde{A}_1)_k C_1^{-1} Q U_2^*(I - E_{yy}E_{yy}^\dagger). \tag{7.39}$$

For the case of the minimal Frobenius norm, the unique solution is given by

$$X = (E_{xy}E_{yy}^{1/2\dagger})_k E_{yy}^{1/2\dagger}$$

and the conditions for uniqueness follow directly from Theorem 53.

Corollary 13. *If the constraints of* $\operatorname{rank} X \leq k$ *and of the minimal* $\|X\|$ *are omitted in the problem* (7.16) *then its solution is not unique and the family of solutions is given by*

$$X = B^\dagger AC^\dagger + K \tag{7.40}$$

where $K \in \mathbb{C}^{m\times n}$ *is given by* (7.28)–(7.29) *with* X_{11} *as in* (7.31)*, and* X_{12}*,* X_{21} *and* X_{22} *arbitrary. The minimum is given by*

$$\min_X \|A - BXC\|^2 = \|A - BB^\dagger AC^\dagger C\|. \tag{7.41}$$

Proof. If the constraint of $\operatorname{rank} X \leq k$ is omitted then the inequality (7.17) turns to the equality

$$\min_{Z_{11}} \|A_{11} - Z_{11}\| = 0$$

with $Z_{11} = A_{11}$ and $(A_{11})_k = A_{11}$. Then it follows from (7.32) that the unconstrained minimum $\min_X \|A - BXC\|^2$ is achieved if

$$X = B^\dagger P_{B,L} A P_{C,R} C^\dagger + K. \tag{7.42}$$

Because the constraint that $\operatorname{rank} X \leq k$ is omitted, matrices X_{12}, X_{21} and X_{22} in the representation of the matrix K need not be defined by (7.30) and indeed are arbitrary.

Next, we have

$$B^\dagger P_{B,L} = B^\dagger \quad \text{and} \quad P_{C,R} C^\dagger = C^\dagger$$

therefore, (7.42) implies (7.40).

The expression (7.41) follows directly from (7.40). □

Corollary 14. *The following is true:*

$$B^\dagger BKCC^\dagger = \mathbb{O}. \tag{7.43}$$

Proof. The proof follows immediately from (7.28). □

Remark 31. *Corollary* 14 *implies an interesting link between* (7.40) *and the-solution to the equation* $A - BXC = \mathbb{O}$. *The equality* (7.41) *implies that if* $BB^{\dagger}AC^{\dagger}C = A$ *then* $\min_X \|A - BXC\|^2 = 0$ *because* $A - BXC = \mathbb{O}$ *for* X *given by* (7.40). *It follows from* (7.40) *and* (7.43) *that* X *can be written as* $X = B^{\dagger}AC^{\dagger} + K - B^{\dagger}BKCC^{\dagger}$ *and that in this particular case,* X *formally coincides with the solution* [6] *to the equation* $A - BXC = \mathbb{O}$. *The equation* $A - BXC = \mathbb{O}$ *is consistent if and only if* $BB^{\dagger}AC^{\dagger}C = A$.

7.4 A Generic Principal Component Analysis and Karhunen-Loève Transform

7.4.1 The generic PCA–KLT

Scharf [134] presented an extension of PCA–KLT for the case of minimizing $\mathcal{J}(P)$ given by

$$\mathcal{J}(P) = E[\|\mathbf{x} - \mathcal{P}(\mathbf{y})\|^2] \tag{7.44}$$

subject to

$$\operatorname{rank}(\mathcal{P}) \leq r \leq m, \tag{7.45}$$

where $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$, $\mathbf{y} \in L^2(\Omega, \mathbb{R}^m)$ and the covariance matrix $E[\mathbf{y}\mathbf{y}^T]$ is nonsingular. The difference from the standard PCA-KLT is that $\mathcal{P}$ transforms an arbitrary $\mathbf{y}$, not $\mathbf{x}$.

Yamashita and Ogawa [183] proposed and justified a version of PCA–KLT for the case where $E[\mathbf{y}\mathbf{y}^T]$ is singular and $\mathbf{y} = \mathbf{x} + \mathbf{w}$ with $\mathbf{w}$ an additive noise.

Hua and Liu [63] considered PCA–KLT with a formal replacement of the inverse of matrix $E[\mathbf{y}\mathbf{y}^T]$ by its pseudo-inverse.

The general form of PCA–KLT in terms of the pseudo-inverse is given in Theorem 55 below.

An attractive feature of the methods [63], [183] is that invertibility of the covariance matrix $E[\mathbf{y}\mathbf{y}^T]$ is not assumed. Some other known extensions of PCA–KLT work under the condition that $E[\mathbf{y}\mathbf{y}^T]$ is nonsingular, and this restriction can impose limitations on the applicability of the method. In many practical situations, the matrix $E[\mathbf{y}\mathbf{y}^T]$ is singular. See, for example, [158, 166] and [170] in this regard.

Here, we give a rigorous generalization of the methods [63, 134, 183].

For $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$ and $\mathbf{y} \in L^2(\Omega, \mathbb{R}^n)$, we wish to find a linear operator $\mathcal{F}^0 : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^m)$ such that

$$J(F^0) = \min_F J(F) \tag{7.46}$$

subject to

$$\text{rank } \mathcal{F} \leq k \leq \min\{m, n\}. \tag{7.47}$$

Here, $F \in \mathbb{R}^{m\times n}$,

$$J(F) = E[\|\mathbf{x} - \mathcal{F}(\mathbf{y})\|^2], \tag{7.48}$$

and we write $\mathcal{F}$ instead of $\mathcal{P}$ in (7.44) to distinguish between the model which follows and the results in [63, 134, 183] associated with the problem (7.44)–(7.45).

Solution to the problem (7.46)–(7.47)

Now, we are in the position to give a solution to the problem (7.46)–(7.47).

Let us denote

$$A = E_{xy}(E_{yy}^{1/2})^{\dagger} \quad \text{and} \quad C = E_{yy}^{1/2}.$$

Let

$$\text{rank } A = l$$

and let the SVD of A be given by

$$A = U_A \Sigma_A V_A^T \tag{7.49}$$

where

$$U_A = [g_1, \ldots, g_n] \in \mathbb{R}^{m\times m} \quad \text{and} \quad V_A = [q_1, \ldots, q_n] \in \mathbb{R}^{n\times n}$$

are orthogonal matrices and

$$\Sigma_A = \text{diag}(\sigma_1(A), \ldots, \sigma_{\min(m,n)}(A)) \in \mathbb{R}^{m\times n}$$

is a generalized diagonal matrix with $\sigma_1(A) \geq \cdots \geq \sigma_l(A) > 0$ and $\sigma_{l+1}(A) = \cdots = \sigma_{\min(m,n)}(A) = 0$ on the main diagonal. Put

$$U_{Ak} = [g_1, \ldots, g_k], \quad V_{Ak} = [q_1, \ldots, q_k] \text{ and } \Sigma_{Ak} = \text{diag}(\sigma_1(A), \ldots, \sigma_k(A)).$$

We write A_k for the truncated SVD defined as

$$A_k = U_{Ak} \Sigma_{Ak} V_{Ak}^T \tag{7.50}$$

and denote

$$A_k = (E_{xy}(E_{yy}^{1/2})^{\dagger})_k.$$

Let

$$\text{rank } C = r.$$

The SVD of matrix C,

$$C = U_C \Sigma_C U_C^T,$$

is defined in a manner similar to the above.

As before, we also denote

$$\mathbb{E}_{xy} = E[\mathbf{x}\mathbf{y}^T] - E[\mathbf{x}]E[\mathbf{y}^T].$$

Lemma 42. *The following is true:*

$$A_k(\mathbb{E}_{yy}^{1/2})^\dagger \mathbb{E}_{yy}^{1/2} = A_k. \tag{7.51}$$

Proof. As an extension of the technique presented in the proving Lemmata 23 and 24 in Section 5.4.2, it can be shown that for any matrices $Q_1, Q_2 \in \mathbb{R}^{m\times n}$,

$$\mathcal{N}(Q_1) \subseteq \mathcal{N}(Q_2) \quad \Rightarrow \quad Q_2(I - Q_1^\dagger Q_1) = \mathbb{O}, \tag{7.52}$$

where $\mathcal{N}(Q_i)$ is the null space of Q_i for $i = 1, 2$. In regard to the equation (7.51),

$$\mathcal{N}([\mathbb{E}_{yy}^{1/2}]^\dagger) \subseteq \mathcal{N}(\mathbb{E}_{xy}[\mathbb{E}_{yy}^{1/2}]^\dagger). \tag{7.53}$$

The definition of A_k implies that

$$\mathcal{N}(\mathbb{E}_{xy}[\mathbb{E}_{yy}^{1/2}]^\dagger) \subseteq \mathcal{N}(A_k)$$

and

$$\mathcal{N}([\mathbb{E}_{yy}^{1/2}]^\dagger) \subseteq \mathcal{N}(A_k).$$

On the basis of (7.52), the latter implies

$$A_k[I - (\mathbb{E}_{yy}^{1/2})^\dagger \mathbb{E}_{yy}^{1/2}] = \mathbb{O},$$

i.e. (7.51) is true. □

Similarly to Remark 30, we write

$$C_1 = \mathrm{diag}(\sigma_1(C), \ldots, \sigma_r(C)) \in \mathbb{R}^{r\times r},$$

$$U_C = [U_1\ U_2] \quad \text{where} \quad U_1 \in \mathbb{R}^{n\times r},$$

and

$$E_{xy}E_{yy}^{1/2\dagger}U_C = [\widetilde{A}_1\ \widetilde{A}_2]$$

where $\widetilde{A}_1 = E_{xy}E_{yy}^{1/2\dagger}U_1 \in \mathbb{R}^{m\times r}$ and $\widetilde{A}_2 = E_{xy}E_{yy}^{1/2\dagger}U_2 \in \mathbb{R}^{m\times(n-r)}$.

Lemma 43. *For any* $\mathbf{v} \in L^2(\Omega, \mathbb{R}^n)$,

$$(\mathbb{E}_{vv}^{1/2})^\dagger \mathbb{E}_{vv}^{1/2} = \mathbb{E}_{vv}^{1/2} (\mathbb{E}_{vv}^{1/2})^\dagger, \tag{7.54}$$

$$\mathbb{E}_{vv}^{1/2} (\mathbb{E}_{vv}^{1/2})^\dagger = \mathbb{E}_{vv} \mathbb{E}_{vv}^\dagger \quad \textit{and} \quad \mathbb{E}_{vv}^\dagger \mathbb{E}_{vv} = \mathbb{E}_{vv} \mathbb{E}_{vv}^\dagger. \tag{7.55}$$

Proof. Let the SVD for $\mathbb{E}_{vv}$ be given by

$$\mathbb{E}_{vv} = V \Sigma V^T,$$

where V is the orthogonal matrix and $\Sigma = \text{diag}(\alpha_1, \ldots, \alpha_q, 0, \ldots, 0)$ with $\alpha_1 \geq \ldots \geq \alpha_q > 0$ and q the rank of $\mathbb{E}_{vv}$. Then

$$\mathbb{E}_{vv}^\dagger = V \Sigma^\dagger V^T, \quad \mathbb{E}_{vv}^{1/2} = V \Sigma^{1/2} V^T \quad \text{and} \quad (\mathbb{E}_{vv}^{1/2})^\dagger = V (\Sigma^{1/2})^\dagger V^T,$$

where

$$\Sigma^\dagger = \text{diag}(\alpha_1^{-1}, \ldots, \alpha_q^{-1}, 0, \ldots, 0), \quad \Sigma^{1/2} = \text{diag}(\alpha_1^{1/2}, \ldots, \alpha_q^{1/2}, 0, \ldots, 0)$$

and

$$(\Sigma^{1/2})^\dagger = \text{diag}(\alpha_1^{-1/2}, \ldots, \alpha_q^{-1/2}, 0, \ldots, 0).$$

Thus, (7.54) follows. The proof of the relationships (7.55) is similar. □

Theorem 55. *The solution to the problem* (7.46)–(7.47) *is not necessarily unique and is given in general by a family of operators* $\{\mathcal{F}^0\}$ *determined by the family* $\{F^0\}$ *of matrices*

$$F^0 = (E_{xy} E_{yy}^{1/2\dagger})_k E_{yy}^{1/2\dagger} + (\widetilde{A}_1)_k C_1^{-1} Q U_2^T (I - E_{yy} E_{yy}^\dagger), \tag{7.56}$$

where $Q \in \mathbb{R}^{r \times (n-r)}$ *is an arbitrary matrix.*

The error associated with the operator $\mathcal{F}^0$ *is given by*

$$\begin{aligned} E[\|\mathbf{x} - \mathcal{F}^0(\mathbf{y})\|^2] &= \|E_{xx}^{1/2}\|^2 + \|A_k - E_{xy}(E_{yy}^{1/2})^\dagger\|^2 \\ &\quad - \|E_{xy}(E_{yy}^{1/2})^\dagger\|^2 \\ &= \|E_{xx}^{1/2}\|^2 - \sum_{j=1}^{k} \sigma_j^2(A). \end{aligned} \tag{7.57}$$

Proof. We have

$$\begin{aligned} E[\|\mathbf{x}-\mathcal{F}_(\mathbf{y})\|^2] &= \mathrm{tr}\{E_{xx}-E_{xy}F^T-FE_{yx}+FE_{yy}F^T\} \\ &= \|E_{xx}^{1/2}\|^2-\|E_{xy}(E_{yy}^{1/2})^\dagger\|^2 \\ &\qquad +\|(F-E_{xy}E_{yy}^\dagger)E_{yy}^{1/2}\|^2 \\ &= \|E_{xx}^{1/2}\|^2-\|E_{xy}(E_{yy}^{1/2})^\dagger\|^2 \\ &\qquad +\|E_{xy}(E_{yy}^{1/2})^\dagger-FE_{yy}^{1/2}\|^2 \quad (7.58) \end{aligned}$$

because

$$E_{yy}^\dagger E_{yy}^{1/2}=(E_{yy}^{1/2})^\dagger$$

and

$$E_{xy}E_{yy}^\dagger E_{yy}=E_{xy} \quad (7.59)$$

by Lemma 24 of Section 5.4.2.

In (7.58), the only term that depends on F is $\|E_{xy}(E_{yy}^{1/2})^\dagger-FE_{yy}^{1/2}\|^2$. By Theorem 54 and Remark 30, its minimum is attained if $F=F^0$. Therefore, (7.56) follows.

The error representation (7.57) is true because of the following observation. On the basis of Lemma 42, we have

$$\begin{aligned} &\|E_{xy}(E_{yy}^{1/2})^\dagger-F^0E_{yy}^{1/2}\|^2 \\ &=\|E_{xy}(E_{yy}^{1/2})^\dagger-A_k(E_{yy}^{1/2})^\dagger E_{yy}^{1/2}\|^2 \\ &=\|E_{xy}(E_{yy}^{1/2})^\dagger-A_k\|^2 \\ &=\sum_{j=r+1}^{l}\sigma_j^2(A). \end{aligned}$$

Since

$$\|E_{xy}(E_{yy}^{1/2})^\dagger\|^2=\sum_{j=1}^{l}\sigma_j^2(A),$$

then (7.58) implies (7.57). □

We note that the crucial issues in proving Theorem 55 are Theorem 54 and the equation (7.59).

Definition 38. *The set $\{\mathcal{F}^0\}$ of operators $\mathcal{F}^0$ is called the family of generic Karhunen-Loève transforms. We also say that $\mathcal{F}^0$ provides the generic Principal Component Analysis.*

The efficiency of PCA–KLT and its generalization (7.56) is characterized by the compression ratio and the accuracy of the estimate of vector $\mathbf{x}$.

Compression of vector $\mathbf{x}$ (in fact, filtering and compression of data $\mathbf{y}$) is provided by the matrix $\Sigma_{Ak} V_{Ak}^T (E_{yy}^{1/2})^\dagger$ or by the matrix $V_{Ak}^T (E_{yy}^{1/2})^\dagger$. *Reconstruction* of the compressed vector is performed by the matrix U_{Ak} or by the matrix $U_{Ak}\Sigma_{Ak}$. Such a transform always exists since $\mathcal{F}^0$ is constructed from pseudo-inverse matrices.

We would like to point out that the operator $\mathcal{F}^0$ is not unique because of the arbitrary matrix Q in (7.56).

The differences between the provided solution in Theorem 55 and those in [63, 134, 183] are that the transform given by (7.56) is rigorously justified, including an analytical representation of non-uniqueness of such a transform.

Simulations which illustrate numerical properties of the transform $\mathcal{F}^0$ are given in Sections 7.5.9 and 7.6.9.

7.4.2 The minimum norm generic PCA–KLT

The generic Principal Component Analysis (or the generic Karhunen-Loève transform) presented by (7.56) depends on an arbitrary matrix Q and therefore, it is not unique. This implies a natural question: *What what kind of condition should be imposed on the statement of the problem and the solution* (7.56) *to make it unique?*

Below, we show that uniqueness is implied if we seek the solution F^0 with minimum norm.

Corollary 15. *The minimum Frobenius norm solution to the problem* (7.46)–(7.47) *is unique and it is given by the operator* $\widetilde{\mathcal{F}}^0$ *determined by the matrix* $\widetilde{F}^0$ *such that*

$$\widetilde{F}^0 = (E_{xy} E_{yy}^{1/2\dagger})_k E_{yy}^{1/2\dagger}. \tag{7.60}$$

The error associated with the operator $\widetilde{\mathcal{F}}^0$ *is given by* (7.57).

Proof. Let $R(m,n,k) \subseteq \mathbb{R}^{m\times n}$ be the variety of all $m \times n$ matrices of rank at most k. By Theorem 53, the minimum Frobenius norm solution to the problem

$$\min_{F\in R(m,n,k)} \|E_{xy}(E_{yy}^{1/2})^\dagger - F E_{yy}^{1/2}\|^2$$

is given by (7.60). This follows in a way which is similar to that used in Remark 30. In (7.58), the term $\|E_{xy}(E_{yy}^{1/2})^\dagger - F E_{yy}^{1/2}\|^2$ is the only one

which depends on F. A representation of the error associated with the operator $\widetilde{\mathcal{F}}^0$ follows directly from the proof of Theorem 55. Therefore, Corollary 15 is true. □

Remark 32. *It is interesting that* $E\|\mathcal{F}^0(\mathbf{y})\|^2$ *does not depend on the arbitrary matrix* Q *in* (7.56). *Indeed, let us denote* (*see* (7.56))

$$L = A_k E_{yy}^{1/2\dagger}, \quad M = (\widetilde{A}_1)_k C_1^{-1} Q U_2^T \ \text{ and } \ N = I - E_{yy}E_{yy}^{\dagger} = I - E_{yy}^{1/2}E_{yy}^{\dagger 1/2}.$$

Then we have

$$\begin{aligned} E\|\mathcal{F}^0(\mathbf{y})\|^2 &= \text{tr } E\{[\mathcal{L}(\mathbf{y}) + \mathcal{M}\mathcal{N}(\mathbf{y})][\mathcal{L}(\mathbf{y}) + \mathcal{M}\mathcal{L}(\mathbf{y})]^T\} \\ &= \text{tr } [LE_{yy}L^T + LE_{yy}N^TM^T + MNE_{yy}L^T \\ &\qquad + MNE_{yy}N^TM^T] \\ &= \text{tr } [LE_{yy}L^T] \\ &= \|LE_{yy}^{1/2}\|^2 \\ &= \|A_k(E_{yy}^{1/2})^{\dagger}E_{yy}^{1/2}\|^2 \\ &= \|A_k\|^2 \end{aligned}$$

because

$$\begin{aligned} E_{yy}N^T &= E_{yy}[I - E_{yy}^{\dagger}E_{yy}] \\ &= \mathbb{O} \end{aligned}$$

and

$$A_k(E_{yy}^{1/2})^{\dagger}E_{yy}^{1/2} = A_k$$

by Lemma 42.

7.5 Optimal Hybrid Transform Based on Hadamard-quadratic Approximation

7.5.1 Motivations

For a given compression ratio, the Karhunen-Loève transform (PCA–KLT) considered in Sections 7.2 and 7.4 minimizes the reconstruction error over the class of all *linear* reduced-rank transforms. Nevertheless, it may happen that the accuracy and compression ratio associated with PCA–KLT are still not satisfactory. In such a case, an improvement in the accuracy and compression ratio can be achieved by a transform with a more general structure than that of PCA–KLT. Special *non-linear* transforms have been

studied, in particular, in [23, 62, 159, 165, 168, 175, 182] using transform structures developed from the generalized Volterra polynomials. Nevertheless, the transforms [23, 62, 159, 165, 168, 175, 182] imply a substantial computational burden associated with the large number N of terms required by the underlying Volterra polynomial structure.

Here, we present the approach to constructing transforms of random vectors which is motivated by the hybrid method considered in Section 5.7 of Chapter 4. Our objective is to justify the transform which has advantages over methods presented in Sections 7.2 and 7.4 and in references [23, 62, 159, 165, 168, 175, 182].

A device for the approach is a s follows. The vector is first pre-estimated from the special iterative procedure such that each iterative loop is aimed at a solution of the *unconstrained* best approximation problem with the approximant given by the Hadamard-quadratic operator. The final estimate follows from a solution of the *constrained* best approximation problem with the Hadamard-quadratic approximant.

We show that the combination of these techniques allows us to build a more efficient and flexible method compared with PCA–KLT and its generalization given in Section 7.4. The estimation accuracy associated with the proposed method can be adjusted by a variation of the three degrees of freedom which are the transform degree, the number of iterations and the rank of the special covariance matrix. In connection with this, see Remark 36 in Section 7.5.8. In contrast, the techniques based on the development of PCA–KLT idea [63, 68, 134, 183] have the rank of covariance matrix as the only degree of freedom.

We establish a quite unrestrictive condition (see inequality (7.134) in Section 7.5.7 below), under which our transform provides a significantly smaller estimation error than the error associated with PCA–KLT's methods of Sections 7.2–7.4 and those presented in [63, 68, 134, 183].

7.5.2 Problem formulation and method description

Let $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$ be an unknown random vector and $\mathbf{y} \in L^2(\Omega, \mathbb{R}^n)$ observable random data such that $\mathbf{x} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(m)})^T$ and $\mathbf{y} = (\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(n)})^T$ where $\mathbf{x}_{(k)}, \mathbf{y}_{(i)} \in L^2(\Omega, \mathbb{R})$ for $k = 1, \ldots, m$ and $i = 1, \ldots, n$.

As before, for every $\omega \in \Omega$, we write

$$x = \mathbf{x}(\omega) \quad \text{and} \quad y = \mathbf{y}(\omega),$$

where $x = (x_{(1)}, \ldots, x_{(m)})^T$, $y = (y_{(1)}, \ldots, y_{(n)})^T$, $x_{(k)} = \mathbf{x}_{(k)}(\omega)$ and $y_{(i)} = \mathbf{y}_{(i)}(\omega)$ for $k = 1, \ldots, m$ and $i = 1, \ldots, n$.

The problem is to find a nonlinear transform $\mathcal{T} : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^m)$ of $\mathbf{x}$ from data $\mathbf{y}$ so that $\mathcal{T}$ provides both

(i) a better associated error of the estimate $\mathcal{T}(\mathbf{y})$, for a given compression ratio, and

(ii) a better compression ratio for the fixed accuracy of the estimate $\mathcal{T}(\mathbf{y})$, compared with the best known fixed rank linear estimator produced by the generic PCA–KLT of Section 7.4 and that in [63, 68, 134, 183].

Hereinafter we use the Hadamard product for vectors to define y^2 as $y^2=(y_{(1)}^2,\ldots,y_{(n)}^2)^T$.

The proposed method of solution consists of the following device.

Let us write $T_j=[U_j\ V_j\ W_j]\in\mathbb{R}^{m\times(2\nu+1)}$, $U_j\in\mathbb{R}^m$, $V_j,W_j\in\mathbb{R}^{m\times\nu}$, $\tilde{T}_j=[\tilde{U}_j\ \tilde{V}_j\ \tilde{W}_j]\in\mathbb{R}^{m\times(2\nu+1)}$, $\tilde{U}_j\in\mathbb{R}^m$, $\tilde{V}_j,\tilde{W}_j\in\mathbb{R}^{m\times\nu}$, $v_j=\begin{bmatrix}1\\ x_j\\ x_j^2\end{bmatrix}$ and $\nu=\begin{cases}n & \text{if}\quad j=0\\ m & \text{if}\quad j=1,2,\ldots p.\end{cases}$

For $j=0,1,\ldots,p$, define an operator $\tilde{\mathcal{T}}_j:L^2(\Omega,\mathbb{R}^{2\nu+1})\to L^2(\Omega,\mathbb{R}^m)$ by

$$[\tilde{\mathcal{T}}_j(\mathbf{v}_j)](\omega)=\tilde{T}_j[\mathbf{v}_j(\omega)]$$

and denote

$$J(T_j)=E[\|\mathbf{x}-\mathcal{T}_j(\mathbf{v}_j)\|^2].$$

Let $x_0=y$ and let $x_1,x_2,\ldots,x_p\in\mathbb{R}^m$ be determined as follows. For $j=0$, we write

$$x_1=\tilde{T}_0(v_0)=[\tilde{U}_0\ \tilde{V}_0\ \tilde{W}_0]\begin{bmatrix}1\\ y\\ y^2\end{bmatrix}$$

and find $\tilde{T}_0$ from a solution of the unconstrained problem

$$J(\tilde{T}_0)=\min_{T_0}J(T_0). \tag{7.61}$$

For $j=p-1$, we write

$$x_p=\tilde{T}_{p-1}(v_{p-1})=[\tilde{U}_{p-1}\ \tilde{V}_{p-1}\ \tilde{W}_{p-1}]\begin{bmatrix}1\\ x_{p-1}\\ x_{p-1}^2\end{bmatrix}$$

and find $\tilde{T}_{p-1}$ from a solution of the unconstrained problem

$$J(\tilde{T}_{p-1})=\min_{T_{p-1}}J(T_{p-1}). \tag{7.62}$$

For $j=p$, we write

$$x_{p+1}=T_{pr}^0(v_p)=[U_p^0\ V_p^0\ W_p^0]\begin{bmatrix}1\\ x_p\\ x_p^2\end{bmatrix}$$

and find T_{pr}^0 from a solution of the constrained problem

$$J(T_{pr}^0) = \min_{T_p} J(T_p) \tag{7.63}$$

subject to

$$\text{rank } T_p \leq r \tag{7.64}$$

with $r < \nu$.

The desired transform $\mathcal{T}$ is defined by

$$[\mathcal{T}(\mathbf{y})](\omega) = T_{pr}^0[\mathbf{v}_p(\omega)]. \tag{7.65}$$

In other words, the desired *nonlinear* transform $\mathcal{T}$ of $\mathbf{x}$, from data $\mathbf{y}$, is reduced to the *linear constrained* estimator $\mathcal{T}_{pr}^0$, with respect to $\mathbf{v}_p$, formed from the pre-estimate $\mathbf{x}_p$ of $\mathbf{x}$. The pre-estimate $\mathbf{x}_p$ follows from a solution of the p unconstrained best approximation problems (7.61)–(7.62). We call this procedure *the method of best hybrid approximations.*

The solutions to problems (7.61)–(7.62) aim to improve the known solution of the customary linear least square problem due to terms $\tilde{U}_j$, $\tilde{W}_j$ and $\mathbf{v}_j$. The iterative procedure (7.61)–(7.62) is to obtain the pre-estimate $\mathbf{x}_{j+1}$ with the accuracy better than the accuracy of pre-estimates from the preceding iterative loops. The terms U_p^0, W_p^0 and v_p in (7.63), (7.64) are used with the purpose of improvement of the linear constrained problem solution given in Sections 7.2 and 7.4.

Note that equation (7.65) can equivalently be rewritten as

$$\mathcal{T}(\mathbf{y}) = U_p^0 + V_p^0 x_p + W_p^0 x_p^2, \tag{7.66}$$

i.e. $\mathcal{T}(y)$ can be interpreted as the second degree estimate (with respect to x_p) of x.

In the next sections, we substantiate that the combination of these new techniques allows us to obtain the nonlinear transform with a considerably better performance in comparison with the generic PCA–KLT of Section 7.4. In particular, it will be shown that the error associated with the proposed transform can be achieved less than the error associated with the transforms in [63, 68, 134, 182] and Section 7.4 by exploiting the second degree terms in (7.61)–(7.62) and by increasing the number of iterations in (7.61)–(7.62).

It will also be shown that the proposed method does not require invertibility of any matrix used for the solution of problems (7.61)–(7.64).

7.5.3 Preliminary results

Let $\mathbf{g}$ and $\mathbf{h}$ be random vectors with realizations in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, and let $\mathbf{q} = \mathbf{h}^2$ and $\mathbf{s} = \begin{bmatrix} 1 \\ \mathbf{h} \\ \mathbf{q} \end{bmatrix}$. Similarly to the preceding Sections, we denote

$$E_{gh} = E[\mathbf{g}\mathbf{h}^T] \quad \text{and} \quad \mathbb{E}_{gh} = E_{gh} - E[\mathbf{g}]E[\mathbf{h}^T]$$

and we write $\mathcal{N}(\mathbb{E}_{gh})$ for the null space of matrix $\mathbb{E}_{gh}$.

For our purposes, we represent Lemmata 24 and 25 from Section 5.4.2 in the form of the following Lemmata 44 and 45, respectively.

Lemma 44. *The following equations hold:*

$$\mathbb{E}_{qh}\mathbb{E}_{hh}^{\dagger}\mathbb{E}_{hh} = \mathbb{E}_{qh} \quad \textit{and} \quad \mathbb{E}_{gq}\mathbb{E}_{qq}^{\dagger}\mathbb{E}_{qq} = \mathbb{E}_{gq}. \tag{7.67}$$

Lemma 45. *Let* $\mathcal{D}_{qh} = \mathbb{E}_{qq} - \mathbb{E}_{qh}\mathbb{E}_{hh}^{\dagger}\mathbb{E}_{hq}$. *Then*

$$\mathbb{E}_{qq}\mathcal{D}_{qh}^{\dagger}\mathcal{D}_{qh} = \mathbb{E}_{qq}, \quad \mathbb{E}_{gq}\mathcal{D}_{qh}^{\dagger}\mathcal{D}_{qh} = \mathbb{E}_{gq} \quad \textit{and} \quad \mathbb{E}_{hq}\mathcal{D}_{qh}^{\dagger}\mathcal{D}_{qh} = \mathbb{E}_{hq}. \tag{7.68}$$

The solution of the problems (7.61)–(7.64) will be given in terms of the $(2\nu+1)\times(2\nu+1)$ matrix $E_{ss}^{\dagger}$. In the next Lemma, we show that this matrix can be calculated via smaller $n \times n$ matrices. As a result, an associated computational load is facilitated.

Lemma 46. *Let*

$$P_{11} = 1 - P_{12}E[h] - P_{13}E[q],\ P_{12} = P_{21}^T,\ P_{13} = -E[h^T]P_{23} - E[q^T]P_{33},$$

$$P_{21} = -P_{22}E[h] - P_{23}E[q],\ P_{22} = \mathbb{E}_{hh}^{\dagger} - P_{23}\mathbb{E}_{qh}\mathbb{E}_{hh}^{\dagger},\ P_{23} = P_{32}^T,$$

$$P_{31} = -P_{33}E[q] - P_{32}E[h], \quad P_{32} = -P_{33}\mathbb{E}_{qh}\mathbb{E}_{hh}^{\dagger}, \quad P_{33} = \mathcal{D}.$$

Then

$$E_{ss}^{\dagger} = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix}. \tag{7.69}$$

Proof. Let

$$t = \begin{bmatrix} 1 \\ h \end{bmatrix}, \quad G_{11} = 1 - G_{12}E[h], \quad G_{12} = -E[h^T]G_{22}, \tag{7.70}$$

$$G_{21} = G_{12}^T, \quad G_{22} = \mathbb{E}_{hh}^{\dagger} \quad \text{and} \quad G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}. \tag{7.71}$$

Then

$$E_{tt}^{\dagger} = G. \tag{7.72}$$

The validity of equation (7.72) is shown by the following observations. We have

$$E_{tt} G E_{tt} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

where

$$\begin{aligned}
Q_{11} &= G_{11} + E[h^T]G_{21} + G_{12}E[h] + E[h^T]G_{22}E[h] \\
&= 1, \\
Q_{12} &= G_{11}E[h^T] + E[h^T]G_{21}E[h^T] + G_{12}E_{hh} + E[h^T]G_{22}E_{hh} \\
&= E[h^T], \\
Q_{21} &= E[h]G_{11} + E_{hh}G_{21} + E[h]G_{12}E[h] + E_{hh}G_{22}E[h] \\
&= E[h], \\
Q_{22} &= E[h]G_{11}E[h^T] + E_{hh}G_{21}E[h^T] + E[h]G_{12}E_{hh} + E_{hh}G_{22}E_{hh}, \\
&= E_{hh}.
\end{aligned}$$

Hence $E_{tt}GE_{tt} = E_{tt}$, i.e. the first Moore-Penrose condition is satisfied. The remaining Moore-Penrose conditions for $E_{tt}^{\dagger}$, defined by (7.72), are easily verified as well, and therefore (7.72) is valid.

Next, let

$$R_{11} = E_{tt}^{\dagger} - R_{12}E_{qt}E_{tt}^{\dagger}, \tag{7.73}$$

$$R_{12} = R_{21}^T, \tag{7.74}$$

$$R_{21} = -R_{22}E_{qt}E_{tt}^{\dagger} \tag{7.75}$$

$$R_{22} = D_{qt}^{\dagger}. \tag{7.76}$$

Similarly to the above and on the basis of Lemmata 44 and 45, it can be shown that

$$E_{ss}^{\dagger} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}, \tag{7.77}$$

where

$$\begin{aligned}
D_{qt} &= E_{qq} - E_{qt}E_{tt}^{\dagger}E_{tq} \\
&= \mathbb{E}_{qq} - \mathbb{E}_{qh}\mathbb{E}_{hh}^{\dagger}\mathbb{E}_{hq} \\
&= \mathcal{D}_{qh}.
\end{aligned}$$

Then (7.69) follows from (7.77) by virtue of (7.70) - (7.76). □

7.5.4 Solution of the problems (7.61)–(7.62)

In this section, we give solutions to the minimization problems posed above and provide the error analysis associated with the solutions.

To make an uniform notation for problems (7.61)–(7.62) we write

$$J(\tilde{T}_j) = \min_{T_j} J(T_j) \tag{7.78}$$

with $j = 0, 1, \ldots, p$.

For any matrix M we write $M(:, n_1 : n_2)$ to denote a matrix consisting of $n_2 - n_1 + 1$ successive columns of M beginning from the column numbered by n_1.

Let

$$\mathcal{K}_j = K_j[I - E_{v_j v_j} E^{\dagger}_{v_j v_j}],$$

where $K_j \in \mathbb{R}^{m \times (2\nu+1)}$ is an arbitrary matrix and I is the identity matrix, and let

$$\mathcal{K}_{Uj} = \mathcal{K}_j(:, 1:1), \quad \mathcal{K}_{Vj} = \mathcal{K}_j(:, 2:\nu{+}1) \quad \text{and} \quad \mathcal{K}_{Wj} = \mathcal{K}_j(:, \nu{+}2 : 2\nu{+}1).$$

We also denote $z_j = x_j^2$,

$$\bar{U}_j = E[x] - \bar{V}_j E[x_j] - \bar{W}_j E[z_j], \quad \bar{V}_j = (\mathbb{E}_{x x_j} - \bar{W}_j \mathbb{E}_{z_j x_j}) \mathbb{E}^{\dagger}_{x_j x_j} \tag{7.79}$$

and

$$\bar{W}_j = (\mathbb{E}_{x z_j} - \mathbb{E}_{x x_j} \mathbb{E}^{\dagger}_{x_j x_j} \mathbb{E}_{x_j z_j}) \mathcal{D}^{\dagger}_{z_j x_j}. \tag{7.80}$$

The following theorem provides the solution to problem (7.78) both in terms of pseudo-inverse matrix $E^{\dagger}_{v_j v_j} \in \mathbb{R}^{(2\nu+1)\times(2\nu+1)}$ and in terms of smaller pseudo-inverse matrices $\mathbb{E}^{\dagger}_{x_j x_j} \in \mathbb{R}^{\nu \times \nu}$, $\mathcal{D}^{\dagger}_{z_j x_j} \in \mathbb{R}^{\nu \times \nu}$. The latter is used for a computation of the alternative representation of the estimate $x_{j+1} = \tilde{T}_j v_j$ given by equation (7.88) below.

Theorem 56. *The unconstrained minimum* (7.78) *is achieved for*

$$\tilde{T}_j = [\tilde{U}_j \;\; \tilde{V}_j \;\; \tilde{W}_j] = E_{x v_j} E^{\dagger}_{v_j v_j} + \mathcal{K}_j \tag{7.81}$$

where

$$\tilde{U}_j = \bar{U}_j + \mathcal{K}_{Uj}, \quad \tilde{V}_j = \bar{V}_j + \mathcal{K}_{Vj} \quad \textit{and} \quad \tilde{W}_j = \bar{W}_j + \mathcal{K}_{Wj}. \tag{7.82}$$

Proof. It follows from Lemma 44 that

$$E_{x v_j} E^{\dagger}_{v_j v_j} E_{v_j v_j} = E_{x v_j} \tag{7.83}$$

and then

$$\begin{aligned} J(T_j) &= \Lambda_j + \mathrm{tr}\{(T_j - E_{xv_j}E^{\dagger}_{v_jv_j})E_{v_jv_j}(T_j - E_{xv_j}E^{\dagger}_{v_jv_j})^T\} \\ &= \Lambda_j + \|(T_j - E_{xv_j}E^{\dagger}_{v_jv_j})E^{1/2}_{v_jv_j}\|^2, \end{aligned}$$

where

$$\Lambda_j = \mathrm{tr}\{E_{xx} - E_{xv_j}E^{\dagger}_{v_jv_j}E_{v_jx}\}. \tag{7.84}$$

The minimum of this functional is achieved if

$$T_jE^{1/2}_{v_jv_j} - E_{xv_j}E^{\dagger}_{v_jv_j}E^{1/2}_{v_jv_j} = \mathbb{O}$$

which is equivalent to the equation (see Section 5.4.2)

$$T_jE_{v_jv_j} - E_{xv_j} = \mathbb{O}. \tag{7.85}$$

The necessary and sufficient condition [6] for the equation (7.85) to have a solution is given by (7.83) which is true by Lemma 44. Therefore, it follows from [6], pp 39-40 that the solution is given by $T_j = \tilde{T}_j$.

Next, on the strength of Lemma 46, it is easy to see that

$$E_{xv_j}E^{\dagger}_{v_jv_j} = [\bar{U}_j\ \bar{V}_j\ \bar{W}_j]. \tag{7.86}$$

Then (7.82) follows from (7.81) and (7.86).

The theorem is proven. □

Corollary 16. *The best estimate of* $\mathbf{x}$ *in the sense* (7.78) *is given by*

$$\mathbf{x}_{j+1} = \tilde{\mathcal{T}}_j(\mathbf{v}_j) \tag{7.87}$$

where $\tilde{\mathcal{T}}_j$ *is defined by* (7.81).

Corollary 17. *The equivalent representation of the estimate* (7.87) *is*

$$\mathbf{x}_{j+1} = \tilde{\mathcal{U}}_j + \tilde{\mathcal{V}}_j(\mathbf{x}_j) + \tilde{\mathcal{W}}_j(\mathbf{z}_j) \tag{7.88}$$

with $\tilde{U}_j$, $\tilde{V}_j$, $\tilde{W}_j$ *defined by* (7.82) *and* $\mathbf{z}_j = \mathbf{x}^2_j$. *The error associated with estimate* (7.87), (7.88) *is*

$$E[\|\mathbf{x} - \mathbf{x}_j\|^2] = \Lambda_{j-1} = \mathrm{tr}\{E_{xx} - E_{xv_{j-1}}E^{\dagger}_{v_{j-1}v_{j-1}}E_{v_{j-1}x}\} \tag{7.89}$$

Proof. Equation (7.89) follows directly from (7.84), (7.84) and (7.85). □

The representation of estimate x_{j+1} in form (7.88) can be computationally more effective compared with the form given in (7.87).

Note, that it is natural to choose $\mathcal{K}_{Uj} = \mathbb{O}$, $\mathcal{K}_{Vj} = \mathbb{O}$ and $\mathcal{K}_{Wj} = \mathbb{O}$ in equations (7.82) and (7.88), where $\mathbb{O}$ is the zero matrix/vector.

Definition 39. *The estimate given by equations* (7.81), (7.87), (7.88) *is called the* $(j+1)-th$ *unconstrained estimate of* $\mathbf{x}$.

Next, to find matrix T^0_{pr} giving the minimum (7.63) subject to constraint (7.64) we use the notation as follows. Let

$$E_{xv_p}(E^{1/2}_{v_pv_p})^\dagger = G\Sigma Q^T \tag{7.90}$$

be the singular value decomposition (SVD) of $E_{xv_p}(E^{1/2}_{v_pv_p})^\dagger$ where

$$G = [g_1, \ldots, g_{2\nu+1}] \in \mathbb{R}^{m\times(2\nu+1)} \text{ and } Q = [q_1, \ldots, q_{2\nu+1}] \in \mathbb{R}^{(2\nu+1)\times(2\nu+1)}$$

are orthogonal matrices and

$$\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{2\nu+1}) \in \mathbb{R}^{(2\nu+1)\times(2\nu+1)}$$

is a diagonal matrix with $\sigma_1 \geq \cdots \geq \sigma_l > 0$ and $\sigma_{l+1} = \cdots = \sigma_{2\nu+1} = 0$. Put $G_r = [g_1, \ldots, g_r]$, $Q_r = [q_1, \ldots, q_r]$ and $\Sigma_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ and define

$$P_r = P_{r(x,v_p)} = G_r\Sigma_r Q_r^T. \tag{7.91}$$

We also denote

$$E^{1/2}_{v_pv_p} = U_C\Sigma_C U_C^T$$

for the SVD of $E^{1/2}_{v_pv_p}$, and write

$$U_C = [U_1U_2], \quad C_1 = \operatorname{diag}(\sigma_1(C), \ldots, \sigma_t(C)) \quad \text{and} \quad E_{xv_p}E^{1/2}_{v_pv_p}U_C = [\widetilde{A}_1\widetilde{A}_2],$$

where

$$\Sigma_C = \operatorname{diag}(\sigma_1(C), \ldots, \sigma_{2\nu+1}(C))$$

with $\sigma_1(C) \geq \cdots \geq \sigma_t(C) > 0$ and $\sigma_{t+1}(C) = \cdots = \sigma_{2\nu+1}(C) = 0$, and

$$U_1 \in \mathbb{R}^{(2\nu+1)\times t} \quad \text{and} \quad \widetilde{A}_1 \in \mathbb{R}^{m\times t}.$$

The desired transform $\mathcal{T}$, given by equation (7.65), is defined by the following theorem.

Theorem 57. *The constrained minimum* (7.63)–(7.64) *is achieved for*

$$T^0_{pr} = P_r(E^{1/2}_{v_pv_p})^\dagger + (\widetilde{A}_1)_rC_1^{-1}Q_pU_2^T[I - E^{1/2}_{v_pv_p}(E^{1/2}_{v_pv_p})^\dagger] \tag{7.92}$$

where $Q_p \in \mathbb{R}^{t\times(2\nu+1-t)}$ *is an arbitrary matrix, and this minimum is*

$$J(T^0_{pr}) = \Lambda_p + \|P_r - E_{xv_p}(E^\dagger_{v_pv_p})^{1/2}\|^2. \tag{7.93}$$

Proof. Similarly to (7.84) we have

$$J(\tilde{T}_p) = \Lambda_p + \|\tilde{T}_p E_{v_p v_p}^{1/2} - E_{x v_p} E_{v_p v_p}^{\dagger} E_{v_p v_p}^{1/2}\|^2. \tag{7.94}$$

By Theorem 54, and Remarks 29 and 30, the functional (7.94) achieves the minimum subject to constraint (7.64) if $\tilde{T}_p = T_{pr}^0$.

Equation (7.93) follows directly from the above.

The theorem is proven. □

The methods of matrices $E_{x v_j}$, $E_{v_j v_j}$ estimation and associated error analysis can be found in Section 5.3.

Remark 33. *The $(j+1)$th unconstrained estimate $\mathbf{x}_{j+1}$ of $\mathbf{x}$ with $j = 0, 1, \ldots, p-1$, and the constrained estimate $\mathcal{T}_{pr}^0(\mathbf{v}_p)$ of $\mathbf{x}$, are not unique because K_j and M_p are arbitrary matrices.*

7.5.5 Error analysis associated with transform $\mathcal{T}$

The optimal transform $\mathcal{T}$ results in the estimate

$$\begin{aligned} \mathbf{x}_{p,r} &= \mathcal{T}(\mathbf{y}) \\ &= \mathcal{T}_{pr}^0(\mathbf{v}_p). \end{aligned} \tag{7.95}$$

Theorem 58. *Let*

$$\Delta(x_j) = \|(\mathbb{E}_{x z_j} - \mathbb{E}_{x x_j} \mathbb{E}_{x_j x_j}^{\dagger} \mathbb{E}_{x_j z_j})(\mathcal{D}_{z_j x_j}^{\dagger})^{\frac{1}{2}}\|^2. \tag{7.96}$$

The error associated with the optimal transform $\mathcal{T}$ is

$$\begin{aligned} E[\|\mathbf{x} - \mathcal{T}(\mathbf{y})\|^2] = \mathrm{tr}\{\mathbb{E}_{xx}\} + \sum_{i=r+1}^{l} \sigma_i^2 - \|\mathbb{E}_{xy}(\mathbb{E}_{yy}^{\dagger})^{1/2}\|^2 \\ - \sum_{j=0}^{p-1} \Delta(x_j). \end{aligned} \tag{7.97}$$

Proof. Let us first show that the error associated with the $p-$th unconstrained estimate $\mathbf{x}_p$ (7.87) is

$$E[\|\mathbf{x} - \mathbf{x}_p\|^2] = \mathrm{tr}\{\mathbb{E}_{xx}\} - \|\mathbb{E}_{xy}(\mathbb{E}_{yy}^{\dagger})^{1/2}\|^2 - \sum_{j=0}^{p-1} \Delta(x_j). \tag{7.98}$$

Indeed it follows from (7.84) and (7.84) that

$$E[\|\mathbf{x} - \mathbf{x}_p\|^2] = \mathrm{tr}\{E_{xx} - E_{x v_{p-1}} E_{v_{p-1} v_{p-1}}^{\dagger} E_{v_{p-1} x}\}, \tag{7.99}$$

where on the strength of Lemma 3,

$$\begin{aligned}\mathrm{tr}\{E_{xv_{p-1}}E^{\dagger}_{v_{p-1}v_{p-1}}E_{v_{p-1}x}\} &= \mathrm{tr}\{E[x]E[x^T]\} \\ &\quad +\|\mathbb{E}_{xx_{p-1}}(\mathbb{E}^{\dagger}_{x_{p-1}x_{p-1}})^{1/2}\|^2+\Delta(x_{p-1}). \quad (7.100)\end{aligned}$$

Therefore

$$E[\|\mathbf{x}-\mathbf{x}_p\|^2]=\mathrm{tr}\{\mathbb{E}_{xx}\}-\|\mathbb{E}_{xx_{p-1}}(\mathbb{E}^{\dagger}_{x_{p-1}x_{p-1}})^{1/2}\|^2-\Delta(x_{p-1}). \quad (7.101)$$

Hence, for $p=1$ equation (7.98) follows directly from (7.101).

Let us assume that (7.98) is true for $p=k$. To prove that (7.98) is now true for $p=k+1$, we need some preliminaries.

Let us denote

$$\tau=\mathbf{x}-E[\mathbf{x}],\quad \tau_i=\mathbf{x}_i-E[\mathbf{x}_i],\quad \vartheta_i=\mathbf{z}_i-E[\mathbf{z}_i],$$

and consider the functional

$$J_\tau(U_i,V_i,W_i)=E[\|\tau-(U_i+V_i\tau_i+W_i\vartheta_i)\|^2]. \quad (7.102)$$

It is easy to see that

$$\min_{U_i,V_i,W_i}J_\tau(U_i,V_i,W_i)=\min_{V_i,W_i}J_\tau(V_i,W_i) \quad (7.103)$$

where

$$J_\tau(V_i,W_i)=J_\tau(\mathbb{O}_m,V_i,W_i).$$

Note that the functional $J_\tau(V_i,W_i)$ can be written as

$$\begin{aligned}J_\tau(V_i,W_i) &= J_\tau(Z_i) \\ &= E[\|\tau-Z_i\theta_i\|^2] \quad (7.104)\end{aligned}$$

where $Z_i=[V_i\ W_i]$ and $\theta_i=\begin{bmatrix}\tau_i\\ \vartheta_i\end{bmatrix}$.

Next, let

$$D_{\vartheta_i\tau_i}=E_{\vartheta_i\vartheta_i}-E_{\vartheta_i\tau_i}E^{\dagger}_{\tau_i\tau_i}E_{\tau_i\vartheta_i}. \quad (7.105)$$

Then matrices

$$\check{V}_i=(E_{\tau\tau_i}-\check{W}_iE_{\vartheta_i\tau_i})E^{\dagger}_{\tau_i\tau_i}+\mathcal{K}_{Vi}, \quad (7.106)$$

$$\check{W}_i=(E_{\tau\vartheta_i}-E_{\tau\tau_i}E^{\dagger}_{\tau_i\tau_i}E_{\tau_i\vartheta_i})D^{\dagger}_{\vartheta_i\tau_i}+\mathcal{K}_{Wi} \quad (7.107)$$

and

$$\check{Z}_i=E_{\tau\theta_i}E^{\dagger}_{\theta_i\theta_i}+M_i[I-E^{1/2}_{\theta_i\theta_i}(E^{1/2}_{\theta_i\theta_i})^{\dagger}] \quad (7.108)$$

where $M_i \in \mathbb{R}^{m\times 2\nu}$ is arbitrary, are such that

$$J_\tau(\check{V}_i, \check{W}_i) = \min_{V_i, W_i} J_\tau(V_i, W_i)$$

and

$$J_\tau(\check{Z}_i) = \min_{Z_i} J_\tau(Z_i).$$

Then for τ_{i+1} defined by

$$\tau_{i+1} = \check{V}_i\tau_i + \check{W}_i\vartheta_i = \check{Z}_i\theta_i, \tag{7.109}$$

we have

$$\begin{aligned} E[\|\tau - \tau_{i+1}\|^2] &= \mathrm{tr}\{E_{\tau\tau} - E_{\tau\theta_i}E^\dagger_{\theta_i\theta_i}E_{\theta_i\tau}\} \\ &= \mathrm{tr}\{E_{\tau\tau} - E_{\tau\tau_i}E^\dagger_{\tau_i\tau_i}E_{\tau_i\tau}\} - \check{\Delta}(\tau_i), \end{aligned} \tag{7.110}$$

where

$$\check{\Delta}(\tau_i) = \|(E_{\tau\vartheta_i} - E_{\tau\tau_i}E^\dagger_{\tau_i\tau_i}E_{\tau_i\vartheta_i})(D^\dagger_{\vartheta_i\tau_i})^{1/2}\|^2.$$

Now, on the strength of (7.98) with $p = k$ and of (7.110) with $i = k-1$,

$$\begin{aligned} E[\|x - x_k\|^2] &= E[\|\tau - \tau_k\|^2] \\ &= \mathrm{tr}\{E_{\tau\tau} - E_{\tau\theta_{k-1}}E^\dagger_{\theta_{k-1}\theta_{k-1}}E_{\theta_{k-1}\tau} \\ &= \mathrm{tr}\{E_{\tau\tau}\} - \|E_{\tau\tau_0}(E^\dagger_{\tau_0\tau_0})^{1/2}\|^2 - \sum_{j=0}^{k-1}\check{\Delta}(\tau_j), \end{aligned} \tag{7.111}$$

since

$$\mathbb{E}_{xx} = E_{\tau\tau},\ \ \mathbb{E}_{z_ix_i} = E_{\vartheta_i\tau_i},\ \ \mathbb{E}_{xz_i} = E_{\tau\vartheta_i}\ \text{ and }\ \mathbb{E}_{x_ix_i} = E_{\tau_i\tau_i}. \tag{7.112}$$

Equations (7.108) and (7.109) imply that

$$\begin{aligned} E_{\tau\tau_k} &= E[\tau\theta^T_{k-1}(\check{Z}_{k-1})^T] \\ &= E_{\tau\theta_{k-1}}E^\dagger_{\theta_{k-1}\theta_{k-1}}E_{\theta_{k-1}\tau} \\ &\qquad + E_{\tau\theta_{k-1}}\{I - (E^\dagger_{\theta_{k-1}\theta_{k-1}})^{1/2}E^{1/2}_{\theta_{k-1}\theta_{k-1}}\}M_i^T \\ &= E_{\tau\theta_{k-1}}E^\dagger_{\theta_{k-1}\theta_{k-1}}E_{\theta_{k-1}\tau}. \end{aligned}$$

Analogously,

$$\begin{aligned} E_{\tau_k\tau_k} &= E[\check{Z}_{k-1}\theta_{k-1}\theta^T_{k-1}(\check{Z}_{k-1})^T] \\ &= E_{\tau\theta_{k-1}}E^\dagger_{\theta_{k-1}\theta_{k-1}}E_{\theta_{k-1}\tau}. \end{aligned}$$

As a result, we have

$$E_{\tau\tau_k}E^\dagger_{\tau_k\tau_k}E_{\tau_k\tau} = E_{\tau\theta_{k-1}}E^\dagger_{\theta_{k-1}\theta_{k-1}}E_{\theta_{k-1}\tau}. \tag{7.113}$$

Thus, on the basis of (7.110) and (7.113),

$$\begin{aligned} E[\|\tau-\tau_{k+1}\|^2] &= \mathrm{tr}\{E_{\tau\tau}-E_{\tau\tau_k}E^{\dagger}_{\tau_k\tau_k}E_{\tau_k\tau}\}-\check{\Delta}(\tau_k) \\ &= \mathrm{tr}\{E_{\tau\tau}-E_{\tau\theta_{k-1}}E^{\dagger}_{\theta_{k-1}\theta_{k-1}}E_{\theta_{k-1}\tau}\}-\check{\Delta}(\tau_k) \quad (7.114) \end{aligned}$$

and therefore (7.111) implies

$$\begin{aligned} E[\|\tau-\tau_{k+1}\|^2] &= E[\|x-x_{k+1}\|^2] \\ &= \mathrm{tr}\{E_{\tau\tau}\}-\|E_{\tau\tau_0}(E^{\dagger}_{\tau_0\tau_0})^{1/2}\|^2 \\ &\qquad -\sum_{j=0}^{k-1}\check{\Delta}(\tau_j)-\check{\Delta}(\tau_k). \quad (7.115) \end{aligned}$$

Then (7.98) with $p=k+1$ follows from (7.115) on the basis of (7.112). By virtue of that, the error estimate (7.98) is proven.

Next, it follows from (7.89), (7.93) that

$$\begin{aligned} E[\|x-x_{p,r}\|^2] &= J(T^0_{pr}) \\ &= E[\|x-x_p\|^2] \\ &\qquad +\|P_r-E_{xv_{p-1}}(E^{\dagger}_{v_{p-1}v_{p-1}})^{1/2}\|^2, \quad (7.116) \end{aligned}$$

where [50]

$$\|P_r-E_{xv_{p-1}}(E^{\dagger}_{v_{p-1}v_{p-1}})^{1/2}\|^2=\sum_{i=r+1}^{l}\sigma_i^2. \quad (7.117)$$

Hence, (7.98), (7.116) and (7.117) prove (7.97). □

Remark 34. *It follows from equation* (7.97) *that the error associated with the proposed transform* $\mathcal{T}$ *is decreasing with an increase in the number of iterations* p.

Remark 35. *The second degree term* Wx_j^2 *in* (7.87), (7.88) *is an important ingredient of the transform* $\mathcal{T}$. *Firstly, the term*

$$\Delta(x_0)=\|(\mathbb{E}_{xz_0}-\mathbb{E}_{xx_0}\mathbb{E}^{\dagger}_{x_0x_0}\mathbb{E}_{x_0z_0})(\mathcal{D}^{\dagger}_{z_0x_0})^{\frac{1}{2}}\|^2$$

which decreases the value $E[\|\mathbf{x}-\mathcal{T}(\mathbf{y})\|^2]$ *in* (7.97), *is a result of implementing the term* Wx_0^2 *in* (7.87), (7.88). *Secondly, if* $Wx_j^2=\mathbb{O}$ *then the procedure* (7.87), (7.95) *gives no decrease in the error* $E[\|\mathbf{x}-\mathcal{T}(\mathbf{y})\|^2]$ *for* $j=1,2,\ldots$ *since in this case,* $\Delta(x_j)=0$ *for* $j=1,2,\ldots$.

7.5.6 Particular cases

The proposed approach generalizes PCA–KLT and the known methods based on modifications of PCA–KLT idea as follows.

If in (7.92),

$$\mathbf{y} = \mathbf{x}, \quad p = 0, \quad U_p^0 = \mathbb{O}, \quad W_p^0 = \mathbb{O}, \quad \text{and} \quad M_p = \mathbb{O}$$

then T_{pr}^0 coincides with PCA–KLT.

The best fixed rank linear transform [134], which generalizes PCA–KLT, follows from (7.92) as a particular case if

$$p = 0, \quad U_p^0 = \mathbb{O}, \quad W_p^0 = \mathbb{O}$$

and if the matrix $E[\mathbf{y}\mathbf{y}^T]$ is invertible.

The transform considered in Section 7.4 where invertibility of $E[\mathbf{y}\mathbf{y}^T]$ is not assumed, also follows from (7.92) if $p = 0$, $U_p^0 = \mathbb{O}$ and $W_p^0 = \mathbb{O}$.

The best unconstrained transform of the second degree [158] produces the estimate which coincides with ((7.88) if $j = 0$ and $\tilde{U}_0 = \mathbb{O}$.

7.5.7 Comparative analysis of errors associated with hybrid Hadamard-quadratic transform and the generic PCA–KLT

Let $S_{\gamma(x,y)}$ be the truncated SVD of $E_{xy}(E_{yy}^{1/2})^\dagger$ defined similarly to equations (7.90), (7.91) but with the replacement of v_p by y and of r by γ such that $\gamma \le m$.

The generic PCA–KLT considered in Section 7.4 is given by

$$H_\gamma = S_{\gamma(x,y)}(E_{yy}^{1/2})^\dagger + K_\gamma[I - E_{yy}^{1/2}(E_{yy}^{1/2})^\dagger], \tag{7.118}$$

where $K_\gamma = (\widetilde{A}_1)_r C_1^{-1} Q U_2^T \in \mathbb{R}^{m\times n}$ is the matrix such that rank $H_\gamma \le \gamma < s$ with s the number of nonzero singular values $\beta_1, \ldots, \beta_{s_k}$ of the matrix $E_{xy}(E_{yy}^{1/2})^\dagger$.

As it has been mentioned before, the transform H_γ is optimal in the class of the linear transforms and it is a particular case of the proposed nonlinear transform $\mathcal{T}$ defined by the equations (7.65), (7.92) when $p = 0$, $U_p^0 = \mathbb{O}$ and $W_p^0 = \mathbb{O}$.

Let us compare the error $E[\|\mathbf{x} - \mathcal{H}_\gamma(\mathbf{y})\|]^2$ associated with the transform H_γ (7.118) and the error $E[\|\mathbf{x} - \check{\mathcal{T}}(\mathbf{y})\|]^2$ associated with a particular case $\check{\mathcal{T}}$ of the proposed transform when $U_j^0 = \mathbb{O}$ in (7.92), (7.95), as in Section 7.4, but for all $j = 1, 2, \ldots, p$ in (7.87), (7.88).

The equations representing transform $\check{\mathcal{T}}$ follow from (7.87), (7.79) - (7.82), (7.88), (7.90) - (7.92), (7.95) when $\tilde{U}_j = \bar{U}_j = U_p^0 = \mathbb{O}$, and they are as follows

$$\begin{aligned} \check{\mathbf{x}}_{p,r} &= \check{\mathcal{T}}(\mathbf{y}) \\ &= \check{\mathcal{T}}_{p,r}(\check{\mathbf{v}})_p, \end{aligned} \tag{7.119}$$

where

$$\check{\mathbf{v}}_j = \begin{bmatrix} \check{\mathbf{x}}_j \\ \check{\mathbf{z}}_j \end{bmatrix}, \quad \check{\mathbf{z}}_j = \check{\mathbf{x}}_j^2,$$

$$\check{T}_{p,r} = P_{r(x,\check{v}_p)}(E_{\check{v}_p\check{v}_p}^{1/2})^\dagger + M_p[I - E_{\check{v}_p\check{v}_p}(E_{\check{v}_p\check{v}_p})^\dagger], \tag{7.120}$$

$$\check{x}_{j+1} = \check{T}_j\check{v}_j \tag{7.121}$$

$$= \check{V}_j\check{x}_j + \check{W}_j\check{z}_j, \tag{7.122}$$

$$\begin{aligned} \check{T}_j &= E_{x\check{v}_j}E_{\check{v}_j\check{v}_j}^\dagger + \check{\mathcal{K}}_j \\ &= [\check{V}_j \ \check{W}_j], \end{aligned} \tag{7.123}$$

$$\check{\mathcal{K}}_j = K_j[I - E_{\check{v}_j\check{v}_j}(E_{\check{v}_j\check{v}_j})^\dagger], \tag{7.124}$$

$$\check{V}_j = (E_{x\check{x}_j} - \check{W}_jE_{\check{z}_j\check{x}_j})E_{\check{x}_j\check{x}_j}^\dagger + \check{\mathcal{K}}_j(:, 1:\nu), \tag{7.125}$$

$$\check{W}_j = (E_{x\check{z}_j} - E_{x\check{x}_j}E_{\check{x}_j\check{x}_j}^\dagger E_{\check{x}_j\check{z}_j})D_{\check{z}_j\check{x}_j}^\dagger + \check{\mathcal{K}}_j(:, \nu+1:2\nu) \tag{7.126}$$

and where $P_{r(x,\check{v}_p)}$ is defined similarly to (7.91),

$$P_{r(x,\check{v}_p)} = \check{G}_r\check{\Sigma}_r\check{Q}_r^T \tag{7.127}$$

with

$$\check{G}_r = [\check{g}_1, \ldots, \check{g}_r], \quad \check{Q}_r = [\check{q}_1, \ldots, \check{q}_r] \quad \text{and} \quad \check{\Sigma}_r = \operatorname{diag}(\check{\sigma}_1, \ldots, \check{\sigma}_r)$$

formed from orthogonal matrices

$$\check{G} = [\check{g}_1, \ldots, \check{g}_{2\nu}] \in \mathbb{R}^{m\times 2\nu}, \quad \check{Q} = [\check{g}_1, \ldots, \check{g}_{2\nu}] \in \mathbb{R}^{2\nu\times 2\nu}$$

and from diagonal matrix

$$\check{\Sigma} = \operatorname{diag}(\check{\sigma}_1, \ldots, \check{\sigma}_{2\nu}) \in \mathbb{R}^{(2\nu)\times(2\nu)}$$

with $\check{\sigma}_1 \geq \cdots \geq \check{\sigma}_l > 0$ and $\check{\sigma}_{l+1} = \cdots = \check{\sigma}_{2\nu} = 0$, respectively, such that

$$\check{G}\check{\Sigma}\check{Q}^T = E_{x\check{v}_p}(E_{\check{v}_p\check{v}_p}^{1/2})^\dagger. \tag{7.128}$$

Matrix $D_{\check{z}_j\check{x}_j}$ in (7.126) is defined in accordance with (7.105). We also denote

$$\check{\Delta}_j = \|(E_{x\check{z}_j} - E_{x\check{x}_j}E_{\check{x}_j\check{x}_j}^\dagger E_{\check{x}_j\check{z}_j})(D_{\check{z}_j\check{x}_j}^\dagger)^{1/2}\|^2$$

and

$$\Xi = \sum_{j=0}^{p-1} \check{\Delta}_j + \sum_{i=\gamma+1}^{s} \beta_i^2 - \sum_{k=r+1}^{l} \check{\sigma}_k^2.$$

Theorem 59. *The error* $E[\|\mathbf{x}-\check{\mathcal{T}}(\mathbf{y})\|]^2$ *associated with the proposed transform* (7.119) - (7.127) *is less than the error* $E[\|\mathbf{x}-\mathcal{H}_\gamma(\mathbf{y})\|]^2$ *associated with the transform* (7.118)*, for* Ξ*, i.e.*

$$E[\|\mathbf{x} - \mathcal{H}_\gamma(\mathbf{y})\|]^2 - E[\|\mathbf{x} - \check{\mathcal{T}}(\mathbf{y})\|]^2 \quad = \tag{7.129}$$

$$\sum_{j=0}^{p-1} \check{\Delta}_j + \sum_{i=\gamma+1}^{s} \beta_i^2 - \sum_{k=r+1}^{l} \check{\sigma}_k^2. \tag{7.130}$$

Proof. Similarly to (7.97) we have

$$E[\|\mathbf{x} - \check{\mathcal{T}}(\mathbf{y})\|^2] = \mathrm{tr}\{E_{xx}\} + \sum_{k=r+1}^{l} \check{\sigma}_k^2 - \|E_{xy}(E_{yy}^{\dagger})^{1/2}\|^2 \tag{7.131}$$

$$- \sum_{j=0}^{p-1} \check{\Delta}_j. \tag{7.132}$$

The equation for the error $E[\|\mathbf{x} - \mathcal{H}_\gamma(\mathbf{y})\|]^2$,

$$E[\|\mathbf{x} - \mathcal{H}_\gamma(\mathbf{y})\|]^2 = \mathrm{tr}\{E_{xx}\} + \sum_{i=\gamma+1}^{l} \beta_i^2 - \|E_{xy}(E_{yy}^{\dagger})^{1/2}\|^2 \tag{7.133}$$

is derived from (7.118) by virtue of Lemma 44, and then (7.129) is obvious. □

Corollary 18. *If*

$$\sum_{k=r+1}^{l} \check{\sigma}_k^2 < \sum_{j=0}^{p-1} \check{\Delta}_j + \sum_{i=\gamma+1}^{s} \beta_i^2 \tag{7.134}$$

then

$$E[\|\mathbf{x} - \check{\mathcal{T}}(\mathbf{y})\|]^2 < E[\|\mathbf{x} - \mathcal{H}_\gamma(\mathbf{y})\|]^2.$$

Proof. The proof follows directly from the above. □

Thus, the inequality (7.134) is the condition for a better performance of the proposed transform compared with transform (7.118), [63]. In particular, the inequality (7.134) takes place for the case when $r = \gamma$, where r and γ are numbers of principal components produced by the methods (7.119) - (7.127) and (7.118), [63], respectively. In other words, if (7.134) is true then for the same r and γ, the error associated with our method (7.119) - (7.127) is less than the error associated with the method (7.118).

Note that the condition (7.134) is not hardly restrictive and is normally satisfied, mainly due to the term $\sum_{j=0}^{p-1} \check{\Delta}_j$.

7.5.8 A special case: the errors $E[\|\mathbf{x} - \mathcal{H}_\gamma(\mathbf{y})\|]^2$ and $E[\|\mathbf{x} - \check{\mathcal{T}}(\mathbf{y})\|]^2$ are the same

Let us now consider the case when the errors (7.131), (7.133) associated with the methods (7.119) - (7.127) and (7.118) are the same, and consider the corresponding rank values r and γ (i.e. the numbers r and γ of the corresponding principal components of the methods (7.119) - (7.127) and (7.118) respectively).

First, note that the RHS's of the expressions (7.131), (7.133) contain the same constant term $\mathrm{tr}\{E_{xx}\} - \|E_{xy}(E_{yy}^{\dagger})^{1/2}\|^2$, and the terms $\sum_{k=r+1}^{l} \check{\sigma}_k^2 - \sum_{j=0}^{p-1} \check{\Delta}_j$ and $\sum_{i=\gamma+1}^{l} \beta_i^2$ which are variable with respect to r, p, and γ.

Let us suppose that

$$E[\|\mathbf{x} - \mathcal{H}_\gamma(\mathbf{y})\|]^2 = E[\|\mathbf{x} - \check{\mathcal{T}}(\mathbf{y})\|]^2 =: \varepsilon.$$

Then equation (7.129) implies that

$$\sum_{i=\gamma+1}^{s} \beta_i^2 = \sum_{k=r+1}^{l} \check{\sigma}_k^2 - \sum_{j=0}^{p-1} \check{\Delta}_j, \tag{7.135}$$

where the LHS and RHS are the variable terms of the errors (7.131), (7.133) associated with the methods (7.119) - (7.127) and (7.118).

We observe that the RHS in (7.135) can be reduced by increasing the number of iterations p in our method (7.87), (7.95). The corresponding reduction of the LHS in (7.135) can only be made by increasing the number γ in the transform (7.118) given in Section 7.4. Hence, to achieve the same accuracy ε, the method (7.119) - (7.127) uses, in general, a smaller number

r of the principal components than PCA–KLT and its modification (7.118), [63].

Moreover, for some $\varepsilon_{\mathcal{T}}$ where

$$\varepsilon_{\mathcal{T}} = E[\|\mathbf{x} - \check{\mathcal{T}}(\mathbf{y})\|]^2,$$

the accuracy $E[\|\mathbf{x} - \check{\mathcal{T}}(\mathbf{y})\|]^2$ can not be achieved by the method (7.118), [63] for any γ in (7.118), (7.133).

Remark 36. *The equations* (7.97), (7.129) - (7.133), (7.135) *substantiate the remark made in the Section* 7.5.1*: the proposed method possesses three degrees of freedom as follows: the number p of iteration loops in* (7.87), (7.88*), the degree of the approximants $T_j v_j$ and $\tilde{T}_p v_p$,*[5] *and the rank r of $T^0_{p,r}$ in* (7.63), (7.64). *In contrast, the performance of the generic PCA–KLT* (7.118) *can be regulated by a variation of the rank γ only.*

Remark 37. *PCA–KLT is a particular case of the generic PCA–KLT (Section* 7.4*) and therefore the results of the comparative analysis above are valid for PCA–KLT as well.*

7.5.9 Numerical example

In our example, we apply the proposed method to simultaneous filtering, compression and consequent reconstruction of a noisy digitized image. The aim of the example is to illustrate the impact on the final image estimate of both the method of hybrid best approximation (7.61)–(7.64), (7.87) and the representation of the estimate as the second degree polynomial (7.66).

The original digitized image 'Lenna' has been given by matrix $X \in \mathbb{R}^{256\times256}$ and observed noisy image has been modeled in the form

$$Y = 150N. * X, \tag{7.136}$$

where $N \in \mathbb{R}^{256\times256}$ is a matrix with normally distributed entries with mean 0 and variance 1. Symbol $.*$ means element-by-element matrix multiplication. The corresponding images are presented in Fig. 7.1.

Matrices X and Y have been partitioned into su-bmatrices

$$X^{(1)} = X(1:85,:), \quad X^{(2)} = X(86:170,:), \quad X^{(3)} = X(171:256,:)$$

and

$$Y^{(1)} = Y(1:85,:), \quad Y^{(2)} = Y(86:170,:), \quad Y^{(3)} = Y(171:256,:),$$

[5]For instance, $\tilde{W}_j = \mathbb{O}$ in (7.87) and $W^0_p \neq \mathbb{O}$ in (7.65) imply the estimates of the first degree and of the second degree correspondingly.

where $X(n_1 : n_2, :)$ means a sub-matrix formed by $n_2 - n_1 + 1$ consequent rows of X beginning with the n_1-th raw. We interpreted $X^{(1)}$, $X^{(2)}$, $X^{(3)}$ as realizations of different random vectors and $Y^{(1)}$, $Y^{(2)}$, $Y^{(3)}$ as corresponding observed data.

The proposed method (7.87), (7.92), (7.95) has been applied four times to the every pair $X^{(k)}$, $Y^{(k)}$, each time with the same $\mathcal{K}_j = \mathbb{O}$, $M_p = \mathbb{O}$, $p = 50$ but with a different value of r such that $r = r_s = 20 + 5s$ for $s = 0, 1, 2, 3$.

The generic PCA–KLT (7.118) (Section 7.4) has also been applied four times to the same pairs $X^{(k)}$, $Y^{(k)}$, with the same rank γ such that $\gamma = r_s$ where $s = 0, 1, 2, 3$. For comparison of the obtained results, the error ratios $\epsilon_H^{(k)} \,/\, \epsilon_T^{(k)}$ are presented in the Table 2 where

$$\epsilon_H^{(k)} = \|X^{(k)} - X_H^{(k)}\|^2 \quad \text{and} \quad \epsilon_T^{(k)} = \|X^{(k)} - X_T^{(k)}\|^2$$

are the errors associated with the estimate $X_H^{(k)}$ by (7.118) and with the estimate $X_T^{(k)}$ by (7.87), (7.81), (7.92), (7.95), respectively.

Matrices E_{xv_j}, $E_{v_jv_j}$, E_{xy}, E_{yy} in (7.87), (7.81), (7.92), (7.95) and (7.118) have been estimated with the known maximum likelihood estimates given in Section 5.3.1.

In this simulations, the ranks r and γ of the both methods are the same, therefore their compression ratios are equal,

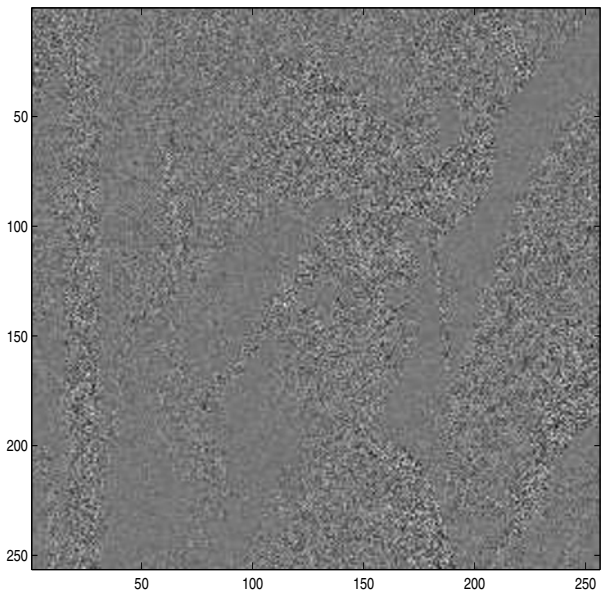
$$c_T = \frac{r}{256} \quad \text{and} \quad c_H = \frac{\gamma}{256}.$$

Hence, it follows from the Table 1 that, for the same compression ratio, the accuracy of the image reconstruction by the proposed method is from 153 to 311 times better that the reconstruction accuracy of the generic PCA–KLT (7.118), depending on the parts $X^{(k)}$, $Y^{(k)}$ of the images.

Table 1: Ratio of errors associated with estmators (7.92) and (7.118) for image portions $X^{(k)}$ with $k = 1, 2, 3$

Error ratios	*Rank r of $T^0_{p,r}$ and rank γ of H_γ*			
	$r = \gamma = 20$	$r = \gamma = 25$	$r = \gamma = 30$	$r = \gamma = 35$
$\epsilon_H^{(1)} \,/\, \epsilon_T^{(1)}$	236.0	268.5	293.0	308.5
$\epsilon_H^{(2)} \,/\, \epsilon_T^{(2)}$	153.3	199.5	245.0	286.6
$\epsilon_H^{(3)} \,/\, \epsilon_T^{(3)}$	157.5	214.9	264.3	311.0

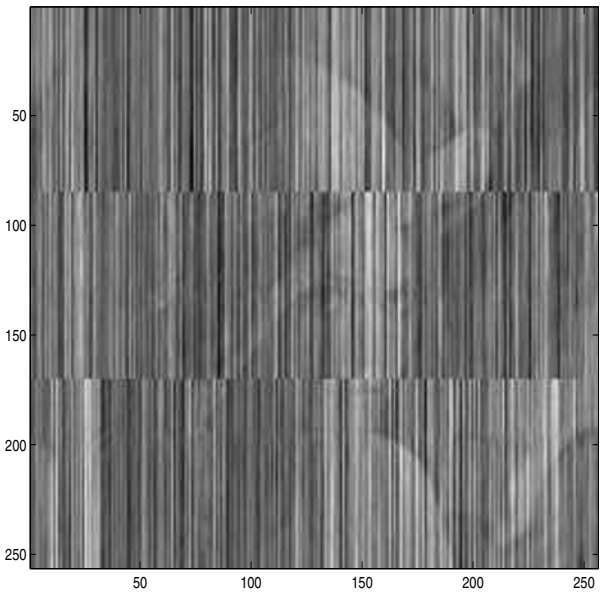
Fig. 7.2 represents the images reconstructed after simultaneous filtering and compression by these methods.



(a) Observed data Y.

(b) Image X to be filtered, compressed and then reconstructed from observed data. This digitized image has been taken from http://sipi.usc.edu/database/.

Figure 7.1: Data used in numerical example to the methods of Sections 7.4 and 7.5.



(a) Image reconstructed after filtering and compression by transform (7.118) with $\gamma = 20$ applied to each pair $X^{(k)}$, $Y^{(k)}$ with $k = 1, 2, 3$.

(b) Image reconstructed after filtering and compression by method (7.87), (7.81), (7.92), (7.95) with $j = 49$ and $r = 20$ applied to each pair $X^{(k)}$, $Y^{(k)}$ with $k = 1, 2, 3$.

Figure 7.2: The performance comparison of the methods (7.118) and (7.87), (7.81), (7.92), (7.95).

The Table 1 represents the values of the ratios ϵ_H / ϵ_T of errors

$$\epsilon_H = \|X - X_H\|^2 \quad \text{and} \quad \epsilon_T = \|X - X_T\|^2$$

associated with estimates

$$X_H = [{X_H^{(1)}}^T \ \ {X_H^{(2)}}^T \ \ {X_H^{(3)}}^T]^T$$

and

$$X_T = [{X_T^{(1)}}^T \ \ {X_T^{(2)}}^T \ \ {X_T^{(3)}}^T]^T$$

of the entire image X. In this case, the error ϵ_T associated with the proposed method is from 179 to 301 times smaller than the error ϵ_H associated with the method of Section 7.4.

Next, we wish to illustrate in a more conspicuous way the impact of the method of hybrid best approximations (7.61)–(7.64), (7.87) and of the second degree transform (7.66) on the superiority of the proposed estimation procedure over the method of Section 7.4.

To do this, we fix p and choose $r = 1$ in (7.92), which is the worst rank value for the quality of estimation by (7.92), (7.95) with fixed p (see the error equation (7.97)), but is the best rank value for the compression ratio range. In other words, in this case each sub-matrix $X^{(k)} \in \mathbb{R}^{m_k \times \nu_k}$ is compressed by the proposed method to a column in $\mathbb{R}^{m_k}$.

We also choose the full rank transform presented in Section 7.4, which gives its best quality of estimation (see the error equation (7.133)), but provides the worst compression ratio $c_H = 1$; this means that in this case the generic PCA–KLT provides no compression. The errors

$$\|X^{(k)} - X_{T,1}^{(k)}\|^2 \quad \text{and} \quad \|X^{(k)} - X_{H,full}^{(k)}\|^2$$

associated with the both transforms are given in Table 3, where $X_{T,1}^{(k)}$ is the estimate of $X^{(k)}$ by (7.95) with $p = 50$ and $r = 1$, and $X_{H,full}^{(k)}$ is the full rank estimate by the transform of Section 7.4.

It follows from Table 3 that the hybrid best approximations (7.87)–(7.78), (7.81) and the second degree transform (7.66) provide the final error $\|X^{(k)} - X_{T,1}^{(k)}\|^2$, which is from 2.6 to 3.5 times smaller than the error $\|X^{(k)} - X_{H,full}^{(k)}\|^2$, even for the extremely worst rank condition ($r = 1$) for our approach.

Summarizing the above, we would like to point out the following.

The error analysis given by the expressions (7.97), (7.129), (7.134), (7.135) demonstrates that the advantages of the proposed approach over PCA–KLT and the generic PCA–KLT of Section 7.4 are as follows:

Table 2: Ratio of errors associated with estimators (7.92) and (7.118) for entire image X

Error ratios	*Rank* r *of* $T^0_{p,r}$ *and rank* γ *of* H_γ			
	$r = \gamma = 20$	$r = \gamma = 25$	$r = \gamma = 30$	$r = \gamma = 35$
$\epsilon_H \ / \ \epsilon_T$	179.0	226.1	267.0	301.7

Table 3: Estimation errors corresponding the extremely distinct values of ranks of transforms (7.92) and (7.118)

Estimation errors	$k = 1$	$k = 2$	$k = 3$
$\|X^{(k)} - X^{(k)}_{T,1}\|^2$ for rank $r = 1$ in (7.92)	0.4667×10^4	0.5035×10^4	0.5500×10^4
$\|X^{(k)} - X^{(k)}_{H,full}\|^2$ for full rank transform (7.118)	1.6387×10^4	1.4757×10^4	1.4431×10^4

(i) under the condition (7.134), for the same rank r (i.e. for the same number of principal components r), the error associated with our method can be made less than the error associated with the generic PCA–KLT (Section 7.4) both by exploiting the second degree term in (7.87), (7.95), (7.119), (7.121) and by increasing the number of iterations p in (7.87), (7.121); and

(ii) for the same errors associated with the method (7.87), (7.95) and the method of Section 7.4, the method (7.87), (7.95) generates a smaller number of principal components.

These features imply that the above technique is preferable for many applied problems of the high dimensionality which have been considered in Section 7.4.

7.6 Optimal Transform Formed by a Combination of Nonlinear Operators

Our objective is to justify a new transform that may have both accuracy and compression ratio better than those of the transforms considered in the preceding sections.

We show that the proposed approach generalizes the Fourier series in Hilbert space, the Wiener filter, the Karhunen-Loève transform and the transforms given in [158, 167, 173].

7.6.1 Method description

Achievement of the above objective is based on the presentation of the proposed transform in the form of a sum with p terms (7.137) where each term is interpreted as a particular rank-reduced transform. Moreover, terms in (7.137) are represented as a combination of three operations $\mathcal{F}_k$, $\mathcal{Q}_k$ and $\boldsymbol{\varphi}_k$ for each $k = 1, \ldots, p$. The prime idea is to determine $\mathcal{F}_k$ separately, for each $k = 1, \ldots, p$, from an associated rank-constrained minimization problem similar to that in PCA–KLT. The operations $\mathcal{Q}_k$ and $\boldsymbol{\varphi}_k$ are auxiliary for finding $\mathcal{F}_k$. It is natural to expect that a contribution of each term in (7.137) will improve the entire transform performance.

To realize such a scheme, we choose the $\mathcal{Q}_k$ as orthogonal/orthonormal operators (see Section 7.6.2). Then each $\mathcal{F}_k$ can be determined independently for each individual problem (7.159) or (7.188) below. Next, operators $\boldsymbol{\varphi}_k$ are used to reduce the number of terms from N (as in [159, 167, 173, 182]) to p with $p << N$. For example, this can be done when we choose $\boldsymbol{\varphi}_k$ in the form presented in Section 7.6.6. Moreover, the composition of operators $\mathcal{Q}_k$ and $\boldsymbol{\varphi}_k$ allows us to reduce the related covari-

ance matrices to the identity matrix or to a block-diagonal form with small blocks. Remark 40 in Section 7.6.4 gives more details in this regard. The computational work associated with such blocks is much less than that for the large covariance matrices in [158, 159, 167, 173, 182].

To regulate accuracy associated with the proposed transform and its compression ratio, we formulate the problem in the form (7.140)–(7.141) where (7.141) consists of p constraints. It is shown in Sections 5.2.1, 5.2.2 and 5.2.4 that such a combination of constraints allows us to equip the proposed transforms with several degrees of freedom.

The structure of our transform is presented in Section 7.6.2 and the formal statement of the problem in Section 7.6.3. In Section 7.6.4, we determine operators $\mathcal{F}_1, \ldots, \mathcal{F}_p$ (Theorems 60 and 61).

As before, we denote

$$\mathbf{x} \in L^2(\Omega, \mathbb{R}^m), \quad \mathbf{y} \in L^2(\Omega, \mathbb{R}^n), \quad x = \mathbf{x}(\omega) \in \mathbb{R}^m, \quad y = \mathbf{y}(\omega) \in \mathbb{R}^n,$$
$$E_x = E[\mathbf{x}] \quad \text{and} \quad \mathbb{E}_{xy} = E[(\mathbf{x} - E_x)(\mathbf{y} - E_y)^T] = E_{xy} - E[\mathbf{x}]E[\mathbf{y}^T].$$

7.6.2 Structure of the proposed transform

Generic form

The proposed transform $\mathcal{T}_p$ is presented in the form

$$\begin{aligned} \mathcal{T}_p(\mathbf{y}) &= f + \sum_{k=1}^{p} \mathcal{F}_k \mathcal{Q}_k \boldsymbol{\varphi}_k(\mathbf{y}) \\ &= f + \mathcal{F}_1 \mathcal{Q}_1 \boldsymbol{\varphi}_1(\mathbf{y}) + \ldots + \mathcal{F}_p \mathcal{Q}_p \boldsymbol{\varphi}_p(\mathbf{y}), \end{aligned} \tag{7.137}$$

where

$$f \in \mathbb{R}^m, \quad \boldsymbol{\varphi}_k : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^n),$$

$$\mathcal{Q}_1, \ldots, \mathcal{Q}_p : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^n) \quad \text{and} \quad \mathcal{F}_k : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^m).$$

In general, one can put

$$\mathbf{x} \in L^2(\Omega, H_X), \quad \mathbf{y} \in L^2(\Omega, H_Y), \quad \boldsymbol{\varphi}_k : L^2(\Omega, H_Y) \to L^2(\Omega, H_k),$$

$$\mathcal{Q}_k : L^2(\Omega, H_k) \to L^2(\Omega, \tilde{H}_k) \quad \text{and} \quad \mathcal{F}_k : L^2(\Omega, \tilde{H}_k) \to L^2(\Omega, H_X)$$

with H_X, H_Y, H_k and $\tilde{H}_k$ separable Hilbert spaces, and $k = 1, \ldots, p$.

In (7.137), the vector f and operators $\mathcal{F}_1, \ldots, \mathcal{F}_p$ are determined from the minimization problem (7.140)-(7.141) given in the Section 7.6.3. Operators $\mathcal{Q}_1, \ldots, \mathcal{Q}_p$ in (7.137) are orthogonal (orthonormal) in the sense of the Definition 1 in Section 7.6.3 (in this regard, see also Remark 3 in Section 5.1).

To demonstrate and justify flexibility of the transform $\mathcal{T}_p$ with respect to the choice of $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_p$ in (7.137), we mainly study the case where $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_p$ are arbitrary. Specifications of $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_p$ are presented in Sections 3.2, 5.2.4 and 5.2.5 where we also discuss the benefits associated with some particular forms of $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_p$.

Some particular cases

Particular cases of the model $\mathcal{T}_p$ are associated with specific choices of $\boldsymbol{\varphi}_k$, $\mathcal{Q}_k$ and $\mathcal{F}_k$. Some examples are given below.

(i) If $H_X = H_Y = \mathbb{R}^n$ and $H_k = \tilde{H}_k = \mathbb{R}^{nk}$ where $\mathbb{R}^{nk}$ is the kth degree of $\mathbb{R}^n$, then (7.137) generalizes the known transform structures [158, 166, 167, 173]. The models [158, 166, 167, 173] follow from (7.137) if

$$\boldsymbol{\varphi}_k(\mathbf{y}) = \mathbf{y}^k \quad \text{where} \quad \mathbf{y}^k = (\mathbf{y}, \ldots, \mathbf{y}) \in L^2(\Omega, \mathbb{R}^{nk}),$$

$$\mathcal{Q}_k = \mathcal{I} \quad \text{where } \mathcal{I} \text{ is the identity operator,}$$

and

$$\text{if } \mathcal{F}_k \text{ is a } k\text{-linear operator.}$$

It has been shown in [158, 166, 167, 173] that such a form of $\boldsymbol{\varphi}_k$ leads to a significant improvement in the associated accuracy. See Section 7.6.6 for more details.

(ii) If $\boldsymbol{\varphi}_k : L^2(\Omega, H_Y) \to L^2(\Omega, H_X)$ and $\{\mathbf{u}_1, \mathbf{u}_2, \ldots\}$ is a basis in $L^2(\Omega, H_X)$ then $\boldsymbol{\varphi}_k$ and $\mathcal{Q}_k$ can be chosen so that

$$\boldsymbol{\varphi}_k(\mathbf{y}) = \mathbf{u}_k \quad \text{and} \quad \mathcal{Q}_k = \mathcal{I},$$

respectively. As a result, in this particular case,

$$\mathcal{T}_p(\mathbf{y}) = f + \sum_{k=1}^{p} \mathcal{F}_k(\mathbf{u}_k).$$

(iii) A similar case follows if $\boldsymbol{\varphi}_k : L^2(\Omega, H_Y) \to L^2(\Omega, H_k)$ is arbitrary but $\mathcal{Q}_k : L^2(\Omega, H_k) \to L^2(\Omega, \tilde{H}_k)$ is defined so that

$$\mathcal{Q}_k[\boldsymbol{\varphi}_k(\mathbf{y})] = \mathbf{v}_k \quad \text{with } k = 1, \ldots, p$$

where $\{\mathbf{v}_1, \mathbf{v}_2, \ldots\}$ is a basis in $L^2(\Omega, \tilde{H}_k)$. Then

$$\mathcal{T}_p(\mathbf{y}) = f + \sum_{k=1}^{p} \mathcal{F}_k(\mathbf{v}_k).$$

(iv) Let $\tilde{\mathbf{x}}^{(1)}, \ldots, \tilde{\mathbf{x}}^{(p)}$ be estimates of $\mathbf{x}$ by the known transforms. For instance, we could use the transforms considered in [63, 162, 169] and the transforms given in Chapter 5. Then we can put

$$\boldsymbol{\varphi}_1(\mathbf{y}) = \tilde{\mathbf{x}}^{(1)}, \quad \ldots, \quad \boldsymbol{\varphi}_p(\mathbf{y}) = \tilde{\mathbf{x}}^{(p)}.$$

In particular, one could choose $\tilde{\mathbf{x}}^{(1)} = \mathbf{y}$. In such a way, the vector $\mathbf{x}$ is pre-estimated from $\mathbf{y}$, and therefore, the overall $\mathbf{x}$ estimate by $\mathcal{T}_p$ will be improved. A new recursive method for finding $\tilde{\mathbf{x}}^{(1)}, \ldots, \tilde{\mathbf{x}}^{(p)}$ is given in Section 7.6.6 below.

Other particular cases of the proposed transform are considered in Sections 7.6.6 and 7.6.7.

Remark 38. *The particular case of $\mathcal{T}_p$ considered in the item (iii) above can be interpreted as an operator form of the Fourier polynomial in Hilbert space* [20]. *The benefits associated with the Fourier polynomials are well known. In item (ii) of Section* 7.6.7, *this case is considered in more detail.*

7.6.3 Statement of the problem

Hereinafter in this section, we suppose that $\mathcal{F}_k$ is linear for all $k = 1, \ldots, p$, the Hilbert spaces are the finite dimensional Eucledian spaces, $H_X = \mathbb{R}^m$ and $H_Y = H_k = \tilde{H}_k = \mathbb{R}^n$, and $\mathcal{Q}_1, \ldots, \mathcal{Q}_p$ are orthogonal operators by Definition 31 and Lemmata 33 and 34 of Section 5.7.4. The latter means that the vectors $\mathbf{v}_1$, ..., $\mathbf{v}_p$ defined by

$$\mathbf{v}_1 = \mathcal{Q}_1[\boldsymbol{\varphi}_1(\mathbf{y})], \quad \ldots, \quad \mathbf{v}_p = \mathcal{Q}_p[\boldsymbol{\varphi}_p(\mathbf{y})] \tag{7.138}$$

are orthogonal.

Let us denote

$$J(f, \mathcal{F}_1, \ldots \mathcal{F}_p) = E[\|\mathbf{x} - \mathcal{T}_p(\mathbf{y})\|^2]. \tag{7.139}$$

The problem is to determine the vector f^0 and operators $\mathcal{F}_1^0, \ldots, \mathcal{F}_p^0$ such that

$$J(f^0, \mathcal{F}_1^0, \ldots \mathcal{F}_p^0) = \min_{f, \mathcal{F}_1, \ldots, \mathcal{F}_p} J(f, \mathcal{F}_1, \ldots, \mathcal{F}_p) \tag{7.140}$$

subject to

$$\text{rank } \mathcal{F}_1 \le \eta_1, \quad \ldots, \quad \text{rank } \mathcal{F}_p \le \eta_p, \tag{7.141}$$

where

$$\eta_1 + \ldots + \eta_p \le \eta \le \min\{m, n\}.$$

We write

$$\mathcal{T}_p^0(\mathbf{y}) = f^0 + \sum_{k=1}^{p} \mathcal{F}_k^0(\mathbf{v}_k) \tag{7.142}$$

with

$$\mathbf{v}_k = \mathcal{Q}_k[\boldsymbol{\varphi}_k(\mathbf{y})].$$

It is supposed that covariance matrices formed from vectors $\mathcal{Q}_1\boldsymbol{\varphi}_1(\mathbf{y})$, ..., $\mathcal{Q}_p\boldsymbol{\varphi}_p(\mathbf{y})$ in (7.137) are known or can be estimated. Various estimation methods can be found in Section 4.3.

Remark 39. *Unlike known rank-constrained problems, we consider p constraints given by* (7.141)*. The number p of the constraints and the ranks* $\eta_1, \ldots, \eta_p$ *form the degrees of freedom for* $\mathcal{T}_p^0$*. Variation of p and* $\eta_1, \ldots, \eta_p$ *allows us to regulate accuracy associated with the transform* $\mathcal{T}_p^0$ *(see* (7.149) *and* (7.178) *in Section* 7.6.4*) and its compression ratio (see* (7.195*) in Section* 6.5.8*).*

7.6.4 Determination of f^0, $\mathcal{F}_1^0, \ldots, \mathcal{F}_p^0$ satisfying (7.140)–(7.141)

The case when matrix $\mathbb{E}_{\mathbf{v_i v_i}}$ is invertible for $i = 1, \ldots, p$

First, we consider the simpler case when $\mathbb{E}_{v_i v_i}$ is invertible for all $i = 1, \ldots, p$. Then the vector f^0 and operators $\mathcal{F}_1^0, \ldots, \mathcal{F}_p^0$ satisfying (7.140)–(7.141) are defined from the following Theorem 60. For each $i = 1, \ldots, p$, let $U_i \Sigma_i V_i^T$ be the SVD of $\mathbb{E}_{x v_i}$,

$$U_i \Sigma_i V_i^T = \mathbb{E}_{x v_i}, \tag{7.143}$$

where $U_i \in \mathbb{R}^{m \times n}$, $V_i \in \mathbb{R}^{n \times n}$ are orthogonal and $\Sigma_i \in \mathbb{R}^{n \times n}$ is diagonal,

$$U_i = [s_{i1}, \ldots, s_{in}], \quad V_i = [d_{i1}, \ldots, d_{in}], \tag{7.144}$$

$$\Sigma_i = \operatorname{diag}(\alpha_{i1}, \ldots, \alpha_{in}) \tag{7.145}$$

with $\alpha_{i1} \geq \cdots \geq \alpha_{ir} > 0$, $\alpha_{i,r+1} = \cdots = \alpha_{in} = 0$ and $r = 1, \ldots, n$ where $r = r(i)$. We set

$$U_{i\eta_i} = [s_{i1}, \ldots, s_{i\eta_i}], \quad V_{i\eta_i} = [d_{i1}, \ldots, d_{i\eta_i}],$$

$$\Sigma_{i\eta_i} = \operatorname{diag}(\alpha_{i1}, \ldots, \alpha_{i\eta_i}),$$

where $U_{i\eta_i} \in \mathbb{R}^{m \times \eta_i}$, $V_{i\eta_i} \in \mathbb{R}^{n \times \eta_i}$ and $\Sigma_{i\eta_i} \in \mathbb{R}^{\eta_i \times \eta_i}$. Now we define $K_{i\eta_i} \in \mathbb{R}^{m \times n}$ and $\mathcal{K}_{i\eta_i} : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^m)$ by

$$K_{i\eta_i} = U_{i\eta_i} \Sigma_{i\eta_i} V_{i\eta_i}^T \quad \text{and} \quad [\mathcal{K}_{i\eta_i}(\mathbf{w}_i)](\omega) = K_{i\eta_i}[\mathbf{w}_i(\omega)], \tag{7.146}$$

respectively, for any $\mathbf{w}_i \in L^2(\Omega, \mathbb{R}^n)$.

Theorem 60. *Let* $\mathbf{v}_1, \dots, \mathbf{v}_p$ *be orthogonal vectors determined by Lemma 33 of Section* 5.7.4. *Then the vector* f^0 *and operators* $\mathcal{F}_1^0, \dots, \mathcal{F}_p^0$, *satisfying* (7.140)–(7.141), *are determined by*

$$f^0 = E[\mathbf{x}] - \sum_{k=1}^{p} F_k^0 E[\mathbf{v}_k], \tag{7.147}$$

$$\mathcal{F}_1^0 = \mathcal{K}_{1\eta_1}, \quad \dots, \quad \mathcal{F}_p^0 = \mathcal{K}_{p\eta_p}. \tag{7.148}$$

The accuracy associated with transform $\mathcal{T}_p^0$, *determined by* (7.142) *and* (7.147)–(7.148), *is given by*

$$E[\|\mathbf{x} - \mathcal{T}_p^0(\mathbf{y})\|^2] = \|\mathbb{E}_{xx}^{1/2}\|^2 - \sum_{k=1}^{p}\sum_{j=1}^{\eta_k} \alpha_{kj}^2. \tag{7.149}$$

Proof. The functional $J(f, \mathcal{F}_1, \dots, \mathcal{F}_p)$ is written as

$$\begin{aligned} J(f, \mathcal{F}_1, \dots, \mathcal{F}_p) = \operatorname{tr}[E_{xx} - E[\mathbf{x}]f^T - \sum_{i=1}^{p} E_{xv_i} F_i^T \\ - fE[\mathbf{x}^T] + ff^T + f\sum_{i=1}^{p} E[\mathbf{v}_i^T]F_i^T - \sum_{i=1}^{p} F_i E_{v_i x} \\ + \sum_{i=1}^{p} F_i E[\mathbf{v}_i] f^T + E(\sum_{i=1}^{p} \mathcal{F}_i(\mathbf{v}_i)[\sum_{k=1}^{p} \mathcal{F}_i(\mathbf{v}_i)]^T)]. \end{aligned} \tag{7.150}$$

We remind that here and below, F_i is defined by $[\mathcal{F}_i(\mathbf{v}_i)](\omega) = F_i[\mathbf{v}_i(\omega)]$ so that, for example,

$$E[\mathcal{F}_k(\mathbf{v}_k)\mathbf{x}_k^T] = F_k E_{v_k x_k}.$$

In other words, the right hand side in (7.150) is a function of $f, \mathcal{F}_1, \dots, \mathcal{F}_p$ indeed.

Let us show that $J(f, \mathcal{F}_1, \dots, \mathcal{F}_p)$ can be represented as

$$J(f, \mathcal{F}_1, \dots, \mathcal{F}_p) = J_0 + J_1 + J_2, \tag{7.151}$$

where

$$J_0 = \|\mathbb{E}_{xx}^{1/2}\|^2 - \sum_{i=1}^{p} \|\mathbb{E}_{xv_i}\|^2, \tag{7.152}$$

$$J_1 = \|f - E[\mathbf{x}] + \sum_{i=1}^{p} F_i E[\mathbf{v}_i]\|^2 \quad \text{and} \quad J_2 = \sum_{i=1}^{p} \|F_i - \mathbb{E}_{xv_i}\|^2.$$

Indeed, J_1 and J_2 are rewritten as follows

$$\begin{aligned} J_1 &= \mathrm{tr}(ff^T - fE[\mathbf{x}^T] + \sum_{i=1}^{p} fE[\mathbf{v}_i^T]F_i + E[\mathbf{x}]E[\mathbf{x}^T] \\ &\quad -E[\mathbf{x}]f^T - \sum_{i=1}^{p} E[\mathbf{x}]E[\mathbf{v}_i^T]F_i^T + \sum_{i=1}^{p} F_iE[\mathbf{v}_i]f^T \\ &\quad - \sum_{i=1}^{p} F_iE[\mathbf{v}_i]E[\mathbf{x}^T] + \sum_{i=1}^{p} F_iE[\mathbf{v}_i]\sum_{k=1}^{p} E[\mathbf{v}_k^T]F_k^T) \end{aligned} \quad (7.153)$$

and

$$\begin{aligned} J_2 &= \sum_{i=1}^{p} \mathrm{tr}(F_i - \mathbb{E}_{xv_i})(F_i^T - \mathbb{E}_{v_ix}) \\ &= \sum_{i=1}^{p} \mathrm{tr}(F_iF_i^T - F_i\mathbb{E}_{v_ix} - \mathbb{E}_{xv_i}F_i^T + \mathbb{E}_{xv_i}\mathbb{E}_{v_ix}). \end{aligned} \quad (7.154)$$

In (7.154), $\sum_{i=1}^{p} \mathrm{tr}(F_iF_i^T)$ can be represented in the form

$$\begin{aligned} \sum_{i=1}^{p} \mathrm{tr}(F_iF_i^T) &= \mathrm{tr}[E(\sum_{i=1}^{p} F_i\mathbf{v}_i \sum_{k=1}^{p} \mathbf{v}_k^T F_k^T)] \\ &\quad -\mathrm{tr}(\sum_{i=1}^{p} F_iE[\mathbf{v}_i]\sum_{k=1}^{p} E[\mathbf{v}_k^T]F_k^T) \end{aligned} \quad (7.155)$$

because

$$E[\mathbf{v}_i\mathbf{v}_k^T] - E[\mathbf{v}_i]E[\mathbf{v}_k^T] = \begin{cases} \mathbb{O}, & i \neq k, \\ I, & i = k \end{cases} \quad (7.156)$$

due to the orthonormality of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$.

Then

$$\begin{aligned} J_0 + J_1 + J_2 &= \mathrm{tr}(E_{xx} - E[\mathbf{x}]E[\mathbf{x}^T]) \\ &\quad - \sum_{i=1}^{p} \mathrm{tr}[\mathbb{E}_{xv_i}\mathbb{E}_{v_ix}] + \mathrm{tr}(ff^T - fE[\mathbf{x}^T] + \sum_{i=1}^{p} fE[\mathbf{v}_i^T]F_i \\ &\quad +E[\mathbf{x}]E[\mathbf{x}^T] - E[\mathbf{x}]f^T - \sum_{i=1}^{p} E[\mathbf{x}]E[\mathbf{v}_i^T]F_i^T \\ &\quad + \sum_{i=1}^{p} F_iE[\mathbf{v}_i]f^T - \sum_{i=1}^{p} F_iE[\mathbf{v}_i]E[\mathbf{x}^T] \end{aligned}$$

$$+\sum_{i=1}^{p} F_i E[\mathbf{v}_i] \sum_{k=1}^{p} E[\mathbf{v}_k^T] F_k^T) + \mathrm{tr}[E(\sum_{i=1}^{p} F_i \mathbf{v}_i \sum_{k=1}^{p} \mathbf{v}_k^T F_k^T)]$$
$$-\mathrm{tr}(\sum_{i=1}^{p} F_i E[\mathbf{v}_i] \sum_{k=1}^{p} E[\mathbf{v}_k^T] F_k^T) - \sum_{i=1}^{p} \mathrm{tr}(F_i E_{v_i x}$$
$$-F_i E[\mathbf{v}_i] E[\mathbf{x}^T] + E_{xv_i} F_i^T - E[\mathbf{x}] E[\mathbf{v}_i^T] F_i^T - \mathbb{E}_{xv_i} \mathbb{E}_{v_i x})$$
$$= J(f, \mathcal{F}_1, \ldots, \mathcal{F}_p). \tag{7.157}$$

Hence, (7.151) is true. Therefore,

$$J(f, \mathcal{F}_1, \ldots, \mathcal{F}_p) = \|\mathbb{E}_{xx}^{1/2}\|^2 - \sum_{k=1}^{p} \|\mathbb{E}_{xv_k}\|^2 \tag{7.158}$$
$$+\|f - E[\mathbf{x}] + \sum_{k=1}^{p} F_k E[\mathbf{v}_k]\|^2 + \sum_{k=1}^{p} \|F_k - \mathbb{E}_{xv_k}\|^2.$$

It follows from (7.158) that the constrained minimum (7.140)–(7.141) is achieved if $f = f^0$ with f^0 given by (7.147), and if F_k^0 is such that

$$J_k(F_k^0) = \min_{F_k} J_k(F_k) \tag{7.159}$$

subject to

$$\mathrm{rank}(F_k) = \eta_k,$$

where

$$J_k(F_k) = \|F_k - \mathbb{E}_{xv_k}\|^2.$$

The solution to (7.159) is given [50] by

$$F_k^0 = K_{k\eta_k}. \tag{7.160}$$

Then

$$E[\|\mathbf{x} - \mathcal{T}_p^0(\mathbf{y})\|^2] = \|\mathbb{E}_{xx}^{1/2}\|^2 - \sum_{k=1}^{p} (\|\mathbb{E}_{xv_k}\|^2 - \|K_{k\eta_k} - \mathbb{E}_{xv_k}\|^2).$$

Here [50],

$$\|\mathbb{E}_{xv_k}\|^2 = \sum_{j=1}^{r} \alpha_{kj}^2 \quad \text{and} \quad \|K_{k\eta_k} - \mathbb{E}_{xv_k}\|^2 = \sum_{j=\eta_k+1}^{r} \alpha_{kj}^2 \tag{7.161}$$

with $r = r(k)$. Thus, (7.149) is true. The theorem is proved. □

Corollary 19. *Let* $\mathbf{v}_1, \ldots, \mathbf{v}_p$ *be determined by Lemma 33 of Section* 5.7.4. *Then the vector* $\hat{f}$ *and operators* $\hat{\mathcal{F}}_1, \ldots, \hat{\mathcal{F}}_p$ *satisfying the unconstrained problem* (7.140)*, are determined by*

$$\hat{f} = E[\mathbf{x}] - \sum_{k=1}^{p} \hat{F}_k E[\mathbf{v}_k] \text{ and } \hat{\mathcal{F}}_1 = \mathcal{E}_{xv_1}, \quad \ldots, \quad \hat{\mathcal{F}}_p = \mathcal{E}_{xv_p} \tag{7.162}$$

with $\hat{\mathcal{F}}_k$ *such that* $[\hat{\mathcal{F}}_k(\mathbf{v}_k)](\omega) = \hat{F}_k \mathbf{v}_k(\omega)$ *where* $\hat{F}_k \in \mathbb{R}^{n\times m}$ *and* $k = 1, \ldots, p$.

The accuracy associated with transform $\hat{\mathcal{T}}_p$ *given by*

$$\hat{\mathcal{T}}_p(\mathbf{y}) = \hat{f} + \sum_{k=1}^{p} \hat{\mathcal{F}}_k(\mathbf{v}_k) \tag{7.163}$$

is such that

$$E[\|\mathbf{x} - \hat{\mathcal{T}}_p(\mathbf{y})\|^2] = \|\mathbb{E}_{xx}^{1/2}\|^2 - \sum_{k=1}^{p} \|\mathbb{E}_{xv_k}\|^2. \tag{7.164}$$

Proof. The proof follows directly from (7.159). □

The case when matrix $\mathbb{E}_{\mathbf{v_i v_i}}$ is not invertible for $i = 1, \ldots, p$

For $\mathbf{u}_k, \mathbf{v}_j, \mathbf{w}_j \in L^2(\Omega, \mathbb{R}^n)$, we define operators $\mathcal{E}_{u_k v_j}, \mathcal{E}^{\dagger}_{v_j v_j}, (\mathcal{E}^{1/2}_{v_k v_k})^{\dagger} : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^n)$ by the expressions

$$[\mathcal{E}_{u_k v_j}(\mathbf{w}_j)](\omega) = \mathbb{E}_{u_k v_j} \mathbf{w}_j(\omega), \quad [\mathcal{E}^{\dagger}_{v_j v_j}(\mathbf{w}_j)](\omega) = \mathbb{E}^{\dagger}_{v_j v_j} \mathbf{w}_j(\omega) \tag{7.165}$$

and

$$[(\mathcal{E}^{1/2}_{v_k v_k})^{\dagger}(\mathbf{w}_j)](\omega) = (\mathbb{E}^{1/2})^{\dagger}_{v_j v_j} \mathbf{w}_j(\omega), \tag{7.166}$$

respectively.

We write $M_k \in \mathbb{R}^{m\times n}$ for an arbitrary matrix, and define operator $\mathcal{M}_k : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^m)$ by $[\mathcal{M}_k(\mathbf{w})](\omega) = M_k \mathbf{w}(\omega)$ for any $\mathbf{w} \in L^2(\Omega, \mathbb{R}^n)$.

For the case under consideration (matrix $\mathbb{E}_{v_k v_k}$ is not invertible), we introduce the SVD of $\mathbb{E}_{xv_k}(\mathbb{E}^{1/2}_{v_k v_k})^{\dagger}$,

$$U_k \Sigma_k V_k^T = \mathbb{E}_{xv_k}(\mathbb{E}^{1/2}_{v_k v_k})^{\dagger}, \tag{7.167}$$

where, as above, $U_k \in \mathbb{R}^{m\times n}$, $V_k \in \mathbb{R}^{n\times n}$ are orthogonal and $\Sigma_k \in \mathbb{R}^{n\times n}$ is diagonal,

$$U_k = [s_{k1}, \ldots, s_{kn}], \quad V_k = [d_{k1}, \ldots, d_{kn}], \tag{7.168}$$

$$\Sigma_k = \mathrm{diag}(\beta_{k1}, \ldots, \beta_{kn}) \tag{7.169}$$

with $\beta_{k1} \geq \cdots \geq \beta_{kr} > 0$, $\beta_{k,r+1} = \beta_{kn} = 0$, $r = 1, \ldots, n$ and $r = r(k)$.

Let us set

$$U_{k\eta_k} = [s_{k1}, \ldots, s_{k\eta_k}], \quad V_{k\eta_k} = [d_{k1}, \ldots, d_{k\eta_k}] \tag{7.170}$$

$$\Sigma_{k\eta_k} = \mathrm{diag}(\beta_{k1}, \ldots, \beta_{k\eta_k}), \tag{7.171}$$

where $U_{k\eta_k} \in \mathbb{R}^{m\times\eta_k}$, $V_{k\eta_k} \in \mathbb{R}^{n\times\eta_k}$ and $\Sigma_{k\eta_k} \in \mathbb{R}^{\eta_k\times\eta_k}$. Now we define $G_{k\eta_k} \in \mathbb{R}^{m\times n}$ and $\mathcal{G}_{k\eta_k} : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^m)$ by

$$G_{k\eta_k} = U_{k\eta_k}\Sigma_{k\eta_k}V_{k\eta_k}^T \quad \text{and} \quad [\mathcal{G}_{k\eta_k}(\mathbf{w}_k)](\omega) = G_{k\eta_k}[\mathbf{w}_k(\omega)], \tag{7.172}$$

respectively, for any $\mathbf{w}_k \in L^2(\Omega, \mathbb{R}^n)$.

Lemma 47. *Let $\mathbf{v}_k \in L^2(\Omega, \mathbb{R}^n)$. Then*

$$G_{\eta_k}(\mathbb{E}_{v_kv_k}^{1/2})^\dagger\mathbb{E}_{v_kv_k}^{1/2} = G_{\eta_k}. \tag{7.173}$$

Proof. This lemma is a different form of Lemma 42 given in Section 7.4.1. □

We also write

$$\mathbb{E}_{v_kv_k}^{1/2} = U_{(k)}\Sigma_{(k)}U_{(k)}^T$$

for the SVD of $\mathbb{E}_{v_kv_k}^{1/2}$, and denote

$$U_{(k)} = [U_{(k)1}\ U_{(k)2}], \quad C_{(k)} = \mathrm{diag}(\sigma_{1(k)}, \ldots, \sigma_{t(k)})$$

and

$$\mathbb{E}_{xv_k}\mathbb{E}_{v_kv_k}^{1/2}U_{(k)} = [\widetilde{A}_{(k)1}\ \widetilde{A}_{(k)2}],$$

where

$$\Sigma_{(k)} = \mathrm{diag}(\sigma_{1(k)}, \ldots, \sigma_{n(k)})$$

with $\sigma_{1(k)} \geq \cdots \geq \sigma_{t(k)} > 0$ and $\sigma_{t+1(k)} = \cdots = \sigma_{n(k)} = 0$, $U_{(k)1} \in \mathbb{R}^{n\times t}$ and $\widetilde{A}_{(k)1} \in \mathbb{R}^{m\times t}$.

If A is any matrix then we write $(A)_{\eta_k}$ for a matrix defined similarly to (7.12).

Theorem 61. *Let* $\mathbf{v}_1, \ldots, \mathbf{v}_p$ *be orthogonal vectors determined by Lemma* 34 *of Section* 5.7.4. *Then* f^0 *and* $\mathcal{F}_1^0, \ldots, \mathcal{F}_p^0$, *satisfying* (7.140)–(7.141), *are determined by*

$$f^0 = E[\mathbf{x}] - \sum_{k=1}^{p} F_k^0 E[\mathbf{v}_k] \tag{7.174}$$

and

$$F_1^0 = G_{\eta_1}(\mathbb{E}_{v_1 v_1}^{1/2})^\dagger + M_1[I - \mathbb{E}_{v_1 v_1}^{1/2}(\mathbb{E}_{v_1 v_1}^{1/2})^\dagger], \tag{7.175}$$

$$\vdots$$

$$F_p^0 = G_{\eta_p}(\mathbb{E}_{v_p v_p}^{1/2})^\dagger + M_p[I - \mathbb{E}_{v_p v_p}^{1/2}(\mathbb{E}_{v_p v_p}^{1/2})^\dagger] \tag{7.176}$$

where for $k = 1, \ldots, p$,

$$M_k = (\widetilde{A}_{(k)1})_{\eta_k} C_{(k)}^{-1} Q_k U_{(k)2}^T \tag{7.177}$$

with an arbitrary matrix $Q_k \in \mathbb{R}^{t \times (n-t)}$.

The accuracy associated with transform $\mathcal{T}_p^0$ *given by* (7.142) *and* (7.174)–(7.176) *is such that*

$$E[\|\mathbf{x} - \mathcal{T}_p^0(\mathbf{y})\|^2] = \|\mathbb{E}_{xx}^{1/2}\|^2 - \sum_{k=1}^{p}\sum_{j=1}^{\eta_k} \beta_{kj}^2. \tag{7.178}$$

Proof. If $\mathbf{v}_1, \ldots, \mathbf{v}_p$ are determined by Lemma 34, then $J(f, \mathcal{F}_1, \ldots, \mathcal{F}_p)$ is still represented by (7.150). Let us consider J_0, J_1 and J_2 given by

$$J_0 = \|\mathbb{E}_{xx}^{1/2}\|^2 - \sum_{k=1}^{p} \|\mathbb{E}_{x v_k}(\mathbb{E}_{v_k v_k}^{1/2})^\dagger\|^2, \tag{7.179}$$

$$J_1 = \|f - E[\mathbf{x}] + \sum_{k=1}^{p} F_k E[\mathbf{v}_k]\|^2 \tag{7.180}$$

and

$$J_2 = \sum_{k=1}^{p} \|F_k \mathbb{E}_{v_k v_k}^{1/2} - \mathbb{E}_{x v_k}(\mathbb{E}_{v_k v_k}^{1/2})^\dagger\|^2. \tag{7.181}$$

To show that

$$J(f, \mathcal{F}_1, \ldots, \mathcal{F}_p) = J_0 + J_1 + J_2 \tag{7.182}$$

with $J(f, \mathcal{F}_1, \ldots, \mathcal{F}_p)$ defined by (7.150), we use the relationships (see Section 4.4.1)

$$\mathbb{E}_{xv_k}\mathbb{E}^{\dagger}_{v_kv_k}\mathbb{E}_{v_kv_k} = \mathbb{E}_{xv_k} \quad \text{and} \quad \mathbb{E}^{\dagger}_{v_kv_k}\mathbb{E}^{1/2}_{v_kv_k} = (\mathbb{E}^{1/2}_{v_kv_k})^{\dagger} \tag{7.183}$$

Then

$$\begin{aligned} J_1 &= \operatorname{tr}(ff^T - fE[\mathbf{x}^T] + \sum_{k=1}^{p} fE[\mathbf{v}_k^T]F_k + E[\mathbf{x}]E[\mathbf{x}^T] \\ &\quad -E[\mathbf{x}]f^T - \sum_{k=1}^{p} E[\mathbf{x}]E[\mathbf{v}_k^T]F_k^T + \sum_{k=1}^{p} F_kE[\mathbf{v}_k]f^T \\ &\quad - \sum_{k=1}^{p} F_kE[\mathbf{v}_k]E[\mathbf{x}^T] + \sum_{k=1}^{p} F_kE[\mathbf{v}_k]\sum_{i=1}^{p} E[\mathbf{v}_i^T]F_i^T) \end{aligned} \tag{7.184}$$

and

$$\begin{aligned} J_2 &= \sum_{k=1}^{p} \operatorname{tr}(F_k - \mathbb{E}_{xv_k}\mathbb{E}^{\dagger}_{v_kv_k})\mathbb{E}_{v_kv_k}(F_k^T - \mathbb{E}^{\dagger}_{v_kv_k}\mathbb{E}_{v_kx}) \\ &= \sum_{k=1}^{p} \operatorname{tr}(F_k\mathbb{E}_{v_kv_k}F_k^T - F_k\mathbb{E}_{v_kx} - \mathbb{E}_{xv_k}F_k^T + \mathbb{E}_{xv_k}\mathbb{E}^{\dagger}_{v_kv_k}\mathbb{E}_{v_kx}), \end{aligned}$$

where

$$\begin{aligned} \sum_{k=1}^{p} \operatorname{tr}(F_k\mathbb{E}_{v_kv_k}F_1^T) &= \operatorname{tr}[E(\sum_{k=1}^{p} F_k\mathbf{v}_k \sum_{i=1}^{p} \mathbf{v}_i^T F_i^T)] \\ &\quad -\operatorname{tr}(\sum_{k=1}^{p} F_kE[\mathbf{v}_k]\sum_{i=1}^{p} E[\mathbf{v}_i^T]F_i^T) \end{aligned} \tag{7.185}$$

because

$$E[\mathbf{v}_i\mathbf{v}_k^T] - E[\mathbf{v}_i]E[\mathbf{v}_k^T] = \mathbb{O} \quad \text{for} \quad i \neq k \tag{7.186}$$

due to orthogonality of the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{s_k}$.

On the basis of (7.183)–(7.185) and similarly to (7.157)–(7.157), we establishe that (7.182) is true. Hence,

$$\begin{aligned} J(f, \mathcal{F}_1, \ldots, \mathcal{F}_p) &= \|\mathbb{E}^{1/2}_{xx}\|^2 - \sum_{k=1}^{p} \|\mathbb{E}_{xv_k}(\mathbb{E}^{1/2}_{v_kv_k})^{\dagger}\|^2 \\ &\quad +\|f - E[\mathbf{x}] + \sum_{k=1}^{p} F_kE[\mathbf{v}_k]\|^2 \end{aligned} \tag{7.187}$$

$$+\sum_{k=1}^{p}\|F_k\mathbb{E}_{v_kv_k}^{1/2}-\mathbb{E}_{xv_k}(\mathbb{E}_{v_kv_k}^{1/2})^{\dagger}\|^2.$$

It follows from the last two terms in (7.187) that the constrained minimum (7.140)–(7.141) is achieved if $f=f^0$ with f^0 given by (7.174), and F_k^0 is such that

$$J_k(F_k^0)=\min_{F_k}J_k(F_k) \tag{7.188}$$

subject to

$$\operatorname{rank}(F_k)=\eta_k,$$

where

$$J_k(F_k)=\|F_k\mathbb{E}_{v_kv_k}^{1/2}-\mathbb{E}_{xv_k}(\mathbb{E}_{v_kv_k}^{1/2})^{\dagger}\|^2.$$

Therefore, the constrained minimum (7.140)–(7.141) is achieved if $f=f^0$ where f^0 is defined by (7.174), and if

$$F_k=F_k^0=G_{\eta_k}(\mathbb{E}_{v_kv_k}^{1/2})^{\dagger}+M_k[I-\mathbb{E}_{v_kv_k}^{1/2}(\mathbb{E}_{v_kv_k}^{1/2})^{\dagger}]. \tag{7.189}$$

The latter follows from Theorem 54 and Remarks 29 and 30. Thus, (7.175)–(7.176) are true.

Next, similar to (7.161),

$$\|\mathbb{E}_{xv_k}(\mathbb{E}_{v_kv_k}^{1/2})^{\dagger}\|^2-\|G_{\eta_k}-\mathbb{E}_{xv_k}(\mathbb{E}_{v_kv_k}^{1/2})^{\dagger}\|^2=\sum_{j=1}^{\eta_k}\beta_{kj}^2. \tag{7.190}$$

Then (7.178) follows from (7.187), (7.189), (7.174) and (7.190). □

Remark 40. *The known reduced-rank transforms based on the Volterra polynomial structure* [159, 167, 173, 182] *require the computation of a covariance matrix similar to* $\mathbb{E}_{vv}$*, where* $\mathbf{v}=[\mathbf{v}_1,\dots,\mathbf{v}_p]^T$*, but for* $p=N$ *where* N *is large (see Section* 4.4.2 *). The relationships* (7.157)–(7.159) *and* (7.184)–(7.188) *illustrate the nature of the proposed method and its difference from the techniques in* [159, 167, 173, 182]*: due to the structure* (7.137) *of the transform* $\mathcal{T}_p$*, the procedure for finding* f^0*,* $\mathcal{F}_1^0,\dots,\mathcal{F}_p^0$ *avoids direct computation of* $\mathbb{E}_{vv}$ *which could be troublesome due to large* N*. If operators* $\mathcal{Q}_1,\dots,\mathcal{Q}_p$ *are orthonormal, as in Theorem* 60*, then* (7.156) *is true and the covariance matrix* $\mathbb{E}_{vv}$ *is reduced to the identity. If operators* $\mathcal{Q}_1,\dots,\mathcal{Q}_p$ *are orthogonal, as in Theorem* 61*, then* (7.186) *holds and the*

covariance matrix $\mathbb{E}_{vv}$ *is reduced to a block-diagonal form with non-zero blocks* $\mathbb{E}_{v_1v_1}$, ..., $\mathbb{E}_{v_pv_p}$ *so that*

$$\mathbb{E}_{vv} = \begin{bmatrix} \mathbb{E}_{v_1v_1} & \mathbb{O} & \dots & \mathbb{O} \\ \mathbb{O} & \mathbb{E}_{v_2v_2} & \dots & \mathbb{O} \\ \dots & \dots & \dots & \dots \\ \mathbb{O} & \mathbb{O} & \dots & \mathbb{E}_{v_pv_p} \end{bmatrix}$$

with $\mathbb{O}$ *denoting the zero block. As a result, the procedure for finding* f^0, $\mathcal{F}_1^0$, ..., $\mathcal{F}_p^0$ *is reduced to* p *separate rank-constrained problems* (7.159) *or* (7.188). *Unlike the methods in* [159, 167, 173, 182], *the operators* $\mathcal{F}_1^0, \dots, \mathcal{F}_0^p$ *are determined with* much smaller $m \times n$ *and* $n \times n$ *matrices given by the simple formulae* (7.147) *and* (7.174)–(7.176). *This implies a reduction in computational work compared with that required by the approach in* [159, 167, 173, 182]. *In Table* 5 *of Section* 7.6.8, *this observation is illustrated with results from numerical simulations.*

Corollary 20. *Let* $\mathbf{v}_1, \dots, \mathbf{v}_p$ *be determined by Lemma* 34 *of Section* 5.7.4. *Then the vector* $\bar{f}$ *and operators* $\bar{\mathcal{F}}_1, \dots, \bar{\mathcal{F}}_p$, *satisfying the unconstrained minimum* (7.140), *are determined by*

$$\bar{f} = E[\mathbf{x}] - \sum_{k=1}^{p} \bar{F}_k E[\mathbf{v}_k] \tag{7.191}$$

and

$$\bar{F}_1 = \mathbb{E}_{xv_1}\mathbb{E}^{\dagger}_{v_1v_1} + M_1[I - \mathbb{E}_{v_1v_1}\mathbb{E}^{\dagger}_{v_1v_1}], \tag{7.192}$$

$$\vdots$$

$$\bar{F}_p = \mathbb{E}_{xv_p}\mathbb{E}^{\dagger}_{v_pv_p} + M_p[I - \mathbb{E}_{v_pv_p}\mathbb{E}^{\dagger}_{v_pv_p}] \tag{7.193}$$

with $M_1, \dots, M_p$ *defined in Theorem* 61. *The associated accuracy for transform* $\bar{\mathcal{T}}_p$, *defined by*

$$\bar{\mathcal{T}}_p(\mathbf{y}) = \bar{f} + \sum_{k=1}^{p} \bar{\mathcal{F}}_k(\mathbf{v}_k),$$

is given by

$$E[\|\mathbf{x} - \bar{\mathcal{T}}_p(\mathbf{y})\|^2] = \|\mathbb{E}_{xx}^{1/2}\|^2 - \sum_{k=1}^{p} \|\mathbb{E}_{xv_k}(\mathbb{E}^{1/2}_{v_kv_k})^{\dagger}\|^2. \tag{7.194}$$

Proof. It follows from (7.187) that the unconstrained minimum (7.140) is achieved if f is defined by (7.191) and if F_k satisfies the equation

$$F_k \mathbb{E}_{v_k v_k}^{1/2} - \mathbb{E}_{x v_k} (\mathbb{E}_{v_k v_k}^{1/2})^\dagger = \mathbb{O}$$

for each $k = 1, \ldots, p$. Similar to (7.189), its general solution is given by

$$F_k = \bar{F}_k = \mathbb{E}_{x v_k} \mathbb{E}_{v_k v_k}^\dagger + M_k [I - \mathbb{E}_{v_k v_k} \mathbb{E}_{v_k v_k}^\dagger].$$

because

$$\mathbb{E}_{v_k v_k}^{1/2} (\mathbb{E}_{v_k v_k}^{1/2})^\dagger = \mathbb{E}_{v_k v_k} \mathbb{E}_{v_k v_k}^\dagger.$$

We define $\bar{\mathcal{F}}_k$ by $[\bar{\mathcal{F}}_k(\mathbf{w}_k)](\omega) = \bar{F}_k[\mathbf{w}_k(\omega)]$ for all $k = 1, \ldots, p$, and then (7.192)–(7.193) are true. The relation (7.194) follows from (7.187) and (7.191)–(7.193). □

Remark 41. *The transforms given by Theorems* 60 *and* 61 *are not unique due to arbitrary operators* $\mathcal{M}_1$, …, $\mathcal{M}_p$. *A natural particular choice is* $\mathcal{M}_1 = \ldots = \mathcal{M}_p = \mathbb{O}$.

7.6.5 Compression procedure by $\mathcal{T}_p^0$

Let us consider transform $\mathcal{T}_p^0$ given by (60), (7.174)–(7.176) with $M_k = \mathbb{O}$ for $k = 1, \ldots, p$ where M_k is the matrix given in (7.189). We write $[\mathcal{T}_p^0(\mathbf{y})](\omega) = T_p^0(y)$ with $T_p^0 : \mathbb{R}^n \to \mathbb{R}^m$.

Let

$$B_k^{(1)} = S_{k\eta_k} V_{k\eta_k} D_{k\eta_k}^T \quad \text{and} \quad B_k^{(2)} = D_{k\eta_k}^T (\mathbb{E}_{v_k v_k}^{1/2})^\dagger$$

so that $B_k^{(1)} \in \mathbb{R}^{m \times \eta_k}$ and $B_k^{(2)} \in \mathbb{R}^{\eta_k \times n}$. Here, η_1, …, η_p are determined by (7.141). Then

$$T_p^0(y) = f + \sum_{k=1}^{p} B_k^{(1)} B_k^{(2)} v_k,$$

where $v_k = \mathbf{v}_k(\omega)$ and $B_k^{(2)} v_k \in \mathbb{R}^{\eta_k}$ for $k = 1, \ldots, p$ with $\eta_1 + \ldots + \eta_p < m$. Hence, matrices $B_1^{(2)}, \ldots, B_p^{(2)}$ perform compression of the data presented by $v_1, \ldots, v_p$. Matrices $B_1^{(1)}, \ldots, B_p^{(1)}$ perform reconstruction of the reference signal from the compressed data.

The compression ratio of transform $\mathcal{T}_p^0$ is given by

$$c^0 = (\eta_1 + \ldots + \eta_p)/m. \tag{7.195}$$

7.6.6 Special cases of transform $\mathcal{T}_p$

Choice of operators $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_p$

The results above have been derived for any operators $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_p$ in the model $\mathcal{T}_p$. Some specializations for $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_p$ were given in Section 7.6.2. Here and in Section 7.6.6, we consider alternative forms for $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_p$.

(i) Operators $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_p$ can be determined by a recursive procedure as follows. First, we set $\boldsymbol{\varphi}_k(\mathbf{y}) = \mathbf{y}$ and determine estimate $\mathbf{x}^{(1)}$ of $\mathbf{x}$ from the solution of problem (7.140) (with no constraints (7.141)) by Corollaries 19 or 20 with $p = 1$. Next, we put

$$\boldsymbol{\varphi}_1(\mathbf{y}) = \mathbf{y} \quad \text{and} \quad \boldsymbol{\varphi}_2(\mathbf{y}) = \mathbf{x}^{(1)},$$

and find estimate $\mathbf{x}^{(2)}$ from the solution of unconstrained problem (7.140) with $p = 2$. In general, for $j = 1, \ldots, p$, we define

$$\boldsymbol{\varphi}_j(\mathbf{y}) = \mathbf{x}^{(j-1)},$$

where $\mathbf{x}^{(j-1)}$ has been determined similarly to $\mathbf{x}^{(2)}$ from the previous steps. In particular, $\mathbf{x}^{(0)} = \mathbf{y}$.

(ii) Operators $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_p$ can also be chosen as elementary functions. In item (i) of Section 7.6.2, $\boldsymbol{\varphi}_k(\mathbf{y})$ was constructed from the power functions. An alternative possibility is to choose trigonometric functions for $\boldsymbol{\varphi}_k(\mathbf{y})$. One can put

$$[\boldsymbol{\varphi}_1(\mathbf{y})](\omega) = y \quad \text{and} \quad [\boldsymbol{\varphi}_{k+1}(\mathbf{y})](\omega) = [\cos(ky_1), \ldots, \cos(ky_n)]^T \tag{7.196}$$

with $y = [y_1, \ldots, y_n]^T$ and $k = 1, \ldots, p-1$.

Special form of the constraint

While the statement of the problem in the form (7.140)–(7.141) with p constraints allows us to facilitate a computational load, other forms of the constraint can lead to some alternative advantages. In particular, if in (7.140)–(7.141), $p = q, q+1, \ldots$, where $q = \min\{m, n\}$, and $F_j \neq \mathbb{O}$ for all $j = 1, \ldots, p$ then even for the minimal possible ranks $\eta_1 = 1, \ldots, \eta_p = 1$ in (7.141), the compression ratio is

$$c = \frac{p}{q} \geq 1,$$

i.e. no compression of $\mathbf{x}$ can be achieved. To avoid such a bottle-neck, we consider the case when p constraints (7.141) are replaced with the one constraint in the form

$$\text{rank } [F_1 \ldots F_p] \leq r \leq q. \tag{7.197}$$

Then the problem is to find $f^0, F_1^0, \ldots, F_p^0$ satisfying (7.140) subject to (7.197). As in Section 7.6.3, it is supposed that the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$ in (7.137), (7.138) and (7.140) are orthogonal.

Let the SVD of the matrix $[E_{xv_1} E_{v_1 v_1}^{\dagger 1/2} \ldots E_{xv_p} E_{v_p v_p}^{\dagger 1/2}]$ be

$$U \Sigma V^T = [E_{xv_1} E_{v_1 v_1}^{\dagger 1/2} \ldots E_{xv_p} E_{v_p v_p}^{\dagger 1/2}], \tag{7.198}$$

where $U \in \mathbb{R}^{m \times n}$, $V \in \mathbb{R}^{n \times n}$ are orthogonal and $\Sigma \in \mathbb{R}^{n \times n}$ is diagonal,

$$U = [s_1, \ldots, s_n], \quad V = [d_1, \ldots, d_n], \tag{7.199}$$

$$\Sigma = \operatorname{diag}(\beta_1, \ldots, \beta_n) \tag{7.200}$$

with $\beta_1 \geq \cdots \geq \beta_l > 0$, $\beta_{l+1} = \beta_n = 0$ and $l = 1, \ldots, n$.

Let us set

$$U_r = [s_1, \ldots, s_r], \quad V_r = [d_1, \ldots, d_r] \tag{7.201}$$

$$\Sigma_r = \operatorname{diag}(\beta_1, \ldots, \beta_r), \tag{7.202}$$

where $U_r \in \mathbb{R}^{m \times r}$, $V_r \in \mathbb{R}^{n \times r}$ and $\Sigma_r \in \mathbb{R}^{r \times r}$. Now we define $G_r \in \mathbb{R}^{m \times n}$ and $\mathcal{G}_r : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^m)$ by

$$G_r = U_r \Sigma_r V_r^T \quad \text{and} \quad [\mathcal{G}_r(\mathbf{w})](\omega) = G_r[\mathbf{w}(\omega)], \tag{7.203}$$

respectively, for any $\mathbf{w} \in L^2(\Omega, \mathbb{R}^n)$. The matrix G_r can be represented in a block form

$$G_r = [B_1 \ldots B_p], \tag{7.204}$$

where $B_j = G_r[:, (j-1)n + 1 : jn] \in \mathbb{R}^{m \times n}$ is a block formed by the n subsequent columns of the matrix G_r beginning from the $((j-1)n+1)$th column.

The family of solutions to the problem (7.140), (7.197) is given by the following theorem.

Theorem 62. *The vector f^0 and matrices F_1^0, ..., F_p^0 that satisfy (7.140) and (7.197) are such that*

$$f^0 = E[\mathbf{x}] - \sum_{j=1}^{p} F_j^0 E[\mathbf{v}_j] \tag{7.205}$$

and

$$F_j^0 = B_j E_{v_j v_j}^{1/2 \dagger} + M_j (I - E_{v_j v_j}^{1/2} E_{v_j v_j}^{1/2 \dagger}), \quad j = 1, \ldots, p, \tag{7.206}$$

where $M_j \in \mathbb{R}^{m\times n}$ *is an arbitrary matrix such that rank* $[F_1^0 \dots F_p^0] \leq r$, *defined similarly to* M_j *in Theorem* 61.

The error associated with the transform defined by (7.137), (7.138), (7.205), (7.206) *is given by*

$$E[\|\mathbf{x} - \mathcal{T}_p^0(\mathbf{y})\|^2] = \|\mathbb{E}_{xx}^{1/2}\|^2 - \sum_{j=1}^{r} \beta_j^2. \tag{7.207}$$

Proof. The proof follows from thew proof of Theorem 61 and from the fact that

$$\mathbb{E}_{vv}^{1/2} = \begin{bmatrix} \mathbb{E}_{v_1v_1}^{1/2} & \mathbb{O} & \dots & \mathbb{O} \\ \mathbb{O} & \mathbb{E}_{v_2v_2}^{1/2} & \dots & \mathbb{O} \\ \dots & \dots & \dots & \dots \\ \mathbb{O} & \mathbb{O} & \dots & \mathbb{E}_{v_pv_p}^{1/2} \end{bmatrix}$$

and

$$\mathbb{E}_{vv}^{1/2\dagger} = \begin{bmatrix} \mathbb{E}_{v_1v_1}^{1/2\dagger} & \mathbb{O} & \dots & \mathbb{O} \\ \mathbb{O} & \mathbb{E}_{v_2v_2}^{1/2\dagger} & \dots & \mathbb{O} \\ \dots & \dots & \dots & \dots \\ \mathbb{O} & \mathbb{O} & \dots & \mathbb{E}_{v_pv_p}^{1/2\dagger} \end{bmatrix}.$$

□

Corollary 21. *The accuracy associated with transform given by Theorem* 62 *is better than that of the transform given by Theorem* 61 *if*

$$\sum_{k=1}^{p}\sum_{j=1}^{\eta_k} \beta_{kj}^2 < \sum_{j=1}^{r} \beta_j^2. \tag{7.208}$$

Proof. The proof follows directly from the proofs of Theorems 61 and 62. □

In Section 7.6.9, this transform is illustrated with numerical simulations.

7.6.7 Other particular cases of transform $\mathcal{T}_p$: comparison with known transforms

Optimal non-linear filtering

.

The transforms $\hat{\mathcal{T}}_p$ (7.162)–(7.163) and $\bar{\mathcal{T}}_p$ (7.191)–(7.193), which are particular cases of the transforms given in Theorems 1 and 2, represent optimal filters that perform pure filtering with no signal compression. Therefore they are important in their own right.

The Fourier series as a particular case of transform $\bar{\mathcal{T}}_p$.

For the case of the minimization problem (7.140) with no constraint (7.141), $\mathcal{F}_1, \ldots, \mathcal{F}_p$ are determined by the expressions (7.162) and (7.191)–(7.193) which are similar to those for the Fourier coefficients [20]. The structure of the model $\mathcal{T}_p$ presented by (7.137) is different, of course, from that for the Fourier series and Fourier polynomial (i.e. a truncated Fourier series) in Hilbert space [20]. The differences are that $\mathcal{T}_p$ transforms $\mathbf{y}$ (not $\mathbf{x}$ as the Fourier polynomial does) and that $\mathcal{T}_p$ consists of a combination of three operators $\boldsymbol{\varphi}_k$, $\mathcal{Q}_k$ and $\mathcal{F}_k$ where

$$\mathcal{F}_k : L^2(\Omega, \tilde{H}_k) \to L^2(\Omega, H_X)$$

is an operator, not a scalar as in the Fourier series [20]. The solutions (7.162) and (7.191)–(7.193) of the unconstrained problem (7.140) are given in terms of the observed vector $\mathbf{y}$, not in terms of the basis of $\mathbf{x}$ as in the Fourier series/polynomial. The special features of $\mathcal{T}_p$ require special computation methods as described above and in Section 7.6.8 below.

Here, we show that the Fourier series is a particular case of the transform $\mathcal{T}_p$.

Let $\mathbf{x} \in L^2(\Omega, H)$ with H a Hilbert space, and let $\{\mathbf{v}_1, \mathbf{v}_2, \ldots\}$ be an orthonormal basis in $L^2(\Omega, H)$. For any $\mathbf{g}, \mathbf{h} \in L^2(\Omega, H)$, we define the scalar product $\langle \cdot, \cdot \rangle$ and the norm $\| \cdot \|_E$ in $L^2(\Omega, H)$ by

$$\langle \mathbf{g}, \mathbf{h} \rangle = \int_\Omega \mathbf{g}(\omega)\mathbf{h}(\omega) d\mu(\omega) \quad \text{and} \quad \|\mathbf{g}\|_E = \langle \mathbf{g}, \mathbf{g} \rangle^{1/2}, \tag{7.209}$$

respectively. In particular, if $H = \mathbb{R}^m$ then

$$\|\mathbf{g}\|_E^2 = \int_\Omega \mathbf{g}(\omega)[\mathbf{g}(\omega)]^T d\mu(\omega) = \int_\Omega \|\mathbf{g}(\omega)\|^2 d\mu(\omega) = E[\|\mathbf{g}\|^2] \tag{7.210}$$

i.e. $E[\|\mathbf{g}\|^2]$ is defined similarly to that in (7.139).

Let us consider the special case of transform $\mathcal{T}_p$ presented in item (iii) of Section 7.6.2 and let us also consider the unconstrained problem (7.140) formulated in terms of such a $\mathcal{T}_p$ where we now assume that $\mathbf{x}$ has the zero mean, $f = \mathbb{O}$, $p = \infty$, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots\}$ is an orthonormal basis in $L^2(\Omega, H)$ and $\mathcal{F}_k$ is a scalar, not an operator as before. We denote $\alpha_k = \mathcal{F}_k$ with $\alpha_k \in \mathbb{R}$. Then similar to (7.162) in Corollary 19, the solution to unconstrained problem (7.140) is defined by $\hat{\alpha}_k$ such that

$$\hat{\alpha}_k = \mathbb{E}_{xv_k} \quad \text{with} \quad k = 1, 2, \ldots.$$

Here,

$$\mathbb{E}_{xv_k} = E[\mathbf{x}\mathbf{v}_k] - E[\mathbf{x}]E[\mathbf{v}_k] = E[\mathbf{x}\mathbf{v}_k] = \langle \mathbf{x}, \mathbf{v}_k \rangle$$

since $E[\mathbf{x}] = 0$ by the assumption. Hence, $\hat{\alpha}_k = \mathbb{E}_{xv_k}$ is the Fourier coefficient and the considered particular case of $\mathcal{T}_p(\mathbf{y})$ with $\mathcal{F}_k$ determined by $\hat{\alpha}_k$ is given by

$$\mathcal{T}_p(\mathbf{y}) = \sum_{k=1}^{\infty} \langle \mathbf{x}, \mathbf{v}_k \rangle \mathbf{v}_k. \tag{7.211}$$

Thus, the Fourier series (7.211) in Hilbert space follows from (7.137), (7.140) and (7.162) when $\mathcal{T}_p$ has the form given in item (iii) of Section 7.6.2 with $\mathbf{x}$, f, p, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots\}$ and $\mathcal{F}_k$ as above.

The Wiener filter as a particular case of transform $\bar{\mathcal{T}}_p$ (7.191)–(7.193)

In the following Corollaries 22 and 23, we show that the filter $\bar{\mathcal{T}}_p$ guarantees better accuracy than that of the Wiener filter.

Corollary 22. *Let* $p = 1$, $E[\mathbf{x}] = 0$, $E[\mathbf{y}] = 0$, $\boldsymbol{\varphi}_1 = \mathcal{I}$, $\mathcal{Q}_1 = \mathcal{I}$ *and* $M_1 = \mathbb{O}$ *or* $M_1 = E_{xy}E_{yy}^{\dagger}$. *Then* $\bar{\mathcal{T}}_p$ *is reduced to the filter* $\check{\mathcal{T}}$ *such that* $[\check{\mathcal{T}}(\mathbf{y})](\omega) = \check{T}[\mathbf{y}(\omega)]$ *with*

$$\check{T} = E_{xy}E_{yy}^{\dagger}. \tag{7.212}$$

Proof. Let $\bar{T}_1$ be such that $[\bar{\mathcal{T}}_1(\mathbf{y})](\omega) = \bar{T}_1[\mathbf{y}(\omega)]$. For $E[\mathbf{x}] = 0$ and $E[\mathbf{y}] = 0$, we have $\bar{f} = \mathbb{O}$. If $\boldsymbol{\varphi}_1 = I$ and $\mathcal{Q}_1 = I$ then $\mathbf{v}_1 = \mathbf{y}$.

Next, for $A_1 = \mathbb{E}_{xv_1}\mathbb{E}_{v_1v_1}^{\dagger}$, on the basis of (7.183), one has

$$\bar{T}_1 = E_{xy}E_{yy}^{\dagger} + E_{xy}E_{yy}^{\dagger} - E_{xy}E_{yy}^{\dagger}E_{yy}E_{yy}^{\dagger} = E_{xy}E_{yy}^{\dagger} = \check{T}.$$

The case when $A_1 = \mathbb{O}$ is obvious. Hence, (7.212) is true. □

Remark 42. *The unconstrained linear filter, given by* (7.212)*, has been proposed in* [63]. *The filter* (7.212) *is treated as a generalization of the Wiener filter.*

Let $\tilde{\mathbf{x}}$, $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_p$ be the zero mean vectors. The transform $\bar{\mathcal{T}}_p$, applied to $\tilde{\mathbf{x}}$, $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_p$, is denoted by $\bar{\mathcal{T}}_{W,p}$.

Corollary 23. *The error* $E[\|\tilde{\mathbf{x}} - \bar{\mathcal{T}}_{W,p}(\tilde{\mathbf{y}})\|^2]$ *associated with the transform* $\bar{\mathcal{T}}_{W,p}$ *is smaller than the error* $E[\|\tilde{\mathbf{x}} - \check{\mathcal{T}}(\tilde{\mathbf{y}})\|^2]$ *associated with the Wiener filter* [63] *by* $\sum_{k=2}^{p} \|E_{\tilde{x}\tilde{v}_k}(E_{\tilde{v}_k\tilde{v}_k}^{1/2})^{\dagger}\|^2$, *i.e.*

$$E[\|\tilde{\mathbf{x}} - \bar{\mathcal{T}}_{W,p}(\tilde{\mathbf{y}})\|^2] = E[\|\tilde{\mathbf{x}} - \check{\mathcal{T}}(\tilde{\mathbf{y}})\|^2] - \sum_{k=2}^{p} \|E_{\tilde{x}\tilde{v}_k}(E_{\tilde{v}_k\tilde{v}_k}^{1/2})^{\dagger}\|^2. \tag{7.213}$$

Proof. It is easy to show that

$$E[\|\tilde{\mathbf{x}} - \check{\mathcal{T}}(\tilde{\mathbf{y}})\|^2] = \|E_{\tilde{x}\tilde{x}}^{1/2}\|^2 - \|E_{\tilde{x}\tilde{v}_1}(E_{\tilde{v}_1\tilde{v}_1}^{1/2})^{\dagger}\|^2, \tag{7.214}$$

and then (7.213) follows from (7.194) and (7.214). □

PCA-KLT as a particular case of transform $\mathcal{T}_p^0$ (7.174)–(7.178).

PCA-KLT [63] follows from (7.174)–(7.178) as a particular case if

$$f = \mathbb{O}, \quad p = 1, \quad \boldsymbol{\varphi}_1 = I, \quad \mathcal{Q}_1 = I \quad \text{and} \quad A_1 = \mathbb{O}.$$

To compare the transform $\mathcal{T}_p^0$ with PCA-KLT [63], we apply $\mathcal{T}_p^0$, represented by (7.174)–(7.178), to the zero mean vectors $\tilde{\mathbf{x}}$, $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_p$ as above. We write $\mathcal{T}_p^*$ for such a version of $\mathcal{T}_p^0$, and $\mathcal{T}_{PCA-KLT}$ for PCA-KLT [63].

Corollary 24. *The error $E[\|\tilde{\mathbf{x}} - \mathcal{T}_p^*(\tilde{\mathbf{y}})\|^2]$ associated with the transform $\mathcal{T}_p^*$ is smaller than the error $E[\|\tilde{\mathbf{x}} - \mathcal{T}_{PCA-KLT}(\tilde{\mathbf{y}})\|^2]$ associated with PCA-KLT* [63] *by $\sum_{k=2}^{p}\sum_{j=1}^{\eta_k}\beta_{kj}^2$, i.e.*

$$E[\|\tilde{\mathbf{x}} - \mathcal{T}_p^*(\tilde{\mathbf{y}})\|^2] = E[\|\tilde{\mathbf{x}} - \mathcal{T}_{PCA-KLT}(\tilde{\mathbf{y}})\|^2] - \sum_{k=2}^{p}\sum_{j=1}^{\eta_k}\beta_{kj}^2. \tag{7.215}$$

Proof. The error associated with $\mathcal{F}_{PCA-KLT}$ [63] is represented by (7.178) for $p = 1$,

$$E[\|\tilde{\mathbf{x}} - \mathcal{T}_{PCA-KLT}(\tilde{\mathbf{y}})\|^2] = \|E_{\tilde{x}\tilde{x}}^{1/2}\|^2 - \sum_{j=1}^{\eta_1}\beta_{1j}^2. \tag{7.216}$$

Then (7.215) follows from (7.178) and (7.216). □

The transform [158] as a particular case of transform $\mathcal{T}_p^0$.

The transform [158] follows from (7.137) as a particular case if

$$f = \mathbb{O}, \quad p = 2, \quad \boldsymbol{\varphi}_1(\mathbf{y}) = \mathbf{y}, \quad \boldsymbol{\varphi}_2(\mathbf{y}) = \mathbf{y}^2 \quad \text{and} \quad \mathcal{Q}_1 = \mathcal{Q}_2 = I$$

where $\mathbf{y}^2$ is defined by $\mathbf{y}^2(\omega) = [y_1^2, \ldots, y_n^2]^T$. We note that transform [158] has been generalized in [173].

The transforms [173] as particular cases of transform $\mathcal{T}_p$.

The transform [173] follows from (7.137) if

$$\mathcal{Q}_k = \mathcal{I}, \quad \boldsymbol{\varphi}_k(\mathbf{y}) = \mathbf{y}^k \quad \text{where} \quad \mathbf{y}^k = (\mathbf{y}, \dots, \mathbf{y}) \in L^2(\Omega, \mathbb{R}^{nk})$$

with $\mathbb{R}^{nk}$ is the kth degree of $\mathbb{R}^n$, and if $\mathcal{F}_k$ is a k-linear operator.

To compare transform $\mathcal{T}_p^0$ and transform $\mathcal{T}_{[173]}$ [173] of rank r, we write

$$z_j = y_j y, \quad z = [z_1, \dots, z_n]^T, \quad s = [1 \; y^T \; z^T]^T$$

and denote by $\alpha_1, \dots, \alpha_r$ the non-zero singular values associated with the truncated SVD for the matrix $\mathbb{E}_{xs}(\mathbb{E}_{ss}^{1/2})^{\dagger}$. Such a SVD is constructed similarly to that in (7.167)–(7.170).

Corollary 25. *Let*

$$\Delta_p = \sum_{k=1}^{p} \sum_{j=1}^{\eta_k} \beta_{kj}^2 - \sum_{j=1}^{r} \alpha_j^2$$

and let $\Delta_p \geq 0$. The error $E[\|\mathbf{x} - \mathcal{T}_p^0(\mathbf{y})\|^2]$ associated with the transform $\mathcal{T}_p^0$ is less than the error $E[\|\mathbf{x} - \mathcal{T}_{[173]}(\mathbf{y})\|^2]$ associated with the transform $\mathcal{T}_{[173]}$ by Δ_p, i.e.

$$E[\|\mathbf{x} - \mathcal{T}_p^0(\mathbf{y})\|^2] = E[\|\mathbf{x} - \mathcal{T}_{[173]}(\mathbf{y})\|^2] - \Delta_p. \tag{7.217}$$

Proof. It follows from [173] that

$$E[\|\mathbf{x} - \mathcal{T}_{[173]}(\mathbf{y})\|^2] = \|E_{xx}^{1/2}\|^2 - \sum_{j=1}^{r} \alpha_j^2. \tag{7.218}$$

Then (7.217) follows from (7.178) and (7.218). □

We note that, in general, a theoretical verification of the condition $\Delta_p \geq 0$ is not straightforward. At the same time, for any particular $\mathbf{x}$ and $\mathbf{y}$, Δ_p can be estimated numerically. In the case when, for a given p, the condition $\Delta_p \geq 0$ is not fulfilled, the accuracy $E[\|\mathbf{x} - \mathcal{T}_p^0(\mathbf{y})\|^2]$ can be improved by increasing p or by applying the hybrid method presented in Chapter 4.

As we have noted before, the method in [173] requires much more computational work than that needed for transform $\mathcal{T}_p^0$.

The results of numerical experiments presented in Section 7.6.9 below demonstrate the superiority of the proposed transforms given in Theorems 60, 61 and Corollaries 19, 20 over transforms [63, 158, 173].

Remark 43. *Unlike the technique presented in* [147]*, the above method implements simultaneous filtering and compression, and provides this data processing in probabilistic setting. The idea of implicitly mapping the data into a high-dimensional feature space* [21] *could be extended to the transform presented in this paper.*

7.6.8 Numerical realization

(i) *Orthogonalization.* Numerical realization of transforms of random vectors implies a representation of observed data and estimates of covariance matrices in the form of associated samples. For the random vector $\mathbf{u}_k$, we have q realizations, which are concatenated into $n \times q$ matrix U_k. A column of U_k is a realization of $\mathbf{u}_k$. Thus, a sequence of vectors $\mathbf{u}_1, \ldots, \mathbf{u}_p$ is represented by a sequence of matrices $U_1, \ldots, U_p$. Therefore the transformation of $\mathbf{u}_1, \ldots, \mathbf{u}_p$ to orthonormal or orthogonal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$ (by Lemmata 33 and 34 in Section 5.7.4) is reduced to a procedure for matrices $U_1, \ldots, U_p$ and $V_1, \ldots, V_p$. Here, $V_k \in \mathbb{R}^{n \times q}$ is a matrix formed from realizations of the random vector $\mathbf{v}_k$ for each $k = 1, \ldots, p$.

Alternatively, matrices $V_1, \ldots, V_p$ can be determined from known procedures for matrix orthogonalization [50]. In particular, the QR decomposition [50] can be exploited in the following way. Let us form a matrix $U = [U_1^T \ldots U_p^T]^T \in \mathbb{R}^{np \times q}$ where p and q are chosen such that $np = q$, i.e. U is square.[6] Let

$$U = VR$$

be the QR decomposition for U with $V \in \mathbb{R}^{np \times q}$ orthogonal and $R \in \mathbb{R}^{np \times q}$ upper triangular. Next, we write $V = [V_1^T \ldots V_p^T]^T \in \mathbb{R}^{mp \times q}$ where $V_k \in \mathbb{R}^{n \times q}$ for $k = 1, \ldots, p$. The sub-matrices $V_1, \ldots, V_p$ of V are orthogonal, i.e. $V_i V_j^T = \begin{cases} \mathbb{O}, & i \neq j, \\ I, & i = j, \end{cases}$, for $i, j = 1, \ldots, p$, as required.

Other known procedures for matrix orthogonalization can be applied to $U_1, \ldots, U_p$ in a similar fashion.

Remark 44. *For the cases when* $\mathbf{v}_1, \ldots, \mathbf{v}_p$ *are orthonormal or orthogonal but not orthonormal, the associated accuracies* (7.149)*,* (7.164)*,* (7.178) *and* (7.194) *differ for the factors depending on* $(\mathbb{E}_{v_k v_k}^{1/2})^\dagger$*. In the case of orthonormal* $\mathbf{v}_1, \ldots, \mathbf{v}_p$*,* $(\mathbb{E}_{v_k v_k}^{1/2})^\dagger = I$ *and this circumstance can lead to an increase in the accuracy.*

(ii) *Covariance matrices.* The expectations and covariance matrices in Theorems 60 and 61 and Corollaries 19 and 20 can be estimated, for example, by the techniques considered in Section 5.3 of Chapter 4.

[6] Matrix U can also be presented as $U = [U_1 \ldots U_p]$ with p and q such that $n = pq$.

(iii) *Transforms $\mathcal{T}_p^0$, $\hat{\mathcal{T}}_{\mathbf{p}}$ and $\bar{\mathcal{T}}_{\mathbf{p}}$ for zero mean vectors.* The computational work for $\mathcal{T}_p^0$ (Theorems 60 and 61), $\hat{\mathcal{T}}_p$ and $\bar{\mathcal{T}}_p$ (Corollaries 19 and 20, respectively) can be reduced if $\mathcal{T}_p^0$, $\hat{\mathcal{T}}_p$ and $\bar{\mathcal{T}}_p$ are applied to the zero mean vectors $\tilde{\mathbf{x}}$, $\tilde{\mathbf{v}}_1$, ..., $\tilde{\mathbf{v}}_p$ given by

$$\tilde{\mathbf{x}} = \mathbf{x} - E[\mathbf{x}], \quad \tilde{\mathbf{v}}_1 = \mathbf{v}_1 - E[\mathbf{v}_1], \ldots, \tilde{\mathbf{v}}_p = \mathbf{v}_p - E[\mathbf{v}_p].$$

Then $f^0 = \mathbb{O}$ and $\bar{f} = \mathbb{O}$. The estimates of the original $\mathbf{x}$ are then given by

$$\check{\mathbf{x}} = E[\mathbf{x}] + \sum_{k=1}^{p} \mathcal{F}_k^0(\tilde{\mathbf{v}}_k), \quad \hat{\mathbf{x}} = E[\mathbf{x}] + \sum_{k=1}^{p} \hat{\mathcal{F}}_k(\tilde{\mathbf{v}}_k)$$

and

$$\bar{\mathbf{x}} = E[\mathbf{x}] + \sum_{k=1}^{p} \bar{\mathcal{F}}_k(\tilde{\mathbf{v}}_k)$$

respectively. Here, $\mathcal{F}_k^0$, $\hat{\mathcal{F}}_k$ and $\bar{\mathcal{F}}_k$ are defined similarly to (7.147), (7.162), (7.175), (7.176), (7.192) and (7.193).

7.6.9 Simulations

Simulations to illustrate transforms by Theorem 61 and Corollary 20

The transforms $\mathcal{T}_2^0$ (Theorem 61), $\tilde{\mathcal{T}}_2$ (Corollary 20) and the known transforms [63], [158], [173] have been applied to compression, filtering and subsequent restoration of the reference signal given by the matrix $X \in \mathbb{R}^{256\times256}$. The matrix X represents the data obtained from an aerial digital photograph of a plant[7].

We divide X into $m \times q$ sub-matrices $X_{ij} \in \mathbb{R}^{m\times q}$ with $i = 1, \ldots, 256/m$ and $j = 1, \ldots, 256/q$. By assumption, the sub-matrix X_{ij} is interpreted as q realizations of a random vector $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$ with each column representing a realization. Observed data were modelled in the form

$$Y_{ij} = \bar{X}_{ij} \bullet \bar{X}_{ij} \bullet \bar{X}_{ij}, \tag{7.219}$$

where the symbol $\bullet$ denotes the Hadamard product and $\bar{X}_{ij}$ represents q realizations of the vector $\bar{\mathbf{x}} = \mathbf{x} - E[\mathbf{x}]$.

The proposed transform T_2^0, given by (7.137) and (7.174)–(7.176) for $p = 2$, and the transforms $T_{[63]}$ [63], $T_{[158]}$ [158] and $T_{[173]}$ [173] have been applied to each pair X_{ij}, Y_{ij}.

[7]The database is available in http://sipi.usc.edu/services/database/Database.html.

Table 4. Performance comparison of transforms T_2^0 and $T_{[63]}$.

$m = 8, \quad q = 16$						
	Accuracy			Flops		
cr	$J_{[63]}$	J_2^0	$J_{[63]/2}$	$K_{[63]}$	$K_{T_2^0}$	$K_{[63]}/K_{T_2^0}$
–	5.95×10^6	2.01×10^{-25}	1.34×10^{33}	1.14×10^7	3.20×10^7	0.36
1/2	5.95×10^6	3.97×10^3	6.64×10^5	2.37×10^7	5.50×10^7	0.43
1/4	5.95×10^6	3.30×10^4	5.50×10^4	2.35×10^7	5.21×10^7	0.43
$m = 16, \quad q = 32$						
	Accuracy			Flops		
cr	$J_{[63]}$	J_2^0	$J_{[63]/2}$	$K_{[63]}$	$K_{T_2^0}$	$K_{[63]}/K_{T_2^0}$
–	9.82×10^6	5.09×10^{-25}	4.1×10^{31}	2.08×10^7	6.03×10^7	0.35
1/2	9.82×10^6	7.10×10^3	1.70×10^4	4.34×10^7	9.81×10^7	0.45
1/4	9.84×10^6	7.50×10^4	2.46×10^3	4.29×10^7	9.81×10^7	0.44
$m = 32, \quad q = 64$						
	Accuracy			Flops		
cr	$J_{[63]}$	J_2^0	$J_{[63]/2}$	$K_{[63]}$	$K_{T_2^0}$	$K_{[63]}/K_{T_2^0}$
–	1.81×10^7	1.16×10^{-24}	8.6×10^{30}	3.95×10^7	1.18×10^8	0.33
1/2	1.82×10^7	1.75×10^4	6.73×10^3	8.21×10^7	1.88×10^8	0.44
1/4	1.83×10^7	2.00×10^5	1.05×10^3	8.12×10^7	1.86×10^8	0.44

Operators $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ in (7.137) have been defined so that

$$\boldsymbol{\varphi}_1(\mathbf{y}) = \mathbf{y} \quad \text{and} \quad \boldsymbol{\varphi}_2(\mathbf{y}) = \mathbf{x}^{(1)}$$

where $\mathbf{x}^{(1)}$ has been determined from the procedure presented in item (i) of Section 7.6.6. Orthogonal matrices V_1, V_2 for the transform T_2^0 have been determined from the QR decomposition as described in Section 7.6.8. Covariance matrices have been estimated from the associated samples with the method given in Section 5.3.1.

We write

$$J_{[63]} = \max_{ij} \|X_{ij} - T_{[63]}Y_{ij}\|^2,$$

$$J_{[158]} = \max_{ij} \|X_{ij} - T_{[158]}Y_{ij}\|^2,$$

$$J_2^0 = \max_{ij} \|X_{ij} - T_2^0 Y_{ij}\|^2,$$

$$J_{[173]} = \max_{ij} \|X_{ij} - T_{[173]}Y_{ij}\|^2,$$

and

$$J_{[63]/2} = \max_{ij}[\|X_{ij} - T_{[63]}Y_{ij}\|^2]/\|X_{ij} - T_2^0 Y_{ij}\|^2],$$

$$J_{[158]/2} = \max_{ij}[\|X_{ij} - T_{[158]}Y_{ij}\|^2/\|X_{ij} - T_2^0 Y_{ij}\|^2],$$

$$J_{[173]/2} = \max_{ij}[\|X_{ij} - T_{[173]}Y_{ij}\|^2/\|X_{ij} - T_2^0 Y_{ij}\|^2].$$

We note that the transform T_p^0 is optimal in the class of transforms defined in Section 7.6.3. The transforms $T_{[63]}$ [63], $T_{[158]}$ [158] and $T_{[173]}$ [173] are each optimal in different transform classes.

The results of simulations are presented in Tables 4–6. For the same compression ratio, the proposed transforms T_2^0, T_3^0 and the known transforms $T_{[63]}$, $T_{[158]}$, $T_{[173]}$ are compared with respect accuracy and computational work.

The symbol '*cr*' denotes the compression ratio. In the first column of each table, the symbol '–' denotes the case of a pure filtering with no compression. In this case, PCA-KLT rank and the ranks of the operators $F_1^0, \ldots, F_p^0$ of the transform T_p^0 are all equal to m, i.e. $\eta_{[7]} = \eta_1 = \ldots = \eta_p = m$.

Tables 4–6 also represent the cumulative number of flops needed for processing of data Y_{ij} for all $i = 1, \ldots, 256/m$ and $j = 1, \ldots, 256/q$. We denote the number by K with a subscript related to the associated transform. We note that the number of flops, needed for the matrix orthogonalization procedure, has been included in the overall number of flops for T_2^0.

Table 5. Performance comparison of transforms T_2^0 and $T_{[158]}$.

$m = 8, \quad q = 16$						
	Accuracy			Flops		
cr	$J_{[158]}$	J_2^0	$J_{[158]/2}$	$K_{[158]}$	$K_{T_2^0}$	$K_{[158]}/K_{T_2^0}$
–	4.97×10^4	2.01×10^{-25}	7.37×10^{32}	4.67×10^7	3.20×10^7	1.46
1/2	2.19×10^6	3.97×10^3	9.53×10^4	9.75×10^7	5.50×10^7	1.74
1/4	1.92×10^6	3.30×10^4	7.03×10^3	9.70×10^7	5.21×10^7	1.86
$m = 16, \quad q = 32$						
	Accuracy			Flops		
cr	$J_{[158]}$	J_2^0	$J_{[158]/2}$	$K_{[158]}$	$K_{T_2^0}$	$K_{[158]}/K_{T_2^0}$
–	1.75×10^5	5.09×10^{-25}	9.7×10^{29}	8.36×10^7	6.03×10^7	1.39
1/2	1.97×10^5	7.10×10^3	9.81×10^3	1.80×10^8	9.81×10^7	1.83
1/4	2.30×10^5	7.50×10^4	1.73×10^3	1.79×10^8	9.81×10^7	1.83
$m = 32, \quad q = 64$						
	Accuracy			Flops		
cr	$J_{[158]}$	J_2^0	$J_{[158]/2}$	$K_{[158]}$	$K_{T_2^0}$	$K_{[158]}/K_{T_2^0}$
–	1.10×10^5	1.16×10^{-24}	9.1×10^{29}	1.54×10^8	1.18×10^8	1.31
1/2	1.30×10^5	1.75×10^4	2.61×10^2	3.45×10^8	1.88×10^8	1.84
1/4	3.69×10^5	2.00×10^4	1.77×10^2	3.43×10^8	1.87×10^8	1.83

Tables 4–6 are to compare accuracies of transforms T_2^0, $T_{[63]}$ and $T_{[158]}$. It follows from the tables that, for the same compression ratio, the accuracy of the transform T_2^0 is substantially better than that for the transforms in [63] and [158]. Although PCA-KLT requires a smaller number $K_{[63]}$ of flops than that of T_2^0, the accuracy of PCA-KLT cannot be improved for the same compression ratio. At the same time, $K_{T_2^0}$ is less than $K_{[158]}$ (see Table 5).

In Figures 7.3–7.5, we present the results of simulations for the case $m = 16$, $q = 32$and $cr = 1/2$ taken from the Tables 4 and 5. In particular, for a more conspicuous illustration of these results, in Fig. 7.5(b) we represent the case of one-dimensional signals that are typical fragments of plots of the same row in matrices $\{X_{ij}\}$ (solid line) and $\{\mathcal{T}_2^0(Y_{ij})\}$ (dashed line with circles).

In Table 6, we compare performance of the proposed transform T_2^0 with that of the transform $T_{[173]}$. The transform $T_{[173]}$ is based on a Volterra polynomial of the second degree which requires $N = 16$ terms for $X_{ij}, Y_{ij} \in \mathbb{R}^{16\times 32}$. This implies a substantial increase in computational work for $T_{[173]}$ in comparison with the proposed transform T_2^0.

The simulations demonstrate the superiority of the considered transform over known transforms $T_{[63]}$, $T_{[158]}$ and $T_{[173]}$ with respect to associated accuracies and computational work. We note that the accuracy of transforms $T_{[63]}$ and $T_{[158]}$ cannot be improved (for a fixed compression ratio) and the computational work for transform $T_{[173]}$ cannot be diminished (for the same associated accuracy and fixed compression ratio). With this method, it is possible that these measures can be improved using the free parameters in the considered transform. This point has been discussed in Remark 39 in Section 7.6.3.

In the case of $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ chosen from (7.196), the accuracy J_2^0 associated with T_0^2 is considerably worse than that for the original choice for $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$.

Simulations to illustrate transform by Theorem 62

Let the tensor $(X^{(1)}, X^{(2)}, X^{(3)})$ be the numerical representation of the known image "Tree",[8] where $X^{(k)} \in \mathbb{R}^{M\times N}$, $k = 1, 2, 3$, and $M = N = 256$. For each $k = 1, 2, 3$, matrix $X^{(k)}$ has been partitioned into 256 sub-matrices $X_{ij}^{(k)} \in \mathbb{R}^{8\times 32}$ with $i = 1, \ldots, 32$ and $j = 1, \ldots, 8$ so that $X^{(k)} = \{X_{ij}^{(k)}\}$.

[8]The database can be found in http://sipi.usc.edu/services/database/Database.html.

Table 6. Performance comparison of transforms T_2^0 and $T_{[173]}$.

$m = 16, \quad q = 32$						
	Accuracy			Flops		
cr	$J_{[173]}$	J_2^0	$J_{[173]/2}$	$K_{[173]}$	$K_{T_2^0}$	$K_{[173]}/K_{T_2^0}$
–	3.67×10^{-5}	5.09×10^{-25}	4.63×10^{20}	6.17×10^{10}	7.01×10^{7}	880.2
1/2	2.45×10^{4}	7.10×10^{3}	91.7	6.24×10^{10}	9.07×10^{7}	688.0
1/4	1.92×10^{5}	7.50×10^{4}	97.3	6.24×10^{10}	9.07×10^{7}	668.0

Each sub-matrix $X_{ij}^{(k)}$ has been interpreted as a set of 32 realizations of a random signal with columns representing realizations.

Observed data has been simulated in the form

$$Y_{ij}^{(k)} = 10\,\mathcal{R}_{ij}^{(k)} \bullet X_{ij}^{(k)} + 500\,\tilde{\mathcal{R}}_{ij}^{(k)}, \tag{7.220}$$

where $\mathcal{R}_{ij}^{(k)}$ and $\tilde{\mathcal{R}}_{ij}^{(k)}$ are matrices with uniformly distributed entries in the interval $(0, 1)$ and normally distributed entries with mean 0 and variance 1, respectively.

In these simulations, we compare the performances of the transform T_2^0 given by (7.137), (7.138), (7.205), (7.206) with the best transform of the second degree P_2 by [158] and the optimal hybrid Hadamard-quadratic transform T_{pr}^0 given by (7.87) and Theorems 56 and 57 for $p = 9$ and $r = 5$ (Section 7.5). We note that the transform P_2^0 generalizes PCA-KLT and is a particular case of T_{pr}^0 (see Section 7.5.6). The rank has been used equal to 5 for all the transforms.

The transforms have been applied to each pair of sub-matrices $X_{ij}^{(k)}$ and $Y_{ij}^{(k)}$. The related covariance matrices have been estimated by the known simple estimates given in Section 5.3.1. As above, operators $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ for the transform T_2^0 have been defined by $\boldsymbol{\varphi}_1(\mathbf{y}) = \mathbf{y}$ and $\boldsymbol{\varphi}_2(\mathbf{y}) = \mathbf{x}^{(1)}$ with $\mathbf{x}^{(1)}$ determined from the procedure provided in Section 7.6.6.

The results of simulations are presented in Table 7 and Figures 7.6–7.7. In the table,

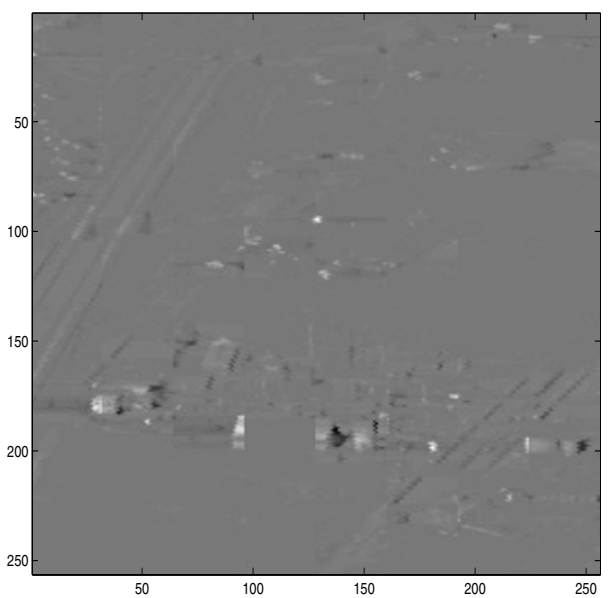
$$\Delta_{P_2^0} = \|X^{(k)} - X_{P_2^0}^{(k)}\|^2, \quad \Delta_{T_{pr}^0} = \|X^{(k)} - X_{T_{pr}^0}^{(k)}\|^2 \text{ and } \Delta_{T_2^0} = \|X^{(k)} - X_{T_2^0}^{(k)}\|^2$$

where $X_{P_2^0}^{(k)}$, $X_{T_{pr}^0}^{(k)}$ and $X_{T_2^0}^{(k)}$ are results of the application of the transform P_2^0, T_{pr}^0 and T_2^0, respectively. The matrices $X_{[152]}^{(k)}$, $X_{[152],9}^{(k)}$, $X_{\check{P}_2}^{(k)}$, $X_{\check{P}_2,9}^{(k)}$, $X_{\hat{P}_2}^{(k)}$ and $X_{\hat{P}_2,9}^{(k)}$ have been composed from the corresponding 8×64 sub-matrices similarly to the matrices from the preceding Section.

Table 7: Errors associated with transforms P_2^0, T_{pr}^0 and T_2^0 of the same rank $r = 5$.

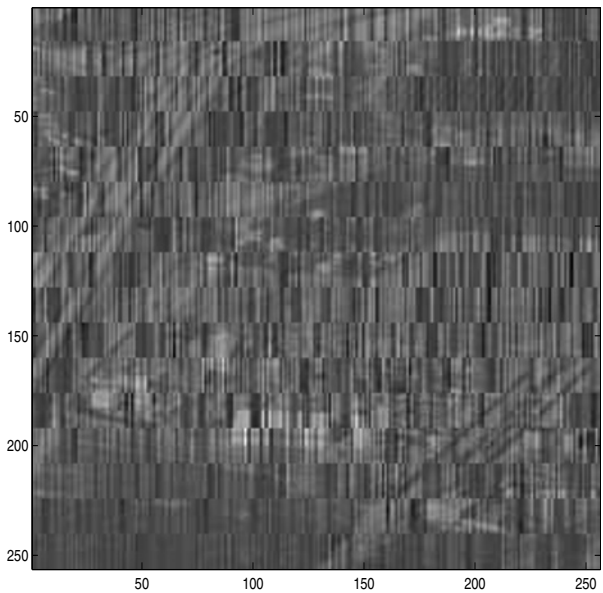
	$\Delta_{P_2^0}$	$\Delta_{T_{pr}^0}$	$\Delta_{T_2^0}$
X_1	2.2×10^8	3.4×10^7	1.1×10^6
X_2	2.2×10^8	3.0×10^7	1.3×10^6
X_3	2.7×10^8	3.8×10^7	1.2×10^6

(a) Given reference signals $\{X_{ij}\}$. This digitized image has been taken from http://sipi.usc.edu/database/.

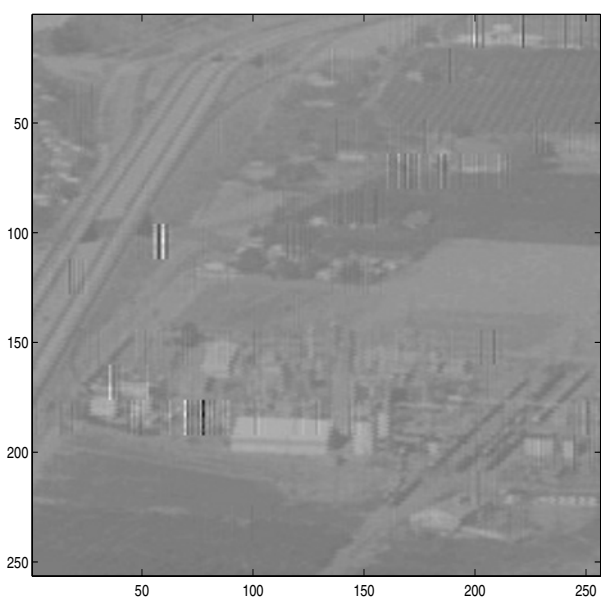


(b) Observed data $\{Y_{ij}\}$.

Figure 7.3: Illustration of simulation results from Tables 4 and 5.



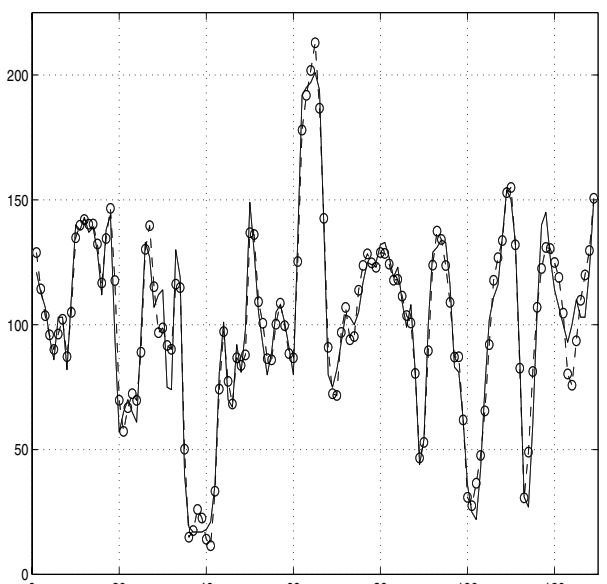
(a) Estimates of $\{X_{ij}\}$ by [63].



(b) Estimates of $\{X_{ij}\}$ by [158].

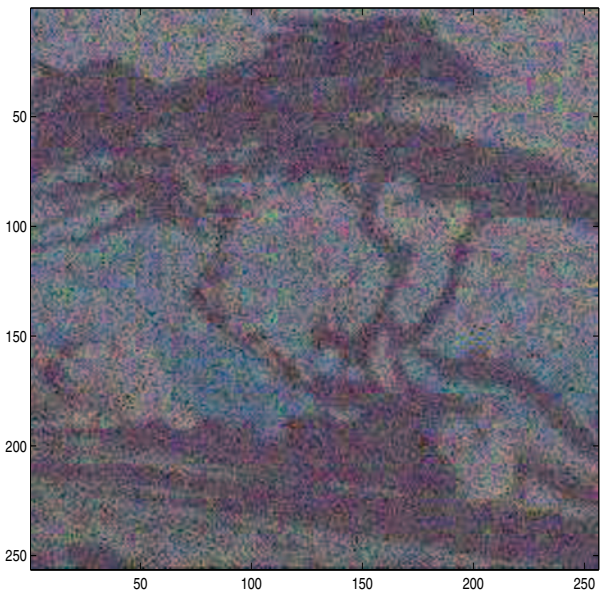
Figure 7.4: Illustration of simulation results from Tables 4 and 5.

(a) Estimates of $\{X_{ij}\}$ by $\mathcal{T}_2^0$.



(b) Plots of related rows in $\{X_{ij}\}$ and $\mathcal{T}_2^0(Y_{ij})$.

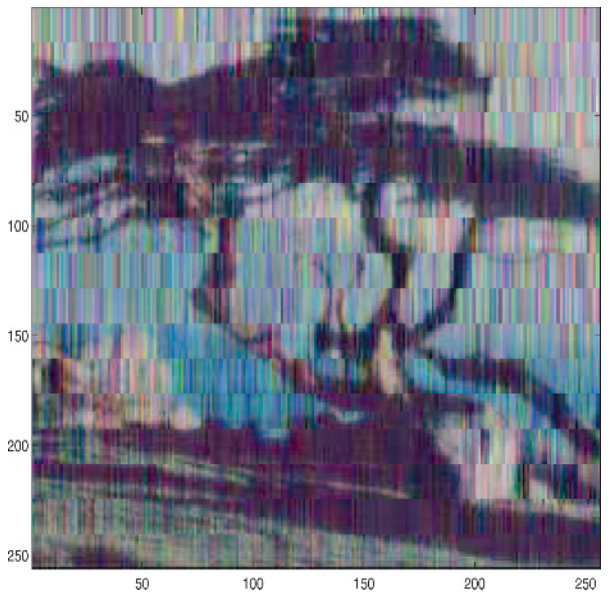
Figure 7.5: Illustration of simulation results from Tables 4 and 5.



(a) Observed data ($Y^{(1)}$, $Y^{(2)}$, $Y^{(3)}$).

(b) Reference signals ($X^{(1)}$, $X^{(2)}$, $X^{(3)}$). This digitized image has been taken from http://sipi.usc.edu/database/.

Figure 7.6: Examples of performance of transforms P_2^0 and T_2^0.



(a) Estimates $(X^{(1)}_{P_2^0}, X^{(2)}_{P_2^0}, X^{(3)}_{P_2^0})$ by transform P_2^0.

(b) Estimates $(X^{(1)}_{T_2^0}, X^{(2)}_{T_2^0}, X^{(3)}_{T_2^0})$ by transform T_2^0.

Figure 7.7: Examples of performance of transforms P_2^0 and T_2^0.

7.6.10 Discussion

Some distinctive features of the above techniques are as follows.

Remark 45. *It follows from Theorems* 60 *and* 61*, and Corollaries* 19 *and* 20 *that the accuracy associated with the considered transform improves when p increases.*

Remark 46. *Unlike the approaches based on Volterra polynomials* [159, 167, 173, 182] *the considered method does not require computation of pseudo-inverses for large $N \times N$ matrices. Instead, the proposed transforms use pseudo-inverses of $n \times n$ matrix $\mathbb{E}_{v_k v_k}$. See Theorems* 60 *and* 61*. This leads to a substantial reduction in computational work.*

In summary, the approach considered in Section 7.6 is based on a representation of a transform in the form of the sum of p reduced-rank transforms. Each particular transform is formed by the linear reduced-rank operator $\mathcal{F}_k$, and by operators $\boldsymbol{\varphi}_k$ and $\mathcal{Q}_k$ with $k = 1 \ldots, p$. Such a device allows us to improve the numerical characteristics (accuracy, compression ration and computational work) of the known transforms based on the Volterra polynomial structure [159, 167, 173, 182]. These objectives are achieved due to the special "intermediate" operators $\boldsymbol{\varphi}_1$, ..., $\boldsymbol{\varphi}_p$ and $\mathcal{Q}_1$, ..., $\mathcal{Q}_p$. Such operators reduce the determination of optimal linear reduced-rank operators $\mathcal{F}_1^0$, ..., $\mathcal{F}_p^0$ to the computation of a sequence of relatively small matrices (Theorems 60 and 61).

The explicit representations of the accuracy associated with the proposed transforms have been rigorously justified.

It has been shown that the proposed approach generalizes the Fourier series in Hilbert space, the Wiener filter, the Karhunen-Loève transform and the transforms [159, 167, 173].

7.7 Optimal Generalized Hybrid Transform

In this section, we consider an extension of the methods of Sections 7.5 and 7.6 to a more general case. An idea is to apply the transform given by Theorems 60 and 61 to each iteration of the recurrent procedure of the hybrid method of Section 7.5. The technique summarized in Theorems 60 and 61 is more general than the Hadamard-quadratic approximation used in Section 7.5 and, therefore, has more degrees of freedom to improve the performance of the transform resulted in such a device.

7.7.1 Method description

Let $\mathbf{x}$ and $\mathbf{y}$ be as in Section 7.5, i.e. $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$ and $\mathbf{y} \in L^2(\Omega, \mathbb{R}^n)$. The proposed transform consists of two stages which we call the recurrent unconstrained error minimization and the constrained error minimization.

Stage one: recurrent unconstrained error minimization. Let $\mathbf{x}_0 = \mathbf{y}$ and let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q \in L^2(\Omega, \mathbb{R}^m)$ be defined by

$$\mathbf{x}_{j+1} = \tilde{\mathcal{T}}_j(\mathbf{x}_j), \tag{7.221}$$

where for $j = 0, 1, \ldots q-1$,

$$\begin{aligned} \tilde{\mathcal{T}}_j(\mathbf{x}_j) &= \tilde{f}_j + \sum_{k=1}^{p} \tilde{\mathcal{F}}_{kj} \mathcal{Q}_{kj} \boldsymbol{\varphi}_{kj}(\mathbf{x}_j) \\ &= \tilde{f}_j + \sum_{k=1}^{p} \tilde{\mathcal{F}}_{kj}(\mathbf{v}_{kj}) \end{aligned} \tag{7.222}$$

with

$$\tilde{\mathcal{T}}_j : L^2(\Omega, \mathbb{R}^\nu) \to L^2(\Omega, \mathbb{R}^m), \quad \nu = \begin{cases} n & \text{if} \quad j = 0, \\ m & \text{if} \quad j = 1, , \ldots, q \end{cases},$$

$$\tilde{f}_j \in \mathbb{R}^m, \quad \boldsymbol{\varphi}_{kj} : L^2(\Omega, \mathbb{R}^\nu) \to L^2(\Omega, \mathbb{R}^\nu), \quad \mathcal{Q}_{kj} : L^2(\Omega, \mathbb{R}^\nu) \to L^2(\Omega, \mathbb{R}^n \nu),$$

$$\tilde{\mathcal{F}}_{kj} : L^2(\Omega, \mathbb{R}^\nu) \to L^2(\Omega, \mathbb{R}^m) \quad \text{and} \quad \mathbf{v}_{kj} = \mathcal{Q}_{kj}[\boldsymbol{\varphi}_{kj}(\mathbf{x}_j)].$$

Here, for each $j = 0, 1, \ldots, q-1$, operators $\boldsymbol{\varphi}_{kj}$ and $\mathcal{Q}_{kj}$ are chosen similarly to those in Section 7.6, and the vector $\tilde{f}_j$ and operators $\tilde{\mathcal{F}}_{kj}$ are determined from the solution of the unconstrained minimization problem

$$J(\tilde{f}_j, \tilde{\mathcal{F}}_{1j}, \ldots \tilde{\mathcal{F}}_{pj}) = \min_{f_j, \mathcal{F}_{1j}, \ldots, \mathcal{F}_{pj}} J(f_j, \mathcal{F}_{1j}, \ldots, \mathcal{F}_{pj}), \tag{7.223}$$

where

$$J(f_j, \mathcal{F}_{1j}, \ldots \mathcal{F}_{pj}) = E[\|\mathbf{x} - \mathcal{T}_j(\mathbf{x}_j)\|^2]. \tag{7.224}$$

The functional $J(f_j, \mathcal{F}_{1j}, \ldots \mathcal{F}_{pj})$ represents the error associated with $\mathbf{x}$ estimation by $\mathcal{T}_j$, therefore, the solution to the problem (7.223) is called the recurrent unconstrained error minimization.

Stage two: constrained error minimization. Next, for $j = q$, the vector f_q^0 and the operators $\mathcal{F}_{kq}^0$ are determined from the solution of the rank-constrained minimization problem

$$J(f_q^0, \mathcal{F}_{1q}^0, \ldots \mathcal{F}_{pq}^0) = \min_{\tilde{f}_q, \tilde{\mathcal{F}}_{1q}, \ldots, \tilde{\mathcal{F}}_{pq}} J(\tilde{f}_q, \tilde{\mathcal{F}}_{1q}, \ldots, \tilde{\mathcal{F}}_{pq}) \tag{7.225}$$

subject to

$$\text{rank } \tilde{\mathcal{F}}_{1q} \leq \eta_1, \quad \ldots, \quad \text{rank } \tilde{\mathcal{F}}_{pq} \leq \eta_p, \tag{7.226}$$

where

$$\eta_1 + \ldots + \eta_p \leq \eta \leq \min\{m, n\}.$$

We write

$$\mathcal{T}_q^0(\mathbf{x}_q) = f_q^0 + \sum_{k=1}^{p} \mathcal{F}_{kq}^0(\mathbf{v}_{kq}). \tag{7.227}$$

7.7.2 Determination of the transform

Theorem 63. *Let* $\mathbf{v}_{1j}, \ldots, \mathbf{v}_{pj}$ *be determined by Lemma* 34 *of Section* 5.7.4 *for each* $j = 0, 1, \ldots, q$. *Then the vector* $\tilde{f}_j$ *and operators* $\tilde{\mathcal{F}}_{1j}, \ldots, \tilde{\mathcal{F}}_{pj}$, *satisfying the unconstrained minimum* (7.223), *are determined by*

$$\tilde{f}_j = E[\mathbf{x}_j] - \sum_{k=1}^{p} \bar{F}_{kj} E[\mathbf{v}_{kj}] \tag{7.228}$$

and

$$\tilde{F}_{1j} = \mathbb{E}_{x_j v_{1j}} \mathbb{E}^{\dagger}_{v_{1j} v_{1j}} + M_1[I - \mathbb{E}_{v_{1j} v_{1j}} \mathbb{E}^{\dagger}_{v_{1j} v_{1j}}], \tag{7.229}$$

$$\vdots$$

$$\tilde{F}_{pj} = \mathbb{E}_{x_j v_{pj}} \mathbb{E}^{\dagger}_{v_{pj} v_{pj}} + M_p[I - \mathbb{E}_{v_{pj} v_{pj}} \mathbb{E}^{\dagger}_{v_{pj} v_{pj}}], \tag{7.230}$$

where $M_j \in \mathbb{R}^{m \times n}$ *is an arbitrary matrix for each* $j = 1, \ldots, p$.

The associated accuracy for transform $\tilde{\mathcal{T}}_j$, *defined by* (7.222) *is given by*[9]

$$E[\|\mathbf{x} - \tilde{\mathcal{T}}_j(\mathbf{x}_j)\|^2] = \|\mathbb{E}_{xx}^{1/2}\|^2 - \sum_{s=0}^{j} \sum_{k=1}^{p} \|\mathbb{E}_{x v_{ks}} (\mathbb{E}^{1/2}_{v_{ks} v_{ks}})^{\dagger}\|^2. \tag{7.231}$$

Proof. The proof of relationships (7.228)–(7.230) follows from the proofs of Theorem 61 and Corollary 20. The proof of the error representation (7.231) follows from the proofs of Theorem 43 (Section 5.7.7), Theorem 61 and Corollary 20. □

[9] *In particular,* $\mathcal{M}_k$ *can be chosen as the zero operator.*

We need more notation as follows.

Let the SVD of the matrix $\mathbb{E}_{x_q v_{kq}}(\mathbb{E}^{1/2}_{v_{kq}v_{kq}})^\dagger$ be given by

$$U_{kq}\Sigma_{kq}V_{kq}^T = \mathbb{E}_{x_q v_{kq}}(\mathbb{E}^{1/2}_{v_{kq}v_{kq}})^\dagger, \tag{7.232}$$

where $U_{kq} \in \mathbb{R}^{m\times n}$, $V_{kq} \in \mathbb{R}^{n\times n}$ are orthogonal and $\Sigma_{kq} \in \mathbb{R}^{n\times n}$ is diagonal,

$$U_{kq} = [s_{k1}, \ldots, s_{kn}], \quad V_{kq} = [d_{k1}, \ldots, d_{kn}], \tag{7.233}$$

$$\Sigma_{kq} = \mathrm{diag}(\beta_{k1}, \ldots, \beta_{kn}) \tag{7.234}$$

with $\beta_{k1} \geq \cdots \geq \beta_{kr} > 0$, $\beta_{k,r+1} = \beta_{kn} = 0$, $r = 1, \ldots, n$ and $r = r(k)$.

Let us set

$$U_{k\eta_k} = [s_{k1}, \ldots, s_{k\eta_k}], \quad V_{k\eta_k} = [d_{k1}, \ldots, d_{k\eta_k}] \tag{7.235}$$

$$\Sigma_{k\eta_k} = \mathrm{diag}(\beta_{k1}, \ldots, \beta_{k\eta_k}), \tag{7.236}$$

where $U_{k\eta_k} \in \mathbb{R}^{m\times\eta_k}$, $V_{k\eta_k} \in \mathbb{R}^{n\times\eta_k}$ and $\Sigma_{k\eta_k} \in \mathbb{R}^{\eta_k\times\eta_k}$. Now we define $G_{k\eta_k} \in \mathbb{R}^{m\times n}$ and $\mathcal{G}_{k\eta_k} : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^m)$ by

$$G_{k\eta_k} = U_{k\eta_k}\Sigma_{k\eta_k}V_{k\eta_k}^T \quad \text{and} \quad [\mathcal{G}_{k\eta_k}(\mathbf{w}_{kq})](\omega) = G_{k\eta_k}[\mathbf{w}_{kq}(\omega)], \tag{7.237}$$

respectively, for any $\mathbf{w}_{kq} \in L^2(\Omega, \mathbb{R}^n)$.

The operators $\mathbb{E}^\dagger_{v_{kq}v_{kq}}, (\mathcal{E}^{1/2}_{v_{kq}v_{kq}})^\dagger : L^2(\Omega, \mathbb{R}^\nu) \to L^2(\Omega, \mathbb{R}^m)$ are defined similarly to the operators $\mathcal{E}^\dagger_{v_k v_k}, (\mathcal{E}^{1/2}_{v_k v_k})^\dagger$ given in Section 7.6.4.

We write $A_k \in \mathbb{R}^{m\times n}$ for an arbitrary matrix, and define operator $\mathcal{M}_k : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^m)$ by $[\mathcal{M}_k(\mathbf{w})](\omega) = A_k\mathbf{w}(\omega)$ for any $\mathbf{w} \in L^2(\Omega, \mathbb{R}^n)$.

Theorem 64. *Let $\mathbf{v}_{1q}, \ldots, \mathbf{v}_{pq}$ be orthogonal vectors determined by Lemma 34 of Section 5.7.4. Then f^0_q and $\mathcal{F}^0_{1q}, \ldots, \mathcal{F}^0_{pq}$, satisfying (7.223)–(7.226), are determined by*

$$f^0_q = E[\mathbf{x}_q] - \sum_{k=1}^{p} F^0_{kq}E[\mathbf{v}_{kq}] \tag{7.238}$$

and

$$F^0_{1q} = G_{1\eta_1}(\mathbb{E}^{1/2}_{v_{1q}v_{1q}})^\dagger + M_1[I - \mathbb{E}^{1/2}_{v_{1q}v_{1q}}(\mathbb{E}^{1/2}_{v_{1q}v_{1q}})^\dagger], \tag{7.239}$$

$$\vdots$$

$$F^0_{pq} = G_{p\eta_p}(\mathbb{E}^{1/2}_{v_{pq}v_{pq}})^\dagger + M_p[I - \mathbb{E}^{1/2}_{v_{pq}v_{pq}}(\mathbb{E}^{1/2}_{v_{pq}v_{pq}})^\dagger] \tag{7.240}$$

where, for each $j = 1, \ldots, p$, $M_j \in \mathbb{R}^{m\times n}$ is an arbitrary matrix defined similarly to M_j in Theorem 61.

The accuracy associated with transform $\mathcal{T}^0_{pq}$ *given by* (7.227) *and* (7.238)–(7.240) *is such that*

$$E[\|\mathbf{x} - \mathcal{T}^0_{pq}(\mathbf{y})\|^2] = \|\mathbb{E}^{1/2}_{xx}\|^2 - \sum_{k=1}^{p}\sum_{i=1}^{\eta_k}\beta^2_{ki} - \sum_{s=0}^{q-1}\sum_{k=1}^{p}\|\mathbb{E}_{xv_{ks}}(\mathbb{E}^{1/2}_{v_{ks}v_{ks}})^{\dagger}\|^2. \quad (7.241)$$

Proof. The proof follows from the proofs of Theorem 43 (Section 5.7.7) and Theorem 61. □

7.7.3 Discussion

Optimal hybrid filtering.

The transform $\tilde{\mathcal{T}}_j$ given by (7.222) and Theorem 63 represents a model of the filter and, therefore, is important in its own right. The transform $\tilde{\mathcal{T}}_j$ can be considered independently from the transform $\mathcal{T}^0_{pq}$. It follows from the representation of the error (7.231) associated with the filter $\tilde{\mathcal{T}}_j$ that its accuracy is improved with an increase in the number of iterations j and the number of terms p. Other advantages and degrees of freedom are associated with the transform given by Theorems 60 and 61 of Section 7.6. This transform is used at each iteration of the recurrent procedure (7.221).

Features associated with the transform of Section 7.6.

Both transform $\tilde{\mathcal{T}}_{pj}$ and transform $\mathcal{T}^0_{pq}$ possesses features inherited from the transform of Section 7.6. In particular, the computational load of both transforms is lessened in comparison with the transforms based on the Volterra polynomial structure (see Remark 40 in Section 7.6).

The compression procedure of the transform $\mathcal{T}^0_{pq}$ is similar to that considered in Section 6.5.8. The compression ratio is given by

$$c^0 = (\eta_1 + \ldots + \eta_p)/m \quad (7.242)$$

and it can be varied according to variations in each η_k for $k = 1, \ldots, p$.

7.8 Concluding Remarks

In this chapter, we have presented the computational methods for data processing of high dimensionality. In Sections 7.2 and 7.4, we have given

a rigorous justification and generalization for the Principal Components Analysis also known as the Karhunen-Loève Transform (PCA–KLT). In Section 7.3, wide generalizations of the Eckart-Young low-rank approximation theorem have been presented. In Sections 7.5–7.7, three computational methods have been provided which are diverse generalizations of PCA–KLT. These methods are united by the idea of increasing the degrees of freedom compared to PCA–KLT. While PCA–KLT has only one degree of freedom (its rank), the methods given in Sections 7.5–7.7 have extra degrees of freedom associated with their specific structures. As a result, the extra degrees of freedom are the number of iterations for the methods in Sections 7.5 and 7.7, and the choice of nonlinear operators which comprise the methods of Sections 7.6 and 7.7. Variations of the degrees of freedom allow us to improve the performance of the methods for data processing presented in this chapter.

Bibliography

[1] T. Anderson, *An Introduction to Multivariate Statistical Analysis,* New York, Wiley, 1984.

[2] Z. D. Bai and Y. Q. Lim, Limit of the smallest eigenvalue of a large dimensional sample covariance matrix, *Ann. Probab.* 21, 1275–1294, 1993.

[3] A. V. Balakrishnan, *Applied Functional Analysis,* 2nd ed., Springer-Verlag, New York, 1981.

[4] A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. Info. Th.*, Vol. 39, 3, pp. 930-945, 1993.

[5] D. M. Bates and D. G. Watts, *Nonlinear Regression Analysis and its Applications,* Wiley, New York, 1988.

[6] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications,* John Wiley & Sons, New York, 1974.

[7] S. P. Berge and T. I. Fossen, On the Properties of the Nonlinear Ship Equations of Motion. *Mathematical and Computer Modelling of Dynamical Systems.* 6(4), pp. 365 –381, 2000.

[8] A. Bertuzzi, A. Gandolfi and A. Germani, Causal Polynomial Approximation for Input-Output Maps on Hilbert Spaces, *Math. Systems Theory,* 14, pp. 339-352, 1981.

[9] H. W. Bode and C. E. Shannon, A Simplified Derivation of Linear Least Square Smoothing and Prediction Theory, *Proc. IRE,* 38, pp. 417–425, 1950.

[10] T. L. Boullion and P. L. Odell, *Generized Inverse Matrices,* John Willey & Sons, Inc., New York, 1972.

[11] E. Brinksma, J-P. Katoen, D. Latella and R. Langerak, A stochastic causality-based process algebra, *The Computer Journal,* 7, pp. 552 - 565, 1995.

[12] V. I. Bruno, An approximate Weierstrass theorem in topological vector space, *J. Approximation Theory*, 42, pp. 1-3, 1984.

[13] J. A. Calvin and R. L. Dykstra, Least Squares Estimation of Covariance Matrices in Balanced Multivariate Variance Components Models, *J. Amer. Stat. Assoc.* 86, no 414, 388–395, 1991.

[14] C. J. Champion, Empirical Bayesian estimation of normal variances and covariances, *J. Multivariate Analysis*, 87, pp 60–79, 2003.

[15] S. Chen and S. A. Billings, Representations of non-linear systems: the NARMAX model, *Int. J. Control*, 49, pp. 1013–1032, 1989.

[16] S. Chen, S. A. Billings and W. Luo, Orthogonal least squares methods and their application to non-linear system identification, *Int. J. Control*, 50, pp. 1873–1896, 1989.

[17] T. Chen and H. Chen, Universal Approximation to Nonlinear Operators by Neural Networks with Arbitrary Activation Functions and Its Application to Dynamical Systems, *IEEE Trans. on Neural Networks*, vol. 6, no. 4, pp. 911–917, 1995.

[18] W. Chen, U. Mitra and P. Schniter, On the equivalence of three reduced rank linear estimators with applications to DS-CDMA, *IEEE Trans. on Information Theory,* 48, no. 9, pp. 2609–2614, 2002.

[19] J. S. Chipman, "Proofs" and proofs of the Eckart-Young theorem, (with Appendix by Heinz Neudecker), pp.71-83, in *Stochastic Processes and Functional Analysis. In Celebration of M.M. Rao's 65th birthday* (J. A. Goldstein, N. E. Gretsky, J. J. Uhl Jr. eds.), Marcel Dekker, 1997.

[20] M. Cotlar and R. Cignoli, *An Introduction to Functional Analysis*, pp. 114–116, North-Holland Publishing Company, Amsterdam, London, 1974.

[21] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods,* Cambridge, Cambridge University Press, 2000.

[22] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Math. Control Signals Systems*, 2, pp. 303-314, 1989.

[23] I. K. Daugavet and A. A. Lanne, On polynomial approximation of non-linear operators in the space $\mathcal{C}$, *Siberian Mathematical Journal*, 26, pp. 44-48, 1985.

[24] I. K. Daugavet, On linear splittings of signals (Russian), *Methods of Optimiz. and their Applic.*, Irkutsk, pp. 175–188, 1982.

[25] I. K. Daugavet, On a class of polinomial operators (Russian), *Methods of Calculations* 15, pp. 113–119, 1988.

[26] I. K. Daugavet, On polynomial approximation of operators (Russian), *Vestnik St Petersburg Univ., Mathematics* 27 (1994), 23–26.

[27] W. B. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, McGraw–Hill, New York, 1958.

[28] J.-P. Delmas, On eigenvalue decomposition estimators of centro-symmetric covariance matrices, *Signal Processing,* 78, pp. 101–116, 1999.

[29] F. Deutsch, *Best Approximation in Inner Product Spaces,* Springer-Verlag, New York, 2001.

[30] N. R. Draper and H. Smith, *Applied Regression Analysis,* Wiley-Interscience, 1998.

[31] N. Dunford and J. T. Schwartz, *Linear Operators, Part 1, General Theory*, Wiley Classics Library, Wiley, New York, 1988.

[32] N. Dunford and J. T. Schwartz, *Linear Operators, Part 2*, Wiley Classics Library, Wiley, New York, 1988.

[33] P. D'Urso and T. Gastaldi, An "orderwise" polynomial regression procedure for fuzzy data, *Fuzzy Sets and Syst.,* 130, pp. 1–19, 2002.

[34] C. Eckart and G. Young, The Approximation of One Matrix by Another of Lower Rank, *Psychometrika,* 1, pp. 211-218, 1936.

[35] M. Eerola, *Probabilistic causality in Longitudinal Studies,* Lecture Notes in Statistics, Springer-Verlag, New-York, 1994.

[36] P. L. Falb and M. I. Freedman, A generalised transform theory for causal operators, *SIAM J. Control*, 7, pp. 452–471. MR 58#33192a, 1969.

[37] V. N. Fomin and M. V. Ruzhansky, Abstract optimal linear filtering, *SIAM J. Control Optim.,* 38, pp. 1334–1352, 2000.

[38] L. Fox and I. B. Parker, *Chebyshev Polynomials in Numerical Analysis*, Oxford Mathematical Handbooks, Oxford University Press, 1979.

[39] Y. Foures and I.E. Segal, *Causality and analyticity*, Transac. Amer. Math. Soc. 78, pp. 385–405. MR 16#1032d, 1985.

[40] L. E. Franks, *Signal Theory*, Prentice–Hall, Englewood Cliffs, N. J., 1966.

[41] S. Friedland, *Personal communication*, December 2005.

[42] S. Friedland and A. P. Torokhti, Generalized rank-constrained matrix approximations, *SIAM J. Matrix Anal. Appl.* (accepted).

[43] A. Frieze, R. Kannan, and S. Vempala, Fast Monte-Carlo Algorithms for Finding Low-Rank Approximations, *Journal of the ACM,* 51, No. 6, 2004.

[44] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* Academic Press, Boston, 1990.

[45] P. G. Gallman and K. S. Narendra, Representations of non–linear systems via the Stone–Weierstrass theorem, *Automatica* 12, pp. 619–622, 1976.

[46] V. Gimeno, Obtaining the EEG envelope in real time: a practical method based on homomorphic filtering, *Neuropsychobiol.,* 18, pp. 110–112, 1987.

[47] I. Z. Gohberg and M. G. Kreĭn, *Theory and Application of Volterra Operators in Hilbert Space*, Transl. of Math. Momographs, Amer. Math. Soc. 24, 1970. MR 41#9041.

[48] J. S. Goldstein, I. Reed, and L. L. Scharf, A Multistage Representation of the Wiener Filter Based on Orthogonal Projections, *IEEE Trans. on Information Theory,* vol. 44, pp. 2943-2959, 1998.

[49] M. Golomb and H. Weinbeerger, Optimal approximation and error bounds, in *On Numerical Approximation,* Univ. of Wisc, pp. 117–190, 1985.

[50] G. H. Golub and C. F. Van Loan, *Matrix Computations,* Baltimore, MD: Johns Hopkins University Press, 1996.

[51] P. R. Halmos, *Measure Theory*, 12th printing, University Series in Higher Mathematics, Van Nostrand, Princeton, 1968.

[52] S. Haykin, *Adaptive Filter Theory*, Prentice–Hall, Englewood Cliffs, N. J., 1991.

[53] M. L. Honig and J. S. Goldstein, Adaptive reduced-rank interference suppression based on multistage Wiener filter, *IEEE Trans. on Communications,* vol. 50, no. 6, pp. 986–994, 2002.

[54] M. L. Honig and W. Xiao, Performance of reduced-rank linear interferrence suppression, *IEEE Trans. on Information Theory,* vol. 47, no. 5, pp. 1928–1946, 2001.

[55] H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components, *J. Educ. Psychol.,* 24, pp. 417-441 and 498-520, 1933.

[56] P. G. Howlett and A. P. Torokhti, A methodology for the constructive approximation of nonlinear operators defined on noncompact sets, *Numer. Funct. Anal. and Optimiz.* 18, pp. 343–365, 1997.

[57] P. G. Howlett and and A. P. Torokhti, Weak interpolation and approximation of non–linear operators on the space $\mathcal{C}([0,1])$, *Numer. Funct. Anal. and Optimiz.* 19, 1025–1043, 1998.

[58] P. G. Howlett, A. P. Torokhti and C. E. M. Pearce, The modelling and numerical simulation of causal non-linear systems, *Nonlinear Analysis: Theory, Methods & Applications,* vol. 47 (8), pp. 5559-5572, 2001.

[59] P. G. Howlett, A. P. Torokhti, & Pearce, C. E. M.: A Philosophy for the Modelling of Realistic Non-linear Systems, *Proc. of Amer. Math. Soc.,* 131, 2, pp. 353–363, 2003.

[60] P. Howlett, C. Pearce and A. Torokhti, On Nonlinear Operator Approximation with Preassigned Accuracy, *J. Comput. Anal. Appl.*, 5, no. 3, pp. 273–297, 2003.

[61] P. Howlett, C. Pearce and A. Torokhti, An optimal linear filter for random signals with realisations in Hilbert Space, *ANZIAM Journal*, 44, pp. 485–500, 2003.

[62] P. Howlett, C. Pearce and A. Torokhti, On Problems of Best Approximation for Nonlinear Operators, *Nonlinear Functional Analysis and Applications*, 6, pp. 351–368, 2001.

[63] Y. Hua and W. Q. Liu, Generalized Karhunen–Loève transform, *IEEE Signal Proc. Lett.* 5 (6), 141–142, 1998.

[64] Y. Hua, M. Nikpour, and P. Stoica, Optimal Reduced-Rank estimation and filtering, *IEEE Trans. on Signal Processing,* vol. 49, pp. 457-469, 2001.

[65] V. I. Istrăţescu, A Weierstrass theorem for real Banach spaces, *J. Approximation Theory,* 19, pp. 118-122, 1977.

[66] W. James and C. Stein, Estimation with quadratic loss, *Proc. the Fourth Berkeley Symposium on Mathematics abd Statistical Probability,* 1, University of California Press, Berkeley, pp. 361-379, 1961.

[67] M. Jansson and P. Stoica, Forward-only and forward-backward sample covariances - A comparative study, *Signal Processing,* vol. 7, pp. 235 - 245, 1999.

[68] I.T. Jolliffe, *Principal Component Analysis,* Springer Verlag, New York, 1986.

[69] R. Jones, Causality and determinism in physics, *Am. J. Phys.,* 64, 208 - 215, 1996.

[70] D. Kazakos, Optimal constrained representation and filtering of signal, *Signal Proc.* 5, 347–353, 1983.

[71] K. Karhunen, Über Lineare Methoden in der Wahrscheinlichkeitsrechnung, *Ann. Acad. Sci. Fennicae,* Ser. A137, 1947.

[72] G. Kauermann and R. J. Carroll, A Note on the Efficiency of Sandwich Covariance Matrix Estimation, *Journal of the American Statistical Association*, 96, pp 1387-1396, 2001.

[73] J. Kelley, *General Topology*, Van Nostrand, Princeton, N. J., 1957.

[74] V. V. Khlobystov, On Extremal Problems on a Set of Interpolation Polynomials in a Hilbert Space, *J. Mathematical Sciences,* 104, Number 6, pp. 1672 - 1676, 2001.

[75] A. Kneip and K. J. Utikal, Inference for density families using functional principal component analysis, *Journal of the American Statistical Association*, vol. 96, pp. 519 – 542, 2001.

[76] T. G. Kolda, Orthogonal tensor decompositions, *SIAM J. Matrix Analysis and Applications,* 23, no. 1, pp. 243 – 255, 2001.

[77] A. N. Kolmogorov, Interpolation and extrapolation of the stationary random sequences, *Izv. AN USSR Math.*, 5, pp. 3–14, 1941.

[78] M. A. Kowalski, K. A. Sikorski, and F. Stenger, *Selected Topics in Approximation and Computations*, New York–Oxford, Oxford University Press, 1995.

[79] S. Kraut, R. H. Anderson, and J. L. Krolik, A generalized Karhunen-Loeve basis for efficient estimation of tropospheric refractivity using radar clutter, *IEEE Trans. on Signal Processing,* vol. 52, pp. 48–60, 2004.

[80] T. Kubokawa and M.S. Srivastava, Estimating the covariance matrix: a new approach, *J. Multivariate Analysis*, 86, pp. 28–47, 2003.

[81] L. De Lathauwer, B. De Moor and J. Vandewalle, A multilinear singular value decomposition, *SIAM J. Matrix Anal. Appl.,* 21, no. 4, pp. 1253-1278, 2000.

[82] L. De Lathauwer, B. De Moor B and J. Vandewalle, On the Best rank-1 and Rank-$(R_1, R_2, ..., R_N)$ Approximation and Applications of Higher-Order Tensors, *SIAM J. Matrix Anal. Appl.,* 21, no. 4, pp. 1324-1342, 2000.

[83] O. Ledoit and M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, *J. Multivariate Analysis* 88, pp. 365–411, 2004.

[84] E.I. Lehmann, *Testing Statistical Hypotheses,* John Wiley, New York, 1986.

[85] P. L. Leung and F. Y. Ng, Improved estimation of a covariance matrix in an elliptically contoured matrix distribution, *J. Multivariate Analysis*, vol. 88, Issue 1, pp. 131–137, 2004.

[86] P. Lévy, *Problèmes concrets d'analyse fonctionnelle*, Villars, Paris, 2nd edn, 1951.

[87] Q. Lin and J. P. Allebach, Combating speckle in SAR images: vector filtering and sequential classification based on a multiplicative noise model, *IEEE Trans. Geoscience and Remote Sensing*, 28, pp. 647–653, 1990.

[88] M. Loève, Fonctions aléatoires de second ordre, In *P. Lévy, Processus Stochastiques et Mouvement Brownien,* Hermann, Paris, 1948.

[89] D. G. Luenberger, *Optomization by vector space methods*, John Wiley, New York, 1969.

[90] W.-S. Lu, S.-C. Pei, and P.-H. Wang, Weighted Low-Rank Approximation of General Complex Matrices and Its Application in the Design of 2-D Digital Filters, *IEEE Trans. on Circuits and Systems-I: Fundamental Theory & Appl.,* 44, No. 7, pp. 650-655, 1997.

[91] V. L. Makarov and V.V. Khlobystov, *Foundations for theory of polynomial operator interpolation*, Naukova Dumka, Kiev, 1999 (In Russian).

[92] V. J. Mathews and G. L. Sicuranza, *Polynomial Signal Processing,* J. Wiley & Sons, 2001.

[93] A. W. Moore, J. Schneider and K. Deng, Efficient locally weighted polynomial regression predictions, *Proc. 14th Intern. Conf. on Machine Learning,* 236–244, 1997.

[94] A. S. Nemirovsky and S. M. Semenov, On functional approximation in Hilbert spaces (Russian), *Mat. Sb.* 92, 257–282, 1973.

[95] R. D. Nowak and B. D. Van Veen, Tensor product basis approximation for Volterra filters, *IEEE Trans. Signal Proc.,* 44, pp. 36–50, 1996.

[96] F. A. Ocaña, A. M. Aguilera, and M. J. Valderrama, Functional Principal Componenets Analysis by Choice of Norm, *J. Multivariate Anal.* 71, pp. 262 -276, 1999.

[97] R. E. A. C. Paley and N. Wiener, Fourier transform in the complex domain, *Amer. Math. Soc. Colloq. Publ.,* 19, 1934.

[98] J. Park and I. W. Sandberg, Universal approximation using radial–basis–function networks, *Neural Computation* 3, pp. 246–257, 1991.

[99] J. Park and I. W. Sandberg, Criteria for the approximation of nonlinear systems, *IEEE Trans. Circuits Syst. Part 1, Fund. Theory and Applic.* 39, pp. 673–676, 1992.

[100] K. Pearson, On lines and planes of closest fit to systems of points in space. *Phil. Mag.,* 6 No. 2, pp. 559 - 572, 1901.

[101] D. Peña and V. Yohai, A fast procedure for outlier diagnostics in large regression problems, *J. Amer. Statist. Assoc.,* 94, pp. 434–445, 1999.

[102] L. I. Perlovsky and T. L. Marzetta, Estimating a Covariance Matrix from Incomplete Realizations of a Random Vector, *IEEE Trans. on Signal Processing*, 40, pp. 2097-2100, 1992.

[103] L. Petroviĉ, Causality and Markovian representation, *Statist. and Prob. Lett.,* 29, 223–227, 1996.

[104] P. M. Prenter, A Weierstrass theorem for real, separable Hilbert spaces, *J. Approximation Theory,* vol. 3, pp. 341-357, 1970.

[105] P. M. Prenter, Lagrange and Hermite interpolation in Banach spaces, *J. Approximation Theory,* 4, pp. 419-432, 1971.

[106] W. A. Porter, Nonlinear systems in Hilber space, *Int. J. Contr,* vol. 13, pp. 593–602, 1971.

[107] W. A. Porter, An overview of polynomic system theory, *Proc. IEEE,* vol. 64, pp. 36–44, 1976.

[108] J. B. Prolla and S. Machado, Weighted Grothendieck subspaces, *Trans. Amer. Soc.*, 186, pp. 247-258, 1973.

[109] L. Rabiner and B. Gould, *Theory and application of digital processing signals*, Prentice-Hall, Englewood Cliffs, N., J., 1975.

[110] C. R. Rao, *Linear Statistical Inference and its Applications,* Wiley, New York, 1973.

[111] D. C. Rife and R. R. Boorstyn, Single-tone parameter estimations from discrete-time observations, *IEEE Transactions on Information Theory*, IT-20, 3, pp. 591-598, 1974.

[112] W. J. Rugh, *Nonlinear System Theory. The Volterra-Wiener Approach.,* The Johns Hopkins University Press, Baltimore, 1981.

[113] W. Rudin, *Principles of Mathematical Analysis*, McGraw-Hill, New York, 1976.

[114] B. Russel, On the notion of cause, *Proc. Aristotelian Soc.,* 13, pp. 1 - 25, 1913.

[115] M. Ruzhanski and V. Fomin, Optimal Filter Construction for a General Quadratic Cost Functional, *Bulletin of St. Petersburg University. Mathematics,* 28, pp. 50–55, 1995.

[116] R. M. De Santis, Causality Theory in Systems Analysis, *Proc. of IEEE,* 64, pp. 36–44, 1976.

[117] I. W. Sandberg, Conditions for the causality of nonlinear operators defined on a linear space, *Quart. Appl. Math.,* 21, pp. 87 - 91, 1965.

[118] I. W. Sandberg, On Volterra expansions for time–varying nonlinear systems, *IEEE Trans. Circuits & Syst.* CAS-30 (2), pp. 61–67, 1983.

[119] I. W. Sandberg, The mathematical foundations of associated expansions for mildly nonlinear systems, *IEEE Trans. Circuits & Syst.* CAS-30 (7), pp. 441–455, 1983.

[120] I. W. Sandberg, Nonlinear input–output maps and approximate representations, *AT & T Tech. J.* 64, pp. 1967–1983, 1985.

[121] I. W. Sandberg, Approximation theorems for discrete–time systems, *IEEE Trans. Circuits Syst. Part 1, Fund. Theory and Applic.* 38, pp. 564–566, 1991.

[122] I. W. Sandberg, Uniform approximation and the circle criterion, *IEEE Trans. Circuits Syst. Part 1, Fund. Theory and Applic.* 38, pp. 1450–1458, 1993.

[123] I. W. Sandberg, Notes on Uniform Approximation of Timre-varying Systems on Finite Time Intervals, *IEEE Trans. Circuits and Syst. : I: Fundamental Theory and Appl.,* 45, pp. 863 - 865, 1998.

[124] I. W. Sandberg, On approximation of linear functionals on L_p spaces, *IEEE Trans. on Circuits and Syst. – I: Fundamental Theory and Applic.* 42, pp. 402–404, 1995.

[125] I. W. Sandberg, Uniform approximation of multidimensional myopic maps, *IEEE Trans. on Circuits and Syst. – I: Fundamental Theory and Applic.* 44, pp. 477–485, 1997.

[126] I. W. Sandberg, A representation theorem for linear systems, *IEEE Trans. on Circuits and Syst. – I: Fundamental Theory and Applic.* 45, pp. 578–583, 1998.

[127] I. W. Sandberg, Separation conditions and approximation of continuous–time approximately finite memory systems, *IEEE Trans. on Circuits and Syst. – I: Fundamental Theory and Applic.* 46, pp. 820–826, 1999.

[128] I.W. Sandberg, Approximately–Finite Memory and Input–Output Maps, *IEEE Trans. on Circuits and Systems. Part 1, Fundamental theory and applications,* vol. 39, pp. 549 - 556, 1992.

[129] I.W. Sandberg, Criteria for iniform approximatiion using dynamical neural networks," in *Proc. 4th IEEE Mediterranean Symp. Control Automation,* Crete, Greece, June 1996, pp. 1-5, Plenary address.

[130] I.W. Sandberg, Time-Delay Polynomial Networks and Quality of Approximation, *IEEE Trans. on Circuits and Systems. Part 1, Fundamental theory and applications,* vol. 47, pp. 40 - 49, Jan. 2000.

[131] I. W. Sandberg, I.W., $\mathbb{R}_+$-fading memory and extensions of input-output maps. IEEE Trans. Circuits and Syst. I: Fundamental Theory and Applic. 49, pp. 1586–1591, 2002.

[132] B. Sazonov, *A remark about characteristic functionals*, Probability Theory and its Applications, 3, pp. 201-205, 1958.

[133] H. S. Schaefer, *Topological vector spaces*, Springer Verlag, Third Printing, 1971.

[134] L. L. Scharf, The SVD and reduced rank signal processing, *Signal Processing*, vol. 25, pp. 113 - 133, 1991.

[135] M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems,* J. Wiley & Sons, New York, 1980.

[136] M. K. Schneider and A. S. Willsky, A Krylov Subspace Method for Covariance Approximation and Simulation of a Random Process and Fields, *J. Multidim. Syst. & Signal Processing*, 14, pp 295-318, 2003.

[137] A. F. G. Seber and C. J. Wild, *Nonlinear Regression,* Wiley, New York, 2003.

[138] R.B. Shorrok and J.V. Zidek, An improved estimator of the generalized invariance, *Ann. Statist.,* 4, pp. 629–638, 1976.

[139] B.K. Sinha, On improved estimators of the generalized variance, *J. Multivariate Anal.,* 6, pp. 617–626, 1976.

[140] D. Slepian, Prolate spheroidal wave functions, Fourier analysis and uncertainty - V: the discrete case, *The Bell System Technical Journal,* vol. 57, no. 5, pp. 1371-1430, 1978.

[141] E. D. Sontag, *Polynomial Response Maps*, Lecture Notes in Control and Information Sciences, 13, 1979.

[142] H. W. Sorenson, *Parameter Estimation, Principles and Problems*, Marcel Dekker, New York, 1980.

[143] G. W. Stewart, *Introduction to Matrix Computations,* Academic Press, New York, 1973.

[144] J. H. Stock and M. W. Watson, Forecasting using principal components from a large number of predictors, *Journal of the American Statistical Association*, vol. 97, pp. 1167-1179, 2002.

[145] P. Stoica and R. Moses, *Introduction to Spectral Analysis*, Prentice Hall, NJ, 1997.

[146] P. Suppes, *A Probabilistic Theory of Causality,* Amsterdam, North-Holland, 1970.

[147] J. B. Tenenbaum, V. de Silva and J. C. Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science,* vol. 290, issue 5500, pp. 2319–2323, 2000.

[148] V. Timofte, Stone-Weierstrass theorem revisited, *J. Approx. Theory*, 136, pp. 45–59, 2005.

[149] M. E. Tipping and C.M. Bishop, Probabilistic principal component analysis, *J. of the Royal Statistical Society. Serie A,* vol. 61, pp. 611 - 619, 1999.

[150] M. E. Tipping and C.M. Bishop, Mixtures of Probabilistic Principal Component Analysers, *Neural Computation,* vol. 11, pp. 443 - 482 , 1999.

[151] A. P. Torokhti, On polynomial synthesis of non-linear systems, *Electronic Modelling, Journal of the USSR Academy of Sciences*, 11, pp. 28-34, 1989 (in Russian).

[152] A. P. Torokhti, Modelling of non-linear dynamical systems, *Electronic Modelling, Journal of the USSR Academy of Sciences*, 6, pp. 10-16, 1990 (in Russian).

[153] A. P. Torokhti, On constructive restoration of continuous maps in Banach spaces, *Sov. Mathematics (Izv. VUZ. Mathematics)*, 1, pp. 86-89, 1990 (in Russian).

[154] A. P. Torokhti, On constructive approximation of non-linear operators, *Sov. Mathematics (Izv. VUZ. Mathematics)*, 7, pp. 24-28, 1991 (in Russian).

[155] A. P. Torokhti, Polynomial synthesis and numerical relization of non-linear systems, *Mathematical Modelling and Computations,* pp. 64–74, Bulgar Acad. Sci., 1992.

[156] A. P. Torokhti and P. G. Howlett, On the constructive approximation of non–linear operators in the modelling of dynamical systems, *J. Austral. Math. Soc. Ser. B.* 39, pp. 1–27, 1997.

[157] A. Torokhti and P. Howlett, On the Best Quadratic Approximation of Nonlinear Systems, *IEEE Trans. on Circuits and Systems. Part I, Fundamental theory and applications*, vol. 48, pp. 595–602, May 2001.

[158] A. Torokhti and P. Howlett, Optimal Fixed Rank Transform of the Second Degree, *IEEE Trans. on Circuits & Systems. Part II, Analog & Digital Signal Processing,* v ol. 48, pp. 309–315, 2001.

[159] A. Torokhti, P. Howlett and C. Pearce, Methods of Constrained and Unconstrained Approximation for Mappings in Probability Spaces, *Modern Applied Mathematics,* Ed. J.C. Misra, Narosa, New Delhi, pp. 83–129, 2005.

[160] A. Torokhti and P. Howlett, On the Best Quadratic Approximation of Nonlinear Systems, *IEEE Trans. on Circuits and Systems. Part I, Fundamental theory and applications* 48(5), pp. 595–602, 2001.

[161] A. P. Torokhti, P. G. Howlett and C. Pearce, Method of Hybrid Approximations for Modeling of Multidimensional Nonlinear Systems, *J. Multidim. Syst. & Signal Processing,* 14, no 4, pp. 397-410, 2003.

[162] A. Torokhti and P. Howlett, An Optimal Filter of the Second Order, *IEEE Trans. on Signal Processing*, 49(5), pp. 1044–1048, 2001.

[163] A. Torokhti, P. Howlett and C. Pearce, Optimal Recursive Estimation of Raw Data, *Annals of Ooperations Research,* 133, pp. 285–302, 2005.

[164] P. Howlett, C. Pearce and A. Torokhti, On Nonlinear Operator Approximation with Preassigned Accuracy, *J. Comput. Anal. Appl.,* 5, no. 3, pp. 273–298, 2003.

[165] Torokhti, A., Howlett, P. and Pearce, C. Method of Best Successive Approximations for Nonlinear Operators, *J. Comput. Anal. Appl.*, 5, no. 3, pp. 299–312, 2003.

[166] A. Torokhti and P. Howlett, Constructing Fixed Rank Optimal Estimators with Method of Recurrent Best Approximations, *J. Multivariate Analysis,* 86, 2, pp. 293-309, 2003.

[167] A. Torokhti and P. Howlett, Best Approximation of Operators in the Modelling of Nonlinear Systems, *IEEE Trans. On Circuits & Systems. Part I, Fundamental Theory & Appl.*, 49, no. 12, 1792-1798, 2002.

[168] A. Torokhti and P. Howlett, Best Causal Mathematical Models for Nonlinear Systems, *IEEE Trans. On Circuits & Systems. Part I, Fundamental Theory & Appl.*, 52, no. 5, 1013-1020, 2005.

[169] A. Torokhti and P. Howlett, Method of recurrent best estimators of second degree for optimal filtering of random signals, *Signal Processing*, vol. 83, pp. 1013–1024, 2003.

[170] A. Torokhti and P. Howlett, Optimal Transform Formed by a Combination of Nonlinear Operators: The Case of Data Dimensionality Reduction, *IEEE Trans. on Signal Processing*, 54, 4, pp. 1431-1444, 2006.

[171] A. Torokhti and P. Howlett, Best approximation of identity mapping: the case of variable memory, *J. Approx. Theory,* 143, 1, pp. 111-123, 2006.

[172] A. Torokhti and P. Howlett, Towards Best Approximation of Nonlinear Systems: A Case of Models with Memory, *J. Numerical Functional Analysis & Optimization,* (accepted).

[173] A. Torokhti, P. Howlett and C. Pearce, New perspectives on optimal transforms of random vectors, *Optimization: Theory and Applications,* Kluwer (to appear).

[174] V. N. Vapnik, *Estimation of dependences based on empirical data,* Springer–Verlag, New York, 1982.

[175] Y. Wan, T. J. Dodd and R. F. Harrison, Infinite degree Volterra series estimation, *CDROM Proc. 2nd Int. Conf. on Comput. Intelligence, Robotics and Autonomous Systems,* Singapore, 2003.

[176] L. Wang, Asymptotics of estimates in constrained nonlinear regression with long-range dependent innovations., *Ann. Inst. Statist. Math.,* 56, pp. 251–264, 2004.

[177] M. Weiss and H. A. Preisig, Structural analysis in the dynamical modelling of chemical engineering systems. *Mathematical and Computer Modelling of Dynamical Systems.* 6(4), pp. 325 –364, 2000.

[178] G. B. Wetherill, *Regression Analysis with Applications,* Chapman and Hall, London, 1986.

[179] N. Wiener, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications,* Academic Press, New York, 1949.

[180] J. C. Willems, Stability, instability, invertibility and causality, *SIAM J. Control*, 7 no. 4, pp. 645–671, 1969.

[181] C. O. Wu, Local polynomial regression with selection biased data, *Statist. Sinica,* 10, pp. 789–817, 2000.

[182] I. Yamada, T. Sekiguchi and K. Sakaniwa, Reduced rank Volterra filter for robust identification of nonlinear systems, *Proc. 2nd Int. Workshop on Multidimensional (ND) Systems - NDS2000,* Czocha Castle, Poland, 171–175, 2000.

[183] Y. Yamashita and H. Ogawa, Relative Karhunen–Loève transform, *IEEE Trans. on Signal Proc.,* 44, pp. 371–378, 1996.

[184] Y. Q. Yin, Limiting spectral disribution for a class of random matrices, *J. Multivariate Anal.*,20, pp. 50–68, 1986.

[185] K. Yosida, *Functional Analysis*, 5th edn, Springer-Verlag, New York, 1978.

[186] L. H. Zou and J. Lu, Linear associative memories with optimal rejection to coloured input noise, *IEEE Trans. on Circuits and Syst. – II: Analogue and Digital Signal Proc.,* 44, pp. 990–1000, 1997.

This page intentionally left blank

Index

This page intentionally left blank

Mathematics in Science and Engineering
Edited by C.K. Chui, Stanford University

Recent titles:
C. De Coster and P. Habets, *Two-Point Boundary Value Problems: Lower and Upper Solutions*
Wei-Bin Zang, *Discrete Dynamical Systems, Bifurcations and Chaos in Economics*
I. Podlubny, *Fractional Differential Equations*
E. Castillo, A. Iglesias, R. Ruíz-Cobo, *Functional Equations in Applied Sciences*
V. Hutson, J.S. Pym, M.J. Cloud, *Applications of Functional Analysis and Operator Theory (Second Edition)*
V. Lakshmikantham and S.K. Sen, *Computational Error and Complexity in Science and Engineering*
T.A. Burton, V*olterra Integral and Differential Equations (Second Edition)*
E.N. Chukwu, *A Mathematical Treatment of Economic Cooperation and Competition Among Nations: with Nigeria, USA, UK, China and Middle East Examples*
V.V. Ivanov and N. Ivanova, *Mathematical Models of the Cell and Cell Associated Objects*
Z. Zong, *Information-Theoretic Methods for Estimating Complicated Probability Distributions*
A.G. Ramm, *Dynamical Systems Method for Solving Operator Equations*
J. Mishra and S.N. Mishra, *L-System Fractals*
I.V. Konnov, *Equilibrium Models and Variational Inequalities*
H.G. Dehling, T. Gottschalk and A.C. Hoffmann, *Stochastic Modeling in Process Technology*