NOTE TO USERS

DISSERTATION

TWO-CHANNEL SIGNAL PROCESSING IN CANONICAL COORDINATES

Submitted by

Ali Pezeshki

Department of Electrical and Computer Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2004

COLORADO STATE UNIVERSITY

October 25, 2004

WE HEREBY RECOMMEND THAT THE **DISSERTATION** PREPARED UNDER OUR SUPERVISION BY **ALI PEZESHKI** ENTITLED **TWO-CHANNEL SIGNAL PROCESSING IN CANONICAL COORDINATES** BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Prof. Anthony A. Maciejewski

Prof. F. Jay Breidt

Prof. Mahmood R. Azimi-Sadjadi
Adviser

Prof. Louis L. Scharf
Co-Adviser

Prof. Anthony A. Maciejewski
Department Head/Director

ABSTRACT OF DISSERTATION

TWO-CHANNEL SIGNAL PROCESSING IN CANONICAL COORDINATES

Canonical coordinates provide an elegant framework for analyzing and solving many two-channel problems in signal processing, communications, radar and sonar, and sensor fusion. This dissertation addresses some of the existing issues in canonical correlation analysis of two-channel data, establishes a direct connection between canonical coordinates and certain two-channel signal processing problems, and exploits canonical correlation analysis to solve some real two-channel signal processing problems.

More specifically, in this dissertation, connections between two-channel constrained least squares (CLS) problems and various canonical coordinate systems are established. It is shown that under certain sets of constraints a two-channel CLS problem will produce one of the important canonical coordinate systems, namely canonical coordinates, half-canonical coordinates, or programmable canonical correlation analysis (PCCA) coordinates. Further, a unified framework for building reduced-rank Wiener filters is developed. It is demonstrated that, depending on the objective of reduced-rank estimation, either canonical coordinates or half-canonical coordinates are optimal for building the reduced-rank Wiener filter. Simple algorithms, called alternating power methods, are also developed that allow for both recursive and real-time computation of canonical coordinates, half-canonical coordinates, and reduced-rank Wiener filters. The developed algorithms may be viewed as two-step decompositions of the standard power method, as they solve a coupled generalized eigenvalue problem through power iterations. In addition, a network structure, with lateral connections

that implement a deflation process, is developed for recursive extraction of canonical coordinates.

This dissertation also addresses the empirical canonical coordinate decompositions of two-channel data, where the channel covariances are estimated from a limited number of data samples and are not necessarily full-rank. It clarifies how the number of samples (sample support) drawn from two-channel data, and the ranks of the data matrices, affect the algebraic and geometric properties of empirical canonical correlations and coordinates. It is shown that empirical canonical correlations are maximal invariants that measure the cosines of the principal angles between the row spaces of the data matrices for the two data channels. When the sample support is smaller than the sum of the ranks of the two data matrices, some of the empirical canonical correlations become one, regardless of the two-channel model that generates the samples. In such cases, the empirical canonical correlations may not be used as estimates of correlation between random variables. This has interesting implications for canonical correlation analysis of nonlinear functions of two-channel data, where the aim is to capture coherence between the two channels by estimating correlation between their high-order attributes. This will be possible only if the sample support is greater than the sum of the ranks of the nonlinearly mapped data matrices. In these cases, however, the so-called kernel formulations of canonical correlation analysis are computationally disadvantageous with respect to the direct formulations.

Finally, canonical correlation analysis is employed to develop a multi-aspect feature extraction method for underwater target classification. The developed feature extraction method exploits the linear dependence or coherence between two consecutive sonar returns. This is accomplished by extracting the dominant canonical correlations between the two sonar returns and using them as features for classifying mine-like objects from non-mine-like objects. The experimental results on a wideband acoustic backscattered data set, which contains sonar returns from several

mine-like and non-mine-like objects in two different environmental conditions, show that canonical correlation features can offer good discrimination between mine-like and non-mine-like objects. Further, the results show that in a fixed bottom condition, canonical correlation features do not vary with changes in aspect angle.

Ali Pezeshki
Department of Electrical and Computer Engineering
Colorado State University
Fort Collins, Colorado 80523
Fall 2004

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deep gratitude to my advisor, Professor Mahmood R. Azimi-Sadjadi, and my co-advisor, Professor Louis L. Scharf, for their generous support, encouragement, and guidance throughout my PhD program. Their constant flow of ideas, thoughtful discussions, and careful comments and criticism have profoundly influenced my way of thinking. I am also grateful for the opportunities they have provided for me to work on several interesting and challenging research problems, and also interact with many distinguished researchers. I have been very fortunate to have studied under their supervision.

I would also like to thank Professors Anthony A. Maciejewski and F. Jay Breidt for serving on my committee and reviewing this thesis.

Special thanks are due to the Office of Naval Research (ONR) for funding this project, under contract number N00014-02-1-0006. The data used in this work were supplied by The Applied Research Laboratories at the University of Texas at Austin under the auspices of the Office of Naval Research, Code 321 Undersea Signal Processing. I would like to thank Drs. Pat Pitt and Ken Scarbrough from the ARL-UT for providing this data set and their technical assistance.

Special thanks are also due to Professor Yingbo Hua, at the University of California at Riverside, for valuable comments and feedback on the results of Chapters 3, 4, and 5, and to Dr. Gerald Dobeck at the Coastal Systems Station for many useful discussions and comments on the results of Chapter 8.

I am grateful to my former and current colleagues, Marc Robinson, Jiame Salazar, Arta Jamshidi, Kishor Saitwal, Jianqi Wang, Amanda Falcone, Kumar Srinivasan, Gordon Wichern, Makoto Yamada, Vincent Wong, and Ramin Zahedi for creating a stimulating work atmosphere.

I am indebted to my friends, Leili Taeb, Sina Farsiu, Bardia Alavi, Azadeh Faridi, and Emad Dolatshhi-Zand for their important, albeit indirect, influence on the development of this thesis.

Finally, I would like to express my appreciation towards my parents, Mohammad and Homa, and my brother, Houman. I owe them all that I am, and all that I have ever accomplished.

To my parents, Homa and Mohammad,

my brother, Houman, and my grandmother, Fakhr Iran.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Two-channel problems find numerous applications in signal processing, communication, sonar, radar, and sensor fusion. In filtering and communication one of the channels contains the unobserved source variables to be estimated and the other channel contains the observed measurement variables. In radar and sonar the two channels may be the outputs of two subarrays in space, or the outputs over two subintervals in time. In sensor fusion, the channels correspond to different sensory measurements of the same process. Typically, the problem is one of estimating a linear function of variables in one channel from a linear combination of variables in the other, or one of measuring the linear dependence or coherence between the elements of the two channels, or the rate at which one channel carries information about the other.

The type of two-channel problem to be solved determines the right coordinate system for obtaining and analyzing the solution. For many two-channel problems in signal processing and communications, canonical coordinates [1]– [7] have been shown to provide the right coordinate system. In [6] and [7] it is shown that all performance measures commonly used for second-order inference and communication over Gaussian channels are determined only by the canonical correlations [1]– [7] of two-channel data. These performance measures are invariant to uncoupled nonsingular transformations of the channels. More specifically, in [6] and [7] it is shown that canonical coordinates decompose the linear minimum mean-squared error (MMSE) estimator

into a parallel set of uncorrelated canonical linear MMSE estimators. The volume of the concentration ellipse[1] of the filtering error, or equivalently determinant of the error covariance matrix of the linear MMSE estimator is found to be multiplicatively decomposed into the product of volumes of canonical concentration ellipses, each of which is determined by a canonical correlation. Additionally, [6] and [7] showed that linear dependence between two data channels is decomposed into the product of the linear dependence between their canonical coordinates, and is determined by the corresponding canonical correlations. It was also demonstrated that the transformation to canonical coordinates transforms a Gaussian channel into a parallel combination of independent Gaussian channels, each of which communicates at a (canonical) rate that depends only on a canonical correlation. Further, the total information rate or mutual information between the channels is the sum of these canonical rates.

In addition, canonical coordinates have been shown [7], [9] to be optimal for reduced-rank estimation [6]– [11], when the objective of estimation is to minimize the volume of the concentration ellipse of the filtering error, and for quantization of noisy sources when the objective is to preserve maximum information rate [12].

In view of the importance of canonical coordinates in two-channel signal processing, in this thesis we aim to address some of the issues in canonical correlation analysis of two-channel data, establish a direct connection between canonical coordinates and certain two-channel signal processing problems, and exploit the use of canonical correlation analysis in some real two-channel signal processing applications. More specifically, the main topics addressed in this thesis are:

1. Connections between canonical coordinates and two-channel constrained least squares (CLS) problems,

---

[1]The concentration ellipse for a zero-mean random vector $\mathbf{e}$ with covariance matrix $E[\mathbf{e}\mathbf{e}^T] = \mathbf{R}_{ee}$ is defined as $\{\mathbf{e} : \mathbf{e}^T\mathbf{R}_{ee}^{-1}\mathbf{e} = 1\}$. The volume of this concentration ellipse is proportional to $\det\{\mathbf{R}_{ee}\}$ [8].

2. Optimal reduced-rank Wiener filtering in canonical coordinates,

3. Recursive computation of canonical coordinates,

4. Extraction of canonical coordinates using a network with lateral connections,

5. Empirical canonical coordinate decompositions of two-channel linear and non-linear maps, and

6. Feature extraction in canonical coordinates.

In what follows, we present a brief introduction to each of these topics.

1.1 Connections Between Canonical Coordinates and Two-Channel CLS Problems

Two-channel CLS problems are concerned with estimating a linear combination of elements in one data channel from a linear combination of the elements in another data channel, under certain constraints. The interpretation of the channels depends on the application. Naturally, the performance objective of the two-channel filtering problem, and the corresponding constraints, determines the coordinate system under which the problem has to be solved. We demonstrate in this thesis (see also [13]) that under certain sets of constraints, two-channel least squares problems lead to various canonical coordinate decompositions, namely canonical coordinate decompositions [7], half-canonical coordinate decompositions [8], [10], and programmable canonical correlation analysis (PCCA) coordinate decompositions [14]- [17]. The half-canonical coordinates are important for designing a class of optimal reduced-rank filters [8]–[9]. The PCCA coordinate system was proposed in [14]- [17] for adaptive source separation. It was this work that motivated us to develop two-channel CLS problems and establish their connections with various canonical coordinate systems.

Apart from establishing a direct connection between various canonical coordinate systems and a general class of two-channel problems, our results on this topic

provide us with a foundation to develop recursive methods for computing canonical coordinates and half-canonical coordinates and also allow us to establish connections between two-channel CLS filters and optimal reduced-rank Wiener filters.

1.2 Optimal Reduced–Rank Wiener Filtering in Canonical Coordinates

Reduced-rank estimation and filtering [6]– [11], [18]– [21] are important for a wide range of signal processing applications where data or model reduction, robustness against noise or model errors, or high computational efficiency is desired. Fundamental results on optimal reduced-rank estimators and filters include the work by Brillinger [11], the reduced-rank Wiener filter (RRWF) by Scharf [6]– [10], and the reduced-rank maximum likelihood estimator (RRMLE) by Stoica and Viberg [18]. Other examples of reduced-rank estimators and filters include the reduced-rank multilayer neural network (RRMNN) by Diamantaras and Kung [19], the relative Karhunen-Loeve transform (RKLT) by Yamashita and Ogawa [20], and the generalized Karhunen-Loeve transform (GKLT) by Hua and Liu [21].

The choice of the coordinate system for building an optimal reduced-rank Wiener filter depends on the measure to be optimized. Common measures for reduced-rank Wiener filtering are (1) mean-squared error (MSE), or trace of error covariance matrix, (2) whitened (weighted) MSE, and (3) volume of the concentration ellipse, or determinant of the error covariance matrix. In [9], reduced-rank Wiener filtering under these measures have been reviewed and the last two measures has been shown to be equivalent.

In this thesis (see also [13]), we intend to review the connections between different classes (each class corresponds to one of the error measures) of optimal reduced-rank Wiener filters and clarify their connections with canonical coordinates, half-canonical coordinates, and two-channel CLS filters. We demonstrate that canonical coordinates

are optimal for reduced-rank Wiener filtering when the objective of estimation is to minimize either the volume of the concentration ellipse of the filtering error or the whitened MSE. Further, we establish that half-canonical coordinates are optimal for reduced-rank Wiener filtering when the objective is to minimize the MSE. Our results reproduce those of [9]. However, we obtain all of these results in a *unified* way using the line of argument given in [8] for optimal reduced-rank Wiener filtering in half-canonical coordinates. In addition, we determine several equivalent representations of reduced-rank Wiener filters in each class, and at the same time clarify the connections between reduced-rank Wiener filters and two-channel CLS filters.

1.3 Recursive Computation of Canonical Coordinates

In many applications such as reduced-rank estimation and low-rank modelling [6]–[11], [18]– [21] only a small subset of canonical coordinates, which have large canonical correlations, need to be extracted. Discarding canonical coordinate pairs with small canonical correlations has little effect on two-channel performance measures such as volume of the concentration ellipse, information rate, and linear dependence. The problem is that a conventional method of canonical coordinate decomposition, e.g. the one in [7], does not offer a simple way for computing a small subset of canonical coordinates. A full singular value decomposition (SVD) for a coherence matrix [7] has to be computed, regardless of the rank-reduction. In addition, the conventional method does not allow an easy update of the canonical coordinate mappings in time, making it intractable for online applications. A similar argument applies to a conventional method of half-canonical coordinate decomposition [8].

To address these shortcomings of the conventional methods of canonical and half-canonical coordinate decomposition, we derive (see also [13]) various algorithms, called *alternating power methods*, with *deflation*, to recursively compute the canonical

coordinate and half-canonical coordinate *mapping vectors*, one-by-one or in groups. Canonical coordinate and half-canonical coordinate mapping vectors are the vectors that transform the channels into their canonical coordinates. The algorithms we develop also allow for updating the mapping vectors in time as new samples of the channels are observed. In addition, provided that the rank-reduction is relatively large and canonical correlations or half-canonical correlations are not close together, the alternating power methods can be computationally more efficient than the conventional methods, as they do not require any matrix square-root operations.

Our alternating power methods are identical in form to the alternating power methods derived in [9] for computing reduced-rank Wiener filters. However, the algorithms in [9] do not yield the canonical or half-canonical coordinate maps. Thus, what is original here is the discovery that alternating power methods may be used to compute canonical and half-canonical coordinate maps as well as canonical and half-canonical correlations, making them more applicable for signal processing problems than they would appear from the work in [9].

1.4 Extraction of Canonical Coordinates using a Network with Lateral Connections

In 1982, Oja [22] showed that a linear network with a single node trained with a normalized Hebbian rule can extract the dominant principal component of a stationary vector process. Sanger [23] and Foldik [24] extended Oja's work to the multi-node case in order to simultaneously extract the first m principal components of a vector process. Diamantaras and Kung [25], [26] exploited the idea of using lateral connections with anti-Hebbian learning to recursively extract the principal components. In a different approach, based on recursive least squares (RLS) learning, Bannour and Azimi-Sadjadi [27] proposed another structure for recursive extraction of principal components.

The interesting fact about the work in [25] is the use of lateral connections for recursively computing the principal components of a data channel. The network proposed in [25] is called an APEX (adaptable principal component extractor) network. It consists of two parts: (1) a simple feedforward network that is updated using a Hebbian-type learning and can extract the dominant principal component of a data channel and (2) a set of lateral connections that connect the outputs together and are trained to deflate the contribution of the already extracted principal components from the input data channel. The combination of these two sets of connections allows for recursively extracting the principal components, one by one. Each time that a new principal component has to be extracted, a new node is added to the APEX network, allowing for extraction of the new principal component, without the need to retrain the previous nodes.

Recently, Lai and Fyfe [28] proposed a network for performing canonical correlations analysis. However, their network only finds the most significant canonical coordinate pair and the corresponding canonical correlation. In this thesis (see also [29] and [30]), motivated by the APEX structure in [25], we develop a network structure with lateral connections for recursively extracting the canonical coordinates of a two-channel vector process. The network is structured to use lateral connections for performing deflations. The network weights are updated using a gradient descent algorithm [31], so they suffer from slow convergence (even as slow as linear convergence) and sometimes instability, depending on initialization and choice of the step size.

1.5 Empirical Canonical Coordinate Decompositions of Two-Channel Linear and Nonlinear Maps

All the developments and results reported to date for canonical correlation analysis of two-channel data, including those in [1]– [7], are based on the assumption that either the theoretical covariance and cross-covariance matrices of the channels are known, or enough independent copies of the channels are available to obtain full-rank estimates of the covariance matrices. Little attention has been devoted to the algebraic limits to canonical correlation analysis, when two-channel data are scarce. That is, just how poor can sample support become before sample canonical correlations cease to carry any information about the true canonical correlations? This is one of the particular questions we address in this thesis.

More generally, we study (see also [32]) the *empirical* canonical correlation analysis of two-channel data. The term empirical implies that the canonical correlation analysis is based on covariance matrices that are estimated from a limited number of samples of the two-channel data, and are not necessarily full-rank. The available data samples may be obtained from linear or nonlinear transformations of a limited number of random samples drawn from a two-channel vector process. The question to be addressed in this thesis is whether or not the canonical correlations and coordinates obtained from sample data have the same algebraic and geometric properties as the underlying theoretical ones. For example, when do empirical canonical correlations estimate theoretical canonical correlations between two data channels? We show that when the sample support (number of data samples drawn from each data channel) is smaller than the sum of the ranks of the two data matrices[2], the empirical canonical correlations are defective and may not be used as estimates of canonical correlations,

[2]The data matrices are column-wise collections of vector data samples that are drawn from the two (vector) data channels.

or coherence, between random variables of the two-channel data.

Our results have interesting implications for canonical correlation analysis of nonlinear functions of two-channel data, where the aim is to capture coherence between the two channels by estimating correlation between their high-order attributes. The basis for considering nonlinearities is that canonical correlation analysis only exploits the second-order statistics of two-channel data. In some applications, however, the information that is obtained from second-order statistics alone, such as linear dependence, may not be sufficient to achieve the desired performance. By pre-processing the data with nonlinear functions, one may be able to capture coherence between the two data channels by estimating the canonical correlations between their high-order attributes. We show that this is possible only when the sample support is greater than the sum of the ranks of the nonlinearly mapped data matrices.

The idea of using nonlinear maps prior to linear processing was first exploited by Vapnik in the theory of support vector machines (SVM) for the design of large margin classifiers [33], [34]. The idea in SVM is to use a nonlinear mapping to map the input space into a high-dimensional feature space, in which the features are linearly separable. Perhaps the most intriguing aspect of SVM is that the high dimensional nonlinear mappings are never *explicitly* computed and all computations are carried out in the original low dimensional space, using the *kernel trick* [35]. Since the development of SVM, numerous results have been reported on kernel-nonlinear counterparts of standard information processing techniques, among which are kernel versions of principal component analysis [35], [36], Fisher discriminant analysis [35], [37], linear least squares estimation [38], Mahalanobis distance [38], and even canonical correlation analysis [39]- [46]

However, our results on the effect of sample support on algebraic and geometric properties of empirical canonical correlations indicate that in cases where empirical canonical correlations can estimate the theoretical canonical correlations the kernel

formulations for canonical correlation analysis are in fact computationally disadvantageous with respect to the direct formulations, in which the nonlinear mappings are explicitly computed.

It should be mentioned that some of our findings, concerning the effect of sample support on empirical canonical correlations, have been reported in [45], without rigorous proof and analysis.

1.6 Feature Extraction in Canonical Coordinates

Measuring linear dependence or coherence between multiple data channels may be used for detecting the presence of an unknown but common signature among the channels. This is the basic idea behind multi-channel tests for linear dependence [3] and the multiple coherence test of [47], [48]. The idea is that linear dependence between the channels is an indication of the presence of a common signature, whereas linear independence indicates the absence of a common signature.

In a two-channel case, the linear dependence between the channels is measured by the canonical correlations of the channels. This implies that canonical correlations can be viewed as features that capture linear dependence or coherence between two data channels, and hence may be used for detection or classification purposes. We intend to exploit this idea for extracting features to classify underwater mine-like objects from non-mine-like objects (see also [49] and [50]). In this application, the channels correspond to acoustic backscattered signals at two consecutive aspect angles, with certain spatial separation. Using canonical correlations, we exploit the linear dependence between two backscattered signals (sonar returns) to determine whether common signatures associated with targets or non-targets are present. We hypothesize that the amount of coherence between the two sonar returns generated by the presence of a mine-like object is different from that caused by the presence of a non-mine-like object. Therefore, the dominant canonical correlations, which capture

most of the coherence between two sonar returns, may be used to classify the objects at the corresponding aspect angles. We test our hypothesis on a subset of a wideband data set [51] that has been collected at the Applied Research Lab (ARL), University of Texas (UT)-Austin, and benchmark our results against those obtained in [52] on the same data set.

Additionally, we follow up on our discussion of the potential use of nonlinearly mapped two-channel data, followed by canonical correlation analysis, with the aim to capture coherence between high-order attributes of the two channels. We use several nonlinearities to map the data samples, extracted from the sonar returns, in order to investigate whether the canonical correlations between the nonlinear functions of the backscattered signals can improve the discrimination of mine-like objects from non-mine-like objects.

1.7 Organization of the Thesis

In Chapter 2, we review canonical coordinates and canonical correlations, highlight their algebraic and geometric properties, and clarify their importance in analyzing the Wiener filter, linear dependence, and information rate. Further, we review half-canonical coordinates to prepare the readers for the developments in other chapters. Therefore, this chapter provides a foundation for our developments throughout this thesis. Chapter 3 establishes the connections between two-channel CLS problems and various canonical coordinate systems, namely canonical coordinates, half-canonical coordinates, and PCCA coordinates. In Chapter 4, reduced-rank Wiener filtering in canonical and half-canonical coordinates is reviewed and connections between reduced-rank Wiener filters and two-channel CLS filters are established. Chapter 5 discusses the recursive computation of canonical coordinates, half-canonical coordinates, and correlations. In Chapter 6, a network structure with lateral connections and a set of learning rules for extracting canonical coordinates is presented. Chapter

7 addresses the empirical canonical coordinate decomposition of two-channel data, in which the channel covariances are not known and need to be estimated from a limited number of data samples. In Chapter 8, we exploit canonical correlations as features for classifying underwater targets from non-targets. Chapter 9 presents a review of some of the existing methods [53]– [56] for selecting appropriate nonlinearities for different nonlinear information processing techniques, in order to achieve a certain performance level. Finally, Chapter 10 draws conclusions and outlines suggestions for future work.

CHAPTER 2

CANONICAL COORDINATES AND THE GEOMETRY OF INFERENCE AND RATE

2.1 Introduction

In this chapter, we review canonical coordinates and canonical correlations, and highlight their algebraic and geometric properties. Our aim is to clarify why the canonical coordinate system is the right coordinate system for analyzing Wiener filters, linear dependence, and information rate. In addition, we review half-canonical coordinates and half-canonical correlations to prepare the readers for the developments in other chapters. Therefore, this chapter provides the foundation for our developments in this thesis. The material presented here, and much of the language and terminology, are drawn from [7] and [8].

2.2 Canonical Coordinates

Let us consider the composite data vector $\mathbf{z}$ consisting of two random vectors $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$, $m \leq n$, i.e.

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \in \mathbb{R}^{(m+n)}. \tag{2.1}$$

We assume that $\mathbf{x}$ and $\mathbf{y}$ have zero means and share the composite covariance matrix

$$\mathbf{R}_{zz} = E[\mathbf{z}\,\mathbf{z}^T] = E\left[\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\begin{pmatrix}\mathbf{x}^T & \mathbf{y}^T\end{pmatrix}\right] = \begin{bmatrix}\mathbf{R}_{xx} & \mathbf{R}_{xy}\\ \mathbf{R}_{yx} & \mathbf{R}_{yy}\end{bmatrix}. \tag{2.2}$$

Whenever it is needed to assign a probability distribution to $\mathbf{z}$, we assume it to be Gaussian and denote it by $\mathbf{z} : N(\mathbf{0}, \mathbf{R}_{zz})$. In such cases, $\mathbf{x}$ and $\mathbf{y}$ will be marginally Gaussian, that is, $\mathbf{x} : N(\mathbf{0}, \mathbf{R}_{xx})$ and $\mathbf{y} : N(\mathbf{0}, \mathbf{R}_{yy})$.

We may think of the elements of the cross-covariance matrix $\mathbf{R}_{xy}$, i.e. $[\mathbf{R}_{xy}]_{ij} = E[x_i y_j]$, as inner products in the Hilbert space of second-order random variables. Here, x_i is the ith element of $\mathbf{x}$ and y_j is the jth element of $\mathbf{y}$.

If $\mathbf{x}$ and $\mathbf{y}$ are now replaced by their corresponding white vectors, then the whitened composite vector $\boldsymbol{\xi}$ is obtained as

$$\boldsymbol{\xi} = \begin{bmatrix}\boldsymbol{\zeta}\\ \boldsymbol{\nu}\end{bmatrix} = \begin{bmatrix}\mathbf{R}_{xx}^{-1/2} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_{yy}^{-1/2}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}, \tag{2.3}$$

where $\mathbf{R}_{xx}^{1/2}$ is a square-root (not necessarily symmetric) of $\mathbf{R}_{xx}$. That is, $\mathbf{R}_{xx}^{1/2}\mathbf{R}_{xx}^{T/2} = \mathbf{R}_{xx}$ and $\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xx}\mathbf{R}_{xx}^{-T/2} = \mathbf{I}$. The covariance matrix of $\boldsymbol{\xi}$ may be written as

$$\mathbf{R}_{\xi\xi} = E[\boldsymbol{\xi}\,\boldsymbol{\xi}^T] = E\left[\begin{pmatrix}\boldsymbol{\zeta}\\ \boldsymbol{\nu}\end{pmatrix}\begin{pmatrix}\boldsymbol{\zeta}^T & \boldsymbol{\nu}^T\end{pmatrix}\right] = \begin{bmatrix}\mathbf{R}_{\zeta\zeta} & \mathbf{R}_{\zeta\nu}\\ \mathbf{R}_{\nu\zeta} & \mathbf{R}_{\nu\nu}\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \mathbf{C}\\ \mathbf{C}^T & \mathbf{I}\end{bmatrix}, \tag{2.4}$$

where

$$\mathbf{C} = E[\boldsymbol{\zeta}\boldsymbol{\nu}^T] = E[(\mathbf{R}_{xx}^{-1/2}\mathbf{x})(\mathbf{R}_{yy}^{-1/2}\mathbf{y})^T] = \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} \tag{2.5}$$

is called the *coherence matrix* of $\mathbf{x}$ and $\mathbf{y}$. Therefore, the coherence matrix $\mathbf{C}$ is the cross-covariance matrix between the white versions of $\mathbf{x}$ and $\mathbf{y}$. Correspondingly, the coordinates $\boldsymbol{\zeta}$ and $\boldsymbol{\nu}$ are called the *coherence coordinates*. Since $\boldsymbol{\zeta}$ and $\boldsymbol{\nu}$ are white vectors, the elements of the coherence matrix $\mathbf{C}$ measure the cosines of the angles between the elements of $\boldsymbol{\zeta}$ and $\boldsymbol{\nu}$. That is, $[\mathbf{C}]_{ij} = E[\zeta_i \nu_j]$ measures the cosine of the angle between ζ_i and ν_j, in the Hilbert space of second-order random variables.

We now determine the singular value decomposition (SVD) of the coherence matrix, namely

$$\begin{aligned} \mathbf{C} &= \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} = \mathbf{F}_c\mathbf{\Sigma}_c\mathbf{G}_c^T \quad \text{and} \\ \mathbf{F}_c^T\mathbf{C}\mathbf{G}_c &= \mathbf{F}_c^T\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}\mathbf{G}_c = \mathbf{\Sigma}_c, \end{aligned} \tag{2.6}$$

where $\mathbf{F}_c \in \mathbb{R}^{m\times m}$ and $\mathbf{G}_c \in \mathbb{R}^{n\times n}$ are orthogonal matrices, i.e.

$$\mathbf{F}_c^T\mathbf{F}_c = \mathbf{F}_c\mathbf{F}_c^T = \mathbf{I}(m) \quad \text{and} \quad \mathbf{G}_c^T\mathbf{G}_c = \mathbf{G}_c\mathbf{G}_c^T = \mathbf{I}(n), \tag{2.7}$$

and

$$\mathbf{\Sigma}_c = [\ \mathbf{\Sigma}_c(m) \quad \mathbf{0}\ ] \in \mathbb{R}^{m\times n} \tag{2.8}$$

is a diagonal singular value matrix, with $\mathbf{\Sigma}_c(m) = \text{diag}[\sigma_1, \sigma_2, \ldots, \sigma_m]$ and $1 \geq \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m > 0$.

Let us use the orthogonal matrices $\mathbf{F}_c$ and $\mathbf{G}_c$ to transform the whitened composite vector $\boldsymbol{\xi}$ into the canonical composite vector $\mathbf{w}$,

$$\mathbf{w} = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{F}_c^T & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_c^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} \mathbf{F}_c^T & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_c^T \end{bmatrix} \begin{bmatrix} \mathbf{R}_{xx}^{-1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{yy}^{-1/2} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}. \tag{2.9}$$

Then, the covariance matrix for the canonical composite vector $\mathbf{w}$ is obtained as

$$\mathbf{R}_{ww} = E[\mathbf{w}\mathbf{w}^T] = E\left[\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \begin{pmatrix} \mathbf{u}^T & \mathbf{v}^T \end{pmatrix} \right] = \begin{bmatrix} \mathbf{R}_{uu} & \mathbf{R}_{uv} \\ \mathbf{R}_{vu} & \mathbf{R}_{vv} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{\Sigma}_c \\ \mathbf{\Sigma}_c^T & \mathbf{I} \end{bmatrix}. \tag{2.10}$$

We refer to the elements of $\mathbf{u} = [u_i]_{i=1}^m \in \mathbb{R}^m$ as the *canonical coordinates* of $\mathbf{x}$ and to the elements of $\mathbf{v} = [v_i]_{i=1}^n \in \mathbb{R}^n$ as the canonical coordinates of $\mathbf{y}$. The diagonal cross-correlation matrix $\mathbf{\Sigma}_c$,

$$\mathbf{\Sigma}_c = E[\mathbf{u}\mathbf{v}^T] = E[(\mathbf{F}_c^T\mathbf{R}_{xx}^{-1/2}\mathbf{x})(\mathbf{G}_c^T\mathbf{R}_{yy}^{-1/2}\mathbf{y})^T] = \mathbf{F}_c^T\mathbf{C}\mathbf{G}_c \tag{2.11}$$

is called the *canonical correlation matrix* of *canonical correlations* σ_i, with $1 \geq \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m > 0$. Correspondingly, $\mathbf{\Sigma}_c\mathbf{\Sigma}_c^T$ is the squared canonical correlation

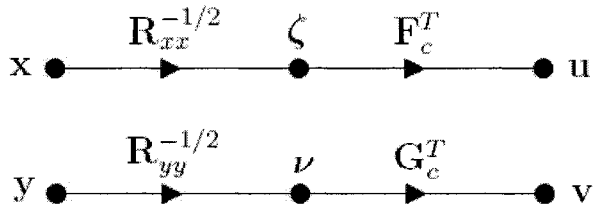

Figure 2.1: Transformation from standard coordinates $\mathbf{x}$ and $\mathbf{y}$ to canonical coordinates $\mathbf{u}$ and $\mathbf{v}$.

matrix of squared canonical correlations σ_i^2. Thus, the canonical correlations measure the correlations between pairs of corresponding canonical coordinates. That is, $E[u_i v_j] = \sigma_i \delta_{ij}$; $i \in [1, m]$, $j \in [1, n]$, with δ_{ij} being the Kronecker delta. The canonical correlations σ_i are also the singular values of the coherence matrix $\mathbf{C}$. Since $\mathbf{F}_c$ and $\mathbf{G}_c$ are orthogonal matrices, we may write the squared coherence matrix $\mathbf{C}\mathbf{C}^T$ as

$$\begin{aligned} \mathbf{C}\mathbf{C}^T &= \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-T/2} \\ &= \mathbf{F}_c\mathbf{\Sigma}_c\mathbf{G}_c^T\mathbf{G}_c\mathbf{\Sigma}_c^T\mathbf{F}_c^T = \mathbf{F}_c\mathbf{\Sigma}_c\mathbf{\Sigma}_c^T\mathbf{F}_c^T. \end{aligned} \tag{2.12}$$

This shows that the squared canonical correlations σ_i^2 are the eigenvalues of the squared coherence matrix $\mathbf{C}\mathbf{C}^T$, or equivalently, of the matrix $\mathbf{R}_{xx}^{-T/2}\mathbf{C}\mathbf{C}^T\mathbf{R}_{xx}^{T/2} = \mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}$. It is interesting to note that these eigenvalues are invariant to the choice of a square-root for $\mathbf{R}_{xx}$.

Figure 2.1 illustrates the transformation from standard coordinates $\mathbf{x}$ and $\mathbf{y}$ to coherence coordinates ζ and ν and then to canonical coordinates $\mathbf{u}$ and $\mathbf{v}$.

2.3 Geometry of Canonical Coordinates

The canonical correlations σ_i are invariant to block-diagonal transformations of $\mathbf{R}_{zz}$ of form

$$\mathbf{T}\mathbf{R}_{zz}\mathbf{T}^T = \begin{bmatrix} \mathbf{T}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xy} \\ \mathbf{R}_{yx} & \mathbf{R}_{yy} \end{bmatrix} \begin{bmatrix} \mathbf{T}_1^T & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2^T \end{bmatrix}, \tag{2.13}$$

where $\mathbf{T}_1 \in \mathbb{R}^{m \times m}$ and $\mathbf{T}_2 \in \mathbb{R}^{n \times n}$ are nonsingular matrices [5]. In other words, the canonical correlations σ_i are invariant under nonsingular transformation of $\mathbf{x}$ by $\mathbf{T}_1\mathbf{x}$

and $\mathbf{y}$ by $\mathbf{T}_2\mathbf{y}$. This may easily be proved by showing that the coherence matrix of $\mathbf{T}_1\mathbf{x}$ and $\mathbf{T}_2\mathbf{y}$ is the same as the coherence matrix of $\mathbf{x}$ and $\mathbf{y}$.

In fact, the canonical correlations σ_i form a *complete* or *maximal* set of invariants [5] for the composite covariance matrix $\mathbf{R}_{zz} = E[\mathbf{z}\mathbf{z}^T]$, under the linear transformation group

$$\mathcal{T} = \left\{ \mathbf{T} = \begin{bmatrix} \mathbf{T}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 \end{bmatrix}, \det\{\mathbf{T}\} \neq 0 \right\}, \tag{2.14}$$

with group action $\mathbf{R}_{zz} \rightarrow \mathbf{T}\mathbf{R}_{zz}\mathbf{T}^T$ [5]. That is, any function of $\mathbf{R}_{zz}$ that is invariant under the transformation group $\mathcal{T}$ is a function of $\mathbf{\Sigma}_c$ [5]. This is the reason that the correlations σ_i and coordinates $\mathbf{u} = [u_i]_{i=1}^m$ and $\mathbf{v} = [v_i]_{i=1}^n$ are called canonical.

The ith canonical correlation $\sigma_i = E[u_i v_i]$ measures the cosine of the angle between u_i, the ith canonical coordinate of $\mathbf{x}$, and v_i, the ith canonical coordinate of $\mathbf{y}$. The angle between u_i and v_i plays the same role as a principal angle between two linear subspaces, but in the Hilbert space of second-order random variables instead of a Euclidean space. A detailed proof for this claim is presented in [5]. Nonetheless, considering the Euclidean case here can be helpful to develop intuition.

Let $\mathbf{M} \in \mathbb{R}^{(m+n)\times m}$ and $\mathbf{N} \in \mathbb{R}^{(m+n)\times n}$ be orthonormal bases for $m-$ and $n-$dimensional subspaces of $\mathbb{R}^{(m+n)}$, then the cosines of principal angles between $< \mathbf{M} >$ and $< \mathbf{N} >$ are measured by the singular values of the matrix $\mathbf{M}^T\mathbf{N}$ [57]:

$$\mathbf{M}^T\mathbf{N} = \tilde{\mathbf{F}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{G}}^T \quad \text{or} \quad (\mathbf{M}\tilde{\mathbf{F}})^T(\mathbf{N}\tilde{\mathbf{G}}) = \tilde{\mathbf{\Sigma}}. \tag{2.15}$$

That is, the principal angles between $< \mathbf{M} >$ and $< \mathbf{N} >$ are the angles between the rotated columns of $\mathbf{M}$ and $\mathbf{N}$, with the rotations implemented by $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{G}}$. Here $\tilde{\mathbf{F}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{G}}^T$ is an SVD of $\mathbf{M}^T\mathbf{N}$. This may be viewed as the Euclidean space analog of

$$\begin{aligned} &E[\boldsymbol{\zeta}\boldsymbol{\nu}^T] = E[(\mathbf{R}_{xx}^{-1/2}\mathbf{x})(\mathbf{R}_{yy}^{-1/2}\mathbf{y})^T] = \mathbf{F}_c\mathbf{\Sigma}_c\mathbf{G}_c^T \text{ or} \\ &E[(\mathbf{F}_c^T\boldsymbol{\zeta})(\mathbf{G}_c\boldsymbol{\nu})^T] = E[(\mathbf{F}_c^T\mathbf{R}_{xx}^{-1/2}\mathbf{x})(\mathbf{G}_c^T\mathbf{R}_{yy}^{-1/2}\mathbf{y})^T], = \mathbf{\Sigma}_c \end{aligned} \tag{2.16}$$

where the elements of $\mathbf{u} = \mathbf{F}_c^T\mathbf{R}_{xx}^{-1/2}\mathbf{x}$ and $\mathbf{v} = \mathbf{G}_c^T\mathbf{R}_{yy}^{-1/2}\mathbf{y}$ are orthonormal bases, with respect to the inner product $E[u_i v_j]$, for $m-$ and $n-$dimensional subspaces

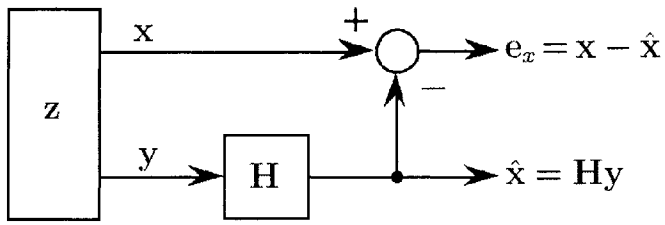

Figure 2.2: Filtering problem.

of the Hilbert space of second-order random variables. In other words, the inner product $E[(\mathbf{R}_{xx}^{-1/2}\mathbf{x})(\mathbf{R}_{yy}^{-1/2}\mathbf{y})^T]$, between the white random variables in $\mathbf{R}_{xx}^{-1/2}\mathbf{x}$ and $\mathbf{R}_{yy}^{-1/2}\mathbf{y}$, is the Hilbert space analog of the Euclidean space inner product $\mathbf{M}^T\mathbf{N}$, between orthonormal vectors in $\mathbf{M}$ and $\mathbf{N}$. Therefore, the ith canonical correlation σ_i measures the cosine of the ith principal angle between u_i and v_i, or equivalently the ith principal angle between the linear subspaces spanned by the canonical coordinates $\mathbf{u} = \mathbf{F}_c^T\mathbf{R}_{xx}^{-1/2}\mathbf{x}$ and $\mathbf{v} = \mathbf{G}_c^T\mathbf{R}_{yy}^{-1/2}\mathbf{y}$. We note that these cosines are invariant to nonsingular transformation of $\mathbf{x}$ by $\mathbf{T}_1\mathbf{x}$ and $\mathbf{y}$ by $\mathbf{T}_2\mathbf{y}$.

2.4 Wiener Filtering in Canonical Coordinates

Let us now consider the filtering problem illustrated in Figure 2.2. In this picture, the linear minimum mean squared error (MMSE) estimator of $\mathbf{x}$ from $\mathbf{y}$ is denoted by $\hat{\mathbf{x}} = \mathbf{H}\mathbf{y}$ and the corresponding (orthogonal) error is $\mathbf{e}_x = \mathbf{x} - \hat{\mathbf{x}}$. In the standard coordinates, the Wiener filter $\mathbf{H}$ and the error covariance matrix $\mathbf{Q}_{xx} = E[\mathbf{e}_x\mathbf{e}_x^T]$ are given by

$$\begin{aligned} \mathbf{H} &= \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} \\ \mathbf{Q}_{xx} &= E\{(\mathbf{x} - \hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}})^T\} = \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}. \end{aligned} \tag{2.17}$$

Using the SVD in (2.6), the Wiener filter and the error covariance matrix, in canonical coordinates, may be written as

$$\mathbf{H} = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} = \mathbf{R}_{xx}^{1/2}\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}\mathbf{R}_{yy}^{-1/2} = \mathbf{R}_{xx}^{1/2}\mathbf{F}_c\mathbf{\Sigma}_c\mathbf{G}_c^T\mathbf{R}_{yy}^{-1/2} \tag{2.18}$$

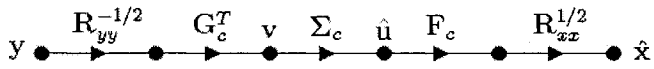

Figure 2.3: The decomposition of the Wiener filter in canonical coordinates.

and

$$\begin{aligned}\mathbf{Q}_{xx} &= \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx} \\ &= \mathbf{R}_{xx}^{1/2}(\mathbf{I} - \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-T/2})\mathbf{R}_{xx}^{T/2} \\ &= \mathbf{R}_{xx}^{1/2}(\mathbf{I} - \mathbf{F}_c\boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_c^T\mathbf{F}_c^T)\mathbf{R}_{xx}^{T/2} \\ &= \mathbf{R}_{xx}^{1/2}\mathbf{F}_c(\mathbf{I} - \boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_c^T)\mathbf{F}_c^T\mathbf{R}_{xx}^{T/2}.\end{aligned} \tag{2.19}$$

Figure 2.3 illustrates this decomposition of the Wiener filter in canonical coordinates. The first stage whitens the data vector $\mathbf{y}$ to produce the corresponding coherence coordinates $\boldsymbol{\nu}$, the second stage transforms the coherence coordinates $\boldsymbol{\nu}$ into the canonical coordinates $\mathbf{v}$, and the third stage filters $\mathbf{v}$ with the *canonical Wiener filter* $\boldsymbol{\Sigma}_c$ to produce the linear MMSE estimator $\hat{\mathbf{u}} = \boldsymbol{\Sigma}_c\mathbf{v}$ of the canonical coordinates of $\mathbf{x}$, with the error covariance matrix

$$\mathbf{Q}_{uu} = E[\mathbf{e}_u\mathbf{e}_u^T] = E[(\mathbf{u} - \hat{\mathbf{u}})(\mathbf{u} - \hat{\mathbf{u}})^T] = \mathbf{I} - \boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_c^T. \tag{2.20}$$

This is indeed the linear MMSE estimator of $\mathbf{u}$ from $\mathbf{v}$, as the composite covariance matrix of $\mathbf{u}$ and $\mathbf{v}$ is of form (2.10). The fourth stage transforms $\hat{\mathbf{u}}$ into the linear MMSE estimator of the coherence coordinates $\boldsymbol{\zeta}$, and the fifth stage re-colors them to produce $\hat{\mathbf{x}}$, the linear MMSE estimator of $\mathbf{x}$ from $\mathbf{y}$.

The concentration ellipse [8] for the filtering error $\mathbf{e}_x = \mathbf{x} - \hat{\mathbf{x}}$, with covariance matrix $\mathbf{Q}_{xx} = E[\mathbf{e}_x\mathbf{e}_x^T]$, has a volume proportional to

$$\begin{aligned}\det\{\mathbf{Q}_{xx}\} &= \det\{\mathbf{R}_{xx}^{1/2}\mathbf{F}_c(\mathbf{I} - \boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_c^T)\mathbf{F}_c^T\mathbf{R}_{xx}^{T/2}\} \\ &= \det\{\mathbf{R}_{xx}\}\det\{\mathbf{I} - \boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_c^T\}.\end{aligned} \tag{2.21}$$

The above identity follows from the decomposition of $\mathbf{Q}_{xx}$ in (2.19) and the fact that $\mathbf{F}_c$ is an orthogonal matrix. Similarly, the concentration ellipse for the data vector $\mathbf{x}$

has a volume proportional to $\det\{\mathbf{R}_{xx}\}$. Therefore, the ratio of these determinants measures the *relative volume* of the concentration ellipses. This ratio is the same as the ratio of the volume of the concentration ellipse of the filtering error $\mathbf{e}_u = \mathbf{u} - \hat{\mathbf{u}}$ to the volume of the concentration ellipse of the canonical coordinates $\mathbf{u}$. That is,

$$\begin{aligned} \frac{\det\{\mathbf{Q}_{xx}\}}{\det\{\mathbf{R}_{xx}\}} &= \det\{\mathbf{I} - \mathbf{\Sigma}_c \mathbf{\Sigma}_c^T\} \\ &= \prod_{i=1}^{m}(1-\sigma_i^2) = \frac{\det\{\mathbf{Q}_{uu}\}}{\det\{\mathbf{R}_{uu}\}}, \end{aligned} \tag{2.22}$$

where the last identity follows from (2.20), recalling that $\mathbf{R}_{uu} = \mathbf{I}$. The fact that the relative volumes in standard coordinates and canonical coordinates are the same fits our intuition, as the relative volume depends only on the canonical correlations and hence are invariant to linear transformations.

A physical interpretation for the decompositions in (2.20) and (2.22) is that the canonical coordinate transformation replaces the original composite data vector $\mathbf{z}$ by a parallel combination of uncorrelated random variables, each of whose error covariance is $1-\sigma_i^2$. The error covariance for the parallel combination is $\mathrm{diag}(1-\sigma_1^2, \ldots, 1-\sigma_m^2)$, with determinant $\prod_{i=1}^{m}(1-\sigma_i^2)$.

2.5 Linear Dependence and Coherence

The standard measure of linear dependence for the composite data vector $\mathbf{z} = [\mathbf{x}^T \quad \mathbf{y}^T]^T$ is the Hadamard ratio, inside the inequality

$$0 \leq \frac{\det\{\mathbf{R}_{zz}\}}{\prod_{i=1}^{m+n}[\mathbf{R}_{zz}]_{ii}} \leq 1, \tag{2.23}$$

where $[\mathbf{R}_{zz}]_{ii}$'s, $i \in [1, m+n]$ are the diagonal elements of $\mathbf{R}_{zz}$. This ratio takes the value 0 iff there is linear dependence among elements of $\mathbf{z}$; it takes the value 1 iff the elements of $\mathbf{z}$ are mutually uncorrelated.

By introducing a block Cholesky factorization for $\mathbf{R}_{zz}$ of the form

$$\mathbf{R}_{zz} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xy} \\ \mathbf{R}_{yx} & \mathbf{R}_{yy} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{yy} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{R}_{yy}^{-1}\mathbf{R}_{yx} & \mathbf{I} \end{bmatrix}, \tag{2.24}$$

we may write $\det\{\mathbf{R}_{zz}\}$ as

$$\begin{aligned}\det\{\mathbf{R}_{zz}\} &= \det\{\mathbf{Q}_{xx}\}\det\{\mathbf{R}_{yy}\} \\ &= \det\{\mathbf{R}_{xx}\}\tfrac{\det\{\mathbf{Q}_{xx}\}}{\det\{\mathbf{R}_{xx}\}}\det\{\mathbf{R}_{yy}\},\end{aligned} \tag{2.25}$$

yielding the following decomposition of the Hadamard ratio:

$$\frac{\det\{\mathbf{R}_{zz}\}}{\prod_{i=1}^{m+n}[\mathbf{R}_{zz}]_{ii}} = \frac{\det\{\mathbf{R}_{xx}\}}{\prod_{i=1}^{m}[\mathbf{R}_{xx}]_{ii}}\det\{\mathbf{I}-\boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_c^T\}\frac{\det\{\mathbf{R}_{yy}\}}{\prod_{i=1}^{n}[\mathbf{R}_{yy}]_{ii}}. \tag{2.26}$$

The first and third terms on the right hand side of (2.26) measure the linear dependence *among* the elements of $\mathbf{x}$ and $\mathbf{y}$, respectively, while the middle term,

$$L = \det(\mathbf{I}-\boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_c^T) = \prod_{i=1}^{m}(1-\sigma_i^2); \quad 0 \le L \le 1, \tag{2.27}$$

measures the linear dependence *between* the elements of $\mathbf{x}$ and $\mathbf{y}$. The measure L takes the value 0 iff there is linear dependence between elements of $\mathbf{x}$ and $\mathbf{y}$; it takes the value 1 iff the elements of $\mathbf{x}$ and $\mathbf{y}$ are mutually uncorrelated. The ith term of the product on the right hand side of (2.27), i.e. $(1-\sigma_i^2)$, measures the linear dependence between the ith canonical coordinate of $\mathbf{x}$ and the ith canonical coordinate of $\mathbf{y}$. This implies that the linear dependence between $\mathbf{x}$ and $\mathbf{y}$ is decomposed into the linear dependence between their canonical coordinates, and is measured only by their canonical correlations or principal cosines.

Correspondingly, we may define the coherence measure between the elements of $\mathbf{x}$ and $\mathbf{y}$ as

$$H = 1 - L = 1 - \det(\mathbf{I}-\boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_c^T) = 1 - \prod_{i=1}^{m}(1-\sigma_i^2); \quad 0 \le H \le 1. \tag{2.28}$$

The elements of $\mathbf{x}$ and $\mathbf{y}$ are perfectly coherent iff $H = 1$; they are mutually non-coherent iff $H = 0$.

2.6 Information Rate and Mutual Information

We now determine the rate at which the $\mathbf{x}-$channel carries information about the $\mathbf{y}-$channel, and vice versa, or simply the mutual information between $\mathbf{x}$ and $\mathbf{y}$. According to Shannon [58], the information rate for the composite data vector $\mathbf{z} =$

$[\mathbf{x}^T \ \mathbf{y}^T]^T$ is defined as

$$R = H_x + H_y - H_z, \tag{2.29}$$

where H_x, H_y, and H_z represent the entropies of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, respectively. In communications, H_x is the entropy of the message $\mathbf{x}$, H_y is the entropy of the measurement $\mathbf{y}$, and H_z is the shared entropy.

For Gaussian composite data vector $\mathbf{z} : N(\mathbf{0}, \mathbf{R}_{zz})$, with distribution function $f(\mathbf{z})$, the entropy is

$$H_z = E[\log f(\mathbf{z})] = \frac{m+n}{2}\log(2\pi e) + \frac{1}{2}\log\det\{\mathbf{R}_{zz}\}, \tag{2.30}$$

Using H_z in (2.30) and similar expressions for H_x and H_y, we may obtain the information rate R as

$$\begin{aligned} R &= \tfrac{m}{2}\log(2\pi e) + \tfrac{1}{2}\log\det\{\mathbf{R}_{xx}\} + \tfrac{n}{2}\log(2\pi e) + \tfrac{1}{2}\log\det\{\mathbf{R}_{yy}\} \\ &\quad -\tfrac{m+n}{2}\log(2\pi e) - \tfrac{1}{2}\log\det\{\mathbf{R}_{zz}\} \\ &= \tfrac{1}{2}\log\det\{\mathbf{R}_{xx}\} + \tfrac{1}{2}\log\det\{\mathbf{R}_{yy}\} - \tfrac{1}{2}\log\det\{\mathbf{R}_{zz}\}. \end{aligned} \tag{2.31}$$

Using (2.21) and (2.25) for $\det\{\mathbf{Q}_{xx}\}$ and $\det\{\mathbf{R}_{zz}\}$, we may write the information rate as

$$\begin{aligned} R &= \tfrac{1}{2}\log\det\{\mathbf{R}_{xx}\} - \tfrac{1}{2}\log\det\{\mathbf{Q}_{xx}\} \\ &= \tfrac{1}{2}\log\det\{\mathbf{R}_{xx}\} - \tfrac{1}{2}\log\det\{\mathbf{R}_{xx}\} - \tfrac{1}{2}\log\det\{\mathbf{I} - \mathbf{\Sigma}_c\mathbf{\Sigma}_c^T\} \\ &= -\tfrac{1}{2}\log\det\{\mathbf{I} - \mathbf{\Sigma}_c\mathbf{\Sigma}_c^T\} = \tfrac{1}{2}\textstyle\sum_{i=1}^{m}\log\frac{1}{1-\sigma_i^2}. \end{aligned} \tag{2.32}$$

The ith term in the above summation, i.e.

$$R_i = \frac{1}{2}\log\frac{1}{1-\sigma_i^2}, \tag{2.33}$$

is the rate at which the ith canonical coordinate of $\mathbf{y}$, i.e. v_i, brings information about the ith canonical coordinate of $\mathbf{x}$, i.e. u_i. Thus, we further refer to the R_i as canonical rates. This result implies that the rate at which the $\mathbf{y}-$channel brings

information about the $\mathbf{x}-$channel is just the sum of the canonical rates at which the canonical coordinates of $\mathbf{y}$ carry information about the canonical coordinates of $\mathbf{x}$:

$$R = \sum_{i=1}^{m} R_i = \frac{1}{2} \sum_{i=1}^{m} \log \frac{1}{1-\sigma_i^2}. \tag{2.34}$$

In communications, a physical interpretation of this result is that the transformation to canonical coordinates transforms the Gaussian channel into a parallel combination of independent Gaussian channels, each of which has canonical rate R_i. Consequently, the total information rate is the sum of these canonical rates. As the total information between the canonical coordinates is determined solely by the canonical correlations or principal cosines, it is invariant to linear transformations, and thus is the same as the rate for the original channels.

2.7 Half-Canonical Coordinates

Another canonical coordinate system that will be used in this thesis is the half-canonical coordinate system [8], [10]. This coordinate system, as will be shown in Chapter 4, is important for reduced-rank estimation where the objective of estimation is to minimize the mean-squared error (MSE). In this section, we review half-canonical coordinates and half-canonical correlations in order to prepare the readers for the forthcoming developments in the subsequent chapters.

In contrast to (2.3), where both $\mathbf{x}$ and $\mathbf{y}$ are whitened, let us only whiten $\mathbf{y}$. Then, the composite vector $\boldsymbol{\xi}$ and its covariance matrix $\mathbf{R}_{\xi\xi}$ become

$$\boldsymbol{\xi} = \begin{bmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{yy}^{-1/2} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \tag{2.35}$$

and

$$\mathbf{R}_{\xi\xi} = E[\boldsymbol{\xi}\,\boldsymbol{\xi}^T] = E\left[\begin{pmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\nu} \end{pmatrix} \begin{pmatrix} \boldsymbol{\zeta}^T & \boldsymbol{\nu}^T \end{pmatrix} \right] = \begin{bmatrix} \mathbf{R}_{\zeta\zeta} & \mathbf{R}_{\zeta\nu} \\ \mathbf{R}_{\nu\zeta} & \mathbf{R}_{\nu\nu} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{C}_h \\ \mathbf{C}_h^T & \mathbf{I} \end{bmatrix}, \tag{2.36}$$

where

$$\mathbf{C}_h = E[\boldsymbol{\zeta}\boldsymbol{\nu}^T] = E[\mathbf{x}(\mathbf{R}_{yy}^{-1/2}\mathbf{y})^T] = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} \tag{2.37}$$

is called the *half-coherence matrix* of $\mathbf{x}$ and $\mathbf{y}$. Therefore, the half-coherence matrix $\mathbf{C}_h$ is the cross-covariance matrix between $\mathbf{x}$ and the whitened version of $\mathbf{y}$.

We now determine the SVD of the half-coherence matrix, namely

$$\begin{aligned} \mathbf{C}_h = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} = \mathbf{U}_h\boldsymbol{\Sigma}_h\mathbf{V}_h^T \quad \text{and} \\ \mathbf{U}_h^T\mathbf{C}_h\mathbf{V}_h = \mathbf{U}_h^T\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}\mathbf{V}_h = \boldsymbol{\Sigma}_h, \end{aligned} \tag{2.38}$$

where $\mathbf{U}_h \in \mathbb{R}^{m\times m}$ and $\mathbf{V}_h \in \mathbb{R}^{n\times n}$ are orthogonal matrices, i.e.

$$\mathbf{U}_h^T\mathbf{U}_h = \mathbf{U}_h\mathbf{U}_h^T = \mathbf{I}(m) \quad \text{and} \quad \mathbf{V}_h^T\mathbf{V}_h = \mathbf{V}_h\mathbf{V}_h^T = \mathbf{I}(n), \tag{2.39}$$

and

$$\boldsymbol{\Sigma}_h = [\ \boldsymbol{\Sigma}_h(m) \quad \mathbf{0}\ ] \in \mathbb{R}^{m\times n} \tag{2.40}$$

is a diagonal singular value matrix, with $\boldsymbol{\Sigma}_h(m) = \text{diag}[\sigma_1, \sigma_2, \ldots, \sigma_m]$ and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m > 0$.

Transforming the composite vector $\boldsymbol{\xi}$ with the orthogonal matrices $\mathbf{U}_h$ and $\mathbf{V}_h$ yields the half-canonical composite vector $\mathbf{w}$,

$$\mathbf{w} = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_h^T & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_h^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_h^T & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_h^T \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{yy}^{-1/2} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \tag{2.41}$$

with the covariance matrix $\mathbf{R}_{ww}$,

$$\begin{aligned} \mathbf{R}_{ww} = E[\mathbf{w}\mathbf{w}^T] &= E\left[\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \begin{pmatrix} \mathbf{u}^T & \mathbf{v}^T \end{pmatrix} \right] \\ &= \begin{bmatrix} \mathbf{R}_{uu} & \mathbf{R}_{uv} \\ \mathbf{R}_{vu} & \mathbf{R}_{vv} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_h^T\mathbf{R}_{xx}\mathbf{U}_h & \boldsymbol{\Sigma}_h \\ \boldsymbol{\Sigma}_h^T & \mathbf{I} \end{bmatrix}. \end{aligned} \tag{2.42}$$

We refer to the elements of $\mathbf{u} = [u_i]_{i=1}^m \in \mathbb{R}^m$ as the *half-canonical coordinates* of $\mathbf{x}$ and to the elements of $\mathbf{v} = [v_i]_{i=1}^n \in \mathbb{R}^n$ as the half-canonical coordinates of $\mathbf{y}$.

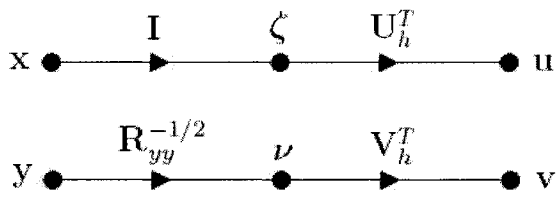

Figure 2.4: Transformation from standard coordinates $\mathbf{x}$ and $\mathbf{y}$ to half-canonical coordinates $\mathbf{u}$ and $\mathbf{v}$.

Correspondingly, the diagonal cross-correlation matrix $\mathbf{\Sigma}_h$,

$$\mathbf{\Sigma}_h = E[\mathbf{u}\mathbf{v}^T] = E[(\mathbf{U}_h^T\mathbf{x})(\mathbf{V}_h^T\mathbf{R}_{yy}^{-1/2}\mathbf{y})^T] = \mathbf{U}_h^T\mathbf{C}_h\mathbf{V}_h \tag{2.43}$$

is called the *half-canonical correlation matrix* of *half-canonical correlations* σ_i, with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m > 0$. Figure 2.4 illustrates the transformation from standard coordinates $\mathbf{x}$ and $\mathbf{y}$ to half-canonical coordinates $\mathbf{u}$ and $\mathbf{v}$.

The half-canonical correlations σ_i are invariant to block-diagonal transformations of $\mathbf{R}_{zz}$ of form (2.13), with $\mathbf{T}_1 = \mathbf{I}$. In other words, the half-canonical correlations σ_i are invariant under nonsingular transformations of $\mathbf{y}$. This may easily be proved by showing that the half-coherence matrix of $\mathbf{x}$ and $\mathbf{T}_2\mathbf{y}$ is the same as the half-coherence matrix of $\mathbf{x}$ and $\mathbf{y}$.

In fact, the half-canonical correlations σ_i form a complete or maximal set of invariants for the composite covariance matrix $\mathbf{R}_{zz} = E[\mathbf{z}\mathbf{z}^T]$, under the linear transformation group of (2.14), with $\mathbf{T}_1 = \mathbf{I}$ and group action $\mathbf{R}_{zz} \rightarrow \mathbf{T}\mathbf{R}_{zz}\mathbf{T}^T$. That is, any function of $\mathbf{R}_{zz}$ that is invariant under the nonsingular transformations of $\mathbf{y}$ is a function of the half-canonical correlation matrix $\mathbf{\Sigma}_h$. This maximal invariance property may easily be proved using a similar proof as that presented in [5] for canonical coordinates. The only difference is that $\mathbf{T}_1$ has to be set to identity. This is the reason that the correlations σ_i and coordinates $\mathbf{u} = [u_i]_{i=1}^m$ and $\mathbf{v} = [v_i]_{i=1}^n$ are called half-canonical.

2.8 Conclusions

In this chapter, canonical coordinates and canonical correlations were introduced and their algebraic and geometric properties were reviewed. It was shown that canonical correlations measure the cosine of principal angles between two linear subspaces in the Hilbert space of second-order random variables. Further, they form a maximal set of invariants for the composite covariance matrix of two-channel data.

Evidently, the canonical coordinate system is the right coordinate system for analyzing second-order filtering and communication over the Gaussian channel. In this coordinate system, the volume of the concentration ellipse is multiplicatively decomposed into the product of the volumes of (canonical) concentration ellipses, and is determined only by the canonical correlations. Additionally, the linear dependence between the channels is decomposed into the product of the linear dependence between the canonical coordinates of the channels, each of which is determined solely by the corresponding canonical correlation. The information rate between the channels is additively decomposed into a sum of canonical rates, each of which measures the rate at which a canonical coordinate of one channel carries information about its corresponding canonical coordinate of the other channel. Furthermore, each canonical rate depends only on the principal cosine between the corresponding pair of canonical coordinates. We also reviewed the half-canonical coordinates and half-canonical correlations to prepare the readers for the developments and discussions in subsequent chapters.

To conclude, the review in this chapter showed that all performance measures of interest for second-order inference and Gaussian communications are determined by the canonical correlations or principal cosines of the two-channel data. Additionally, these performance measures are invariant to uncoupled nonsingular transformations of the channels.

CHAPTER 3

TWO-CHANNEL CONSTRAINED LEAST SQUARES PROBLEMS AND CONNECTIONS WITH CANONICAL COORDINATES

3.1 Introduction

In this chapter, we consider two-channel constrained least squares (CLS) problems, where the objective is to estimate a linear function of elements in one channel from a linear combination of elements in the other, under certain constraints. By imposing various constraints we aim to derive a general set of solutions to the two-channel CLS problem, and at the same time clarify the connections between two-channel CLS filtering and various canonical coordinate systems. The material presented in this chapter are also reported in [13].

Referring to Figure 3.1, the most general form of a two-channel least squares problem one may consider may be posed as $\min_{\mathbf{D}_x,\mathbf{D}_y} J$, where J is the quadratic function

$$J = \mathrm{tr}\{E[(\mathbf{D}_x^T\mathbf{x} - \mathbf{D}_y^T\mathbf{y})(\mathbf{D}_x^T\mathbf{x} - \mathbf{D}_y^T\mathbf{y})^T]\}. \tag{3.1}$$

The matrix $\mathbf{D}_x^T \in \mathbb{R}^{m\times m}$ is a linear map that transforms $\mathbf{x} \in \mathbb{R}^m$ to $\mathbf{u} = \mathbf{D}_x^T\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{D}_y^T \in \mathbb{R}^{m\times n}$ is a linear map that transforms $\mathbf{y} \in \mathbb{R}^n$ to $\mathbf{v} = \mathbf{D}_y^T\mathbf{y} \in \mathbb{R}^m$, $m \leq n$. Without any constraints on $\mathbf{D}_x$ and $\mathbf{D}_y$, the optimization problem of (3.1) yields the trivial solution $\mathbf{D}_x = \mathbf{0}$, $\mathbf{D}_y = \mathbf{0}$. However, we shall show in this chapter

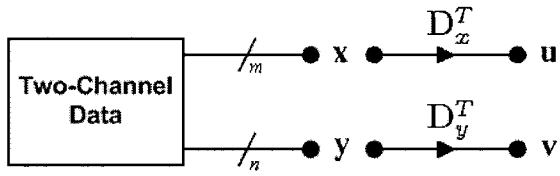

Figure 3.1: Two-channel filtering problem.

that under special constraints the optimization problem in (3.1) leads to canonical coordinate decompositions [1]– [7], half-canonical coordinate decompositions [8], [10], or programmable canonical correlation analysis (PCCA) coordinate decompositions [14]– [17]. The PCCA coordinate system was proposed in [14]– [17] for adaptive source separation. In fact, the work in [14]– [17] motivated us to develop various two-channel constrained least squares problems and establish their connections with canonical coordinate systems. Later in this thesis, we use the results of this chapter to develop simple methods for recursive computation of canonical coordinates, half-canonical coordinates, and also to establish connections between two-channel CLS filters and reduced-rank Wiener filters [6]– [11].

3.2 Two-Channel CLS Problems and Solutions

Consider the two-channel problem of Figure 3.1, with two random vectors, $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$, and linear maps $\mathbf{D}_x \in \mathbb{R}^{m \times m}$ and $\mathbf{D}_y \in \mathbb{R}^{n \times m}$, $m \leq n$. If $m > n$, then we would simply reverse the roles of $\mathbf{x}$ and $\mathbf{y}$. We assume that $\mathbf{x}$ and $\mathbf{y}$ have zero means and share the composite covariance matrix in (2.2).

The two-channel CLS problem may be defined as $\min_{\mathbf{D}_x, \mathbf{D}_y} J$, subject to constraints on $\mathbf{D}_x$ and $\mathbf{D}_y$, where J is the scalar objective function

$$\begin{aligned} J &= E[\|\mathbf{D}_x^T\mathbf{x} - \mathbf{D}_y^T\mathbf{y}\|^2] \\ &= \mathrm{tr}\{E[(\mathbf{D}_x^T\mathbf{x} - \mathbf{D}_y^T\mathbf{y})(\mathbf{D}_x^T\mathbf{x} - \mathbf{D}_y^T\mathbf{y})^T]\} \\ &= \mathrm{tr}\{\mathbf{D}_x^T\mathbf{R}_{xx}\mathbf{D}_x - \mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y - \mathbf{D}_y^T\mathbf{R}_{yx}\mathbf{D}_x + \mathbf{D}_y^T\mathbf{R}_{yy}\mathbf{D}_y\}. \end{aligned} \tag{3.2}$$

Here, the matrices $\mathbf{D}_x \in \mathbb{R}^{m\times m}$ and $\mathbf{D}_y \in \mathbb{R}^{n\times m}$ have equal column dimensions, $\mathrm{tr}\{\cdot\}$ denotes trace of a matrix, and $E[\cdot]$ denotes statistical expectation.

Alternatively, we may rewrite the objective function J as

$$J = \mathrm{tr}\{\mathbf{D}_x^T\mathbf{Q}_{xx}\mathbf{D}_x + (\mathbf{R}_{yy}\mathbf{D}_y - \mathbf{R}_{yx}\mathbf{D}_x)^T\mathbf{R}_{yy}^{-1}(\mathbf{R}_{yy}\mathbf{D}_y - \mathbf{R}_{yx}\mathbf{D}_x)\}, \tag{3.3}$$

where $\mathbf{Q}_{xx} = \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}$ is the Schur complement of $\mathbf{R}_{xx}$, or as

$$J = \mathrm{tr}\{\mathbf{D}_y^T\mathbf{Q}_{yy}\mathbf{D}_y + (\mathbf{R}_{xx}\mathbf{D}_x - \mathbf{R}_{xy}\mathbf{D}_y)^T\mathbf{R}_{xx}^{-1}(\mathbf{R}_{xx}\mathbf{D}_x - \mathbf{R}_{xy}\mathbf{D}_y)\}, \tag{3.4}$$

where $\mathbf{Q}_{yy} = \mathbf{R}_{yy} - \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$ is the Schur complement of $\mathbf{R}_{yy}$.

Let us now consider the constraints that relate the two-channel CLS problems to various canonical coordinate systems.

3.2.1 Case 1: Canonical Coordinates

Referring to Figure 3.1, the objective here is to whiten $\mathbf{u} = \mathbf{D}_x^T\mathbf{x}$ and $\mathbf{v} = \mathbf{D}_y^T\mathbf{y}$, and diagonally cross-correlate them, while minimizing J. Thus, the suitable constraints are

$$\begin{gathered}\mathbf{R}_{uu} = \mathbf{D}_x^T\mathbf{R}_{xx}\mathbf{D}_x = \mathbf{I}, \quad \mathbf{R}_{vv} = \mathbf{D}_y^T\mathbf{R}_{yy}\mathbf{D}_y = \mathbf{I}, \quad \text{and} \\ \mathbf{R}_{uv} = \mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y = \mathbf{\Sigma} = \mathrm{diag}[\sigma_1, \sigma_2, \ldots, \sigma_m],\end{gathered} \tag{3.5}$$

where $\mathbf{I}$ is the $m \times m$ identity matrix. The diagonal matrix $\mathbf{\Sigma}$ is not known *a priori*. However, it may be assumed, without loss of generality, that the diagonal elements of $\mathbf{\Sigma}$ are arranged in descending order. That is,

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m > 0. \tag{3.6}$$

Using the method of Lagrange multipliers, the constrained minimization problem may be written as $\min_{\mathbf{D}_x,\mathbf{D}_y} J_C$, where J_C is the scalar objective function

$$\begin{aligned}J_C = \mathrm{tr}\{&\mathbf{D}_x^T\mathbf{Q}_{xx}\mathbf{D}_x + (\mathbf{R}_{yy}\mathbf{D}_y - \mathbf{R}_{yx}\mathbf{D}_x)^T\mathbf{R}_{yy}^{-1} \\ &\times(\mathbf{R}_{yy}\mathbf{D}_y - \mathbf{R}_{yx}\mathbf{D}_x)\} + \mathrm{tr}\{(\mathbf{D}_x^T\mathbf{R}_{xx}\mathbf{D}_x - \mathbf{I})\mathbf{\Lambda}_1\} \\ &+\mathrm{tr}\{(\mathbf{D}_y^T\mathbf{R}_{yy}\mathbf{D}_y - \mathbf{I})\mathbf{\Lambda}_2\} + 2\mathrm{tr}\{(\mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y - \mathbf{\Sigma})\mathbf{\Lambda}_3\}\end{aligned} \tag{3.7}$$

or equivalently as

$$\begin{aligned} J_C = \mathrm{tr}\{&\mathbf{D}_y^T\mathbf{Q}_{yy}\mathbf{D}_y + (\mathbf{R}_{xx}\mathbf{D}_x - \mathbf{R}_{xy}\mathbf{D}_y)^T\mathbf{R}_{xx}^{-1} \\ &\times(\mathbf{R}_{xx}\mathbf{D}_x - \mathbf{R}_{xy}\mathbf{D}_y)\} + \mathrm{tr}\{(\mathbf{D}_x^T\mathbf{R}_{xx}\mathbf{D}_x - \mathbf{I})\mathbf{\Lambda}_1\} \\ &+\mathrm{tr}\{(\mathbf{D}_y^T\mathbf{R}_{yy}\mathbf{D}_y - \mathbf{I})\mathbf{\Lambda}_2\} + 2\mathrm{tr}\{(\mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y - \mathbf{\Sigma})\mathbf{\Lambda}_3\} \end{aligned} \tag{3.8}$$

where $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$, and $\mathbf{\Lambda}_3$ are $m \times m$ Lagrange multipliers. It can easily be verified that the trace constraints in (3.7) and (3.8) indeed impose the actual constraints in (3.5). Taking the derivative of J_C with respect to $\mathbf{D}_x$ and $\mathbf{D}_y$ and setting the results to zero yields the coupled equations

$$\mathbf{R}_{xy}\mathbf{D}_y(\mathbf{I} - \mathbf{\Lambda}_3) = \mathbf{R}_{xx}\mathbf{D}_x(\mathbf{I} + \mathbf{\Lambda}_1) \tag{3.9}$$

$$\mathbf{R}_{yx}\mathbf{D}_x(\mathbf{I} - \mathbf{\Lambda}_3) = \mathbf{R}_{yy}\mathbf{D}_y(\mathbf{I} + \mathbf{\Lambda}_2). \tag{3.10}$$

Pre-multiplying (3.9) by $\mathbf{D}_x^T$ and (3.10) by $\mathbf{D}_y^T$ yields the coupled equations

$$\begin{aligned} \mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y(\mathbf{I} - \mathbf{\Lambda}_3) &= \mathbf{D}_x^T\mathbf{R}_{xx}\mathbf{D}_x(\mathbf{I} + \mathbf{\Lambda}_1) \\ \mathbf{D}_y^T\mathbf{R}_{yx}\mathbf{D}_x(\mathbf{I} - \mathbf{\Lambda}_3) &= \mathbf{D}_y^T\mathbf{R}_{yy}\mathbf{D}_y(\mathbf{I} + \mathbf{\Lambda}_2). \end{aligned} \tag{3.11}$$

At the solution, the Lagrange multipliers force the constraints in (3.5), and thus, (3.11) reduces to

$$\begin{aligned} \mathbf{\Sigma}(\mathbf{I} - \mathbf{\Lambda}_3) &= (\mathbf{I} + \mathbf{\Lambda}_1) \\ \mathbf{\Sigma}(\mathbf{I} - \mathbf{\Lambda}_3) &= (\mathbf{I} + \mathbf{\Lambda}_2) \end{aligned} \tag{3.12}$$

which implies that $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2$, assuming that $\mathbf{\Sigma}$ and $(\mathbf{I} - \mathbf{\Lambda}_3)$ are nonsingular. Using (3.12), we may rewrite (3.9) and (3.10) as the coupled system

$$\begin{aligned} \mathbf{R}_{xy}\mathbf{D}_y &= \mathbf{R}_{xx}\mathbf{D}_x\mathbf{\Sigma} \\ \mathbf{R}_{yx}\mathbf{D}_x &= \mathbf{R}_{yy}\mathbf{D}_y\mathbf{\Sigma}. \end{aligned} \tag{3.13}$$

Equivalently, we may combine these two equations as

$$\begin{aligned} \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{D}_x &= \mathbf{R}_{xx}\mathbf{D}_x\mathbf{\Sigma}^2 \\ \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{D}_y &= \mathbf{R}_{yy}\mathbf{D}_y\mathbf{\Sigma}^2. \end{aligned} \tag{3.14}$$

The set of equations in (3.13) and (3.14) are key results, for they characterize the solutions for $\mathbf{D}_x$ and $\mathbf{D}_y$ to minimize J under the constraints in (3.5). These solutions, in turn, produce the coordinates $\mathbf{u} = \mathbf{D}_x^T\mathbf{x}$ and $\mathbf{v} = \mathbf{D}_y^T\mathbf{y}$, and correlations $\mathbf{\Sigma} = \mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y$ of $\mathbf{x}$ and $\mathbf{y}$. The equations in (3.14) are (symmetric) generalized eigenvalue problems for $\mathbf{D}_x$ and $\mathbf{D}_y$, with the shared eigenvalue matrix $\mathbf{\Sigma}^2$. Therefore, we refer to (3.14) as a "coupled (symmetric) generalized eigenvalue problem". Correspondingly, (3.13) may be viewed as a coupled (asymmetric) generalized eigenvalue problem. In Section 3.3.1, we shall establish that $\mathbf{u} = \mathbf{D}_x^T\mathbf{x}$, $\mathbf{v} = \mathbf{D}_y^T\mathbf{y}$, and $\mathbf{\Sigma} = \text{diag}[\sigma_1, \ldots, \sigma_m] = \mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y$ are, indeed, the standard canonical coordinates and canonical correlations of $\mathbf{x}$ and $\mathbf{y}$, thereby justifying our terminology for this case.[1] In Chapter 5, we shall give an alternating power method for recursively solving (3.13) for columns of $\mathbf{D}_x = [\mathbf{d}_{x,1}, \ldots, \mathbf{d}_{x,m}]$ and $\mathbf{D}_y = [\mathbf{d}_{y,1}, \ldots, \mathbf{d}_{y,m}]$, and diagonal elements of $\mathbf{\Sigma}$, one by one or in groups.

3.2.2 Case 2: Half-Canonical Coordinates

Referring again to Figure 3.1, the objective is now to whiten $\mathbf{v} = \mathbf{D}_y^T\mathbf{y}$ only, and diagonally cross-correlate $\mathbf{u} = \mathbf{D}_x^T\mathbf{x}$ and $\mathbf{v}$, while minimizing J. The relevant constraints in this case are

$$\begin{aligned} &\mathbf{D}_x^T\mathbf{D}_x = \mathbf{I}, \quad \mathbf{R}_{vv} = \mathbf{D}_y^T\mathbf{R}_{yy}\mathbf{D}_y = \mathbf{I}, \quad \text{and} \\ &\mathbf{R}_{uv} = \mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y = \mathbf{\Sigma} = \text{diag}[\sigma_1, \sigma_2, \ldots, \sigma_m]. \end{aligned} \tag{3.15}$$

The diagonal matrix $\mathbf{\Sigma}$ is not known *a priori*. However, similar to Case 1, it is assumed that its diagonal elements satisfy (3.6). Since the matrix $\mathbf{D}_x$ is square and full-rank, it follows that $\mathbf{D}_x\mathbf{D}_x^T = \mathbf{I}$, as well.

---

[1] If instead of the set of constraints in (3.5), we had only constrained the *diagonal* elements of $\mathbf{D}_x^T\mathbf{R}_{xx}\mathbf{D}_x$ and $\mathbf{D}_y^T\mathbf{R}_{yy}\mathbf{D}_y$ to be unity, we could still have obtained the generalized eigenvalue problems in (3.13) and (3.14) for $\mathbf{D}_x$ and $\mathbf{D}_y$. However, solving (3.13) and (3.14) under this new set of constraints would not have guaranteed that $\mathbf{D}_x^T\mathbf{R}_{xx}\mathbf{D}_x = \mathbf{I}$, $\mathbf{D}_y^T\mathbf{R}_{yy}\mathbf{D}_y = \mathbf{I}$, or $\mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y$ diagonal. Consequently, $\mathbf{u} = \mathbf{D}_x^T\mathbf{x}$ and $\mathbf{v} = \mathbf{D}_y^T\mathbf{y}$ would not have been the canonical coordinates of $\mathbf{x}$ and $\mathbf{y}$.

The Lagrange multiplier method for the objective function in (3.2) and constraints in (3.15) yields the coupled equations

$$\mathbf{R}_{xy}\mathbf{D}_y(\mathbf{I}-\mathbf{\Lambda}_3) = \mathbf{R}_{xx}\mathbf{D}_x + \mathbf{D}_x\mathbf{\Lambda}_1 \tag{3.16}$$

$$\mathbf{R}_{yx}\mathbf{D}_x(\mathbf{I}-\mathbf{\Lambda}_3) = \mathbf{R}_{yy}\mathbf{D}_y(\mathbf{I}+\mathbf{\Lambda}_2) \tag{3.17}$$

where $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$, and $\mathbf{\Lambda}_3$ are $m \times m$ Lagrange multipliers. Pre-multiplying (3.16) by $\mathbf{D}_x^T$ and (3.17) by $\mathbf{D}_y^T$, and enforcing the constraints in (3.15) yields

$$\begin{aligned}\mathbf{\Sigma}(\mathbf{I}-\mathbf{\Lambda}_3) &= \mathbf{D}_x{}^T\mathbf{R}_{xx}\mathbf{D}_x + \mathbf{\Lambda}_1 \\ \mathbf{\Sigma}(\mathbf{I}-\mathbf{\Lambda}_3) &= \mathbf{I} + \mathbf{\Lambda}_2.\end{aligned} \tag{3.18}$$

Using these solutions for $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ and assuming that $\mathbf{\Sigma}$ and $(\mathbf{I}-\mathbf{\Lambda}_3)$ are nonsingular, we may rewrite (3.16) and (3.17) as the coupled equations

$$\begin{aligned}\mathbf{R}_{xy}\mathbf{D}_y &= \mathbf{D}_x\mathbf{\Sigma} \\ \mathbf{R}_{yx}\mathbf{D}_x &= \mathbf{R}_{yy}\mathbf{D}_y\mathbf{\Sigma}\end{aligned} \tag{3.19}$$

or equivalently as

$$\begin{aligned}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{D}_x &= \mathbf{D}_x\mathbf{\Sigma}^2 \\ \mathbf{R}_{yx}\mathbf{R}_{xy}\mathbf{D}_y &= \mathbf{R}_{yy}\mathbf{D}_y\mathbf{\Sigma}^2\end{aligned} \tag{3.20}$$

Therefore, the set of equations in (3.19) and (3.20) characterize the solutions for $\mathbf{D}_x$ and $\mathbf{D}_y$ to minimize J under the constraints in (3.15). Equation (3.20) is a coupled (symmetric) generalized eigenvalue problem for $\mathbf{D}_x$ and $\mathbf{D}_y$, with the shared eigenvalue matrix $\mathbf{\Sigma}^2$. Correspondingly, (3.19) is a coupled (asymmetric) generalized eigenvalue problem. In Section 3.3.2, we shall establish that $\mathbf{u} = \mathbf{D}_x^T\mathbf{x}$, $\mathbf{v} = \mathbf{D}_y^T\mathbf{y}$, and $\mathbf{\Sigma} = \text{diag}[\sigma_1, \ldots, \sigma_m] = \mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y$ are, indeed, half-canonical coordinates and half-canonical correlations of $\mathbf{x}$ and $\mathbf{y}$. Later in Chapter 5, we shall give alternating power methods for recursively solving (3.19) for columns of $\mathbf{D}_x = [\mathbf{d}_{x,1}, \ldots, \mathbf{d}_{x,m}]$ and $\mathbf{D}_y = [\mathbf{d}_{y,1}, \ldots, \mathbf{d}_{y,m}]$, and diagonal elements of $\mathbf{\Sigma}$, one by one or in groups.

3.2.3 Case 3: Programmable Canonical Correlation Analysis

In this case, the objective is to whiten $\mathbf{u} = \mathbf{D}_x^T\mathbf{x}$ and $\mathbf{v} = \mathbf{D}_y^T\mathbf{y}$, while minimizing J [14]. The suitable constraints in this case are

$$\mathbf{D}_x^T\mathbf{R}_{xx}\mathbf{D}_x = \mathbf{I}, \quad \text{and} \quad \mathbf{D}_y^T\mathbf{R}_{yy}\mathbf{D}_y = \mathbf{I}. \tag{3.21}$$

Comparing to Case 1, the constraint on the diagonal cross-covariance is not imposed. The term PCCA [14] suggests that $\mathbf{u} = \mathbf{D}_x^T\mathbf{x}$ and $\mathbf{v} = \mathbf{D}_y^T\mathbf{y}$ are canonical coordinates *programmed* by $\mathbf{D}_x$ and $\mathbf{D}_y$. We shall show that this constrained minimization problem is not well-posed, because the solution given in [14] is really the unique solution for Case 1, which happens to be just one of an infinite number of solutions to the PCCA problem.

Contrary to the two-channel CLS problem in Case 1, in PCCA the constraint that ${\mathbf{D}_x}^T\mathbf{R}_{xy}\mathbf{D}_y$ be diagonal is relaxed. Nonetheless, in [14], the coupled (symmetric) generalized eigenvalue problem (3.14) is solved for $\mathbf{D}_x$ and $\mathbf{D}_y$. Thus, the solution of [14] for the PCCA coordinates is actually a solution for Case 1, canonical coordinates, and not for the problem posed. We wish to clarify this point by contrasting the two solutions.

In the case of PCCA, the Lagrange multiplier method for (3.2), with constraints in (3.21), yields the equations

$$\begin{aligned} \mathbf{R}_{xy}\mathbf{D}_y &= \mathbf{R}_{xx}\mathbf{D}_x(\mathbf{I} + \mathbf{\Lambda}_1) \\ \mathbf{R}_{yx}\mathbf{D}_x &= \mathbf{R}_{yy}\mathbf{D}_y(\mathbf{I} + \mathbf{\Lambda}_2) \end{aligned} \tag{3.22}$$

where $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ are $m \times m$ Lagrange multipliers. At the solution, (3.21) is satisfied and thus we have

$$(\mathbf{I} + \mathbf{\Lambda}_1) = (\mathbf{I} + \mathbf{\Lambda}_2)^T = {\mathbf{D}_x}^T\mathbf{R}_{xy}\mathbf{D}_y. \tag{3.23}$$

Without constraints on $\mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y$ there is no way to determine $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$. Nonetheless, we may proceed with them undetermined to rewrite (3.22) as

$$\begin{aligned}\mathbf{R}_{xy}\mathbf{D}_y &= \mathbf{R}_{xx}\mathbf{D}_x(\mathbf{I}+\mathbf{\Lambda}_1)\\ \mathbf{R}_{yx}\mathbf{D}_x &= \mathbf{R}_{yy}\mathbf{D}_y(\mathbf{I}+\mathbf{\Lambda}_1)^T\end{aligned} \tag{3.24}$$

or equivalently as

$$\begin{aligned}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{D}_x &= \mathbf{R}_{xx}\mathbf{D}_x(\mathbf{I}+\mathbf{\Lambda}_1)(\mathbf{I}+\mathbf{\Lambda}_1)^T\\ \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{D}_y &= \mathbf{R}_{yy}\mathbf{D}_y(\mathbf{I}+\mathbf{\Lambda}_1)^T(\mathbf{I}+\mathbf{\Lambda}_1).\end{aligned} \tag{3.25}$$

Since $(\mathbf{I}+\mathbf{\Lambda}_1)(\mathbf{I}+\mathbf{\Lambda}_1)^T$ is not necessarily diagonal, neither equation in (3.25) is a generalized eigenvalue problem. However, noting that $(\mathbf{I}+\mathbf{\Lambda}_1)(\mathbf{I}+\mathbf{\Lambda}_1)^T$ and $(\mathbf{I}+\mathbf{\Lambda}_1)^T(\mathbf{I}+\mathbf{\Lambda}_1)$ are symmetric and have the same eigenvalues, we may consider the eigenvalue decompositions

$$\begin{aligned}(\mathbf{I}+\mathbf{\Lambda}_1)(\mathbf{I}+\mathbf{\Lambda}_1)^T &= \mathbf{A}\mathbf{Q}^2\mathbf{A}^T\\ (\mathbf{I}+\mathbf{\Lambda}_1)^T(\mathbf{I}+\mathbf{\Lambda}_1) &= \mathbf{B}\mathbf{Q}^2\mathbf{B}^T\end{aligned} \tag{3.26}$$

where $\mathbf{A}$ and $\mathbf{B}$ are orthogonal matrices, i.e. $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$, $\mathbf{B}^T\mathbf{B} = \mathbf{B}\mathbf{B}^T = \mathbf{I}$, and $\mathbf{Q}^2$ is a diagonal eigenvalue matrix. Using (3.26), we may rewrite (3.25) as

$$\begin{aligned}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}(\mathbf{D}_x\mathbf{A}) &= \mathbf{R}_{xx}(\mathbf{D}_x\mathbf{A})\mathbf{Q}^2\\ \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}(\mathbf{D}_y\mathbf{B}) &= \mathbf{R}_{yy}(\mathbf{D}_y\mathbf{B})\mathbf{Q}^2.\end{aligned} \tag{3.27}$$

This equation is a coupled (symmetric) generalized eigenvalue problem for $\mathbf{D}_x\mathbf{A}$ and $\mathbf{D}_y\mathbf{B}$. The corresponding coupled (asymmetric) generalized eigenvalue problem is

$$\begin{aligned}\mathbf{R}_{xy}(\mathbf{D}_y\mathbf{B}) &= \mathbf{R}_{xx}(\mathbf{D}_x\mathbf{A})\mathbf{Q}\\ \mathbf{R}_{yx}(\mathbf{D}_x\mathbf{A}) &= \mathbf{R}_{yy}(\mathbf{D}_y\mathbf{B})\mathbf{Q}.\end{aligned} \tag{3.28}$$

These generalized eigenvalue problems for $\mathbf{D}_x\mathbf{A}$ and $\mathbf{D}_y\mathbf{B}$ are the same as those for $\mathbf{D}_x$ and $\mathbf{D}_y$ in Case 1, i.e. (3.13) and (3.14). Therefore, we may write

$$\mathbf{D}_{x,\mathrm{PCCA}}\mathbf{A} = \mathbf{D}_{x,\mathrm{CC}} \quad \text{and} \quad \mathbf{D}_{y,\mathrm{PCCA}}\mathbf{B} = \mathbf{D}_{y,\mathrm{CC}} \tag{3.29}$$

where subscript CC stands for Canonical Coordinates (Case 1), and PCCA for Programmable Canonical Correlation Analysis (Case 3). That is, in the PCCA case, solving (3.28) determines $\mathbf{D}_x$ and $\mathbf{D}_y$ up to unknown (right) orthogonal matrices $\mathbf{A}$ and $\mathbf{B}$. The ambiguity in the solution, however, does not affect the minimum value of (3.2). The solution for Case 1, $\mathbf{D}_x = \mathbf{D}_{x,\mathrm{CC}}$ and $\mathbf{D}_y = \mathbf{D}_{y,\mathrm{CC}}$, would solve the PCCA problem, corresponding to $\mathbf{D}_x^T\mathbf{R}_{xx}\mathbf{D}_x = \mathbf{D}_{x,\mathrm{CC}}^T\mathbf{R}_{xx}\mathbf{D}_{x,\mathrm{CC}} = \mathbf{I}$, $\mathbf{D}_y^T\mathbf{R}_{yy}\mathbf{D}_y = \mathbf{D}_{y,\mathrm{CC}}^T\mathbf{R}_{yy}\mathbf{D}_{y,\mathrm{CC}} = \mathbf{I}$, and $\mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y = \mathbf{D}_{x,\mathrm{CC}}^T\mathbf{R}_{xy}\mathbf{D}_{y,\mathrm{CC}} = \mathbf{\Sigma}$, with $\mathbf{\Sigma}$ diagonal. However, so would $\mathbf{D}_x = \mathbf{D}_{x,\mathrm{CC}}\mathbf{M}$ and $\mathbf{D}_y = \mathbf{D}_{y,\mathrm{CC}}\mathbf{M}$, where $\mathbf{M} \in \mathbb{R}^{m\times m}$ is any orthogonal matrix. In this latter case, $\mathbf{D}_x^T\mathbf{R}_{xx}\mathbf{D}_x = \mathbf{M}^T\mathbf{D}_{x,\mathrm{CC}}^T\mathbf{R}_{xx}\mathbf{D}_{x,\mathrm{CC}}\mathbf{M} = \mathbf{I}$, $\mathbf{D}_y^T\mathbf{R}_{yy}\mathbf{D}_y = \mathbf{M}^T\mathbf{D}_{y,\mathrm{CC}}^T\mathbf{R}_{yy}\mathbf{D}_{y,\mathrm{CC}}\mathbf{M} = \mathbf{I}$, but $\mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y = \mathbf{M}^T\mathbf{D}_{x,\mathrm{CC}}^T\mathbf{R}_{xy}\mathbf{D}_{y,\mathrm{CC}}\mathbf{M} = \mathbf{M}^T\mathbf{\Sigma}\mathbf{M}$ is not diagonal. In summary, a non-unique solution to the PCCA problem, as originally posed in [14], is made unique by imposing the additional constraint of Case 1, i.e. canonical coordinates.

3.3 Two-Channel CLS and Various Canonical Coordinate Systems

In Section 3.2, the two-channel CLS problems were given what might have appeared to be arbitrary names. In this section, we legitimize these names by establishing the connections between the two-channel CLS problems in Cases 1 and 2 and various well-established canonical coordinate systems.

3.3.1 Canonical Coordinates

Consider the constraints in (3.5) for the two-channel CLS problem of Case 1. These constraints may be rewritten as

$$\mathbf{F}^T\mathbf{F} = \mathbf{F}\mathbf{F}^T = \mathbf{I},\ \mathbf{G}^T\mathbf{G} = \mathbf{I},\ \text{and}\ \mathbf{F}^T\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}\mathbf{G} = \mathbf{\Sigma}, \tag{3.30}$$

where $\mathbf{F}^T = \mathbf{D}_x^T\mathbf{R}_{xx}^{1/2}$, $\mathbf{G}^T = \mathbf{D}_y^T\mathbf{R}_{yy}^{1/2}$, $\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xx}\mathbf{R}_{xx}^{-T/2} = \mathbf{I}$, and $\mathbf{R}_{xx}^{1/2}\mathbf{R}_{xx}^{T/2} = \mathbf{R}_{xx}$. Thus, $\mathbf{F}\mathbf{\Sigma}\mathbf{G}^T = \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} = \mathbf{C}$ is a thin SVD [57] of the coherence matrix $\mathbf{C}$.

The thin SVD of a rectangular matrix $\mathbf{A}$ is a truncated version of the SVD of $\mathbf{A}$, in which the zero singular values of $\mathbf{A}$ and their corresponding singular vectors are discarded in forming the SVD. Thus, the thin SVD matrices $\mathbf{F} \in \mathbb{R}^{m\times m}$, $\mathbf{G} \in \mathbb{R}^{n\times m}$, and $\mathbf{\Sigma} \in \mathbb{R}^{m\times m}$ have the following relations with the full SVD matrices $\mathbf{F}_c \in \mathbb{R}^{m\times m}$, $\mathbf{G}_c \in \mathbb{R}^{n\times n}$, and $\mathbf{\Sigma}_c \in \mathbb{R}^{m\times n}$ in (2.6):

$$\mathbf{F} = \mathbf{F}_c, \quad \mathbf{G} = \mathbf{G}_{c,m}, \quad \text{and} \quad \mathbf{\Sigma} = \mathbf{\Sigma}_c(m) = \text{diag}[\sigma_1, \ldots, \sigma_m]. \tag{3.31}$$

Here $\mathbf{G}_{c,m} \in \mathbb{R}^{n\times m}$ denotes the matrix that carries the first m columns of $\mathbf{G}_c \in \mathbb{R}^{n\times n}$. Since we have assumed $m \leq n$, $\mathbf{G}^T\mathbf{G} = \mathbf{G}_{c,m}^T\mathbf{G}_{c,m} = \mathbf{I}(m)$ but $\mathbf{G}\mathbf{G}^T = \mathbf{G}_{c,m}\mathbf{G}_{c,m}^T$ is an orthogonal projection onto the subspace $< \mathbf{G} >$.

Therefore, the diagonal matrix $\mathbf{\Sigma} = \mathbf{F}^T\mathbf{C}\mathbf{G} = \mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y$ is in fact the canonical correlation matrix of canonical correlations σ_i, $i \in [1, m]$. Correspondingly, the matrices

$$\mathbf{D}_x^T = \mathbf{F}^T\mathbf{R}_{xx}^{-1/2} \quad \text{and} \quad \mathbf{D}_y^T = \mathbf{G}^T\mathbf{R}_{yy}^{-1/2}, \tag{3.32}$$

which are solutions to the two-channel CLS problem of Case 1, map $\mathbf{x}$ and $\mathbf{y}$ to their respective canonical coordinates $\mathbf{u} = \mathbf{D}_x^T\mathbf{x}$ and $\mathbf{v} = \mathbf{D}_y^T\mathbf{y}$. Thus, we refer to $\mathbf{D}_x^T$ and $\mathbf{D}_y^T$, as *canonical coordinate maps.* As opposed to the canonical coordinate vector $\mathbf{v} = \mathbf{G}_c^T\mathbf{R}_{yy}^{-1/2}\mathbf{y} \in \mathbb{R}^n$ in Chapter 2, here the vector $\mathbf{v} = \mathbf{G}^T\mathbf{R}_{yy}^{-1/2}\mathbf{y} \in \mathbb{R}^m$ contains only the first m canonical coordinates of $\mathbf{y}$. However, for simplicity in notation in this chapter and Chapters 4, 5, and 6 we use $\mathbf{v}$ to denote the first m canonical coordinates of $\mathbf{y}$. Note that since $m \leq n$, only the first m canonical coordinates of $\mathbf{y}$ are important, as there are only m nonzero canonical correlations.

This connection between two-channel CLS filters $\mathbf{D}_x^T$ and $\mathbf{D}_y^T$ and canonical coordinate decomposition implies that the set of constraints in (3.5) are the right constraints when the objective is to carry the linear dependence or information rate between $\mathbf{x}$ and $\mathbf{y}$ in pairwise dependence or rates between the elements of $\mathbf{u}$ and $\mathbf{v}$.

Using (3.30) and (3.32), along with the cyclic property of trace, the minimum value of J for Case 1, denoted by J_{uu}, may be written as

$$\begin{aligned} J_{uu} &= \mathrm{tr}\{\mathbf{F}^T\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}\mathbf{R}_{xx}^{-T/2}\mathbf{F}\} \\ &\quad +\mathrm{tr}\{\mathbf{I}-\mathbf{\Sigma}-\mathbf{\Sigma}^T+\mathbf{F}^T\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-T/2}\mathbf{F}\} \\ &= \mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}\mathbf{R}_{xx}^{-T/2}\}+\mathrm{tr}\{\mathbf{I}-\mathbf{\Sigma}-\mathbf{\Sigma}^T+\mathbf{\Sigma}\mathbf{\Sigma}^T\} \\ &= \mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}\mathbf{R}_{xx}^{-T/2}\}+\mathrm{tr}\{(\mathbf{I}-\mathbf{\Sigma})(\mathbf{I}-\mathbf{\Sigma}^T)\} \\ &= \mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}\mathbf{R}_{xx}^{-T/2}\}+\textstyle\sum_{i=1}^{m}(1-\sigma_i)^2. \end{aligned} \tag{3.33}$$

It is interesting to compare this index with the MSE in estimating $\mathbf{u}$ from $\mathbf{v}$. From (2.10), the linear MMSE estimator of $\mathbf{u}$ from $\mathbf{v}$ is $\hat{\mathbf{u}} = \mathbf{\Sigma}\mathbf{v}$ and its corresponding error covariance matrix is $\mathbf{Q}_{uu} = \mathbf{E}[(\mathbf{u}-\hat{\mathbf{u}})(\mathbf{u}-\hat{\mathbf{u}})^T] = \mathbf{I}-\mathbf{\Sigma}\mathbf{\Sigma}^T$. Thus, the MSE in estimating $\mathbf{u}$ from $\mathbf{v}$ is

$$\begin{aligned} \mathrm{MSE}_{uu} &= \mathrm{tr}\{\mathbf{Q}_{uu}\} = \mathrm{tr}\{E[(\mathbf{u}-\hat{\mathbf{u}})(\mathbf{u}-\hat{\mathbf{u}})^T]\} \\ &= \mathrm{tr}\{(\mathbf{I}-\mathbf{\Sigma}\mathbf{\Sigma}^T)\} = \textstyle\sum_{i=1}^{m}(1-\sigma_i^2) = \mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}\mathbf{R}_{xx}^{-T/2}\}. \end{aligned} \tag{3.34}$$

The last equality directly follows from the decomposition of $\mathbf{Q}_{xx}$ in (2.19), using the cyclic property trace. Thus, the connection between J_{uu} and MSE_{uu} is

$$\begin{aligned} J_{uu} &= \mathrm{MSE}_{uu}+\mathrm{tr}\{(\mathbf{I}-\mathbf{\Sigma})(\mathbf{I}-\mathbf{\Sigma})^T\} \\ &= \mathrm{MSE}_{uu}+\textstyle\sum_{i=1}^{m}(1-\sigma_i)^2. \end{aligned} \tag{3.35}$$

We see that J_{uu} is within $\mathrm{tr}\{(\mathbf{I}-\mathbf{\Sigma})(\mathbf{I}-\mathbf{\Sigma})^T\} = \sum_{i=1}^{m}(1-\sigma_i)^2$, an invariant for this two-channel problem, of the MSE for estimating $\mathbf{u}$ from $\mathbf{v}$, further clarifying the connection between two-channel CLS filtering and canonical correlation analysis.

3.3.2 Half-Canonical Coordinates

For the two-channel CLS problem of Case 2, the constraints in (3.15) may be written as

$$\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}, \quad \text{and} \quad \mathbf{U}^T\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}\mathbf{V} = \mathbf{\Sigma}, \tag{3.36}$$

where $\mathbf{U}^T = \mathbf{D}_x^T$, and $\mathbf{V}^T = \mathbf{D}_y^T \mathbf{R}_{yy}^{1/2}$. Thus, $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} = \mathbf{C}_h$ is a thin SVD of the half-coherence matrix $\mathbf{C}_h = E[\mathbf{x}\mathbf{y}^T\mathbf{R}_{yy}^{-T/2}] = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}$. Comparing the thin SVD $\mathbf{C}_h$ with its full SVD in (2.38), we have

$$\mathbf{U} = \mathbf{U}_h, \quad \mathbf{V} = \mathbf{V}_{h,m}, \quad \text{and} \quad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_h(m) = \text{diag}[\sigma_1, \ldots, \sigma_m] \tag{3.37}$$

where $\mathbf{V}_{h,m} \in \mathbb{R}^{n \times m}$ contains the first m columns of $\mathbf{V}_h \in \mathbb{R}^{n \times n}$. Note that since $m \leq n$, $\mathbf{V}^T\mathbf{V} = \mathbf{V}_{h,m}^T\mathbf{V}_{h,m} = \mathbf{I}(m)$ but $\mathbf{V}\mathbf{V}^T = \mathbf{V}_{h,m}\mathbf{V}_{h,m}^T$ is an orthogonal projection onto the subspace $< \mathbf{V} >$.

Therefore, the diagonal matrix $\boldsymbol{\Sigma} = \mathbf{U}^T\mathbf{C}_h\mathbf{V} = \mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y$ is in fact the half-canonical correlation matrix of half-canonical correlations σ_i, $i \in [1, m]$. In this case, the matrices

$$\mathbf{D}_x^T = \mathbf{U}^T, \quad \text{and} \quad \mathbf{D}_y^T = \mathbf{V}^T\mathbf{R}_{yy}^{-1/2}, \tag{3.38}$$

which are solutions to the two-channel CLS problem of Case 2, map $\mathbf{x}$ and $\mathbf{y}$ to their respective half-canonical coordinates $\mathbf{u} = \mathbf{D}_x^T\mathbf{x}$ and $\mathbf{v} = \mathbf{D}_y^T\mathbf{y}$. For this reason, we refer to $\mathbf{D}_x^T$ and $\mathbf{D}_y^T$, as *half-canonical coordinate maps.* As opposed to the half-canonical coordinate vector $\mathbf{v} = \mathbf{V}_h^T\mathbf{R}_{yy}^{-1/2}\mathbf{y} \in \mathbb{R}^n$ in Chapter 2, here the vector $\mathbf{v} = \mathbf{V}^T\mathbf{R}_{yy}^{-1/2}\mathbf{y} \in \mathbb{R}^m$ contains only the first m half-canonical coordinates of $\mathbf{y}$. However, for simplicity in notation in this chapter and Chapters 4, 5, and 6 we use $\mathbf{v}$ to denote the first m half-canonical coordinates of $\mathbf{y}$. Again note that since $m \leq n$, only the first m half-canonical coordinates of $\mathbf{y}$ are important, as there are only m nonzero half-canonical correlations.

Using (3.36) and (3.38), along with the cyclic property of trace, the minimum value of J for Case 2, denoted by J_{xx}, may be written as

$$\begin{aligned} J_{xx} &= \text{tr}\{\mathbf{U}^T\mathbf{Q}_{xx}\mathbf{U}\} + \text{tr}\{\mathbf{I} - \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^T + \mathbf{U}^T\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{U}\} \\ &= \text{tr}\{\mathbf{U}^T\mathbf{Q}_{xx}\mathbf{U}\} + \text{tr}\{\mathbf{I} - \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^T + \mathbf{U}^T\mathbf{C}_h\mathbf{C}_h^T\mathbf{U}\} \\ &= \text{tr}\{\mathbf{Q}_{xx}\} + \text{tr}\{(\mathbf{I} - \boldsymbol{\Sigma})(\mathbf{I} - \boldsymbol{\Sigma}^T)\} \\ &= \text{tr}\{\mathbf{Q}_{xx}\} + \textstyle\sum_{i=1}^m (1 - \sigma_i)^2. \end{aligned} \tag{3.39}$$

The first term on the right side of (3.39), $\mathrm{tr}\{\mathbf{Q}_{xx}\} = \mathrm{tr}\{\mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\} = \mathrm{MSE}_{xx}$, is the MSE in estimating $\mathbf{x}$ from $\mathbf{y}$, using the Wiener filter $\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}$. Therefore, we have

$$\begin{aligned} J_{xx} &= \mathrm{MSE}_{xx} + \mathrm{tr}\{(\mathbf{I} - \mathbf{\Sigma})(\mathbf{I} - \mathbf{\Sigma})^T\} \\ &= \mathrm{MSE}_{xx} + \textstyle\sum_{i=1}^{m}(1 - \sigma_i)^2 \end{aligned} \tag{3.40}$$

Thus, J_{xx} is within $\mathrm{tr}\{(\mathbf{I} - \mathbf{\Sigma})(\mathbf{I} - \mathbf{\Sigma})^T\} = \sum_{i=1}^{m}(1 - \sigma_i)^2$, an invariant for this two-channel problem, of the MSE in estimating $\mathbf{x}$ from $\mathbf{y}$, further clarifying the connection between two-channel CLS filtering and half-canonical correlation analysis.

3.4 Conclusions

In this chapter, a general class of two-channel CLS problems, with various constraints, was introduced and the corresponding solutions were derived. We showed that the solution to each two-channel CLS problem is determined from a coupled (asymmetric) generalized eigenvalue problem. Furthermore, it was shown that depending upon the constraints, the two-channel CLS solution decomposes the two data channels into one of three important coordinate systems, namely canonical coordinates, half-canonical coordinates, or PCCA coordinates. In addition, we clarified that the PCCA problem, as originally posed in [14], has an infinite number of solutions, the most compelling of which is the canonical coordinate solution.

CHAPTER 4

OPTIMAL REDUCED-RANK FILTERING IN FULL- AND HALF-CANONICAL COORDINATES

4.1 Introduction

Reduced-rank estimation and filtering [6]– [11], [18]– [21] are important for a wide range of signal processing applications where data or model reduction, robustness against noise or model errors, or high computational efficiency is desired. Generally, rank-reduction is performed by discarding the subdominant modes (subdominant singular values and their corresponding singular vectors) of a covariance or cross-covariance matrix. This results in a dimensionality reduction and hence may lower computational load. When the discarded modes correspond to the noise subspace, rank-reduction can provide robustness against noise. Fundamental results on optimal reduced-rank estimators and filters include [11], the reduced-rank Wiener filter (RRWF) [6]– [10], and the reduced-rank maximum likelihood estimator (RRMLE) [18]. Other examples of reduced-rank estimators and filters include the reduced-rank multilayer neural network (RRMNN) [19], the relative Karhunen-Loeve transform (RKLT) [20], and the generalized Karhunen-Loeve transform (GKLT) [21].

In this chapter, we wish to establish a unified framework for deriving and implementing three classes of reduced-rank Wiener filters, where each class corresponds to a particular error measure. We intend to clarify the connections between reduced-rank Wiener filters, canonical coordinates, half-canonical coordinates, and two-channel CLS filters. In Section 4.2.1, we will show that two of the classes of reduced-rank Wiener filters are equivalent and canonical coordinates are optimal for reduced-rank Wiener filtering under their corresponding error measures. Then, in Section 4.2.2, we establish that half-canonical coordinates are optimal for reduced-rank Wiener filtering under the third error measure. Our results reproduce what is known from [9], but our method follows the line of argument given in [8] for deriving the reduced-rank Wiener filter in half-canonical coordinates. The material presented in this chapter are also reported in [13].

4.2 Reduced-Rank Filtering

Consider again the composite vector $\mathbf{z} = [\mathbf{x}^T \ \mathbf{y}^T]$, with zero-mean random vectors, $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$, $m \leq n$, and composite covariance matrix $\mathbf{R}_{zz} = E[\mathbf{z}\mathbf{z}^T]$ of (2.2). Let $\hat{\mathbf{x}} = \mathbf{H}[r]\mathbf{y}$ be a rank $r < m$ estimate of $\mathbf{x}$ for some rank-r linear transformation of $\mathbf{y}$. Then, the covariance matrix of the filtering error $\mathbf{e}_x[r] = \mathbf{x} - \hat{\mathbf{x}} = \mathbf{x} - \mathbf{H}[r]\mathbf{y}$ may be written as

$$\begin{aligned}
\mathbf{Q}_{xx}[r] &= E[\mathbf{e}_x[r]\mathbf{e}_x[r]^T] \\
&= E[(\mathbf{x} - \mathbf{H}[r]\mathbf{y})(\mathbf{x} - \mathbf{H}[r]\mathbf{y})^T] \\
&= \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx} + [\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} - \mathbf{H}[r]]\mathbf{R}_{yy}[\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} - \mathbf{H}[r]]^T \\
&= \mathbf{Q}_{xx} + [\mathbf{H} - \mathbf{H}[r]]\mathbf{R}_{yy}[\mathbf{H} - \mathbf{H}[r]]^T
\end{aligned} \tag{4.1}$$

where $\mathbf{Q}_{xx} = E[(\mathbf{x} - \mathbf{H}\mathbf{y})(\mathbf{x} - \mathbf{H}\mathbf{y})^T] = \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}$ is the error covariance matrix in estimating $\mathbf{x}$ from $\mathbf{y}$, using the full-rank Wiener filter $\mathbf{H} = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}$.

The choice of coordinate system for building the optimal rank-r Wiener filter $\mathbf{H}[r] \in \mathbb{R}^{m \times n}$ depends on the measure to be optimized. The common measures for

reduced-rank Wiener filtering are

$$\mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}[r]\mathbf{R}_{xx}^{-T/2}\},\ \det\{\mathbf{Q}_{xx}[r]\},\ \text{and}\ \mathrm{tr}\{\mathbf{Q}_{xx}[r]\}.$$

The first measure, $\mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}[r]\mathbf{R}_{xx}^{-T/2}\}$, is a whitened MSE measure, the second measure, $\det\{\mathbf{Q}_{xx}[r]\}$, is proportional to the volume of the concentration ellipse of the filtering error $\mathbf{e}_x[r]$, and the third measure, $\mathrm{tr}\{\mathbf{Q}_{xx}[r]\}$, measures the MSE. In [9], these error measures have been reviewed and the first two measures have been shown to be equivalent. In what follows we reproduce the results of [9], in a unified way, using the line of argument given in [8] for deriving the reduced-rank Wiener filter in half-canonical coordinates. We demonstrate that canonical coordinates are optimal for rank reduction based on the first two error measures, $\mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}[r]\mathbf{R}_{xx}^{-T/2}\}$ and $\det\{\mathbf{Q}_{xx}[r]\}$, while half-canonical coordinates are optimal for rank reduction based on the third measure, $\mathrm{tr}\{\mathbf{Q}_{xx}[r]\}$. Additionally, we present several equivalent implementations of reduced-rank Wiener filters, in canonical and half-canonical coordinates, and establish connections between reduced-rank Wiener filters and the two-channel CLS filters of Chapter 3.

4.2.1 Optimal Reduced-Rank Filtering in Canonical Coordinates

Here the objective is to find the rank-r filter $\mathbf{H}[r]$ that minimizes the trace of the weighted error covariance matrix $\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}[r]\mathbf{R}_{xx}^{-T/2}$ [9]. Note that $\mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}[r]\mathbf{R}_{xx}^{-T/2}\}$ $= \mathrm{tr}\{\mathbf{F}^T\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}[r]\mathbf{R}_{xx}^{-T/2}\mathbf{F}\} = \mathrm{tr}\{\mathbf{Q}_{uu}[r]\}$ is the MSE for the rank-r Wiener estimator of the canonical coordinates $\mathbf{u} = \mathbf{F}^T\mathbf{R}_{xx}^{-1/2}\mathbf{x}$ from the canonical coordinates $\mathbf{v} = \mathbf{G}^T\mathbf{R}_{yy}^{-1/2}\mathbf{y}$ [6], so we denote it by $\mathrm{MSE}_{uu}[r]$:

$$\begin{aligned}\mathrm{MSE}_{uu}[r] &= \mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}[r]\mathbf{R}_{xx}^{-T/2}\}\\ &= \mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}(\mathbf{Q}_{xx} + [\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} - \mathbf{H}[r]]\mathbf{R}_{yy}\\ &\qquad \times[\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} - \mathbf{H}[r]]^T)\mathbf{R}_{xx}^{-T/2}\}.\end{aligned} \tag{4.2}$$

Here $\mathbf{F} \in \mathbb{R}^{m\times m}$ is the orthogonal matrix in the thin SVD of the coherence matrix $\mathbf{C} = \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} = \mathbf{F}\mathbf{\Sigma}\mathbf{G}^T$ in (3.30). Inserting $\mathbf{C} = \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}$ in (4.2), we

may rewrite $\mathrm{MSE}_{uu}[r]$ as

$$\begin{aligned}\mathrm{MSE}_{uu}[r] &= \mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}\mathbf{R}_{xx}^{-T/2}\} \\ &\quad +\mathrm{tr}\{\left[\mathbf{C}-\mathbf{R}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2}\right]\left[\mathbf{C}-\mathbf{R}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2}\right]^T\}.\end{aligned} \tag{4.3}$$

The first term on the right side of this equation is fixed. Thus, minimizing $\mathrm{MSE}_{uu}[r]$ is equivalent to minimizing the second term,

$$\epsilon_{uu}^2 = \mathrm{tr}\{[\mathbf{C}-\mathbf{R}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2}][\mathbf{C}-\mathbf{R}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2}]^T\} \tag{4.4}$$

which is the Frobenius norm of the matrix $\mathbf{C}-\mathbf{R}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2}$. It measures the extra MSE introduced by rank reduction. The optimum choice for the rank-r Wiener filter $\mathbf{H}[r]$ is the rank-r matrix that best approximates the coherence matrix $\mathbf{C} = \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} = \mathbf{F}\mathbf{\Sigma}\mathbf{G}^T$, by minimizing ϵ_{uu}^2 [59]. Thus, it is given by $\mathbf{R}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2} = \mathbf{F}\mathbf{\Sigma}[r]\mathbf{G}^T$, or

$$\mathbf{H}[r] = \mathbf{R}_{xx}^{1/2}\mathbf{F}\mathbf{\Sigma}[r]\mathbf{G}^T\mathbf{R}_{yy}^{-1/2};\quad \mathbf{\Sigma}[r] = \begin{bmatrix}\mathbf{\Sigma}(r) & 0\\ 0 & 0\end{bmatrix} \in \mathbb{R}^{m\times m}, \tag{4.5}$$

where $\mathbf{\Sigma}(r) = \mathrm{diag}[\sigma_1,\ldots,\sigma_r]$ is the first $r\times r$ block of the diagonal matrix $\mathbf{\Sigma}$, i.e. the canonical correlation matrix of the canonical correlations σ_i. Using (4.5), the optimal value of $\mathrm{MSE}_{uu}[r]$ is

$$\begin{aligned}\mathrm{MSE}_{uu}[r] &= \mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}\mathbf{R}_{xx}^{-T/2}\} + \mathrm{tr}\{\left[\mathbf{C}-\mathbf{R}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2}\right]\left[\mathbf{C}-\mathbf{R}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2}\right]^T\} \\ &= \mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}\mathbf{R}_{xx}^{-T/2}\} + \mathrm{tr}\{(\mathbf{F}\mathbf{\Sigma}\mathbf{G}^T-\mathbf{F}\mathbf{\Sigma}[r]\mathbf{G}^T)(\mathbf{F}\mathbf{\Sigma}\mathbf{G}^T-\mathbf{F}\mathbf{\Sigma}[r]\mathbf{G}^T)^T\} \\ &= \mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}\mathbf{R}_{xx}^{-T/2}\} + \mathrm{tr}\{\mathbf{F}(\mathbf{\Sigma}-\mathbf{\Sigma}[r])\mathbf{G}^T\mathbf{G}(\mathbf{\Sigma}-\mathbf{\Sigma}[r])^T\mathbf{F}^T\} \\ &= \mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}\mathbf{R}_{xx}^{-T/2}\} + \mathrm{tr}\{\mathbf{\Sigma}\mathbf{\Sigma}^T-\mathbf{\Sigma}\mathbf{\Sigma}[r]^T-\mathbf{\Sigma}[r]\mathbf{\Sigma}^T+\mathbf{\Sigma}^T[r]\mathbf{\Sigma}[r]^T\} \\ &= \mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}\mathbf{R}_{xx}^{-T/2}\} + \mathrm{tr}\{\mathbf{\Sigma}\mathbf{\Sigma}^T-\mathbf{\Sigma}[r]\mathbf{\Sigma}[r]^T\} \\ &= \mathrm{MSE}_{uu} + \textstyle\sum_{i=r+1}^{m}\sigma_i^2.\end{aligned} \tag{4.6}$$

The first term on the right hand side of (4.6) is the minimum MSE for a full-rank estimator of $\mathbf{u}$ from $\mathbf{v}$ and the second term is the extra MSE due to rank reduction.

We now show that the minimization of $\mathrm{MSE}_{uu}[r]$ in (4.2) is equivalent to minimization of the volume of the concentration ellipse $\{\mathbf{e}_x[r] \in \mathbb{R}^m : \mathbf{e}_x^T[r]\mathbf{Q}_{xx}[r]^{-1}\mathbf{e}_x[r] = 1\}$. This volume is proportional to determinant of the error covariance matrix $\mathbf{Q}_{xx}[r]$ [8], which we may write as

$$\begin{aligned}\det\{\mathbf{Q}_{xx}[r]\} &= \det\{\mathbf{Q}_{xx} + [\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} - \mathbf{H}[r]]\mathbf{R}_{yy}[\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} - \mathbf{H}[r]]^T\} \\ &= \det\{\mathbf{Q}_{xx}^{1/2}\}\det\{\mathbf{I} + [\mathbf{Q}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} - \mathbf{Q}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2}] \\ &\quad \times[\mathbf{Q}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} - \mathbf{Q}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2}]^T\}\det\{\mathbf{Q}_{xx}^{T/2}\}.\end{aligned} \tag{4.7}$$

The terms $\det\{\mathbf{Q}_{xx}^{1/2}\}$ and $\det\{\mathbf{Q}_{xx}^{T/2}\}$ do not affect minimization of $\det\{\mathbf{Q}_{xx}[r]\}$ and thus may be dropped. We may then rewrite the middle determinant as

$$\begin{aligned}&\det\{\mathbf{I} + \left[\mathbf{Q}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} - \mathbf{Q}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2}\right] \\ &\times \left[\mathbf{Q}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} - \mathbf{Q}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2}\right]^T\} = \textstyle\prod_{i=1}^{m}(1+\gamma_i^2)\end{aligned} \tag{4.8}$$

where the γ_i's are the singular values of $\mathbf{Q}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} - \mathbf{Q}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2}$. Thus, minimization of $\det\{\mathbf{Q}_{xx}[r]\}$ reduces to minimization of the γ_i's. Given a fixed matrix $\mathbf{M}$ and a rank-r matrix $\mathbf{N}$ of the same dimension as $\mathbf{M}$, the singular values of $\mathbf{M}-\mathbf{N}$ are minimized if $\mathbf{N}$ is the rank-r approximation of $\mathbf{M}$ that is computed from the SVD of $\mathbf{M}$ [11], [60]. Thus, the optimal rank-r Wiener filter $\mathbf{H}[r]$ must satisfy

$$\mathbf{Q}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2} = [\mathbf{Q}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}]_r \tag{4.9}$$

where $[\mathbf{Q}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}]_r$ is the rank-r approximation of the matrix $\mathbf{Q}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}$ computed from the SVD of $\mathbf{Q}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}$.

The matrix $\mathbf{Q}_{xx}^{1/2}$ is a square-root (not necessarily symmetric) of $\mathbf{Q}_{xx} = \mathbf{Q}_{xx}^{1/2}\mathbf{Q}_{xx}^{T/2} = \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}$. Therefore, we may write $\mathbf{Q}_{xx}^{1/2}$ as[1]

$$\begin{aligned}
\mathbf{Q}_{xx}^{1/2} &= (\mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx})^{1/2} \\
&= (\mathbf{R}_{xx}^{1/2}(\mathbf{I} - \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-T/2})\mathbf{R}_{xx}^{T/2})^{1/2} \\
&= (\mathbf{R}_{xx}^{1/2}(\mathbf{I} - \mathbf{C}\mathbf{C}^T)^{1/2}(\mathbf{I} - \mathbf{C}\mathbf{C}^T)^{T/2}\mathbf{R}_{xx}^{T/2})^{1/2} \\
&= \mathbf{R}_{xx}^{1/2}(\mathbf{I} - \mathbf{C}\mathbf{C}^T)^{1/2}.
\end{aligned} \tag{4.10}$$

Plugging $\mathbf{Q}_{xx}^{1/2}$ in the right hand side of (4.9) and using the thin SVD in (3.30) for the coherence matrix $\mathbf{C} = \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} = \mathbf{F}\mathbf{\Sigma}\mathbf{G}^T$ reduces (4.9) to

$$\begin{aligned}
\mathbf{Q}_{xx}^{-1/2}\mathbf{H}[r]\mathbf{R}_{yy}^{1/2} &= [(\mathbf{I} - \mathbf{C}\mathbf{C}^T)^{-1/2}\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}]_r \\
&= [(\mathbf{I} - \mathbf{F}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{F}^T)^{-1/2}\mathbf{F}\mathbf{\Sigma}\mathbf{G}^T]_r \\
&= [(\mathbf{F}(\mathbf{I} - \mathbf{\Sigma}\mathbf{\Sigma}^T)^{1/2})^{-1}\mathbf{F}\mathbf{\Sigma}\mathbf{G}^T]_r \\
&= ([(\mathbf{I} - \mathbf{\Sigma}\mathbf{\Sigma}^T)^{-1/2}\mathbf{\Sigma}\mathbf{G}^T]_r \\
&= (\mathbf{I} - \mathbf{\Sigma}\mathbf{\Sigma}^T)^{-1/2}\mathbf{\Sigma}[r]\mathbf{G}^T.
\end{aligned} \tag{4.11}$$

Pre-multiplying (4.11) by $\mathbf{Q}_{xx}^{1/2} = \mathbf{R}_{xx}^{1/2}(\mathbf{I} - \mathbf{F}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{F}^T)^{1/2}$ and post-multiplying by $\mathbf{R}_{yy}^{-1/2}$ yields

$$\begin{aligned}
\mathbf{H}[r] &= \mathbf{R}_{xx}^{1/2}(\mathbf{I} - \mathbf{F}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{F}^T)^{1/2}(\mathbf{I} - \mathbf{\Sigma}\mathbf{\Sigma}^T)^{-1/2}\mathbf{\Sigma}[r]\mathbf{G}^T\mathbf{R}_{yy}^{-1/2} \\
&= \mathbf{R}_{xx}^{1/2}\mathbf{F}(\mathbf{I} - \mathbf{\Sigma}\mathbf{\Sigma}^T)^{1/2}(\mathbf{I} - \mathbf{\Sigma}\mathbf{\Sigma}^T)^{-1/2}\mathbf{\Sigma}[r]\mathbf{G}^T\mathbf{R}_{yy}^{-1/2} \\
&= \mathbf{R}_{xx}^{1/2}\mathbf{F}\mathbf{\Sigma}[r]\mathbf{G}^T\mathbf{R}_{yy}^{-1/2}.
\end{aligned} \tag{4.12}$$

This is the optimal rank-r Wiener filter $\mathbf{H}[r]$ that minimizes (4.7) and is equivalent to the optimal rank-r Wiener filter in (4.5) that minimizes the measure $\mathrm{MSE}_{uu}[r]$ in (4.2). Therefore, canonical coordinates are the right coordinate system for reduced-rank Wiener filtering when the objective is to minimize either the determinant of the

[1]We note that the square-root matrix $\mathbf{Q}_{xx}^{1/2}$ is arbitrary up to a right orthogonal matrix. That is, $\mathbf{Q}_{xx}^{1/2}\mathbf{T}^T$, with $\mathbf{T}^T\mathbf{T} = \mathbf{T}\mathbf{T}^T = \mathbf{I}$, is also a square-root of $\mathbf{Q}_{xx}$. Nonetheless, the orthogonal matrix $\mathbf{T}$ does not affect our discussion, and hence we consider $\mathbf{Q}_{xx}^{1/2}$ to be of form (4.10).

error covariance matrix, $\det\{\mathbf{Q}_{xx}[r]\}$, or the whitened MSE, $\mathrm{tr}\{\mathbf{R}_{xx}^{-1/2}\mathbf{Q}_{xx}[r]\mathbf{R}_{xx}^{-T/2}\}$. The corresponding rank-r Wiener estimate of $\mathbf{x}$ from $\mathbf{y}$ is given by $\hat{\mathbf{x}} = \mathbf{H}[r]\mathbf{y}$. The optimal value of $\det\{\mathbf{Q}_{xx}[r]\}$ in (4.7) is

$$\begin{aligned}\det\{\mathbf{Q}_{xx}[r]\} &= \det\{\mathbf{R}_{xx}\}\det\{\mathbf{I} - \mathbf{F}\boldsymbol{\Sigma}[r]\boldsymbol{\Sigma}[r]^T\mathbf{F}^T - \mathbf{F}\boldsymbol{\Sigma}[r]\boldsymbol{\Sigma}[r]^T\mathbf{F}^T + \mathbf{F}\boldsymbol{\Sigma}[r]\boldsymbol{\Sigma}[r]^T\mathbf{F}^T\} \\ &= \det\{\mathbf{R}_{xx}\}\det\{\mathbf{I} - \boldsymbol{\Sigma}[r]\boldsymbol{\Sigma}[r]^T\} \\ &= \det\{\mathbf{R}_{xx}\}\textstyle\prod_{i=1}^{r}(1-\sigma_i^2) \\ &= \det\{\mathbf{Q}_{xx}\}\frac{1}{\prod_{i=r+1}^{m}(1-\sigma_i^2)}\end{aligned} \tag{4.13}$$

which is proportional to the volume of the concentration ellipse for the rank-r estimator $\mathbf{H}[r]$. The last equality (4.13) follows from (2.22).

In analogy to (2.23), the linear dependence for the composite vector $\mathbf{s} = [\mathbf{x}^T \quad \hat{\mathbf{x}}^T]^T$ may be measured by the Hadamard ratio inside the inequality

$$0 \le \frac{\det\{\mathbf{R}_{ss}\}}{\prod_{i=1}^{2m}[\mathbf{R}_{ss}]_{ii}} \le 1 \tag{4.14}$$

where

$$\begin{aligned}\mathbf{R}_{ss} &= E\left[\begin{pmatrix}\mathbf{x}\\ \hat{\mathbf{x}}\end{pmatrix}(\mathbf{x}^T \quad \hat{\mathbf{x}}^T)\right] = \begin{bmatrix}\mathbf{R}_{xx} & \mathbf{R}_{x\hat{x}}\\ \mathbf{R}_{\hat{x}x} & \mathbf{R}_{\hat{x}\hat{x}}\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{R}_{\hat{x}x}\mathbf{R}_{xx}^{-1} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{R}_{xx} & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}_{xx}[r]\mathbf{R}_{xx}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{R}_{xx}^{-1}\mathbf{R}_{x\hat{x}}\\ \mathbf{0} & \mathbf{R}_{\hat{x}\hat{x}}\end{bmatrix}\end{aligned} \tag{4.15}$$

is the composite covariance matrix for the composite vector $\mathbf{s}$, and $[\mathbf{R}_{ss}]_{ii}$'s are the diagonal elements of $\mathbf{R}_{ss}$. The ratio takes the value 0 iff there is linear dependence among elements of $\mathbf{s}$; it takes the value 1 iff elements of $\mathbf{s}$ are mutually uncorrelated. Using (4.13) and (4.15), the Hadamard ratio in (4.14) may be written as

$$\begin{aligned}\frac{\det\{\mathbf{R}_{ss}\}}{\prod_{i=1}^{2m}[\mathbf{R}_{ss}]_{ii}} &= \frac{\det\{\mathbf{Q}_{xx}[r]\}\det\{\mathbf{R}_{\hat{x}\hat{x}}\}}{\prod_{i=1}^{m}[\mathbf{R}_{xx}]_{ii}\prod_{i=1}^{m}[\mathbf{R}_{\hat{x}\hat{x}}]_{ii}} \\ &= \frac{\det\{\mathbf{R}_{xx}\}}{\prod_{i=1}^{m}[\mathbf{R}_{xx}]_{ii}}\det\{\mathbf{I} - \boldsymbol{\Sigma}[r]\boldsymbol{\Sigma}[r]^T\}\frac{\det\{\mathbf{R}_{\hat{x}\hat{x}}\}}{\prod_{i=1}^{m}[\mathbf{R}_{\hat{x}\hat{x}}]_{ii}}.\end{aligned} \tag{4.16}$$

The first term on the right hand side measures the linear dependence among the elements of $\mathbf{x}$, and the third term measures the linear dependence among the elements of $\hat{\mathbf{x}}$. The middle term $L[r] = \det\{\mathbf{I} - \mathbf{\Sigma}[r]\mathbf{\Sigma}[r]^T\}$ measures the linear dependence between the elements of $\mathbf{x}$ and $\hat{\mathbf{x}}$. It is seen that the linear dependence $L[r]$ is proportional to the volume of the concentration ellipse of the filtering error $\mathbf{e}_x[r]$. This linear dependence may also be written as

$$L[r] = \det\{\mathbf{I} - \mathbf{\Sigma}[r]\mathbf{\Sigma}[r]^T\} = \prod_{i=1}^{r}(1-\sigma_i^2) = \frac{L}{\prod_{i=r+1}^{m}(1-\sigma_i^2)} \tag{4.17}$$

where $L = \prod_{i=1}^{m}(1-\sigma_i^2)$ measures the linear dependence between elements of $\mathbf{x}$ and $\mathbf{y}$.

When the composite vector $\mathbf{z} = [\mathbf{x}^T \ \mathbf{y}^T]^T$ is normally distributed, the rate at which the rank-r estimate $\hat{\mathbf{x}}$ carries information about $\mathbf{x}$ is

$$\begin{aligned} R[r] &= \tfrac{1}{2}\log\det\{\mathbf{R}_{xx}\} - \tfrac{1}{2}\log\det\{\mathbf{Q}_{xx}[r]\} \\ &= -\tfrac{1}{2}\textstyle\sum_{i=1}^{r}\log(1-\sigma_i^2) = R + \tfrac{1}{2}\sum_{i=r+1}^{m}\log(1-\sigma_i^2) \end{aligned} \tag{4.18}$$

where $R = -\frac{1}{2}\sum_{i=1}^{m}\log(1-\sigma_i^2)$ is the rate at which $\mathbf{y}$ (or alternatively, the full-rank estimate of $\mathbf{x}$ from $\mathbf{y}$) carries information about $\mathbf{x}$. The derivation of the information rate $R[r]$ is similar to that of the information rate R in (2.32). As can be seen, the rate at which $\hat{\mathbf{x}}$ carries information about $\mathbf{x}$ is decomposed into the sum of the r largest canonical rates $R_i = -\frac{1}{2}\log(1-\sigma_i^2)$, $i \in [1, r]$.

Various Implementations: We now determine various implementations for the rank-r Wiener filer $\mathbf{H}[r]$ in canonical coordinates and establish the connections between reduced-rank Wiener filtering in canonical coordinates and two-channel CLS filtering.

Naturally, one implementation for the rank-r Wiener filter $\mathbf{H}[r]$ is the one in (4.12), i.e. $\mathbf{H}[r] = \mathbf{R}_{xx}^{1/2}\mathbf{F}\mathbf{\Sigma}[r]\mathbf{G}^T\mathbf{R}_{yy}^{-1/2}$. The interpretation here is that the measurement vector $\mathbf{y}$ is whitened by $\mathbf{R}_{yy}^{-1/2}$, transformed by $\mathbf{G}^T$ to its canonical coordinates $\mathbf{v}$, filtered by the rank-r canonical Wiener filter $\mathbf{\Sigma}[r]$ to produce a rank-r estimate of

canonical coordinates of $\mathbf{x}$, transformed back by $\mathbf{F}$, and finally recolored by $\mathbf{R}_{xx}^{1/2}$ to produce the rank-r estimate $\hat{\mathbf{x}}$. Alternatively, we may write $\mathbf{H}[r]$ as

$$\begin{aligned}\mathbf{H}[r] &= \mathbf{R}_{xx}\mathbf{R}_{xx}^{-T/2}\mathbf{F}\mathbf{\Sigma}[r]\mathbf{G}^T\mathbf{R}_{yy}^{-1/2} \\ &= \mathbf{R}_{xx}\mathbf{D}_x\mathbf{\Sigma}[r]\mathbf{D}_y^T\end{aligned} \tag{4.19}$$

where $\mathbf{D}_x^T = \mathbf{F}^T\mathbf{R}_{xx}^{-1/2}$ and $\mathbf{D}_y^T = \mathbf{G}^T\mathbf{R}_{yy}^{-1/2}$ are the canonical coordinates maps in (3.32), or equivalently two-channel CLS filters in canonical coordinates. This time, the interpretation is that the measurement vector $\mathbf{y}$ is transformed by $\mathbf{D}_y^T$ to its canonical coordinates $\mathbf{v}$, filtered by the rank-r canonical Wiener filter $\mathbf{\Sigma}[r]$ to produce a rank-r estimate of the canonical coordinates of $\mathbf{x}$, transformed back by $\mathbf{D}_x$, and finally colored by $\mathbf{R}_{xx}$ to produce the rank-r estimate $\hat{\mathbf{x}}$.

The rank-r Wiener filter $\mathbf{H}[r]$ may also be implemented with projection matrices, as we now show. Let us partition $\mathbf{F}$ and $\mathbf{G}$ into $\mathbf{F} = [\ \mathbf{F}_r \quad \mathbf{F}_\star\ ]$ and $\mathbf{G} = [\ \mathbf{G}_r \quad \mathbf{G}_\star\ ]$, and define

$$\mathbf{I}[r] = \begin{bmatrix} \mathbf{I}(r) & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{m\times m}, \tag{4.20}$$

where $\mathbf{I}(r)$ is an $r \times r$ identity matrix. Pre-multiplying the thin SVD of the coherence matrix, i.e. $\mathbf{F}^T\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}\mathbf{G} = \mathbf{\Sigma}$, by $\mathbf{I}[r]$ and post-multiplying it by $\mathbf{G}^T\mathbf{R}_{yy}^{-1/2}$ yields

$$\begin{bmatrix} \mathbf{F}_r^T \\ 0 \end{bmatrix} \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} = \mathbf{\Sigma}[r]\mathbf{G}^T\mathbf{R}_{yy}^{-1/2}. \tag{4.21}$$

Substituting for $\mathbf{\Sigma}[r]\mathbf{G}^T\mathbf{R}_{yy}^{-1/2}$ in (4.5) yields the implementation

$$\mathbf{H}[r] = \mathbf{R}_{xx}^{1/2}\mathbf{F}_r\mathbf{F}_r^T\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} = \mathbf{R}_{xx}^{1/2}\mathbf{P}_{\mathbf{F}_r}\mathbf{R}_{xx}^{-1/2}\mathbf{H}, \tag{4.22}$$

where $\mathbf{P}_{\mathbf{F}_r} = \mathbf{F}_r\mathbf{F}_r^T$ is the orthogonal projection onto the span of $\mathbf{F}_r$ and $\mathbf{H} = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}$ is the full-rank Wiener filter. The interpretation here is that the output of the full-rank Wiener filter $\mathbf{H}$ is whitened by $\mathbf{R}_{xx}^{-1/2}$, projected onto the subspace $< \mathbf{F}_r >$, and re-colored by $\mathbf{R}_{xx}^{1/2}$. Alternatively, pre-multiplying the thin SVD $\mathbf{F}^T\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}\mathbf{G} =$

$\mathbf{\Sigma}$ by $\mathbf{R}_{xx}^{1/2}\mathbf{F}$ and post-multiplying it by $\mathbf{I}[r]$ yields

$$\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}[\ \mathbf{G}_r \quad 0\ ] = \mathbf{R}_{xx}^{1/2}\mathbf{F}\mathbf{\Sigma}[r]. \tag{4.23}$$

Substituting for $\mathbf{R}_{xx}^{1/2}\mathbf{F}\mathbf{\Sigma}[r]$ in (4.5) yields the implementation

$$\mathbf{H}[r] = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}\mathbf{G}_r\mathbf{G}_r^T\mathbf{R}_{yy}^{-1/2} = \mathbf{H}\mathbf{R}_{yy}^{1/2}\mathbf{P}_{\mathbf{G}_r}\mathbf{R}_{yy}^{-1/2}, \tag{4.24}$$

where $\mathbf{P}_{\mathbf{G}_r} = \mathbf{G}_r\mathbf{G}_r^T$ is the orthogonal projection onto the span of $\mathbf{G}_r$. The interpretation here is that the measurement vector $\mathbf{y}$ is whitened by $\mathbf{R}_{yy}^{-1/2}$, projected onto the subspace $< \mathbf{G}_r >$, re-colored by $\mathbf{R}_{yy}^{1/2}$, and then filtered by the full-rank Wiener filter $\mathbf{H}$.

Partitioning $\mathbf{D}_x = \mathbf{R}_{xx}^{-T/2}\mathbf{F}$ and $\mathbf{D}_y = \mathbf{R}_{yy}^{-T/2}\mathbf{G}$ into $\mathbf{D}_x = \begin{bmatrix} \mathbf{D}_{x,r} & \mathbf{D}_{x,\star} \end{bmatrix} = \mathbf{R}_{xx}^{-T/2}[\ \mathbf{F}_r \quad \mathbf{F}_\star\ ]$ and $\mathbf{D}_y = \begin{bmatrix} \mathbf{D}_{y,r} & \mathbf{D}_{y,\star} \end{bmatrix} = \mathbf{R}_{yy}^{-T/2}[\ \mathbf{G}_r \quad \mathbf{G}_\star\ ]$ yields

$$\mathbf{R}_{xx}^{-T/2}\mathbf{F}_r = \mathbf{D}_{x,r} \quad \text{and} \quad \mathbf{R}_{yy}^{-T/2}\mathbf{G}_r = \mathbf{D}_{y,r}. \tag{4.25}$$

Using (4.25), the rank-r Wiener filter in (4.5) and (4.24) may be implemented as

$$\begin{aligned} \mathbf{H}[r] &= \mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T\mathbf{H} = \mathbf{P}_{\mathbf{D}_{x,r}}\mathbf{H}, \quad \text{and} \\ \mathbf{H}[r] &= \mathbf{H}\mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T = \mathbf{H}\mathbf{P}_{\mathbf{D}_{y,r}} \end{aligned} \tag{4.26}$$

where $\mathbf{P}_{\mathbf{D}_{x,r}}$ and $\mathbf{P}_{\mathbf{D}_{y,r}}$ are the following oblique projection [61], [62] operators:

$$\mathbf{P}_{\mathbf{D}_{x,r}} = \mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T \quad \text{and} \quad \mathbf{P}_{\mathbf{D}_{y,r}} = \mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T. \tag{4.27}$$

In the first implementation in (4.26), the output of the full-rank Wiener filter $\mathbf{H}$ is obliquely projected, using $\mathbf{P}_{\mathbf{D}_{x,r}}$, onto the subspace $< \mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T >$, while in the second implementation the measurement $\mathbf{y}$ is obliquely projected, using $\mathbf{P}_{\mathbf{D}_{y,r}}$, onto the subspace $< \mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T >$ and then filtered by the full-rank Wiener filter $\mathbf{H}$.

Therefore, the rank-r Wiener filter $\mathbf{H}[r]$ has six equivalent representations, depicted in Figure 4.1. Reading down the left hand side, from (a) to (c) to (e), produces various implementations for $\mathbf{H}[r]$ in the *orthogonal* coordinates of $\mathbf{F}_r$ and $\mathbf{G}_r$, using

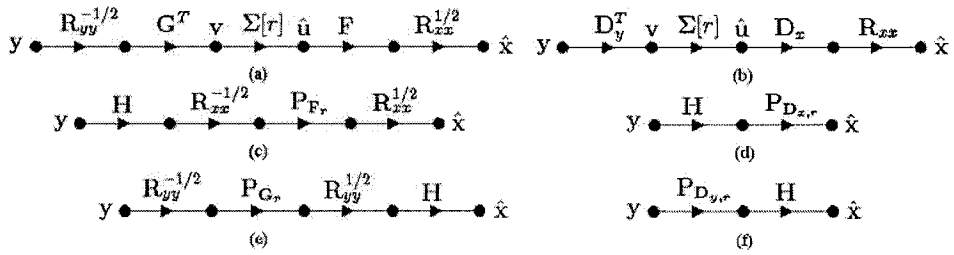

Figure 4.1: Equivalent representations of the rank-r Wiener filter in canonical coordinates. (a) $\mathbf{H}[r] = \mathbf{R}_{xx}^{1/2}\mathbf{F}\mathbf{\Sigma}[r]\mathbf{G}^T\mathbf{R}_{yy}^{-1/2}$. (b) $\mathbf{H}[r] = \mathbf{R}_{xx}\mathbf{D}_x\mathbf{\Sigma}[r]\mathbf{D}_y^T$. (c) $\mathbf{H}[r] = \mathbf{R}_{xx}^{1/2}\mathbf{P}_{\mathbf{F}_r}\mathbf{R}_{xx}^{-1/2}\mathbf{H}$. (d) $\mathbf{H}[r] = \mathbf{P}_{\mathbf{D}_{x,r}}\mathbf{H}$. (e) $\mathbf{H}\mathbf{R}_{yy}^{1/2}\mathbf{P}_{\mathbf{G}_r}\mathbf{R}_{yy}^{-1/2}$. (f) $\mathbf{H}[r] = \mathbf{H}\mathbf{P}_{\mathbf{D}_{y,r}}$.

the *orthogonal* projections $\mathbf{P}_{\mathbf{F}_r}$ and $\mathbf{P}_{\mathbf{G}_r}$. Reading down the right hand side, from (b) to (d) to (f), produces various implementations for $\mathbf{H}[r]$ in the *nonorthogonal* coordinates of $\mathbf{D}_{x,r}$ and $\mathbf{D}_{y,r}$, using the *oblique* projections $\mathbf{P}_{\mathbf{D}_{x,r}}$ and $\mathbf{P}_{\mathbf{D}_{y,r}}$.

4.2.2 Optimal Reduced-Rank Filtering in Half-Canonical Coordinates

Here the objective is to find the rank-r filter $\mathbf{H}[r]$ that minimizes the trace of the error covariance matrix $\mathbf{Q}_{xx}[r]$ in (4.1). Thus, the measure to be minimized is

$$\begin{aligned} \mathrm{MSE}_{xx}[r] &= \mathrm{tr}\{\mathbf{Q}_{xx}[r]\} \\ &= \mathrm{tr}\{\mathbf{Q}_{xx} + [\mathbf{H} - \mathbf{H}[r]]\mathbf{R}_{yy}[\mathbf{H} - \mathbf{H}[r]]^T\} \end{aligned} \tag{4.28}$$

which we may rewrite as

$$\begin{aligned} \mathrm{MSE}_{xx}[r] &= \mathrm{tr}\{\mathbf{Q}_{xx}\} + \mathrm{tr}\{[\mathbf{H}\mathbf{R}_{yy}^{1/2} - \mathbf{H}[r]\mathbf{R}_{yy}^{1/2}][\mathbf{H}\mathbf{R}_{yy}^{1/2} - \mathbf{H}[r]\mathbf{R}_{yy}^{1/2}]^T\} \\ &= \mathrm{tr}\{\mathbf{Q}_{xx}\} + \mathrm{tr}\{[\mathbf{C}_h - \mathbf{H}[r]\mathbf{R}_{yy}^{1/2}][\mathbf{C}_h - \mathbf{H}[r]\mathbf{R}_{yy}^{1/2}]^T\} \end{aligned} \tag{4.29}$$

with $\mathbf{C}_h = \mathbf{H}\mathbf{R}_{yy}^{1/2} = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}$ being the half-coherence matrix defined in Chapter 2. The first term on the right side of this equation, i.e. $\mathrm{tr}\{\mathbf{Q}_{xx}\}$, is the MSE in estimating $\mathbf{x}$ from $\mathbf{y}$ with the full-rank Wiener filter $\mathbf{H} = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}$, and is fixed. The second term,

$$\epsilon_{xx}^2 = \mathrm{tr}\{[\mathbf{C}_h - \mathbf{H}[r]\mathbf{R}_{yy}^{1/2}][\mathbf{C}_h - \mathbf{H}[r]\mathbf{R}_{yy}^{1/2}]^T\} \tag{4.30}$$

is the Frobenius norm of the matrix $\mathbf{C}_h - \mathbf{H}[r]\mathbf{R}_{yy}^{1/2}$, which measures the extra variance introduced by rank reduction [8], [10]. The optimum choice for the rank-r Wiener

filter $\mathbf{H}[r]$ is the rank-r matrix that best approximates the half-coherence matrix $\mathbf{C}_h = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, by minimizing ϵ_{xx}^2 [8], [10]. Thus, it is given by $\mathbf{H}[r]\mathbf{R}_{yy}^{1/2} = \mathbf{U}\mathbf{\Sigma}[r]\mathbf{V}^T$ or,

$$\mathbf{H}[r] = \mathbf{U}\mathbf{\Sigma}[r]\mathbf{V}^T\mathbf{R}_{yy}^{-1/2}; \quad \mathbf{\Sigma}[r] = \begin{bmatrix} \mathbf{\Sigma}(r) & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{m\times m}, \tag{4.31}$$

where $\mathbf{\Sigma}(r) = \text{diag}[\sigma_1, \ldots, \sigma_r]$ is the first $r \times r$ block of the diagonal matrix $\mathbf{\Sigma}$ in the thin SVD $\mathbf{C}_h = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. We note that the diagonal matrix $\mathbf{\Sigma}$ is the half-canonical correlation matrix of the half-canonical correlations σ_i.

Correspondingly, $\hat{\mathbf{x}} = \mathbf{H}[r]\mathbf{y}$ is the rank-r Wiener estimate of $\mathbf{x}$ from $\mathbf{y}$. Using (4.31), the optimal value of $\text{MSE}_{xx}[r]$ is

$$\begin{aligned} \text{MSE}_{xx}[r] &= \text{tr}\{\mathbf{Q}_{xx}\} + \text{tr}\{(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T - \mathbf{U}\mathbf{\Sigma}[r]\mathbf{V}^T)(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T - \mathbf{U}\mathbf{\Sigma}[r]\mathbf{V}^T)^T\} \\ &= \text{tr}\{\mathbf{Q}_{xx}\} + \text{tr}\{\mathbf{U}(\mathbf{\Sigma} - \mathbf{\Sigma}[r])\mathbf{V}^T\mathbf{V}(\mathbf{\Sigma} - \mathbf{\Sigma}[r])^T\mathbf{U}^T\} \\ &= \text{tr}\{\mathbf{Q}_{xx}\} + \underbrace{\text{tr}\{\mathbf{\Sigma}\mathbf{\Sigma}^T - \mathbf{\Sigma}\mathbf{\Sigma}[r]^T - \mathbf{\Sigma}[r]\mathbf{\Sigma}^T + \mathbf{\Sigma}[r]\mathbf{\Sigma}[r]^T\}}_{\sum_{i=r+1}^{m}\sigma_i^2} \\ &= \text{tr}\{\mathbf{Q}_{xx}\} + \text{tr}\{\mathbf{\Sigma}\mathbf{\Sigma}^T - \mathbf{\Sigma}[r]\mathbf{\Sigma}[r]^T\} \\ &= \text{MSE}_{xx} + \textstyle\sum_{i=r+1}^{m}\sigma_i^2 \end{aligned} \tag{4.32}$$

which is the MSE of the rank-r Wiener estimator $\mathbf{H}[r]$. The first term on the right hand side of (4.32) is the minimum MSE for the full-rank estimator of $\mathbf{x}$ from $\mathbf{y}$, i.e. $\hat{\mathbf{x}} = \mathbf{H}\mathbf{y}$, and the second term is the extra MSE ϵ_{xx}^2 due to rank reduction.

Various Implementations: We now determine various implementations for the rank-r Wiener filer $\mathbf{H}[r]$ in half-canonical coordinates, clarifying in the process the connections between reduced-rank Wiener filtering in half-canonical coordinates and two-channel CLS filtering.

Naturally, one implementation for the rank-r Wiener filter $\mathbf{H}[r]$ is the one in (4.31), i.e. $\mathbf{H}[r] = \mathbf{U}\mathbf{\Sigma}[r]\mathbf{V}^T\mathbf{R}_{yy}^{-1/2}$. Here the interpretation is that the measurement vector $\mathbf{y}$ is whitened by $\mathbf{R}_{yy}^{-1/2}$, transformed by $\mathbf{V}^T$ to its half-canonical coordinates $\mathbf{v}$, filtered by the rank-r (half-canonical) Wiener filter $\mathbf{\Sigma}[r]$ to produce a rank-r estimate

of half-canonical coordinates of $\mathbf{x}$, and finally transformed back by $\mathbf{U}$ to produce the rank-r estimate $\hat{\mathbf{x}}$. Alternatively, we may write $\mathbf{H}[r]$ as

$$\begin{aligned}\mathbf{H}[r] &= \mathbf{U}\mathbf{\Sigma}[r]\mathbf{V}^T\mathbf{R}_{yy}^{-1/2} \\ &= \mathbf{D}_x\mathbf{\Sigma}[r]\mathbf{D}_y^T\end{aligned} \tag{4.33}$$

where $\mathbf{D}_x^T = \mathbf{U}^T$ and $\mathbf{D}_y^T = \mathbf{V}^T\mathbf{R}_{yy}^{-1/2}$ are the half-canonical coordinates maps in (3.38), or equivalently two-channel CLS filters in half-canonical coordinates. This time, the interpretation is that the measurement vector $\mathbf{y}$ is transformed by $\mathbf{D}_y^T$ to its canonical coordinates $\mathbf{v}$, filtered by the rank-r half-canonical Wiener filter $\mathbf{\Sigma}[r]$ to produce a rank-r estimate of half-canonical coordinates of $\mathbf{x}$, and then transformed back by $\mathbf{D}_x$, producing the rank-r estimate $\hat{\mathbf{x}}$.

Similar to Section 4.2.1, we may implement $\mathbf{H}[r]$ with projection matrices. Let us partition $\mathbf{U}$ and $\mathbf{V}$ into $\mathbf{U} = [\ \mathbf{U}_r \quad \mathbf{U}_\star\ ]$ and $\mathbf{V} = [\ \mathbf{V}_r \quad \mathbf{V}_\star\ ]$. Pre-multiplying the thin SVD $\mathbf{U}^T\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}\mathbf{V} = \mathbf{\Sigma}$ by $\mathbf{I}[r]$ and post-multiplying it by $\mathbf{V}^T\mathbf{R}_{yy}^{-1/2}$ yields

$$\begin{bmatrix} \mathbf{U}_r^T \\ 0 \end{bmatrix} \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} = \begin{bmatrix} \mathbf{U}_r^T \\ 0 \end{bmatrix} \mathbf{H} = \mathbf{\Sigma}[r]\mathbf{V}^T\mathbf{R}_{yy}^{-1/2}. \tag{4.34}$$

Substituting for $\mathbf{\Sigma}[r]\mathbf{V}^T\mathbf{R}_{yy}^{-1/2}$ in (4.31) yields the implementation

$$\mathbf{H}[r] = \mathbf{U}_r\mathbf{U}_r^T\mathbf{H} = \mathbf{P}_{\mathbf{U}_r}\mathbf{H}, \tag{4.35}$$

where $\mathbf{P}_{\mathbf{U}_r} = \mathbf{U}_r\mathbf{U}_r^T$ is the orthogonal projection onto the span of $\mathbf{U}_r$. The interpretation is that a full-rank Wiener filter $\mathbf{H}$ is followed by a projection onto the subspace $< \mathbf{U}_r >$. Alternatively, pre-multiplying the thin SVD $\mathbf{U}^T\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}\mathbf{V} = \mathbf{\Sigma}$ by $\mathbf{U}$ and post-multiplying it by $\mathbf{I}[r]$ yields

$$\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}[\ \mathbf{V}_r \quad 0\ ] = \mathbf{U}\mathbf{\Sigma}[r]. \tag{4.36}$$

Substituting for $\mathbf{U}\mathbf{\Sigma}[r]$ in (4.31) yields the implementation

$$\mathbf{H}[r] = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}\mathbf{V}_r\mathbf{V}_r^T\mathbf{R}_{yy}^{-1/2} = \mathbf{H}\mathbf{R}_{yy}^{1/2}\mathbf{P}_{\mathbf{V}_r}\mathbf{R}_{yy}^{-1/2}, \tag{4.37}$$

where $\mathbf{P}_{\mathbf{V}_r} = \mathbf{V}_r\mathbf{V}_r^T$ is the orthogonal projection onto the subspace $< \mathbf{V}_r >$. The interpretation here is that the measurement vector $\mathbf{y}$ is whitened by $\mathbf{R}_{yy}^{-1/2}$, projected onto the subspace $< \mathbf{V}_r >$, re-colored by $\mathbf{R}_{yy}^{1/2}$, and then filtered by the full-rank Wiener filter $\mathbf{H}$.

Partitioning $\mathbf{D}_x = \mathbf{U}$ and $\mathbf{D}_y = \mathbf{R}_{yy}^{-T/2}\mathbf{V}$ into $\mathbf{D}_x = \begin{bmatrix} \mathbf{D}_{x,r} & \mathbf{D}_{x,\star} \end{bmatrix} = [\ \mathbf{U}_r \quad \mathbf{U}_\star\ ]$ and $\mathbf{D}_y = \begin{bmatrix} \mathbf{D}_{y,r} & \mathbf{D}_{y,\star} \end{bmatrix} = \mathbf{R}_{yy}^{-T/2}[\ \mathbf{V}_r \quad \mathbf{V}_\star\ ]$ yields

$$\mathbf{U}_r = \mathbf{D}_{x,r} \quad \text{and} \quad \mathbf{R}_{yy}^{-T/2}\mathbf{V}_r = \mathbf{D}_{y,r}. \tag{4.38}$$

Using (4.38), the rank-r Wiener filter in (4.35) and (4.37) may be implemented as

$$\begin{aligned} \mathbf{H}[r] &= \mathbf{D}_{x,r}\mathbf{D}_{x,r}^T\mathbf{H} = \mathbf{P}_{\mathbf{D}_{x,r}}\mathbf{H} \text{ and} \\ \mathbf{H}[r] &= \mathbf{H}\mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T = \mathbf{H}\mathbf{P}_{\mathbf{D}_{y,r}} \end{aligned} \tag{4.39}$$

where $\mathbf{P}_{\mathbf{D}_{x,r}}$ and $\mathbf{P}_{\mathbf{D}_{y,r}}$ are the following orthogonal and oblique projection operators:

$$\mathbf{P}_{\mathbf{D}_{x,r}} = \mathbf{D}_{x,r}\mathbf{D}_{x,r}^T \quad \text{and} \quad \mathbf{P}_{\mathbf{D}_{y,r}} = \mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T. \tag{4.40}$$

In the first implementation in (4.39), the output of the full-rank Wiener filter $\mathbf{H}$ is projected, using the orthogonal projection $\mathbf{P}_{\mathbf{D}_{x,r}}$, onto the subspace $< \mathbf{D}_{x,r} >$, while in the second implementation the measurement vector $\mathbf{y}$ is obliquely projected, using $\mathbf{P}_{\mathbf{D}_{y,r}}$, onto the subspace $< \mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T >$ and then filtered by the full-rank Wiener filter $\mathbf{H}$.

Therefore, similar to Section 4.2.1, the rank-r Wiener filter $\mathbf{H}[r]$ associated with the measure in (4.28) has six equivalent representations. These representations are depicted in Figure 4.2. Reading down the left hand side, from (a) to (c) to (e), shows various implementations for $\mathbf{H}[r]$ in the coordinates of $\mathbf{U}_r$ and $\mathbf{V}_r$. Reading down the right hand side, from (b) to (d) to (f), shows various implementations for $\mathbf{H}[r]$ in the coordinates of $\mathbf{D}_{x,r}$ and $\mathbf{D}_{y,r}$.

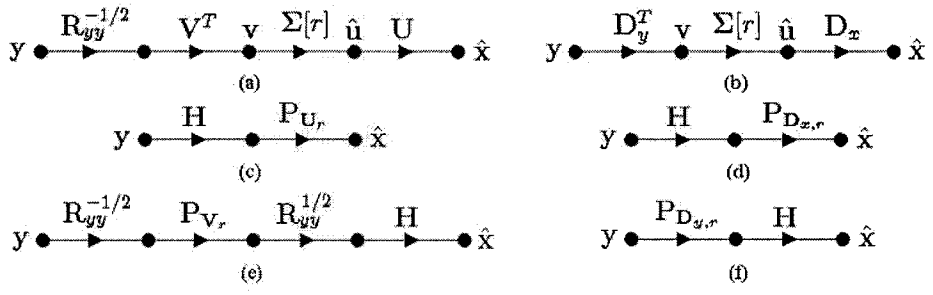

Figure 4.2: Equivalent representations of the rank-r Wiener filter in half-canonical coordinates. (a) $\mathbf{H}[r] = \mathbf{U}\mathbf{\Sigma}[r]\mathbf{V}^T\mathbf{R}_{yy}^{-1/2}$. (b) $\mathbf{H}[r] = \mathbf{D}_x\mathbf{\Sigma}[r]\mathbf{D}_y^T$. (c) $\mathbf{H}[r] = \mathbf{P}_{\mathbf{U}_r}\mathbf{H}$. (d) $\mathbf{H}[r] = \mathbf{P}_{\mathbf{D}_{x,r}}\mathbf{H}$. (e) $\mathbf{H}[r] = \mathbf{H}\mathbf{R}_{yy}^{1/2}\mathbf{P}_{\mathbf{V}_r}\mathbf{R}_{yy}^{-1/2}$. (f) $\mathbf{H}[r] = \mathbf{H}\mathbf{P}_{\mathbf{D}_{y,r}}$.

4.3 Conclusions

In this chapter, a unified framework for deriving three different classes of reduced-rank Wiener filters was presented, with each class corresponding to a particular error measure for reduced-rank estimation. Two of the classes, corresponding to whitened MSE and volume of the concentration ellipse, are equivalent and canonical coordinates are optimal for reduced-rank Wiener filtering under their corresponding error measures. For the third class, which corresponds to MSE estimation, half-canonical coordinates are optimal for reduced-rank Wiener filtering [8]. Our results reproduce what is known from [9]. However, we have derived all of these results in a unified way, using the line of argument presented in [8] for reduced-rank Wiener filtering in half-canonical coordinates. Additionally, we have presented several implementations for reduced-rank Wiener filters in each class, and clarified the connections between reduced-rank Wiener filters and two-channel CLS filters.

CHAPTER 5

COMPUTING CANONICAL COORDINATE AND HALF- CANONICAL COORDINATE MAPPINGS

5.1 Introduction

A conventional method for canonical coordinate decomposition, like the one presented in Chapter 2, does not offer a simple way for computing a small subset of canonical coordinates for reduced-rank estimation or low-rank modelling. A full SVD of the coherence matrix has to be computed, regardless of the rank-reduction. There are indeed simple and fast algorithms to compute the principal singular vectors of a coherence matrix, e.g. [57], [63]– [67], but the coherence matrix itself requires the computation of the square-root-inverses of the channel covariance matrices $\mathbf{R}_{xx}$ and $\mathbf{R}_{yy}$. In addition, the conventional method does not allow an easy update of the canonical coordinate mappings in time for online applications. A similar argument may be made about a conventional method of half-canonical coordinate decomposition.

In this chapter, our goal is to develop simple methods for recursive computation of canonical coordinates and half-canonical coordinates. In Section 5.2, we derive various *alternating power methods*, with *deflation*, to *recursively* compute the canonical coordinate mapping vectors (columns of the canonical coordinate maps

$\mathbf{D}_x = [\mathbf{d}_{x,1}, \ldots, \mathbf{d}_{x,m}]$ and $\mathbf{D}_y = [\mathbf{d}_{y,1}, \ldots, \mathbf{d}_{y,m}]$), one-by-one or in groups. These algorithms also allow for updating the mapping vectors in time as new samples of the channels are observed. Our alternating power methods may be viewed as *two-step* decompositions of the standard power method [57], [68], [69] as they solve the coupled (asymmetric) generalized eigenvalue problem for the canonical coordinate maps $\mathbf{D}_x$ and $\mathbf{D}_y$ through power iterations. One of the algorithms presented in this section, an alternating *block* power method, has been reported in [15] for solving the PCCA problem discussed in Chapter 3. This algorithm has been generalized here to incorporate deflation by blocks, resulting in an order recursive alternating power method. In Section 5.3, we derive similar alternating power methods for recursively solving the coupled (asymmetric) generalized eigenvalue problem for half-canonical coordinate maps. Provided that the rank-reduction is relatively large and the singular values of the coherence matrix or half-coherence matrix are not close together, the alternating power methods can be more efficient in computation than the conventional methods, as they do not require any matrix square-roots.

It must be mentioned that the alternating power methods presented in this chapter are identical in form to those derived in [9] for computing the $\mathbf{AB}^T$ factorization of the rank-r Wiener filter $\mathbf{H}[r] = \mathbf{AB}^T$. However, the algorithms in [9] do not yield the canonical and half-canonical coordinate maps, and the corresponding canonical and half-canonical correlation matrices. Thus, the original contribution here is the discovery that alternating power methods may be used to compute canonical and half-canonical coordinate maps and correlations, making them more applicable in signal processing problems than they would appear from the work in [9]. We note that the material presented in this chapter are also reported in [13].

5.2 Computing Canonical Coordinate Mappings

The power method [57], [68], [69] may be the simplest and oldest method for computing the principal eigenvectors of a matrix. It is also a natural choice for computing the principal subspace of a matrix [67]. Our aim here is to adapt it to the coupled generalized (symmetric) eigenvalue problem in (3.14),

$$\begin{aligned} \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{D}_x &= \mathbf{R}_{xx}\mathbf{D}_x\boldsymbol{\Sigma}^2 \\ \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{D}_y &= \mathbf{R}_{yy}\mathbf{D}_y\boldsymbol{\Sigma}^2 \end{aligned} \tag{5.1}$$

or equivalently to the coupled (asymmetric) generalized eigenvalue problem in (3.13),

$$\begin{aligned} \mathbf{R}_{xy}\mathbf{D}_y &= \mathbf{R}_{xx}\mathbf{D}_x\boldsymbol{\Sigma} \\ \mathbf{R}_{yx}\mathbf{D}_x &= \mathbf{R}_{yy}\mathbf{D}_y\boldsymbol{\Sigma} \end{aligned} \tag{5.2}$$

to solve for the canonical coordinate maps $\mathbf{D}_x$ and $\mathbf{D}_y$.

A standard power method [57], [68], [69] for computing the first columns of $\mathbf{D}_x$ and $\mathbf{D}_y$, i.e. $\mathbf{d}_{x,1}$ and $\mathbf{d}_{y,1}$, associated with the dominant eigenvalue of (5.1), may be summarized as follows. Let k denote the index of iteration, start with a random choice for $\mathbf{d}_{x,1}(0) \in \mathbb{R}^m$ and $\mathbf{d}_{y,1}(0) \in \mathbb{R}^n$, and iterate the following equations on k until convergence:

$$\begin{cases} \bar{\mathbf{d}}_{x,1}(k+1) = (\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx})^{k+1}\mathbf{d}_{x,1}(k) \\ \mathbf{d}_{x,1}(k+1) = \bar{\mathbf{d}}_{x,1}(k+1)(\bar{\mathbf{d}}_{x,1}(k+1)^T\mathbf{R}_{xx}\bar{\mathbf{d}}_{x,1}(k+1))^{-1/2} \\ \bar{\mathbf{d}}_{y,1}(k+1) = (\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy})^{k+1}\mathbf{d}_{y,1}(k) \\ \mathbf{d}_{y,1}(k+1) = \bar{\mathbf{d}}_{y,1}(k+1)(\bar{\mathbf{d}}_{y,1}(k+1)^T\mathbf{R}_{yy}\bar{\mathbf{d}}_{y,1}(k+1))^{-1/2}. \end{cases} \tag{5.3}$$

The normalization to obtain $\mathbf{d}_{x,1}$ from $\bar{\mathbf{d}}_{x,1}$ and $\mathbf{d}_{y,1}$ from $\bar{\mathbf{d}}_{y,1}$ ensures that $\mathbf{d}_{x,1}^T\mathbf{R}_{xx}\mathbf{d}_{x,1}$ $= 1$ and $\mathbf{d}_{y,1}^T\mathbf{R}_{yy}\mathbf{d}_{y,1} = 1$ for each iteration k.

5.2.1 Alternating Power Method

A simpler algorithm can be developed based on the coupled (asymmetric) generalized eigenvalue problem in (5.2) to find $\mathbf{d}_{x,1}$ and $\mathbf{d}_{y,1}$. Let us rewrite (5.2) as

$$\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{D}_y = \mathbf{D}_x\boldsymbol{\Sigma} \tag{5.4}$$

$$\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{D}_x = \mathbf{D}_y\boldsymbol{\Sigma}. \tag{5.5}$$

These equations suggest an alternating sequence of approximations to $\mathbf{D}_x$ and $\mathbf{D}_y$. Given a random initial guess $\mathbf{d}_{y,1}(0)$, the first estimate of $\mathbf{d}_{x,1}$ is computed from (5.4). With this estimate of $\mathbf{d}_{x,1}$, (5.5) is used to compute a new estimate of $\mathbf{d}_{y,1}$. This iterative alternation between (5.4) and (5.5) continues until convergence. We can summarize the algorithm as follows. Randomly select $\mathbf{d}_{y,1}(0) \in \mathbb{R}^n$, start with $k = 0$, and iterate the following equations on k until convergence:

$$\begin{cases} \bar{\mathbf{d}}_{x,1}(k+1) = (\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy})\mathbf{d}_{y,1}(k) \\ \mathbf{d}_{x,1}(k+1) = \bar{\mathbf{d}}_{x,1}(k+1)(\bar{\mathbf{d}}_{x,1}(k+1)^T\mathbf{R}_{xx}\bar{\mathbf{d}}_{x,1}(k+1))^{-1/2} \\ \bar{\mathbf{d}}_{y,1}(k+1) = (\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx})\mathbf{d}_{x,1}(k+1) \\ \mathbf{d}_{y,1}(k+1) = \bar{\mathbf{d}}_{y,1}(k+1)(\bar{\mathbf{d}}_{y,1}(k+1)^T\mathbf{R}_{yy}\bar{\mathbf{d}}_{y,1}(k+1))^{-1/2}. \end{cases} \tag{5.6}$$

This algorithm is a *two-step* decomposition of the standard power method of (5.3). Therefore, the convergence of it follows from the convergence of the standard power method. Provided that the first eigenvalue of (5.2) is larger than the second one (i.e. $\sigma_1 > \sigma_2$) and the initial guess of $\mathbf{d}_{y,1}$ is not orthogonal to $\mathbf{d}_{y,1}$ (i.e. $\mathbf{d}_{y,1}(0)^T\mathbf{R}_{yy}\mathbf{d}_{y,1} \neq 0$), the estimates $\mathbf{d}_{y,1}(k)$ and $\mathbf{d}_{x,1}(k)$ converge to the true $\mathbf{d}_{y,1}$ and $\mathbf{d}_{x,1}$. The estimation error at iteration k converges to zero as an exponential function of the ratio of the second eigenvalue to the first one, i.e. as $(\sigma_2/\sigma_1)^k$ [57], [68], [69].

The algorithm in (5.6) may be rewritten as

$$\begin{cases} \mathbf{R}_{xx}\bar{\mathbf{d}}_{x,1}(k+1) = \mathbf{R}_{xy}\mathbf{d}_{y,1}(k) \\ \mathbf{d}_{x,1}(k+1) = \bar{\mathbf{d}}_{x,1}(k+1)(\bar{\mathbf{d}}_{x,1}(k+1)^T\mathbf{R}_{xx}\bar{\mathbf{d}}_{x,1}(k+1))^{-1/2} \\ \mathbf{R}_{yy}\bar{\mathbf{d}}_{y,1}(k+1) = \mathbf{R}_{yx}\mathbf{d}_{x,1}(k+1) \\ \mathbf{d}_{y,1}(k+1) = \bar{\mathbf{d}}_{y,1}(k+1)(\bar{\mathbf{d}}_{y,1}(k+1)^T\mathbf{R}_{yy}\bar{\mathbf{d}}_{y,1}(k+1))^{-1/2}. \end{cases} \tag{5.7}$$

So, at each iteration k, the vectors $\bar{\mathbf{d}}_{x,1}(k+1)$ and $\bar{\mathbf{d}}_{y,1}(k+1)$ are determined by solving the linear systems of equations $\mathbf{R}_{xx}\bar{\mathbf{d}}_{x,1}(k+1) = \mathbf{R}_{xy}\mathbf{d}_{y,1}(k)$ and $\mathbf{R}_{yy}\bar{\mathbf{d}}_{y,1}(k+1) = \mathbf{R}_{yx}\mathbf{d}_{x,1}(k+1)$, respectively. Any standard method for solving a linear system of equations (see e.g. [57]) may be used here. We call the algorithm in (5.7) an *alternating power method*, in the sense that it solves a coupled generalized eigenvalue problem using alternating iterations. As mentioned earlier, it may be viewed as a two-step decomposition of the standard power iterations for computing the principal eigenvector of a matrix.

The alternating power method reported and analyzed in [9][2] for computing the canonical components of a reduced-rank Wiener filter is a generalization of an Iterative Quadratic Minimum Distance (IQMD) method [70]. It furnishes matrices $\mathbf{A}$ and $\mathbf{B}$ that decompose the rank-r Wiener filter $\mathbf{H}[r]$ as $\mathbf{H}[r] = \mathbf{A}\mathbf{B}^T$. The matrices $\mathbf{A}$ and $\mathbf{B}$ are related to the canonical coordinate maps $\mathbf{D}_x$ and $\mathbf{D}_y$ as $\mathbf{A} = \mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{\Sigma}(r)\mathbf{M}$ and $\mathbf{B} = \mathbf{D}_{y,r}\mathbf{M}^{-T}$, where $\mathbf{D}_{x,r}$ and $\mathbf{D}_{y,r}$ contain the first r columns of $\mathbf{D}_x$ and $\mathbf{D}_y$, $\mathbf{\Sigma}(r) = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ contains the first r diagonal elements of $\mathbf{\Sigma}$, and $\mathbf{M}$ is any $r \times r$ nonsingular matrix. Because of the ambiguous matrix $\mathbf{M}$, the algorithm cannot be used to compute the canonical coordinate maps, or their corresponding canonical correlations $\sigma_i = \mathbf{d}_{x,i}^T\mathbf{R}_{xy}\mathbf{d}_{y,i}$.

---

[2]The alternating power method of [9] is not a simple two-step decomposition of the standard power method, and therefore its convergence analysis requires special treatment.

5.2.2 Alternating Block Power Method

If the ratio (σ_2/σ_1) is close to one, then the convergence rate is very slow. One way to address this problem is to combine the block power method of [68] and [69] with the above alternating procedure to solve for several columns of $\mathbf{D}_x$ and $\mathbf{D}_y$. The idea is to start with $l \leq m$ orthogonal vectors and after each iteration use a Gram-Schmidt orthogonalization procedure [57] to guarantee that the constraints $\mathbf{D}_{x,l}^T \mathbf{R}_{xx} \mathbf{D}_{x,l} = \mathbf{I}$ and $\mathbf{D}_{y,l}^T \mathbf{R}_{yy} \mathbf{D}_{y,l} = \mathbf{I}$ are satisfied. This algorithm may be summarized as follows. Initialize $\mathbf{D}_{y,l}(0)$ with orthogonal columns, start with $k = 0$, and iterate the following equations on k until convergence:

$$\begin{cases} \text{Solve } \mathbf{R}_{xx}\bar{\mathbf{D}}_{x,l}(k+1) = \mathbf{R}_{xy}\mathbf{D}_{y,l}(k) \text{ for } \bar{\mathbf{D}}_{x,l}(k+1) \\ \bar{\mathbf{D}}_{x,l}(k+1) \xrightarrow{GSO} \mathbf{D}_{x,l}(k+1) \text{ such that } \mathbf{D}_{x,l}^T(k+1)\mathbf{R}_{xx}\mathbf{D}_{x,l}(k+1) = \mathbf{I} \\ \text{Solve } \mathbf{R}_{yy}\bar{\mathbf{D}}_{y,l}(k+1) = \mathbf{R}_{yx}\mathbf{D}_{x,l}(k+1) \text{ for } \bar{\mathbf{D}}_{y,l}(k+1) \\ \bar{\mathbf{D}}_{y,l}(k+1) \xrightarrow{GSO} \mathbf{D}_{y,l}(k+1) \text{ such that } \mathbf{D}_{y,l}^T(k+1)\mathbf{R}_{yy}\mathbf{D}_{y,l}(k+1) = \mathbf{I} \end{cases} \tag{5.8}$$

where GSO stands for Gram-Schmidt Orthogonalization. The GSO for $\mathbf{D}_{x,l}$ may be summarized as follows. At each iteration k, do the following for $i = 1, 2, \ldots, l$,

$$\begin{aligned} \tilde{\mathbf{d}}_{x,i}(k+1) &= [\mathbf{I} - \tilde{\mathbf{D}}_{x,i-1}(k+1)\tilde{\mathbf{D}}_{x,i-1}^T(k+1)\mathbf{R}_{xx}]\bar{\mathbf{d}}_{x,i}(k+1) \\ \mathbf{d}_{x,i}(k+1) &= \tilde{\mathbf{d}}_{x,i}(k+1)(\tilde{\mathbf{d}}_{x,i}^T(k+1)\mathbf{R}_{xx}\tilde{\mathbf{d}}_{x,i}(k+1))^{-1/2} \end{aligned} \tag{5.9}$$

where $\tilde{\mathbf{D}}_{x,i-1}(k+1) = [\tilde{\mathbf{d}}_{x,1}(k+1), \ldots, \tilde{\mathbf{d}}_{x,i-1}(k+1)]$. A similar set of equations may be written for the GSO for $\mathbf{D}_{y,l}$. In [15], this algorithm was used to iteratively compute the solutions to the PCCA problem.

5.2.3 Alternating Power Method With Deflation

We now extend the previous algorithm by introducing an alternating power method with deflation for computing the canonical coordinate maps. Assume that the first $r < m$ canonical coordinate mappings (the first r columns of $\mathbf{D}_x$ and $\mathbf{D}_y$) have already

been found. Partition $\mathbf{D}_x$, $\mathbf{D}_y$ and $\mathbf{\Sigma}$ into

$$\begin{aligned}\mathbf{D}_x &= \begin{bmatrix} \mathbf{D}_{x,r} & \mathbf{D}_{x,\star} \end{bmatrix}, \ \mathbf{D}_y = \begin{bmatrix} \mathbf{D}_{y,r} & \mathbf{D}_{y,\star} \end{bmatrix}, \text{ and} \\ \mathbf{\Sigma} &= \begin{bmatrix} \mathbf{\Sigma}(r) & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}(\star) \end{bmatrix}. \end{aligned} \tag{5.10}$$

where $\mathbf{D}_{x,r} \in \mathbb{R}^{m \times r}$ and $\mathbf{D}_{y,r} \in \mathbb{R}^{n \times r}$ carry the first r columns of $\mathbf{D}_x \in \mathbb{R}^{m \times m}$ and $\mathbf{D}_y \in \mathbb{R}^{m \times m}$, and $\mathbf{D}_{x,\star} \in \mathbb{R}^{m \times (m-r)}$ and $\mathbf{D}_{y,\star} \in \mathbb{R}^{n \times (m-r)}$ carry their last $m - r$ columns. The diagonal matrices $\mathbf{\Sigma}(r) \in \mathbb{R}^{r \times r}$ and $\mathbf{\Sigma}(\star) \in \mathbb{R}^{(m-r) \times (m-r)}$ carry the first r and the last $m-r$ diagonal elements of the canonical correlation matrix $\mathbf{\Sigma} \in \mathbb{R}^{m \times m}$, respectively.

To compute the $(r+1)$th pair of canonical coordinate mappings, deflate the first r canonical coordinate mappings from the left hand sides of (5.2) to get

$$\begin{aligned}(\mathbf{I} - \mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T)\mathbf{R}_{xy}\mathbf{D}_{y,\star} &= \mathbf{R}_{xx}\mathbf{D}_{x,\star}\mathbf{\Sigma}(\star) \\ (\mathbf{I} - \mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T)\mathbf{R}_{yx}\mathbf{D}_{x,\star} &= \mathbf{R}_{yy}\mathbf{D}_{y,\star}\mathbf{\Sigma}(\star). \end{aligned} \tag{5.11}$$

The $(r+1)$th canonical coordinate mappings $\mathbf{d}_{x,r+1}$ and $\mathbf{d}_{y,r+1}$ are the first columns of $\mathbf{D}_{x,\star}$ and $\mathbf{D}_{y,\star}$, which are now associated with the dominant eigenvalue of (5.11). Appendix A makes this claim precise. Thus, these mappings can be computed by iterating between the two equations in (5.11) with a random initialization. The dominant eigenvalue of (5.11), i.e. $\sigma_{r+1} = \mathbf{d}_{x,r+1}^T\mathbf{R}_{xy}\mathbf{d}_{y,r+1}$, is the $(r+1)$th canonical correlation of $\mathbf{x}$ and $\mathbf{y}$.

The deflation process may also be implemented with projection matrices. Using (4.27) we may rewrite (5.11) as

$$\begin{aligned}(\mathbf{I} - \mathbf{P}_{\mathbf{D}_{x,r}})\mathbf{R}_{xy}\mathbf{D}_{y,\star} &= \mathbf{R}_{xx}\mathbf{D}_{x,\star}\mathbf{\Sigma}(\star) \\ (\mathbf{I} - \mathbf{P}_{\mathbf{D}_{y,r}})\mathbf{R}_{yx}\mathbf{D}_{x,\star} &= \mathbf{R}_{yy}\mathbf{D}_{y,\star}\mathbf{\Sigma}(\star). \end{aligned} \tag{5.12}$$

Thus, the alternating power method for finding $\mathbf{d}_{x,r+1}$ and $\mathbf{d}_{y,r+1}$ may be summarized as follows. Randomly select $\mathbf{d}_{y,r+1}(0) \in \mathbb{R}^n$, set $k = 0$, and iterate the following

equations on k until convergence:

$$\begin{cases} \text{Solve } \mathbf{R}_{xx}\bar{\mathbf{d}}_{x,r+1}(k+1) = (\mathbf{I} - \mathbf{P}_{\mathbf{D}_{x,r}})\mathbf{R}_{xy}\mathbf{d}_{y,r+1}(k) \text{ for } \bar{\mathbf{d}}_{x,r+1}(k+1) \\ \mathbf{d}_{x,r+1}(k+1) = \bar{\mathbf{d}}_{x,r+1}(k+1)(\bar{\mathbf{d}}^T_{x,r+1}(k+1)\mathbf{R}_{xx}\bar{\mathbf{d}}_{x,r+1}(k+1))^{-1/2} \\ \text{Solve } \mathbf{R}_{yy}\bar{\mathbf{d}}_{y,r+1}(k+1) = (\mathbf{I} - \mathbf{P}_{\mathbf{D}_{y,r}})\mathbf{R}_{yx}\mathbf{d}_{x,r+1}(k+1) \text{ for } \bar{\mathbf{d}}_{y,r+1}(k+1) \\ \mathbf{d}_{y,r+1}(k+1) = \bar{\mathbf{d}}_{y,r+1}(k+1)(\bar{\mathbf{d}}^T_{y,r+1}(k+1)\mathbf{R}_{yy}\bar{\mathbf{d}}_{y,r+1}(k+1))^{-1/2}. \end{cases} \tag{5.13}$$

Provided that $\sigma_{r+1} < \sigma_r$ this algorithm yields the $(r+1)$th pair of canonical coordinate mappings at the rate of $(\sigma_{r+1}/\sigma_r)^k$. We note that the alternating block power method in (5.8) may be used to simultaneously compute several columns of $\mathbf{D}_{x,\star}$ and $\mathbf{D}_{y,\star}$, associated with the dominant eigenvalues of the deflated coupled (asymmetric) generalized eigenvalue problem in (5.12). This produces an alternating block power method with deflation.

Unlike the conventional method of canonical coordinate decomposition, the alternating power method, with deflation, in (5.13) does not require any matrix square-roots. Moreover, all operations are matrix-vector multiplications where the number of vectors might be much smaller than the number of columns of the matrix. Therefore, provided that the eigenvalues associated with the desired canonical coordinates are not close to each other, the alternating power method, with deflation, in (5.13) is an efficient algorithm for practical extraction of a few dominant canonical coordinate mappings and canonical correlations.

5.2.4 Order Recursive Alternating Power Method

In most applications, the number of canonical coordinate pairs to be extracted is not known *a priori* or it may vary with time. One may run a test of information rate or linear dependence based on the measure in (4.17) and (4.18) to determine if a pre-specified threshold is met. If the threshold is not reached, additional canonical coordinate pairs must be extracted. However, if the threshold is exceeded, computation of the mapping vectors associated with the less significant canonical coordinate

pairs may be stopped to reduce the computational load. Thus, the alternating power method shall be modified in the next paragraph to allow for changes in the number of columns of $\mathbf{D}_x$ and $\mathbf{D}_y$ to be computed, during the iterations of the algorithm.

The alternating block power method in (5.8) yields $\mathbf{D}_{x,r}$ and $\mathbf{D}_{y,r}$ asymptotically. Thus, (5.8) and (5.13) can be run at the same time, with $\mathbf{D}_{x,r}$ and $\mathbf{D}_{y,r}$ in (5.13) being replaced by $\mathbf{D}_{x,r}(k)$ and $\mathbf{D}_{y,r}(k)$. At each iteration k, the combination of (5.8) and (5.13) can be run successively for $r = 0, 1, \ldots, p-1$, $p \leq m$ to extract up to p columns of $\mathbf{D}_x$ and $\mathbf{D}_y$. This algorithm may be summarized as follows. At each iteration k, do the following for $r = 0, 1, \ldots, p-1$:

$$
\begin{cases}
\mathbf{a}_{r+1}(k+1) = \mathbf{R}_{xy}\mathbf{d}_{y,r+1}(k) \\
\boldsymbol{\alpha}_{r+1}(k+1) = \mathbf{D}_{x,r}^T(k+1)\mathbf{a}_{r+1}(k+1) \\
\mathbf{b}_{r+1}(k+1) = \mathbf{D}_{x,r}(k+1)\boldsymbol{\alpha}_{r+1}(k+1) \\
\boldsymbol{\beta}_{r+1}(k+1) = \mathbf{a}_{r+1}(k+1) - \mathbf{R}_{xx}\mathbf{b}_{r+1}(k+1) \\
\text{Solve } \mathbf{R}_{xx}\bar{\mathbf{d}}_{x,r+1}(k+1) = \boldsymbol{\beta}_{r+1}(k+1) \text{ for } \bar{\mathbf{d}}_{x,r+1}(k+1) \\
\mathbf{d}_{x,r+1}(k+1) = \bar{\mathbf{d}}_{x,r+1}(k+1)(\bar{\mathbf{d}}_{x,r+1}^T(k+1)\boldsymbol{\beta}_{r+1}(k+1))^{-1/2} \\
\mathbf{q}_{r+1}(k+1) = \mathbf{R}_{yx}\mathbf{d}_{x,r+1}(k+1) \\
\boldsymbol{\gamma}_{r+1}(k) = \mathbf{D}_{y,r}^T(k+1)\mathbf{q}_{r+1}(k+1) \\
\mathbf{s}_{r+1}(k+1) = \mathbf{D}_{y,r}(k+1)\boldsymbol{\gamma}_{r+1}(k+1) \\
\boldsymbol{\theta}_{r+1}(k+1) = \mathbf{q}_{r+1}(k+1) - \mathbf{R}_{yy}\mathbf{s}_{r+1}(k+1) \\
\text{Solve } \mathbf{R}_{yy}\bar{\mathbf{d}}_{y,r+1}(k+1) = \boldsymbol{\theta}_{r+1}(k+1) \text{ for } \bar{\mathbf{d}}_{y,r+1}(k+1) \\
\mathbf{d}_{y,r+1}(k+1) = \bar{\mathbf{d}}_{y,r+1}(k+1)(\bar{\mathbf{d}}_{y,r+1}^T(k+1)\boldsymbol{\theta}_{r+1}(k+1))^{-1/2} \\
\mathbf{D}_{x,r+1}(k+1) = [\ \mathbf{D}_{x,r}(k+1) \quad \mathbf{d}_{x,r+1}(k+1)\ ] \\
\mathbf{D}_{y,r+1}(k+1) = [\ \mathbf{D}_{y,r}(k+1) \quad \mathbf{d}_{y,r+1}(k+1)\ ].
\end{cases}
\tag{5.14}
$$

Note that the computations that require $\mathbf{D}_{x,0}$ and $\mathbf{D}_{y,0}$ must be ignored. The value of p may be changed during the iterations of the algorithm to meet the pre-specified

criterion. The above algorithm only involves scalar-vector and vector-matrix multiplications and no matrix-matrix multiplication is required.

5.2.5 Online Implementation

For online implementation the idea is simply to allow the correlation matrices $\mathbf{R}_{xx}$, $\mathbf{R}_{yy}$, and $\mathbf{R}_{xy}$ to be updated as new data become available during the iteration of the alternating power method. This may be done using the standard rank-one update equation

$$\mathbf{R}_{xx}(j) = \delta \mathbf{R}_{xx}(j-1) + \mathbf{x}(j)\mathbf{x}^T(j) \tag{5.15}$$

where $\delta \in (0,1)$ is a forgetting factor. To prevent $\mathbf{R}_{xx}$ from becoming singular at early iterations, $\mathbf{R}_{xx}(0)$ may be chosen $\mathbf{R}_{xx} = \rho^2\mathbf{I}$, where $|\rho|$ is small. After each rank-one update of the covariance matrices, the alternating power method (any version) may be iterated for one or more iterations. During the iterations, the covariance matrices are kept fixed. When the next sample pair of the data is observed, covariance matrices are updated again, and the procedure is repeated. The number of times that the equations in an alternating power method are iterated introduces a trade-off between the accuracy of the algorithm and computational load. This trade-off is illustrated in the simulation examples in Section 5.4.

5.3 Computing Half-Canonical Coordinate Mappings

Similar to Section 5.2, we may derive alternating power methods for solving the coupled (asymmetric) generalized eigenvalue problem in (3.19), i.e.

$$\begin{aligned} \mathbf{R}_{xy}\mathbf{D}_y &= \mathbf{D}_x\boldsymbol{\Sigma} \\ \mathbf{R}_{yx}\mathbf{D}_x &= \mathbf{R}_{yy}\mathbf{D}_y\boldsymbol{\Sigma} \end{aligned} \tag{5.16}$$

to find the half-canonical coordinate mappings. However, it is interesting to note that when $\mathbf{R}_{xx}$ is set to $\mathbf{R}_{xx} = \mathbf{I}$, the generalized eigenvalue problem of (5.2) for

canonical coordinate maps reduces to the generalized eigenvalue problem of (5.16) for half-canonical coordinate maps. Thus, all variations of the alternating power methods introduced for computing the canonical coordinate mappings, may be used to compute the half-canonical coordinate mappings by replacing $\mathbf{R}_{xx}$ with $\mathbf{R}_{xx} = \mathbf{I}$. Note that in the alternating power method with deflation, (5.13), the oblique projection matrix $\mathbf{P}_{\mathbf{D}_{x,r}} = \mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T$ is replaced by the orthogonal projection $\mathbf{P}_{\mathbf{D}_{x,r}} = \mathbf{D}_{x,r}\mathbf{D}_{x,r}^T$.

5.4 Simulation Results

This section demonstrates the correctness of the alternating power method for extracting canonical coordinate mappings on a synthesized data set. The data set is constructed from the channel models

$$\mathbf{x} = \mathbf{H}_{xx}\boldsymbol{\eta}_x$$
$$\mathbf{y} = \mathbf{H}_{yx}\mathbf{x} + \mathbf{H}_{yy}\boldsymbol{\eta}_y$$

where $\mathbf{x} \in \mathbb{R}^4$, and $\mathbf{y} \in \mathbb{R}^5$. The matrices $\mathbf{H}_{xx} \in \mathbb{R}^{4\times4}$, $\mathbf{H}_{yy} \in \mathbb{R}^{5\times5}$ and $\mathbf{H}_{yx} \in \mathbb{R}^{5\times4}$ are known, and $\boldsymbol{\eta}_x \in \mathbb{R}^4$ and $\boldsymbol{\eta}_y \in \mathbb{R}^5$ are two independent white Gaussian vectors.

Let $\hat{\mathbf{d}}_{x,i}$ and $\hat{\mathbf{d}}_{y,i}$ denote the estimates of the ith canonical coordinate mappings $\mathbf{d}_{x,i}$ and $\mathbf{d}_{y,i}$. We define the normalized error norm of the ith estimated canonical coordinate mapping as

$$e_{\mathbf{d}_{x,i}} = \frac{\|\mathbf{d}_{x,i} - \hat{\mathbf{d}}_{x,i}\|}{\|\mathbf{d}_{x,i}\|} \quad \text{and} \quad e_{\mathbf{d}_{y,i}} = \frac{\|\mathbf{d}_{y,i} - \hat{\mathbf{d}}_{y,i}\|}{\|\mathbf{d}_{y,i}\|}.$$

We also define the rank-r group errors of the ideal canonical coordinate mappings $\mathbf{D}_{x,r}$ and $\mathbf{D}_{y,r}$ as

$$E_{\mathbf{D}_{x,r}} = \frac{\|\mathbf{D}_{x,r} - \hat{\mathbf{D}}_{x,r}\|}{\|\mathbf{D}_{x,r}\|} \quad \text{and} \quad E_{\mathbf{D}_{y,r}} = \frac{\|\mathbf{D}_{y,r} - \hat{\mathbf{D}}_{y,r}\|}{\|\mathbf{D}_{y,r}\|}$$

where $\|\mathbf{D}_x\|$ denotes the Frobenius norm $\|\mathbf{D}_x\|^2 = \mathrm{tr}\{\mathbf{D}_x^T\mathbf{D}_x\}$. In these definitions, the vectors $\mathbf{d}_{x,i}$ and $\mathbf{d}_{y,i}$, their estimates $\hat{\mathbf{d}}_{x,i}$ and $\hat{\mathbf{d}}_{y,i}$, the matrices $\mathbf{D}_{x,r}$ and $\mathbf{D}_{y,r}$ and their estimates $\hat{\mathbf{D}}_{x,r}$ and $\hat{\mathbf{D}}_{y,r}$ are normalized in sign.

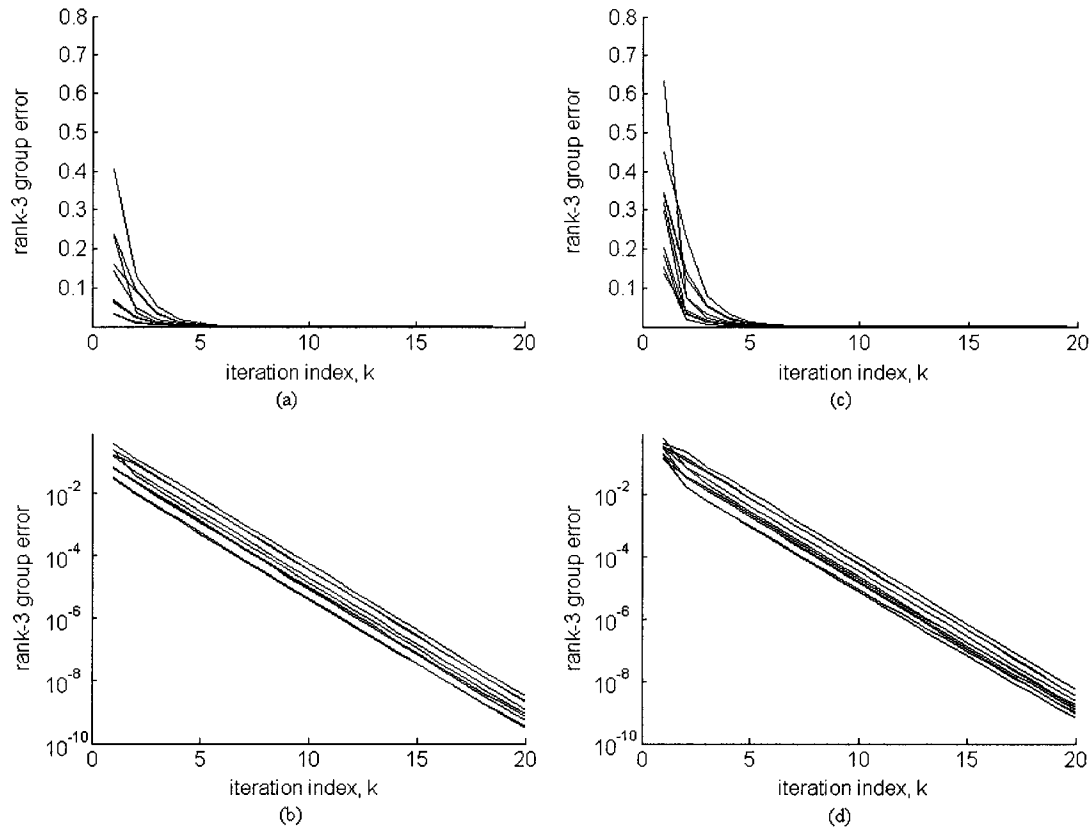

Figure 5.1: Rank-3 group errors for the alternating block power method, in batch mode, with ten independent initializations: (a) $E_{\mathbf{D}_{x,3}}$, linear scale (b) $E_{\mathbf{D}_{x,3}}$, logarithmic scale (c) $E_{\mathbf{D}_{y,3}}$, linear scale (d) $E_{\mathbf{D}_{y,3}}$, logarithmic scale. The results confirm that convergence of the alternating block power method is exponential in iteration number.

Alternating Block Power Method in Batch Mode: The covariance matrices are computed from $N = 500$ samples and kept constant during the iteration of the alternating block power method in (5.8). Figures 5.1(a) and (b) show the rank-3 group errors associated with $\mathbf{D}_x$ and $\mathbf{D}_y$ when ten independent initializations are used. The errors are very small after the seventh iteration. The logarithmic versions of these plots are given in Figures 5.1(c) and (d). The exponential convergence of the algorithm is prominent in these figures.

Alternating Block Power Method in Online Mode: For this case the covariance matrices are updated using the rank-one time updating equation in (5.15) as a new

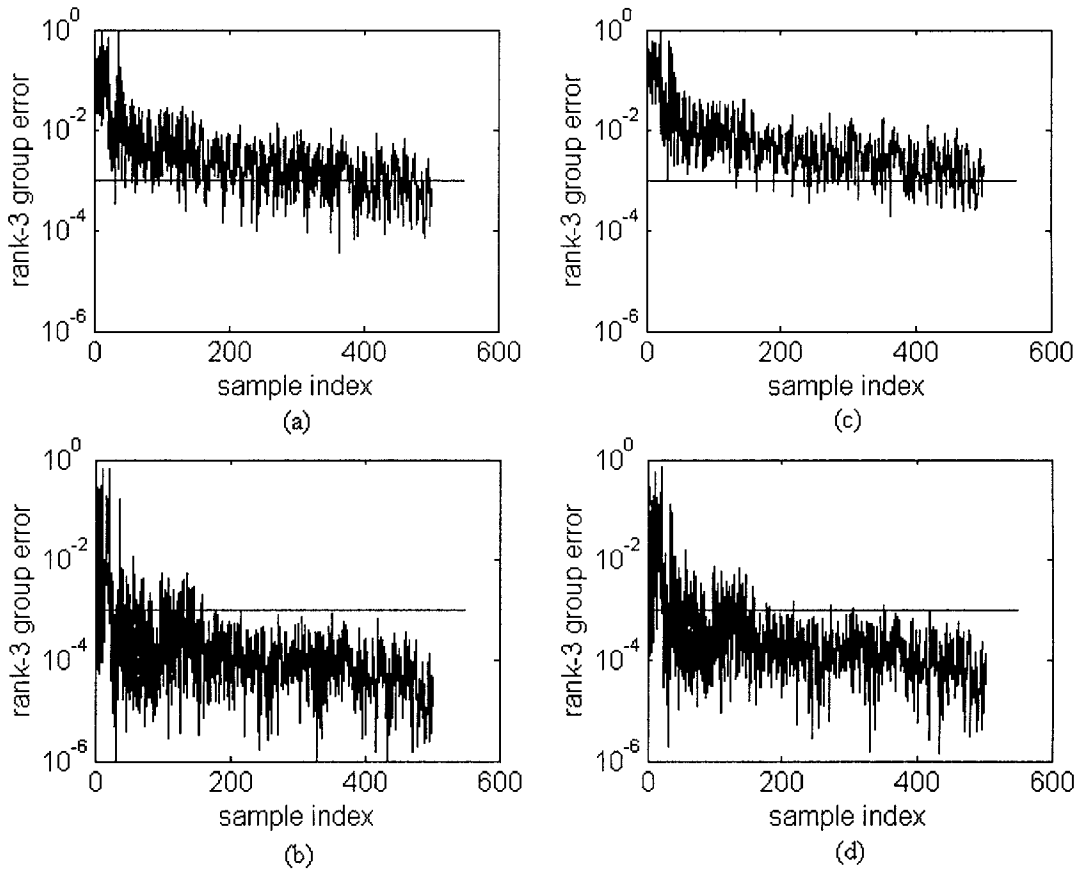

Figure 5.2: Rank-3 group errors for the alternating block power method, in online mode, with a variable number of iterations per sample index: (a) $E_{\mathbf{D}_{x,3}}$, one iteration. (b) $E_{\mathbf{D}_{x,3}}$, four iterations. (c) $E_{\mathbf{D}_{y,3}}$, one iteration. (d) $E_{\mathbf{D}_{y,3}}$, four iterations.

sample pair $\mathbf{x}(j)$ and $\mathbf{y}(j)$ becomes available. After each time (sample) update, the covariance matrices may be used in one or more iterations of the alternating block power method. The forgetting factor used for updating the covariance matrices is chosen to be $\delta = 0.99$, and $\mathbf{R}_{xx}(0)$ and $\mathbf{R}_{yy}(0)$ are set to $0.01\mathbf{I}$. All other assumptions are as in the previous case. Figures 5.2(a)-(d) show the rank-3 group errors associated with $\mathbf{D}_x$ and $\mathbf{D}_y$ versus the iteration index (for ten independent initializations). In these plots, the number of iterations of the algorithm for each new sample pair is one, and four respectively.

Alternating Power Method With Deflation in Batch Mode: The covariance matrices are computed in batch mode from the 500 samples of the data channels and kept fixed during the iterations of the alternating power method in (5.13). We use the algorithm in (5.13) to compute the first three pair of canonical coordinate mappings. The computation of the ith canonical coordinate mappings $\mathbf{d}_{x,i}$ and $\mathbf{d}_{y,i}$, $i \in [1,3]$ is started after the estimates of the $(i-1)$th canonical coordinate mappings have converged. In computing $\mathbf{d}_{x,1}$ and $\mathbf{d}_{y,1}$ the matrices $\mathbf{P}_{\mathbf{D}_{x,0}}$ and $\mathbf{P}_{\mathbf{D}_{y,0}}$ are set to zero. Figures 5.3(a)-(f) show the normalized error norms of the estimated canonical coordinate mappings associated with $\mathbf{d}_{x,i}$ and $\mathbf{d}_{y,i}$, $i \in [1,3]$ vs. iteration number, when ten independent initializations are used. The plots are given in logarithmic scale. The straight decaying lines show that the convergence of the algorithm is exponential in iteration number. The constant error levels in Figures 5.3(c) and (f), after approximately seven iterations, are due to the numerical precision of MATLAB, and the error propagation caused by deflation.

Order Recursive Alternating Power Method in Online Mode: Here the covariance matrices are updated using the rank-one time updating equation (5.15) as a new sample pair $\mathbf{x}(j)$ and $\mathbf{y}(j)$ becomes available. Each updated covariance matrix is used for four iterations in the alternating power method in (5.14) before it is updated again. The forgetting factor used for updating the covariance matrices is $\delta = 0.99$, and $\mathbf{R}_{xx}(0)$ and $\mathbf{R}_{yy}(0)$ are set to $0.01\mathbf{I}$. Figures 5.4(a)-(f) show the normalized error norms of the estimated canonical coordinate mappings $e_{\mathbf{d}_{x,i}}$ and $e_{\mathbf{d}_{y,i}}$, $i \in [1,3]$ versus sample index for ten initializations of the algorithm.

5.5 Conclusions

In this chapter various alternating power methods have been developed which provide simple methods for recursive computation of the canonical coordinate and half-canonical coordinate mapping vectors. In essence, these alternating power methods

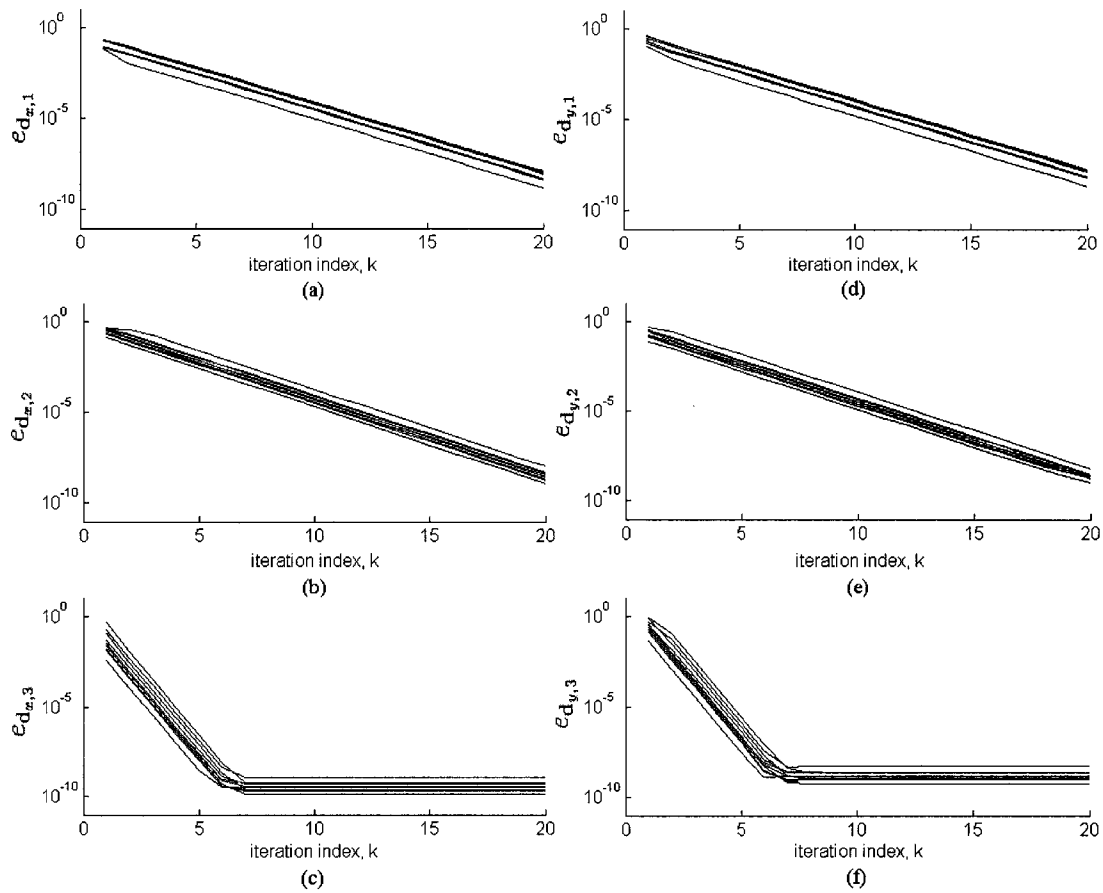

Figure 5.3: Normalized error norms of the estimated canonical coordinate mappings for the alternating power method with deflation, in batch mode, in logarithmic scale with ten independent initializations: (a) $e_{\mathbf{d}_{x,1}}$. (b) $e_{\mathbf{d}_{x,2}}$. (c) $e_{\mathbf{d}_{x,3}}$. (d) $e_{\mathbf{d}_{y,1}}$. (e) $e_{\mathbf{d}_{y,2}}$. (f) $e_{\mathbf{d}_{y,3}}$. The results confirm that convergence of the alternating power method with deflation is exponential in iteration number.

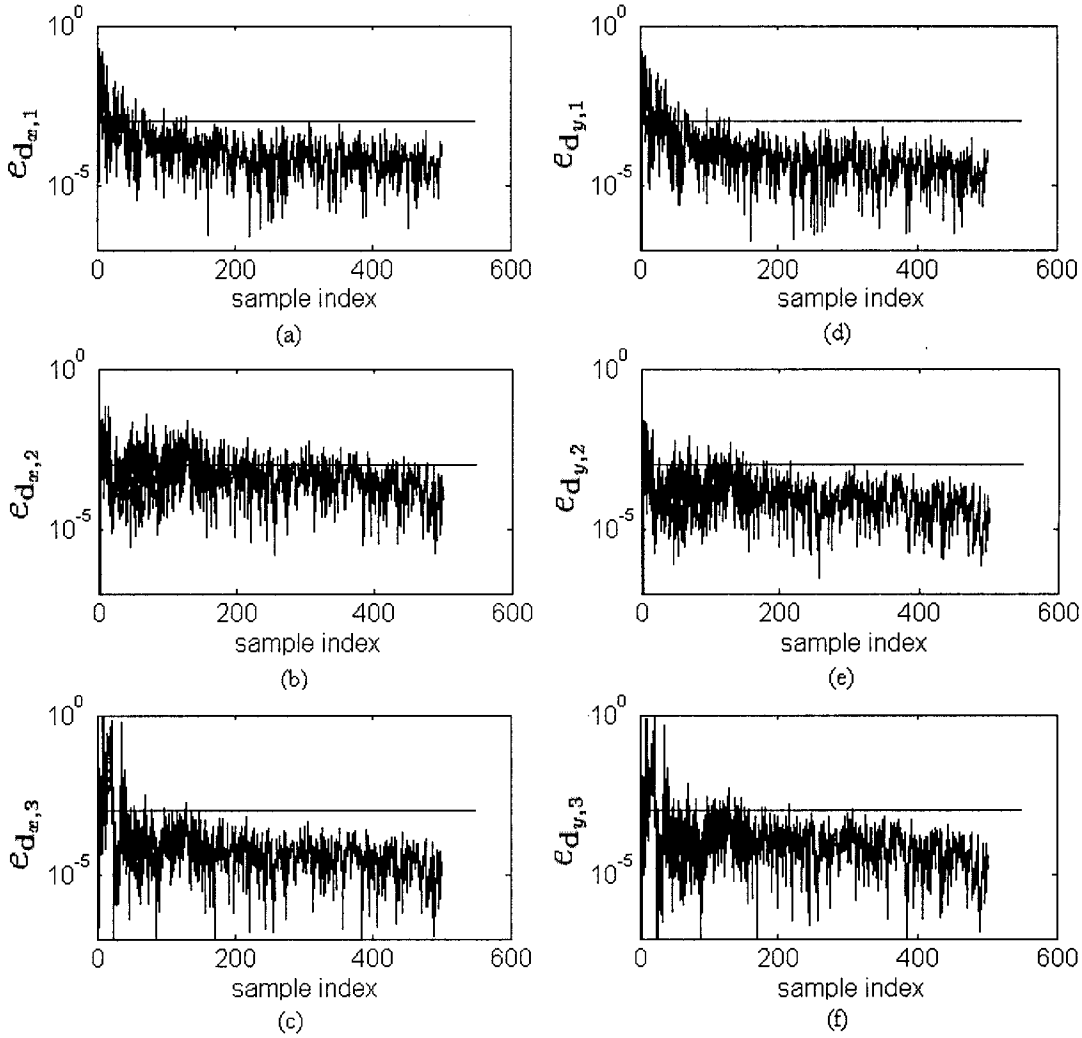

Figure 5.4: Normalized error norms of the estimated canonical coordinate mappings for the order recursive alternating power method, in online mode, in logarithmic scale with ten independent initializations: (a) $e_{\mathbf{d}_{x,1}}$. (b) $e_{\mathbf{d}_{x,2}}$. (c) $e_{\mathbf{d}_{x,3}}$. (d) $e_{\mathbf{d}_{y,1}}$. (e) $e_{\mathbf{d}_{y,2}}$. (f) $e_{\mathbf{d}_{y,3}}$.

may be viewed as *two-step* decompositions of the standard power method, as they solve a coupled (asymmetric) generalized eigenvalue problem through power iterations. They may be used in deflation, block, or block-deflation mode, allowing for computation of the canonical coordinate and half-canonical coordinate mapping vectors, one by one, or in groups. Moreover, they may be used in batch mode on a fixed data sample, or in online mode for updating the mapping vectors in time. Provided that the rank-reduction is relatively large and the singular values of the coherence or half-coherence matrix are not close together, the alternating power methods can be more efficient in computation than the conventional methods, as they do not require any matrix square-roots. As established in Chapter 4, reduced-rank Wiener filters may be implemented in terms of canonical coordinate and half-canonical coordinate maps. Thus, the alternating power methods developed here are also simple algorithms for computing reduced-rank Wiener filters, regardless of the coordinate system.

CHAPTER 6

A NETWORK FOR RECURSIVE EXTRACTION OF CANONICAL COORDINATES

6.1 Introduction

In 1982, Oja [22] showed that a linear network with a single node trained with a normalized Hebbian-type rule can extract the dominant principal component of a stationary vector process. Sanger [23] and Foldik [24] extended Oja's work to the multi-node case to simultaneously extract the first m principal components of a vector process. Diamantaras and Kung [25], [26] exploited the idea of using lateral connections with anti-Hebbian learning to recursively extract the principal components. In a different approach, based on recursive least squares (RLS) learning, Bannour and Azimi-Sadjadi [27] proposed another structure for recursive extraction of principal components.

Perhaps the most interesting aspect of the work in [25] is the use of lateral connections for recursively computing the principal components of a data channel. The network proposed in [25] is called an APEX (adaptable principal component extractor) network. It consists of two parts: (1) a simple feedforward network that is updated using a Hebbian-type learning to extract the dominant principal component

of a data channel, and (2) a set of lateral connections that connect the outputs together and are trained to deflate the contribution of the already extracted principal components from the original data channel. The combination of these two sets of connections allows for recursively extracting the principal components, one by one. Each time that a new principal component has to be extracted, simply a new node is added to the APEX network, allowing for extraction of the new principal component, without the need to retrain the previous nodes.

Recently, Lai and Fyfe [28] proposed a network for performing canonical correlations analysis. However, their network only finds the first canonical coordinate pair and the corresponding canonical correlation. In this chapter, motivated by the APEX structure in [25], we develop a network structure with lateral connections and a set of updating rules for recursively extracting the canonical coordinates of a two-channel vector process. These developments are also reported in [29] and [30]. We start by formulating the problem of finding the first canonical coordinate pair as a constrained minimization problem, based on what we have established in Chapter 3. Given the first r canonical coordinate pairs, we formulate the problem of finding the $(r+1)$th pair as one of finding the first canonical coordinate pair of a deflated version of the two-channel data. We then use this formulation to develop a network structure for extracting canonical coordinates. This network consists of two independent sub-networks, each of which has a set of feedforward connections and a set of lateral connections. In fact, each sub-network has the exact same structure as an APEX network. The feedforward weights of the network are updated using a gradient descent algorithm [31] to solve the constrained minimization problem mentioned above. The lateral connections are updated to perform a deflation process that subtracts the contributions of the first r canonical coordinates from the original data channels.

6.2 Network Structure and Updating Rules

Consider the two-channel data vector $\mathbf{z} = [\mathbf{x}^T \ \mathbf{y}^T]^T \in \mathbb{R}^{m+n}$ consisting of the zero-mean random vectors $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$, $m \leq n$, with the composite covariance matrix in (2.2). Let $\mathbf{d}_{x,i} \in \mathbb{R}^m$ and $\mathbf{d}_{y,i} \in \mathbb{R}^n$ denote the ith canonical coordinate mapping vectors, defined in Chapter 3. That is, $\mathbf{d}_{x,i}$ and $\mathbf{d}_{y,i}$ are the ith columns of the canonical coordinate maps $\mathbf{D}_x$ and $\mathbf{D}_y$ in (3.32). Then, the ith canonical coordinates of $\mathbf{x}$ and $\mathbf{y}$ and their corresponding canonical correlation are

$$u_i = \mathbf{d}_{x,i}^T \mathbf{x}, \quad v_i = \mathbf{d}_{y,i}^T \mathbf{y}, \quad \text{and} \quad \sigma_i = E\{u_i v_i\} = \mathbf{d}_{x,i}^T \mathbf{R}_{xy} \mathbf{d}_{y,i}. \tag{6.1}$$

Further, we have

$$\begin{aligned} E[u_i u_j] &= \mathbf{d}_{x,i}^T \mathbf{R}_{xx} \mathbf{d}_{x,j} = \delta(i-j) \\ E[v_i v_j] &= \mathbf{d}_{y,i}^T \mathbf{R}_{yy} \mathbf{d}_{y,j} = \delta(i-j) \\ E[u_i v_j] &= \mathbf{d}_{x,i}^T \mathbf{R}_{xy} \mathbf{d}_{y,j} = \sigma_i \delta(i-j) \end{aligned} \tag{6.2}$$

where $\delta(\cdot)$ is the Kronecker delta.

Noting that $\sigma_1 = \mathbf{d}_{x,1}^T \mathbf{R}_{xy} \mathbf{d}_{y,1}$ is the largest canonical correlation, the problem of finding the first canonical coordinate mapping vectors $\mathbf{d}_{x,1}$ and $\mathbf{d}_{y,1}$ may be formulated as the maximization problem

$$\max_{\mathbf{d}_{x,1}, \mathbf{d}_{y,1}} \ \mathbf{d}_{x,1}^T \mathbf{R}_{xy} \mathbf{d}_{y,1} \tag{6.3}$$

subject to the constraints

$$\mathbf{d}_{x,1}^T \mathbf{R}_{xx} \mathbf{d}_{x,1} = 1 \quad \text{and} \quad \mathbf{d}_{y,1}^T \mathbf{R}_{yy} \mathbf{d}_{y,1} = 1. \tag{6.4}$$

This formulation also directly follows from the formulation of the two-channel CLS filtering problem of Chapter 3 in the canonical coordinate case.

Using the method of Lagrange multipliers, we may rewrite the constrained optimization problem defined by (6.3) and (6.4) as minimizing the objective function J_1 of the form

$$J_1 = -\mathbf{d}_{x,1}^T \mathbf{R}_{xy} \mathbf{d}_1 + (\mathbf{d}_{x,1}^T \mathbf{R}_{xx} \mathbf{d}_{x,1} - 1)\frac{\lambda_{1,1}}{2} + (\mathbf{d}_{y,1}^T \mathbf{R}_{yy} \mathbf{d}_{y,1} - 1)\frac{\lambda_{1,2}}{2}, \tag{6.5}$$

where $\lambda_{1,1}$ and $\lambda_{1,2}$ are Lagrange multipliers that enforce the constraints in (6.4).

Now, assume that the first $r < m$ columns of $\mathbf{D}_x$ and $\mathbf{D}_y$ have already been found. Let $\mathbf{D}_{x,r} \in \mathbb{R}^{m\times r}$ and $\mathbf{D}_{y,r} \in \mathbb{R}^{n\times r}$ be the matrices that contain the first r columns of $\mathbf{D}_x \in \mathbb{R}^{m\times m}$ and $\mathbf{D}_y \in \mathbb{R}^{n\times m}$, respectively. That is,

$$\mathbf{D}_{x,r} = [\mathbf{d}_{x,1}, \ldots, \mathbf{d}_{x,r}] \quad \text{and} \quad \mathbf{D}_{y,r} = [\mathbf{d}_{y,1}, \ldots, \mathbf{d}_{y,r}]. \tag{6.6}$$

The first r canonical coordinates of $\mathbf{x}$ and $\mathbf{y}$ are then given by

$$\begin{aligned} \mathbf{u}_r &= [u_1, \ldots, u_r]^T = \mathbf{D}_{x,r}^T \mathbf{x} \\ \mathbf{v}_r &= [v_1, \ldots, v_r]^T = \mathbf{D}_{y,r}^T \mathbf{y}. \end{aligned} \tag{6.7}$$

By deflating the contribution of the first r canonical coordinates $\mathbf{u}_r$ and $\mathbf{v}_r$ from the input channels, we may formulate the problem of finding the $(r+1)$th pair of canonical coordinates as one of finding the first canonical coordinate pair of the deflated input channels. This may be done by replacing $\mathbf{R}_{xy}$ in (6.5) with its deflated version $(\mathbf{I}-\mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T)\mathbf{R}_{xy}(\mathbf{I}-\mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T)^T$. The proof of this deflation argument is very similar to the proof presented in Appendix A for the deflated generalized eigenvalue problem in (5.11). However, for the sake of completeness a separate proof is given in Appendix B.

Therefore, the $(r+1)$th pair of canonical coordinate mapping vectors $\mathbf{d}_{x,r+1}$ and $\mathbf{d}_{y,r+1}$ may be found by minimizing the objective function J_{r+1} of the form

$$\begin{aligned} J_{r+1} = &-\mathbf{d}_{x,r+1}^T(\mathbf{I} - \mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T)\mathbf{R}_{xy}(\mathbf{I} - \mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T)^T\mathbf{d}_{y,r+1} \\ &+(\mathbf{d}_{x,r+1}^T\mathbf{R}_{xx}\mathbf{d}_{x,r+1} - 1)\tfrac{\lambda_{r+1,1}}{2} + (\mathbf{d}_{y,r+1}^T\mathbf{R}_{yy}\mathbf{d}_{y,r+1} - 1)\tfrac{\lambda_{r+1,2}}{2}, \end{aligned} \tag{6.8}$$

where $\lambda_{r+1,1}$ and $\lambda_{r+1,2}$ are Lagrange multipliers that guarantee the unit variance property of the new pair of coordinates:

$$\begin{aligned} E\{u_{r+1}^2\} &= \mathbf{d}_{x,r+1}^T\mathbf{R}_{xx}\mathbf{d}_{x,r+1} = 1 \\ E\{v_{r+1}^2\} &= \mathbf{d}_{y,r+1}^T\mathbf{R}_{yy}\mathbf{d}_{y,r+1} = 1. \end{aligned} \tag{6.9}$$

Taking the partial derivatives of J_{r+1} with respect to $\mathbf{d}_{x,r+1}$ and $\mathbf{d}_{y,r+1}$ yields

$$\begin{aligned}\frac{\partial J_{r+1}}{\partial \mathbf{d}_{x,r+1}} &= -(\mathbf{I}-\mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T)\mathbf{R}_{xy}(\mathbf{I}-\mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T)^T\mathbf{d}_{y,r+1}+\mathbf{R}_{xx}\mathbf{d}_{x,r+1}\lambda_{r+1,1}\\ \frac{\partial J_{r+1}}{\partial \mathbf{d}_{y,r+1}} &= -(\mathbf{I}-\mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T)\mathbf{R}_{yx}(\mathbf{I}-\mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T)^T\mathbf{d}_{x,r+1}+\mathbf{R}_{yy}\mathbf{d}_{y,r+1}\lambda_{r+1,2}.\end{aligned} \tag{6.10}$$

At the solution the constraints in (6.9) are satisfied. Moreover, $\mathbf{d}_{x,r+1}$ and $\mathbf{d}_{y,r+1}$ are respectively, orthogonal to $\mathbf{R}_{xx}\mathbf{D}_{x,r}$ and $\mathbf{R}_{yy}\mathbf{D}_{y,r}$. That is,

$$\mathbf{d}_{x,r+1}^T\mathbf{R}_{xx}\mathbf{D}_{x,r}=\mathbf{0} \quad \text{and} \quad \mathbf{d}_{y,r+1}^T\mathbf{R}_{yy}\mathbf{D}_{y,r}=\mathbf{0}. \tag{6.11}$$

Using (6.9) and (6.11), the optimal values of Lagrange multipliers in (6.8) are found to be

$$\lambda_{r+1,1}=\lambda_{r+1,2}=\lambda_{r+1}=\mathbf{d}_{x,r+1}^T\mathbf{R}_{xy}\mathbf{d}_{y,r+1}=\sigma_{r+1}, \tag{6.12}$$

where σ_{r+1} is the $(r+1)$th canonical correlation of $\mathbf{x}$ and $\mathbf{y}$. Correspondingly, the $(r+1)$th canonical coordinate pair of $\mathbf{x}$ and $\mathbf{y}$ is given by

$$\begin{aligned}u_{r+1} &= \mathbf{d}_{x,r+1}^T\mathbf{x}\\ v_{r+1} &= \mathbf{d}_{y,r+1}^T\mathbf{y}.\end{aligned} \tag{6.13}$$

Using (6.7) and (6.11), we may rewrite (6.13) as

$$\begin{aligned}u_{r+1} &= \mathbf{d}_{x,r+1}^T(\mathbf{I}-\mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T)\mathbf{x}=\mathbf{d}_{x,r+1}^T\mathbf{x}-\mathbf{q}_r^T\mathbf{u}_r\\ v_{r+1} &= \mathbf{d}_{y,r+1}^T(\mathbf{I}-\mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T)\mathbf{y}=\mathbf{d}_{y,r+1}^T\mathbf{x}-\mathbf{p}_r^T\mathbf{v}_r\end{aligned} \tag{6.14}$$

where

$$\begin{aligned}\mathbf{q}_r^T &= \mathbf{d}_{x,r+1}^T\mathbf{R}_{xx}\mathbf{D}_{x,r}\\ \mathbf{p}_r^T &= \mathbf{d}_{y,r+1}^T\mathbf{R}_{yy}\mathbf{D}_{y,r}.\end{aligned} \tag{6.15}$$

The pair of equations in (6.14) may be used to define a network structure for extracting the $(r+1)$th pair of canonical coordinates, given the first r pairs. Each equation in (6.14) defines a single layer sub-network that features a feedforward set of weights from the input to the output and a set of lateral connections that connects the first r nodes to the $(r+1)$th node. Figure 6.1 shows the structure of this network.

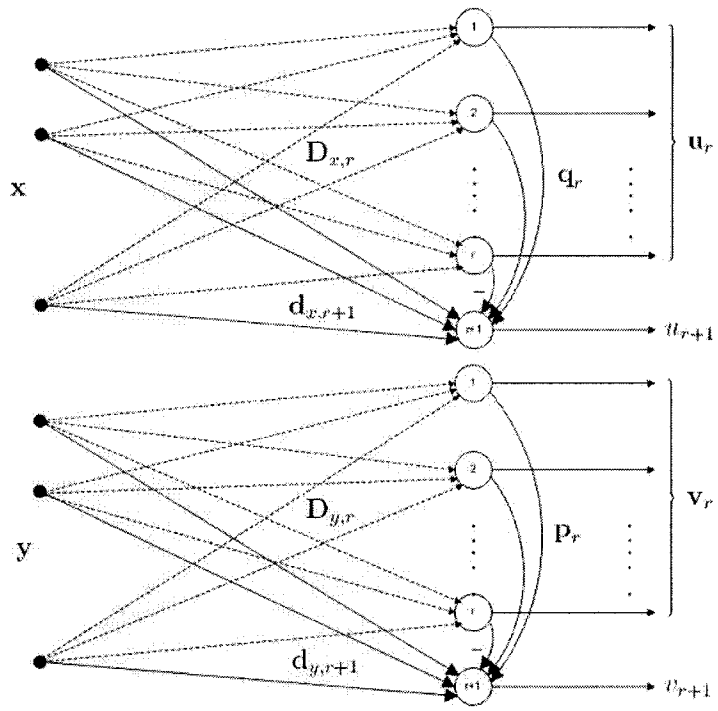

Figure 6.1: The structure of the network for recursive extraction of canonical coordinates of $\mathbf{x}$ and $\mathbf{y}$.

In this structure, $\mathbf{D}_{x,r} \in \mathbb{R}^{m\times r}$ and $\mathbf{D}_{y,r} \in \mathbb{R}^{n\times r}$ are the weight matrices that map $\mathbf{x}$ and $\mathbf{y}$ to their first r canonical coordinates $\mathbf{u}_r$ and $\mathbf{v}_r$. Given these weights, the network may be trained, by minimizing J_{r+1} in (6.8), to extract the $(r+1)$th canonical coordinate pair and the corresponding mapping vectors. The weight vectors $\mathbf{d}_{x,r+1} \in \mathbb{R}^m$ and $\mathbf{d}_{y,r+1} \in \mathbb{R}^n$ are trained to maximize the correlation between the outputs u_{r+1} and v_{r+1}, and make them unit variance. The lateral weight vector $\mathbf{q}_r \in \mathbb{R}^r$ is trained to orthogonalize $\mathbf{u}_r$ (the first r canonical coordinates of $\mathbf{x}$) to u_{r+1} (the $(r+1)$th canonical coordinate of $\mathbf{x}$.) Similarly, the lateral weight vector $\mathbf{p}_r \in \mathbb{R}^r$ is trained to orthogonalize $\mathbf{v}_r$ to v_{r+1}. The lateral connections perform a deflation process that subtracts the contributions of the already extracted coordinates from the linear subspaces spanned by the elements of $\mathbf{x}$ and $\mathbf{y}$. This structure allows for adding new nodes for extracting additional canonical coordinates, without the need for retraining the previous nodes. Thus, an adequate number of canonical coordinates

can be recursively extracted to capture a desired percentage of coherence or mutual information.

Using the gradient descent learning algorithm, with instantaneous values of covariance matrices inserted into (6.10), we may derive the following updating rules for $\mathbf{d}_{x,r+1}$, and $\mathbf{d}_{y,r+1}$:

$$\begin{aligned}\mathbf{d}_{x,r+1}(j+1) &= \mathbf{d}_{x,r+1}(j) + [(\mathbf{x}(j+1) - \mathbf{S}_r(j+1)\mathbf{u}_r(j+1))v_r(j+1) \\ &\quad -\mathbf{x}(j+1)\mathbf{x}(j+1)^T\mathbf{d}_{x,r+1}(j)\lambda_{r+1}(j+1)]\beta(j+1) \\ \mathbf{d}_{y,r+1}(j+1) &= \mathbf{d}_{y,r+1}(j) + [(\mathbf{y}(j+1) - \mathbf{T}_r(j+1)\mathbf{v}_r(j+1))u_r(j+1) \\ &\quad -\mathbf{y}(j+1)\mathbf{y}(j+1)^T\mathbf{d}_{y,r+1}(j)\lambda_{r+1}(j+1)]\beta(j+1)\end{aligned} \tag{6.16}$$

where j is the index of iteration and $\beta(j+1)$ is the learning rate (step size) at iteration $j+1$. Matrices $\mathbf{S}_r \in \mathbb{R}^{m\times r}$ and $\mathbf{T}_r \in \mathbb{R}^{n\times r}$ are updated to asymptotically approximate $\mathbf{R}_{xx}\mathbf{D}_{x,r}$ and $\mathbf{R}_{yy}\mathbf{D}_{y,r}$, respectively. From (6.12), the Lagrange multiplier $\lambda_{r+1} = \lambda_{r+1,1} = \lambda_{r+1,2}$ shall be updated to asymptotically approximate $\mathbf{d}_{x,r+1}^T\mathbf{R}_{xy}\mathbf{d}_{y,r+1} = \sigma_{r+1}$, the $(r+1)$th canonical correlation. Thus, the updating rules for $\mathbf{S}_r$, $\mathbf{T}_r$, and λ_{r+1} are

$$\begin{aligned}\mathbf{S}_r(j+1) &= \tfrac{j}{j+1}\mathbf{S}_r(j) + \tfrac{1}{j+1}\mathbf{x}(j+1)\mathbf{u}_r^T(j+1) \\ \mathbf{T}_r(j+1) &= \tfrac{j}{j+1}\mathbf{T}_r(j) + \tfrac{1}{j+1}\mathbf{y}(j+1)\mathbf{v}_r^T(j+1) \\ \lambda_{r+1}(j+1) &= \tfrac{j}{j+1}\lambda_{r+1}(j) + \tfrac{1}{j+1}\mathbf{d}_{x,r+1}^T(j)\mathbf{x}(j+1)\mathbf{y}^T(j+1)\mathbf{d}_{y,r+1}(j).\end{aligned} \tag{6.17}$$

Finally using (6.15), the learning rules for the lateral connection weight vectors $\mathbf{q}_r$ and $\mathbf{p}_r$ may be written as

$$\begin{aligned}\mathbf{q}_r(j+1) &= \mathbf{S}_r^T(j+1)\mathbf{d}_{x,r+1}(j+1) \\ \mathbf{p}_r(j+1) &= \mathbf{T}_r^T(j+1)\mathbf{d}_{y,r+1}(j+1).\end{aligned} \tag{6.18}$$

Thus, we may summarize the step-by-step training algorithm for extracting the $(r+1)$th canonical coordinate pair for $r = 0, 1, \ldots, m-1$, and the corresponding

mapping vectors as

$$
\begin{aligned}
&\mathbf{u}_r(j+1) = \mathbf{D}_{x,r}^T \mathbf{x}(j+1) \\
&\mathbf{v}_r(j+1) = \mathbf{D}_{y,r}^T \mathbf{y}(j+1) \\
&u_{r+1}(j+1) = \mathbf{d}_{x,r+1}^T(j+1)\mathbf{x}(j+1) - \mathbf{q}_r^T(j)\mathbf{u}_r(j+1) \\
&v_{r+1}(j+1) = \mathbf{d}_{y,r+1}^T(j+1)\mathbf{y}(j+1) - \mathbf{p}_r^T(j)\mathbf{v}_r(j+1) \\
&\lambda_{r+1}(j+1) = \tfrac{1}{j+1}[j\lambda_{r+1}(j) + \mathbf{d}_{x,r+1}^T(j)\mathbf{x}(j+1)\mathbf{y}^T(j+1)\mathbf{d}_{y,r+1}(j)] \\
&\mathbf{S}_r(j+1) = \tfrac{1}{j+1}[j\mathbf{S}_r(j) + \mathbf{x}(j+1)\mathbf{u}_r^T(j+1)] \\
&\mathbf{T}_r(j+1) = \tfrac{1}{j+1}[j\mathbf{T}_r(j) + \mathbf{y}(j+1)\mathbf{v}_r^T(j+1)] \\
&\mathbf{d}_{x,r+1}(j+1) = \mathbf{d}_{x,r+1}(j) + [(\mathbf{x}(j+1) - \mathbf{S}_r(j+1)\mathbf{u}_r(j+1))v_r(j+1) \\
&\qquad\qquad -\mathbf{x}(j+1)\mathbf{x}(j+1)^T\mathbf{d}_{x,r+1}(j)\lambda_{r+1}(j+1)]\beta(j+1) \\
&\mathbf{d}_{y,r+1}(j+1) = \mathbf{d}_{y,r+1}(j) + [(\mathbf{y}(j+1) - \mathbf{T}_r(j+1)\mathbf{v}_r(j+1))u_r(j+1) \\
&\qquad\qquad -\mathbf{y}(j+1)\mathbf{y}(j+1)^T\mathbf{d}_{y,r+1}(j)\lambda_{r+1}(j+1)]\beta(j+1) \\
&\mathbf{q}_r(j+1) = \mathbf{S}_r^T(j+1)\mathbf{d}_{x,r+1}(j+1) \\
&\mathbf{p}_r(j+1) = \mathbf{T}_r^T(j+1)\mathbf{d}_{y,r+1}(j+1).
\end{aligned}
\tag{6.19}
$$

The initial values $\mathbf{d}_{x,r+1}(0) \in \mathbb{R}^m$, $\mathbf{d}_{y,r+1}(0) \in \mathbb{R}^n$, $\mathbf{q}_r(0) \in \mathbb{R}^r$, $\mathbf{p}_r(0) \in \mathbb{R}^r$, $\mathbf{S}_r(0) \in \mathbb{R}^{m\times r}$, $\mathbf{T}_r(0) \in \mathbb{R}^{n\times r}$, and λ_{r+1} may be chosen randomly. The learning rate β may be varied or kept fixed [31]. It should be pointed out that since the network is updated using a gradient descent algorithm, it suffers from slow convergence (even as slow as linear convergence) and sometimes instability, depending on the initialization and choice of the step size. Readers are referred to [31] for a detailed discussion about the convergence behavior of the gradient descent algorithm.

In most applications, the number of canonical coordinate pairs to be extracted is not known *a priori*. However, each time a new canonical coordinate pair is extracted, we may run a test based on (2.27) and (2.32) to determine if the amount of linear dependence or mutual information captured, meets a pre-specified threshold. If the

threshold is not reached, we may add another node to the network to extract the next pair of canonical coordinates.

6.3 Simulation Results

In this section, the proposed network is used to recursively extract the canonical coordinate mappings for a synthesized data set. The performance of the network is demonstrated by presenting the plots of squared error between the actual canonical coordinate mappings, computed using the conventional method in (2.9), and the ones estimated by the network, along with the plots of the squared error for canonical correlations.

Let $\hat{\mathbf{d}}_{x,i} \in \mathbb{R}^m$ and $\hat{\mathbf{d}}_{y,i} \in \mathbb{R}^n$ denote the estimate of the ith pair of the actual canonical coordinate mappings $\mathbf{d}_{x,i}$ and $\mathbf{d}_{y,i}$, respectively,. We define $e^2_{\mathbf{d}_{x,i}}$ and $e^2_{\mathbf{d}_{y,i}}$ as the squared estimation error of the ith canonical coordinate mappings $\mathbf{d}_{x,i}$ and $\mathbf{d}_{y,i}$. That is,

$$e^2_{\mathbf{d}_{x,i}} = \|\mathbf{d}_{x,i} - \hat{\mathbf{d}}_{x,i}\|^2 \quad \text{and} \quad e^2_{\mathbf{d}_{y,i}} = \|\mathbf{d}_{y,i} - \hat{\mathbf{d}}_{y,i}\|^2.$$

Also, $e^2_{\sigma_i} = (\sigma_i - \hat{\sigma}_i)^2$ is defined as the squared estimation error of the ith canonical correlation σ_i. The actual canonical correlation σ_i is found from the SVD in (2.6). From (6.12), it is seen that the ith canonical correlation σ_i is estimated by the Lagrange multiplier λ_i.

The data set is formed from 500 samples of two data channels $\mathbf{x} \in \mathbb{R}^4$ and $\mathbf{y} \in \mathbb{R}^5$, governed by the linear model

$$\mathbf{x} = \mathbf{H}_x \boldsymbol{\eta}_x$$

$$\mathbf{y} = \mathbf{H}_y \boldsymbol{\eta}_y + \mathbf{H}_{yx}\mathbf{x}$$

where $\mathbf{H}_x \in \mathbb{R}^{4\times4}$, $\mathbf{H}_y \in \mathbb{R}^{5\times5}$, and $\mathbf{H}_{yx} \in \mathbb{R}^{5\times4}$ are used to synthesize $\mathbf{x}$ and $\mathbf{y}$ from $\boldsymbol{\eta}_x \in \mathbb{R}^4$ and $\boldsymbol{\eta}_y \in \mathbb{R}^5$, which are two independent white Gaussian vector processes. The network is trained for 2500 epochs using the algorithm in (6.19), without knowledge of the generating mechanism for $\mathbf{x}$ and $\mathbf{y}$. An epoch is a complete sweep over the

500 samples. The learning rate is varied linearly from $\beta = 5 \times 10^{-3}$ to $\beta = 5 \times 10^{-6}$ in 2500 steps. All the initial values in (6.19) are randomly selected.

Figures 6.2(a)-(d) show the squared estimation errors $e^2_{\mathbf{d}_{x,i}}$, $i \in [1, 4]$ vs. the epoch index for ten independent initializations of the network. It is seen that in all the cases the squared error approaches zero within a misadjustment error [31], and thus the weights of the upper sub-network in Figure 6.1 converge to the actual canonical coordinate mapping vectors that map the first data channel $\mathbf{x}$ into its canonical coordinates $\mathbf{u}$. The plots of the squared estimation errors $e^2_{\mathbf{d}_{y,i}}$, $i \in [1, 4]$ vs. epoch index for the ten initializations are shown in Figures 6.3(a)-(d). The convergence behaviors are very similar to those in Figures 6.2(a)-(d). In all the cases the squared error approaches zero within a misadjustment error, and thus the weights of the lower sub-network in Figure 6.1 converge to the actual canonical coordinate mapping vectors that map the second data channel $\mathbf{y}$ into its canonical coordinates $\mathbf{v}$.

Figures 6.4(a)-(d) show the squared estimation errors $e^2_{\sigma_i}$, $i \in [1, 4]$ vs. the epoch index for the ten initializations. These plots show that the squared error decays to zero in all the cases. The estimate of the ith canonical correlation σ_i is given by the Lagrange multiplier λ_i. These plots indicate that λ_i's converge to the actual canonical correlations σ_i's.

6.4 Conclusions

In this chapter, a network structure and a set of updating rules for recursive extraction of canonical coordinates of two data channels was developed. The network was built based on a constrained minimization problem that exploits a deflation process. The deflation process is performed by incorporating lateral connections into the network in a manner similar to the APEX network in [25]. The structure of the network, along with the updating rules, allows for adding a new node to the network in order to extract a new canonical coordinate pair, without the need to retrain the previous

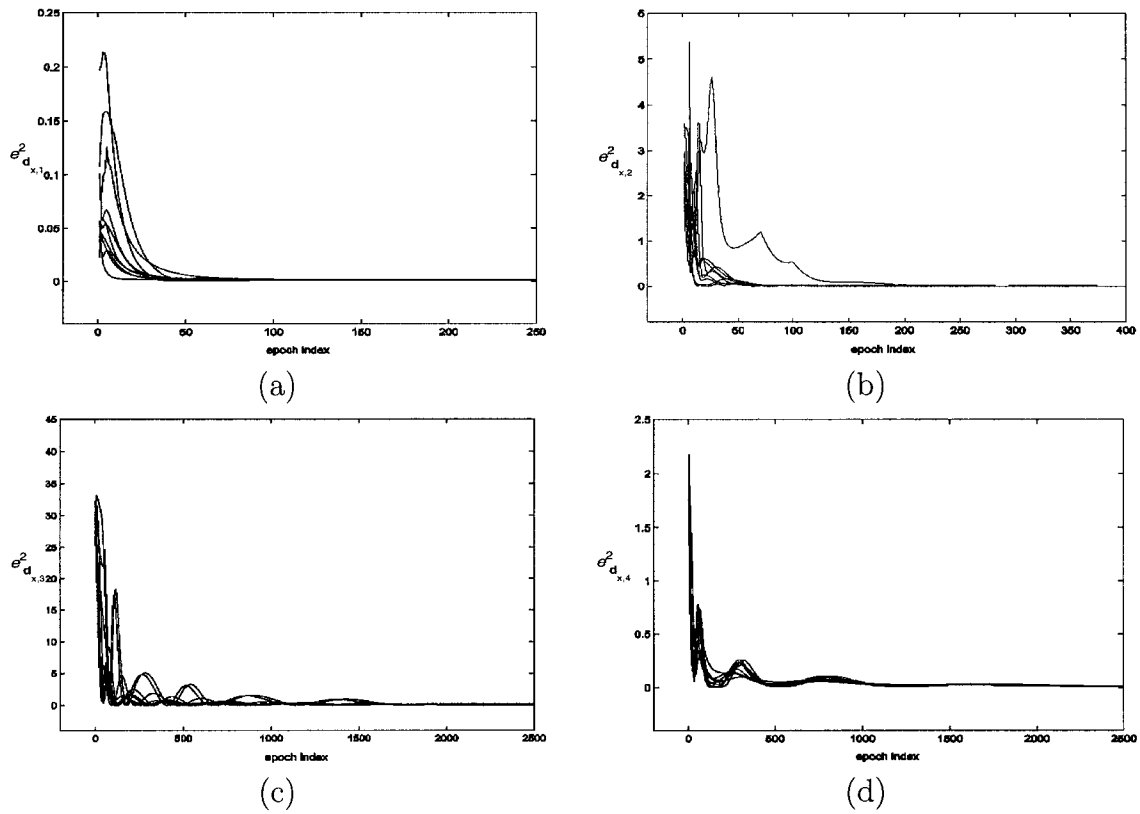

Figure 6.2: The squared error for $\mathbf{d}_{x,i}$'s, $i \in [1,4]$ vs. the epoch index for ten independent initializations of the network: (a) $i = 1$, $e^2_{\mathbf{d}_{x,1}} = \|\mathbf{d}_{x,1} - \hat{\mathbf{d}}_{x,1}\|^2$. (b) $i = 2$, $e^2_{\mathbf{d}_{x,2}} = \|\mathbf{d}_{x,2} - \hat{\mathbf{d}}_{x,2}\|^2$. (c) $i = 3$, $e^2_{\mathbf{d}_{x,3}} = \|\mathbf{d}_{x,3} - \hat{\mathbf{d}}_{x,3}\|^2$. (d) $i = 4$, $e^2_{\mathbf{d}_{x,4}} = \|\mathbf{d}_{x,4} - \hat{\mathbf{d}}_{x,4}\|^2$. In all cases the squared error approaches zero and the weights of the upper sub-network in Figure 6.1 converge to the actual canonical coordinate mapping vectors that map the first data channel $\mathbf{x}$ into its canonical coordinates $\mathbf{u}$.

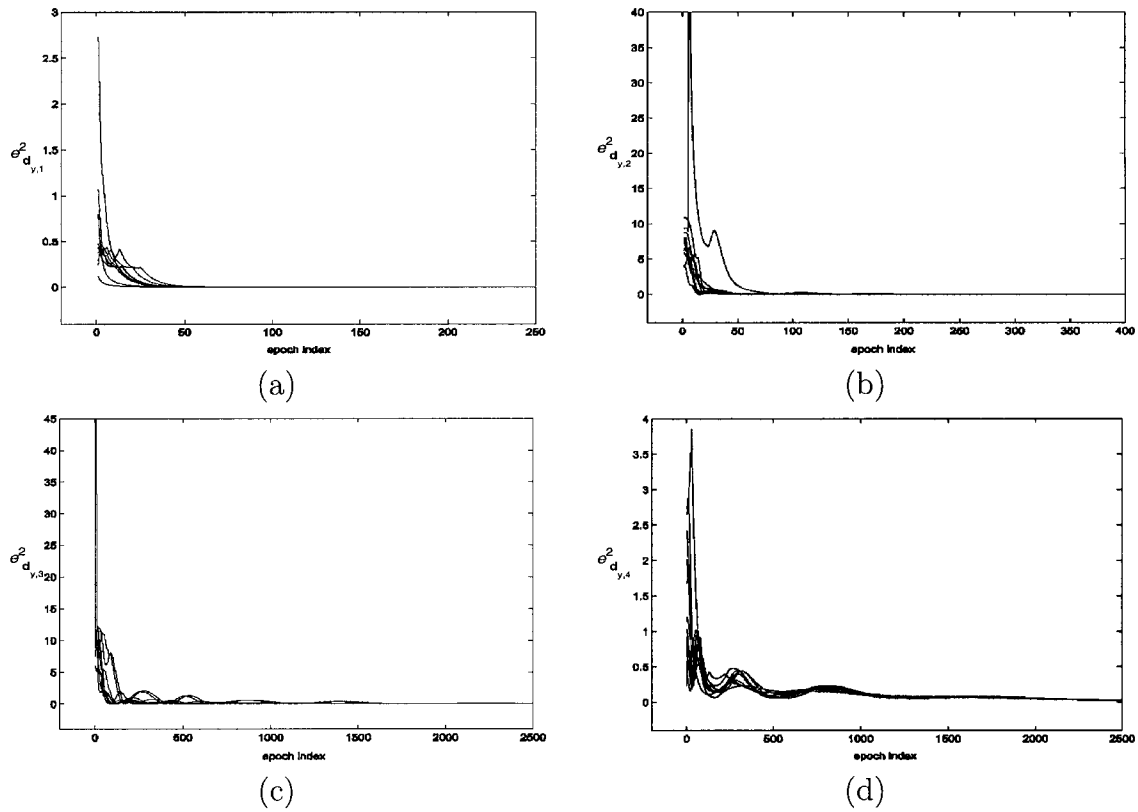

Figure 6.3: The squared error for $\mathbf{d}_{y,i}$'s, $i \in [1,4]$ vs. the epoch index for ten independent initializations of the network: (a) $i = 1$, $e^2_{\mathbf{d}_{y,1}} = \|\mathbf{d}_{y,1} - \hat{\mathbf{d}}_{y,1}\|^2$. (b) $i = 2$, $e^2_{\mathbf{d}_{y,2}} = \|\mathbf{d}_{y,2} - \hat{\mathbf{d}}_{y,2}\|^2$. (c) $i = 3$, $e^2_{\mathbf{d}_{y,3}} = \|\mathbf{d}_{y,3} - \hat{\mathbf{d}}_{y,3}\|^2$. (d) $i = 4$, $e^2_{\mathbf{d}_{y,4}} = \|\mathbf{d}_{y,4} - \hat{\mathbf{d}}_{y,4}\|^2$. In all cases the squared error approaches zero and the weights of the lower sub-network in Figure 6.1 converge to the actual canonical coordinate mapping vectors that map the second data channel $\mathbf{y}$ into its canonical coordinates $\mathbf{v}$.

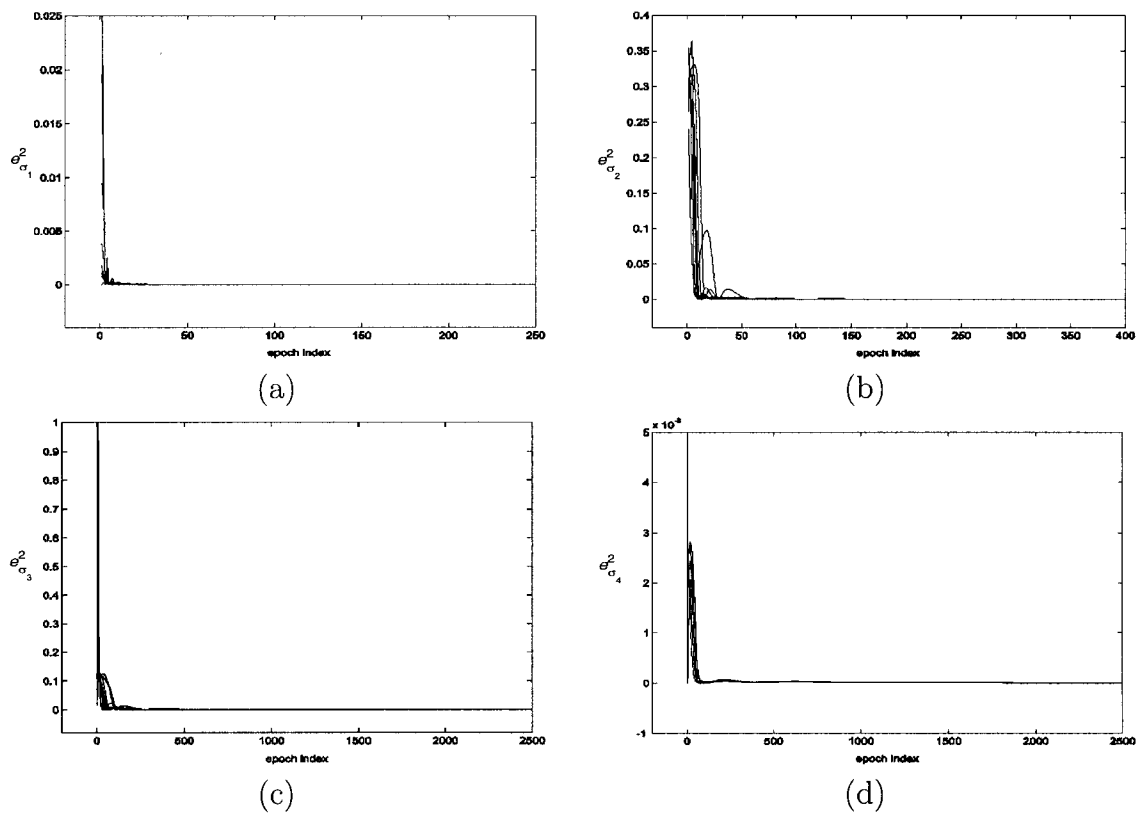

Figure 6.4: The squared error for σ_i's, $i \in [1,4]$ vs. the epoch index for ten independent initializations of the network: (a) $i = 1$, $e^2_{\sigma_1} = (\sigma_1 - \hat{\sigma}_1)^2$. (b) $i = 2$, $e^2_{\sigma_2} = (\sigma_2 - \hat{\sigma}_2)^2$. (c) $i = 3$, $e^2_{\sigma_3} = (\sigma_3 - \hat{\sigma}_3)^2$. (d) $i = 4$, $e^2_{\sigma_4} = (\sigma_4 - \hat{\sigma}_4)^2$. The estimate of σ_i is given by the Lagrange multiplier λ_i. The plots show that λ_i converges to the actual canonical correlation σ_i in all the cases.

nodes. The extraction of canonical coordinate pairs may be stopped when a pre-specified percentage of coherence or mutual information is captured. A simulation example was presented to demonstrate the validity of the proposed network and updating rules.

We note that what is interesting about this work is the use of lateral connections for performing the deflation process. However, since the network is updated using a gradient descent algorithm, it suffers from slow convergence (even as slow as linear convergence) and sometimes instability, depending on the initialization and choice of the step size. Thus, in two-channel signal processing applications, where recursive extraction of canonical coordinates is required, the alternating power methods developed in Chapter 5 are preferred.

CHAPTER 7

EMPIRICAL CANONICAL COORDINATE DECOMPOSITIONS IN SUBSPACES FOR TWO-CHANNEL LINEAR AND NONLINEAR MAPS

7.1 Introduction

All the developments and results reported to date for canonical correlation analysis of two-channel data, including those presented in the previous chapters, are based on the assumption that either the theoretical covariance and cross-covariance matrices of the channels are known, or enough independent copies of the channels are available to obtain full-rank estimates of the covariance matrices. However, little attention has been devoted to the algebraic limits to canonical correlation analysis. That is, just how poor can sample support become before sample canonical correlations cease to carry any information about the true canonical correlations? This is one of the particular questions we address in this chapter.

More generally, we study the empirical canonical correlation analysis of two-channel data. The term empirical implies that the canonical correlation analysis is based on covariance matrices that are estimated from a limited number of samples

of the two-channel data, and are not necessarily full-rank. The available data samples may be obtained from linear or nonlinear transformations of a limited number of random samples drawn from a two-channel vector process. The question to be addressed in this chapter is whether or not the canonical correlations and coordinates obtained from sample data have the same algebraic and geometric properties as the underlying theoretical ones. For example, when do empirical canonical correlations estimate theoretical canonical correlations between two data channels?

It must be mentioned that the reason for considering nonlinearities is that canonical correlation analysis only exploits the second-order statistics of two-channel data. In some applications, such as detection, classification, and feature extraction, however, the information that is obtained from second-order statistics alone, such as linear dependence, may not be sufficient to achieve the desired performance. One possible solution, in such cases, is to map the channels using nonlinear mappings prior to canonical correlation analysis. The canonical correlations extracted from the mapped data channels may then reveal coherence between high-order attributes of the original data channels.

The idea of using nonlinear maps prior to linear processing was first exploited by Vapnik in the theory of support vector machines (SVM) for the design of large margin classifiers [33], [34]. The idea in SVM is to use a nonlinear mapping to map the input space into a high-dimensional feature space, in which the features are linearly separable. Perhaps the most intriguing aspect of SVM is that the high dimensional nonlinear mappings are never *explicitly* computed and all computations are carried out in the original low dimensional space. The trick, known as the *kernel trick* [35], cleverly reformulates the problem in such a way that only inner products of the nonlinear mappings appear in the equations. These inner products are then replaced by kernel functions that may be computed in the input space. Since the development of SVM, numerous results have been reported on kernel-nonlinear counterparts of

standard information processing techniques, among which are kernel versions of principal component analysis [35], [36], Fisher discriminant analysis [35], [37], linear least squares estimation [38], and Mahalanobis distance [38].

Several kernel formulations have also been reported for canonical correlation analysis of nonlinear maps of two-channel data, among which are [39]- [46]. However, in all of them a very important question has been ignored: Do canonical correlations and coordinates, extracted for nonlinearly mapped data channels, possess the algebraic and geometric properties of the theoretical canonical correlations and coordinates, and can the empirical canonical correlations measure high-order coherence between two data channels?

The basis for this question is that the nonlinear mappings used in the kernel-nonlinear methods are often of dimension higher than the number of available samples, resulting in a paucity of samples after nonlinear mapping, which in turn produces rank-deficient sample covariance matrices. Even when low-dimensional nonlinear mappings are used, the mapped data vectors for each channel may become linearly dependent, resulting in rank-deficient sample covariance matrices in the mapped domain. In the kernel formulation in [39], it is suggested that regularization may be used to estimate the covariance matrices in order to obtain full-rank sample covariance matrices. However, the empirical canonical correlations and coordinates obtained with regularized covariance matrices do not share the algebraic and geometric properties of the true canonical correlations and canonical coordinates.

In this chapter, we shall clarify how the number of samples drawn from two-channel data and the ranks of the data matrices[1] affect the algebraic and geometric properties of empirical canonical correlations and canonical coordinates. We establish

[1]The data matrices are column-wise collections of vector data samples that are drawn from the two vector data channels.

that empirical canonical correlations do form a maximal set of invariants for the composite sample covariance matrix of two-channel data. Further, we demonstrate that empirical canonical correlations measure the cosines of the principal angles between the row spaces of the two data matrices in Euclidean space, and hence are experimental surrogates for the true canonical correlations, which measure the principal cosines between subspaces of the Hilbert space of second-order random variables. These results have been reported in [5] for the case where sample covariance matrices are full-rank.

We establish that when the number of vector samples drawn from each vector channel is smaller than the sum of the ranks of the two data matrices, some of the empirical canonical correlations become one, regardless of the two-channel model that generates the samples. In such cases, the empirical canonical correlations should be interpreted solely as principal cosines between two linear subspaces of a Euclidean space, and not as estimates of canonical correlations or principal cosines between subspaces of the Hilbert space of second-order random variables.

When the number of samples is greater than the sum of the ranks of the two data matrices, it may be possible for the empirical canonical correlations to estimate canonical correlations and principal cosines between random variables, and hence be used as estimates of coherence between two data channels. This implies that the sum of the ranks of the two data matrices determines the minimum number of data samples (sample support) required to estimate the theoretical canonical correlations.

Our results have interesting implications for canonical correlation analysis of nonlinear functions of two-channel data. That is, only when the number of data samples is greater than the sum of the ranks of the two data matrices do the empirical canonical correlations estimate the correlation between high-order attributes of the original channels. In such cases, however, the kernel formulations are computationally disadvantageous with respect to the direct formulations.

The material presented in this chapter are also reported in [32]. Some of the findings of this chapter, concerning the effect of sample support on empirical canonical correlations, have been reported in [45], without rigorous proof and analysis.

Notations and typographic conventions used in this chapter: Given a matrix $\mathbf{A}$, we denote the ijth element of $\mathbf{A}$ by $[\mathbf{A}]_{ij}$, the ith column of $\mathbf{A}$ by either $\mathbf{a}_i$ or $\mathbf{A}(:,i)$, the ith row of $\mathbf{A}$ by $\mathbf{A}(i,:)$, the matrix consisting of the first p columns of $\mathbf{A}$ by $\mathbf{A}(:,1:p)$, and the matrix consisting of the first p rows of $\mathbf{A}$ by $\mathbf{A}(1:p,:)$. We use Col-Span$\{\mathbf{A}\}$ and Row-Span$\{\mathbf{A}\}$ to denote the linear subspaces spanned by the columns and rows of $\mathbf{A}$, respectively. Whenever the inverse of a rank-deficient matrix is needed, the Moore-Penrose pseudo inverse [57] is used. However, for the sake of simplicity of notation, we use $\mathbf{A}^{-1}$ for both inverse and pseudo inverse of $\mathbf{A}$.

7.2 Empirical Canonical Coordinate Decomposition in Subspaces

In order to make our discussion general for both linear and nonlinear cases, we assume that the vector data samples are drawn from the two-channel data model

$$\mathbf{x} = \boldsymbol{\phi}(\boldsymbol{\theta}) \quad \text{and} \quad \mathbf{y} = \boldsymbol{\psi}(\boldsymbol{\upsilon}) \tag{7.1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\upsilon} \in \mathbb{R}^l$ are two zero-mean random vectors, sharing the nonsingular composite covariance matrix

$$\mathbf{R}_{\mu\mu} = E[\boldsymbol{\mu}\boldsymbol{\mu}^T] = E\left[\begin{pmatrix}\boldsymbol{\theta}\\ \boldsymbol{\upsilon}\end{pmatrix}\begin{pmatrix}\boldsymbol{\theta}^T & \boldsymbol{\upsilon}^T\end{pmatrix}\right] = \begin{bmatrix}\mathbf{R}_{\theta\theta} & \mathbf{R}_{\theta\upsilon}\\ \mathbf{R}_{\upsilon\theta} & \mathbf{R}_{\upsilon\upsilon}\end{bmatrix}, \tag{7.2}$$

and $\boldsymbol{\phi} : \mathbb{R}^d \longrightarrow \mathbb{R}^m$ and $\boldsymbol{\psi} : \mathbb{R}^l \longrightarrow \mathbb{R}^n$ are linear or nonlinear transformations that transform $\boldsymbol{\theta}$ and $\boldsymbol{\upsilon}$ into $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$, respectively. Without loss of generality, we assume $m \leq n$.

Let us now consider the sample data matrices

$$\begin{aligned}\mathbf{X} &= [\mathbf{x}_1, \ldots, \mathbf{x}_M] = [\boldsymbol{\phi}(\boldsymbol{\theta}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{\theta}_M)]\\ \mathbf{Y} &= [\mathbf{y}_1, \ldots, \mathbf{y}_M] = [\boldsymbol{\psi}(\boldsymbol{\upsilon}_1), \ldots, \boldsymbol{\psi}(\boldsymbol{\upsilon}_M)]\end{aligned} \tag{7.3}$$

obtained from M two-channel data samples $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M] \in \mathbb{R}^{d \times M}$ and $\boldsymbol{\Upsilon} = [\boldsymbol{\upsilon}_1, \ldots, \boldsymbol{\upsilon}_M] \in \mathbb{R}^{l \times M}$. Since the $\boldsymbol{\theta_i}$ and the $\boldsymbol{\upsilon_i}$ are random samples drawn from the two-channel process $\boldsymbol{\mu} = [\boldsymbol{\theta}^T \ \boldsymbol{\upsilon}^T]^T$, with probability one, $\text{Rank}\{\boldsymbol{\Theta}\} = \min(d, M)$ and $\text{Rank}\{\boldsymbol{\Upsilon}\} = \min(l, M)$.

For linear transformations in (7.1), we may write $\mathbf{x}_i = \mathbf{A}\boldsymbol{\theta_i}$ and $\mathbf{y}_i = \mathbf{B}\boldsymbol{\upsilon_i}$, with $\mathbf{A} \in \mathbb{R}^{m \times d}$ and $\mathbf{B} \in \mathbb{R}^{n \times l}$. Then, (7.3) becomes

$$\mathbf{X} = \mathbf{A}\boldsymbol{\Theta} \quad \text{and} \quad \mathbf{Y} = \mathbf{B}\boldsymbol{\Upsilon}. \tag{7.4}$$

In this linear case, the rows of $\mathbf{X} \in \mathbb{R}^{m \times M}$ and $\mathbf{Y} \in \mathbb{R}^{n \times M}$ are linear combinations of rows of $\boldsymbol{\Theta}$ and $\boldsymbol{\Upsilon}$, respectively. This is only an example. In what follows, no assumption of linearity is made.

Without loss of generality, we assume that the columns of the sample data matrices $\mathbf{X}$ and $\mathbf{Y}$ are centered, i.e. $\sum_{i=1}^{M} \mathbf{x}_i = \mathbf{0}$ and $\sum_{i=1}^{M} \mathbf{y}_i = \mathbf{0}$. Then, we may write the composite sample covariance matrix of $\mathbf{z} = [\mathbf{x}^T \ \mathbf{y}^T]^T$ as

$$\mathbf{S}_{zz} = \mathbf{Z}\mathbf{Z}^T = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} [\mathbf{X}^T \ \mathbf{Y}^T] = \begin{bmatrix} \mathbf{X}\mathbf{X}^T & \mathbf{X}\mathbf{Y}^T \\ \mathbf{Y}\mathbf{X}^T & \mathbf{Y}\mathbf{Y}^T \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{S}_{yy} \end{bmatrix}, \tag{7.5}$$

where $\mathbf{S}_{xx} = \mathbf{X}\mathbf{X}^T$, $\mathbf{S}_{yy} = \mathbf{Y}\mathbf{Y}^T$, and $\mathbf{S}_{xy} = \mathbf{X}\mathbf{Y}^T$ are sample auto-covariance and cross-covariance matrices of $\mathbf{x}$ and $\mathbf{y}$, computed from the sample data matrices $\mathbf{X}$ and $\mathbf{Y}$. Note that the normalization by M in the sample covariance matrices is ignored, as it does not affect the discussion.

There are two ways to think of the elements of $\mathbf{S}_{xx}$, $\mathbf{S}_{yy}$, and $\mathbf{S}_{xy}$. For example, the ijth element of $\mathbf{S}_{xy}$ may be written as

$$[\mathbf{S}_{xy}]_{ij} = [\mathbf{X}\mathbf{Y}^T]_{ij} = [\sum_{k=1}^{M} \mathbf{X}(:,k)\mathbf{Y}(:,k)^T]_{ij} \tag{7.6}$$

or alternatively as

$$[\mathbf{S}_{xy}]_{ij} = [\mathbf{X}\mathbf{Y}^T]_{ij} = \mathbf{X}(i,:)\mathbf{Y}(j,:)^T. \tag{7.7}$$

In the first representation, $[\mathbf{S}_{xy}]_{ij}$ is the ijth element of the sum of the outer products of columns of $\mathbf{X}$ and $\mathbf{Y}$. However, in the second representation, $[\mathbf{S}_{xy}]_{ij}$ is the inner

product between the ith row of $\mathbf{X}$, i.e. $\mathbf{X}(i,:)$, and the jth row of $\mathbf{Y}$, i.e. $\mathbf{Y}(j,:)$. In analogy with the Hilbert space case, where $[\mathbf{R}_{xy}]_{ij} = E[x_i y_j]$ is the inner product between random varibles x_i and y_j, we may think of the ith row of $\mathbf{X}$, i.e. $\mathbf{X}(i,:)$, as an experimental *surrogate* for x_i, and $\mathbf{Y}(j,:)$ as an experimental surrogate for y_j. Correspondingly, the inner product $[\mathbf{X}\mathbf{Y}^T]_{ij} = \mathbf{X}(i,:)\mathbf{Y}(j,:)^T$ in the Euclidean space of $M-$vectors may be viewed as a stand-in for the inner product $[\mathbf{R}_{xy}]_{ij} = E[x_i y_j]$ in the Hilbert space of second-order random variables.

Let us assume that $\text{Rank}\{\mathbf{X}\} = p$ and $\text{Rank}\{\mathbf{Y}\} = q$. This implies that the dimension of the subspaces spanned by rows of $\mathbf{X}$ and $\mathbf{Y}$, i.e. Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$, which are subspaces of $\mathbb{R}^M$, are p and q, respectively. Note that since the row spaces of $\mathbf{X}$ and $\mathbf{Y}$ are subspaces of $\mathbb{R}^M$ and columns of $\mathbf{X}$ and $\mathbf{Y}$ are centered, we have $p \leq M-1$ and $q \leq M-1$. Therefore, we may write the SVD's of the sample data matrices $\mathbf{X}$ and $\mathbf{Y}$ as,

$$\begin{aligned} \mathbf{X} &= \mathbf{U}_x \mathbf{\Sigma}_x \mathbf{V}_x^T \quad \text{and} \quad \mathbf{U}_x^T \mathbf{X} \mathbf{V}_x = \mathbf{\Sigma}_x, \\ \mathbf{Y} &= \mathbf{U}_y \mathbf{\Sigma}_y \mathbf{V}_y^T \quad \text{and} \quad \mathbf{U}_y^T \mathbf{Y} \mathbf{V}_y = \mathbf{\Sigma}_y, \end{aligned} \tag{7.8}$$

where $\mathbf{U}_x \in \mathbb{R}^{m\times m}$, $\mathbf{V}_x \in \mathbb{R}^{M\times M}$, $\mathbf{U}_y \in \mathbb{R}^{n\times n}$, and $\mathbf{V}_y \in \mathbb{R}^{M\times M}$ are orthogonal matrices and $\mathbf{\Sigma}_x \in \mathbb{R}^{m\times M}$ and $\mathbf{\Sigma}_y \in \mathbb{R}^{n\times M}$ are diagonal:

$$\mathbf{\Sigma}_x = \begin{bmatrix} \mathbf{\Sigma}_x(p) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma}_y = \begin{bmatrix} \mathbf{\Sigma}_y(q) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{7.9}$$

The diagonal matrices $\mathbf{\Sigma}_x(p) \in \mathbb{R}^{p\times p}$ and $\mathbf{\Sigma}_y(q) \in \mathbb{R}^{q\times q}$ contain the nonzero singular values of $\mathbf{X}$ and $\mathbf{Y}$.

Using the SVD's in (7.8), we may rewrite the composite sample covariance matrix $\mathbf{S}_{zz}$ as

$$\begin{aligned} \mathbf{S}_{zz} = \begin{bmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{S}_{yy} \end{bmatrix} &= \begin{bmatrix} \mathbf{X}\mathbf{X}^T & \mathbf{X}\mathbf{Y}^T \\ \mathbf{Y}\mathbf{X}^T & \mathbf{Y}\mathbf{Y}^T \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{U}_x \mathbf{\Sigma}_x \mathbf{\Sigma}_x^T \mathbf{U}_x^T & \mathbf{U}_x \mathbf{\Sigma}_x \mathbf{V}_x^T \mathbf{V}_y \mathbf{\Sigma}_y^T \mathbf{U}_y^T \\ \mathbf{U}_y \mathbf{\Sigma}_y \mathbf{V}_y^T \mathbf{V}_x \mathbf{\Sigma}_x^T \mathbf{U}_x^T & \mathbf{U}_y \mathbf{\Sigma}_y \mathbf{\Sigma}_y^T \mathbf{U}_y^T \end{bmatrix} \end{aligned} \tag{7.10}$$

or alternatively

$$\begin{bmatrix} \mathbf{U}_x^T & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_y^T \end{bmatrix} \mathbf{S}_{zz} \begin{bmatrix} \mathbf{U}_x & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_y \end{bmatrix} = \begin{bmatrix} \mathbf{U}_x^T & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_y^T \end{bmatrix} \begin{bmatrix} \mathbf{X}\mathbf{X}^T & \mathbf{X}\mathbf{Y}^T \\ \mathbf{Y}\mathbf{X}^T & \mathbf{Y}\mathbf{Y}^T \end{bmatrix} \begin{bmatrix} \mathbf{U}_x & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_y \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{\Sigma}_x\mathbf{\Sigma}_x^T & \mathbf{\Sigma}_x\mathbf{V}_x^T\mathbf{V}_y\mathbf{\Sigma}_y^T \\ \mathbf{\Sigma}_y\mathbf{V}_y^T\mathbf{V}_x\mathbf{\Sigma}_x^T & \mathbf{\Sigma}_y\mathbf{\Sigma}_y^T \end{bmatrix}. \tag{7.11}$$

This formula once again shows that rows of $\mathbf{X}$ and $\mathbf{Y}$ are experimental surrogates for the elements of $\mathbf{x}$ and $\mathbf{y}$. In the SVD coordinates, it is the rows of $\mathbf{U}_x$ and $\mathbf{U}_y$ that serve as surrogates, but the weighting of inner products depends on $\mathbf{\Sigma}_x\mathbf{V}_x^T$ and $\mathbf{\Sigma}_y\mathbf{V}_y^T$.

In analogy to Chapter 2, we define the *sample coherence matrix* $\hat{\mathbf{C}}$ for $\mathbf{X}$ and $\mathbf{Y}$ as[2]

$$\hat{\mathbf{C}} = \mathbf{S}_{xx}^{-1/2}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-T/2} = (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-T/2}. \tag{7.12}$$

From here on it is understood that all the subsequent developments are for sample data cases, and hence the '$\hat{\ }$' notation is dropped, for the sake of simplicity.

Using (7.10), and plugging in for $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_y$ from (7.9), we may write $\mathbf{C}$ as

$$\begin{aligned} \mathbf{C} &= (\mathbf{U}_x\mathbf{\Sigma}_x\mathbf{\Sigma}_x^T\mathbf{U}_x^T)^{-1/2}\mathbf{U}_x\mathbf{\Sigma}_x\mathbf{V}_x^T\mathbf{V}_y\mathbf{\Sigma}_y^T\mathbf{U}_y^T(\mathbf{U}_y\mathbf{\Sigma}_y\mathbf{\Sigma}_y^T\mathbf{U}_y^T)^{-T/2} \\ &= \mathbf{U}_x(\mathbf{\Sigma}_x\mathbf{\Sigma}_x^T)^{-1/2}\mathbf{\Sigma}_x\mathbf{V}_x^T\mathbf{V}_y\mathbf{\Sigma}_y^T(\mathbf{\Sigma}_y\mathbf{\Sigma}_y^T)^{-T/2}\mathbf{U}_y^T \end{aligned} \tag{7.13}$$

and then simplify it to

$$\begin{aligned} \mathbf{C} &= \mathbf{U}_x \begin{bmatrix} \mathbf{\Sigma}_x^2(p) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^{-1/2} \begin{bmatrix} \mathbf{\Sigma}_x(p) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}_x^T\mathbf{V}_y \begin{bmatrix} \mathbf{\Sigma}_y(q)^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_y^2(q) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^{-T/2} \mathbf{U}_y^T \\ &= \mathbf{U}_x \begin{bmatrix} \mathbf{I}(p) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}_x^T\mathbf{V}_y \begin{bmatrix} \mathbf{I}(q) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}_y^T \end{aligned} \tag{7.14}$$

---

[2]Note that when $\mathbf{X}$ and $\mathbf{Y}$ are not full-rank, $(\mathbf{X}\mathbf{X}^T)^{-1/2}$ and $(\mathbf{Y}\mathbf{Y}^T)^{-1/2}$ denote Moore-Penrose pseudo inverses of $(\mathbf{X}\mathbf{X}^T)^{1/2}$ and $(\mathbf{Y}\mathbf{Y}^T)^{1/2}$.

with $\mathbf{\Sigma}_x^2(p) = \mathbf{\Sigma}_x(p)\mathbf{\Sigma}_x^T(p)$ and $\mathbf{\Sigma}_y^2(q) = \mathbf{\Sigma}_y(q)\mathbf{\Sigma}_y^T(q)$.

We now consider a full SVD for the sample coherence matrix $\mathbf{C}$ of the form

$$\mathbf{C} = \mathbf{F}_c\mathbf{\Sigma}_c\mathbf{G}_c^T \quad \text{and} \quad \mathbf{F}_c^T\mathbf{C}\mathbf{G}_c = \mathbf{\Sigma}_c \tag{7.15}$$

where $\mathbf{F}_c \in \mathbb{R}^{m\times m}$ and $\mathbf{G}_c \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\mathbf{\Sigma}_c \in \mathbb{R}^{m\times n}$ is diagonal. Note that since $\text{Rank}\{\mathbf{X}\} = p$ and $\text{Rank}\{\mathbf{Y}\} = q$, the sample coherence matrix $\mathbf{C}$ is of rank $r = \min(p, q)$. Thus, the singular value matrix $\mathbf{\Sigma}_c$ has only r nonzero elements $\sigma_i > 0$, $i \in [1, r]$.

Matching (7.14) and (7.15), we may write

$$\begin{bmatrix} \mathbf{I}(p) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}_x^T\mathbf{V}_y \begin{bmatrix} \mathbf{I}(q) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{U}_x^T\mathbf{F}_c\mathbf{\Sigma}_c\mathbf{G}_c^T\mathbf{U}_y = \mathbf{U}_x^T\mathbf{F}_c \begin{bmatrix} \mathbf{\Sigma}_c(r) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{G}_c^T\mathbf{U}_y \tag{7.16}$$

or alternatively

$$\begin{bmatrix} \mathbf{V}_x(:, 1:p)^T\mathbf{V}_y(:, 1:q) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{U}_x^T\mathbf{F}_c \begin{bmatrix} \mathbf{\Sigma}_c(r) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{G}_c^T\mathbf{U}_y, \tag{7.17}$$

where $\mathbf{\Sigma}_c(r) = \text{diag}(\sigma_1, \ldots, \sigma_r)$ is the diagonal matrix that carries the nonzero singular values of $\mathbf{C}$, and $\mathbf{V}_x(:, 1:p) \in \mathbb{R}^{M\times p}$ and $\mathbf{V}_y(:, 1:q) \in \mathbb{R}^{M\times q}$ are column-wise matrices that carry the first p columns of $\mathbf{V}_x$ and the first q columns of $\mathbf{V}_y$.

The column-wise matrices $\mathbf{V}_x(:, 1:p) \in \mathbb{R}^{M\times p}$ and $\mathbf{V}_y(:, 1:q) \in \mathbb{R}^{M\times q}$ are orthonormal bases for $p-$ and $q-$dimensional subspaces of $\mathbb{R}^M$. Therefore, singular values of $\mathbf{V}_x(:, 1:p)^T\mathbf{V}_y(:, 1:q)$ measure the cosines of principal angles between Col-Span$\{\mathbf{V}_x(:, 1:p)\}$ and Col-Span$\{\mathbf{V}_y(:, 1:q)\}$. Since $\mathbf{F}_c$, $\mathbf{G}_c$, $\mathbf{U}_x$, and $\mathbf{U}_y$ are orthogonal matrices, the rank-r matrices $\mathbf{V}_x(:, 1:p)^T\mathbf{V}_y(:, 1:q)$ and $\mathbf{\Sigma}_c(r)$ in (7.17) are similar. That is, the diagonal elements of $\mathbf{\Sigma}_c(r)$, i.e. σ_i, $i \in [1, r]$, are the singular values of $\mathbf{V}_x(:, 1:p)^T\mathbf{V}_y(:, 1:q)$, and hence σ_i measures the cosine of the ith principal angle between Col-Span$\{\mathbf{V}_x(:, 1:p)\}$ and Col-Span$\{\mathbf{V}_y(:, 1:q)\}$. However, the SVD's in (7.8) show that Col-Span$\{\mathbf{V}_x(:, 1:p)\}$ and Col-Span$\{\mathbf{V}_y(:, 1:q)\}$ are

the same as Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$, respectively. Thus, the σ_i measure the cosines of the principal angles between the row spaces of $\mathbf{X}$ and $\mathbf{Y}$.

Let us define the composite matrix $\mathbf{W}$ consisting of matrices $\mathbf{U}$ and $\mathbf{V}$ as

$$\begin{aligned}\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} &= \begin{bmatrix} \mathbf{F}_c^T \mathbf{S}_{xx}^{-1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_c^T \mathbf{S}_{yy}^{-1/2} \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{F}_c^T (\mathbf{X}\mathbf{X}^T)^{-1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_c^T (\mathbf{Y}\mathbf{Y}^T)^{-1/2} \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}.\end{aligned} \tag{7.18}$$

Then, the composite sample covariance matrix for $\mathbf{W}$ may be written as

$$\begin{aligned}\mathbf{S}_{ww} = \mathbf{W}\mathbf{W}^T = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} [\mathbf{U}^T \ \mathbf{V}^T] &= \begin{bmatrix} \mathbf{S}_{uu} & \mathbf{S}_{uv} \\ \mathbf{S}_{vu} & \mathbf{S}_{vv} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{F}_c^T \mathbf{P}_{\mathbf{U}_x(:,1:p)} \mathbf{F}_c & \mathbf{\Sigma}_c \\ \mathbf{\Sigma}_c^T & \mathbf{G}_c^T \mathbf{P}_{\mathbf{U}_y(:,1:q)} \mathbf{G}_c \end{bmatrix}\end{aligned} \tag{7.19}$$

where $\mathbf{P}_{\mathbf{U}_x(:,1:p)} = \mathbf{U}_x(:, 1:p)\mathbf{U}_x(:, 1:p)^T$ and $\mathbf{P}_{\mathbf{U}_y(:,1:q)} = \mathbf{U}_y(:, 1:q)\mathbf{U}_y(:, 1:q)^T$ are orthogonal projection matrices onto the $p-$ and $q-$dimensional subspaces spanned by the first p and q columns of $\mathbf{U}_x$ and $\mathbf{U}_y$, respectively. This equation shows that the diagonal singular value matrix $\mathbf{\Sigma}_c$ is the sample cross-covariance matrix between the matrices $\mathbf{U}$ and $\mathbf{V}$, i.e. $\mathbf{\Sigma}_c = \mathbf{U}\mathbf{V}^T$.

In Appendix C, it is shown that the diagonal elements of $\mathbf{\Sigma}_c$ form a maximal set of invariants for the composite covariance matrix $\mathbf{S}_{zz} = \mathbf{Z}\mathbf{Z}^T$ under the transformation group

$$\mathcal{T} = \left\{ \mathbf{T} = \begin{bmatrix} \mathbf{T}_1 \mathbf{U}_x(:, 1:p)^T & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 \mathbf{U}_y(:, 1:q)^T \end{bmatrix} \in \mathbb{R}^{(p+q)\times(p+q)}, \quad \begin{matrix} \det\{\mathbf{T}_1\} \neq 0 \\ \det\{\mathbf{T}_2\} \neq 0 \end{matrix} \right\} \tag{7.20}$$

with group action $\mathbf{S}_{zz} \rightarrow \mathbf{T}\mathbf{S}_{zz}\mathbf{T}^T$. Therefore, the cross-covariance matrix $\mathbf{\Sigma}_c$ is indeed the *sample canonical correlation matrix*, consisting of the *sample canonical correlations* σ_i, with $1 \geq \sigma_1 \geq \cdots \geq \sigma_m > 0$. These sample canonical correlations

mimic corresponding algebraic and geometric properties of canonical correlations in a Hilbert space.

The sample canonical correlation $\sigma_i = [\mathbf{U}\mathbf{V}^T]_{ii}$ is the experimental surrogate for the ith canonical correlation $\sigma_i = E[u_i v_i]$, where the standard inner product $[\mathbf{S}_{uv}]_{ij} = [\mathbf{U}\mathbf{V}^T]_{ij} = \mathbf{U}(i,:)\mathbf{V}(j,:)^T$ stands for the inner product $[\mathbf{R}_{uv}]_{ij} = E[u_i v_j]$ in the Hilbert space of second-order random variables. Correspondingly, the rows of the matrices $\mathbf{U}$ and $\mathbf{V}$ in (7.18) are surrogates for the canonical coordinates $\mathbf{u}$ and $\mathbf{v}$, respectively. However, contrasting to the Hilbert space case in Chapter 2, i.e. (2.10), the sample auto-covariance matrices for $\mathbf{U}$ and $\mathbf{V}$ in (7.19) are not necessarily identity. Rather they are orthogonal projection matrices given by $\mathbf{U}\mathbf{U}^T = \mathbf{F}_c^T \mathbf{P}_{\mathbf{U}_x(:,1:p)} \mathbf{F}_c$ and $\mathbf{V}\mathbf{V}^T = \mathbf{G}_c^T \mathbf{P}_{\mathbf{U}_y(:,1:q)} \mathbf{G}_c$. Only when $p = m$ and $q = n$, which yields full-rank sample covariance matrices for $\mathbf{X}$ and $\mathbf{Y}$, are the sample covariance matrices for $\mathbf{U}$ and $\mathbf{V}$ identity, as $\mathbf{P}_{\mathbf{U}_x(:,1:m)} = \mathbf{U}_x \mathbf{U}_x^T = \mathbf{I}$ and $\mathbf{P}_{\mathbf{U}_y(:,1:n)} = \mathbf{U}_y \mathbf{U}_y^T = \mathbf{I}$. Note that in this case, the transformation group in (7.20), under which the sample canonical correlations σ_i are maximal invariants, is of form (2.14), as in this case $\mathbf{U}_x(:, 1:p)^T = \mathbf{U}_x^T$ and $\mathbf{U}_y(:, 1:q)^T = \mathbf{U}_y^T$ are orthogonal matrices and act like nonsingular transformations on $\mathbf{X}$ and $\mathbf{Y}$.

From the discussion following (7.17), sample canonical correlations σ_i are cosines of the principal angles between Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$. Therefore, it is the *row spaces* of $\mathbf{X}$ and $\mathbf{Y}$ that determine the sample canonical correlations. This fits intuition, as in the sample data case the ith rows of $\mathbf{X}$ and $\mathbf{Y}$ are experimental surrogates for x_i and y_i, where the Euclidean space inner product $[\mathbf{S}_{xy}]_{ij} = [\mathbf{X}\mathbf{Y}^T]_{ij} = \mathbf{X}(i,:)\mathbf{Y}(j,:)^T$ stands for the Hilbert space inner product $[\mathbf{R}_{xy}]_{ij} = E[x_i y_j]$.

So far, we have made no assumption about the number of samples drawn from each channel, i.e M, and the dimension of the row spaces of $\mathbf{X}$ and $\mathbf{Y}$. We have established here that the empirical canonical correlations measure the cosines of principal angles in Euclidean space, where row vectors are surrogate for random variables. The

question is then, can these principal angles in Euclidean space estimate the principal angles in the Hilbert space of second-order random variables. In what follows we shall discuss how the algebraic and geometric properties of the empirical canonical correlations, carried by the surrogate rows of $\mathbf{U}$ and $\mathbf{V}$, may be affected by M, the number of samples, and p and q, the dimensions of the row spaces $\mathbf{X}$ and $\mathbf{Y}$. For this purpose, we consider two different cases; namely *sample-poor* and *sample-rich.*

7.2.1 Case 1: Sample-Poor ($M < p + q$)

In this case, the number of data samples M drawn from each channel is smaller than the sum of the dimensions of the row spaces of $\mathbf{X}$ and $\mathbf{Y}$, i.e. $M < p + q$. Referring to the two-channel linear model in (7.4), assume, without loss of generality, that the column mean vectors of $\boldsymbol{\Theta}$ and $\boldsymbol{\Upsilon}$ are zero, i.e. $\sum_{i=1}^{M} \boldsymbol{\theta}_i = \mathbf{0}$ and $\sum_{i=1}^{M} \boldsymbol{\upsilon}_i = \mathbf{0}$. Then, a sample-poor case occurs when the number of samples M is smaller than the sum of the ranks of the matrices $\mathbf{A} \in \mathbb{R}^{m \times d}$ and $\mathbf{B} \in \mathbb{R}^{n \times l}$. Note that the matrices $\mathbf{A}$ and $\mathbf{B}$ may be chosen so that they transform the data samples $\boldsymbol{\theta}_i$ and $\boldsymbol{\upsilon}_i$ to a higher, lower, or same dimensions. All that matters is the relationship between the ranks of $\mathbf{A}$ and $\mathbf{B}$, and the number of samples M. For nonlinear function $\boldsymbol{\phi}(\cdot)$ and $\boldsymbol{\psi}(\cdot)$ of the two-channel data $\boldsymbol{\mu} = [\boldsymbol{\theta}^T \ \boldsymbol{\upsilon}^T]^T$, a sample-poor case may occur when mapping the data samples $\boldsymbol{\theta}_i$ and $\boldsymbol{\upsilon}_i$ to a higher, lower, or same dimensions, as long as after the mapping M is smaller than the sum of the dimensions of the row spaces of $\mathbf{X}$ and $\mathbf{Y}$. Usually, when m and n, the dimensions of the nonlinearly mapped data samples $\mathbf{x}_i = \boldsymbol{\phi}(\boldsymbol{\theta}_i)$ and $\mathbf{y}_i = \boldsymbol{\psi}(\boldsymbol{\upsilon}_i)$, are larger than the number of samples M, the sample data matrices $\mathbf{X}$ and $\mathbf{Y}$ are sample-poor. This, typically occurs in kernel-nonlinear information processing methods, where the original data samples $\boldsymbol{\theta}_i$ and $\boldsymbol{\upsilon}_i$ are implicitly mapped, using nonlinear mappings, into high-dimensional spaces. This is also the case in most of the kernel approaches to canonical correlation analysis [39]- [45].

As established earlier in this section, the sample canonical correlations σ_i measure the cosines of the principal angles between Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$. However, the row-space of $\mathbf{X} \in \mathbb{R}^{m\times M}$ is a $p-$dimensional subspace of $\mathbb{R}^M$, and the row space of $\mathbf{Y} \in \mathbb{R}^{n\times M}$ is a $q-$dimensional subspace of $\mathbb{R}^M$. Therefore, when $p+q > M$, Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$ have to share a subspace of $\mathbb{R}^M$ of dimension at least $d = (p+q) - M$. Since these two subspaces overlap in at least d dimensions, the cosines of at least d principal angles, or equivalently d sample canonical correlations will be equal to one, regardless of the two-channel vector process that generates the samples. This result has also been reported in [45], but no rigorous proof or analysis for the geometric properties of sample canonical correlations is given. Therefore, in this case, the sample canonical correlation matrix $\mathbf{\Sigma}_c$ may be expressed as

$$\mathbf{\Sigma}_c = \mathbf{U}\mathbf{V}^T = \begin{bmatrix} \mathbf{I}(d) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_c(\star) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{7.21}$$

where $\mathbf{\Sigma}_c(\star) = \text{diag}(\sigma_{d+1}, \ldots, \sigma_r)$, with $1 \geq \sigma_{d+1} \geq \cdots \geq \sigma_r > 0$ and $r = \min(p, q)$. This shows that even when the samples are drawn from a two-channel process in which the elements of $\mathbf{x}$ and $\mathbf{y}$ are mutually uncorrelated, some empirical canonical correlations become one as a result of poor sampling. Therefore, in this case the empirical canonical correlations between the data matrices $\mathbf{X}$ and $\mathbf{Y}$ definitely do not estimate the canonical correlations between the random vectors $\mathbf{x}$ and $\mathbf{y}$. Equivalently, the principal angles between the Euclidean spaces spanned by rows of $\mathbf{X}$ and $\mathbf{Y}$ *do not estimate* the principal angles between the Hilbert spaces spanned by the random variables in $\mathbf{x}$ and $\mathbf{y}$.

Equation (7.21) also shows that the $d-$dimensional subspace of $\mathbb{R}^M$ that is shared by Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$ is spanned by the first d surrogate rows of $\mathbf{U}$, or alternatively by the first d surrogate rows of $\mathbf{V}$, as illustrated in Figure 7.1. In this figure, the planes show the row spaces of $\mathbf{X}$ and $\mathbf{Y}$ and the intersection line shows

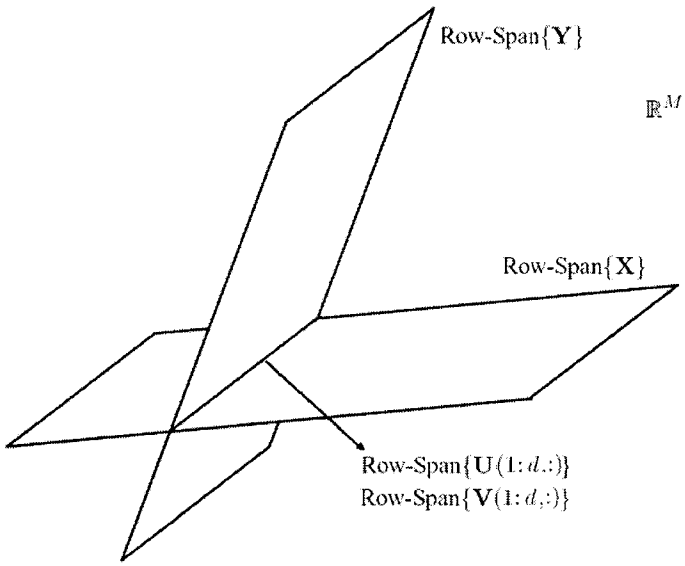

Figure 7.1: Geometry of empirical canonical correlations in a sample-poor case.

the $d-$dimensional subspace shared between Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$.

Since the p columns of $\mathbf{V}_x(:, 1:p)$ and the q columns of $\mathbf{V}_y(:, 1:q)$ are, respectively, orthonormal bases for Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$, the fact that at least d of the principal cosines between Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$ are equal to one shows that at least $d = (p+q) - M$ columns of $\mathbf{V}_x(:, 1:p)$ may always be rotated so that they will be perfectly aligned with d columns of $\mathbf{V}_y(:, 1:q)$. The SVD of $\mathbf{V}_x(:, 1:p)^T\mathbf{V}_y(:, 1:q)$ in (7.17) shows that the rotation matrices that perform this alignment are implemented by $\mathbf{U}_x(:, 1:p)^T\mathbf{F}_c(:, 1:p)$ and $\mathbf{U}_y(:, 1:q)^T\mathbf{G}_c(:, 1:q)$, yielding the row-wise basis vectors $\mathbf{U}(1:p, :)$ and $\mathbf{V}(1:q, :)$.

There are two special scenarios of a sample-poor case, the studies of which are particularly illuminating.

Scenario 1: Consider a sample-poor case in which the dimensions of the row spaces of $\mathbf{X}$ and $\mathbf{Y}$ are $p = q = M - 1$, with $M > 2$. This, obviously is a sample-poor case, as $p + q = 2M - 2 > M$ for $M > 2$. It must be pointed out that this scenario occurs only when M, the number of samples, is smaller than both m and n, the channel dimensions.

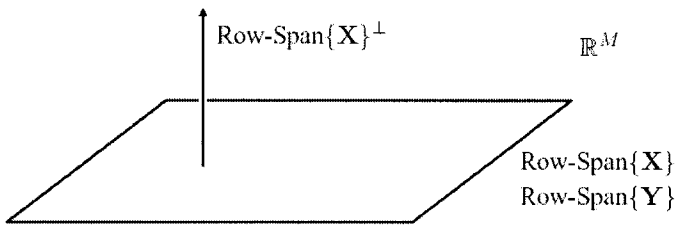

Figure 7.2: Geometry of empirical canonical correlations in a sample-poor case, where row spaces of $\mathbf{X}$ and $\mathbf{Y}$ are identical.

In this scenario, Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$ are both $(M-1)-$dimensional subspaces of $\mathbb{R}^M$. Therefore, the row spaces of $\mathbf{X}$ and $\mathbf{Y}$ have to share a subspace of $\mathbb{R}^M$ of dimension at least $d = 2(M-1) - M = M - 2$. Consequently, at least the first $d = M - 2$ (out of $M - 1$) principal cosines or equivalently the first $d = M - 2$ sample canonical correlations, i.e. σ_i, $i \in [1, M-2]$ are equal to one. When the dimension of the shared subspace is exactly $d = M - 2$, the unshared dimensions of the Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$ have to be orthogonal to each other. In such case, the $(M-1)$th principal cosine or equivalently the $(M-1)$th sample canonical correlation σ_{M-1} will be zero. Thus, we can say that in a sample-poor case where the dimension of the row spaces of $\mathbf{X}$ and $\mathbf{Y}$ are $p = q = M - 1$, all the nonzero sample canonical correlations are equal to one, regardless of the two-channel vector process that generates the samples. That is, the sample canonical correlation matrix $\mathbf{\Sigma}_c$ may be expressed as

$$\mathbf{\Sigma}_c = \mathbf{U}\mathbf{V}^T = \begin{bmatrix} \mathbf{I}(d) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{m \times n}, \tag{7.22}$$

where d is the dimension of the shared subspace between Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$. When all the $M-1$ sample canonical correlations are equal to one, the row spaces of $\mathbf{X}$ and $\mathbf{Y}$ are identical, as illustrated in Figure 7.2. In such case, the surrogate rows of $\mathbf{X}$ may be perfectly estimated from the surrogate rows of $\mathbf{Y}$, or vice versa, with MSE of zero.

Therefore, in this scenario, the $p = M - 1$ columns of $\mathbf{V}_x(:, 1 : M - 1)$ and the $q = M - 1$ columns of $\mathbf{V}_y(:, 1 : M - 1)$ may always be rotated so that at least $d = M - 2$ of them are perfectly aligned. When the alignment is possible for only $M-2$ of these basis vectors, the $(M-1)$th pair of basis vectors that cannot be aligned are orthogonal to each other. The alignment of all the columns of $\mathbf{V}_x(:, 1 : M - 1)$ and $\mathbf{V}_y(:, 1 : M - 1)$ is possible only when all the $M - 1$ principal cosines are equal to one, or equivalently when Row-Span$\{\mathbf{X}\}$ = Row-Span$\{\mathbf{Y}\}$.

Scenario 2: The other interesting scenario of the sample-poor case is when the dimension of the row spaces of $\mathbf{X} \in \mathbb{R}^{m \times M}$ and $\mathbf{Y} \in \mathbb{R}^{n \times M}$ are $p = m$ and $q = n$, respectively, and further $p + q = m + n > M$. Obviously, this case may occur only when the number of samples M is greater than m and greater than n, the channel dimensions, and Rank$\{\mathbf{X}\} = p = m$ and Rank$\{\mathbf{Y}\} = q = n$.

Since in this scenario the sample data matrices $\mathbf{X}$ and $\mathbf{Y}$ are full-rank, the sample covariance matrices $\mathbf{S}_{xx} = \mathbf{X}\mathbf{X}^T$ and $\mathbf{S}_{yy} = \mathbf{Y}\mathbf{Y}^T$ are nonsingular. Consequently, the sample covariance matrices $\mathbf{S}_{uu} = \mathbf{U}\mathbf{U}^T$ and $\mathbf{S}_{vv} = \mathbf{V}\mathbf{V}^T$ in (7.19) will be equal to identity, as the orthogonal projection matrices $\mathbf{P}_{\mathbf{U}_x(:,1:m)} = \mathbf{U}_x\mathbf{U}_x^T$ and $\mathbf{P}_{\mathbf{U}_y(:,1:n)} = \mathbf{U}_y\mathbf{U}_y^T$ reduce to identity matrices. Additionally, as the matrices $\mathbf{U}_x(:, 1 : m) = \mathbf{U}_x$ and $\mathbf{U}_y(:, 1 : n) = \mathbf{U}_y$ are nonsingular, the transformation group in (7.20), reduces to the one in (2.14). We must note that this is the only scenario of a sample-poor case in which the sample covariance matrices $\mathbf{S}_{xx}$ and $\mathbf{S}_{yy}$ are nonsingular.

In this scenario, the samples drawn from the channels are enough to form full-rank sample covariance matrices. Nonetheless, since $p+q = m+n > M$, the row spaces of $\mathbf{X}$ and $\mathbf{Y}$ still share a subspace of $\mathbb{R}^M$ of dimension at least $d = (p+q)-M = (m+n)-M$. In other words, at least d of the principal cosines or sample canonical correlations are equal to one, regardless of the two-channel vector process that the samples are drawn from. This shows that nonsingular sample covariance matrices do not guarantee that sample canonical correlations σ_i estimate the actual correlation or coherence between

the two-channel processes $\mathbf{x}$ and $\mathbf{y}$. Rather they may still have to be interpreted solely as principal cosines between two linear subspaces of Euclidean space and nothing more.

Summarizing our findings, in a sample-poor case, where $M < p + q$, the sample canonical correlations only explain the underlying Euclidean geometry of the two-channel sample data matrices. The sample canonical correlations σ_i should solely be interpreted as principal cosines between two linear subspaces of Euclidean space and definitely not as estimates of canonical correlations or principal cosines in the Hilbert space of second-order random variables. Consequently, in a sample-poor case (linear or nonlinear) the empirical canonical correlations are defective and may not be used in any inference based on estimates of theoretical canonical correlations. This result implies that when the pre-processing of two-channel data samples $\boldsymbol{\Theta}$ and $\boldsymbol{\Upsilon}$ results in sample-poor data matrices $\mathbf{X}$ and $\mathbf{Y}$, the sample canonical correlations between the surrogate rows of $\mathbf{X}$ and $\mathbf{Y}$ may not be used as estimates of coherence between the nonlinear functions (high-order attributes) of the random variables in $\boldsymbol{\theta}$ and $\boldsymbol{\upsilon}$.

7.2.2 Case 2: Sample-Rich ($M \geq p + q$)

In this case the number of samples drawn from each channel is greater than or equal to the sum of the dimensions of the row spaces of $\mathbf{X}$ and $\mathbf{Y}$, i.e. $M \geq p + q$. Note that being sample-rich does not guarantee that the sample covariance matrices $\mathbf{S}_{xx} = \mathbf{X}\mathbf{X}^T$, $\mathbf{S}_{xy} = \mathbf{X}\mathbf{Y}^T$, and $\mathbf{S}_{yy} = \mathbf{Y}\mathbf{Y}^T$ are nonsingular, as $p = \text{Rank}\{\mathbf{X}\}$ and $q = \text{Rank}\{\mathbf{Y}\}$ may be smaller than m and n.

Referring again to the linear two–channel model in (7.4), assume, without loss of generality, that the column mean vectors of $\boldsymbol{\Theta}$ and $\boldsymbol{\Upsilon}$ are zero. Then, a sample-rich case occurs when the sum of the ranks of the matrices $\mathbf{A} \in \mathbb{R}^{m \times d}$ and $\mathbf{B} \in \mathbb{R}^{n \times l}$ is smaller than M. The matrices $\mathbf{A}$ and $\mathbf{B}$ may be chosen so that they transform the data samples $\boldsymbol{\theta}_i$ and $\boldsymbol{\upsilon}_i$ to a higher, lower, or same dimensions. All that matters is the relationship between the ranks of $\mathbf{A}$ and $\mathbf{B}$ and the number of samples M. For

nonlinear mappings $\boldsymbol{\phi}(\cdot)$ and $\boldsymbol{\psi}(\cdot)$, a sample-rich case may occur when mapping the data samples $\boldsymbol{\theta}_i$ and $\boldsymbol{\upsilon}_i$ to a higher, lower, or same dimensions, provided that after the mapping M is greater than or equal to the sum of the dimensions of the row spaces of $\mathbf{X}$ and $\mathbf{Y}$.

In this case, the row spaces of $\mathbf{X}$ and $\mathbf{Y}$ are still $p-$ and $q-$dimensional subspaces of $\mathbb{R}^M$. However, as $p+q \leq M$, Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$ do not necessarily share a subspace of $\mathbb{R}^M$. Therefore, in this case, the first principal cosine between Row-Span$\{\mathbf{X}\}$ and Row-Span$\{\mathbf{Y}\}$, or equivalently the first sample canonical correlation σ_1, is not necessarily equal to one. This implies that, when $M \geq p+q$ it is possible to use sample canonical correlations or principal cosines between $\mathbf{X}$ and $\mathbf{Y}$ to estimate the canonical correlations or principal cosines between the random variables in $\mathbf{x}$ and $\mathbf{y}$. However, these sample canonical correlations may provide poor estimates of the theoretical ones, as we typically need a large number of samples to estimate the theoretical covariance matrices. Thus, all we say here is that contrasting to the sample-poor case, in a sample-rich case the sample canonical correlations or principal cosines in Euclidean space are no longer defective for estimating canonical correlations or principal cosines in the corresponding Hilbert space of second-order random variables. In other words when the number of samples drawn from each channel exceeds the sum of the ranks of the two data matrices, it is possible to use the sample canonical correlations as estimates of the theoretical ones.

Therefore, when the mappings $\boldsymbol{\phi} : \mathbb{R}^d \longrightarrow \mathbb{R}^m$ and $\boldsymbol{\psi} : \mathbb{R}^l \longrightarrow \mathbb{R}^n$ are nonlinear, the sample data matrices $\mathbf{X} \in \mathbb{R}^{m \times M}$ and $\mathbf{Y} \in \mathbb{R}^{n \times M}$ must be sample-rich for the sample canonical correlations to be estimates of canonical correlations between nonlinear functions of the elements of the two-channel process $\boldsymbol{\mu} = [\boldsymbol{\theta}^T \ \boldsymbol{\upsilon}^T]^T$.

7.3 Implications for Kernel Canonical Correlation Analysis

In this section, based on our findings about empirical canonical correlations, we clarify whether or not kernel formulations for canonical correlation analysis [39]- [46] are indeed useful. We wish to determine whether or not the kernel formulations have less computational complexity than the direct formulations, particularly in cases where the empirical canonical correlations of the nonlinearly mapped data samples are to be used as estimates of coherence between the random variables in the original data channels.

In a kernel approach, the surrogate rows of $\mathbf{U}$ and $\mathbf{V}$ are written in such a way that only the inner products of the mapped samples $\mathbf{x}_i = \boldsymbol{\phi}(\boldsymbol{\theta}_i)$ and $\mathbf{y}_i = \boldsymbol{\psi}(\boldsymbol{\upsilon}_i)$ appear in the formulation. Further, these formulations use only nonlinearities for which their inner products build kernel functions satisfying the Mercer condition [35]. Then inner products $\mathbf{x}_i^T\mathbf{x}_j = \boldsymbol{\phi}(\boldsymbol{\theta}_i)^T\boldsymbol{\phi}(\boldsymbol{\theta}_j)$ and $\mathbf{y}_i^T\mathbf{y}_j = \boldsymbol{\psi}(\boldsymbol{\upsilon}_i)^T\boldsymbol{\psi}(\boldsymbol{\upsilon}_j)$ with the corresponding Mercer kernels $k_x(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ and $k_y(\boldsymbol{\upsilon}_i, \boldsymbol{\upsilon}_j)$ are computed directly using the original sample pairs $\{\boldsymbol{\theta}_i, \boldsymbol{\theta}_j\}$ and $\{\boldsymbol{\upsilon}_i, \boldsymbol{\upsilon}_j\}$, hence obviating the need to explicitly compute the nonlinear mappings $\mathbf{x}_i = \boldsymbol{\phi}(\boldsymbol{\theta}_i)$ and $\mathbf{y}_i = \boldsymbol{\psi}(\boldsymbol{\upsilon}_i)$. With this so called "kernel trick" [35], the kernel Gram matrices $\mathbf{K}_x = \mathbf{X}^T\mathbf{X} = [k_x(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)]_{i,j=1}^M \in \mathbb{R}^{M\times M}$ and $\mathbf{K}_y = \mathbf{Y}^T\mathbf{Y} = [k_y(\boldsymbol{\upsilon}_i, \boldsymbol{\upsilon}_j)]_{i,j=1}^M \in \mathbb{R}^{M\times M}$ appear in the formulations, instead of the sample covariance matrices $\mathbf{S}_{xx} = \mathbf{X}\mathbf{X}^T \in \mathbb{R}^{m\times m}$, $\mathbf{S}_{yy} = \mathbf{Y}\mathbf{Y}^T \in \mathbb{R}^{n\times n}$, and $\mathbf{S}_{xy} = \mathbf{X}\mathbf{Y}^T \in \mathbb{R}^{m\times n}$.

To further clarify, we derive a kernel formulation for canonical correlation analysis as follows. Pre-multiply $(\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-T/2} = \mathbf{F}_c\boldsymbol{\Sigma}_c\mathbf{G}_c^T$ by $(\mathbf{X}\mathbf{X}^T)^{1/2}$ and post-multiply it by $\mathbf{G}_c$ to obtain

$$\mathbf{P}_{\mathbf{X}}\mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-T/2}\mathbf{G}_c = (\mathbf{X}\mathbf{X}^T)^{1/2}\mathbf{F}_c\boldsymbol{\Sigma}_c. \tag{7.23}$$

Here, since $\mathbf{X}$ is not in general full-rank, the matrix $\mathbf{P}_{\mathbf{X}} = (\mathbf{X}\mathbf{X}^T)^{1/2}(\mathbf{X}\mathbf{X}^T)^{-1/2} = \mathbf{U}_x(:, 1:p)\mathbf{U}_x(:, 1:p)^T$ is an orthogonal projection.

Similarly, pre-multiply $(\mathbf{Y}\mathbf{Y}^T)^{-1/2}\mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-T/2} = \mathbf{G}_c\mathbf{\Sigma}_c^T\mathbf{F}_c^T$ by $(\mathbf{Y}\mathbf{Y}^T)^{1/2}$ and post-multiply it by $\mathbf{F}_c$ to obtain

$$\mathbf{P}_\mathbf{Y}\mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-T/2}\mathbf{F}_c = (\mathbf{Y}\mathbf{Y}^T)^{1/2}\mathbf{G}_c\mathbf{\Sigma}_c^T, \tag{7.24}$$

where $\mathbf{P}_\mathbf{Y} = (\mathbf{Y}\mathbf{Y}^T)^{1/2}(\mathbf{Y}\mathbf{Y}^T)^{-1/2} = \mathbf{U}_y(:, 1:q)\mathbf{U}_y(:, 1:q)^T$ is an orthogonal projection. Define

$$\mathbf{D}_x = (\mathbf{X}\mathbf{X}^T)^{-T/2}\mathbf{F}_c \quad \text{and} \quad \mathbf{D}_y = (\mathbf{Y}\mathbf{Y}^T)^{-T/2}\mathbf{G}_c, \tag{7.25}$$

and note that $\mathbf{P}_\mathbf{X}\mathbf{X} = \mathbf{X}$ and $\mathbf{P}_\mathbf{Y}\mathbf{Y} = \mathbf{Y}$. Then, we may rewrite (7.23) and (7.24) as

$$\begin{aligned} \mathbf{X}\mathbf{Y}^T\mathbf{D}_y &= \mathbf{X}\mathbf{X}^T\mathbf{D}_x\mathbf{\Sigma}_c \\ \mathbf{Y}\mathbf{X}^T\mathbf{D}_x &= \mathbf{Y}\mathbf{Y}^T\mathbf{D}_y\mathbf{\Sigma}_c^T. \end{aligned} \tag{7.26}$$

Equation (7.26) is a coupled asymmetric generalized eigenvalue problem that may be solved for $\mathbf{D}_x$ and $\mathbf{D}_y$, and the shared eigenvalue matrix $\mathbf{\Sigma}_c\mathbf{\Sigma}_c^T$. We note that this generalized eigenvalue problem is identical in form with the generalized eigenvalue problem in (3.13). However, in deriving the generalized eigenvalue problem of (3.13) we had assumed that channel covariance matrices were known and full-rank. Therefore, we couldn't simply replace the theoretical covariance matrices in (3.13) with their corresponding sample estimates, which may be rank-deficient. Thus, we had to re-derive this generalized eigenvalue problem here for the sample data case.

The sample canonical correlation matrix $\mathbf{\Sigma}_c$ and the surrogate rows of $\mathbf{U}$ and $\mathbf{V}$ may now be determined from

$$\begin{aligned} \mathbf{U} &= \mathbf{F}_c^T(\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{X} = \mathbf{D}_x^T\mathbf{X} \\ \mathbf{V} &= \mathbf{G}_c^T(\mathbf{Y}\mathbf{Y}^T)^{-1/2}\mathbf{Y} = \mathbf{D}_y^T\mathbf{Y} \\ \mathbf{\Sigma}_c &= \mathbf{U}\mathbf{V}^T = \mathbf{D}_x^T\mathbf{X}\mathbf{Y}^T\mathbf{D}_y. \end{aligned} \tag{7.27}$$

From (7.25), it may be shown (see Appendix D) that Col-Span$\{\mathbf{D}_x\}$ = Col-Span$\{\mathbf{X}\}$ and Col-Span$\{\mathbf{D}_y\}$ = Col-Span$\{\mathbf{Y}\}$. Thus, the matrices $\mathbf{D}_x$ and $\mathbf{D}_y$ may be written as

$$\mathbf{D}_x = \mathbf{X}\mathbf{Q}_x \quad \text{and} \quad \mathbf{D}_y = \mathbf{Y}\mathbf{Q}_y, \tag{7.28}$$

where $\mathbf{Q}_x \in \mathbb{R}^{M\times m}$ and $\mathbf{Q}_y \in \mathbb{R}^{M\times n}$ are full-rank matrices.

Pre-multiplying the first equation in (7.26) by $\mathbf{X}^T$ and the second one by $\mathbf{Y}^T$, and then using (7.28) yields the following coupled generalized asymmetric eigenvalue problem for $\mathbf{Q}_x$, $\mathbf{Q}_y$, and $\mathbf{\Sigma}_c$:

$$\begin{aligned} \mathbf{K}_x\mathbf{K}_y\mathbf{Q}_y &= \mathbf{K}_x\mathbf{Q}_x\mathbf{\Sigma}_c \\ \mathbf{K}_y\mathbf{K}_x\mathbf{Q}_x &= \mathbf{K}_y\mathbf{Q}_y\mathbf{\Sigma}_c^T \end{aligned} \tag{7.29}$$

where $\mathbf{K}_x = \mathbf{X}^T\mathbf{X}$ and $\mathbf{K}_y = \mathbf{Y}^T\mathbf{Y}$ are the kernel Gram matrices for $\mathbf{X}$ and $\mathbf{Y}$, respectively. Plugging $\mathbf{D}_x = \mathbf{X}\mathbf{Q}_x$ and $\mathbf{D}_y = \mathbf{Y}\mathbf{Q}_y$ into (7.27), the solutions for $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{\Sigma}_c$ are

$$\begin{aligned} \mathbf{U} &= \mathbf{D}_x^T\mathbf{X} = \mathbf{Q}_x^T\mathbf{X}^T\mathbf{X} = \mathbf{Q}_x^T\mathbf{K}_x \\ \mathbf{V} &= \mathbf{D}_y^T\mathbf{Y} = \mathbf{Q}_y^T\mathbf{Y}^T\mathbf{Y} = \mathbf{Q}_y^T\mathbf{K}_y \\ \mathbf{\Sigma}_c &= \mathbf{U}\mathbf{V}^T = \mathbf{Q}_x^T\mathbf{K}_x\mathbf{K}_y^T\mathbf{Q}_y. \end{aligned} \tag{7.30}$$

The generalized eigenvalue problem of (7.29) for $\mathbf{Q}_x$ and $\mathbf{Q}_y$, and the expressions for $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{\Sigma}_c$ in (7.30) only depend on the kernel Gram matrices $\mathbf{K}_x$ and $\mathbf{K}_y$. The elements of $\mathbf{K}_x = [k_x(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)]_{i,j=1}^M$ and $\mathbf{K}_y = [k_y(\boldsymbol{\upsilon}_i, \boldsymbol{\upsilon}_j)]_{i,j=1}^M$ are the Mercer kernel functions that may be directly computed from the data samples in $\mathbf{\Theta}$ and $\mathbf{\Upsilon}$. Consequently, in computing the sample canonical coordinate matrices $\mathbf{U}$ and $\mathbf{V}$, and the sample canonical correlation matrix $\mathbf{\Sigma}_c$, only the inner products $k_x(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ and $k_y(\boldsymbol{\upsilon}_i, \boldsymbol{\upsilon}_j)$ are needed, and the explicit computation of the nonlinear mappings $\mathbf{x}_i = \boldsymbol{\phi}(\boldsymbol{\theta}_i)$ and $\mathbf{y}_i = \boldsymbol{\psi}(\boldsymbol{\upsilon}_i)$ is not required.

Remark: Here, we have assumed that the columns of the mapped sample data matrices $\mathbf{X}$ and $\mathbf{Y}$ are centered, which seems to require the explicit computation of columns of $\mathbf{X}$ and $\mathbf{Y}$. However, it may easily be shown that this mean correction may be accounted for by simply replacing $\mathbf{K}_x$ and $\mathbf{K}_y$ with $\mathbf{P}_\mathbf{1}^\perp\mathbf{K}_x\mathbf{P}_\mathbf{1}^\perp$ and $\mathbf{P}_\mathbf{1}^\perp\mathbf{K}_y\mathbf{P}_\mathbf{1}^\perp$, where $\mathbf{P}_\mathbf{1}^\perp$ is the (centering) orthogonal projection matrix $\mathbf{P}_\mathbf{1}^\perp = \mathbf{I} - \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T$ and $\mathbf{1} = [1, 1, \ldots, 1]^T \in \mathbb{R}^M$.

Comparing (7.26) and (7.29), the generalized eigenvalue problem in (7.26) requires the computation and manipulation of the sample covariance matrices $\mathbf{S}_{xx} = \mathbf{X}\mathbf{X}^T$, $\mathbf{S}_{xy} = \mathbf{X}\mathbf{Y}^T$, and $\mathbf{S}_{yy} = \mathbf{Y}\mathbf{Y}^T$ of dimensions $m \times m$, $m \times n$, and $n \times n$, respectively, whereas the formulation in (7.29) requires computation and manipulation of the kernel Gram matrices $\mathbf{K}_x = \mathbf{X}^T\mathbf{X}$ and $\mathbf{K}_y = \mathbf{Y}^T\mathbf{Y}$ of dimension $M \times M$.

When the nonlinearities $\boldsymbol{\phi}(\cdot)$ and $\boldsymbol{\psi}(\cdot)$ are chosen so that m and n, the dimensions of the mapped data samples $\mathbf{x}_i = \boldsymbol{\phi}(\boldsymbol{\theta}_i)$ and $\mathbf{y}_i = \boldsymbol{\psi}(\boldsymbol{\upsilon}_i)$, are greater than the number of samples M, i.e. $M < m \leq n$, the kernel formulation in (7.29) is computationally more efficient than the direct formulation in (7.26), as the dimensions of the kernel Gram matrices $\mathbf{K}_x \in \mathbb{R}^{M \times M}$ and $\mathbf{K}_y \in \mathbb{R}^{M \times M}$ are smaller than the dimensions of the sample covariance matrices $\mathbf{S}_{xx} \in \mathbb{R}^{m \times m}$, $\mathbf{S}_{xy} \in \mathbb{R}^{m \times n}$, and $\mathbf{S}_{yy} \in \mathbb{R}^{n \times n}$. This is the case that is typically considered in kernel canonical correlation analysis, as it is only in this case that the formulation in (7.29) can offer a computational advantage with respect to the direct formulation in (7.26). However, when $M < m \leq n$, the sample data matrices $\mathbf{X}$ and $\mathbf{Y}$ are typically sample-poor, i.e. $M < p + q$. Consequently, the sample canonical correlations are defective and do not estimate coherence between the nonlinear functions of the random variables in $\boldsymbol{\theta}$ and $\boldsymbol{\upsilon}$. When $\mathbf{X}$ and $\mathbf{Y}$ are sample-rich, i.e. $M \geq p+q$, but $M < m \leq n$, the empirical canonical correlations will be poor estimates of the theoretical canonical correlations. Therefore, in cases where the kernel formulation is computationally advantageous with respect to the direct formulation, the sample canonical correlations between the nonlinearly mapped data matrices do not usefully estimate coherence between high-order attributes of the random variables in the original data channels.

7.4 Simulation Results

We begin this section with a simple simulation to demonstrate the effect of sample support on canonical correlation analysis. Then, we present three simulation examples to show that pre-processing two-channel data with nonlinear mappings prior to canonical correlation analysis may reveal coherence between high-order attributes of the original data channels, even when second-order analysis of coherence would show them to be non-coherent.

The sample data matrices $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M] \in \mathbb{R}^{4\times M}$ and $\boldsymbol{\Upsilon} = [\boldsymbol{\upsilon}_1, \ldots, \boldsymbol{\upsilon}_M] \in \mathbb{R}^{4\times M}$ are formed from column-wise collections of M random samples drawn from the two-channel model

$$\boldsymbol{\theta} = \boldsymbol{\eta}_\theta$$
$$\boldsymbol{\upsilon} = \boldsymbol{\theta} \circ \boldsymbol{\theta} + \boldsymbol{\eta}_\upsilon.$$

In this model $\boldsymbol{\eta}_\theta \in \mathbb{R}^4$ and $\boldsymbol{\eta}_\upsilon \in \mathbb{R}^4$ are two independent white Gaussian vectors with densities $N(\mathbf{0}, \mathbf{I})$ and $N(\mathbf{0}, 0.1\mathbf{I})$, respectively, and $\boldsymbol{\theta} \circ \boldsymbol{\theta}$ is the Schur product of $\boldsymbol{\theta}$. That is, the ith element of $\boldsymbol{\upsilon} = \boldsymbol{\theta} \circ \boldsymbol{\theta} + \boldsymbol{\eta}_\upsilon$ is $\upsilon_i = \theta_i^2 + \eta_{\upsilon i}$. It is easy to verify that the elements of $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_4]^T$ and those of $\boldsymbol{\theta} \circ \boldsymbol{\theta} = [\theta_1^2, \ldots, \theta_4^2]^T$ are mutually uncorrelated, as the elements of the cross-covariance matrix between $\boldsymbol{\theta}$ and $\boldsymbol{\theta} \circ \boldsymbol{\theta}$ are the third-order moments of $\boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{I})$, which are all zero. Consequently, the elements of $\boldsymbol{\theta}$ and $\boldsymbol{\upsilon}$ are mutually uncorrelated, and in fact independent.

7.4.1 Effect of Sample Support

Table 7.1 lists the theoretical canonical correlations between $\boldsymbol{\theta}$ and $\boldsymbol{\upsilon}$, and the empirical canonical correlations of the sample data matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Upsilon}$ for $M = 100$, 50, 25, 12, and 6. For the first four values of M, the sample data matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Upsilon}$ are sample-rich, whereas for $M = 6$, $\boldsymbol{\Theta}$ and $\boldsymbol{\Upsilon}$ are sample-poor. The coherence between $\boldsymbol{\theta}$ and $\boldsymbol{\upsilon}$, measured by the first two canonical correlations, i.e. $H = 1 - (1 - \sigma_1^2)(1 - \sigma_2^2)$, are also listed in this table for both theoretical and empirical cases. For $M = 100$, the empirical canonical correlations are small and coherence is near zero, showing that

Table 7.1: Effect of sample support on empirical canonical correlations and coherence.

	Theoretical	$M = 100$	$M = 50$	$M = 25$	$M = 12$	$M = 6$
σ_1	0	0.2727	0.5706	0.7247	0.9118	1
σ_2	0	0.1575	0.3852	0.5493	0.7839	1
σ_3	0	0.1305	0.2189	0.3563	0.5383	0.8293
σ_4	0	0.0492	0.0674	0.0933	0.1892	0.3414
H	0	0.0973	0.4257	0.6684	0.9350	1

elements of $\boldsymbol{\theta}$ and $\boldsymbol{\upsilon}$ are mutually uncorrelated. However, as the number of samples M decreases the empirical canonical correlations and coherence increase, making the false impression that $\boldsymbol{\theta}$ and $\boldsymbol{\upsilon}$ are coherent. This shows that as the sample support becomes smaller the empirical canonical correlations carry less and less information about the theoretical ones, and finally when $\boldsymbol{\Theta}$ and $\boldsymbol{\Upsilon}$ become sample-poor ($M = 6$) the empirical canonical correlations cease to carry any information about the theoretical canonical correlations.

Figure 7.3 shows the concentration ellipses of the error in estimating the first two surrogate rows of $\mathbf{U}$ from the first two surrogate rows of $\mathbf{V}$. The theoretical concentration ellipse of the error in estimating the first two canonical coordinates of $\mathbf{x}$, i.e. u_1 and u_2, from the first two canonical coordinates of $\mathbf{y}$, i.e. v_1 and v_2, is also plotted. This concentration ellipse has its largest possible volume, i.e. one, as $\mathbf{u}$ and $\mathbf{v}$ are mutually uncorrelated. Among the empirical cases, the case with $M = 100$ has the closest concentration ellipse to the theoretical one. As the number of samples M decreases the concentration ellipses become smaller, misleading us to think that v_1 and v_2 carry a lot of information about u_1 and u_2. For the case where $M = 6$ and $\boldsymbol{\Theta}$ and $\boldsymbol{\Upsilon}$ are sample-poor, the volume of concentration ellipse is zero, suggesting that u_1 and u_2 can be perfectly estimated from v_1 and v_2. In this case the sample support is so poor that the empirical canonical correlations carry absolutely no information about the theoretical ones.

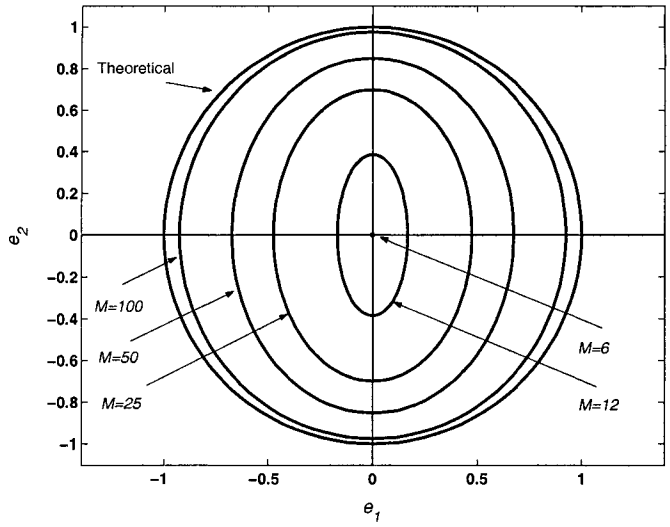

Figure 7.3: Concentration ellipses in canonical coordinates for various sample support sizes.

7.4.2 Pre-Processing Two-Channel Data with Nonlinear Mappings

We now clarify whether or not nonlinearities $\phi : \mathbb{R}^4 \longrightarrow \mathbb{R}^m$ and $\psi : \mathbb{R}^4 \longrightarrow \mathbb{R}^n$ may be chosen such that the empirical canonical correlations of the mapped sample data matrices

$$\begin{aligned} \mathbf{X} &= [\mathbf{x}_1, \ldots, \mathbf{x}_M] = [\phi(\boldsymbol{\theta}_1), \ldots, \phi(\boldsymbol{\theta}_M)] \\ \mathbf{Y} &= [\mathbf{y}_1, \ldots, \mathbf{y}_M] = [\psi(\boldsymbol{\upsilon}_1), \ldots, \psi(\boldsymbol{\upsilon}_M)] \end{aligned} \tag{7.31}$$

reveal coherence between high-order attributes of the elements of $\boldsymbol{\theta}$ and $\boldsymbol{\upsilon}$. In our experiments the mapped sample data matrices $\mathbf{X}$ and $\mathbf{Y}$ will be sample-rich. Based on the results of the previous simulation, the number of samples is chosen to be $M = 100$.

Roughly speaking, if $\phi(\cdot)$ and $\psi(\cdot)$ are chosen to be polynomial type nonlinearities, the elements of the cross-covariance matrix $\mathbf{S}_{xy}$ become estimates of high-order moments of the Gaussian distribution $N(\mathbf{0}, \mathbf{I})$. This gives us some intuition for selecting $\phi(\cdot)$ and $\psi(\cdot)$ to obtain large canonical correlations after the mappings. However, we note that the simulation examples presented in this section are for the

sole purpose of demonstrating that pre-processing two-channel data with nonlinearities prior to canonical correlation analysis can reveal coherence between high-order attributes of the original channels. These simulations are not intended to assess the extent of suitability of different nonlinearities for canonical correlation analysis. Systematic selection of nonlinearities for canonical correlation analysis is an open research area.

Example 1: Looking at the two-channel model in (7.1) for $\boldsymbol{\theta}$ and $\boldsymbol{\upsilon}$, an obvious choice is to select $\boldsymbol{\phi} : \mathbb{R}^4 \longrightarrow \mathbb{R}^4$ and $\boldsymbol{\psi} : \mathbb{R}^4 \longrightarrow \mathbb{R}^4$ so that

$$\begin{aligned} \mathbf{x}_i &= \boldsymbol{\phi}(\boldsymbol{\theta}_i) = [\theta_{i1}^2, \ldots, \theta_{i4}^2]^T \\ \mathbf{y}_i &= \boldsymbol{\psi}(\boldsymbol{\upsilon}_i) = [\upsilon_{i1}, \ldots, \upsilon_{i4}]^T \end{aligned} \tag{7.32}$$

with θ_{ij} being the jth element of $\boldsymbol{\theta}_i$, and υ_{ij} being the jth element of $\boldsymbol{\upsilon}_i$. With this choice of $\boldsymbol{\phi}(\cdot)$ and $\boldsymbol{\psi}(\cdot)$, the mapped sample data matrices $\mathbf{X} \in \mathbb{R}^{4\times 100}$ and $\mathbf{Y} \in \mathbb{R}^{4\times 100}$ are full-rank and sample-rich. In this case, the dimension of the data vectors before and after the mapping are the same.

Table 7.2 lists the empirical canonical correlations of the sample data matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Upsilon}$, and those of the mapped sample data matrices $\mathbf{X}$ and $\mathbf{Y}$. The empirical canonical correlations of the sample data matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Upsilon}$ are very small, showing that the elements of $\boldsymbol{\theta}$ are mutually uncorrelated with elements of $\boldsymbol{\upsilon}$. However, after pre-processing these rows with nonlinear mappings $\boldsymbol{\phi}(\cdot)$ and $\boldsymbol{\psi}(\cdot)$, the empirical canonical correlations of the mapped sample data matrices $\mathbf{X}$ and $\mathbf{Y}$ become very close to one, showing that nonlinear functions of $\boldsymbol{\theta}$ and $\boldsymbol{\upsilon}$ are nearly co-linear.

The coherence between $\boldsymbol{\theta}$ and $\boldsymbol{\upsilon}$, estimated by the first two empirical canonical correlations, is $H = 1 - (1 - \sigma_1^2)(1 - \sigma_2^2) = 0.0973$, whereas the coherence between $\mathbf{x} = \boldsymbol{\phi}(\boldsymbol{\theta})$ and $\mathbf{y} = \boldsymbol{\psi}(\boldsymbol{\upsilon})$ is $H = 0.9972$. Thus, before the nonlinear mappings the coherence is near zero, whereas after the nonlinear mappings the coherence is near one. This result shows that pre-processing the data channels with nonlinear functions

Table 7.2: Empirical canonical correlations between Θ and Υ, and $\mathbf{X}$ and $\mathbf{Y}$ in Example 1.

	Linear	Nonlinear
σ_1	0.2727	0.9758
σ_2	0.1575	0.9702
σ_3	0.1305	0.9589
σ_4	0.0492	0.9554

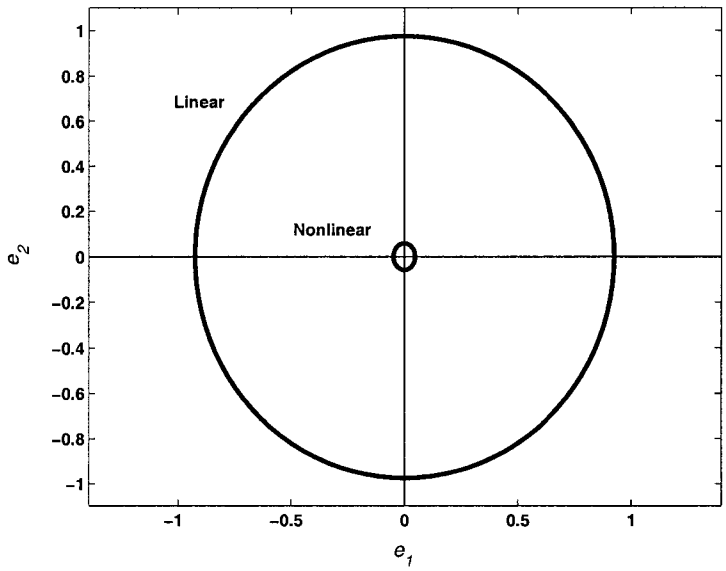

Figure 7.4: Concentration ellipses in canonical coordinates for Example 1.

prior to canonical correlation analysis may indeed reveal coherence between high-order attributes of the channels, even when the original data channels are mutually non-coherent.

Figure 7.4 shows the concentration ellipse of the error $\mathbf{e} = [e_1 \ e_2]^T$ in estimating the first two surrogate rows of $\mathbf{U}$ from the first two surrogate rows of $\mathbf{V}$, before and after applying the nonlinear mappings to Θ and Υ. The considerably smaller concentration ellipse in the nonlinear case suggests that the rate at which the surrogate rows of $\mathbf{V}$ carry information about the surrogate rows of $\mathbf{U}$ has increased significantly, as a result of applying nonlinear mappings $\phi(\cdot)$ and $\psi(\cdot)$.

Table 7.3: Empirical canonical correlations between Θ and Υ, and $\mathbf{X}$ and $\mathbf{Y}$ in Example 2.

	Linear	Nonlinear
σ_1	0.2727	0.8270
σ_2	0.1575	0.7641
σ_3	0.1305	0.4763
σ_4	0.0492	–

Example 2: In the second example, the data samples $\boldsymbol{\theta}_i = [\theta_{i1}, \ldots, \theta_{i4}]^T$ and $\boldsymbol{\upsilon}_i = [\upsilon_{i1}, \ldots, \upsilon_{i4}]^T$ are transformed to a lower dimensional space, using $\boldsymbol{\phi} : \mathbb{R}^4 \longrightarrow \mathbb{R}^3$ and $\boldsymbol{\psi} : \mathbb{R}^4 \longrightarrow \mathbb{R}^3$, where

$$\begin{aligned} \mathbf{x}_i &= \boldsymbol{\phi}(\boldsymbol{\theta}_i) = [\theta_{i1}^2\theta_{i2}^2, \theta_{i2}^2\theta_{i3}^2, \theta_{i3}^2\theta_{i4}^2]^T \\ \mathbf{y}_i &= \boldsymbol{\psi}(\boldsymbol{\upsilon}_i) = [\upsilon_{i1}^2\upsilon_{i2}^2, \upsilon_{i2}^2\upsilon_{i3}^2, \upsilon_{i3}^2\upsilon_{i4}^2]^T. \end{aligned} \tag{7.33}$$

For these choices of $\boldsymbol{\phi}(\cdot)$ and $\boldsymbol{\psi}(\cdot)$, the mapped sample data matrices $\mathbf{X} \in \mathbb{R}^{3\times 100}$ and $\mathbf{Y} \in \mathbb{R}^{3\times 100}$ are full-rank and sample-rich.

Table 7.3 lists the empirical canonical correlations of the sample data matrices Θ and Υ, and those of the mapped sample data matrices $\mathbf{X}$ and $\mathbf{Y}$. In this case, the estimated coherence between $\mathbf{x} = \boldsymbol{\phi}(\boldsymbol{\theta})$ and $\mathbf{y} = \boldsymbol{\psi}(\boldsymbol{\upsilon})$ is $H = 0.8696$. The empirical canonical correlations and the coherence between the surrogate rows of $\mathbf{X}$ and $\mathbf{Y}$ are considerably larger than those between the surrogate rows of Θ and Υ. Again, this implies that pre-processing the data samples with nonlinear functions prior to canonical correlation analysis may reveal coherence between high-order attributes of the channels.

Figure 7.5 shows the concentration ellipse of the error $e = [e_1 \;\; e_2]^T$ in estimating the first two surrogate rows of $\mathbf{U}$ from the first two surrogate rows of $\mathbf{V}$, before and after applying nonlinear mappings to Θ and Υ. Again, the considerably smaller concentration ellipse in the nonlinear case indicates that the rate at which the surrogate rows of $\mathbf{V}$ carry information about the surrogate rows of $\mathbf{U}$ has increased significantly, after applying nonlinear mappings $\boldsymbol{\phi}(\cdot)$ and $\boldsymbol{\psi}(\cdot)$.

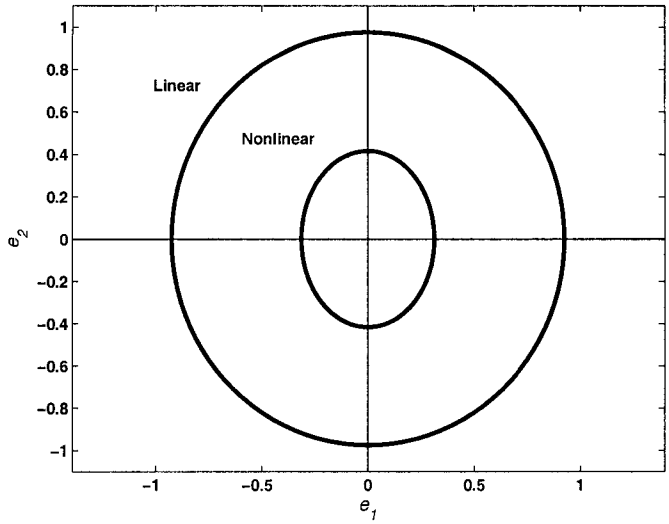

Figure 7.5: Concentration ellipses in canonical coordinates for Example 2.

Example 3: In the last example, the nonlinearities $\boldsymbol{\phi}(\,\cdot\,)$ and $\boldsymbol{\psi}(\,\cdot\,)$ are chosen so that they transform the data samples $\boldsymbol{\theta}_i$ and $\boldsymbol{\upsilon}_i$ to

$$\begin{aligned} \mathbf{x}_i &= \boldsymbol{\phi}(\boldsymbol{\theta}_i) = [\theta_{ij}\theta_{ik}]_{j,k=1}^{4};\ \ j \leq k \\ \mathbf{y}_i &= \boldsymbol{\psi}(\boldsymbol{\upsilon}_i) = [\upsilon_{ij}\upsilon_{ik}]_{j,k=1}^{4};\ \ j \leq k. \end{aligned} \tag{7.34}$$

The elements of $\boldsymbol{\phi}(\theta_i) \in \mathbb{R}^{10}$ and $\boldsymbol{\psi}(\upsilon_i) \in \mathbb{R}^{10}$ are the elements in the upper triangle of the outer products $\boldsymbol{\theta}_i\boldsymbol{\theta}_i^T$ and $\boldsymbol{\upsilon}_i\boldsymbol{\upsilon}_i^T$, respectively. That is, data samples $\boldsymbol{\theta}_i$ and $\boldsymbol{\upsilon}_i$ are replaced by the rank-one correlation matrices $\boldsymbol{\theta}_i\boldsymbol{\theta}_i^T$ and $\boldsymbol{\upsilon}_i\boldsymbol{\upsilon}_i^T$, and only unique elements of these matrices are retained. Here the nonlinearities transform the data samples to a higher dimensional space, but the mapped sample data matrices $\mathbf{X}$ and $\mathbf{Y}$ remain sample-rich.

For the choice of nonlinearities in (7.34), the inner products $\mathbf{x}_i^T\mathbf{x}_j = \boldsymbol{\phi}(\boldsymbol{\theta}_i)^T\boldsymbol{\phi}(\boldsymbol{\theta}_j)$ and $\mathbf{y}_i^T\mathbf{y}_j = \boldsymbol{\psi}(\boldsymbol{\upsilon}_i)^T\boldsymbol{\psi}(\boldsymbol{\upsilon}_j)$ may be written as second-order homogenous kernel functions [35]. That is,

$$\begin{aligned} \mathbf{x}_i^T\mathbf{x}_j &= \boldsymbol{\phi}(\boldsymbol{\theta}_i)^T\boldsymbol{\phi}(\boldsymbol{\theta}_j) = (\boldsymbol{\theta}_i^T\boldsymbol{\theta}_j)^2 = k_x(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \\ \mathbf{y}_i^T\mathbf{y}_j &= \boldsymbol{\psi}(\boldsymbol{\upsilon}_i)^T\boldsymbol{\psi}(\boldsymbol{\upsilon}_j) = (\boldsymbol{\upsilon}_i^T\boldsymbol{\upsilon}_j)^2 = k_y(\boldsymbol{\upsilon}_i, \boldsymbol{\upsilon}_j). \end{aligned} \tag{7.35}$$

Table 7.4: Empirical canonical correlations between Θ and Υ, and $\mathbf{X}$ and $\mathbf{Y}$ in Example 3.

	Linear	Nonlinear (Direct)	Nonlinear (Kernel)
σ_1	0.2727	0.9735	0.9735
σ_2	0.1575	0.9582	0.9582
σ_3	0.1305	0.9500	0.9500
σ_4	0.0492	0.9116	0.9116
σ_5	–	0.6000	0.6000
σ_6	–	0.4264	0.4264
σ_7	–	0.2989	0.2989
σ_8	–	0.1647	0.1647
σ_9	–	0.0937	0.0937
σ_{10}	–	0.0468	0.0468

This example demonstrates the application of kernel nonlinearities for canonical correlation analysis. However, since $M > n \geq m$ the kernel formulation is not computationally advantageous with respect to the direct formulation.

Table 7.4 lists the empirical canonical correlations of the sample data matrices Θ and Υ, and those of the mapped sample data matrices $\mathbf{X}$ and $\mathbf{Y}$, obtained using the direct and kernel formulations. Similar to the previous examples, as a result of nonlinear mappings, the empirical canonical correlations reveal high-order coherence between the original channels. In addition, it is seen that the empirical canonical correlations of $\mathbf{X}$ and $\mathbf{Y}$, computed from the kernel formulation are identical with those computed from the direct formulation.

In this case, the estimated coherence between $\boldsymbol{\theta}$ and $\boldsymbol{\upsilon}$ is $H = 0.9957$. The concentration ellipses of the error $\mathbf{e} = [e_1 \ e_2]^T$ in estimating the first two surrogate rows of $\mathbf{U}$ from the first two surrogate rows of $\mathbf{V}$, before and after applying the nonlinear mappings to Θ and Υ are shown in Figure 7.6. Again, the results indicate that preprocessing the data samples with nonlinear functions prior to canonical correlation analysis may reveal coherence between high-order attributes of the channels.

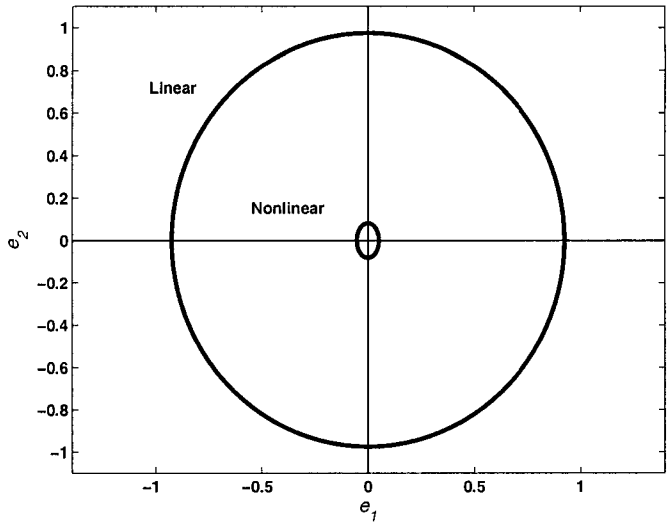

Figure 7.6: Concentration ellipses in canonical coordinates for Example 3.

7.5 Conclusions

In this chapter, we have studied the canonical coordinate decomposition of two-channel data, when the channel covariances are estimated from a limited number of samples. Depending on the number of samples drawn from each channel, and the ranks of the sample data matrices, two different cases emerge: the sample-poor case, in which the number of data samples is smaller than the sum of the ranks of the data matrices, and a sample-rich case, in which the number of data samples is greater than the sum of the ranks of the data matrices. This chapter shows that, in either case, it is the rows of the sample data matrices that determine the empirical canonical correlations, and that the empirical canonical correlations measure the cosines of the principal angles between the row spaces of the two data matrices. Further, we have shown that the empirical canonical correlations form a maximal set of invariants for the composite sample covariance matrix of two-channel data.

We have established that in a sample-poor case some of the empirical canonical correlations or principal cosines are always equal to one, regardless of the two-channel model that generates the data samples. This result has also been reported in [45],

without a rigorous analysis. Therefore, the empirical canonical correlations are defective and may not be used as estimates of canonical correlations between random variables. Geometrically, this means that principal angles between linear subspaces of Euclidean space can not be used as estimates of principal angles between corresponding linear subspaces of the Hilbert space of second-order random variables. In a sample-rich case, however, the empirical canonical correlations do estimate the canonical correlations and principal cosines between the random variables that generate the samples, and hence may be used for estimating coherence between two data channels.

These results imply that sample data matrices of nonlinearly mapped data must remain sample–rich for empirical canonical correlations to estimate coherence between high-order attributes of the original channels. Three simulation examples have demonstrated that empirical canonical correlations extracted from nonlinearly mapped, sample–rich, data samples can estimate coherence between high-order attributes of the original channels. Additionally, we argued that in cases where the kernel formulation of canonical correlation analysis is computationally advantageous with respect to the direct formulation, the empirical canonical correlations between two data matrices do not usefully estimate coherence between the corresponding data channels. Therefore, this computational advantage is superficial and does not have any practical value.

CHAPTER 8

CANONICAL CORRELATIONS FOR CLASSIFICATION OF UNDERWATER TARGETS

8.1 Introduction

The problem of classifying underwater targets using active sonar has attracted a lot of attention in recent years [52], [71]– [85]. This problem involves discrimination between targets and non-targets, as well as characterization of background clutter. Some of the factors that complicate this process include: non-repeatability and variation of target signatures with aspect angle, range, and grazing angle; diverse sizes, shapes, and scattering properties of the targets; presence of natural and man-made clutter; and a highly variable and reverberant operating environment. The problem is even more complicated when bottom targets are encountered, especially if they are buried or obscured by bottom features.

Due to the above factors, it is often difficult to detect and classify objects of interest (target/non-target) based on the measurement from a single object-sensor orientation (aspect/view angle). There are often orientations at which different objects may look nearly identical. Consequently, in real-life situations, the decision about the presence and type of an object is generally made based upon observations of the received

signals at several aspect angles. This is due to the fact that multi-aspect classification typically provides better resolution and sensing of the 3-D properties of the object, in the changing environment.

In recent years, several different multi-aspect classification methods for detection and classification of underwater targets from acoustic backscattered signals have been developed [52], [71]– [81]. A good review of these methods is provided in [52]. However, in all these methods, multi-aspect classification is performed through one of the following classification fusion methods: *decision-level* fusion [52], *feature-level* fusion [52], or a combination of decision-level and feature-level fusion [52]. In the decision-level fusion, after a single sonar return at a particular aspect angle is observed, a *preliminary* decision about the presence and type of the object (target or non-target) is made. The final decision is made at the fusion center, typically a neural network, based upon multiple of these single-aspect preliminary decisions. In the feature-level fusion, on the other hand, a feature vector is extracted from every single sonar return. Then, multiple of these feature vectors are simultaneously applied to a decision making system, e.g. a detector or classifier, to determine the final decision.

In this chapter (see also [49] and [50]), we take a different approach to multi-aspect underwater target classification. In this approach multi-aspect classification is performed by extracting features that capture common target/non-target attributes among two sonar returns. In other words, multi-aspect classification is performed via multi-aspect feature extraction. This is accomplished by exploiting the linear dependence, or coherence, between two consecutive sonar returns, with certain aspect separation. The idea here is that linear dependence between the sonar returns is an indication of the presence of a common signature, whereas linear independence indicates the absence of a common signature. This is the basic idea behind multi-channel tests for linear dependence [3] and the multiple coherence test of [47], [48].

In Chapter 2, we established that the linear dependence between two data channels is measured by the canonical correlations of the channels. This implies that canonical correlations can be viewed as features that capture linear dependence or coherence between the two data channels, and hence may be used for detection or classification. We intend to exploit this idea for classifying underwater mine-like objects (targets) from non-mine-like objects (non-targets). In this approach, the channels correspond to acoustic backscattered signals at two consecutive aspect angles, with certain aspect/ping separation.

Using canonical correlations, we exploit the linear dependence between two backscattered signals or sonar returns to determine whether common signatures associated with targets or non-targets are present. We hypothesize that the amount of coherence between the two sonar returns generated by the presence of a mine-like object is different from that caused by the presence of a non-mine-like object. Therefore, the dominant canonical correlations, which capture most of the coherence between the two sonar returns, may be used to classify the objects at the corresponding aspect angles. We test our hypothesis on a subset of a wideband data set that was collected at the Applied Research Lab (ARL), University of Texas (UT)-Austin, and benchmark our results against those obtained in [52] on the same data set.

Additionally, we shall investigate the potential use of nonlinearly mapped two-channel data, followed by canonical correlation analysis, with the aim to capture coherence between high-order attributes of the two sonar returns. We use several nonlinearities to map the data samples extracted from the sonar returns in order to investigate whether or not the canonical correlations between the nonlinear functions of the backscattered signals can improve the discrimination of mine-like objects from non-mine-like objects. Our results show that not only the canonical correlation features extracted from the nonlinearly mapped sonar returns do not improve the discrimination between targets and non-targets, they impair it compared to the

canonical correlation features extracted from the original (linear) sonar returns.

8.2 Wideband Sonar Data Set

The sonar data set used in this study is a subset of a wideband acoustic backscattered data set collected at the ARL-UT, Lake Travis test facility [51], [52]. This subset contains acoustic backscattered signals from three mine-like and three non-mine-like objects, resting on a solid interface, in two different environmental conditions, namely smooth and rough bottom. For the rough bottom condition the sand was raked, giving it rippled effects. The mine-like objects are two cylindrical steel objects (Targets 6 and 7), and a truncated cone shape plastic object (Target 2). The non-mine-like objects are a water-filled steel drum, a concrete pipe, and a telephone pole.

Figure 8.1 shows the experimental setup used for collecting the ARL-UT data set [51]. During the data collection, the objects were placed on a rotating seabed, 25-30 ft below the lake surface, with minimal embedding/scouring. The diameter of the seabed was 25 ft. The center of the object was positioned as near to the center of the circular platter (seabed) as possible. Certain straps and parts of the supporting barge were also in the water in the vicinity of the seabed. These parts can cause returns separate from the main return from the object. Attempt was made in the experimental setup to baffle these artifacts, using either steel or plastic baffles, coated with 5/8 inch construction styrofoam.

An acoustic panel was set at depression/elevation (D/E) angle of 13.5 degrees, and range of 105 ft from the center of the seabed. The transmit signal was a linear frequency modulated waveform with a bandwidth of 85 kHz in the range of 15-100 kHz, and was approximately 7 msec long. While the seabed was rotated for 360 degrees, acoustic backscattered signals were collected at nearly uniform separations of 0.2 degree, using a receiver array of 64 elements. The backscattered signals were recorded for approximately 16 msec at 500 kHz sampling rate, resulting in 8192 samples.

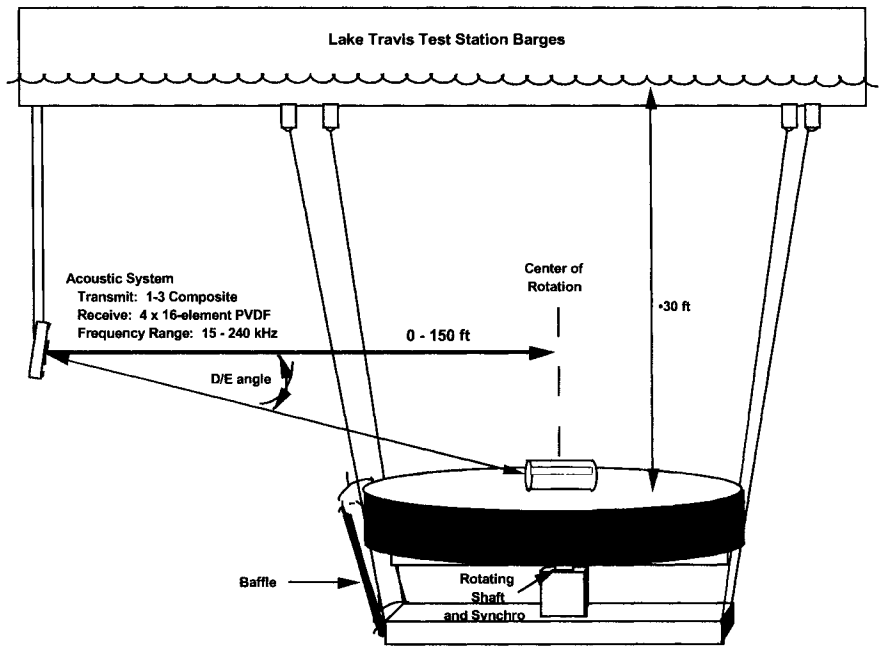

Figure 8.1: ARL-UT experimental setup for bottom target/non-target data collection.

Figure 8.2 shows the acoustic panel, used in collection of the ARL-UT data set [51]. The transmit array is divided into several separate sections, providing various sub-apertures, in order to allow for total insonification of mine-like and non-mine-like objects at any particular frequency between 15 kHz and 100 kHz. The receiver array consists of four horizontal arrays, each with 16 elements (channels). Each receiver element is 2 inches tall and 1 inch wide. Preamps mounted behind the receiving elements allow the use of these elements between 1 kHz and 600 kHz. During the data collection, several different (receiver) array configurations were used: (a) a 16 channel configuration, consisting of all 16 channels of array A, (b) a 16 (4×4) channel configuration, consisting of channels 7 to 10 (counting from the left) of all the arrays (A to D), and (c) a 32 channel configuration, consisting of all 16 channels of array A, channels 6 to 11 of array B, and channels 6 to 10 of arrays C and D. Since in all of these configurations channels 7 to 10 of array A are common, in our experiments, we use the averaged data of these four channels. This averaging results in a beam that

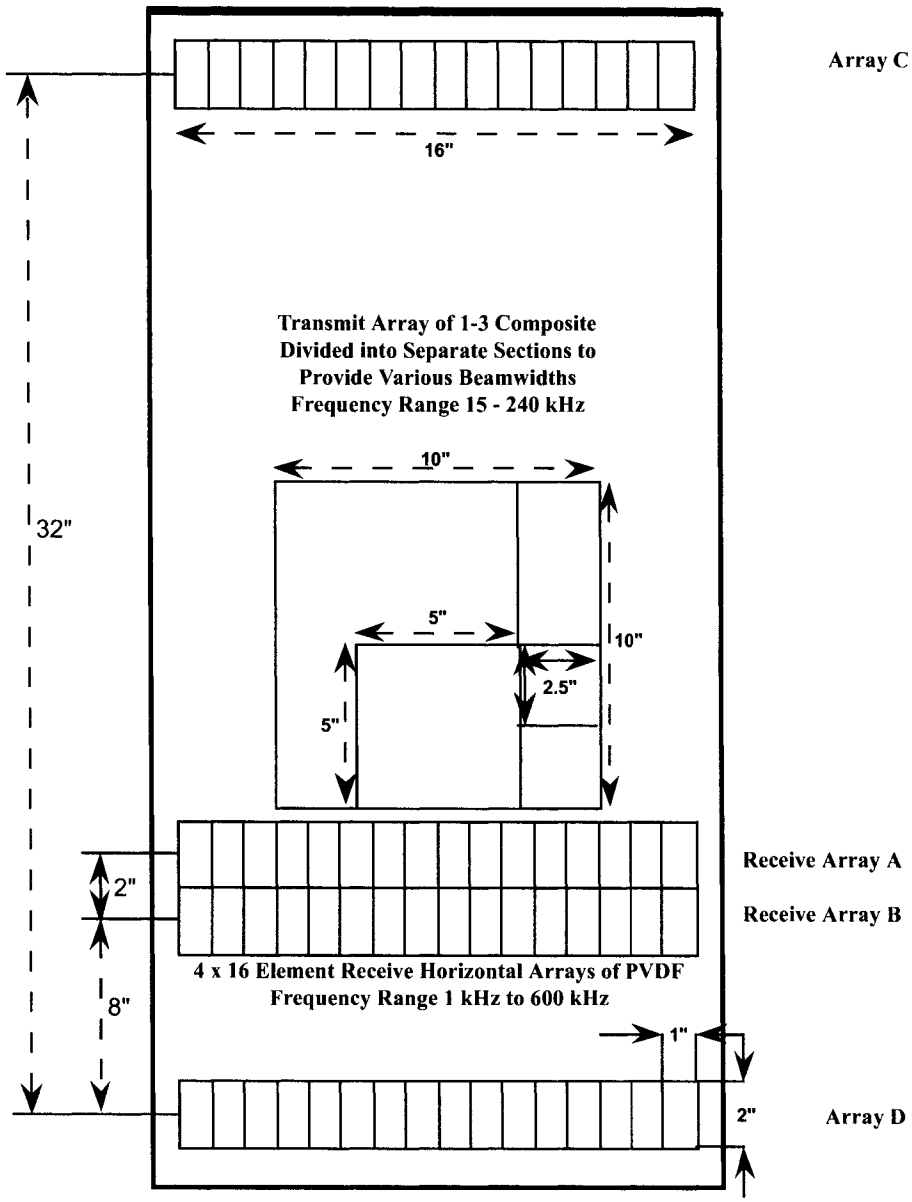

Figure 8.2: Acoustic panel for wideband data acquisition.

mainly insonifies the objects and not much of their surroundings. At 100 kHz the beamwidth is approximately 7.6 degrees, resulting in a coverage width of 14 ft. This coverage is wide enough to cover the entire length of the objects and narrow enough to avoid the side edges of the seabed. Finally, we note that although the backscattered signals were recorded at every 0.2 degrees, only the backscattered signals at every 1 degree are used in our experiments.

8.3 Feature Extraction Process

In this section, we describe how canonical correlation analysis may be used to extract a set of features that capture common target/non-target attributes among two consecutive sonar returns, with certain aspect separation. Later, in Section 8.4, we will apply the feature extraction method presented here to the wideband ARL-UT data set, and use the extracted features for classifying mine-like objects from non-mine-like objects.

To build the ensembles of the two channels ($\mathbf{x}$ and $\mathbf{y}$) for canonical correlation analysis, we partition two backscattered signals from an object, with certain aspect separation, into overlapping blocks, as illustrated in Figure 8.3. In this figure, sonar return 1 is the backscattered signal from an object at aspect angle, say β_1, and sonar return 2 is the backscattered signal from the same object at aspect angle $\beta_2 = \beta_1 + \Delta\beta$, where $\Delta\beta$ is the aspect separation between these two returns. The blocks of sonar return 1 are taken as the samples of the first channel (the $\mathbf{x}$-channel) and the blocks of sonar return 2 are taken as the samples of the second channel (the $\mathbf{y}$-channel). That is, referring to Figure 8.3, the data sample $\mathbf{x}_i$ is the vector of the time series associated with the ith block of range cells in sonar return 1, and $\mathbf{y}_i$ is the vector of the time series associated with the ith block of range cells in sonar return 2. The collections of these data vector samples form the sample data matrices $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M]$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_M]$ for canonical correlation analysis. The dominant canonical correlations between these two sample data matrices, which capture most of the coherence, will be used as features to represent the backscattered signal at the aspect angle β_1, i.e. the first sonar return in the pair. The alternating power methods developed in Chapter 5 may be used here for extracting the dominant canonical correlations.

We note that the aspect separation $\Delta\beta$ should be large enough so that the reverberation effects are almost uncorrelated, but not too large so that the returns from the objects remain coherent. For the ARL-UT data set, we have experimentally

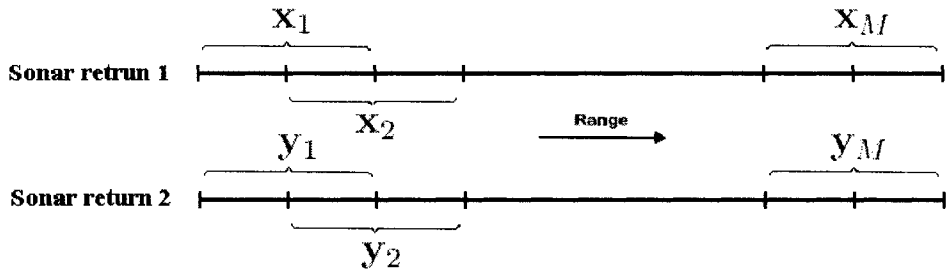

Figure 8.3: Building the ensembles of the two channels ($\mathbf{x}$ and $\mathbf{y}$) for canonical correlation analysis from two consecutive sonar returns.

determined that an aspect separation of $\Delta\beta = 16$ degrees is a reasonable choice. Therefore, in the simulations preformed in Section 8.4, the sonar returns from an object in the ARL-UT data set are paired according to aspect angles as follows: $\{1,17\}$, $\{2,18\}\cdots,\{344,360\},\{345,1\},\cdots,\{360,16\}$. Considering the pair $\{2,18\}$ as an example, sonar return 1 in Figure 8.3 corresponds to the aspect angle $\beta_1 = 2°$, while sonar return 2 corresponds to the aspect angle $\beta_2 = 18°$. Consequently, the dominant canonical correlations extracted from the pair $\{2,18\}$ will be used as features at the aspect angle $\beta_1 = 2°$.

In situations where linear dependence between the sonar returns is not adequate to discriminate targets from non-targets, the data samples obtained from a pair of sonar returns, i.e. $\mathbf{x}_i$'s and $\mathbf{y}_i$'s, may be mapped using nonlinear functions $\boldsymbol{\phi}(\cdot)$ and $\boldsymbol{\psi}(\cdot)$, prior to canonical correlation analysis. The dominant canonical correlations extracted from the nonlinearly mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ may then be used as features to represent the first backscattered signal in the corresponding pair, as in the linear case.

8.4 Canonical Correlation Features and Classification Results

In this section, the feature extraction process described in Section 8.3 is applied to the sonar returns in the ARL-UT data set, and the extracted features are used to classify

mine-like objects from non-mine-like objects. In the experiments performed here, the backscattered signals are partitioned into blocks of size 50 samples, with 50% (25 samples) overlap. This value for block size has been determined experimentally.

8.4.1 Original Two-Channel Sonar Data

Here, the canonical correlation analysis is directly performed between the original sample data matrices $\mathbf{X}$ and $\mathbf{Y}$, formed by range partitioning in two sonar returns with 16 degrees aspect separation. Therefore, each data vector sample for $\mathbf{x}$ and $\mathbf{y}$ channels is 50-dimensional. We extract the first 15 out 50 canonical correlations of the two sample data matrices and use them as features to represent the first aspect angle in the pair. These features capture most of the coherence between the sonar returns. The alternating block power method, developed in Chapter 5, is used here for extracting the first 15 canonical correlations of $\mathbf{X}$ and $\mathbf{Y}$.

Experiment 1: The objective here is to demonstrate the usefulness of canonical correlation features for classifying targets from non-targets in both smooth and rough bottom conditions. The training data set for classification is formed from the feature vectors extracted at 90 aspect angles of the *smooth bottom* data, at aspect increments of 4°. The feature vectors extracted from the rest of the aspect angles of the smooth bottom data (270 aspect angles) are kept to validate the trained classifier. We refer to this set as the validation data set. This is the set that is used to select the best trained classifier. To see how well the trained classifier generalizes, the features extracted from the backscattered signals in the *rough bottom* condition are used as the testing data set.

Figures 8.4 and 8.5 show the scatter plots of the first four features (the first four canonical correlations) for the training, validation, and testing data sets. As can be seen, for the training data set (Figures 8.4(a) and 8.5(a)), and the validation data set (Figures 8.4(b) and 8.5(b)), the features of mine-like objects (Targets 2, 6, and 7) are packed together and almost completely separated from those of the non-mine-like

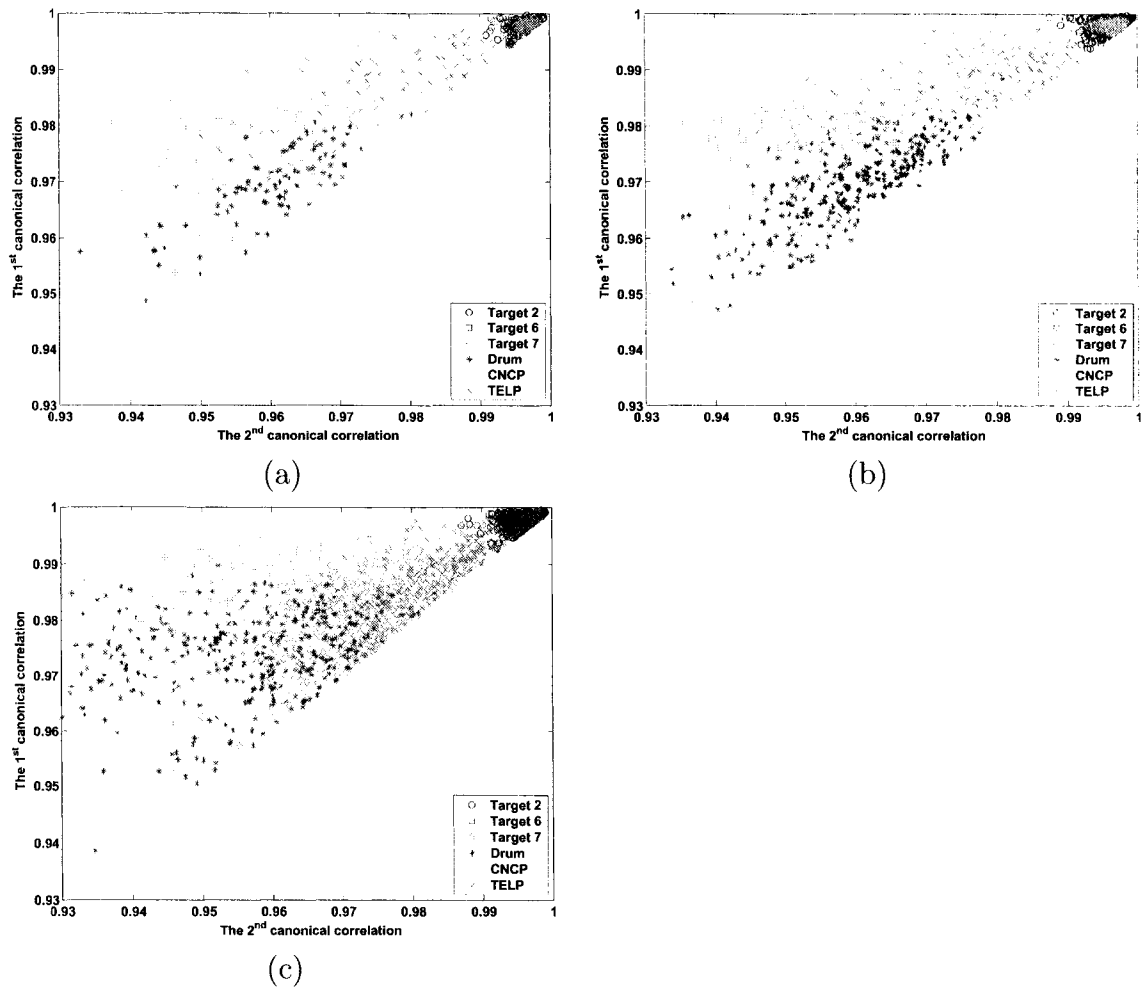

Figure 8.4: Scatter plots of the first two canonical correlation features for (a) training, (b) validation, and (c) testing data sets. The scatter plots show that, for five out of six objects canonical correlations are fairly robust (only slightly change) to the changes in the bottom condition.

objects (steel drum, concrete pipe, and telephone pole). Additionally, the extracted features for the training and validation data sets, for the objects in the smooth bottom condition, are consistent (occupy the same regions in the scatter plots).

In the rough bottom test condition (Figures 8.4(c) and 8.5(c)), features of Target 2, Target 6, and the telephone pole stay in the same regions as in the smooth bottom condition, while those of the drum and concrete pipe become more compact and move slightly towards the right side of the plot. Nonetheless, they still occupy almost the same regions as in training and validation data sets. Features of Target 7, however,

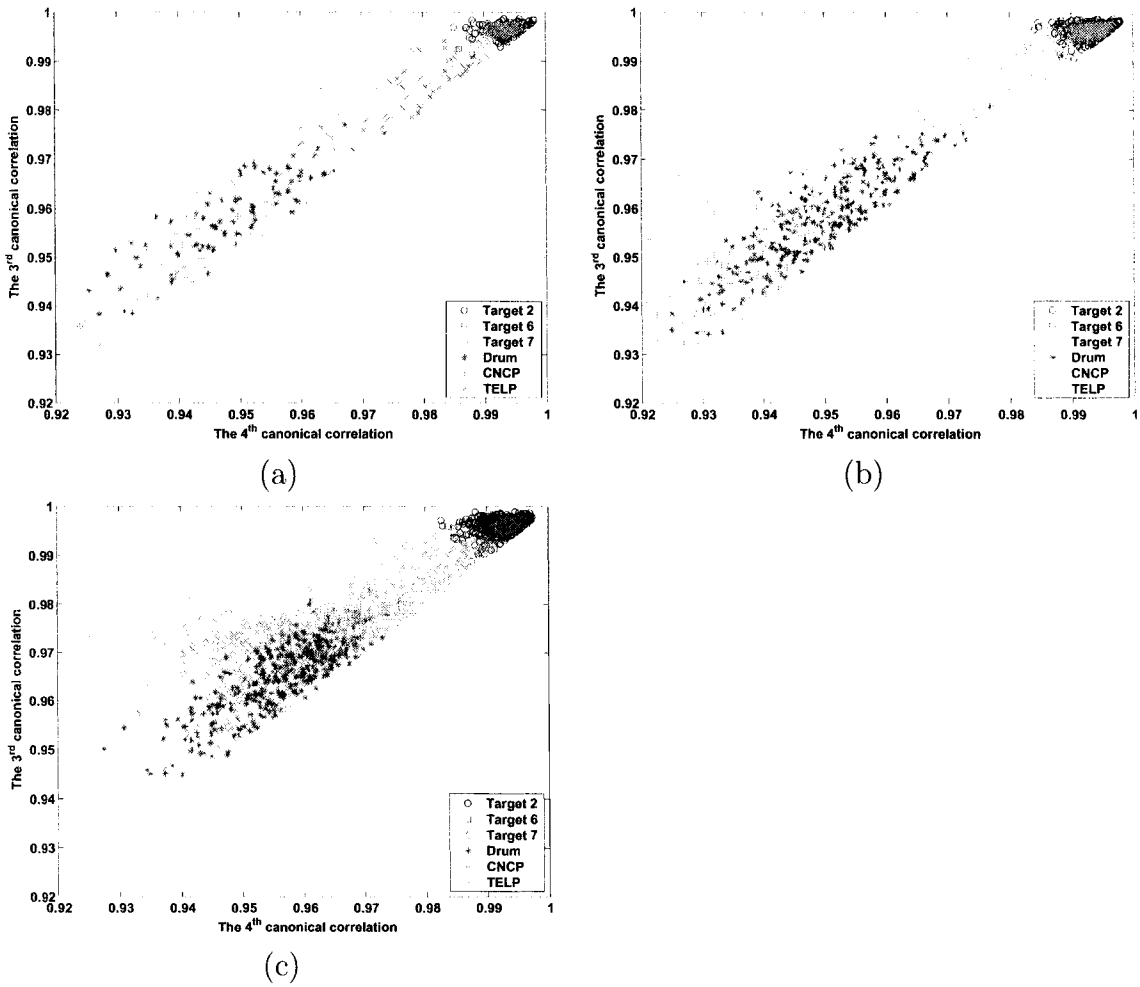

Figure 8.5: Scatter plots of the third and fourth canonical correlation features for (a) training, (b) validation, and (c) testing data sets. The scatter plots show that, for five out of six objects, canonical correlations are fairly robust (only slightly change) to the changes in the bottom condition.

Table 8.1: Classification rates obtained using canonical correlation features versus those of the LPC subband features.

Features	Training	Validation	Testing
Canonical correlation	99.1%	98.6%	81.0%
LPC subband	99.6%	82.5%	75.2%

move from the upper right corner and spread out to the left side and mix with those of the steel drum and concrete pipe. The reason for changes in features of Target 7 may be attributed to the secondary reflections between this rather large cylindrical target, rough sand, and edges of the rotating seabed. Clearly, as will be shown shortly, this leads to some degradation in classification performance in the rough bottom condition. These scatter plots clearly show that for five out of six objects, canonical correlations are fairly robust (only slightly change) to the changes in the bottom condition.

Subsequently, the extracted canonical correlation features are used to train a back-propagation neural network (BPNN) [86] to classify the mine-like objects (Targets 2, 6, and 7) from non-mine-like objects (steel drum, concrete pipe, and telephone pole). To find a good network structure, eight different two-layer BPNN structures were tried. The number of hidden layer neurons was varied from 26 to 46. Each network was trained for ten different weight initializations. The training was performed for 10000 epochs, where an epoch was a complete sweep over the entire training data set. The best BPNN classifier, which was selected based on the average classfication rates on training and validation data sets, gave a correct classification rate of 99.1% on the training data set, 98.6% on the validation data set, and 81.0% on the testing data set. These percentages are obtained based on a hard-limiting decision threshold. Table 8.1 benchmarks our classification results against those in [52], which uses linear predictive coding (LPC) subband features and decision-level fusion.

Table 8.2: Confusion matrices of the BPNN classifier trained with canonical correlation features.

	Validation Data Set		Testing Data Set	
Object	Target	Non-Target	Target	Non-Target
Target 6	270	0	360	0
Target 7	270	0	2	358
Target 2	270	0	360	0
Drum	0	270	0	360
Concrete Pipe	0	270	0	360
Telephone Pole	22	248	52	308

Table 8.3: Confusion matrices of the BPNN classifier trained with LPC subband features.

	Validation Data Set		Testing Data Set	
Object	Target	Non-Target	Target	Non-Target
Target 6	215	55	224	136
Target 7	177	93	214	146
Target 2	265	5	349	11
Drum	57	213	128	232
Concrete Pipe	56	214	92	268
Telephone Pole	17	253	22	338

The confusion matrices obtained for the classifiers, trained using these two feature types, are shown in Table 8.2 (for the canonical correlation features) and Table 8.3 (for the LPC subband features). The results clearly demonstrate the promise of the canonical correlation features for classifying targets from non-targets in different bottom conditions.

It is interesting to note that the classifier trained using canonical correlation features has correctly classified Targets 2 and 6 at all aspect angles in both smooth and rough bottom conditions, while the classifier trained using the LPC subband

features has poor performance on these targets. Additionally the canonical correlation features provide substantially lower false alarm[1] rates (2.7% for validation and 4.8% for testing) compared to the LPC subband features (16.1% for validation and 22.4% for testing). However, the classifier trained with the LPC subband features provides better performance for Target 7 in the rough bottom condition compared to the canonical correlation-based classifier.

Experiment 2: Our goal in this experiment is to investigate the robustness of the canonical correlation features with respect to aspect angle variation in a fixed bottom condition, namely the smooth bottom. The training data set for each object is formed from the feature vectors extracted for 1/4 of the aspect angles that correspond to one side of the objects (aspect angles 0 to 179 degrees) only. The feature vectors extracted from the rest of the aspect angles between 0 to 179 degrees in the smooth bottom condition are kept to validate the trained classifier. The generalization and robustness of the trained classifier is tested, in the same bottom condition, using the features extracted from sonar returns from the other side of the objects, at aspect angles 180 to 359 degrees. Clearly, in this experiment the classifier is not exposed to the information on the other side of the objects during the training and validation process.

Figures 8.6(a)-(c) show the scatter plots of the first two canonical correlation features for the training, validation, and testing data sets. As can be seen, the canonical correlation features for targets are similar and almost completely separated from those of the non-targets. Additionally, the extracted features for the training, validation, and testing data sets for the objects are fairly consistent, implying that the canonical correlation features are indeed robust with respect to aspect angle variation.

---

[1]False alarm is defined as misclassification of a non-target as a target.

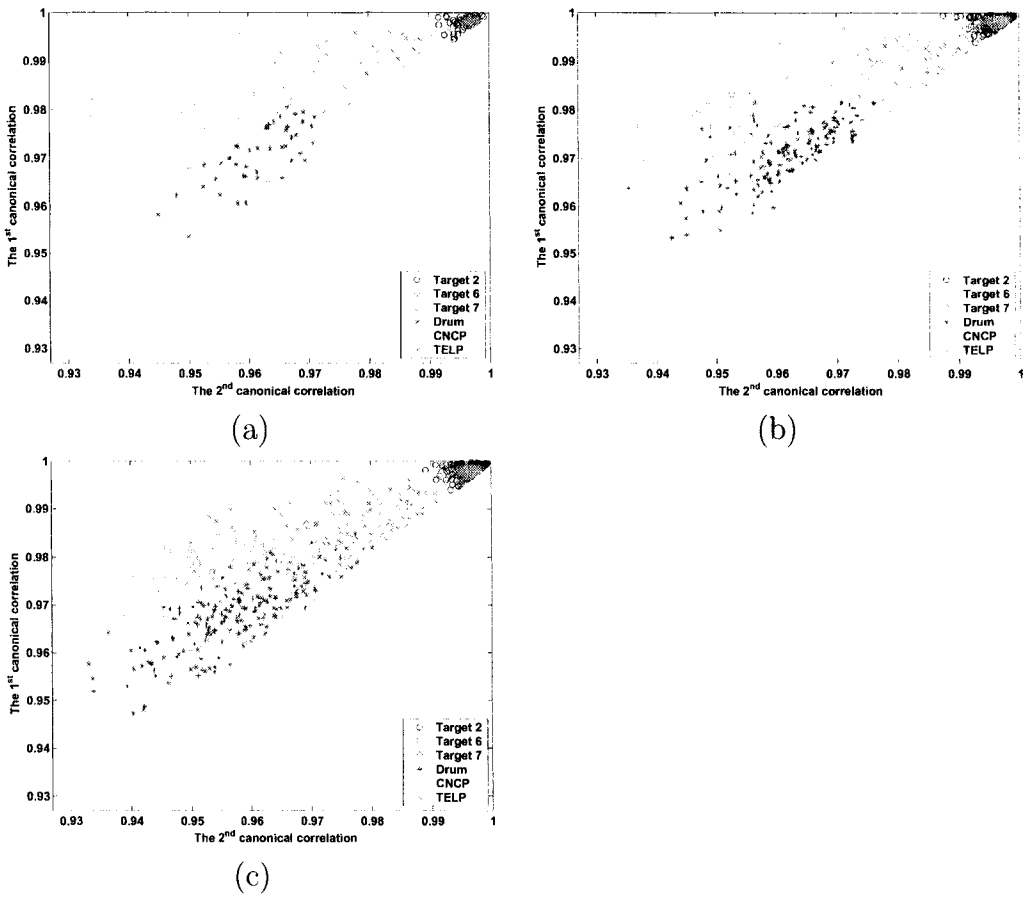

Figure 8.6: Scatter plots of the first two canonical correlations for (a) training, (b) validation, and (c) testing data sets. The plots show that canonical correlation features are indeed robust with respect to aspect angle variation.

Table 8.4: Confusion matrices of the BPNN classifier trained with the canonical correlation features that are extracted from one side of the objects.

	Validation Data Set		Testing Data Set	
Object	Target	Non-Target	Target	Non-Target
Target 6	135	0	180	0
Target 7	135	0	179	1
Target 2	135	0	178	2
Drum	0	135	0	180
Concrete Pipe	0	135	0	180
Telephone Pole	15	120	1	179

In this case, the best two-layer BPNN classifier, trained with the extracted canonical correlation features, yields a correct classification rate of 99.6% on the training data set, 98.1% on the validation data set, and 99.8% on the testing data set. The confusion matrices of this classifier for validation and testing data sets are shown in Table 8.4. It is seen that only at a few aspect angles in the validation data set the telephone pole is misclassified as a mine-like-object. Similarly, in the testing data set, there are only four misclassifications. These results clearly demonstrate that canonical correlation features for an object at only a few aspect angles can be representative of that object at almost all aspect angles, provided that the environmental condition remains unchanged.

8.4.2 Nonlinearly Mapped Two-Channel Sonar Data

We now investigate whether or not pre-processing the sonar returns using nonlinear mappings, prior to canonical correlation analysis, can in fact yield coherent high-order attributes of the original returns and improve the discrimination between targets and non-targets. For this purpose, after partitioning the two sonar returns in a pair (with 16 degrees separation) into blocks of size 50 samples, with 50% overlap, we nonlinearly map the corresponding data samples, i.e. $\mathbf{x}_i$'s and $\mathbf{y}_i$'s, using the nonlinear mapping functions $\boldsymbol{\phi}(\cdot)$ and $\boldsymbol{\psi}(\cdot)$, to $\boldsymbol{\phi}(\mathbf{x}_i)$'s and $\boldsymbol{\psi}(\mathbf{y}_i)$'s.

We have tried several different nonlinearities $\phi(\cdot)$ and $\psi(\cdot)$, though we only present the results of four representative cases. In each case, we extract the first 15 canonical correlations of the mapped data samples $\phi(\mathbf{x}_i)$ and $\psi(\mathbf{y}_i)$. The training, validation, and testing data sets are formed in the same way as in Experiment 1 in Section 8.4.1.

The nonlinearities used here have not been selected in any systematic way. The intuition was to choose the nonlinearities so that the elements of the cross-covariance matrix of the nonlinearly mapped data samples $\phi(\mathbf{x}_i)$ and $\psi(\mathbf{y}_i)$ estimate the high-order moments between the elements of $\mathbf{x}$ and $\mathbf{y}$. Further, in order to allow for a fair comparison with the linear experiments, we choose the nonlinear mappings so that the dimensions of the data samples before and after the mappings remain almost the same. Increasing the dimensions causes the mapped sample data matrices to become closer to a sample-poor case, compared to the original sample data matrices, making the comparison of the linear and nonlinear experiments difficult.

We note that systematic selection of nonlinearities for canonical correlation analysis is an open research area, and to the best of our knowledge no work has been reported on this topic. However, there are a few articles [53]– [56] that discuss the selection of nonlinearities for other information processing methods. We defer the discussion of these methods to Chapter 9. For the cases presented here, the choices of the nonlinearities $\phi(\cdot)$ and $\psi(\cdot)$ are as follows:

Case 1: In the first case, the nonlinear mappings are performed by simply raising every element of the original data samples $\mathbf{x}_i$ and $\mathbf{y}_i$ to the power of two, i.e.

$$\mathbf{x}_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,50} \end{bmatrix} \in \mathbb{R}^{50} \xrightarrow{\boldsymbol{\phi}(\cdot)} \boldsymbol{\phi}(\mathbf{x}_i) = \begin{bmatrix} x_{i,1}^2 \\ x_{i,2}^2 \\ \vdots \\ x_{i,50}^2 \end{bmatrix} \in \mathbb{R}^{50}$$
$$\mathbf{y}_i = \begin{bmatrix} y_{i,1} \\ y_{i,2} \\ \vdots \\ y_{i,50} \end{bmatrix} \in \mathbb{R}^{50} \xrightarrow{\boldsymbol{\psi}(\cdot)} \boldsymbol{\phi}(\mathbf{y}_i) = \begin{bmatrix} y_{i,1}^2 \\ y_{i,2}^2 \\ \vdots \\ y_{i,50}^2 \end{bmatrix} \in \mathbb{R}^{50}. \tag{8.1}$$

In this case, the dimensions of the data samples before and after the mappings are the same.

Figures 8.7(a)-(c) show the scatter plots of the first two canonical correlation features extracted from the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ for training, validation, and testing data sets, respectively. As can be seen from Figures 8.7(a) and (b), in the training and validation data sets (smooth bottom) features of Targets 6 and 7 are somewhat separated from those of the non-targets, though they have some overlap with those of the telephone pole. However, features of Target 2 have a relatively large overlap with those of the steel drum and concrete pipe. Clearly, this overlap leads to some degradation in the classification performance. In the testing data set (rough bottom) in Figure 8.7(c), features of all objects except Target 7 stay almost at the same regions compared to the training and validation data sets. However, features of Target 7 undergo considerable change and mix with those of the steel drum, concrete pipe, and Target 2. We observed similar behavior for features of Target 7 in the rough bottom condition in the first experiment in Section 8.4.1.

We now use the canonical correlation features extracted from the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ for classification. Due to the large amount of overlap between

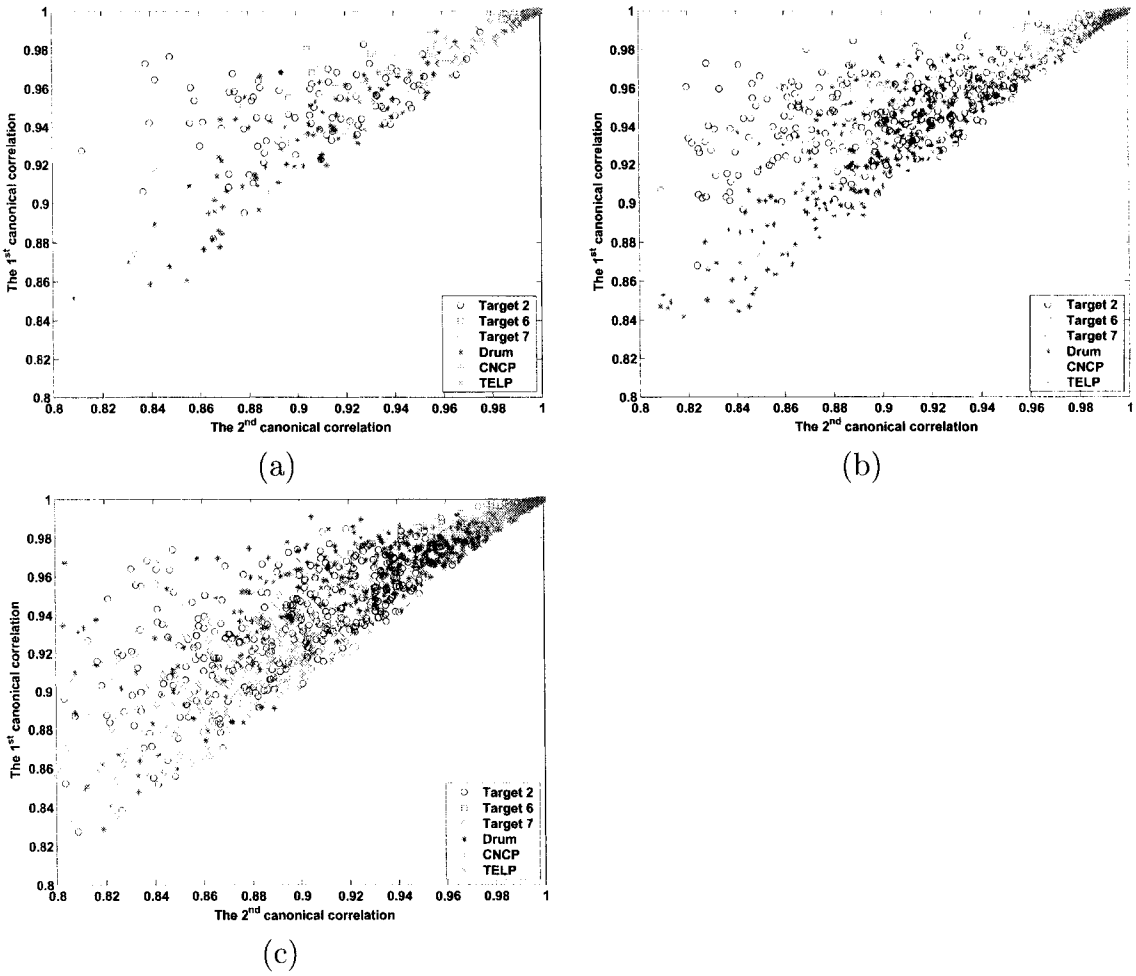

Figure 8.7: Scatter plots of the first two canonical correlations between the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ in Case 1, for (a) training, (b) validation, and (c) testing data sets.

the features of targets and non-targets, the two-layer BPNN classifiers trained with these features did not yield high correct classification rates. Therefore, we tried eight different three-layer BPNN structures, where each network was trained for ten different initializations. The training was performed for 15000 epochs. Similar to the experiments in Section 8.4.1, the best BPNN classifier was selected based on the average of correct classification rates on training and validation data sets.

The best BPNN classifier yields a correct classification rate of 91.1% for the training, 85.3% for the validation, and only 65.4% for the testing data set. The confusion matrices of this classifier are shown in Table 8.5 for both validation and testing data sets. The confusion matrix for the validation data set shows that most of the misclassifications have occurred for Target 2. This was expected, as in the scatter plot in Figure 8.7(b), features of this target are completely mixed with those of the steel drum and concrete pipe. The misclassifications for Targets 6 and 7 in the smooth bottom condition (validation data set) are due to the small overlap between features of these targets and those of the telephone pole. In the testing data set, Target 7 has been misclassified at most of the aspect angles. This was also expected, as the scatter plot in Figure 8.7(c) shows that, in the rough bottom condition, features of this target undergo considerable change and mix with those of the non-targets. Comparing the results of this case with those obtained in the first experiment in Section 8.4.1, it is clearly seen that the nonlinearities in (8.1) impair the discrimination between targets and non-targets.

Table 8.5: Confusion matrices of the BPNN classifier trained with the canonical correlation features of the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ in Case 1.

	Validation Data Set		Testing Data Set	
Object	Target	Non-Target	Target	Non-Target
Target 6	262	8	344	16
Target 7	259	11	128	232
Target 2	131	139	139	221
Drum	32	238	65	295
Concrete Pipe	32	238	154	206
Telephone Pole	16	254	59	301

Case 2: This time, the nonlinear mappings are performed by raising every element of the original data samples $\mathbf{x}_i$ and $\mathbf{y}_i$ to the power of three, i.e

$$\begin{aligned}
\mathbf{x}_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,50} \end{bmatrix} \in \mathbb{R}^{50} \xrightarrow{\boldsymbol{\phi}(\cdot)} \boldsymbol{\phi}(\mathbf{x}_i) = \begin{bmatrix} x_{i,1}^3 \\ x_{i,2}^3 \\ \vdots \\ x_{i,50}^3 \end{bmatrix} \in \mathbb{R}^{50} \\
\mathbf{y}_i = \begin{bmatrix} y_{i,1} \\ y_{i,2} \\ \vdots \\ y_{i,50} \end{bmatrix} \in \mathbb{R}^{50} \xrightarrow{\boldsymbol{\psi}(\cdot)} \boldsymbol{\phi}(\mathbf{y}_i) = \begin{bmatrix} y_{i,1}^3 \\ y_{i,2}^3 \\ \vdots \\ y_{i,50}^3 \end{bmatrix} \in \mathbb{R}^{50}.
\end{aligned} \tag{8.2}$$

Similar to the previous case, the dimensions of the data samples before and after the mappings are the same.

Figures 8.8(a)-(c) show the scatter plots of the first two canonical correlation features extracted from the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ for training, validation, and testing data sets, respectively. In these plots, features of targets and non-targets are completely mixed together in both smooth and rough bottom conditions. Clearly, this leads to poor discrimination between targets and non-targets. Furthermore, it is interesting to note that in the testing data set (rough bottom) in

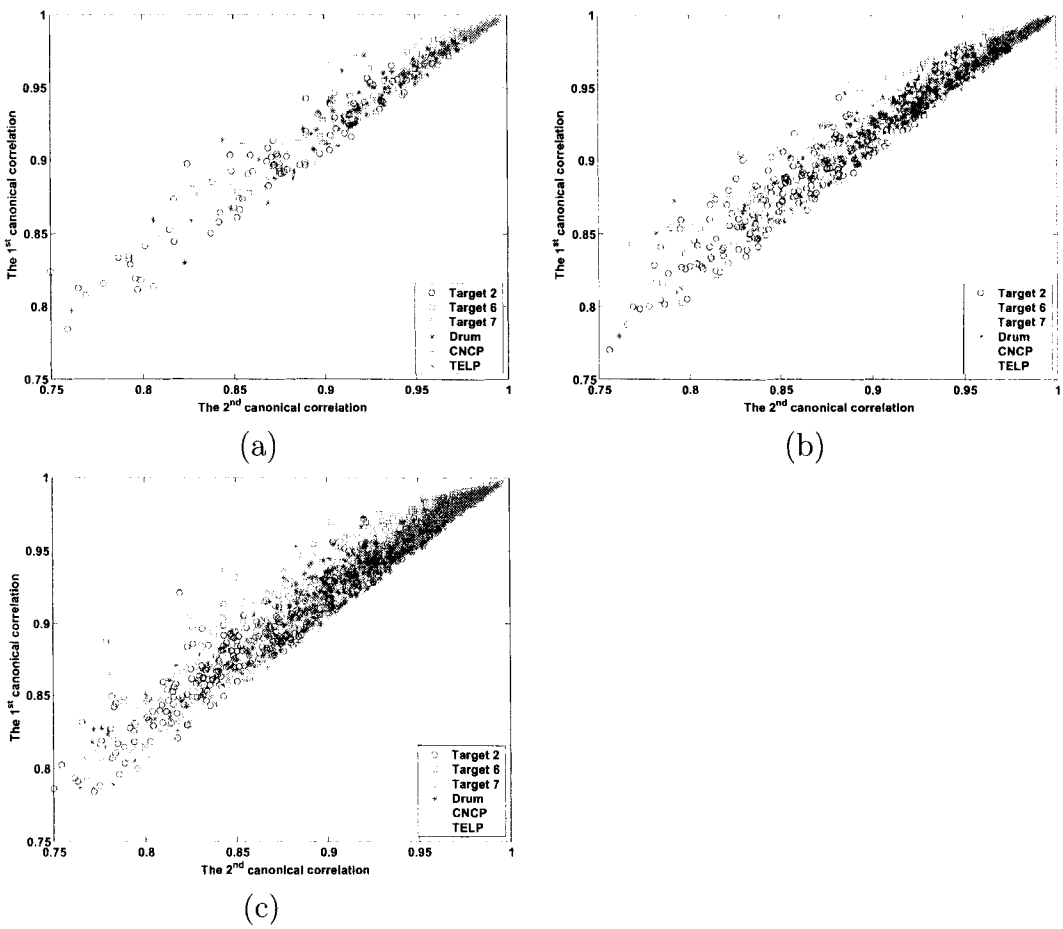

Figure 8.8: Scatter plots of the first two canonical correlations between the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ in Case 2, for (a) training, (b) validation, and (c) testing data sets.

Figure 8.8(c) again features of Target 7 undergo considerable change, while features of other objects remain almost unchanged. This suggests that the changes in features of Target 7 in the rough bottom condition are even prevalent in the nonlinear case, where canonical correlations estimate the coherence between high-order attributes of the sonar returns.

In this case, the best three-layer BPNN classifier yields a correct classification rate of 79.1% for the training, 70.0% for the validation, and 62.6% for the testing data set. The confusion matrices of this classifier are shown in Table 8.6. These results show that in this case, the canonical correlation features extracted from the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ offer very poor discrimination between targets and non-targets.

Case 3: In this case, the nonlinear mapping function $\boldsymbol{\phi}(\cdot)$ is chosen such that each element of the mapped data sample $\boldsymbol{\phi}(\mathbf{x}_i)$ is the product of the corresponding element in the original data sample $\mathbf{x}_i$ by its next element. This leads to the 49-dimensional mapped data vector $\boldsymbol{\phi}(\mathbf{x}_i) \in \mathbb{R}^{49}$. The nonlinear mapping $\boldsymbol{\psi}(\cdot)$ is chosen as the identity map:

$$\begin{aligned}
\mathbf{x}_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,50} \end{bmatrix} \in \mathbb{R}^{50} \xrightarrow{\boldsymbol{\phi}(\cdot)} \boldsymbol{\phi}(\mathbf{x}_i) = \begin{bmatrix} x_{i,1}x_{i,2} \\ x_{i,2}x_{i,3} \\ \vdots \\ x_{i,49}x_{i,50} \end{bmatrix} \in \mathbb{R}^{49} \\
\mathbf{y}_i = \begin{bmatrix} y_{i,1} \\ y_{i,2} \\ \vdots \\ y_{i,50} \end{bmatrix} \in \mathbb{R}^{50} \xrightarrow{\boldsymbol{\psi}(\cdot)} \boldsymbol{\phi}(\mathbf{y}_i) = \begin{bmatrix} y_{i,1} \\ y_{i,2} \\ \vdots \\ y_{i,50} \end{bmatrix} \in \mathbb{R}^{50}.
\end{aligned} \tag{8.3}$$

Therefore, in this case, only the first sonar return, which corresponds to the $\mathbf{x}-$channel, is processed with a nonlinear mapping, while the data samples of the $\mathbf{y}-$channel are kept unchanged.

Table 8.6: Confusion matrices of the BPNN classifier trained with the canonical correlation features of the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ in Case 2.

	Validation Data Set		Testing Data Set	
Object	Target	Non-Target	Target	Non-Target
Target 6	183	87	217	143
Target 7	222	48	186	174
Target 2	161	109	204	156
Drum	91	179	140	220
Concrete Pipe	107	163	127	233
Telephone Pole	44	226	67	293

Figures 8.9(a)-(c) show the scatter plots of the first two canonical correlation features extracted from the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and the original data samples $\mathbf{y}_i$ for the training, validation, and testing data sets, respectively. As can be seen from Figures 8.9(a) and (b), in the training and validation data sets (smooth bottom), features of Targets 6 and 7 exhibit some overlap with those of the concrete pipe and telephone pole, though they are still separable at most aspect angles. Features of Target 2 have some overlap with those of the steel drum and concrete pipe. In the testing data set (rough bottom) in Figure 8.9(c), features of all objects except Target 7 stay almost at the same regions compared to the training and validation data sets. However, again features of Target 7 undergo considerable change, similar to the previous cases.

In this case, the best three-layer BPNN classifier yields a correct classification rate of 95.2% for the training, 90.2% for the validation, and 66.9% for the testing data set. The confusion matrices of this classifier are shown in Table 8.7 for both validation and testing data sets. Compared to Case 1 (Table 8.5), in this case correct classification of Target 2 in both validation and testing data sets has improved. However, Case 1 offers better classification rates for the non-targets in the testing data set (rough bottom). Nonetheless, compared to the first experiment in Section 8.4.1, the nonlinearities in both cases impair the classification results.

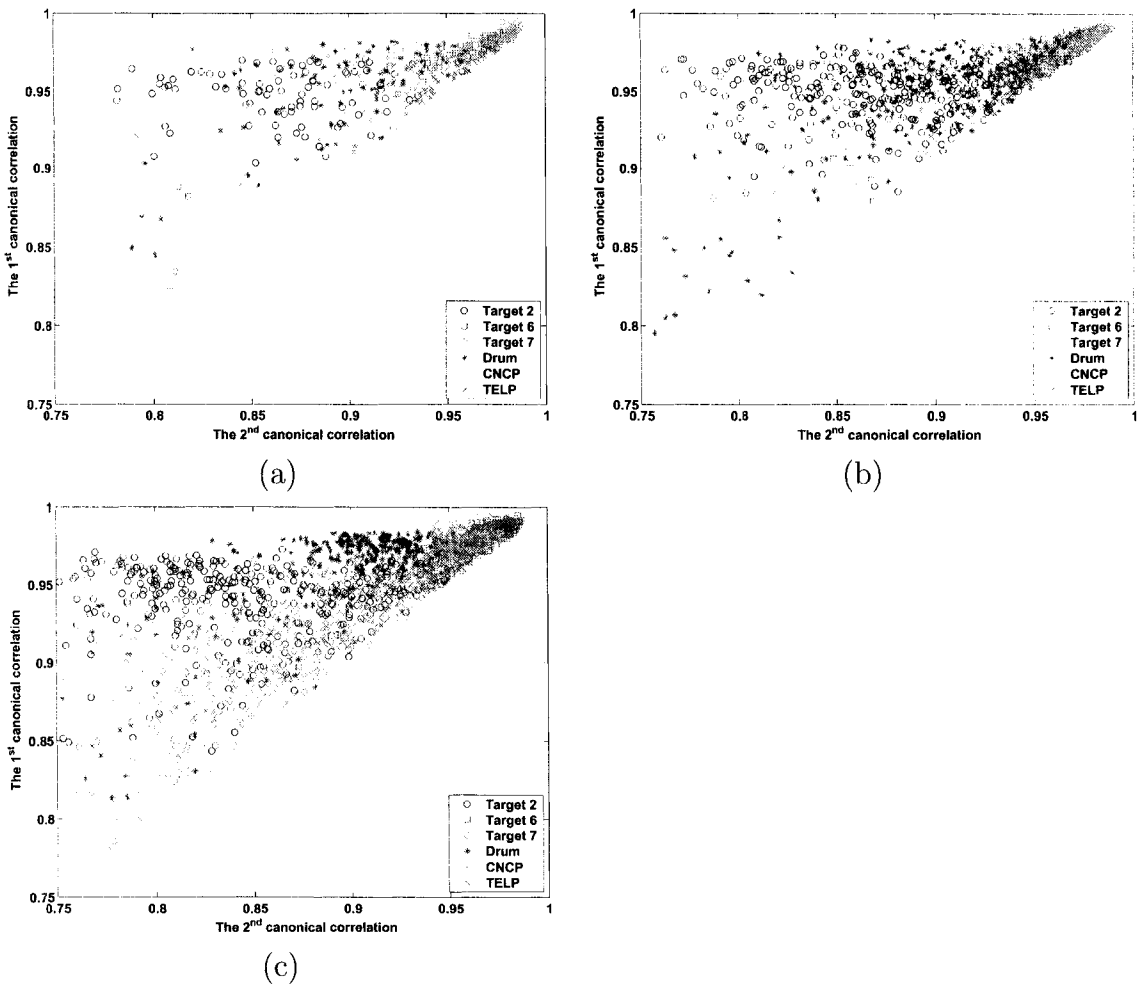

Figure 8.9: Scatter plots of the first two canonical correlations between the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ in Case 3, for (a) training, (b) validation, and (c) testing data sets.

Table 8.7: Confusion matrices of the BPNN classifier trained with the canonical correlation features of the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ in Case 3.

	Validation Data Set		Testing Data Set	
Object	Target	Non-Target	Target	Non-Target
Target 6	254	16	333	27
Target 7	262	8	213	147
Target 2	214	56	285	75
Drum	36	234	105	255
Concrete Pipe	31	239	255	105
Telephone Pole	12	258	106	254

Case 4: Here, the nonlinear mapping $\boldsymbol{\phi}(\cdot)$ is chosen as in Case 3. The nonlinear mapping $\boldsymbol{\psi}(\cdot)$ is chosen so that each element of the mapped data sample $\boldsymbol{\psi}(\mathbf{y}_i)$ is the product of the corresponding element in the original data sample $\mathbf{y}_i$ by its next two elements. This leads to the 48-dimensional mapped data vector $\boldsymbol{\psi}(\mathbf{y}_i) \in \mathbb{R}^{48}$. That is,

$$\mathbf{x}_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,50} \end{bmatrix} \in \mathbb{R}^{50} \xrightarrow{\boldsymbol{\phi}(\cdot)} \boldsymbol{\phi}(\mathbf{x}_i) = \begin{bmatrix} x_{i,1}x_{i,2} \\ x_{i,2}x_{i,3} \\ \vdots \\ x_{i,49}x_{i,50} \end{bmatrix} \in \mathbb{R}^{49}$$

$$\mathbf{y}_i = \begin{bmatrix} y_{i,1} \\ y_{i,2} \\ \vdots \\ y_{i,50} \end{bmatrix} \in \mathbb{R}^{50} \xrightarrow{\boldsymbol{\psi}(\cdot)} \boldsymbol{\phi}(\mathbf{y}_i) = \begin{bmatrix} y_{i,1}y_{i,2}y_{i,3} \\ y_{i,2}y_{i,3}y_{i,4} \\ \vdots \\ y_{i,48}y_{i,49}y_{i,50} \end{bmatrix} \in \mathbb{R}^{48}. \tag{8.4}$$

Figures 8.10(a)-(c) show the scatter plots of the first two canonical correlation features extracted from the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ for the training, validation, and testing data sets, respectively. These plots show that features of targets and non-targets are completely mixed together in both smooth and rough bottom conditions, and again similar to the previous cases, features of Target 7 undergo considerable change in the rough bottom condition.

In this case, the best three-layer BPNN classifier yields a correct classification rate of 84.3% on the training, 72.3% on the validation, and 64.4% on the testing data set. The confusion matrices of this classifier are shown in Table 8.8. As can be seen, in this case the canonical correlation features extracted from the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ offer very poor discrimination between targets and non-targets.

Table 8.9 summarizes all the classification rates obtained in the nonlinear experiments in this section and compares them with those obtained in the first experiment

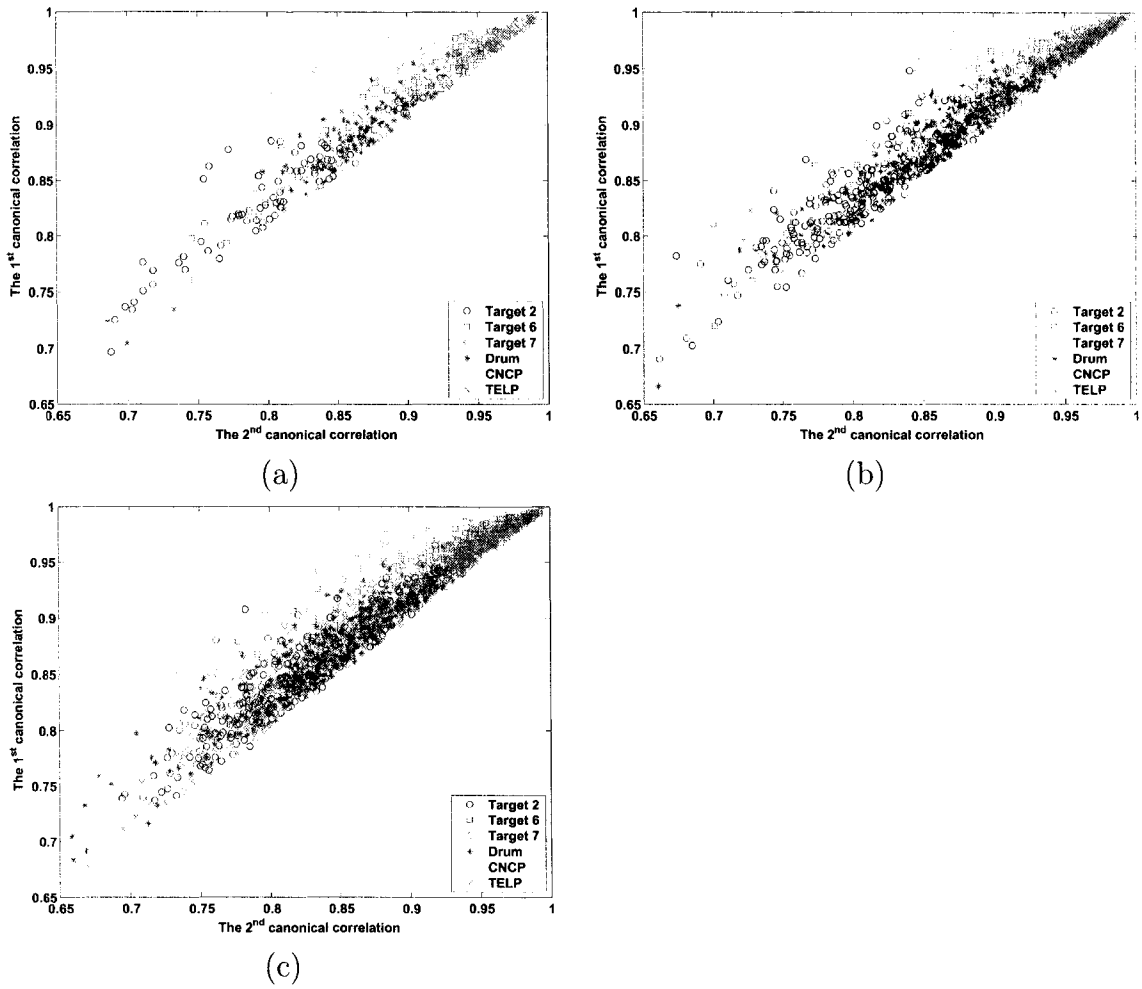

Figure 8.10: Scatter plots of the first two canonical correlations between the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ in Case 4, for (a) training, (b) validation, and (c) testing data sets.

Table 8.8: Confusion matrices of the BPNN classifier trained with the canonical correlation features of the mapped data samples $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\psi}(\mathbf{y}_i)$ in Case 4.

	Validation Data Set		Testing Data Set	
Object	Target	Non-Target	Target	Non-Target
Target 6	189	81	250	110
Target 7	214	56	206	154
Target 2	170	100	227	133
Drum	87	183	160	200
Concrete Pipe	84	186	141	219
Telephone Pole	41	229	71	289

Table 8.9: Comparison of the classification rates in nonlinear Cases 1 to 4 with those in the first experiment in Section 8.4.1.

Features	Training	Validation	Testing
Linear case	99.1%	98.6%	81.0%
Nonlinear Case 1	91.1%	85.3%	65.4%
Nonlinear Case 2	79.1%	70.0%	62.6%
Nonlinear Case 3	95.2%	90.2%	66.9%
Nonlinear Case 4	84.3%	72.3%	64.4%

in Section 8.4.1. Overall, the experiments show that the canonical correlation features extracted from the nonlinearly mapped sonar returns impair the discrimination between targets and non-targets in comparison with the linear case. This suggests that, for the ARL-UT data set, the second-order statistical features carry more discriminatory information than high-order ones.

8.5 Conclusions

In this chapter, canonical correlation analysis was exploited to develop a multi-aspect feature extraction method for underwater target classification from a wideband sonar data set. The basic idea was that in the presence of an object (target or non-target) consecutive sonar returns exhibit linear dependence or coherence, whereas in the absence of an object, the sonar returns are not coherent. Further, we hypothesized that the degree of coherence between the two sonar returns generated by the presence of a mine-like object is different from that caused by the presence of a non-mine-like object. To estimate this coherence, the dominant canonical correlations between the two returns were extracted and used as features for classification.

Our experiments on the wideband ARL-UT data set demonstrate that canonical correlation features can indeed offer good separation between mine-like and non-mine-like objects. The results show that except for one of the objects (Target 7), the canonical correlation features are robust to changes in the bottom condition. Moreover, we showed that in a fixed bottom condition, canonical correlation features do not vary with changes in aspect angle. Several experiments were also conducted to determine whether or not pr-processing the sonar returns with nonlinear functions, prior to canonical correlation analysis, can improve the discrimination between targets and non-targets. The results showed that not only the canonical correlation features extracted from the nonlinearly mapped sonar returns do not improve the discrimination between targets and non-targets, they impair it compared to the canonical correlation features extracted from the original (not mapped) sonar returns.

CHAPTER 9

REVIEW OF EXISTING METHODS FOR SELECTING NONLINEAR FUNCTIONS FOR NONLINEAR INFORMATION PROCESSING

9.1 Introduction

In Chapters 7 and 8, we exploited the idea of pre-processing two-channel data with nonlinear mappings, prior to canonical correlation analysis, with the aim to capture coherence between high-order attributes of the original channels. However, we selected the nonlinearities in a non-systematic way. The problem is that, in practice, there are potentially infinite number of nonlinearities to choose from. For some nonlinearities the mapped data channels may become more coherent, whereas for others they may become less coherent, according to our definition of coherence in (2.28).

Naturally, the question that arises is, is there a systematic way for selecting nonlinearities in order to obtain the most coherent high-order attributes of the original data channels? To the best of our knowledge, no work has been reported on this topic. In fact, the lack of systematic methods for selection of nonlinearities is not limited to canonical correlation analysis. All nonlinear information processing methods, including SVM's [33], [34] and other kernel machines [35], are plagued by the same problem.

For kernel-based methods, selection of a good kernel function is typically performed in a rather heuristic way. The kernel nonlinear information processing is performed for several different classes of kernel functions, and the one that results in the best performance is selected as an appropriate kernel. Even within every kernel class, several kernel functions with different choices of parameters need to be tried, in order to identify the best set of parameters. The reason is that, even for a given class of kernel functions, appropriate selection of the kernel parameters is usually difficult.

In this chapter, we present a review of some of the existing methods for selection of kernel functions for nonlinear information processing. As mentioned earlier, the literature in this area is very limited. In fact, so far, only a few methods [53]- [56] have been reported for selection of kernel functions, which are mainly concerned with kernel selection for SVM's and kernel-based classifiers. Nonetheless, a review of these methods may be insightful for future research.

9.2 Definitions, Notation, and Terminology

We define $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_M] \in \mathbb{R}^{n \times M}$ as the feature matrix of feature vectors $\mathbf{y}_i \in \mathbb{R}^n$ for an m-class classification problem. Correspondingly, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M] \in \mathbb{R}^{m \times M}$ is defined as the class membership matrix of the known class membership vectors $\mathbf{x}_i \in \mathbb{R}^m$. The class membership vector $\mathbf{x}_i$ consists of $(m-1)$ elements that are -1, and one element that is equal to one, the position of which determines the class membership of the feature vector $\mathbf{y}_i$. The feature matrix $\mathbf{Y}$ and class membership matrix $\mathbf{X}$ together build the training set for the classification problem. We define the input space as the subspace spanned by the feature vectors $\mathbf{y}_i$. In cases where a validation set is also required for kernel selection, we assume that a validation feature matrix $\mathbf{Y}_\star = [\mathbf{y}_{M+1}, \ldots, \mathbf{y}_N] \in \mathbb{R}^{n \times (N-M)}$, with known class membership matrix $\mathbf{X}_\star = [\mathbf{x}_{M+1}, \ldots, \mathbf{x}_N] \in \mathbb{R}^{m \times (N-M)}$, is also available. The kernel function $k(\mathbf{y}_i, \mathbf{y}_j) = \boldsymbol{\psi}(\mathbf{y}_i)^T \boldsymbol{\psi}(\mathbf{y}_j)$ is a function that acts on the feature vectors $\mathbf{y}_i$ and $\mathbf{y}_j$, but

measures the inner product between the *(implicitly) mapped feature vectors* $\psi(\mathbf{y}_i)$ and $\psi(\mathbf{y}_j)$. We refer to the subspace spanned by the mapped feature vectors $\psi(\mathbf{y}_i)$, $i \in [1, M]$, as the *feature space*. This is the subspace that is spanned by the kernel Gram matrix $\mathbf{K} = [k(\mathbf{y}_i, \mathbf{y}_j)]_{i,j=1}^{M}$, and hence we denote it by $< \mathbf{K} >$.

9.3 Review of Kernel Selection Methods

The kernel selection approaches developed in [53]– [56] exploit two different basic ideas: (1) adapting Gaussian kernels in SVM [53], [54] and (2) kernel-target alignment [55], [56]. In what follows, we review these two methods. The purpose of this review is to familiarize the readers with the basic idea behind each approach, and hence details are not presented. It is assumed that the readers are familiar with the theory of support vector and kernel machines, and their terminology.

9.3.1 Adapting Gaussian Kernels in SVM

In this approach [53], [54], the idea is to adjust the parameter of a Gaussian kernel in order to minimize the upper bound on the generalization error [33], [34] of an SVM classifier. Considering a Gaussian kernel of the form

$$k(\mathbf{y}_i, \mathbf{y}_j) = e^{-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{2\sigma^2}} \tag{9.1}$$

the free design parameter is σ^2, which determines the spread of the Gaussian kernel function. The idea in [53] and [54] is to adjust σ^2 to minimize the upper bound ϵ on the generalization error of the SVM classifier. To accomplish this goal, in [53] the upper bound ϵ is shown to be a smooth function of the kernel parameter σ^2. This means that when the upper bound ϵ is minimum, a small variation in the kernel parameter σ^2 will produce a small variation in ϵ, hence the upper bound remains near its optimal value. Based on this idea, a procedure is suggested [53] for updating the kernel parameter σ^2. The idea is to start with a very small initial value for σ^2, and use the training data samples $\mathbf{y}_i$ with known class label $\mathbf{x}_i$ to train the SVM classifier.

At each training step after a training feature vector is given to the SVM classifier and the optimal hyperplane is determined, the validation set $\mathbf{Y}_\star$ is used to evaluate the generalization ability of the trained SVM classifier, by computing the generalization error. If the generalization error is smaller than a pre-determined bound the training may be stopped, and the value of σ^2 at that training iteration is selected as the best value. But if the generalization error is larger than the pre-determined bound, the value of σ^2 is changed to $\sigma^2 + \delta\sigma^2$ and the training continues. The adjustment in σ^2 is performed using a gradient descent-type algorithm, called Kernel-Adatron [53], [54], that minimizes the generalization error of the SVM classifier for the validation set.

In [53], this kernel selection method has been applied to a breast cancer classification data set. The results show that the kernel adaptation procedure can improve the generalization ability of the SVM classifier, compared to the cases where σ^2 is tuned by a trial and error procedure. However, the problem remains that the main idea behind the adjusting mechanism is applicable only when the performance measure to be optimized is a smooth function of the kernel parameters. As a result, this idea has only been employed for adapting Gaussian kernels to minimize the upper bound on the generalization error of SVM classifiers.

9.3.2 Kernel-Target Alignment

In the kernel-target alignment approach [55], [56] the basic idea is that a good kernel function is the one that results in an (implicit) feature space that matches the target function to be learned by the classifier. The target function of a classifier is the function that converts every (mapped) feature vector to its true class membership vector. In a classification problem, the target function is unknown, and hence in [55] and [56] the class membership vectors are used instead of the target function. That is, the matching is performed between kernel functions and class membership vectors.

To measure the degree of match between a kernel function and the class membership vectors, [55] and [56] define an alignment measure, called the *empirical kernel-target alignment* measure. For the kernel function $k(\mathbf{y}_i, \mathbf{y}_j) = \boldsymbol{\psi}(\mathbf{y}_i)^T \boldsymbol{\psi}(\mathbf{y}_j)$, this alignment measure is defined as [55], [56]

$$A = \frac{< \mathbf{K}, \mathbf{X}^T\mathbf{X} >}{\sqrt{< \mathbf{K}, \mathbf{K} >< \mathbf{X}^T\mathbf{X}, \mathbf{X}^T\mathbf{X} >}} \tag{9.2}$$

where $\mathbf{K} = [k(\mathbf{y}_i, \mathbf{y}_j)]_{i,j=1}^{M}$ is the kernel Gram matrix of the (implicitly) mapped feature vectors $\boldsymbol{\psi}(\mathbf{y}_i)$, $i \in [1, M]$ and $\mathbf{X}^T\mathbf{X}$ is the Gram matrix of the class membership vectors $\mathbf{x}_i$. The term $< \mathbf{K}, \mathbf{X}^T\mathbf{X} >$ is the inner product between the Gram matrices $\mathbf{K}$ and $\mathbf{X}^T\mathbf{X}$, and is defined as [55], [56]

$$< \mathbf{K}, \mathbf{X}^T\mathbf{X} >= \sum_{i,j=1}^{M} [\mathbf{K}]_{i,j}[\mathbf{X}^T\mathbf{X}]_{i,j} = \sum_{\mathbf{x}_i = \mathbf{x}_j} k(\mathbf{y}_i, \mathbf{y}_j) - \sum_{\mathbf{x}_i \neq \mathbf{x}_j} k(\mathbf{y}_i, \mathbf{y}_j). \tag{9.3}$$

The last equality follows from the fact that the class membership vectors $\mathbf{x}_i$ contain $(m-1)$ elements equal to -1 and one element equal to one. Similar expressions may be written for $< \mathbf{K}, \mathbf{K} >$ and $< \mathbf{X}^T\mathbf{X}, \mathbf{X}^T\mathbf{X} >$. In [55], the terms $\sum_{\mathbf{x}_i = \mathbf{x}_j} k(\mathbf{y}_i, \mathbf{y}_j)$ and $\sum_{\mathbf{x}_i \neq \mathbf{x}_j} k(\mathbf{y}_i, \mathbf{y}_j)$ are interpreted as measures of "within class" and "between class" distance, respectively. Thus, $< \mathbf{K}, \mathbf{X}^T\mathbf{X} >$ is viewed as a measure of clustering of the classes.

The idea is that the best kernel function is the one that results in maximum alignment between $\mathbf{K}$ and $\mathbf{X}^T\mathbf{X}$. The intuition behind this stems from the fact that the alignment measure A takes its maximum value, i.e. one, when the implicitly mapped feature vectors $\boldsymbol{\psi}(\mathbf{y}_i)$ are equal to the true class membership vectors $\mathbf{x}_i$. The reason is that when $\boldsymbol{\psi}(\mathbf{y}_i) = \mathbf{x}_i$, the kernel Gram matrix $\mathbf{K}$ becomes $\mathbf{K} = \mathbf{X}^T\mathbf{X}$. In other words, the idea in [55] and [56] is that, in a classification problem, the best set of mapped feature vectors are the true class membership vectors. Thus, the best kernel function is the one that is associated with the nonlinear mapping functions that map the original feature vectors $\mathbf{y}_i$ to the true class membership vectors $\mathbf{x}_i$. However,

in practice the nonlinear mapping functions that perform such transformation are not known. Thus, one should try to find the nonlinear mappings that result in mapped feature vectors $\psi(\mathbf{y}_i)$ that are closest (most aligned) to the class membership vectors $\mathbf{x}_i$. Note that this is equivalent to finding the kernel Gram matrix $\mathbf{K}$ that is most aligned with the Gram matrix $\mathbf{X}^T\mathbf{X}$, as the the kernel function $k(\mathbf{y}_i, \mathbf{y}_j) = \psi(\mathbf{y}_i)^T\psi(\mathbf{y}_j)$ determines the nonlinear mapping $\psi(\cdot)$.

In order to select a kernel function with a good alignment, [55] and [56] propose to start with an initial choice of kernel function and then adjust the kernel function to maximize the alignment. Consider $k(\mathbf{y}_i, \mathbf{y}_j) = \psi(\mathbf{y}_i)^T\psi(\mathbf{y}_j)$ as the initial kernel function, yielding the initial kernel Gram matrix $\mathbf{K} = [k(\mathbf{y}_i, \mathbf{y}_j)]_{i,j=1}^M$. Also consider $\mathbf{K} = \sum_{i=1}^M \lambda_i \mathbf{q}_i \mathbf{q}_i^T$ as the SVD of $\mathbf{K}$, with the singular vectors $\mathbf{q}_i$ and the singular values λ_i. Having these, [55] and [56] define the parameterized class of (modified) kernel Gram matrices $\hat{\mathbf{K}}$ as

$$\hat{\mathbf{K}} = \sum_{i=1}^{M} \alpha_i \mathbf{q}_i \mathbf{q}_i^T \tag{9.4}$$

where α_i's are some scalar weights, yet to be determined. The kernel selection problem now becomes one of finding the parameters α_i such that the kernel-target alignment measure A between $\hat{\mathbf{K}}$ and $\mathbf{X}^T\mathbf{X}$ is maximized [55], [56].

Note that in this approach, the feature space $< \mathbf{K} >$ will never change, since the modified kernel Gram matrix $\hat{\mathbf{K}}$ is built in the subspace $< \mathbf{K} >$, which is the initial feature space. Therefore, if the initial feature space $< \mathbf{K} >$ does not include feature vectors that are fairly aligned with the class membership vectors, the kernel-target alignment method will never result in a kernel with good alignment, as the feature space is always the same. Thus, the question is, is there a way to combine several feature spaces to improve the alignment?

To address this question, [55] and [56] propose the idea of *kernel combination*, by defining a *kernel-kernel alignment* measure. This measure determines the alignment

between two kernel Gram matrices $\mathbf{K}_1$ and $\mathbf{K}_2$, each of which defines an initial feature space. The kernel-kernel alignment measure is very similar to the kernel-target alignment measure A in (9.2), except that the Gram matrices $\mathbf{K}$ and $\mathbf{X}^T\mathbf{X}$ in (9.2) are replaced with the kernel Gram matrices $\mathbf{K}_1$ and $\mathbf{K}_2$. Using this kernel-kernel alignment measure, along with the Cauchy-Schwartz inequality [8], it is shown in [55] and [56] that when the kernel Gram matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ are equally aligned with $\mathbf{X}^T\mathbf{X}$, but are poorly aligned with each other, the combined kernel Gram matrix $\mathbf{K}_1 + \mathbf{K}_2$ is more aligned with $\mathbf{X}^T\mathbf{X}$ than $\mathbf{K}_1$ and $\mathbf{K}_2$, individually. Thus, by combining two kernel functions, one can construct a new kernel function that is more aligned with the class membership vectors.

The kernel-target alignment and kernel combination methods described here have been applied to a text classification data set in [55], and an Ionosphere classification data set in [56]. The results on these two data sets show that optimizing the kernel-target alignment measure and using the kernel combination idea can indeed improve the classification performance, when compared to the cases where kernel selection is performed by a trial and error procedure.

Compared to the kernel selection method in Section 9.3.1, which is only applicable to Gaussian kernels and support vector machines, the kernel-target alignment method of [55] and [56] is more general, as its applicability is not limited to a special class of kernel functions or support vector machines. It can be applied to any class of kernel functions and kernel classification machines. However, the success of this method depends on the initial guess for the kernel function, which in turn requires some prior knowledge about how to choose a good (highly aligned) feature space. Clearly, this is not easily discernable in practice.

9.4 Conclusions and Discussion

In this chapter, a review of some of the existing methods for selecting nonlinear functions for nonlinear information processing was presented. These methods were: (1) adapting Gaussian kernels in SVM [53], [54] and (2) kernel-target alignment [55], [56]. In the first method the parameter of a Gaussian kernel function is adjusted iteratively to minimize the upper bound on the generalization error of an SVM classifier. The limitation of this approach is that the performance index to be optimized (e.g. the upper bound) must be a smooth function of the kernel parameters. As a result, this method has remained limited to support vector machines and Gaussian kernels.

The second approach, i.e. kernel-target alignment, is more general and is not limited to support vector machines or a specific class of kernel functions. However, it requires some prior knowledge about how to choose an aligned feature space, which is not easily discernable in practice. In this approach, kernel selection is performed by selecting an initial feature space, embedded in the initial choice for the kernel function, and then adjusting the kernel function by maximizing an alignment measure between the kernel function and class membership vectors. By introducing a kernel-kernel alignment measure it is also possible to combine different kernel functions to obtain a kernel function which is more aligned with the class membership vectors.

In general, the applicability of the kernel selection methods reviewed in this chapter are limited to pattern classification applications. Selecting nonlinearities for other (kernel) nonlinear information processing methods remains an open research problem. For canonical correlation analysis, the problem is even more challenging, as the type of nonlinearity to be considered is no longer limited to the kernel-producing ones. The reason is that, as established in Chapter 7, using high-dimensional kernel-producing nonlinearities typically results in sample-poor data matrices, which in turn produce defective empirical canonical correlations. This means that the search has to be conducted over possibly infinite classes of nonlinearities.

CHAPTER 10

CONCLUSIONS AND FUTURE WORK

10.1 Conclusions

In this thesis, we have addressed some of the issues in canonical correlation analysis of two-channel data, established a direct connection between canonical coordinates and certain two-channel signal processing problems, and exploited canonical correlations for classification of underwater targets. The major contributions of this thesis may be summarized as follows:

1. A general class of two-channel CLS problems, with various constraints, has been introduced in Chapter 3, and a general set of solutions has been derived. The solution to each two-channel CLS problem is determined from a coupled (asymmetric) generalized eigenvalue problem. Further, the connections between two-channel CLS problems and various canonical coordinate systems have been established. It has been shown that depending upon the constraints, the two-channel CLS solution decomposes the two data channels into one of three important coordinate systems: canonical coordinates, half-canonical coordinates, or PCCA coordinates.

2. A unified framework for deriving three different classes of reduced-rank Wiener filters has been developed in Chapter 4, with each class corresponding to a particular error measure for reduced-rank estimation. We have established that two

of the classes, corresponding to the whitened MSE and the volume of the concentration ellipse, are equivalent. For these two classes, canonical coordinates are optimal for reduced-rank Wiener filtering. For the third class, which corresponds to MSE estimation, half-canonical coordinates are optimal for reduced-rank Wiener filtering. Our results reproduce those of [8], [9]. However, we have derived all of these results in a unified way, using the line of argument presented in [8] for reduced-rank Wiener filtering in half-canonical coordinates. Additionally, we have presented several implementations of the reduced-rank Wiener filter in each class, and clarified the connections between reduced-rank Wiener filters and two-channel CLS filters.

3. Various alternating power methods have been developed in Chapter 5, which provide simple methods for recursively computing the canonical coordinate and half-canonical coordinate mapping vectors, addressing the deficiencies of conventional methods. These alternating power methods may be viewed as *two-step* decompositions of the standard power method, as they solve a coupled (asymmetric) generalized eigenvalue problem through power iterations. They may be used in deflation, block, or block-deflation mode, allowing for computation of the canonical coordinate and half-canonical coordinate mapping vectors, one by one, or in groups. Moreover, they may be used in batch mode on a fixed data sample, or in online mode for updating the mapping vectors in time. Provided that the rank-reduction is relatively large and the singular values of the coherence or half-coherence matrix are not close together, the alternating power methods can be more efficient in computation than the conventional methods, as they do not require any matrix square-roots.

   The alternating power methods developed in Chapter 5 are identical in form to those derived in [9] for computing reduced-rank Wiener filters. However, the algorithms in [9] do not yield the canonical and half-canonical coordinate maps,

and the corresponding canonical and half-canonical correlation matrices. Thus, the main contribution here is the discovery that alternating power methods may be used to compute canonical and half-canonical coordinate maps and correlations, making them more applicable in a wider variety of signal processing problems.

4. A network structure and a set of updating rules for recursive extraction of canonical coordinates of two data channels have been developed in Chapter 6. The network is based on a constrained minimization problem that exploits a deflation process. The deflation process is performed by incorporating lateral connections into the network. The structure of the network, along with the updating rules, allows a new node to be added to the network in order to extract a new canonical coordinate pair, without the need to retrain the previous nodes. The main contribution here is the use of lateral connections for performing the deflation process that subtracts the contributions of the already extracted canonical coordinates from the original two-channel data. However, since the network is updated using a gradient descent algorithm, it suffers from slow convergence (even as slow as linear convergence) and sometimes instability, depending on the weight initialization and choice of the step size. Thus, in two-channel signal processing applications, where recursive extraction of canonical coordinates is required, the alternating power methods developed in this thesis are preferred.

5. In Chapter 7, we have studied the canonical coordinate decomposition of two-channel data, when the channel covariances are estimated from a limited number of data samples. Depending on the number of samples drawn from each channel, and the ranks of the sample data matrices, two different cases emerge: the sample-poor case, in which the number of data samples is smaller than the sum of the ranks of the data matrices, and a sample-rich case, in which the number

of data samples is greater than the sum of the ranks of the data matrices. It has been shown that, in either case, it is the rows of the sample data matrices that determines the empirical canonical correlations, and that the empirical canonical correlations measure the cosines of the principal angles between the row spaces of the two data matrices. Further, we have shown that the empirical canonical correlations form a maximal set of invariants for the composite sample covariance matrix of two-channel data.

We have established, in Chapter 7, that in a sample-poor case some of the empirical canonical correlations or principal cosines are always equal to one, regardless of the two-channel model that generates the data samples. Therefore, the empirical canonical correlations are defective and may not be used as estimates of canonical correlations between random variables. Geometrically, this means that principal angles between linear subspaces of Euclidean space can not be used as estimates of the principal angles between corresponding linear subspaces of the Hilbert space of second-order random variables. In a sample-rich case, however, the empirical canonical correlations do estimate the canonical correlations and principal cosines between the random variables that generate the samples, and hence may be used for estimating coherence between two data channels.

These results have interesting implications for canonical correlation analysis of nonlinear functions of two-channel data. They imply that sample data matrices of nonlinearly mapped data must remain sample–rich for empirical canonical correlations to estimate coherence between high-order attributes of the original channels. Additionally, we have argued that in cases where the kernel formulation of canonical correlation analysis is computationally advantageous with respect to the direct formulation, the empirical canonical correlations between

two data matrices do not usefully estimate coherence between the corresponding data channels. Therefore, this computational advantage is superficial and does not have any practical value.

6. A new multi-aspect feature extraction method for underwater target classification, from acoustic backscattered data, has been developed in Chapter 8. This method exploits linear dependence or coherence between two consecutive sonar returns with certain aspect separation. This has been accomplished by extracting the dominant canonical correlations between the two sonar returns. The idea is that linear dependence between the sonar returns is an indication of the presence of a common signature, whereas linear independence indicates the absence of a common signature. Further, the amount of coherence between the two sonar returns induced by the presence of a mine-like object is different from that caused by the presence of a non-mine-like object.

   The developed feature extraction method was tested on the ARL-UT wideband data set, which contained acoustic backcattered signals from several mine-like and non-mine-like objects in two different bottom conditions, namely smooth and rough. The results demonstrated that canonical correlation features can offer very good discrimination between mine-like and non-mine-like objects, and are fairly robust to changes in the bottom condition. Moreover, in a fixed bottom condition, canonical correlation features are robust against in aspect angle. In addition, several experiments were conducted to determine whether or not pre-processing sonar returns with nonlinear functions can indeed result in better discrimination between targets and non-targets in the ARL-UT data set. The results showed that not only the canonical correlation features extracted from the nonlinearly mapped sonar returns do not improve the discrimination between targets and non-targets, they impair it compared to those extracted from the original (linear) sonar returns.

10.2 Future Research

The developments in this thesis provide new insights for solving and analyzing two-channel signal processing problems in canonical coordinates. In light of these developments, there are several problems that merit further research:

1. **Investigating canonical correlations for underwater target classification**

   In Chapter 8, we have demonstrated the promise of canonical correlations for discriminating mine-like-objects from non-mine-like objects in the wideband ARL-UT data set. It would be interesting to further evaluate our hypothesis and results by experimenting on other sonar data sets with more diverse object types, and environmental conditions. For instance, experimenting with data sets, such as buried object scanning sonar (BOSS) data set [87], that contain sonar returns from buried objects would be particularly illuminating.

2. **Analysis of coherence between high-order attributes of two data channels**

   The question that needs to be addressed here is, given two sample-rich data matrices drawn from a two-channel vector process, is there a way to determine which high-order attributes of the two data channels are most coherent? The answer to this question can be helpful in developing a systematic way for designing nonlinearities for processing the data channels prior to canonical correlation analysis, with the aim to capture coherence between high-order attributes of the original data channels. Further, for a given classification problem, it would be interesting to determine which high-order attributes of the two data channels can be most helpful to discriminate between different classes, and whether or not there is a connection between the most coherent and the most discriminant high-order attributes.

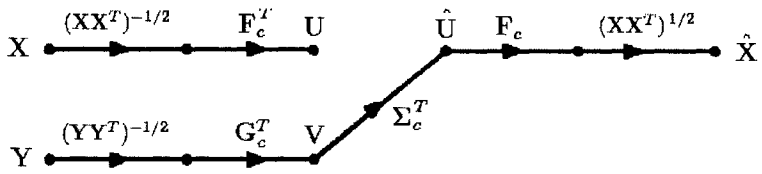

Figure 10.1: Classification in canonical coordinates.

3. **Building classifiers in canonical and half-canonical coordinates**

Referring to the two-channel filtering problem in Figure 10.1, assume that the sample data matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_M] \in \mathbb{R}^{n \times M}$ consists of a column-wise collection of feature vectors $\mathbf{y}_i \in \mathbb{R}^n$ for an $m-$class classification problem, and the sample data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M] \in \mathbb{R}^{m \times M}$ of a column-wise collection of class membership vectors $\mathbf{x}_i \in \mathbb{R}^m$. The class membership vector $\mathbf{x}_i$ is an $m-$dimensional vector with $(m-1)$ elements that are zero and one element that is equal to one, the position of which determines the class membership of the feature vector $\mathbf{y}_i$. Therefore, the filtering problem in Figure 10.1 is one of estimating the class membership vectors $\mathbf{x}_i$ from the feature vectors $\mathbf{y}_i$. Here, the matrices $\mathbf{F}_c$, $\mathbf{G}_c$, and $\mathbf{\Sigma}_c$ form the SVD of the sample coherence matrix $\mathbf{C} = (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-T/2} = \mathbf{F}_c\mathbf{\Sigma}_c\mathbf{G}_c^T$. The idea is to transform the feature vectors $\mathbf{y}_i$ to their canonical coordinates $\mathbf{v}_i = \mathbf{G}_c^T(\mathbf{Y}\mathbf{Y}^T)^{-1/2}\mathbf{y}_i$ and then use the developments in Chapter 4 to build a reduced-rank estimator of the class membership vector $\mathbf{x}_i$.

In practice, a training data set of feature vectors, with known class membership vectors, is available. Using this training set, we may find the matrices $\mathbf{F}_c$, $\mathbf{G}_c$, and $\mathbf{\Sigma}_c$. For a new feature vector with an unknown class membership, we may use these matrices, which are obtained based upon the training set, to build a reduced-rank estimator of the unknown class membership vector. As established in Chapter 4, this is the reduced-rank estimator with the smallest volume of the concentration ellipse. Alternatively, we may build a reduced-rank estimator

of the class membership vector in half-canonical coordinates, minimizing the MSE. The question is, which reduced-rank estimation measure, the volume of the concentration ellipse or the MSE, will yield a higher correct classification rate?

It has been shown [3] that in a two-class classification problem, the Fisher distance between the two classes is determined by the first canonical correlation between the matrix of feature vectors $\mathbf{Y}$, and the matrix of class membership vectors $\mathbf{X}$. This may be extended to a multi-class case. In fact, it may be shown that the multi-class Fisher distance can be decomposed into sum of the terms that are determined by the canonical correlations between the feature matrix $\mathbf{Y}$ and the class membership matrix $\mathbf{X}$. The multi-class Fisher distance [3] provides an overall distance measure between all the classes, but does not show how each class is separated from the others. Thus, it would be interesting to study the connection between the decomposition of multi-class Fisher distance in canonical coordinates and the pairwise two-class Fisher distances.

More generally, it would be interesting to study the connection between the classifiers built in canonical coordinates and half-canonical coordinates with some of the well-known classifiers, such as minimum distance [88], Bayes [88], support vector machines [33]– [35], least squares support vector machines [89], and large margin [33]– [35] classifiers.

4. **Beamforming in canonical coordinates**

Referring to the sensor array in Figure 10.2, assume that $\mathbf{x}_i$ and $\mathbf{y}_i$ are measurement vectors recorded, at the ith snapshot, by subarrays 1 and 2, respectively. The question is, can analysis of coherence between the data samples $\mathbf{x}_i$ and $\mathbf{y}_i$ be used for beamforming? The idea is to decompose the measurement vectors at each subarray into their canonical coordinates. The time series of the first

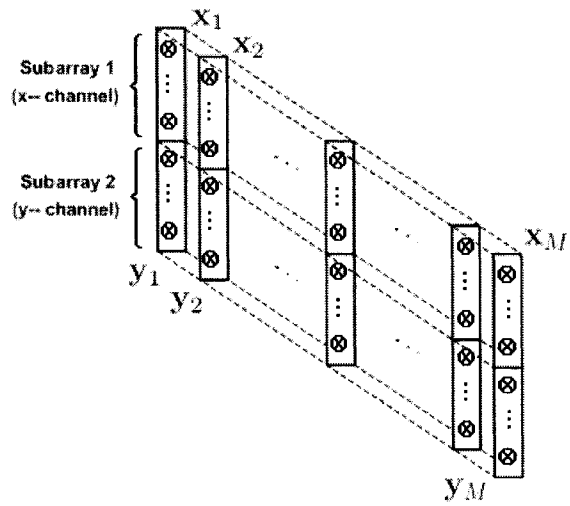

Figure 10.2: Beamforming in canonical coordinates

canonical coordinate pair, which contributes the most to the coherence between the subarrays, is then used as the beamformed versions of the data. Unlike the standard beamforming methods [90], this approach does not rely on the planar wave assumption. As a result, it may be less sensitive to deviations of propagating wavefront from planar wave and uncertainty in the array geometry, compared to the standard beamforming methods. Therefore, it would be very interesting to investigate this idea for beamforming and compare the properties of the canonical coordinate beamformers against those of the standard ones.

REFERENCES

[1] H. Hotelling, "Relations between two sets of variates," *Biometrica*, vol. 28, pp. 321–377, 1936.

[2] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis.* New York: Wiley, 1958.

[3] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, ch. 8-10. Academic Press, 1979.

[4] R. J. Muirhead, *Aspects of Multivariate Statistical Theory.* Wiley, 1982.

[5] M. L. Eaton, *Multivariate Statistics: A Vector Space Approach*, ch. 10. New York: Wiley, 1983.

[6] L. L. Scharf and J. T. Thomas, "Wiener filters in canonical coordinates for transform coding, filtering, and quantizing," *IEEE Trans. Signal Processing*, vol. 46, pp. 647–654, Mar. 1998.

[7] L. L. Scharf and C. T. Mullis, "Canonical coordinates and the geometry of inference, rate and capacity," *IEEE Trans. Signal Processing*, vol. 48, pp. 824–831, Mar. 2000.

[8] L. L. Scharf, *Statistical Signal Processing*, pp. 330–331. MA: Addison-Wesley, 1991.

[9] Y. Hua, M. Nikpour, and P. Stoica, "Optimal reduced-rank estimation and filtering," *IEEE Trans. Signal Processing*, vol. 49, pp. 457–469, Mar. 2001.

[10] L. L. Scharf, "The SVD and reduced rank signal processing," *Signal Processing*, vol. 25, pp. 113–133, 1991.

[11] D. R. Brillinger, *Time Series: Data Analysis and Theory.* SIAM, 2001.

[12] P. J. Schreier and L. L. Scharf, "Canonical coordinates for transform coding of random signals from noisy observations," submitted to *IEEE Trans. Signal Processing*, June 2003.

[13] A. Pezeshki, L. L. Scharf, M. R. Azimi-Sadjadi, and Y. Hua, "Two-channel constrained least squares problems: Solutions using power methods and connections with canonical coordinates," to appear *IEEE Trans. Signal Processing*, vol. 52, pp. 1–15, Dec. 2004.

[14] S. V. Schell and W. A. Gardner, "Programmable canonical correlation analysis: a flexible framework for blind adaptive spatial filtering," *IEEE Trans. Signal Processing*, vol. 43, pp. 2898–2908, Dec. 1995.

[15] M. F. Kahn, W. A. Gardner, and M. A. Mow, "Programmable canonical correlation analyzers with recursion and feedback," *Conf. Rec. Twenty-Ninth Asilomar Conf. Signals, Syst., Comput.*, vol. 1, pp. 351–356, Oct. 1995.

[16] M. F. Kahn and W. A. Gardner, "A time-channelized programmable canonical correlation analyzer," *Conf. Rec. Twenty-Ninth Asilomar Conf. Signals, Syst., Comput.*, vol. 1, pp. 346–350, Oct. 1995.

[17] W. A. Gardner, J. L. Schenck, and S. V. Schell, "Programmable blind adaptive spatial filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, pp. 19–22, April 1994.

[18] P. Stoica and M. Viberg, "Maximum likelihood parameter and rank estimation in reduced-rank multivariate linear regressions," *IEEE Trans. Signal Processing*, vol. 44, pp. 3069–3078, Dec. 1996.

[19] K. I. Diamantaras and S. Y. Kung, "Multilayer neural networks for reduced-rank approximation," *IEEE Trans. Neural Networks*, vol. 5, pp. 684–697, Sept. 1994.

[20] Y. Yamashita and H. Ogawa, "Relative Karhunen-Loeve transform," *IEEE Trans. Signal Processing*, vol. 44, pp. 371–378, Feb. 1996.

[21] Y. Hua and W. Liu, "Generalized Karhunen-Loeve transform," *IEEE Signal Processing Lett.*, vol. 5, pp. 141–142, June 1998.

[22] E. Oja, "A simplified neuron model as principal component analyzer," *J. Math. Biology*, vol. 15, pp. 267–273, 1982.

[23] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neural Networks*, vol. 2, pp. 459–473, 1989.

[24] P. Foldik, "Adaptive network for optimal linear feature extraction," in *Proc. Int. Joint Conf. Neural Networks*, vol. 1, pp. 401–405, June 1989.

[25] S. Y. Kung and K. I. Diamantaras, "Adaptive principal component extraction (APEX) and applications," *IEEE Trans. Signal Processing*, vol. 42, pp. 1202–1217, May 1994.

[26] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks: Theory and Applications.* Wiley, 1996.

[27] S. Bannour and M. R. Azimi-Sadjadi, "Principal component extraction using recursive least squares learning," *IEEE Trans. Neural Networks*, vol. 6, pp. 457–469, Mar. 1995.

[28] P. L. Lai and C. Fyfe, "A neural network implementation of canonical correlation analysis," *Neural Networks*, vol. 12, pp. 1391–1397, 1999.

[29] A. Pezeshki, M. R. Azimi-Sadjadi, and L. L. Scharf, "A network for recursive extraction of canonical coordinates," *Neural Networks*, vol. 16, pp. 801–808, July 2003.

[30] A. Pezeshki, M. R. Azimi-Sadjadi, and L. L. Scharf, "A canonical coordinate decomposition network," in *Proc. IEEE Int. Joint Conf. Neural Networks*, (Portland, OR), pp. 1313–1317, July 20-24 2003.

[31] S. Haykin, *Adaptive Filter Theory.* Upper Saddle River, NJ: Printice Hall, 1996.

[32] A. Pezeshki, L. L. Scharf, and M. R. Azimi-Sadjadi, "Empirical canonical coordinate decompositions in subspaces for two-channel linear and nonlinear maps," submitted to *IEEE Trans. Signal Processing*, Sept. 2004.

[33] V. N. Vapnik, *The Nature of Statistical Learning Theory.* Springer, 1995.

[34] V. N. Vapnik, *Statistical Learning Theory.* Wiley, 1998.

[35] B. Schölkopf and A. J. Smola, *Learning with Kernels, Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge, MA: MIT Press, 2002.

[36] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1998.

[37] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Proc. IEEE Neural Networks for Signal Processing Workshop*, 1999.

[38] A. Ruiz and P. E. López-de-Teruel, "Nonlinear kernel-based statistical pattern analysis," *IEEE Trans. Neural Networks*, vol. 12, pp. 16–32, Jan. 2001.

[39] T. Van Gestel, J. A. K. Suykens, J. De Brabanter, B. De Moor, and J. Vandewalle, "Kernel canonical correlation analysis and least squares support vector machines," in *Proc. Int. Conf. Artificial Neural Networks*, pp. 384–389, 2001.

[40] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *Int. J. Neural Systems*, vol. 10, pp. 365–377, 2000.

[41] T. Melzer, M. Reiter, and H. Bischof, "Nonlinear feature extraction using generalized canonical correlation analysis," in *Proc. Int. Conf. Artificial Neural Networks*, pp. 353–360, 2001.

[42] F. Bach and M. Jordan, "Kernel independent component analysis," *J. Machine Learning Res.*, vol. 3, pp. 1–48, 2002.

[43] A. Gretton, R. Herbrich, and A. J. Smola, "The kernel mutual information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, pp. 880–883, 2003.

[44] M. Kuss and T. Graepel, "The geometry of kernel canonical correlation analysis," Technical Report 108, Max Planck Institute for Biological Cybernetics, May 2003.

[45] M. Kuss, "Kernel multivariate analysis," Master's thesis, Technical University of Berlin, 2001.

[46] A. Pezeshki, M. R. Azimi-Sadjadi, and L. L. Scharf, "Kernel-based canonical coordinate decomposition of two-channel nonlinear maps," in *Proc. IEEE Int. Joint Conf. Neural Networks*, (Budapest, Hungary), pp. 3019–3024, July 25-29 2004.

[47] D. Cochran, H. Gish, and D. Sinno, "A geometric approach to multiple channel signal detection," *IEEE Trans. Signal Processing*, vol. 43, pp. 2049–2057, Sept. 1995.

[48] H. Gish and D. Cochran, "Generalized coherence," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, pp. 2745–2748, April 1987.

[49] A. Pezeshki, M. R. Azimi-Sadjadi, L. L. Scharf, and M. Robinson, "Underwater target classification using canonical correlations," in *Proc. MTS/IEEE Oceans'03*, (San Diego, CA), pp. 1906–1911, Sept. 22-26 2003.

[50] A. Pezeshki, M. R. Azimi-Sadjadi, and L. L. Scharf, "Classification of underwater mine-like and non-mine-like objects using canonical correlations," in *Detection and Remediation Technologies for Mines and Minelike Targets IX*, April 2004.

[51] ARL-UT, "Target data acquisition setup." Manual.

[52] M. Robinson, "Different multi-aspect fusion systems for underwater target classification," Master's thesis, Colorado State University, Fort Collins, CO, Feb. 2003.

[53] N. Cristianini, C. Campbell, and J. Shawe-Taylor, "Dynamically adapting kernels in support vector machines," in *Advances in Neural Information Processing Systems*, pp. 204–210, 1998.

[54] T. Friess, N. Cristianini, and C. Campbell, "The Kernel-Adatron: a fast and simple learning procedure for support vector machines," in *Proc. 15th Int. Conf. Machine Learning*, pp. 188–196, 1998.

[55] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," *J. Machine Learning Res.*, vol. 2, pp. 419–444, 2002.

[56] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, "On kernel-target alignment," in *Proc. Neural Information Processing Systems*, pp. 367–373, 2001.

[57] G. H. Golub and C. F. Van Loan, *Matrix Computations.* Baltimore, MD: John Hopkins Univ. Press, 3rd ed., 1996.

[58] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication.* IL: Univ. of Illinois Press, 1949.

[59] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrica*, vol. 1, pp. 211–218, 1936.

[60] L. Mirsky, "Symmetric gauge functions and unitary invaraint norms," *Quart. J. Math.*, vol. 11, pp. 50–59, 1960.

[61] S. Kayalar and H. L. Weinert, "Oblique projections: Formulas, algorithms and error bounds," *Math. Contr. Signals Syst.*, vol. 2, pp. 33–45, 1989.

[62] R. T. Behrens and L. L. Scharf, "Signal processing applications of oblique projection operators," *IEEE Trans. Signal Processing*, vol. 42, pp. 1413–1424, June 1994.

[63] G. H. Golub and W. Kahan, "Calculating the singular values and pseduo-inverse of a matrix," *SIAM J. Num. Anal.*, vol. 2, pp. 205–224, 1965.

[64] G. H. Golub, F. T. Luk, and M. Overton, "A block Lanczos method for computing the singular values and corresponding singular vectors of a matrix," *ACM Trans. Math. Soft.*, vol. 7, pp. 149–169, 1981.

[65] T. F. Chan, "An improved algorithm for computing the singular value decomposition," *ACM Trans. Math. Soft.*, vol. 8, pp. 72–83, 1982.

[66] G. H. Golub, "Tracking a few extreme singular values and vectors in signal processing," *Proc. IEEE*, pp. 1327–1343, Aug. 1990.

[67] D. W. Tufts and C. D. Mellissinos, "Simple, effective computation of principal eigenvectors and their eigenvalues and application to high-resolution estimation

of frequencies," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1046–1053, Oct. 1986.

[68] G. Strang, *Linear Algebra and Its Applications.* Orlando, FL: Academic Press, 2nd ed., 1980.

[69] J. H. Wilkinson, *The Algebraic Eigenvalue Problem.* Amen house, London: Oxford Univ. Press, 1965.

[70] Y. Hua and M. Nikpour, "Computing the reduced-rank Wiener filter by IQMD," *IEEE Signal Processing Lett.*, vol. 6, pp. 240–242, Sept. 1999.

[71] M. R. Azimi-Sadjadi, D. Yao, Q. Huang, and G. J. Dobeck, "Underwater target classification using wavelet packets and neural networks," *IEEE Trans. Neural Networks*, vol. 11, pp. 784–794, May 2000.

[72] D. Yao, M. R. Azimi-Sadjadi, A. A. Jamshidi, and G. J. Dobeck, "A study of effects of sonar bandwidth for underwater target classification," *IEEE J. Oceanic Eng.*, vol. 27, pp. 619–627, July 2002.

[73] M. R. Azimi-Sadjadi, D. Yao, A. A. Jamshidi, and G. Dobeck, "Underwater target classification in changing environments using adaptive feature mapping," *IEEE Trans. Neural Networks*, vol. 13, pp. 1099–1111, May 2002.

[74] D. Li, M. R. Azimi-Sadjadi, and M. Robinson, "Comparison of different classification algorithms for underwater target discrimination," *IEEE Trans. Neural Networks*, vol. 15, pp. 189–194, Jan. 2004.

[75] M. Robinson, M. R. Azimi-Sadjadi, D. D. Sternlicht, and D. Lemonds, "Multi-aspect acoustic classification of buried objects," in *Proc. MTS/IEEE Oceans'03*, pp. 478–484, 2003.

[76] N. Dasgupta, P. Runkle, L. Couchman, and L. Carin, "Dual hidden Markov model characterization of wavelet coefficients from multi-aspect scattering data," in *Detection and Remediation Technologies for Mines and Minelike Targets II*, vol. 4038, pp. 954–964, April 2000.

[77] L. L. Burton and H. Lai, "Active sonar target imaging and classification system," in *Detection and Remediation Technologies for Mines and Minelike Targets II*, vol. 3079, pp. 19–33, April 1997.

[78] G. A. Carpenter and W. W. Streilein, "ARTMAP–FTR: a neural network for fusion target recognition, with application to sonar classification," in *Detection and Remediation Technologies for Mines and Minelike Targets III*, vol. 3392, pp. 342–356, April 1998.

[79] C. F. Barnes, "Acoustic backscatter classification for mine detection using multiple fused aspects and novel database classification rules," in *Detection and Remediation Technologies for Mines and Minelike Targets III*, vol. 3392, pp. 357–368, April 1998.

[80] D. Casasent and N. Kuljanyavivat, "Mine detection from multiple acoustic backscatter data," in *Detection and Remediation Technologies for Mines and Minelike Targets III*, vol. 3392, pp. 370–381, April 1998.

[81] G. Okimoto and D. Lemonds, "Principal component analysis in the wavelet domain: new features for underwater object recognition," in *Detection and Remediation Technologies for Mines and Minelike Targets IV*, vol. 3710, pp. 697–708, April 1999.

[82] C. Yuan, M. R. Azimi-Sadjadi, J. Wilbur, and G. J. Dobeck, "Underwater target detection using multichannel subband adaptive filtering and high order correlation schemes," *IEEE J. Oceanic Eng.*, vol. 25, pp. 192–205, Jan. 2000.

[83] M. Hasan and M. R. Azimi-Sadjadi, "A modified block FTF adaptive algorithm with applications to underwater target detection," *IEEE Trans. Signal Processing*, vol. 44, pp. 2172–2185, Sept. 1996.

[84] M. Hasan, M. R. Azimi-Sadjadi, and G. J. Dobeck, "Multiple time delay estimation using new spectral estimation schemes," *IEEE Trans. Signal Processing*, vol. 46, pp. 1580–1590, June 1998.

[85] M. R. Azimi-Sadjadi, S. Charleston, J. Wilbur, and G. J. Dobeck, "A new time delay estimation in subbands for resolving multiple specular reflections," *IEEE Trans. Signal Processing*, vol. 46, pp. 3398–3403, Dec. 1998.

[86] S. Haykin, *Neural Networks: A Comprehensive Foundation.* Upper Saddle River, NJ: Printice Hall, 1996.

[87] S. G. Schock, A. Tellier, J. Wulf, J. Sara, and M. Ericksen, "Buried object scanning sonar," *IEEE J. Oceanic Eng.*, vol. 26, pp. 677–689, Oct. 2001.

[88] R. O. Duda, P. E. Hart, and D. Stork, *Pattern Classification.* Wiley Interscience, 2001.

[89] J. A. K. Suykens, T. Van Gestel, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines.* Singapore: World Scientific Pub. Co., 2002.

[90] H. L. Van Trees, *Optimum Array Processing.* Wiley Interscience, 2002.

APPENDIX A

SOLUTION TO THE DEFLATED COUPLED GENERALIZED EIGENVALUE PROBLEM

From (3.30) and (3.32), we have the thin SVD

$$\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} = \mathbf{R}_{xx}^{T/2}\mathbf{D}_x\mathbf{\Sigma}\mathbf{D}_y^T\mathbf{R}_{yy}^{1/2}. \tag{A.1}$$

Pre-multiplying (A.1) by $\mathbf{R}_{xx}^{1/2}$, post-multiplying it by $\mathbf{R}_{yy}^{T/2}$, and using $\mathbf{\Sigma} = \mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y$ yields

$$\mathbf{R}_{xy} = \mathbf{R}_{xx}\mathbf{D}_x\mathbf{\Sigma}\mathbf{D}_y^T\mathbf{R}_{yy} = \mathbf{R}_{xx}\mathbf{D}_x\mathbf{D}_x^T\mathbf{R}_{xy}\mathbf{D}_y\mathbf{D}_y^T\mathbf{R}_{yy}. \tag{A.2}$$

Assume that the first $r < m$ columns of $\mathbf{D}_x$ and $\mathbf{D}_y$ and their corresponding σ_i's have already been found. Rewrite (A.2) using (5.10) as

$$\mathbf{R}_{xy} = \mathbf{R}_{xx}\left[\begin{array}{cc}\mathbf{D}_{x,r} & \mathbf{D}_{x,\star}\end{array}\right]\left[\begin{array}{c}\mathbf{D}_{x,r}^T \\ \mathbf{D}_{x,\star}^T\end{array}\right]\mathbf{R}_{xy}\left[\begin{array}{cc}\mathbf{D}_{y,r} & \mathbf{D}_{y,\star}\end{array}\right]\left[\begin{array}{c}\mathbf{D}_{y,r}^T \\ \mathbf{D}_{y,\star}^T\end{array}\right]\mathbf{R}_{yy}. \tag{A.3}$$

Post-multiplying (A.3) by $\mathbf{D}_{y,\star}$ and recalling that $\mathbf{D}_y^T\mathbf{R}_{yy}\mathbf{D}_y = \mathbf{I}$ gives

$$\begin{aligned}\mathbf{R}_{xy}\mathbf{D}_{y,\star} &= \mathbf{R}_{xx}\left[\begin{array}{cc}\mathbf{D}_{x,r} & \mathbf{D}_{x,\star}\end{array}\right]\left[\begin{array}{c}\mathbf{D}_{x,r}^T \\ \mathbf{D}_{x,\star}^T\end{array}\right]\mathbf{R}_{xy}\left[\begin{array}{cc}\mathbf{D}_{y,r} & \mathbf{D}_{y,\star}\end{array}\right]\left[\begin{array}{c}\mathbf{0} \\ \mathbf{I}(\star)\end{array}\right] \\ &= \mathbf{R}_{xx}(\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T\mathbf{R}_{xy}\mathbf{D}_{y,\star} + \mathbf{D}_{x,\star}\mathbf{D}_{x,\star}^T\mathbf{R}_{xy}\mathbf{D}_{y,\star})\end{aligned} \tag{A.4}$$

where $\mathbf{I}(\star)$ is the $(m-r)\times(m-r)$ identity matrix. Rearranging (A.4) yields

$$(\mathbf{I} - \mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T)\mathbf{R}_{xy}\mathbf{D}_{y,\star} = \mathbf{R}_{xx}\mathbf{D}_{x,\star}\mathbf{D}_{x,\star}^T\mathbf{R}_{xy}\mathbf{D}_{y,\star} = \mathbf{R}_{xx}\mathbf{D}_{x,\star}\mathbf{\Sigma}(\star). \tag{A.5}$$

Similarly, starting with $\mathbf{C}^T = \mathbf{R}_{yy}^{-1/2}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-T/2}$ and following a similar procedure results in

$$(\mathbf{I} - \mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T)\mathbf{R}_{yx}\mathbf{D}_{x,\star} = \mathbf{R}_{yy}\mathbf{D}_{y,\star}\mathbf{\Sigma}(\star). \tag{A.6}$$

Equations (A.5) and (A.6) introduce a coupled asymmetric generalized eigenvalue problem for $\mathbf{D}_{x,\star}$, $\mathbf{D}_{y,\star}$, and $\mathbf{\Sigma}(\star)$, wherein $\mathbf{R}_{xy}$ is deflated by $\mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T\mathbf{R}_{xy}$ and $\mathbf{R}_{yx}$ by $\mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T\mathbf{R}_{yx}$. Thus, $\mathbf{d}_{x,r+1}$ and $\mathbf{d}_{y,r+1}$ are now the generalized eigenvectors associated with the dominant eigenvalue σ_{r+1} of (A.5) and (A.6).

APPENDIX B

PROOF OF DEFLATION IN (6.8)

Here we show that the minimization problem in (6.8) indeed formulates the problem of finding the $(r+1)$th pair of canonical coordinate mappings $\mathbf{d}_{x,r+1}$ and $\mathbf{d}_{y,r+1}$. The assumption is that the matrices $\mathbf{D}_{x,r}$ and $\mathbf{D}_{y,r}$, which contain the first r columns of $\mathbf{D}_x$ and $\mathbf{D}_y$ (the first r pairs of canonical coordinate mappings), have already been found.

Consider the thin SVD of the coherence matrix $\mathbf{C} = \mathbf{F}\mathbf{\Sigma}\mathbf{G}^T$ in (3.30). Partition $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{\Sigma}$ into

$$\mathbf{F} = [\mathbf{F}_r \ \mathbf{F}_\star], \quad \mathbf{G} = [\mathbf{G}_r \ \mathbf{G}_\star], \quad \text{and} \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}(r) & 0 \\ 0 & \mathbf{\Sigma}(\star) \end{bmatrix} \tag{B.1}$$

where the matrices $\mathbf{F}_r \in \mathbb{R}^{m\times r}$ and $\mathbf{G}_r \in \mathbb{R}^{n\times r}$ contain the first r and the matrices $\mathbf{F}_\star \in \mathbb{R}^{m\times(m-r)}$ and $\mathbf{G}_\star \in \mathbb{R}^{n\times(m-r)}$ the last $m-r$ columns of $\mathbf{F} \in \mathbb{R}^{m\times m}$ and $\mathbf{G} \in \mathbb{R}^{n\times m}$. The diagonal matrices $\mathbf{\Sigma}(r) = \text{diag}[\sigma_1, \ldots, \sigma_r]$ and $\mathbf{\Sigma}(\star) = \text{diag}[\sigma_{r+1}, \ldots, \sigma_m]$, respectively, contain the first r and the last $m-r$ canonical correlations. Then, we may rewrite the thin SVD in (3.30) as

$$\mathbf{C} = \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2} = \mathbf{F}\mathbf{\Sigma}\mathbf{G}^T = \mathbf{F}_r\mathbf{\Sigma}(r)\mathbf{G}_r^T + \mathbf{F}_\star\mathbf{\Sigma}(\star)\mathbf{G}_\star^T, \tag{B.2}$$

and

$$\begin{bmatrix} \mathbf{F}_r^T\mathbf{F}_r & \mathbf{F}_\star^T\mathbf{F}_r \\ \mathbf{F}_r^T\mathbf{F}_\star & \mathbf{F}_\star^T\mathbf{F}_\star \end{bmatrix} = \begin{bmatrix} \mathbf{I}(r) & 0 \\ 0 & \mathbf{I}(\star) \end{bmatrix} \text{ and } \begin{bmatrix} \mathbf{G}_r^T\mathbf{G}_r & \mathbf{G}_\star^T\mathbf{G}_r \\ \mathbf{G}_r^T\mathbf{G}_\star & \mathbf{G}_\star^T\mathbf{G}_\star \end{bmatrix} = \begin{bmatrix} \mathbf{I}(r) & 0 \\ 0 & \mathbf{I}(\star) \end{bmatrix} \tag{B.3}$$

where $\mathbf{I}(r)$ and $\mathbf{I}(\star)$ are the $r \times r$ and $(m-r) \times (m-r)$ identity matrices. The SVD in (B.2), may be rewritten as

$$\mathbf{R}_{xy} = \mathbf{R}_{xx}^{1/2}\mathbf{F}_r\mathbf{\Sigma}(r)\mathbf{G}_r^T\mathbf{R}_{yy}^{T/2} + \mathbf{R}_{xx}^{1/2}\mathbf{F}_\star\mathbf{\Sigma}(\star)\mathbf{G}_\star^T\mathbf{R}_{yy}^{T/2}. \tag{B.4}$$

Using (B.2) and (B.3), it may easily be verified that the first term on the right hand side of (B.4) has three equivalent representations. That is,

$$\begin{aligned} \mathbf{R}_{xx}^{1/2}\mathbf{F}_r\mathbf{\Sigma}_r\mathbf{G}_r^T\mathbf{R}_{yy}^{1/2} &= \mathbf{R}_{xx}^{1/2}\mathbf{F}_r\mathbf{F}_r^T\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy} \\ &= \mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}\mathbf{G}_r\mathbf{G}_r^T\mathbf{R}_{yy}^{T/2} \\ &= \mathbf{R}_{xx}^{1/2}\mathbf{F}_r\mathbf{F}_r^T\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-T/2}\mathbf{G}_r\mathbf{G}_r^T\mathbf{R}_{yy}^{T/2}. \end{aligned} \tag{B.5}$$

Using this property, we may rewrite (B.4) as

$$(\mathbf{I} - \mathbf{R}_{xx}^{1/2}\mathbf{F}_r\mathbf{F}_r^T\mathbf{R}_{xx}^{-1/2})\mathbf{R}_{xy}(\mathbf{I} - \mathbf{R}_{yy}^{1/2}\mathbf{G}_r\mathbf{G}_r^T\mathbf{R}_{yy}^{-1/2})^T = \mathbf{R}_{xx}^{1/2}\mathbf{F}_\star\mathbf{\Sigma}(\star)\mathbf{G}_\star^T\mathbf{R}_{yy}^{T/2}. \tag{B.6}$$

We now partition $\mathbf{D}_x$ and $\mathbf{D}_y$ into $\mathbf{D}_x = [\mathbf{D}_{x,r}\ \ \mathbf{D}_{x,\star}]$ and $\mathbf{D}_y = [\mathbf{D}_{y,r}\ \ \mathbf{D}_{y,\star}]$, where the matrices $\mathbf{D}_{x,\star} = [\mathbf{d}_{x,r+1}, \ldots, \mathbf{d}_{x,m}]$ and $\mathbf{D}_{y,\star} = [\mathbf{d}_{y,r+1}, \ldots, \mathbf{d}_{y,m}]$ contain the last $m-r$ columns of $\mathbf{D}_x$ and $\mathbf{D}_y$. Then, from (3.32) and (B.1), we have

$$\begin{aligned} \mathbf{F}_r &= \mathbf{R}_{xx}^{T/2}\mathbf{D}_{x,r}, \quad \mathbf{F}_\star = \mathbf{R}_{xx}^{T/2}\mathbf{D}_{x,\star}, \\ \mathbf{G}_r &= \mathbf{R}_{yy}^{T/2}\mathbf{D}_{y,r}, \quad \mathbf{G}_\star = \mathbf{R}_{yy}^{T/2}\mathbf{D}_{y,\star}. \end{aligned} \tag{B.7}$$

Using (B.7), we may rewrite (B.6) as

$$(\mathbf{I} - \mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T)\mathbf{R}_{xy}(\mathbf{I} - \mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T)^T = \mathbf{R}_{xx}\mathbf{D}_{x,\star}\mathbf{\Sigma}(\star)\mathbf{D}_{y,\star}^T\mathbf{R}_{yy}. \tag{B.8}$$

Pre-multiplying (B.8) by $\mathbf{d}_{x,r+1}^T$, post-multiplying it by $\mathbf{d}_{y,r+1}$, and using $\mathbf{d}_{x,r+1}^T\mathbf{R}_{xx}\mathbf{D}_{x,\star}$ $= [1, 0, \ldots, 0]$ and $\mathbf{d}_{y,r+1}^T\mathbf{R}_{yy}\mathbf{D}_{y,\star} = [1, 0, \ldots, 0]$ results in

$$\mathbf{d}_{x,r+1}^T(\mathbf{I} - \mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T)\mathbf{R}_{xy}(\mathbf{I} - \mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T)^T\mathbf{d}_{y,r+1} = \sigma_{r+1}. \tag{B.9}$$

Considering that σ_{r+1} is the largest diagonal element of $\mathbf{\Sigma}(\star)$, we may formulate the problem of finding $\mathbf{d}_{x,r+1}$, $\mathbf{d}_{y,r+1}$ as

$$\max_{\mathbf{d}_{x,r+1},\mathbf{d}_{y,r+1}} \mathbf{d}_{x,r+1}^T(\mathbf{I}-\mathbf{R}_{xx}\mathbf{D}_{x,r}\mathbf{D}_{x,r}^T)\mathbf{R}_{xy}(\mathbf{I}-\mathbf{R}_{yy}\mathbf{D}_{y,r}\mathbf{D}_{y,r}^T)^T\mathbf{d}_{y,r+1}$$

subject to the constraints

$$\mathbf{d}_{x,r+1}^T\mathbf{R}_{xx}\mathbf{d}_{x,r+1}=1 \quad \text{and} \quad \mathbf{d}_{y,r+1}^T\mathbf{R}_{yy}\mathbf{d}_{y,r+1}=1.$$

APPENDIX C

PROOF OF MAXIMAL INVARIANCE PROPERTY OF EMPIRICAL CANONICAL CORRELATIONS

We first prove that the nonzero empirical canonical correlations of the composite two-channel data matrix $\mathbf{Z} = [\mathbf{X}^T \ \mathbf{Y}^T]^T$ are identical to those of the composite two-channel data matrix $\tilde{\mathbf{Z}}$, defined as

$$\tilde{\mathbf{Z}} = \begin{bmatrix} \tilde{\mathbf{X}} \\ \tilde{\mathbf{Y}} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_x(:, 1:p)^T & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_y(:, 1:q)^T \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}. \tag{C.1}$$

The rows of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ may be viewed as experimental surrogates for the principal components of $\mathbf{x}$ and $\mathbf{y}$.

The coherence matrix for $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ may be written as

$$\begin{aligned} \tilde{\mathbf{C}} &= (\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T)^{-1/2}\tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T)^{-T/2} \\ &= (\mathbf{U}_x(:, 1:p)^T\mathbf{X}\mathbf{X}^T\mathbf{U}_x(:, 1:p))^{-1/2}\mathbf{U}_x(:, 1:p)^T\mathbf{X} \\ &\quad \times\mathbf{Y}^T\mathbf{U}_y(:, 1:q)(\mathbf{U}_y(:, 1:q)^T\mathbf{Y}\mathbf{Y}^T\mathbf{U}_y(:, 1:q))^{-T/2}. \end{aligned} \tag{C.2}$$

Using $\mathbf{X} = \mathbf{U}_x\mathbf{\Sigma}_x\mathbf{V}_x^T$ and $\mathbf{Y} = \mathbf{U}_y\mathbf{\Sigma}_y\mathbf{V}_y^T$, with $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_y$ as in (7.9), we may simplify

(C.2) to

$$\begin{aligned}\tilde{\mathbf{C}} &= \boldsymbol{\Sigma}_x(p)^{-1}\begin{bmatrix}\boldsymbol{\Sigma}_x(p) & \mathbf{0}\end{bmatrix}\mathbf{V}_x^T\mathbf{V}_y\begin{bmatrix}\boldsymbol{\Sigma}_y(q)\\ \mathbf{0}\end{bmatrix}\boldsymbol{\Sigma}_y(q)^{-1} \\ &= \begin{bmatrix}\mathbf{I}(p) & \mathbf{0}\end{bmatrix}\mathbf{V}_x^T\mathbf{V}_y\begin{bmatrix}\mathbf{I}(q) & \mathbf{0}\end{bmatrix}^T = \mathbf{V}_x(:,1:p)^T\mathbf{V}_y(1,:q)\end{aligned} \tag{C.3}$$

However, from the SVD in (7.17), singular values of $\mathbf{V}_x(:,1:p)^T\mathbf{V}_y(:,1:q)$ are the empirical canonical correlations σ_i. This establishes that the non-zero empirical canonical correlations of $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are identical.

Consider the transformation group in (7.20). Each element of group $\mathcal{T}$ may be decomposed as

$$\mathbf{T} = \begin{bmatrix}\mathbf{T}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{T}_2\end{bmatrix}\begin{bmatrix}\mathbf{U}_x(:,1:p)^T & \mathbf{0}\\ \mathbf{0} & \mathbf{U}_y(:,1:q)^T\end{bmatrix}. \tag{C.4}$$

The second matrix on the right hand side of (C.4) first transforms the data matrices $\mathbf{X}$ and $\mathbf{Y}$ to the matrices $\tilde{\mathbf{X}} = \mathbf{U}_x(:,1:p)^T\mathbf{X}$ and $\tilde{\mathbf{Y}} = \mathbf{U}_y(:,1:q)^T\mathbf{Y}$. The matrices $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are then transformed into $\mathbf{T}_1\tilde{\mathbf{X}}$ and $\mathbf{T}_2\tilde{\mathbf{Y}}$, using nonsingular transformations $\mathbf{T}_1$ and $\mathbf{T}_2$. The nonzero empirical canonical correlations of $\tilde{\mathbf{X}} \in \mathbb{R}^{p\times M}$ and $\tilde{\mathbf{Y}} \in \mathbb{R}^{q\times M}$, are maximal invariants for the composite covariance matrix $\tilde{\mathbf{S}}_{zz} = \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T$ under the transformation group of (2.14). However, the nonzero empirical canonical correlations of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are the same as those of $\mathbf{X}$ and $\mathbf{Y}$. Therefore, the empirical canonical correlations σ_i are maximal invariants for $\mathbf{S}_{zz} = \mathbf{Z}\mathbf{Z}^T$ under the transformation group of (7.20).

APPENDIX D

PROOF OF EQUATION (7.28)

Using the SVD of $\mathbf{X}$, i.e. $\mathbf{X} = \mathbf{U}_x \mathbf{\Sigma}_x \mathbf{V}_x^T$, we may write (7.25) as

$$\mathbf{D}_x = (\mathbf{X}\mathbf{X}^T)^{-T/2}\mathbf{F}_c = (\mathbf{U}_x \mathbf{\Sigma}_x \mathbf{\Sigma}_x^T \mathbf{U}_x^T)^{-T/2}\mathbf{F}_c. \tag{D.1}$$

Note that in general, $\mathbf{X}$ is rank-deficient and hence the inverse in (D.1) is a pseudo-inverse. Since $\mathbf{U}_x$ is an orthogonal matrix, we may simplify (D.1) to

$$\mathbf{D}_x = \mathbf{U}_x(\mathbf{\Sigma}_x \mathbf{\Sigma}_x^T)^{-T/2}\mathbf{U}_x^T\mathbf{F}_c = \mathbf{U}_x\mathbf{Q} \tag{D.2}$$

with $\mathbf{Q} = (\mathbf{\Sigma}_x \mathbf{\Sigma}_x^T)^{-T/2}\mathbf{U}_x^T\mathbf{F}_c$. Since $\mathbf{\Sigma}_x$ is of form (7.9), Col-Span$\{\mathbf{D}_x\}$ = Col-Span $\{\mathbf{U}_x(:, 1:p)\}$. However, from the SVD of $\mathbf{X}$, Col-Span$\{\mathbf{U}_x(:, 1:p)\}$ = Col-Span$\{\mathbf{X}\}$, and thereby Col- Span$\{\mathbf{D}_x\}$ = Col-Span$\{\mathbf{X}\}$.