INFINITE DIMENSIONAL DISCRIMINATION AND CLASSIFICATION

A Dissertation

by

HYEJIN SHIN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2006

Major Subject: Statistics

INFINITE DIMENSIONAL DISCRIMINATION AND CLASSIFICATION

A Dissertation

by

HYEJIN SHIN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Randall L. Eubank
	Emanuel Parzen
Committee Members,	Marc G. Genton
	Joseph D. Ward
Head of Department,	Simon J. Sheather

May 2006

Major Subject: Statistics

ABSTRACT

Infinite Dimensional Discrimination and Classification. (May 2006)

Hyejin Shin, B.S., Chonnam National University;

M.S., Seoul National University

Co-Chairs of Advisory Committee: Dr. Randall L. Eubank
Dr. Emanuel Parzen

Modern data collection methods are now frequently returning observations that should be viewed as the result of digitized recording or sampling from stochastic processes rather than vectors of finite length. In spite of great demands, only a few classification methodologies for such data have been suggested and supporting theory is quite limited. The focus of this dissertation is on discrimination and classification in this infinite dimensional setting. The methodology and theory we develop are based on the abstract canonical correlation concept of Eubank and Hsing (2005), and motivated by the fact that Fisher's discriminant analysis method is intimately tied to canonical correlation analysis. Specifically, we have developed a theoretical framework for discrimination and classification of sample paths from stochastic processes through use of the Loève-Parzen isomorphism that connects a second order process to the reproducing kernel Hilbert space generated by its covariance kernel. This approach provides a seamless transition between the finite and infinite dimensional settings and lends itself well to computation via smoothing and regularization. In addition, we have developed a new computational procedure and illustrated it with simulated data and Canadian weather data.

To my parents, grandparents and husband

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisors, Dr. Randy Eubank and Dr. Emanuel Parzen, for their continuous support and encouragement throughout the years. I am especially grateful to Dr. Randy Eubank. This dissertation would not have been possible without his guidance and support. I also would like to extend my thanks to my committee members, Dr. Marc Genton and Dr. Joe Ward, for their valuable suggestions and comments. Many thanks also to Dr. Dahm, Dr. Longnecker and Marilyn Randall for their kindness and help.

I also thank my friends in College Station for their help and friendship. Finally, I give my greatest appreciation to my parents, Hyunwoo Shin and Seoksoon Jung, who have loved and supported me immeasurably. I sincerely thank my grandparents, brothers, sister and sister-in-law for their tremendous support. My special thanks go to Seokho, who is not only my best friend but also my strongest supporter in my whole life. I am also thankful to my parents-in-law.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

Discrimination methods for data classification are one of the most widely used statistical tools in various fields. Traditional statistical methods for solving discrimination problems include linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), multiple logistic regression, nearest neighbor methods, nonparametric function estimation methods, classification trees and neural network classifiers. In recent years, several techniques have been proposed for analyzing observations with more complex structure (e.g., see Hastie et al., 2001).

In many real-life situations, observed data are continuous functions sampled at discrete points. In that case, we should view the observations as the result of digitized recording or sampling from a stochastic process rather than vectors of finite length. However, most current classification methods ignore the inherent nature of functional type data and simply treat it as readings on a high dimensional multivariate vector.

Recent work that actually treats functional data from a random curve perspective includes Hall, Poskitt, and Presnell (2001). They studied signal discrimination using finite-dimensional basis representations and then employ classical discrimination methods like nonparametric kernel methods, LDA and QDA on the basis coefficients. In a similar vein, James and Hastie (2001) proposed likelihood-based functional linear discriminant analysis treating the observations as samples from underlying smoothed curves.

The focus of this dissertation will be on the formulation of a reliable discrimination

---

The format and style follow that of *Biometrics*.

method specially developed from the idea of Fisher's discrimination approach for classifying functions. Our motivation arises from the fact that Fisher's discriminant analysis method is intimately tied to canonical correlation analysis.

Most of the methodologies for functional data parallel those in multivariate data analysis. Accordingly, in Chapter II we will start with the concepts of classical multivariate canonical correlation analysis and discriminant analysis. The ideas that underly discriminant analysis detailed in Chapter VI have their roots in discriminant analysis where covariance matrices are less than full rank. So, we will first investigate theory of canonical correlation analysis and discriminant analysis in the finite dimensional, less than full rank, scenario in Chapter III.

The formulation of canonical correlation analysis and discriminant analysis in the infinite dimensional setting requires a background in functional analysis and the theory of reproducing kernel Hilbert space. We therefore briefly summarize the mathematical preliminaries that are needed for Chapters V–VI in Chapter IV. As we emphasized before, this research is motivated by the fact that Fisher's discriminant analysis and canonical correlation are connected with each other. Thus, we study the abstract canonical correlation concept in Eubank and Hsing (2005) in Chapter V. We will then solve the Fisher's discriminant problem in the infinite dimensional setting and develop the computational algorithm for its application to simulated data and real data in Chapter VI. Chapter VII provides a summary of the results in this dissertation. Some remaining questions are also posed for future research.

CHAPTER II

REVIEW OF SELECTED LITERATURE

We begin with this chapter with an overview of the classical multivariate canonical correlation analysis and discriminant analysis. Then some more current developments in canonical correlation analysis and discriminant analysis for functional data that are germane for subsequent developments are considered.

2.1 Finite Dimensional Canonical Correlation Analysis

Canonical correlation analysis (CCA, hereafter) is a classical multivariate method that is employed for situations where each subject in a sample is measured on two sets of random variables. The goal of this methodology is to provide an understanding of the relationships between the two sets of variables.

CCA was initially developed by Hotelling (1936) as the answer to a problem of finding the linear combination of a set of variables which is most highly correlated with any linear combination of another set of variables. Several generalizations of canonical correlation analysis to $k > 2$ sets of random variables were proposed by Kettenring (1971). Extensions of CCA to time series were developed by Jewell and Bloomfield (1983), Tsay and Tiao (1985) and Tiao and Tsay (1989). Also, Leurgans, Moyeed, and Silverman (1993), Ramsay and Silverman (1997), and He, Müller, and Wang (2002) extended CCA to functional data analysis. A general and unified notion of CCA has been developed by Eubank and Hsing (2005) whose work will be reviewed in Chapter V.

2.1.1 Population Canonical Correlations and Canonical Variables

In this section we provide a discussion of the classical multivariate canonical correlation analysis concept. In what follows bold letters will be used for matrices and column vectors.

Let $\mathbf{X}$ be a p-dimensional random vector and let $\mathbf{Y}$ be a q-dimensional random vector with $\text{Var}(\mathbf{X}) = \mathbf{K}_X,\ \text{Var}(\mathbf{Y}) = \mathbf{K}_Y$, and $\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{K}_{XY} = \mathbf{K}_{YX}^T$. Assume that both $\mathbf{K}_X$ and $\mathbf{K}_Y$ are positive definite.

Now, given $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$ consider the linear combinations $\mathbf{a}^T\mathbf{X}$ and $\mathbf{b}^T\mathbf{Y}$. The squared correlation between these two random variables is

$$\rho^2(\mathbf{a}, \mathbf{b}) = \frac{\text{Cov}^2\left(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}\right)}{\text{Var}\left(\mathbf{a}^T\mathbf{X}\right)\text{Var}\left(\mathbf{b}^T\mathbf{Y}\right)} = \frac{(\mathbf{a}^T\mathbf{K}_{XY}\mathbf{b})^2}{(\mathbf{a}^T\mathbf{K}_X\mathbf{a})(\mathbf{b}^T\mathbf{K}_Y\mathbf{b})} \tag{2.1}$$

provided that $\mathbf{a} \neq \mathbf{0}$ and $\mathbf{b} \neq \mathbf{0}$. Then we may ask what values of $\mathbf{a}$ and $\mathbf{b}$ maximize (2.1). Equivalently we can solve the problem

$$\max_{\mathbf{a}\neq\mathbf{0},\mathbf{b}\neq\mathbf{0}} \text{Cov}^2(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}) \tag{2.2}$$

subject to

$$\text{Var}(\mathbf{a}^T\mathbf{X}) = \text{Var}(\mathbf{b}^T\mathbf{Y}) = 1. \tag{2.3}$$

Now define the first canonical correlation ρ_1 and the associated weight vectors $\mathbf{a}_1, \mathbf{b}_1$ as

$$\rho_1^2 = \text{Cov}^2\left(\mathbf{a}_1^T\mathbf{X}, \mathbf{b}_1^T\mathbf{Y}\right) = \max_{\mathbf{a}\neq\mathbf{0},\mathbf{b}\neq\mathbf{0}} \text{Cov}^2\left(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}\right), \tag{2.4}$$

where $\mathbf{a}, \mathbf{b}$ are subject to (2.3). Similarly, for $i > 1$, the ith canonical correlation ρ_i and associated weight vectors $\mathbf{a}_i, \mathbf{b}_i$ are defined by

$$\rho_i^2 = \text{Cov}^2\left(\mathbf{a}_i^T\mathbf{X}, \mathbf{b}_i^T\mathbf{Y}\right) = \max_{\mathbf{a}\neq\mathbf{0},\mathbf{b}\neq\mathbf{0}} \text{Cov}^2\left(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}\right), \tag{2.5}$$

where $\mathbf{a}, \mathbf{b}$ are subject to (2.3) and

$$\text{Cov}\left(\mathbf{a}^T\mathbf{X}, \mathbf{a}_j^T\mathbf{Y}\right) = \text{Cov}\left(\mathbf{b}^T\mathbf{X}, \mathbf{b}_j^T\mathbf{Y}\right) = 0, \quad j < i. \tag{2.6}$$

Explicit formulae for the canonical correlations and variables can be obtained as follows. Let

$$\mathbf{u} = \mathbf{K}_X^{1/2}\mathbf{a}$$

and set

$$\mathbf{v} = \mathbf{K}_Y^{1/2}\mathbf{b}.$$

Then, solving problem (2.2) and (2.3) is equivalent to solving the problem

$$\max_{\substack{\mathbf{u}\neq\mathbf{0},\mathbf{v}\neq\mathbf{0} \\ ||\mathbf{u}||_{\mathbb{R}^p}=||\mathbf{v}||_{\mathbb{R}^q}=1}} (\mathbf{u}^T\mathbf{K}_X^{-1/2}\mathbf{K}_{XY}\mathbf{K}_Y^{-1/2}\mathbf{v})^2, \tag{2.7}$$

where $\|\cdot\|_{\mathbb{R}^p}$ is the standard Euclidean norm. But, using the singular value decomposition (SVD) of a matrix, $\mathbf{K}_X^{-1/2}\mathbf{K}_{XY}\mathbf{K}_Y^{-1/2}$ can be written in the form

$$\mathbf{K}_X^{-1/2}\mathbf{K}_{XY}\mathbf{K}_Y^{-1/2} = \sum_{i=1}^{\min(p,q)} \rho_i \mathbf{u}_i \mathbf{v}_i^T,$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the eigenvectors of

$$\mathbf{K}_X^{-1/2}\mathbf{K}_{XY}\mathbf{K}_Y^{-1}\mathbf{K}_{YX}\mathbf{K}_X^{-1/2} \text{ and } \mathbf{K}_Y^{-1/2}\mathbf{K}_{YX}\mathbf{K}_X^{-1}\mathbf{K}_{XY}\mathbf{K}_Y^{-1/2},$$

respectively, corresponding to the eigenvalues $\rho_1^2, \ldots, \rho_{\min(p,q)}^2$.

Suppose that $\rho_1^2 \geq \cdots \geq \rho_{\min(p,q)}^2 > 0$. Then, $\mathbf{a}_i = \mathbf{K}_X^{-1/2}\mathbf{u}_i$ and $\mathbf{b}_i = \mathbf{K}_Y^{-1/2}\mathbf{v}_i$ solve problem (2.5) subject to (2.3) and (2.6) with corresponding canonical correlation ρ_i. Note that $\mathbf{a}_i$ and $\mathbf{b}_i$ can be obtained directly from

$$\mathbf{K}_X^{-1}\mathbf{K}_{XY}\mathbf{K}_Y^{-1}\mathbf{K}_{YX}\mathbf{a}_i = \rho_i^2\mathbf{a}_i, \tag{2.8}$$

and

$$\mathbf{K}_Y^{-1}\mathbf{K}_{YX}\mathbf{K}_X^{-1}\mathbf{K}_{XY}\mathbf{b}_i = \rho_i^2\mathbf{b}_i. \tag{2.9}$$

2.1.2 Sample Canonical Correlations and Canonical Variables

Suppose that we observe N iid copies $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_N, \mathbf{Y}_N)$ of $(\mathbf{X}, \mathbf{Y})$. We now consider the sample-based counterpart of the developments in the previous section. For this purpose, we estimate the population variances and covariances by their corresponding sample moments producing the matrices

$$\widehat{\mathbf{K}}_X = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T \quad \text{and} \quad \widehat{\mathbf{K}}_Y = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T \tag{2.10}$$

and

$$\widehat{\mathbf{K}}_{XY} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T, \tag{2.11}$$

where $\bar{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i$ and $\bar{\mathbf{Y}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i$.

Similar to the definitions in the population setting of the previous section, we now take the ith sample canonical variables to be

$$\hat{\mathbf{a}}_i^T\mathbf{X} \quad \text{and} \quad \hat{\mathbf{b}}_i^T\mathbf{Y}$$

with $\hat{\mathbf{a}}_i = \widehat{\mathbf{K}}_X^{-1/2}\hat{\mathbf{u}}_i$ and $\hat{\mathbf{b}}_i = \widehat{\mathbf{K}}_Y^{-1/2}\hat{\mathbf{v}}_i$ for $\hat{\mathbf{u}}_i$ and $\hat{\mathbf{v}}_i$ the eigenvectors of

$$\widehat{\mathbf{K}}_X^{-1/2}\widehat{\mathbf{K}}_{XY}\widehat{\mathbf{K}}_Y^{-1}\widehat{\mathbf{K}}_{YX}\widehat{\mathbf{K}}_X^{-1/2}$$

and

$$\widehat{\mathbf{K}}_Y^{-1/2}\widehat{\mathbf{K}}_{YX}\widehat{\mathbf{K}}_X^{-1}\widehat{\mathbf{K}}_{XY}\widehat{\mathbf{K}}_Y^{-1/2}$$

corresponding to the eigenvalues $\hat{\rho}_1^2 \geq \cdots \geq \hat{\rho}_{\min(p,q)}^2 > 0$. The corresponding estimated ith canonical correlation is $\hat{\rho}_i$.

2.2 Finite Dimensional Discriminant Analysis

The focus of this dissertation is on classification via discriminant analysis. The two standard multivariate methods for discrimination are the Bayesian approach and Fisher's linear discriminant analysis. The latter method is intimately tied to canonical correlation analysis.

In this section we provide a review of multivariate discriminant analysis methods. In particular, we will detail the relationship between Fisher's approach and the canonical correlation analysis technique for discrimination.

Let us now consider a discrimination problem with J classes or populations. We observe $(\mathbf{X}, G)$, where $\mathbf{X} \in \mathbb{R}^p$ is a predictor vector and $G \in \{1, \ldots, J\}$ is a categorical response variable representing the class memberships. We are interested in predicting the class membership G based on the p variables in the vector of predictors $\mathbf{X}$. This is an important practical problem with applications in many fields.

Suppose that class j has the density f_j with the class mean $\boldsymbol{\mu}_j$, covariance matrix $\mathbf{K}_j$ and associated class probability π_j. That is,

$$\mathrm{E}[\mathbf{X}|G = j] = \boldsymbol{\mu}_j,$$

$$\mathrm{Var}(\mathbf{X}|G = j) = \mathrm{E}[(\mathbf{X} - \boldsymbol{\mu}_j)(\mathbf{X} - \boldsymbol{\mu}_j)^T|G = j] = \mathbf{K}_j$$

and $P(G = j) = \pi_j$. Under this formulation there are two basic approaches to the development of discrimination methods: a Bayesian classifier and Fisher's method. We discuss each of these methods, in turn, below.

2.2.1 Bayes Procedure: Linear Discriminant Analysis

Assume that the density of class j is normal with mean $\boldsymbol{\mu}_j$ and a common within class covariance matrix $\mathbf{K}_W$: i.e., $\mathbf{K}_j = \mathbf{K}_W$ for $j = 1, \ldots, J$. Also, assume that $\mathbf{K}_W$ is positive definite.

A Bayesian classifier assigns an observation to the group with the largest posterior probability. Then, the Bayes linear discriminant rule allocates an observation $\mathbf{x}$ to the class for which

$$d_j(\mathbf{x}) = \boldsymbol{\mu}_j^T \mathbf{K}_W^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_j^T \mathbf{K}_W^{-1} \boldsymbol{\mu}_j + \log \pi_j \tag{2.12}$$

is maximized. In the case where we have equal class probabilities, an observation is classified to the class with the smallest squared Mahalanobis distance

$$(\mathbf{x} - \boldsymbol{\mu}_j)^T \mathbf{K}_W^{-1} (\mathbf{x} - \boldsymbol{\mu}_j).$$

2.2.2 Bayes Procedure: Quadratic Discriminant Analysis

The linear discriminant functions in (2.12) create linear boundaries which lead to a simple and easily implementable classification rule. However, these discriminant functions can perform badly when the assumption of a common covariance matrix is not true and often linear decision boundaries do not adequately separate the classes.

Thus, let us allow for different covariance matrices $\mathbf{K}_1, \ldots, \mathbf{K}_J$ for each class with $\mathbf{K}_j$ being positive definite for each $j = 1, \ldots, J$. Then, the Bayes quadratic discriminant rule allocates an observation $\mathbf{x}$ to the class which minimizes

$$d_j^Q(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_j)^T \mathbf{K}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) + \log |\mathbf{K}_j| - 2 \log \pi_j.$$

Quadratic discriminant analysis (QDA) provides more complex decision boundaries and often leads to a classification rule that performs better than the discriminant functions obtained from a linear classifier.

2.2.3 Fisher's Linear Discriminant Analysis

Fisher's linear discriminant analysis is a popular data analytic tool for studying the relationship between a set of predictors and a categorical response as well as a prevalent dimensional reduction tool. The primary purpose of Fisher's discriminant analysis is to separate classes. So we now use this perspective to formulate discriminant functions and to build a corresponding rule for predicting class membership of new observations.

Fisher's approach employs only second order properties of the random variables. Thus, unlike the Bayesian development, it is not necessary to assume any particular para-

metric form for the distribution of the J classes. However, we do assume that, as for LDA, the classes have a common (within-class) covariance matrix $\mathbf{K}_W$.

2.2.3.1 Fisher's linear discriminant function

Fisher's linear discriminant function is defined to be the linear function $\boldsymbol{l}^T\mathbf{X}$ which maximizes the ratio of the between-class variance to the within-class variance. Specifically, let $\mathbf{K}_B$ be the between-class covariance matrix defined by

$$\mathbf{K}_B = \text{Var}_G(\text{E}[\mathbf{X}|G]) = \sum_{j=1}^{J} \pi_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T$$

for

$$\boldsymbol{\mu} = \text{E}[\mathbf{X}] = \text{E}_G[\text{E}(\mathbf{X}|G)] = \sum_{j=1}^{J} \pi_j \boldsymbol{\mu}_j,$$

and similarly let

$$\text{E}_G[\text{Var}(\mathbf{X}|G)] = \sum_{j=1}^{J} \pi_j \mathbf{K}_j = \mathbf{K}_W.$$

Then, the between to within class variance ratio is given by

$$\frac{\text{Var}_G(\text{E}[\boldsymbol{l}^T\mathbf{X}|G])}{\text{E}_G[\text{Var}(\boldsymbol{l}^T\mathbf{X}|G)]} = \frac{\boldsymbol{l}^T\mathbf{K}_B\boldsymbol{l}}{\boldsymbol{l}^T\mathbf{K}_W\boldsymbol{l}} \tag{2.13}$$

with $\boldsymbol{l} = (l_1, \ldots, l_p)^T \neq \mathbf{0}$.

If $\boldsymbol{l}_1$ is the vector which maximizes (2.13) we call the corresponding linear function $\boldsymbol{l}_1^T\mathbf{X}$, Fisher's linear discriminant function or the first canonical variate. Note that the vector $\boldsymbol{l}_1$ in Fisher's linear discriminant function is obtained by solving

$$\max_{\boldsymbol{l}\neq\mathbf{0}} \boldsymbol{l}^T\mathbf{K}_B\boldsymbol{l}, \tag{2.14}$$

where $\boldsymbol{l}$ is subject to

$$\boldsymbol{l}^T\mathbf{K}_W\boldsymbol{l} = 1. \tag{2.15}$$

Thus, $\boldsymbol{l}_1$ is the eigenvector of $\mathbf{K}_W^{-1}\mathbf{K}_B$ corresponding to its largest eigenvalue. In general, $\mathbf{K}_W^{-1}\mathbf{K}_B$ has $\min(p, J-1)$ non-zero eigenvalues. The corresponding eigenvectors define

the second, third, and subsequent linear discriminant functions and we denote these vectors by $\boldsymbol{l}_2, \ldots, \boldsymbol{l}_{\min(p,J-1)}$ in what follows.

Fisher's discriminant analysis is well known as a dimension reduction tool. So, we now consider the case of $s \leq \min(p, J-1)$. Then, Fisher's discrimination rule based on the discriminant function subset $\boldsymbol{l}_1^T\mathbf{X}, \ldots, \boldsymbol{l}_s^T\mathbf{X}$ assigns an observation $\mathbf{x}$ to the class for which the squared Mahalanobis distance

$$\sum_{k=1}^{s}(\boldsymbol{l}_k^T\mathbf{x} - \boldsymbol{l}_k^T\boldsymbol{\mu}_j)^2$$

is minimized over $j = 1, \ldots, J$.

2.2.3.2 Fisher's discriminant function via canonical correlation analysis

In this section we will demonstrate that Fisher's LDA is a special case of canonical correlation. To establish this we will take $\mathbf{X}$ to be a $p \times 1$ random vector representing an observation from one of the J classes as before. To represent the class membership corresponding to $\mathbf{X}$, we then define the dummy variables $Y_j,\ j = 1, \ldots, J-1$, as

$$Y_j = \begin{cases} 1, & \text{if } G = j, \\ 0, & \text{otherwise.} \end{cases}$$

Let $\mathbf{Y} = (Y_1, \ldots, Y_{J-1})^T$ be the resulting $(J-1) \times 1$ indicator response vector.

We are interested in predicting the class membership of an item based on the predictors $\mathbf{X}$. That is, we wish to predict the vector $\mathbf{Y}$ from $\mathbf{X}$ and then use the predicted value to assign the individual to one of the J classes. CCA provides one possible approach to this problem since it generalizes regression methodology.

We now give a result that relates Fisher's linear discriminant analysis to CCA.

THEOREM **II.1**. *Let $\mathbf{K}_B, \mathbf{K}_W$ be the between-class covariance matrix and a common within-class covariance matrix, respectively, defined in Section 2.2.3.1. Let $\mathbf{a}_i, i = 1,$*

$\ldots, \min(p, J-1)$, *be the coefficient vectors of the canonical variables of the* $\mathbf{X}$ *space. Then, the canonical vectors* $\mathbf{a}_i$ *are the eigenvectors of* $\mathbf{K}_W^{-1}\mathbf{K}_B$.

Proof. Set $\text{Var}(\mathbf{X}) = \mathbf{K}_X, \text{Var}(\mathbf{Y}) = \mathbf{K}_Y$ and $\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{K}_{XY} = \mathbf{K}_{YX}^T$. Then, we know that the vectors $\mathbf{a}_i$ of the canonical variables for $\mathbf{X}$ are obtained from

$$\mathbf{K}_X^{-1}\mathbf{K}_{XY}\mathbf{K}_Y^{-1}\mathbf{K}_{YX}\mathbf{a}_i = \rho_i^2\mathbf{a}_i. \tag{2.16}$$

We now show that an application of this result to the present setting gives

$$\mathbf{K}_X = \text{Var}(\mathbf{X}) = \mathbf{K}_B + \mathbf{K}_W, \tag{2.17}$$

$$\mathbf{K}_Y = \text{Var}(\mathbf{Y}) = \text{diag}(\pi_1, \ldots, \pi_{J-1}) - \boldsymbol{\pi}_A\boldsymbol{\pi}_A^T, \tag{2.18}$$

and

$$\mathbf{K}_{XY} = \text{Cov}(\mathbf{X}, \mathbf{Y}) = (\pi_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_{J-1}(\boldsymbol{\mu}_{J-1} - \boldsymbol{\mu})), \tag{2.19}$$

where $\boldsymbol{\pi}_A = (\pi_1, \ldots, \pi_{J-1})^T$.

To verify (2.18) and (2.19), first let us create a $J \times 1$ vector $\mathbf{Y}_A = \mathbf{Y}_A(G)$ from the categorical response G, such that $\mathbf{Y}_A = \mathbf{e}_j$ if $G = j$ for $j = 1, \ldots, J$, with $\mathbf{e}_j$ an elementary vector consisting of all 0's except for a 1 in its jth entry. Then, $\mathbf{Y}_A$ has a multinomial distribution with cell probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)^T$ from which we see that

$$\text{E}[\mathbf{Y}_A] = \boldsymbol{\pi}, \quad \text{Var}(\mathbf{Y}_A) = \text{diag}(\pi_1, \ldots, \pi_J) - \boldsymbol{\pi}\boldsymbol{\pi}^T.$$

Because $\mathbf{Y} = \mathbf{A}\mathbf{Y}_A$ with $\mathbf{A}$ the $(J-1) \times J$ matrix $[\mathbf{I}_{J-1} : \mathbf{0}]$ for $\mathbf{I}_{J-1}$ a $(J-1) \times (J-1)$ identity matrix,

$$\text{E}[\mathbf{Y}] = \mathbf{A}\text{E}[\mathbf{Y}_A] = \mathbf{A}\boldsymbol{\pi} = \boldsymbol{\pi}_A,$$

and

$$\text{Var}(\mathbf{Y}) = \mathbf{A}\text{Var}(\mathbf{Y}_A)\mathbf{A}^T = \text{diag}(\pi_1, \ldots, \pi_{J-1}) - \boldsymbol{\pi}_A\boldsymbol{\pi}_A^T.$$

Also, we can show that

$$\begin{aligned}\mathrm{E}[\mathbf{X}\mathbf{Y}^T] &= \mathrm{E}_G[\mathrm{E}(\mathbf{X}\mathbf{Y}_A{}^T|G)]\mathbf{A}^T = \Big\{\sum_{j=1}^{J}\mathrm{E}[\mathbf{X}|G=j]P(G=j)\mathbf{Y}_A^T(G=j)\Big\}\mathbf{A}^T \\ &= \Big\{\sum_{j=1}^{J}\pi_j\boldsymbol{\mu}_j\mathbf{e}_j^T\Big\}\mathbf{A}^T = (\pi_1\boldsymbol{\mu}_1,\ldots,\pi_J\boldsymbol{\mu}_J)\mathbf{A}^T \\ &= (\pi_1\boldsymbol{\mu}_1,\ldots,\pi_{J-1}\boldsymbol{\mu}_{J-1}).\end{aligned}$$

So we now see that

$$\mathbf{K}_{XY} = \mathrm{E}[\mathbf{X}\mathbf{Y}^T] - \mathrm{E}[\mathbf{X}]\mathrm{E}[\mathbf{Y}]^T = (\pi_1(\boldsymbol{\mu}_1-\boldsymbol{\mu}),\ldots,\pi_{J-1}(\boldsymbol{\mu}_{J-1}-\boldsymbol{\mu})).$$

Now observe that

$$\mathbf{K}_X = \mathrm{Var}(\mathbf{X}) = \mathrm{Var}_G(\mathrm{E}[\mathbf{X}|G]) + \mathrm{E}_G[\mathrm{Var}(\mathbf{X}|G)] = \mathbf{K}_B + \mathbf{K}_W \tag{2.20}$$

as before, and that

$$\mathbf{K}_{XY}\mathbf{K}_Y^{-1}\mathbf{K}_{YX} = \sum_{j=1}^{J}\pi_j(\boldsymbol{\mu}_j-\boldsymbol{\mu})(\boldsymbol{\mu}_j-\boldsymbol{\mu})^T = \mathbf{K}_B \tag{2.21}$$

since $\mathbf{K}_Y^{-1} = \mathrm{diag}\left(\pi_1^{-1},\ldots,\pi_{J-1}^{-1}\right) + \pi_J^{-1}\mathbf{1}\mathbf{1}^T$. Therefore, (2.16) is equivalent to

$$\mathbf{K}_X^{-1}\mathbf{K}_B\mathbf{a}_i = \rho_i^2\mathbf{a}_i.$$

as was to be shown and the desired result

$$\mathbf{K}_W^{-1}\mathbf{K}_B\mathbf{a}_i = \frac{\rho_i^2}{1-\rho_i^2}\mathbf{a}_i$$

is implied by the fact that $\mathbf{K}_X = \mathbf{K}_B + \mathbf{K}_W$.

□

Theorem II.1 tells us that the canonical variables of the $\mathbf{X}$ space are proportionally the same as Fisher's linear discriminant functions in Section 2.2.3.1. The two sets of vectors

differ by a proportionality factor because they are subject to different normalization: i.e., the vectors $\boldsymbol{l}_i$ in Fisher's approach satisfy

$$\boldsymbol{l}_i^T \mathbf{K}_W \boldsymbol{l}_i = 1, \quad \boldsymbol{l}_i^T \mathbf{K}_W \boldsymbol{l}_k \neq 0 \text{ for } j \neq i,\ i,k = 1,\ldots,\min(p, J-1)$$

while the vectors $\mathbf{a}_i$ in the canonical variables of the $\mathbf{X}$ space are normalized via the conditions

$$\mathbf{a}_i^T \mathbf{K}_X \mathbf{a}_i = 1, \quad \mathbf{a}_i^T \mathbf{K}_X \mathbf{a}_k \neq 0 \text{ for } k \neq i,\ i,k = 1,\ldots,\min(p, J-1).$$

2.2.4 Sample Linear Discriminant Functions

Let $(\mathbf{X}_1, G_1),\ldots,(\mathbf{X}_N, G_N)$ be iid copies of $(\mathbf{X}, G)$. Also, for $i = 1,\ldots,N, j = 1,\ldots,J$, let $p_j = \frac{N_j}{N}$ and $\mathbf{X}_{ij} = \mathbf{X}_i I(G_i = j)$, where $I(G_i = j)$ is 1 if $G_i = j$ and otherwise is 0. Then, $\bar{\mathbf{X}}_j = \frac{1}{N_j}\sum_{i=1}^{N} \mathbf{X}_{ij}$ and $\bar{\mathbf{X}} = \sum_{j=1}^{J} p_j \bar{\mathbf{X}}_j$ with $N_j = \sum_{i=1}^{N} I(G_i = j)$ and $N = \sum_{j=1}^{J} N_j$.

As in canonical correlation analysis, we will use

$$\widehat{\mathbf{K}}_W = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_j)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_j)^T,\ \widehat{\mathbf{K}}_B = \sum_{j=1}^{J} p_j(\bar{\mathbf{X}}_j - \bar{\mathbf{X}})(\bar{\mathbf{X}}_j - \bar{\mathbf{X}})^T$$

and $\widehat{\mathbf{K}}_X = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T = \widehat{\mathbf{K}}_B + \widehat{\mathbf{K}}_W$. The sample linear discriminant function based on Bayes' classifier is then

$$\widehat{d}_j(\mathbf{x}) = \bar{\mathbf{X}}_j^T \widehat{\mathbf{K}}_W^{-1}\mathbf{x} - \frac{1}{2}\bar{\mathbf{X}}_j^T \widehat{\mathbf{K}}_W^{-1}\bar{\mathbf{X}}_j + \log p_j$$

and the resulting Bayes linear discriminant rule assigns $\mathbf{x}$ to the population where $\widehat{d}_j(\mathbf{x})$ is largest.

The optimal coefficient vectors in the sample Fisher's discriminant functions are the eigenvector of

$$\widehat{\mathbf{K}}_W^{-1}\widehat{\mathbf{K}}_B.$$

If $\hat{\boldsymbol{l}}_1, \ldots, \hat{\boldsymbol{l}}_s$ are the eigenvectors of $\widehat{\mathbf{K}}_W^{-1}\widehat{\mathbf{K}}_B$ corresponding to the first s largest eigenvalues, then $\mathbf{x}$ is classified into the population whose index minimizes

$$\sum_{k=1}^{s}(\hat{\boldsymbol{l}}_k^T\mathbf{x} - \hat{\boldsymbol{l}}_k^T\bar{\mathbf{X}}_j)^2.$$

2.2.5 Discrimination and Multivariate Analysis of Variance

Discriminant analysis and multivariate analysis of variance (MANOVA) are closely related concepts that, in a sense, represent different sides of the same coin. While discriminant analysis tries to find linear functions that can separate the population mean vectors, MANOVA asks the question of whether discrimination is even feasible. In this section we will discuss some of the connections between Fisher's discriminant analysis and MANOVA.

Consider the situation where we are discriminating between J normal populations with the same covariance matrix. If all the means are equal, that is, $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_J$, then it is meaningless to even attempt to discriminate between the populations. So, to check whether or not discriminant analysis is worthwhile, we are interested in testing the hypothesis $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_J$ given a common within class covariance matrix $\mathbf{K}_j = \mathbf{K}_W, j = 1, \ldots, J$. This is the problem addressed by the one-way multivariate analysis of variance.

Let $(\mathbf{X}_1, G_1), \ldots, (\mathbf{X}_N, G_N)$ be a random sample as in Section 2.2.4. Then the log likelihood is

$$l(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_J, \mathbf{K}_W) = -\frac{N}{2}\log|2\pi\mathbf{K}_W| - \frac{N}{2}\text{tr}(\mathbf{K}_W^{-1}\widehat{\mathbf{K}}_W) - \frac{1}{2}\sum_{j=1}^{J} N_j(\bar{\mathbf{X}}_j - \boldsymbol{\mu}_j)^T\mathbf{K}_W^{-1}(\bar{\mathbf{X}}_j - \boldsymbol{\mu}_j).$$

So, the maximum likelihood estimates (m.l.e.) of $\boldsymbol{\mu}_j$ and $\mathbf{K}_W$ are $\bar{\mathbf{x}}_j$ and $\widehat{\mathbf{K}}_W$, respectively. Thus, the maximized log likelihood is

$$l_1 = -\frac{N}{2}\log|2\pi\widehat{\mathbf{K}}_W| - \frac{Np}{2}.$$

The log likelihood under the null hypothesis is

$$l(\boldsymbol{\mu}, \mathbf{K}_W) = -\frac{N}{2}\log|2\pi\mathbf{K}_W| - \frac{N}{2}\text{tr}(\mathbf{K}_W^{-1}\widehat{\mathbf{K}}_X) - \frac{N}{2}(\bar{\mathbf{X}} - \boldsymbol{\mu})^T\mathbf{K}_W^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$$

and hence the m.l.e's of $\boldsymbol{\mu}$ and $\mathbf{K}_W$ are $\bar{\mathbf{X}}$ and $\widehat{\mathbf{K}}_X$, respectively. Thus, the maximized log likelihood under the null hypothesis is

$$l_0 = -\frac{N}{2}\log|2\pi\widehat{\mathbf{K}}_X| - \frac{Np}{2}.$$

Combining l_0 and l_1 we obtain the likelihood ratio given by

$$(|\widehat{\mathbf{K}}_W|/|\widehat{\mathbf{K}}_X|)^{-N/2}.$$

The corresponding test statistic is referred to as Wilk's Λ. Note that the statistic is

$$|\widehat{\mathbf{K}}_W|/|\widehat{\mathbf{K}}_X| = |\mathbf{I} + \widehat{\mathbf{K}}_W^{-1}\widehat{\mathbf{K}}_B|^{-1} = \prod_{i=1}^{\min(p,J-1)}(1+\hat{\gamma}_i)^{-1},$$

where $\hat{\gamma}_1, \ldots, \hat{\gamma}_{\min(p,J-1)}$ are the eigenvalues of $\widehat{\mathbf{K}}_W^{-1}\widehat{\mathbf{K}}_B$. In fact the Λ statistic is based on $\prod_i(1-\hat{\rho}_i^2)$ due to $\hat{\gamma}_i = \hat{\rho}_i^2/(1-\hat{\rho}_i^2)$, where $\hat{\rho}_1^2, \ldots, \hat{\rho}_{\min(p,J-1)}^2$ are the eigenvalues of $\widehat{\mathbf{K}}_X^{-1}\widehat{\mathbf{K}}_B$. Thus, rejection of H_0 will occur when the estimated canonical correlations are large.

2.3 Functional Canonical Correlation Analysis

In this section we discuss how canonical correlation analysis is implemented when the data are random curves or can be viewed as deriving from random curves. Data of this type arise in many real-life situations, where the observed data represents continuous functions sampled at discrete points.

Smoothed functional canonical correlations have been proposed by Leurgans et al. (1993), who demonstrated the need for regularization in functional canonical correlation

analysis. They assume that the observed curves $\{X_i(t), Y_i(t), i = 1, \ldots, N\}$ are independent realizations of a bivariate second-order stochastic process with zero mean functions and covariance functions $K_X(s,t) = \mathrm{E}[X(s)X(t)]$, $K_Y(s,t) = \mathrm{E}[Y(s)Y(t)]$ and $K_{XY}(s,t) = \mathrm{E}[X(s)Y(t)]$. Suppose that sample covariance functions are given as

$$\widehat{K}_X(s,t) = \frac{1}{N}\sum_{i=1}^{N} X_i(s)X_i(t), \quad \widehat{K}_Y(s,t) = \frac{1}{N}\sum_{i=1}^{N} Y_i(s)Y_i(t),$$

and $\widehat{K}_{XY}(s,t) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} X_i(s)Y_i(t)$.

Let $L^2[0,1]$ be the Hilbert space of square integrable functions on $[0,1]$ with associated inner product

$$\langle f, g\rangle_{L^2[0,1]} = \int_0^1 f(s)g(s)ds.$$

Also, let T_X, T_Y and T_{XY} be the covariance operators defined by

$$(T_X f)(\cdot) = \int_0^1 K_X(\cdot, t)f(t)dt, \quad (T_Y g)(\cdot) = \int_0^1 K_Y(\cdot, t)g(t)dt,$$

and $(T_{XY}g)(\cdot) = \displaystyle\int_0^1 K_{XY}(\cdot, t)g(t)dt$, respectively. Then, canonical correlation analysis finds $\langle f, X\rangle_{L^2[0,1]}$ and $\langle g, Y\rangle_{L^2[0,1]}$ with $f, g \in L^2[0,1]$ maximizing

$$\frac{\langle f, T_{XY}g\rangle^2_{L^2[0,1]}}{\langle f, T_X f\rangle_{L^2[0,1]}\langle g, T_Y g\rangle_{L^2[0,1]}}.$$

Now define the operators V_X, V_Y and V_{XY} by writing $V_X f$ for the function

$$(V_X f)(\cdot) = \int_0^1 \widehat{K}_X(\cdot, t)f(t)dt,$$

and correspondingly for V_Y, V_{XY}. Then, Leurgans et al. (1993) find $\langle f, X\rangle_{L^2[0,1]}$ and $\langle g, Y\rangle_{L^2[0,1]}$ that maximize the penalized sample squared correlation defined by

$$\frac{\langle f, V_{XY}g\rangle^2_{L^2[0,1]}}{\left\{\langle f, V_X f\rangle_{L^2[0,1]} + \vartheta_1\|f''\|^2_{L^2[0,1]}\right\}\left\{\langle g, V_Y g\rangle_{L^2[0,1]} + \vartheta_2\|g''\|^2_{L^2[0,1]}\right\}}, \tag{2.22}$$

where ϑ_1 and ϑ_2 are positive smoothing parameters. This procedure is referred to as smoothed canonical correlation analysis (SCCA). They implemented regularization in the criterion (2.22) via cubic smoothing splines and demonstrated their technique with an application to the study of human gait movement data. The effect of the roughness penalty in the denominator of the squared correlation is that both variances and roughness of canonical variables are considered.

He et al. (2002) developed canonical correlation analysis methodology for functional data using a direct parallel of the finite dimensional multivariate analysis technique applied to covariance operators. For their approach, the auto and cross covariance functions of the processes are assumed to be square integrable. This allows them to define covariance operators on $L^2[0,1]$ as

$$T_X = \sum_i \lambda_i \phi_i \otimes_{L^2[0,1]} \phi_i,\ T_Y = \sum_j \nu_j \theta_j \otimes_{L^2[0,1]} \theta_j \text{ and } T_{XY} = \sum_{i,j} \gamma_{ij} \phi_i \otimes_{L^2[0,1]} \theta_j,$$

where $\{\phi_i\}$ and $\{\theta_j\}$ are orthonormal bases for two Hilbert spaces of square integrable functions on $[0,1]$ and the tensor operator is defined by $(\phi \otimes_{L^2[0,1]} \theta)h = \langle \phi, h \rangle_{L^2[0,1]} \theta$. Then, under certain restrictions they obtain canonical correlations as singular values of

$$C = T_X^{-1/2} T_{XY} T_Y^{-1/2} = \sum_i \rho_i \phi_i \otimes_{L^2[0,1]} \theta_i.$$

The difficulty with this development is that the covariance operators T_X and T_Y are not invertible in $L^2[0,1]$. To circumvent this problem they restrict attention to the sets F_{XX} and F_{YY} that represent orthogonal complements of their null spaces in $L^2[0,1]$. For example, they define

$$F_{XX} = \{f \in L^2[0,1] : \sum_{i=1}^{\infty} \lambda_i^{-1} |\langle f, \phi_i \rangle_{L^2[0,1]}|^2 < \infty,\ f \perp \text{Ker}(T_X)\}$$

with $\text{Ker}(T_X) = \{h \in L^2[0,1] : T_X h = 0\}$.

2.4 Functional Discriminant Analysis

Hall et al. (2001) treated signal discrimination by finding a finite dimensional representation via the Karhunen-Loève basis expansion and employing nonparametric kernel methods on the basis coefficients. Let X be a zero-mean, second-order stochastic process. Provided that the covariance function is continuous on $[0,1] \times [0,1]$, the Karhunen-Lòeve expansion gives

$$X(\cdot) = \sum_{j=1}^{\infty} \lambda_j \phi_j(\cdot),$$

where $\lambda_j = \langle \phi_j, X \rangle_{L^2[0,1]}$ and $\{\phi_j\}$ are the eigenvalues and eigenvector sequence of the covariance operator corresponding to the covariance function of the X process. The λ_j's and ϕ_j's are referred to as the principal component scores and principal component basis functions.

Given a random sample $X_1, \ldots, X_N$ of the process X, the scores $\lambda_{ij} = \langle \phi_j, X_i \rangle_{L^2[0,1]}$, $j \geq 1$, serve as surrogates for the observation X_i, for purpose of density estimation and classification. Taking m principal component scores, they observe data

$$\mathbf{X}_i^{(m)} = (\lambda_{i1}, \ldots, \lambda_{im})^T, \ i = 1, \ldots, N.$$

A kernel estimator of the density of $\mathbf{X}_i^{(m)}$ at $\mathbf{x}^{(m)} = (\xi_1, \ldots, \xi_m)$ with $\xi_j = \langle \phi_j, x \rangle_{L^2[0,1]}$ is given by

$$\hat{f}_m(\mathbf{x}^{(m)}) = \frac{1}{N} \sum_{i=1}^{N} K(h^{-1} \| \mathbf{x}^{(m)} - \mathbf{X}_i^{(m)} \|_{\mathbb{R}^2}),$$

where $\| \mathbf{x}^{(m)} - \mathbf{X}_i^{(m)} \|_{\mathbb{R}^2}^2 = \sum_{j=1}^{m} (\lambda_{ij} - \xi_j)^2$, h is a bandwidth, and K is a compactly supported univariate kernel function. Given training data, they estimate the true class densities by the proposed kernel estimator and classify a new signal x to the class with the largest kernel density estimate on $\mathbf{x}^{(m)}$.

James and Hastie (2001) proposed a functional linear discriminant analysis method derived from treating the longitudinal observations as samples from underlying smoothed

curves. They used natural cubic spline functions to model these curves in a similar way as in Rice and Wu (2001). Also, such curves and measurement errors are assumed to be Gaussian for the standard LDA (Bayes' classifier). Then, for the ith curve from the jth class, their proposed model is

$$\mathbf{Y}_{ij} = \mathbf{S}_{ij}(\boldsymbol{\mu}_j + \boldsymbol{\gamma}_{ij}) + \boldsymbol{\epsilon}_{ij},\ j = 1, \ldots, J,\ i = 1, \ldots, N_j,$$
$$\boldsymbol{\epsilon}_{ij} \sim N_{n_{ij}}(\mathbf{0}, \sigma^2\mathbf{I}),\ \boldsymbol{\gamma}_{ij} \sim N_q(\mathbf{0}, \boldsymbol{\Gamma}),$$

where $\mathbf{Y}_{ij}$ and $\boldsymbol{\epsilon}_{ij}$ are the corresponding vectors of observations and measurement errors at times $t_{ij1}, \ldots, t_{ijn_{ij}}$, $\mathbf{S}_{ij} = (\mathbf{s}(t_{ij1}), \ldots, \mathbf{s}(t_{ijn_{ij}}))^T$ with $\mathbf{s}(\cdot)$ a spline function from a spline basis with dimension q, J is the number of classes and N_j is the number of individuals in the jth class.

In particular, James and Hastie (2001) develop a reduced rank model for sparsely sampled curves via use of Fisher's discriminant analysis method. The reduced rank model has the form

$$\mathbf{Y}_{ij} = \mathbf{S}_{ij}(\boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_j + \boldsymbol{\gamma}_{ij}) + \boldsymbol{\epsilon}_{ij},\ j = 1, \ldots, J,\ i = 1, \ldots, N_j,$$
$$\boldsymbol{\epsilon}_{ij} \sim N_{n_{ij}}(\mathbf{0}, \sigma^2\mathbf{I}),\ \boldsymbol{\gamma}_{ij} \sim N_q(\mathbf{0}, \boldsymbol{\Gamma}),$$

where $\boldsymbol{\lambda}_0$ and $\boldsymbol{\alpha}_j$ are q- and r-dimensional vectors and $\boldsymbol{\Lambda}$ is a $q \times r$ matrix with $r \leq \min(q, J-1)$ satisfying the restrictions $\boldsymbol{\Lambda}^T\mathbf{S}_{ij}^T(\sigma^2\mathbf{I} + \mathbf{S}_{ij}\boldsymbol{\Gamma}\mathbf{S}_{ij}^T)^{-1}\mathbf{S}_{ij}\boldsymbol{\Lambda} = \mathbf{I}, \sum_j \boldsymbol{\alpha}_j = \mathbf{0}$. The fixed-effect term $\mathbf{S}_{ij}(\boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_j)$ models the class mean curves and the random-effects term $\mathbf{S}_{ij}\boldsymbol{\gamma}_{ij}$ allows for individual variation within each class. They fit this model using the EM algorithm and then classify a new observation to the class with the largest posterior probability as in the ordinary multivariate analysis case.

CHAPTER III

CANONICAL CORRELATION ANALYSIS AND DISCRIMINANT ANALYSIS

We present an overview of the foundation of multivariate canonical correlation analysis and discriminant analysis in this chapter. Our treatment of this topic differs somewhat from the classical approach in that we explicitly treat the less than full rank scenario. This opens the door to transactions in infinite dimensions through the reproducing kernel Hilbert space perspective of the next chapters.

3.1 Canonical Correlation Analysis with Less Than Full Rank Covariance Matrices

For a $m \times n$ matrix $\mathbf{A}$, we denote its rank by $r(\mathbf{A})$, define its null space as $\text{Ker}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{0}\}$ and indicate its range by $\text{Im}(\mathbf{A}) = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} = \mathbf{A}\mathbf{x},\ \mathbf{x} \in \mathbb{R}^n\}$. The notation $\perp$ indicates orthogonal complement.

3.1.1 Population Canonical Correlations and Canonical Variables

Suppose that $\mathbf{X}$ is a p-dimensional random vector and that $\mathbf{Y}$ is a q-dimensional random vector with $\text{Var}(\mathbf{X}) = \mathbf{K}_X,\ \text{Var}(\mathbf{Y}) = \mathbf{K}_Y$, and $\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{K}_{XY} = \mathbf{K}_{YX}^T$. In what follows both $\mathbf{K}_X$ and $\mathbf{K}_Y$ may have less than full rank.

Now, we wish to find $\mathbf{a}^T\mathbf{X}$ and $\mathbf{b}^T\mathbf{Y}$ with $\mathbf{a} = (a_1, \ldots, a_p)^T$ and $\mathbf{b} = (b_1, \ldots, b_q)^T$ having the largest possible correlation with one another. For this purpose, let us write the squared correlation between two linear combinations as

$$\rho^2(\mathbf{a}, \mathbf{b}) = \frac{\text{Cov}^2\left(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}\right)}{\text{Var}\left(\mathbf{a}^T\mathbf{X}\right)\text{Var}\left(\mathbf{b}^T\mathbf{Y}\right)} = \frac{(\mathbf{a}^T\mathbf{K}_{XY}\mathbf{b})^2}{(\mathbf{a}^T\mathbf{K}_X\mathbf{a})(\mathbf{b}^T\mathbf{K}_Y\mathbf{b})}$$

when $\mathbf{K}_X\mathbf{a} \neq \mathbf{0}$ and $\mathbf{K}_Y\mathbf{b} \neq \mathbf{0}$.

PROPOSITION **III.1**. *If* $\boldsymbol{l} \in \text{Ker}(\mathbf{K}_X)$ *then* $\mathbf{K}_{YX}\boldsymbol{l} = \mathbf{0}$ *and if* $\boldsymbol{m} \in \text{Ker}(\mathbf{K}_Y)$ *then* $\mathbf{K}_{XY}\boldsymbol{m} = \mathbf{0}$.

Proof. Suppose that $\boldsymbol{l} \in \text{Ker}(\mathbf{K}_X)$. Then, by the Cauchy-Schwartz inequality,

$$(\mathbf{e}_j^T \mathbf{K}_{YX} \boldsymbol{l})^2 = \text{Cov}^2(\boldsymbol{l}^T \mathbf{X}, \mathbf{e}_j^T \mathbf{Y}) \leq \text{Var}(\boldsymbol{l}^T \mathbf{X}) \text{Var}(\mathbf{e}_j^T \mathbf{Y}) = 0$$

for $\mathbf{e}_j$ a $q \times 1$ elementary vector consisting of all 0's except for a 1 in its jth entry. Thus, $\mathbf{e}_j^T \mathbf{K}_{YX} \boldsymbol{l} = 0$ for every $\mathbf{e}_j$. So, $\mathbf{K}_{YX} \boldsymbol{l} = 0$ for $\boldsymbol{l} \in \text{Ker}(\mathbf{K}_X)$. Similarly, if $\boldsymbol{m} \in \text{Ker}(\mathbf{K}_Y)$ then $\mathbf{K}_{XY} \boldsymbol{m} = 0$.

□

For $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$, observe that $\mathbf{a} = \mathbf{a}_* + \mathbf{a}_0$ with $\mathbf{a}_* \in \text{Ker}(\mathbf{K}_X)^\perp, \mathbf{a}_0 \in \text{Ker}(\mathbf{K}_X)$ and $\mathbf{b} = \mathbf{b}_* + \mathbf{b}_0$ with $\mathbf{b}_* \in \text{Ker}(\mathbf{K}_Y)^\perp, \mathbf{b}_0 \in \text{Ker}(\mathbf{K}_Y)$. We now observe from Proposition III.1 that $\mathbf{a}^T \mathbf{K}_{XY} \mathbf{b} = (\mathbf{a}_* + \mathbf{a}_0)^T \mathbf{K}_{XY} (\mathbf{b}_* + \mathbf{b}_0) = \mathbf{a}_*^T \mathbf{K}_{XY} \mathbf{b}_*$, $\mathbf{a}^T \mathbf{K}_X \mathbf{a} = \mathbf{a}_*^T \mathbf{K}_X \mathbf{a}_*$ and $\mathbf{b}^T \mathbf{K}_Y \mathbf{b} = \mathbf{b}_*^T \mathbf{K}_Y \mathbf{b}_*$. So, maximizing $\rho^2(\mathbf{a}, \mathbf{b})$ with $\mathbf{K}_X \mathbf{a} \neq \mathbf{0}$ and $\mathbf{K}_Y \mathbf{b} \neq \mathbf{0}$ is equivalent to maximizing $\rho^2(\mathbf{a}_*, \mathbf{b}_*)$ with $\mathbf{a}_* \in \text{Ker}(\mathbf{K}_X)^\perp$ and $\mathbf{b}_* \in \text{Ker}(\mathbf{K}_X)^\perp$. Thus, equivalently, we may find $\mathbf{a}$ and $\mathbf{b}$ by maximizing

$$\frac{(\mathbf{a}^T \mathbf{K}_{XY} \mathbf{b})^2}{(\mathbf{a}^T \mathbf{K}_X \mathbf{a})\,(\mathbf{b}^T \mathbf{K}_Y \mathbf{b})}$$

over $\mathbf{a} \in \text{Ker}(\mathbf{K}_X)^\perp$ and $\mathbf{b} \in \text{Ker}(\mathbf{K}_Y)^\perp$. Consequently finding the linear combinations of $\mathbf{X}$ and $\mathbf{Y}$ that are most highly correlated is equivalent to finding $\mathbf{a} \in \text{Ker}(\mathbf{K}_X)^\perp$ and $\mathbf{b} \in \text{Ker}(\mathbf{K}_Y)^\perp$ to maximize

$$\text{Cov}^2 \left(\mathbf{a}^T \mathbf{X}, \mathbf{b}^T \mathbf{Y}\right) \tag{3.1}$$

subject to

$$\text{Var}\left(\mathbf{a}^T \mathbf{X}\right) = \text{Var}\left(\mathbf{b}^T \mathbf{Y}\right) = 1. \tag{3.2}$$

Let

$$\mathbf{K}_X = \sum_{i=1}^{r_X} \lambda_{Xi} \mathbf{e}_{Xi} \mathbf{e}_{Xi}^T,$$

where $r_X = r(\mathbf{K}_X) \leq p$ and $(\lambda_{X1}, \mathbf{e}_{X1}), \ldots, (\lambda_{Xr_X}, \mathbf{e}_{Xr_X})$ are the nonzero eigenvalues and associated eigenvectors of $\mathbf{K}_X$. Define the matrices

$$\mathbf{K}_X^{1/2} = \sum_{i=1}^{r_X} \lambda_{Xi}^{1/2} \mathbf{e}_{Xi} \mathbf{e}_{Xi}^T,$$

$$\mathbf{K}_X^{-1/2} = \sum_{i=1}^{r_X} \lambda_{Xi}^{-1/2} \mathbf{e}_{Xi} \mathbf{e}_{Xi}^T$$

and

$$\mathbf{K}_X^{-} = \sum_{i=1}^{r_X} \lambda_{Xi}^{-1} \mathbf{e}_{Xi} \mathbf{e}_{Xi}^T.$$

Also, define $\mathbf{K}_Y, \mathbf{K}_Y^{1/2}, \mathbf{K}_Y^{-1/2}$ and $\mathbf{K}_Y^{-}$ similarly.

PROPOSITION **III.2**. *$\mathbf{K}_X^{-}$ and $\mathbf{K}_X^{-1/2}$ are the Moore-Penrose generalized inverses of $\mathbf{K}_X$ and $\mathbf{K}_X^{1/2}$, respectively. Also,* (i) $\text{Ker}(\mathbf{K}_X) = \text{Ker}(\mathbf{K}_X^{1/2})$, (ii) $\text{Im}(\mathbf{K}_X) = \text{Ker}(\mathbf{K}_X)^{\perp}$, (iii) $\text{Ker}(\mathbf{K}_X^{-})^{\perp} = \text{Ker}(\mathbf{K}_X)^{\perp}$. *Thus, the matrix $\mathbf{K}_X^{-}$ is a one-to-one linear mapping from* $\text{Im}(\mathbf{K}_X)$ *onto* $\text{Ker}(\mathbf{K}_X)^{\perp}$ *and $\mathbf{K}_X^{-1/2}$ is a one-to-one linear mapping from* $\text{Im}(\mathbf{K}_X^{1/2})$ *onto* $\text{Ker}(\mathbf{K}_X^{1/2})^{\perp}$.

Proof. Let $\mathbf{\Lambda}_X = \text{diag}(\lambda_{X1}, \ldots, \lambda_{Xr_X})$ and $\mathbf{P}_X = [\mathbf{e}_{X1}, \ldots, \mathbf{e}_{Xr_X}]$. Then $\mathbf{P}_X^T \mathbf{P}_X = \mathbf{I}_{r_X}$ and $\mathbf{K}_X = \mathbf{P}_X \mathbf{\Lambda}_X \mathbf{P}_X^T$. Then, we can see that

$$\mathbf{K}_X \mathbf{K}_X^{-} \mathbf{K}_X = (\mathbf{P}_X \mathbf{\Lambda}_X \mathbf{P}_X^T)(\mathbf{P}_X \mathbf{\Lambda}_X^{-1} \mathbf{P}_X^T)(\mathbf{P}_X \mathbf{\Lambda}_X \mathbf{P}_X^T) = \mathbf{P}_X \mathbf{\Lambda}_X \mathbf{P}_X^T = \mathbf{K}_X$$

and, similarly, $\mathbf{K}_X^{-} \mathbf{K}_X \mathbf{K}_X^{-} = \mathbf{P}_X \mathbf{\Lambda}_X^{-1} \mathbf{P}_X^T = \mathbf{K}_X^{-}$. Also, $\mathbf{K}_X \mathbf{K}_X^{-}$ and $\mathbf{K}_X^{-} \mathbf{K}_X$ are symmetric which follows from $\mathbf{K}_X \mathbf{K}_X^{-} = \mathbf{K}_X^{-} \mathbf{K}_X = \mathbf{P}_X \mathbf{P}_X^T$. So, $\mathbf{K}_X^{-}$ is the Moore-Penrose generalized inverse of $\mathbf{K}_X$.

Now observe that $\mathbf{K}_X \boldsymbol{l} = \mathbf{0}$ if and only if $\mathbf{e}_{Xi}^T \boldsymbol{l} = 0$ for all $i = 1, \ldots, r_X$ because the vectors $\{\mathbf{e}_{X1}, \ldots, \mathbf{e}_{Xr_X}\}$ are linearly independent. Also, $\mathbf{e}_{Xi}^T \boldsymbol{l} = 0$ for all i if and only if $\mathbf{K}_X^{1/2} \boldsymbol{l} = \mathbf{0}$. Thus, $\text{Ker}(\mathbf{K}_X) = \text{Ker}(\mathbf{K}_X^{1/2})$.

Suppose that $\boldsymbol{l} \in \text{Ker}(\mathbf{K}_X)$ and $\boldsymbol{z} \in \text{Im}(\mathbf{K}_X)$. Then, $\boldsymbol{z} = \mathbf{P}_X \boldsymbol{c}$ for $\boldsymbol{c} \in \mathbb{R}^{r_X}$ because $\text{Im}(\mathbf{K}_X)$ is the space spanned by $\{\mathbf{e}_{X1}, \ldots, \mathbf{e}_{Xr_X}\}$. So, $\boldsymbol{l}^T \boldsymbol{z} = \boldsymbol{l}^T \mathbf{P}_X \boldsymbol{c} = 0$ since $\boldsymbol{l} \in \text{Ker}(\mathbf{K}_X)$ has the consequence that $\mathbf{e}_{Xi}^T \boldsymbol{l} = 0$ for all i. Hence $\boldsymbol{l} \in \text{Im}(\mathbf{K}_X)^{\perp}$ and so $\text{Ker}(\mathbf{K}_X) \subset \text{Im}(\mathbf{K}_X)^{\perp}$. Conversely, if $\boldsymbol{h} \in \text{Im}(\mathbf{K}_X)^{\perp}$ then $0 = \boldsymbol{h}^T (\mathbf{K}_X \boldsymbol{h}) = (\mathbf{K}_X^{1/2} \boldsymbol{h})^T (\mathbf{K}_X^{1/2} \boldsymbol{h})$ and so $\mathbf{K}_X^{1/2} \boldsymbol{h} = \mathbf{0}$: i.e., $\boldsymbol{h} \in \text{Ker}(\mathbf{K}_X^{1/2}) = \text{Ker}(\mathbf{K}_X)$. Therefore, $\text{Ker}(\mathbf{K}_X) = \text{Im}(\mathbf{K}_X)^{\perp}$ and $\text{Ker}(\mathbf{K}_X)^{\perp} = (\text{Im}(\mathbf{K}_X)^{\perp})^{\perp} = \text{Im}(\mathbf{K}_X)$. Since both $\text{Im}(\mathbf{K}_X)$

and $\text{Im}(\mathbf{K}_X^-)$ are spanned by $\mathbf{e}_{X1}, \ldots, \mathbf{e}_{Xr_X}$, $\text{Im}(\mathbf{K}_X) = \text{Im}(\mathbf{K}_X^-)$ and so $\text{Ker}(\mathbf{K}_X)^\perp = \text{Im}(\mathbf{K}_X) = \text{Im}(\mathbf{K}_X^-) = \text{Ker}(\mathbf{K}_X^-)^\perp$.

For $\boldsymbol{z} \in \text{Im}(\mathbf{K}_X)$, observe that $\boldsymbol{z} = \mathbf{P}_X \boldsymbol{c}$ and that $\mathbf{K}_X^- \boldsymbol{z} = \mathbf{0}$ if and only if $\mathbf{P}_X^T \boldsymbol{z} = \mathbf{0}$. Thus, $\mathbf{K}_X^- \boldsymbol{z} = \mathbf{0}$ implies that

$$\mathbf{0} = \mathbf{P}_X^T \boldsymbol{z} = \mathbf{P}_X^T \mathbf{P}_X \boldsymbol{c} = \boldsymbol{c}$$

and we conclude that $\mathbf{K}_X^- \boldsymbol{z} = \mathbf{0}$ if and only if $\boldsymbol{z} = \mathbf{0}$. Moreover, we have $\text{Im}(\mathbf{K}_X^-) = \text{Ker}(\mathbf{K}_X)^\perp$. Therefore, $\mathbf{K}_X^-$ is a one-to-one linear mapping from $\text{Im}(\mathbf{K}_X)$ onto $\text{Ker}(\mathbf{K}_X)^\perp$. Similarly, it can be shown that $\mathbf{K}_X^{-1/2}$ is the Moore-Penrose generalized inverse of $\mathbf{K}_X^{1/2}$ and it is a one-to-one linear mapping from $\text{Im}(\mathbf{K}_X^{1/2})$ onto $\text{Ker}(\mathbf{K}_X^{1/2})^\perp$.

□

To solve problem (3.1) and (3.2), let $\mathbf{u} = \mathbf{K}_X^{1/2}\mathbf{a}$ and $\mathbf{v} = \mathbf{K}_Y^{1/2}\mathbf{b}$. Then, for $\mathbf{a} \in \text{Ker}(\mathbf{K}_X)^\perp$ and $\mathbf{b} \in \text{Ker}(\mathbf{K}_Y)^\perp$, we see that $\mathbf{u} \in \text{Ker}(\mathbf{K}_X^{1/2})^\perp$ and, also, $\mathbf{u} \in \text{Ker}(\mathbf{K}_X)^\perp$ because $\text{Ker}(\mathbf{K}_X^{1/2})^\perp = \text{Ker}(\mathbf{K}_X)^\perp$. Similarly, $\mathbf{v} \in \text{Ker}(\mathbf{K}_Y)^\perp$. It now becomes clear that

$$\frac{(\mathbf{a}^T \mathbf{K}_{XY} \mathbf{b})^2}{(\mathbf{a}^T \mathbf{K}_X \mathbf{a})(\mathbf{b}^T \mathbf{K}_Y \mathbf{b})} = \frac{(\mathbf{u}^T \mathbf{K}_X^{-1/2} \mathbf{K}_{XY} \mathbf{K}_Y^{-1/2} \mathbf{v})^2}{(\mathbf{u}^T \mathbf{u})(\mathbf{v}^T \mathbf{v})}$$

for $\mathbf{a} \in \text{Ker}(\mathbf{K}_X)^\perp$ and $\mathbf{b} \in \text{Ker}(\mathbf{K}_X)^\perp$ because $\mathbf{u} = \mathbf{K}_X^{1/2}\mathbf{a}$ becomes $\mathbf{a} = \mathbf{K}_X^{-1/2}\mathbf{u}$ when $\mathbf{a} \in \text{Ker}(\mathbf{K}_X)^\perp = \text{Ker}(\mathbf{K}_X^{1/2})^\perp$. Thus, in turn, solving problem (3.1) and (3.2) is equivalent to solving the problem

$$\max_{\substack{\mathbf{u} \in \text{Ker}(\mathbf{K}_X)^\perp, \mathbf{v} \in \text{Ker}(\mathbf{K}_Y)^\perp \\ ||\mathbf{u}||_{\mathbb{R}^p} = ||\mathbf{v}||_{\mathbb{R}^q} = 1}} (\mathbf{u}^T \mathbf{K}_X^{-1/2} \mathbf{K}_{XY} \mathbf{K}_Y^{-1/2} \mathbf{v})^2. \tag{3.3}$$

The formulation in (3.3) has the important implication that the optimal $\mathbf{u}$ and $\mathbf{v}$ can be obtained from the singular value decomposition (SVD) of the matrix $\mathbf{K}_X^{-1/2} \mathbf{K}_{XY} \mathbf{K}_Y^{-1/2}$ to produce the weight vectors

$$\mathbf{a} = \mathbf{K}_X^{-1/2}\mathbf{u} \quad \text{and} \quad \mathbf{b} = \mathbf{K}_Y^{-1/2}\mathbf{v}.$$

We can now define the first canonical correlation ρ_1 and the associated weight vectors $\mathbf{a}_1, \mathbf{b}_1$ as

$$\rho_1^2 = \text{Cov}^2\left(\mathbf{a}_1^T\mathbf{X}, \mathbf{b}_1^T\mathbf{Y}\right) = \max_{\mathbf{a}\in\text{Ker}(\mathbf{K}_X)^\perp, \mathbf{b}\in\text{Ker}(\mathbf{K}_Y)^\perp} \text{Cov}^2\left(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}\right), \tag{3.4}$$

where $\mathbf{a}, \mathbf{b}$ are subject to (3.2). For $i > 1$, the ith canonical correlation ρ_i and the associated weight vectors $\mathbf{a}_i, \mathbf{b}_i$ can be defined similarly as

$$\rho_i^2 = \text{Cov}^2\left(\mathbf{a}_i^T\mathbf{X}, \mathbf{b}_i^T\mathbf{Y}\right) = \max_{\mathbf{a}\in\text{Ker}(\mathbf{K}_X)^\perp, \mathbf{b}\in\text{Ker}(\mathbf{K}_Y)^\perp} \text{Cov}^2\left(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}\right) \tag{3.5}$$

where $\mathbf{a}, \mathbf{b}$ are subject to (3.2) and

$$\text{Cov}\left(\mathbf{a}^T\mathbf{X}, \mathbf{a}_j^T\mathbf{Y}\right) = \text{Cov}\left(\mathbf{b}^T\mathbf{X}, \mathbf{b}_j^T\mathbf{Y}\right) = 0, \quad j < i. \tag{3.6}$$

When a solution exists to problem (3.4), ρ_1 is called the first canonical correlation and $\mathbf{a}_1^T\mathbf{X}, \mathbf{b}_1^T\mathbf{Y}$ are referred to as the first canonical variables of the $\mathbf{X}$ and $\mathbf{Y}$ spaces, respectively. Similarly, ρ_i in (3.5) is termed the ith canonical correlation with associated canonical variables of the $\mathbf{X}$ and $\mathbf{Y}$ spaces given by $\mathbf{a}_i^T\mathbf{X}$ and $\mathbf{b}_i^T\mathbf{Y}$.

Suppose that $\mathbf{K}_X^{-1/2}\mathbf{K}_{XY}\mathbf{K}_Y^{-1/2}$ has rank $r \leq \min(r_X, r_Y)$ with $r_Y = r(\mathbf{K}_Y)$. Then the singular value decomposition for $\mathbf{K}_X^{-1/2}\mathbf{K}_{XY}\mathbf{K}_Y^{-1/2}$ is

$$\mathbf{K}_X^{-1/2}\mathbf{K}_{XY}\mathbf{K}_Y^{-1/2} = \mathbf{U}\begin{pmatrix} \mathbf{D}_{r\times r} & \mathbf{O}_{r\times(q-r)} \\ \mathbf{O}_{(p-r)\times r} & \mathbf{O}_{(p-r)\times(q-r)} \end{pmatrix}\mathbf{V}^T, \tag{3.7}$$

where $\mathbf{O}_{k_1\times k_2}$ is a $k_1 \times k_2$ matrix of all zeros, $\mathbf{U}$ is a $p \times p$ orthogonal matrix of eigenvectors corresponding to the eigenvalues $\rho_1^2, \ldots, \rho_r^2$ of

$$\mathbf{K}_X^{-1/2}\mathbf{K}_{XY}\mathbf{K}_Y^{-}\mathbf{K}_{YX}\mathbf{K}_X^{-1/2}$$

and $\mathbf{V}$ is a $q \times q$ orthogonal matrix of eigenvectors corresponding to the eigenvalues $\rho_1^2, \ldots, \rho_r^2$ of

$$\mathbf{K}_Y^{-1/2}\mathbf{K}_{YX}\mathbf{K}_X^{-}\mathbf{K}_{XY}\mathbf{K}_Y^{-1/2}$$

and $\mathbf{D} = \text{diag}(\rho_1, \ldots, \rho_r)$.

THEOREM **III.1**. *Let* $\rho_1^2 \geq \cdots \geq \rho_r^2 > 0$ *and let* $\mathbf{u}_i, \mathbf{v}_i$ *be the columns of* $\mathbf{U}$ *and* $\mathbf{V}$ *that correspond to* ρ_i*. Then,* $\mathbf{a}_i = \mathbf{K}_X^{-1/2}\mathbf{u}_i$ *and* $\mathbf{b}_i = \mathbf{K}_Y^{-1/2}\mathbf{v}_i$ *solve problems (3.4) – (3.5) subject to (3.2) and (3.6) with corresponding canonical correlation* ρ_i*.*

Proof. From (3.7), we have

$$\mathbf{K}_X^{-1/2}\mathbf{K}_{XY}\mathbf{K}_Y^{-1/2} = \sum_{i=1}^{r} \rho_i \mathbf{u}_i \mathbf{v}_i^T.$$

Then, observe that

$$(\mathbf{u}^T\mathbf{K}_X^{-1/2}\mathbf{K}_{XY}\mathbf{K}_Y^{-1/2}\mathbf{v})^2 \leq \rho_1^2 \left(\sum_{i=1}^{r}(\mathbf{u}^T\mathbf{u}_i)(\mathbf{v}^T\mathbf{v}_i)\right)^2 \leq \rho_1^2 \sum_{i=1}^{r}(\mathbf{u}^T\mathbf{u}_i)^2 \sum_{i=1}^{r}(\mathbf{v}^T\mathbf{v}_i)^2.$$

by the Cauchy-Schwarz inequality. Since $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ are orthonormal vectors in $\text{Ker}(\mathbf{K}_X)^\perp$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ are orthonormal vectors in $\text{Ker}(\mathbf{K}_Y)^\perp$, we obtain from Bessel's inequality that

$$(\mathbf{u}^T\mathbf{K}_X^{-1/2}\mathbf{K}_{XY}\mathbf{K}_Y^{-1/2}\mathbf{v})^2 \leq \rho_1^2 \sum_{i=1}^{r}(\mathbf{u}^T\mathbf{u}_i)^2 \sum_{i=1}^{r}(\mathbf{v}^T\mathbf{v}_i)^2 \leq \rho_1^2(\mathbf{u}^T\mathbf{u})(\mathbf{v}^T\mathbf{v}),$$

where equality holds if and only if $\mathbf{u} = \mathbf{u}_1$ and $\mathbf{v} = \mathbf{v}_1$. For the general case we have $\mathbf{u} \perp \mathbf{u}_i$ and $\mathbf{v} \perp \mathbf{v}_i$ for $1 \leq i \leq j-1$ and

$$\text{Cov}^2(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}) = (\mathbf{u}^T\mathbf{K}_X^{-1/2}\mathbf{K}_{XY}\mathbf{K}_Y^{-1/2}\mathbf{v})^2 \leq \rho_j^2(\mathbf{u}^T\mathbf{u})(\mathbf{v}^T\mathbf{v}),$$

with equality if and only if $\mathbf{u} = \mathbf{u}_j$ and $\mathbf{v} = \mathbf{v}_j$.

□

It now follows that $\mathbf{a}_i$ and $\mathbf{b}_i$ can be obtained via solution of the eigenvalue problems

$$\mathbf{K}_X^-\mathbf{K}_{XY}\mathbf{K}_Y^-\mathbf{K}_{YX}\mathbf{a}_i = \rho_i^2\mathbf{a}_i, \tag{3.8}$$

and

$$\mathbf{K}_Y^-\mathbf{K}_{YX}\mathbf{K}_X^-\mathbf{K}_{XY}\mathbf{b}_i = \rho_i^2\mathbf{b}_i. \tag{3.9}$$

Also, the relationship between the coefficient vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ can be derived from (3.7). Specifically, upon premultiplying by $\mathbf{U}^T$ and postmultiplying by $\mathbf{K}_Y^{-1/2}$ we obtain

$$\mathbf{K}_Y^{-}\mathbf{K}_{YX}\mathbf{a}_i = \rho_i \mathbf{b}_i,\ i = 1, \ldots, r, \tag{3.10}$$

and, similarly,

$$\mathbf{K}_X^{-}\mathbf{K}_{XY}\mathbf{b}_i = \rho_i \mathbf{a}_i,\ i = 1, \ldots, r. \tag{3.11}$$

3.1.2 Canonical Correlation Analysis and Regression

CCA can be viewed as an essential technique for carrying out regression of one vector on another vector. To see the connection to ordinary linear regression with one independent variable, suppose that we observe $(\mathbf{X}, Y)$, where $\mathbf{X}$ is a p-variate predictor vector and Y is a scalar response. In this situation, we may be interested in finding the linear combination $\mathbf{a}^T\mathbf{X}$ which is most highly correlated with Y. For this purpose, we can first think of the regression of Y on $\mathbf{X}$.

Set $\mathrm{E}[\mathbf{X}] = \boldsymbol{\mu}_X, \mathrm{E}[Y] = \mu_Y, \mathrm{Var}(\mathbf{X}) = \mathbf{K}_X, \mathrm{Cov}(\mathbf{X}, Y) = \mathbf{K}_{XY}$ and assume that $\mathrm{Var}(Y) = \sigma^2$. When we minimize

$$\mathrm{E}|Y - m(\mathbf{X})|^2$$

over all functions m this provides us with an approximation to Y. More precisely, Y is actually a function on a probability space $(\Omega, \mathcal{B}, P)$ and the best least-squares approximation to $Y(\omega), \omega \in \Omega$, under certain restrictions, is

$$g(\omega) = \mathrm{E}[Y|\mathbf{X}(\omega)],\ \omega \in \Omega.$$

Now, for linear regression we restrict the optimization of

$$\mathrm{E}[Y - m(\mathbf{X})]^2$$

to functions of the form

$$m^*(\omega) = (m \circ \mathbf{X})(\omega) = \alpha + \boldsymbol{\beta}^T \mathbf{X}(\omega)$$

with $\mathbf{X}(\omega)$ the value of the random vector $\mathbf{X}$ for outcome $\omega \in \Omega$ and $\boldsymbol{\beta} \in \text{Ker}(\mathbf{K}_X)^\perp$. The population linear regression plane is the result of this optimization. Specifically, the best least-squares approximation of Y is

$$\left(\mu_Y - \mathbf{K}_{YX}\mathbf{K}_X^- \boldsymbol{\mu}_X\right) + \mathbf{K}_{YX}\mathbf{K}_X^- \mathbf{X}. \tag{3.12}$$

So, $\widetilde{\alpha} + \widetilde{\boldsymbol{\beta}}^T \mathbf{X}(\omega)$ with $\widetilde{\alpha} = \mu_Y - \mathbf{K}_{YX}\mathbf{K}_X^- \boldsymbol{\mu}_X$ and $\widetilde{\boldsymbol{\beta}} = \mathbf{K}_{YX}\mathbf{K}_X^-$ approximates the function $Y(\omega)$ on Ω.

Observe that $\widetilde{\boldsymbol{\beta}}$ can also be obtained as the solution of

$$\min_{\boldsymbol{\beta} \in \text{Ker}(\mathbf{K}_X)^\perp} \left\{\sigma^2 - 2\boldsymbol{\beta}^T \mathbf{K}_{XY} + \boldsymbol{\beta}^T \mathbf{K}_X \boldsymbol{\beta}\right\}. \tag{3.13}$$

Then, solving (3.13) is equivalent to solving

$$\max_{\mathbf{a} \in \text{Ker}(\mathbf{K}_X)^\perp} \text{Cov}^2\left(\mathbf{a}^T \mathbf{X}, Y\right) \tag{3.14}$$

subject to $\text{Var}\left(\mathbf{a}^T \mathbf{X}\right) = 1$.

The weight vector $\mathbf{a}$ in (3.14) and the coefficient vector $\widetilde{\boldsymbol{\beta}}$ are related by

$$\widetilde{\boldsymbol{\beta}} = \sigma \rho \mathbf{a}.$$

This follows from observing that $\mathbf{a}$ is the solution of the problem $\sigma^{-2}\mathbf{K}_X^- \mathbf{K}_{XY}\mathbf{K}_{YX}\mathbf{a} = \rho^2 \mathbf{a}$. So $\sigma^{-1}\mathbf{a}^T \mathbf{K}_{XY} = \text{Corr}(\mathbf{a}^T \mathbf{X}, Y) = \rho$ and moreover we know $\widetilde{\boldsymbol{\beta}} = \mathbf{K}_X^- \mathbf{K}_{XY}$.

3.1.3 Sample Canonical Correlations and Canonical Variables

Suppose now that we observe N iid copies $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_N, \mathbf{Y}_N)$ of $(\mathbf{X}, \mathbf{Y})$. In this section we will discuss how such data can be used to produce consistent estimators of the canonical correlations and variables.

A parallel of the development in the previous section can be followed for the analysis of sample data. All that is needed is that $\mathbf{K}_X, \mathbf{K}_Y, \mathbf{K}_{XY}$ are replaced by their estimators. Specifically, we estimate the population variance and covariances by their corresponding sample moments in (2.10) and (2.11) as in Section 2.1.2.

As in the population setting, we define

$$\hat{\rho}_1^2 = \max_{\mathbf{a}\in \text{Ker}(\widehat{\mathbf{K}}_X)^\perp, \mathbf{b}\in \text{Ker}(\widehat{\mathbf{K}}_Y)^\perp} \hat{\rho}^2(\mathbf{a}, \mathbf{b}) \tag{3.15}$$

with $\hat{\rho}$ the sample correlation between $\mathbf{a}^T\mathbf{X}$ and $\mathbf{b}^T\mathbf{Y}$. That is,

$$\hat{\rho}(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^N \mathbf{a}^T\mathbf{X}_i\mathbf{b}^T\mathbf{Y}_i - \frac{1}{N}\sum_{i=1}^N \mathbf{a}^T\mathbf{X}_i \sum_{i=1}^N \mathbf{b}^T\mathbf{Y}_i}{(SS_{\mathbf{a}^T\mathbf{X}} SS_{\mathbf{b}^T\mathbf{Y}})^{1/2}}$$

with

$$SS_{\mathbf{a}^T\mathbf{X}} = \sum_{i=1}^N (\mathbf{a}^T\mathbf{X}_i)^2 - \frac{1}{N}\left(\sum_{i=1}^N \mathbf{a}^T\mathbf{X}_i\right)^2, \; SS_{\mathbf{b}^T\mathbf{Y}} = \sum_{i=1}^N (\mathbf{b}^T\mathbf{Y}_i)^2 - \frac{1}{N}\left(\sum_{i=1}^N \mathbf{b}^T\mathbf{Y}_i\right)^2.$$

We can then go through exactly the same arguments as for the population case to find $\mathbf{a}^T\mathbf{X}$ and $\mathbf{b}^T\mathbf{Y}$ with $\mathbf{a} \in \text{Ker}(\widehat{\mathbf{K}}_X)^\perp, \; \mathbf{b} \in \text{Ker}(\widehat{\mathbf{K}}_Y)^\perp$, for $\widehat{\mathbf{K}}_X$ and $\widehat{\mathbf{K}}_Y$ defined in (2.10), such that

$$\hat{\rho}^2(\mathbf{a}, \mathbf{b}) = \frac{(\mathbf{a}^T\widehat{\mathbf{K}}_{XY}\mathbf{b})^2}{(\mathbf{a}^T\widehat{\mathbf{K}}_X\mathbf{a})(\mathbf{b}^T\widehat{\mathbf{K}}_Y\mathbf{b})}$$

is maximized. As before, such $\mathbf{a} \in \text{Ker}(\widehat{\mathbf{K}}_X)^\perp$ and $\mathbf{b} \in \text{Ker}(\widehat{\mathbf{K}}_Y)^\perp$ can be obtained by solving

$$\max_{\mathbf{a}\in \text{Ker}(\widehat{\mathbf{K}}_X)^\perp, \mathbf{b}\in \text{Ker}(\widehat{\mathbf{K}}_Y)^\perp} (\mathbf{a}^T\widehat{\mathbf{K}}_{XY}\mathbf{b})^2 \tag{3.16}$$

subject to

$$\mathbf{a}^T\widehat{\mathbf{K}}_X\mathbf{a} = \mathbf{b}^T\widehat{\mathbf{K}}_Y\mathbf{b} = 1. \tag{3.17}$$

Let

$$\widehat{\mathbf{K}}_X = \sum_{i=1}^{r(\widehat{\mathbf{K}}_X)} \hat{\lambda}_{Xi}\hat{\mathbf{e}}_{Xi}\hat{\mathbf{e}}_{Xi}^T,$$

where $(\hat{\lambda}_{Xi}, \hat{\mathbf{e}}_{Xi}),\ i = 1, \ldots, r(\widehat{\mathbf{K}}_X)$, are the nonzero eigenvalues and associated vectors of $\widehat{\mathbf{K}}_X$. Just like in the population setting, define

$$\widehat{\mathbf{K}}_X^{1/2} = \sum_{i=1}^{r(\widehat{\mathbf{K}}_X)} \hat{\lambda}_{Xi}^{1/2} \hat{\mathbf{e}}_{Xi} \hat{\mathbf{e}}_{Xi}^T, \quad \widehat{\mathbf{K}}_X^{-1/2} = \sum_{i=1}^{r(\widehat{\mathbf{K}}_X)} \hat{\lambda}_{Xi}^{-1/2} \hat{\mathbf{e}}_{Xi} \hat{\mathbf{e}}_{Xi}^T$$

and

$$\widehat{\mathbf{K}}_X^{-} = \sum_{i=1}^{r(\widehat{\mathbf{K}}_X)} \hat{\lambda}_{Xi}^{-1} \hat{\mathbf{e}}_{Xi} \hat{\mathbf{e}}_{Xi}^T.$$

Then, we can easily see that $\widehat{\mathbf{K}}_X^{-}$ and $\widehat{\mathbf{K}}_X^{-1/2}$ are the Moore-Penrose generalized inverses of $\widehat{\mathbf{K}}_X$ and $\widehat{\mathbf{K}}_X^{-1/2}$, respectively. Now, letting $\mathbf{u} = \widehat{\mathbf{K}}_X^{1/2}\mathbf{a}$ and $\mathbf{v} = \widehat{\mathbf{K}}_Y^{1/2}\mathbf{b}$ makes solving the problem (3.16) and (3.17) equivalent to solving

$$\max_{\mathbf{u} \in \mathrm{Ker}(\widehat{\mathbf{K}}_X)^{\perp}, \mathbf{v} \in \mathrm{Ker}(\widehat{\mathbf{K}}_Y)^{\perp}} (\mathbf{u}^T \widehat{\mathbf{K}}_X^{-1/2} \widehat{\mathbf{K}}_{XY} \widehat{\mathbf{K}}_Y^{-1/2} \mathbf{v})^2$$

subject to $\mathbf{u}^T\mathbf{u} = \mathbf{v}^T\mathbf{v} = 1$. Such $\mathbf{u}$ and $\mathbf{v}$ are obtained from the SVD of $\widehat{\mathbf{K}}_X^{-1/2} \widehat{\mathbf{K}}_{XY} \widehat{\mathbf{K}}_Y^{-1/2}$.

Define the first sample canonical correlation $\hat{\rho}_1$ and the associated weight vectors $\hat{\mathbf{a}}_1, \hat{\mathbf{b}}_1$ as

$$\hat{\rho}_1^2 = (\hat{\mathbf{a}}_1^T \widehat{\mathbf{K}}_{XY} \hat{\mathbf{b}}_1)^2 = \max_{\mathbf{a} \in \mathrm{Ker}(\widehat{\mathbf{K}}_X), \mathbf{b} \in \mathrm{Ker}(\widehat{\mathbf{K}}_Y)} (\mathbf{a}^T \widehat{\mathbf{K}}_{XY} \mathbf{b})^2, \tag{3.18}$$

where $\mathbf{a}, \mathbf{b}$ are subject to (3.17). For $i > 1$, define the ith sample canonical correlation $\hat{\rho}_i$ and the associated weight vectors $\hat{\mathbf{a}}_i, \hat{\mathbf{b}}_i$ as

$$\hat{\rho}_i^2 = (\hat{\mathbf{a}}_i^T \widehat{\mathbf{K}}_{XY} \hat{\mathbf{b}}_i)^2 = \max_{\mathbf{a} \in \mathrm{Ker}(\widehat{\mathbf{K}}_X), \mathbf{b} \in \mathrm{Ker}(\widehat{\mathbf{K}}_Y)} (\mathbf{a}^T \widehat{\mathbf{K}}_{XY} \mathbf{b})^2, \tag{3.19}$$

where $\mathbf{a}, \mathbf{b}$ are subject to (3.17) and

$$\mathbf{a}^T \widehat{\mathbf{K}}_X \hat{\mathbf{a}}_j = \mathbf{b}^T \widehat{\mathbf{K}}_Y \hat{\mathbf{b}}_j = 0,\ j < i. \tag{3.20}$$

Let $r = r(\widehat{\mathbf{K}}_X^{-1/2} \widehat{\mathbf{K}}_{XY} \widehat{\mathbf{K}}_Y^{-1/2}) \leq \min(r(\widehat{\mathbf{K}}_X), r(\widehat{\mathbf{K}}_Y))$. Then the SVD of the matrix $\widehat{\mathbf{K}}_X^{-1/2} \widehat{\mathbf{K}}_{XY} \widehat{\mathbf{K}}_Y^{-1/2}$ gives

$$\widehat{\mathbf{K}}_X^{-1/2} \widehat{\mathbf{K}}_{XY} \widehat{\mathbf{K}}_Y^{-1/2} = \sum_{i=1}^{r} \hat{\rho}_i^2 \hat{\mathbf{u}}_i \hat{\mathbf{v}}_i^T,$$

where $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_s$ are the eigenvectors corresponding to the eigenvalues $\hat{\rho}_1^2, \ldots, \hat{\rho}_s^2$ of

$$\widehat{\mathbf{K}}_X^{-1/2}\widehat{\mathbf{K}}_{XY}\widehat{\mathbf{K}}_Y^{-}\widehat{\mathbf{K}}_{YX}\widehat{\mathbf{K}}_X^{-1/2}$$

and $\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_s$ are the eigenvectors corresponding to the eigenvalues $\hat{\rho}_1^2, \ldots, \hat{\rho}_s^2$ of

$$\widehat{\mathbf{K}}_Y^{-1/2}\widehat{\mathbf{K}}_{YX}\widehat{\mathbf{K}}_X^{-}\widehat{\mathbf{K}}_{XY}\widehat{\mathbf{K}}_Y^{-1/2}.$$

Suppose that $\hat{\rho}_1^2 \geq \cdots \geq \hat{\rho}_s^2 > 0$. Then, $\hat{\mathbf{a}}_i = \widehat{\mathbf{K}}_X^{-1/2}\hat{\mathbf{u}}_i$ and $\hat{\mathbf{b}}_i = \widehat{\mathbf{K}}_Y^{-1/2}\hat{\mathbf{v}}_i$ solve the problem (3.16) subject to (3.17) and (3.20) with corresponding canonical correlation $\hat{\rho}_i$. The estimated canonical variables $\hat{\mathbf{a}}_i^T\mathbf{X},\ \hat{\mathbf{b}}_i^T\mathbf{Y}$ have maximum sample correlation with one another.

3.2 Discriminant Analysis with Less Than Full Rank Covariance Matrices

Let us now return to J population discriminant analysis problem of Section 2.2. In this setting we observe $(\mathbf{X}, G)$, where $\mathbf{X} \in \mathbb{R}^p$ is a predictor vector and $G \in \{1, \ldots, J\}$ is a categorical response variable representing the class memberships. Recall that class j has density f_j with class mean $\boldsymbol{\mu}_j$, covariance matrix $\mathbf{K}_j$ and associated class probability π_j.

3.2.1 Bayes Procedure: Linear Discriminant Analysis

Assume that the density of class j is normal with mean $\boldsymbol{\mu}_j$ and a common within class covariance matrix $\mathbf{K}_W$: i.e., $\mathbf{K}_j = \mathbf{K}_W$ for $j = 1, \ldots, J$. We will allow $\mathbf{K}_W$ to have less than full rank. This means that $r_W = r(\mathbf{K}_W) \leq p$.

Let $\mathbf{P}$ be an orthogonal matrix such that

$$\mathbf{K}_W = \mathbf{P}\begin{pmatrix} \mathbf{D} & \mathbf{O}_{r_W \times (p-r_W)} \\ \mathbf{O}_{(p-r_W)\times r_W} & \mathbf{O}_{(p-r_W)\times(p-r_W)} \end{pmatrix}\mathbf{P}^T$$

with $\mathbf{D} = \text{diag}(\lambda_{W1}, \ldots, \lambda_{Wr_W})$. Then $\mathbf{P} = [\mathbf{P}_1, \mathbf{P}_2]$ with $\mathbf{P}_1 = [\mathbf{e}_{W1}, \ldots, \mathbf{e}_{Wr_W}]$ a $p \times r_W$ matrix consisting of eigenvectors corresponding to $\lambda_{W1} \geq \cdots \geq \lambda_{Wr_W} > 0$ and $\mathbf{P}_1^T\mathbf{P}_2 =$

$\mathbf{O}_{r_W \times (p-r_W)}$, $\mathbf{P}_2^T\mathbf{P}_2 = \mathbf{I}_{p-r_W}$. Let $\boldsymbol{Z} = \mathbf{P}_1^T\mathbf{X}$ for an observation vector $\mathbf{X}$ from class j. Then, a Bayes discrimination paradigm can be developed by assuming that

$$\boldsymbol{Z}|G = j \sim N_{r_W}(\boldsymbol{\upsilon}_j, \mathbf{D})$$

with $\boldsymbol{\upsilon}_j = \mathbf{P}_1^T\boldsymbol{\mu}_j$.

Let $\mathbf{K}_W^-$ be the Moore-Penrose generalized inverse of $\mathbf{K}_W$ defined by

$$\mathbf{K}_W^- = \sum_{i=1}^{r_W} \lambda_{Wi}^{-1}\mathbf{e}_{Wi}\mathbf{e}_{Wi}^T = \mathbf{P}_1\mathbf{D}^{-1}\mathbf{P}_1^T.$$

Since a Bayesian classifier assigns a new observation to the group with the largest posterior probability, we classify a new observation $\mathbf{x}$ to population i if

$$P(G = i|\boldsymbol{z}) = \max_j P(G = j|\boldsymbol{z}), \tag{3.21}$$

where

$$\begin{aligned}
P(G = j|\boldsymbol{z}) &\propto \exp\left[-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\upsilon}_j)^T\mathbf{D}^{-1}(\boldsymbol{z} - \boldsymbol{\upsilon}_j) + \log\pi_j\right] \\
&\propto \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^T\mathbf{P}_1\mathbf{D}^{-1}\mathbf{P}_1^T(\mathbf{x} - \boldsymbol{\mu}_j) + \log\pi_j\right] \\
&\propto \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^T\mathbf{K}_W^-(\mathbf{x} - \boldsymbol{\mu}_j) + \log\pi_j\right] \\
&\propto \exp\left[\boldsymbol{\mu}_j^T\mathbf{K}_W^-\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j^T\mathbf{K}_W^-\boldsymbol{\mu}_j + \log\pi_j\right].
\end{aligned}$$

Alternatively, we can define the discriminant function for class j to be

$$d_j(\mathbf{x}) = \boldsymbol{\mu}_j^T\mathbf{K}_W^-\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j^T\mathbf{K}_W^-\boldsymbol{\mu}_j + \log\pi_j. \tag{3.22}$$

Then, an equivalent rule to (3.21) is to classify $\mathbf{x}$ to the class for which $d_j(\mathbf{x})$ is largest. Note that this has the consequence that, in the case where we have equal class probabilities, a new observation is classified to the class with the closest centroid or mean vector using the squared generalized Mahalanobis distance

$$D_j(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_j)^T\mathbf{K}_W^-(\mathbf{x} - \boldsymbol{\mu}_j). \tag{3.23}$$

3.2.2 Bayes Procedure: Quadratic Discriminant Analysis

We now allow for different covariance matrices $\mathbf{K}_1, \ldots, \mathbf{K}_J$ for each class with $\mathbf{K}_j$ having rank $r_j = r(\mathbf{K}_j) \leq p$. Let $\mathbf{P}_{1j}$ be a $p \times r_j$ matrix such that $\mathbf{P}_{1j}^T \mathbf{P}_{1j} = \mathbf{I}_{r_j}$ and $\mathbf{K}_j = \mathbf{P}_{1j} \mathbf{D}_j \mathbf{P}_{1j}^T$ with $\mathbf{D}_j$ a diagonal matrix whose elements are the r_j positive eigenvalues of $\mathbf{K}_j$. Now let $\boldsymbol{Z}_j = \mathbf{P}_{1j}^T \mathbf{X}$. Then, if

$$\boldsymbol{Z}_j | G = j \sim N_{r_j}(\boldsymbol{\upsilon}_j, \mathbf{D}_j)$$

with $\boldsymbol{\upsilon}_j = \mathbf{P}_{1j}^T \boldsymbol{\mu}_j$, the corresponding Bayesian classification rule follows from (3.21).

We know that for $\boldsymbol{z}_j = \mathbf{P}_{1j}^T \mathbf{x}$

$$\begin{aligned} P(G = j | \boldsymbol{z}_j) &\propto \exp\left[-\frac{1}{2} (\boldsymbol{z}_j - \boldsymbol{\upsilon}_j)^T \mathbf{D}_j^{-1} (\boldsymbol{z}_j - \boldsymbol{\upsilon}_j) - \frac{1}{2} \log |\mathbf{D}_j| + \log \pi_j \right] \\ &\propto \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)^T \mathbf{K}_j^- (\mathbf{x} - \boldsymbol{\mu}_j) - \frac{1}{2} \log |\mathbf{D}_j| + \log \pi_j \right] \end{aligned}$$

with $\mathbf{K}_j^-$ the Moore-Penrose generalized inverses of $\mathbf{K}_j$. Hence, we classify $\mathbf{x}$ to class i if

$$d_i^Q(\mathbf{x}) = \min_j d_j^Q(\mathbf{x}), \tag{3.24}$$

where the quadratic discriminant function is

$$d_j^Q(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_j)^T \mathbf{K}_j^- (\mathbf{x} - \boldsymbol{\mu}_j) + \log |\mathbf{D}_j| - 2 \log \pi_j. \tag{3.25}$$

3.2.3 Fisher's Linear Discriminant Analysis

The next steps in our development involve the extension of our less than full rank developments to discriminant analysis via Fisher's method and, eventually, with canonical correlation analysis. We begin with how to formulate Fisher's linear discriminant function in the case that $\mathbf{K}_W$ has less than full rank.

3.2.3.1 Population linear discriminant function

Let $\mathbf{K}_B$ be the between-class covariance matrix as defined in Section 2.2.3.1. Recall that

$$\mathbf{K}_B = \text{Var}_G(\text{E}[\mathbf{X}|G]) = \sum_{j=1}^{J} \pi_j (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T$$

for

$$\boldsymbol{\mu} = \mathrm{E}[\mathbf{X}] = \sum_{j=1}^{J} \pi_j \boldsymbol{\mu}_j.$$

Then, Fisher's linear discriminant function is defined to be the linear function $\boldsymbol{l}^T\mathbf{X}$ which maximizes the ratio of the between-class variance to the within-class variance given by

$$\frac{\mathrm{Var}_G(\mathrm{E}[\boldsymbol{l}^T\mathbf{X}|G])}{\mathrm{E}_G[\mathrm{Var}(\boldsymbol{l}^T\mathbf{X}|G)]} = \frac{\boldsymbol{l}^T\mathbf{K}_B\boldsymbol{l}}{\boldsymbol{l}^T\mathbf{K}_W\boldsymbol{l}} \tag{3.26}$$

provided that $\mathbf{K}_W\boldsymbol{l} \neq \mathbf{0}$.

Assume that $\boldsymbol{\mu}_j \in \mathrm{Ker}(\mathbf{K}_W)^\perp$ for all j. Then, the columns and rows of $\mathbf{K}_B$ belong to $\mathrm{Ker}(\mathbf{K}_W)^\perp$ and hence, for $\boldsymbol{l} = \boldsymbol{l}_* + \boldsymbol{l}_0$ with $\boldsymbol{l}_* \in \mathrm{Ker}(\mathbf{K}_W)^\perp$ and $\boldsymbol{l}_0 \in \mathrm{Ker}(\mathbf{K}_W)$, (3.26) becomes

$$\frac{\boldsymbol{l}_*^T\mathbf{K}_B\boldsymbol{l}_*}{\boldsymbol{l}_*^T\mathbf{K}_W\boldsymbol{l}_*}$$

as in Section 3.1.1. Thus, we now wish to find $\boldsymbol{l} = (l_1, \ldots, l_p)^T$ satisfying

$$\max_{\boldsymbol{l} \in \mathrm{Ker}(\mathbf{K}_W)^\perp} \boldsymbol{l}^T\mathbf{K}_B\boldsymbol{l}, \tag{3.27}$$

where $\boldsymbol{l}$ is subject to

$$\boldsymbol{l}^T\mathbf{K}_W\boldsymbol{l} = 1. \tag{3.28}$$

This is equivalent to solving

$$\max_{\substack{\mathbf{u} \in \mathrm{Ker}(\mathbf{K}_W)^\perp \\ ||\mathbf{u}||_{\mathbb{R}^p}=1}} \mathbf{u}^T\mathbf{K}_W^{-1/2}\mathbf{K}_B\mathbf{K}_W^{-1/2}\mathbf{u},$$

where $\mathbf{K}_W^{-1/2}$ is defined as

$$\mathbf{K}_W^{-1/2} = \sum_{j=1}^{r_W} \lambda_{Wi}^{-1/2}\mathbf{e}_{Wi}\mathbf{e}_{Wi}^T$$

with $r_W = r(\mathbf{K}_W) \leq p$ and $(\lambda_{Wi}, \mathbf{e}_{Wi})$ the pairs of positive eigenvalues and associated eigenvectors for $\mathbf{K}_W$.

The optimal $\mathbf{u}$ can be obtained from the spectral decomposition of the matrix

$$\mathbf{K}_W^{-1/2}\mathbf{K}_B\mathbf{K}_W^{-1/2},$$

which produces the optimal $\boldsymbol{l}$ vector

$$\boldsymbol{l} = \mathbf{K}_W^{-1/2}\mathbf{u}.$$

Alternatively $\boldsymbol{l}$ may be characterized directly as the solution of

$$\mathbf{K}_W^{-}\mathbf{K}_B\boldsymbol{l} = \gamma\boldsymbol{l} \tag{3.29}$$

subject to condition (3.28).

Let r be the rank of $\mathbf{K}_W^{-}\mathbf{K}_B$. Then, $r = r(\mathbf{K}_W^{-}\mathbf{K}_B) \leq \min(r(\mathbf{K}_W), r(\mathbf{K}_B)) = \min(r_W, J-1)$. Also, let $\boldsymbol{l}_i, i = 1, \ldots, r$, be the solutions to (3.29) corresponding to the eigenvalues $\gamma_1 \geq \cdots \geq \gamma_r > 0$. We will refer to $\boldsymbol{l}_i^T\mathbf{X}$ as discriminators or discriminant functions.

Take the first s ($\leq r$) discriminators corresponding to the first s largest eigenvalues of $\mathbf{K}_W^{-}\mathbf{K}_B$. Then the classification rule based on a subset $\boldsymbol{l}_1^T\mathbf{X}, \ldots, \boldsymbol{l}_s^T\mathbf{X}$ of the discriminant functions is to classify an observation $\mathbf{x}$ to class i if

$$Dist_i^s(\mathbf{x}) = \min_j Dist_j^s(\mathbf{x}), \tag{3.30}$$

where the squared Mahalanobis distance $Dist_j^s$ is given by

$$Dist_j^s(\mathbf{x}) = \sum_{k=1}^{s}(\boldsymbol{l}_k^T\mathbf{x} - \boldsymbol{l}_k^T\boldsymbol{\mu}_j)^2. \tag{3.31}$$

The assumption $\boldsymbol{\mu}_j \in \text{Ker}(\mathbf{K}_W)^{\perp}$ for all j implies that

$$\text{Ker}(\mathbf{K}_W)^{\perp} = \text{Ker}(\mathbf{K}_X)^{\perp}.$$

We can easily see that $\text{Ker}(\mathbf{K}_X) \subset \text{Ker}(\mathbf{K}_W)$ and, hence, $\text{Ker}(\mathbf{K}_W)^{\perp} \subset \text{Ker}(\mathbf{K}_X)^{\perp}$. Conversely, we observe that for $\mathbf{c} \in \mathbb{R}^p$,

$$\mathbf{K}_X\mathbf{c} = \mathbf{K}_B\mathbf{c} + \mathbf{K}_W\mathbf{c} = \sum_{j=1}^{J}\{\pi_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T\mathbf{c}\}\boldsymbol{\mu}_j + \mathbf{K}_W\mathbf{c} \in \text{Ker}(\mathbf{K}_W)^{\perp}$$

and so $\text{Ker}(\mathbf{K}_X)^\perp \subset \text{Ker}(\mathbf{K}_W)^\perp$.

We now introduce another formulation of Fisher's discrimination method. Let us recall that $\mathbf{K}_X$ is the covariance matrix representing total variability and

$$\mathbf{K}_X = \text{Var}(\mathbf{X}) = \text{Var}_G(\text{E}[\mathbf{X}|G]) + \text{E}_G[\text{Var}(\mathbf{X}|G)] = \mathbf{K}_B + \mathbf{K}_W.$$

Thus, let us consider optimization with respect to $\boldsymbol{l}$ of the ratio

$$\boldsymbol{l}^T\mathbf{K}_B\boldsymbol{l}/\boldsymbol{l}^T\mathbf{K}_X\boldsymbol{l} \tag{3.32}$$

when $\mathbf{K}_X\boldsymbol{l} \neq \mathbf{0}$. In this regards we claim that maximizing (3.26) over $\boldsymbol{l} \in \text{Ker}(\mathbf{K}_W)^\perp$ is equivalent to maximizing (3.32) over $\boldsymbol{l} \in \text{Ker}(\mathbf{K}_X)^\perp$. The validity of this contention is established by first noting that since $\text{Ker}(\mathbf{K}_W)^\perp = \text{Ker}(\mathbf{K}_X)^\perp$, (3.26) becomes

$$\frac{\boldsymbol{l}^T\mathbf{K}_B\boldsymbol{l}}{\boldsymbol{l}^T\mathbf{K}_W\boldsymbol{l}} = \frac{\boldsymbol{l}^T\mathbf{K}_B\boldsymbol{l}/\boldsymbol{l}^T\mathbf{K}_X\boldsymbol{l}}{1 - \boldsymbol{l}^T\mathbf{K}_B\boldsymbol{l}/\boldsymbol{l}^T\mathbf{K}_X\boldsymbol{l}}$$

and then recognizing that $h(x) = \frac{x}{1-x}$ is an increasing function for $0 \leq x < 1$.

As in the Fisher's discriminant problem in (3.26), the optimal $\boldsymbol{l}$ can be characterized as the solution of

$$\mathbf{K}_X^-\mathbf{K}_B\boldsymbol{l} = \lambda\boldsymbol{l} \tag{3.33}$$

and $\boldsymbol{l}$ must satisfy $\boldsymbol{l}^T\mathbf{K}_X\boldsymbol{l} = 1$. Now (3.33) is equivalent to

$$\mathbf{K}_B\boldsymbol{l} = \lambda\mathbf{K}_X\boldsymbol{l} = \lambda(\mathbf{K}_B + \mathbf{K}_W)\boldsymbol{l} \tag{3.34}$$

or

$$\mathbf{K}_W^-\mathbf{K}_B\boldsymbol{l} = \frac{\lambda}{1-\lambda}\boldsymbol{l}, \tag{3.35}$$

because $\boldsymbol{l} \in \text{Ker}(\mathbf{K}_X)^\perp = \text{Ker}(\mathbf{K}_W)^\perp$. So, the solutions in (3.29) and (3.35) are the same apart from a normalizing factor. Since the solutions of (3.35) satisfy $\boldsymbol{l}^T\mathbf{K}_X\boldsymbol{l} = 1$, we have

$$\boldsymbol{l}^T\mathbf{K}_X\boldsymbol{l} = \boldsymbol{l}^T\mathbf{K}_W\boldsymbol{l} + \boldsymbol{l}^T\mathbf{K}_B\boldsymbol{l} = 1$$

so that

$$\boldsymbol{l}^T\mathbf{K}_W\boldsymbol{l} = 1 - \boldsymbol{l}^T\mathbf{K}_B\boldsymbol{l} = 1 - \lambda\boldsymbol{l}^T\mathbf{K}_X\boldsymbol{l} = 1 - \lambda.$$

Let $\boldsymbol{l}_{\text{Fisher}}$ and $\boldsymbol{l}$ be the solutions of (3.29) and (3.35), respectively. Then, $\boldsymbol{l}_{\text{Fisher}}$ is related to $\boldsymbol{l}$ in that

$$\boldsymbol{l}_{\text{Fisher}} = \frac{\boldsymbol{l}}{(\boldsymbol{l}^T\mathbf{K}_W\boldsymbol{l})^{1/2}} = (1-\lambda)^{-1/2}\boldsymbol{l}.$$

Thus, if $\boldsymbol{l}_1, \ldots, \boldsymbol{l}_r$ are solutions of (3.33) or (3.35) corresponding to eigenvalues $\lambda_1, \ldots, \lambda_r$ then given $s \leq r$,

$$Dist_j^s(\mathbf{x}) = \sum_{k=1}^{s} \frac{1}{1-\lambda_k}\left(\boldsymbol{l}_k^T\mathbf{x} - \boldsymbol{l}_k^T\boldsymbol{\mu}_j\right)^2. \tag{3.36}$$

We have shown that the vectors that maximize

$$\boldsymbol{l}^T\mathbf{K}_B\boldsymbol{l}/\boldsymbol{l}^T\mathbf{K}_W\boldsymbol{l}$$

and

$$\boldsymbol{l}^T\mathbf{K}_B\boldsymbol{l}/\boldsymbol{l}^T\mathbf{K}_X\boldsymbol{l}$$

are identical apart from scaling factors. We also have shown that the vector that maximize (3.32) in $\text{Ker}(\mathbf{K}_W)^\perp$ is identical to the vector that maximize (3.32) in $\text{Ker}(\mathbf{K}_X)^\perp$. We further view $\boldsymbol{l}^T\mathbf{K}_B\boldsymbol{l}/\boldsymbol{l}^T\mathbf{K}_X\boldsymbol{l}$ as more interpretable of the two criteria since it is similar in nature to a coefficient of determination. So, we now name the optimization problem in (3.32) a generalized Fisher's linear discriminant analysis.

The condition that $\boldsymbol{\mu}_j \in \text{Ker}(\mathbf{K}_W)^\perp$ for all j is connected to "estimability" of linear functionals $\boldsymbol{l}^T\boldsymbol{\mu}_j$ for $\boldsymbol{l} \in \text{Ker}(\mathbf{K}_W)^T$. Indeed, for $\boldsymbol{l} \in \text{Ker}(\mathbf{K}_W)^\perp$,

$$\text{E}[\boldsymbol{l}^T\mathbf{X}|G = j] = \boldsymbol{l}^T\boldsymbol{\mu}_j$$

is unique if and only if $\boldsymbol{\mu}_j \in \text{Ker}(\mathbf{K}_W)^\perp$. To see this, suppose that for some $\boldsymbol{\mu}_j^{(1)}, \boldsymbol{\mu}_j^{(2)}$ in $\text{Ker}(\mathbf{K}_W)^\perp$ with $\boldsymbol{\mu}_j^{(1)} \neq \boldsymbol{\mu}_j^{(2)}$ we had $\boldsymbol{l}^T\boldsymbol{\mu}_j^{(1)} = \boldsymbol{l}^T\boldsymbol{\mu}_j^{(2)}$ for $\boldsymbol{l} \in \text{Ker}(\mathbf{K}_W)^\perp$. Then, this would produce the contradiction that $\boldsymbol{\mu}_j^{(1)} - \boldsymbol{\mu}_j^{(2)} \in \text{Ker}(\mathbf{K}_W)$. But, $\boldsymbol{\mu}_j^{(1)} - \boldsymbol{\mu}_j^{(2)} \in \text{Ker}(\mathbf{K}_W)^\perp$ and so $\boldsymbol{\mu}_j^{(1)} = \boldsymbol{\mu}_j^{(2)}$.

3.2.3.2 Sample linear discriminant analysis

Let $(\mathbf{X}_1, G_1), \ldots, (\mathbf{X}_N, G_N)$ be iid copies of $(\mathbf{X}, G)$. Much like our approach for canonical correlation analysis, sample discriminant functions can be obtained from estimators of $\mathbf{K}_W$ and $\mathbf{K}_B$. For this purpose we will use $\widehat{\mathbf{K}}_W, \widehat{\mathbf{K}}_B$ and $\widehat{\mathbf{K}}_X$ as defined in Section 2.2.4.

The sample linear discriminant function based on the Bayes' classifier is

$$\widehat{d}_j(\mathbf{x}) = \bar{\mathbf{x}}_j^T \widehat{\mathbf{K}}_W^- \mathbf{x} - \frac{1}{2} \bar{\mathbf{x}}_j^T \widehat{\mathbf{K}}_W^- \bar{\mathbf{x}}_j + \log p_j.$$

One then classifies $\mathbf{x}$ to the population where $\widehat{d}_j(\mathbf{x})$ is largest.

For Fisher's discriminant functions we use the solutions of

$$\widehat{\mathbf{K}}_W^- \widehat{\mathbf{K}}_B \boldsymbol{l} = \gamma \boldsymbol{l}$$

subject to $\boldsymbol{l}^T \widehat{\mathbf{K}}_W \boldsymbol{l} = 1$. If $(\hat{\gamma}_i, \hat{\boldsymbol{l}}_i), i = 1, \ldots, s$, are the solutions corresponding to the first s largest eigenvalues, then $\mathbf{x}$ is classified into population i if

$$\widehat{Dist}_i^s(\mathbf{x}) = \min_j \widehat{Dist}_j^s(\mathbf{x})$$

for

$$\widehat{Dist}_j^s(\mathbf{x}) = \sum_{k=1}^{s} (\hat{\boldsymbol{l}}_k^T \mathbf{x} - \hat{\boldsymbol{l}}_k^T \bar{\mathbf{X}}_j)^2.$$

3.2.4 Fisher's LDA and Bayes Procedures

Suppose $J = 2$ and the class probabilities are equal. Assume, also, that $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are in $\text{Ker}(\mathbf{K}_W)^\perp$. Then, the Bayesian classification rule in (3.21) is equivalent to classifying $\mathbf{x}$ to class 1 if

$$d_1(\mathbf{x}) - d_2(\mathbf{x}) > 0.$$

Otherwise, it is classified to class 2. Since

$$d_1(\mathbf{x}) - d_2(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{K}_W^- \mathbf{x} - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{K}_W^- (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2),$$

the classification rule is equivalent to classifying $\mathbf{x}$ to class 1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{K}_W^- (\mathbf{x} - \boldsymbol{\mu}) > 0,$$

with $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$ and to class 2 otherwise.

For $J = 2$, $\boldsymbol{\mu} = \pi_1 \boldsymbol{\mu}_1 + \pi_2 \boldsymbol{\mu}_2$ so that $\mathbf{K}_B = \sum_{j=1}^{2} \pi_j (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T = \pi_1 \pi_2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$. Thus, Fisher's linear discriminant function is obtained by maximizing

$$\frac{\left\{\boldsymbol{l}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right\}^2}{\boldsymbol{l}^T \mathbf{K}_W \boldsymbol{l}}$$

over $\boldsymbol{l}$ in $\text{Ker}(\mathbf{K}_W)^\perp$. An application of the Cauchy-Schwarz inequality reveals that the maximum of the above ratio is $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{K}_W^- (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and the maximum is attained at $\boldsymbol{l} = \mathbf{K}_W^- (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Thus, Fisher's linear discriminant function is $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{K}_W^- \mathbf{X}$. In this instance, the classification rule is to classify $\mathbf{x}$ to class 1 if

$$\left|(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{K}_W^- \mathbf{x} - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{K}_W^- \boldsymbol{\mu}_1\right| < \left|(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{K}_W^- \mathbf{x} - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{K}_W^- \boldsymbol{\mu}_2\right|,$$

which is exactly the same as the rule obtained from the Bayes procedure.

Similarly, the generalized Fisher's linear discriminant function is obtained by solving

$$\max_{\boldsymbol{l} \in \text{Ker}(\mathbf{K}_X)^\perp} \frac{\left\{\boldsymbol{l}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right\}^2}{\boldsymbol{l}^T \mathbf{K}_X \boldsymbol{l}}.$$

So, the generalized Fisher's linear discriminant function is $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{K}_X^- \mathbf{X}$.

Now let us recall that under the assumption $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \text{Ker}(\mathbf{K}_W)^\perp$, $\text{Ker}(\mathbf{K}_X)^\perp = \text{Ker}(\mathbf{K}_W)^\perp$. Let $\boldsymbol{l} = \mathbf{K}_X^- (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Then, we have

$$\mathbf{K}_X \boldsymbol{l} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$$

and hence

$$\mathbf{K}_W \boldsymbol{l} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 - \mathbf{K}_B \boldsymbol{l} = (1 - 0.25(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{l})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

which implies

$$\mathbf{K}_W \boldsymbol{l} \propto \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2.$$

Since in this case $\text{Ker}(\mathbf{K}_X)^{\perp} = \text{Ker}(\mathbf{K}_W)^{\perp}$, we see that

$$\boldsymbol{l} \propto \mathbf{K}_W^{-}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Then, the corresponding classification rule for generalized Fisher's discrimination is to classify $\mathbf{x}$ to class 1 if

$$\left|(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{K}_X^{-}(\mathbf{x} - \boldsymbol{\mu}_1)\right| < \left|(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{K}_X^{-}(\mathbf{x} - \boldsymbol{\mu}_2)\right|.$$

Thus, the result is that classification via generalized Fisher's discriminant analysis is exactly the same as that by the Bayes classifier because $\boldsymbol{l} \propto \mathbf{K}_W^{-}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

3.2.5 Fisher's Discriminant Function via Regression

Let us recall the generalized Fisher's linear discriminant function, $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{K}_X^{-}\mathbf{X}$, in the case of $J = 2$ with $\pi_1 = \pi_2 = 0.5$ from Section 3.2.4. Also, recall the form of linear regression for a binary response Y coded as 1 for class 1 and 0 for class 2 on a vector $\mathbf{X} \in \mathbb{R}^p$. Specifically, we have the following regression line:

$$\mathbf{K}_{YX}\mathbf{K}_X^{-}\mathbf{X} + (\mu_Y - \mathbf{K}_{YX}\mathbf{K}_X^{-}\boldsymbol{\mu}_X).$$

Then, we can observe that

$$\mathbf{K}_{XY} = \text{E}[\mathbf{X}Y] - \text{E}[\mathbf{X}]\text{E}[Y] = \pi_1\boldsymbol{\mu}_1 - \boldsymbol{\mu}\pi_1 = \pi_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}) = \pi_1\pi_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

since Y is a Bernoulli random variable with $P(Y = 1) = P(G = 1) = \pi_1$. So the slope of the regression line is proportional to $\mathbf{K}_X^{-}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and this quantity is exactly the same as the generalized Fisher's linear discriminant function in the case of two classes.

Suppose that the classification rule is defined to allocate $\mathbf{x}$ to class 1 if

$$\mathbf{K}_{YX}\mathbf{K}_X^{-}\mathbf{x} + (\mu_Y - \mathbf{K}_{YX}\mathbf{K}_X^{-}\boldsymbol{\mu}) > .5$$

and to class 2 otherwise. Then, in the case that $\pi_1 = \pi_2$, this rule becomes exactly the same as the rule from generalized Fisher's discriminant analysis. Finally, by assuming $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \text{Ker}(\mathbf{K}_W)^{\perp}$, this leads to the result that the rule obtained from linear regression is also the same as the rule from Fisher's discriminant analysis or from the Bayes classifier.

3.2.6 Fisher's Approach via Canonical Correlation Analysis

As in Section 2.2.3.2, we will demonstrate the connection between Fisher's LDA and canonical correlation analysis under the less than full rank scenario in this section. For this purpose, let $\mathbf{Y} = (Y_1, \ldots, Y_J)^T$ with $Y_j = I(G = j)$ being indicator response variables. Then, we will prove the following result.

THEOREM **III.2**. *Let $\mathbf{K}_B, \mathbf{K}_W$ be the between-class covariance matrix and a common within-class covariance matrix, respectively, defined in Section 2.2.3.1. Let $\mathbf{a}_i, i = 1, \ldots, r$, be the coefficient vectors of the canonical variables of the $\mathbf{X}$ space. Then, the canonical vectors $\mathbf{a}_i$ are the eigenvectors of $\mathbf{K}_W^{-}\mathbf{K}_B$ and the canonical correlations ρ_i are precisely square roots of the eigenvalues obtained from (3.33).*

Proof. Set $\text{Var}(\mathbf{X}) = \mathbf{K}_X, \text{Var}(\mathbf{Y}) = \mathbf{K}_Y$ and $\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{K}_{XY} = \mathbf{K}_{YX}^T$. Since $\mathbf{Y} = \mathbf{Y}(G)$ from the categorical response variable G is such that $\mathbf{Y} = \mathbf{e}_j$ if $G = j$ for $j = 1, \ldots, J$, with $\mathbf{e}_j$ an elementary vector consisting of all 0's except for a 1 in its jth entry, $\mathbf{Y}$ has a multinomial distribution with cell probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)^T$. Consequently,

$$\text{E}[\mathbf{Y}] = \boldsymbol{\pi}, \quad \mathbf{K}_Y = \text{diag}(\pi_1, \ldots, \pi_J) - \boldsymbol{\pi}\boldsymbol{\pi}^T.$$

Also, we can show that

$$\begin{aligned}\text{E}[\mathbf{X}\mathbf{Y}^T] = \text{E}_G[\text{E}(\mathbf{X}\mathbf{Y}^T|G)] &= \sum_{j=1}^{J} \text{E}[\mathbf{X}|G = j]\mathbf{Y}(G = j)^T P(G = j) \\ &= \sum_{j=1}^{J} \pi_j \boldsymbol{\mu}_j \mathbf{e}_j^T = [\pi_1\boldsymbol{\mu}_1, \ldots, \pi_J\boldsymbol{\mu}_J]\end{aligned}$$

and hence we have

$$\mathbf{K}_{XY} = \mathrm{E}[\mathbf{X}\mathbf{Y}^T] - \mathrm{E}[\mathbf{X}]\mathrm{E}[\mathbf{Y}]^T = [\pi_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J(\boldsymbol{\mu}_J - \boldsymbol{\mu})].$$

Now, the canonical correlation problem is to find $\mathbf{a} \in \mathrm{Ker}(\mathbf{K}_X)^\perp$ and $\mathbf{b} \in \mathrm{Ker}(\mathbf{K}_Y)^\perp$ that maximize

$$\frac{(\mathbf{a}^T\mathbf{K}_{XY}\mathbf{b})^2}{(\mathbf{a}^T\mathbf{K}_X\mathbf{a})(\mathbf{b}^T\mathbf{K}_Y\mathbf{b})} = \frac{\left\{\mathbf{a}^T\left[\pi_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J(\boldsymbol{\mu}_J - \boldsymbol{\mu})\right]\mathbf{b}\right\}^2}{(\mathbf{a}^T\mathbf{K}_X\mathbf{a})(\mathbf{b}^T\mathbf{K}_Y\mathbf{b})}. \tag{3.37}$$

To accomplish this set $\mathbf{c} = \mathrm{diag}(\pi_1^{1/2}, \ldots, \pi_J^{1/2})\mathbf{b} = (\pi_1^{1/2}b_1, \ldots, \pi_J^{1/2}b_J)^T$ and observe that

$$\mathbf{b}^T\mathbf{K}_Y\mathbf{b} = \mathbf{c}^T(\mathbf{I} - \mathbf{d}\mathbf{d}^T)\mathbf{c}$$

with $\mathbf{d} = (\pi_1^{1/2}, \ldots, \pi_J^{1/2})^T$. So, (3.37) becomes

$$\frac{\left\{\mathbf{a}^T\left[\pi_1^{1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J^{1/2}(\boldsymbol{\mu}_J - \boldsymbol{\mu})\right]\mathbf{c}\right\}^2}{(\mathbf{a}^T\mathbf{K}_X\mathbf{a})(\mathbf{c}^T(\mathbf{I} - \mathbf{d}\mathbf{d}^T)\mathbf{c})}.$$

Since $\mathbf{b} \in \mathrm{Ker}(\mathbf{K}_Y)^\perp$ is equivalent to $\mathbf{c} \in \mathrm{Ker}(\mathbf{I} - \mathbf{d}\mathbf{d}^T)^\perp$ and $\mathbf{d} \in \mathrm{Ker}(\mathbf{I} - \mathbf{d}\mathbf{d}^T)$,

$$\mathbf{d}^T\mathbf{c} = 0. \tag{3.38}$$

Thus, in words, finding $\mathbf{a}$ and $\mathbf{b}$ such that $\mathbf{a} \in \mathrm{Ker}(\mathbf{K}_X)^\perp,\ \mathbf{b} \in \mathrm{Ker}(\mathbf{K}_Y)^\perp$ maximize (3.37) is equivalent to finding $\mathbf{a}$ and $\mathbf{c}$ such that

$$\frac{\left\{\mathbf{a}^T\left[\pi_1^{1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J^{1/2}(\boldsymbol{\mu}_J - \boldsymbol{\mu})\right]\mathbf{c}\right\}^2}{(\mathbf{a}^T\mathbf{K}_X\mathbf{a})(\mathbf{c}^T\mathbf{c})}. \tag{3.39}$$

is maximized over $\mathbf{a} \in \mathrm{Ker}(\mathbf{K}_X)^\perp,\ \mathbf{c} \in \mathrm{Ker}(\mathbf{I} - \mathbf{d}\mathbf{d}^T)^\perp$. As in Section 2.1.1, the coefficient vectors $\mathbf{a}_1, \ldots, \mathbf{a}_r$ of the canonical variables of the $\mathbf{X}$ space are then obtained from

$$\mathbf{K}_X^-\left[\pi_1^{1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J^{1/2}(\boldsymbol{\mu}_J - \boldsymbol{\mu})\right]\left[\pi_1^{1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J^{1/2}(\boldsymbol{\mu}_J - \boldsymbol{\mu})\right]^T\mathbf{a} = \rho^2\mathbf{a}. \tag{3.40}$$

Since $\mathbf{K}_B = \sum_{j=1}^J \pi_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T$, (3.40) simplifies to

$$\mathbf{K}_X^-\mathbf{K}_B\mathbf{a} = \rho^2\mathbf{a}$$

or

$$\mathbf{K}_W^- \mathbf{K}_B \mathbf{a} = \frac{\rho^2}{1-\rho^2}\mathbf{a}$$

as was to be shown.

□

We also can find $\mathbf{c}$ such that $\mathbf{d}^T\mathbf{c} = 0$ to maximize (3.39). The corresponding $\mathbf{c}$'s are obtained from

$$\left[\pi_1^{1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J^{1/2}(\boldsymbol{\mu}_J - \boldsymbol{\mu})\right]^T \mathbf{K}_X^- \left[\pi_1^{1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J^{1/2}(\boldsymbol{\mu}_J - \boldsymbol{\mu})\right] \mathbf{c} = \rho^2 \mathbf{c}. \tag{3.41}$$

The moral of this is that the canonical variables of the $\mathbf{Y}$ space look like $\mathbf{b}^T\mathbf{Y}$ with

$$\mathbf{b} = (\pi_1^{-1/2}c_1, \ldots, \pi_J^{-1/2}c_J)^T$$

for $\mathbf{c} = (c_1, \ldots, c_J)^T$ such that $\sum_{j=1}^J \pi_j^{1/2} c_j = 0$ and $\mathbf{c}^T\mathbf{c} = 1$. Also, because $\boldsymbol{\pi}^T\mathbf{b} = \mathbf{d}^T\mathbf{c} = 0$, we can see that

$$[\pi_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J(\boldsymbol{\mu}_J - \boldsymbol{\mu})]\,\mathbf{b} = \sum_{j=1}^J \pi_j b_j \boldsymbol{\mu}_j - \boldsymbol{\mu}\sum_{J=1}^J \pi_j b_j = \sum_{j=1}^J \pi_j b_j \boldsymbol{\mu}_j,$$

which is a contrast among the population means. So, the numerator in (3.37) is simplified to

$$\mathbf{a}^T\,[\pi_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J(\boldsymbol{\mu}_J - \boldsymbol{\mu})]\,\mathbf{b} = \sum_{j=1}^J \pi_j b_j \mathbf{a}^T\boldsymbol{\mu}_j,$$

which is a contrast among transformed means $m_j = \mathbf{a}^T\boldsymbol{\mu}_j$. Moreover, from $\boldsymbol{\pi}^T\mathbf{b} = 0$, we see that

$$\mathbf{b}^T\mathbf{K}_Y\mathbf{b} = \mathbf{b}^T\left(\text{diag}(\pi_1, \ldots, \pi_J) - \boldsymbol{\pi}\boldsymbol{\pi}^T\right)\mathbf{b} = \mathbf{b}^T\text{diag}(\pi_1, \ldots, \pi_J)\mathbf{b} = \mathbf{c}^T\mathbf{c} = 1.$$

Now, premultiplying (3.41) by $\mathbf{K}_X^-\left[\pi_1^{1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J^{1/2}(\boldsymbol{\mu}_J - \boldsymbol{\mu})\right]$ reveals that

$$\mathbf{K}_X^-\mathbf{K}_B(\mathbf{K}_X^-\mathbf{z}) = \rho^2(\mathbf{K}_X^-\mathbf{z})$$

with $\mathbf{z} = \left[\pi_1^{1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J^{1/2}(\boldsymbol{\mu}_J - \boldsymbol{\mu})\right]^T \mathbf{c} = [\pi_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J(\boldsymbol{\mu}_J - \boldsymbol{\mu})]\, \mathbf{b}$ which is a contrast among the population mean vectors. Thus,

$$\mathbf{a} \propto \mathbf{K}_X^- \mathbf{z}$$

and so the discriminant function $(\mathbf{K}_X^- \mathbf{z})^T \mathbf{X}$ is exactly the same as (apart from a constant of proportionality) $\mathbf{a}^T \mathbf{X}$ obtained earlier by Fisher's approach.

Let $\mathbf{a}_1^T \mathbf{X}$ and $\mathbf{b}_1^T \mathbf{Y}$ be the first canonical variables of the $\mathbf{X}$ and $\mathbf{Y}$ space. Then, $\mathbf{a}_1$ and $\mathbf{b}_1$ solve the problem of finding the linear contrast of transformed means that is largest in magnitude. That is, $\mathbf{a}_1$ and $\mathbf{b}_1$ are maximizing

$$\left| \sum_{j=1}^{J} \pi_j b_j \mathbf{a}^T \boldsymbol{\mu}_j \right|$$

subject to $\mathbf{a}^T \mathbf{K}_X \mathbf{a} = 1$, $\sum_{j=1}^{J} \pi_j b_j = 0$ and $\sum_{j=1}^{J} \pi_j b_j^2 = 1$.

The b_j's measure the importance of the transformed mean $m_j = \mathbf{a}^T \boldsymbol{\mu}_j$ in the contrast. So, if b_j is small, m_j does not contribute much to the contrast and conversely. But the b_j's are all the coefficients for the $Y_j = I(G = j)$. These provide information about how important Y_j is to the random variable $\mathbf{b}^T \mathbf{Y}$. Clearly, if $b_j = 0$ then Y_j does not contribute. Also, when the class probabilities are equal, the b_j are the coefficients of the contrast among the transformed means.

Now $\mathbf{a}_1^T \mathbf{X}$ and $\mathbf{b}_1^T \mathbf{Y}$ are the transformed variables with the most correlation. Thus, we are using $\mathbf{a}_1^T \mathbf{X}$ to predict $\mathbf{b}_1^T \mathbf{Y}$. However, $\mathbf{b}_1^T \mathbf{Y}$ is discrete with

$$\mathbf{b}_1^T \mathbf{Y} = b_{1j}$$

with probability π_j.

Our ability to predict $\mathbf{b}_1^T \mathbf{Y}$ is clearly related to how the b_{1j} fall. If, for example, $b_{1i} \neq b_{1j}$ for all $i, j = 1, \ldots, J$, then there are distinct scores associated with each population and we can expect $\mathbf{a}_1^T \mathbf{X}$ to be able to distinguish between each of the J populations. However,

if $b_{11} = b_{12}$ for examples, the $\mathbf{a}_1^T\mathbf{X}$ will not be able to tell populations 1 and 2 apart. So, $\mathbf{a}_1^T\mathbf{X}$ should only be useful in predicting membership in population whose b_{1j}'s are large and distinct.

In practice, we will need to estimate $\mathbf{a}_1$ and $\mathbf{b}_1$ to obtain $\hat{\mathbf{a}}_1^T\mathbf{X}$ and $\hat{\mathbf{b}}_1^T\mathbf{Y}$. The first thing one should do at that point is to look at $\hat{\mathbf{b}}_1$. The coefficients here will tell the populations for which $\hat{\mathbf{a}}_1^T\mathbf{X}$ will be able to serve as a discriminator. If $\hat{\mathbf{b}}_1$ has some small or almost equal coefficients, then another discriminator is needed. So, we go to $\hat{\mathbf{a}}_2^T\mathbf{X},\ \hat{\mathbf{b}}_2^T\mathbf{Y}$ and hope that $\hat{\mathbf{a}}_2^T\mathbf{X}$ will help with the populations that $\hat{\mathbf{a}}_1^T\mathbf{X}$ could not separate. This process can then be repeated, etc.

3.2.7 Classification

Our goal in this section is to formulate the classification rules based on the canonical variables of the $\mathbf{X}$ and $\mathbf{Y}$ spaces. Prior to achieving this aim, we know the fact that, in CCA, if $\mathbf{X}$ is interpreted as causing $\mathbf{Y}$, then $\mathbf{a}^T\mathbf{X}$ may be called the best predictor of $\mathbf{Y}$ and $\mathbf{b}^T\mathbf{Y}$ the most predictable criterion and vice versa.

Let $\eta = \mathbf{a}^T\mathbf{X}$ and $\xi = \mathbf{b}^T\mathbf{Y} = \mathbf{b}^T\mathbf{Y}(G)$ be a pair of the canonical variables of the $\mathbf{X}$ and $\mathbf{Y}$ spaces corresponding to the canonical correlation ρ. Since we know that η is the best predictor of ξ from CCA, we can predict ξ via η using the regression of ξ on η. Then, the predicted score is given by

$$\begin{aligned}\mathrm{E}[\xi] + \frac{\mathrm{Cov}(\xi,\eta)}{\mathrm{Var}(\eta)}(\eta - \mathrm{E}[\eta]) &= \mathrm{E}[\xi] + \rho(\eta - \mathrm{E}[\eta]) \\ &= \rho(\mathbf{a}^T\mathbf{X} - \mathbf{a}^T\boldsymbol{\mu})\end{aligned}$$

as $\mathrm{E}[\xi] = \mathbf{b}^T\boldsymbol{\pi} = \sum_{j=1}^{J}\pi_j b_j = 0$ and $\mathrm{Var}(\eta) = \mathrm{Var}(\mathbf{a}^T\mathbf{X}) = 1$.

We first can think of using a distance measure to compare the predicted scores of the first s canonical $\mathbf{X}$ variables to the class centroid of those scores. For this purpose, set

$\tilde{\xi}_i = \tilde{\xi}_i(\mathbf{x}) = \rho_i \mathbf{a}_i^T \mathbf{x} - \rho_i \mathbf{a}_i^T \boldsymbol{\mu}$. Then, given $s \leq \min(p, J-1)$, define

$$\sum_{k=1}^{s} \frac{1}{\rho_k^2(1-\rho_k^2)} (\tilde{\xi}_k(\mathbf{x}) - \bar{\tilde{\xi}}_{kj})^2 \tag{3.42}$$

with $\bar{\tilde{\xi}}_{kj} = \mathrm{E}[\theta_k(\mathbf{x})|G = j]$.

We can also can consider a distance measure which compares the predicted scores of the first s canonical $\mathbf{X}$ variables to representative points for J classes rather than the class centroids. Natural points to use for this purpose are provided by the canonical $\mathbf{Y}$ variable since the canonical $\mathbf{Y}$ variable corresponding to the population j has the value b_j. So, we could use the following distance measure for classification:

$$\sum_{k=1}^{s} \frac{1}{1-\rho_k^2} (\tilde{\xi}_k(\mathbf{x}) - b_{kj})^2 - \sum_{k=1}^{s} b_{kj}^2. \tag{3.43}$$

Yet, another option is to consider a distance measure which compares the canonical $\mathbf{X}$ scores to the predicted scores of the canonical $\mathbf{X}$ variable via the canonical $\mathbf{Y}$ variable. We can predict the scores to be assigned to the J classes using the regression of η on ξ. To predict η via ξ we use

$$\mathrm{E}[\eta] + \frac{\mathrm{Cov}(\eta, \xi)}{\mathrm{Var}(\xi)} (\xi - \mathrm{E}[\xi]) = \mathbf{a}^T \boldsymbol{\mu} + \rho \mathbf{b}^T \mathbf{Y}$$

since $\mathrm{E}[\xi] = 0$ and $\mathrm{Var}(\xi) = 1$. Set $\tilde{\eta}_{kj} = \mathbf{a}_k^T \boldsymbol{\mu} + \rho_k b_{kj}$. This leads to a distance such as

$$\sum_{k=1}^{s} \frac{1}{1-\rho_k^2} (\mathbf{a}_k^T \mathbf{x} - \tilde{\eta}_{kj})^2. \tag{3.44}$$

We have now introduced several distances: namely, (3.31), (3.42), (3.43) and (3.44). The relationship between these distances is the subject of the next theorem.

THEOREM **III.3**. *The distances in (3.31), (3.42) and (3.44) are the same.*

Proof. We have seen that the distances in (3.31) and (3.36) are identical from the relationship between the vectors that maximizes (3.26) and (3.32) in Section 2.2.3.1. We can easily see from Theorem 4 that the distance in (3.42) is the same as in (3.36) since $\lambda_k = \rho_k^2$.

Now we start with

$$Dist_j^s(\mathbf{x}) = \sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\mathbf{a}_k^T\mathbf{x} - \mathbf{a}_k^T\boldsymbol{\mu}_j)^2 = \sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\mathbf{a}_k^T\mathbf{x} - \mathbf{a}_k^T\boldsymbol{\mu} + \mathbf{a}_k^T\boldsymbol{\mu} - \mathbf{a}_k^T\boldsymbol{\mu}_j)^2.$$

Since

$$\mathbf{K}_{XY} = [\pi_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \pi_J(\boldsymbol{\mu}_J - \boldsymbol{\mu})] = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_J]\mathbf{K}_Y,$$

premultiplication by $\mathbf{K}_Y$ in (3.10) produces

$$\mathbf{K}_{YX}\mathbf{a}_k = \rho_k\mathbf{K}_Y\mathbf{b}_k.$$

This is equivalent to

$$\text{diag}(\pi_1, \ldots, \pi_J)[\boldsymbol{\mu}_1 - \boldsymbol{\mu}, \ldots, \boldsymbol{\mu}_J - \boldsymbol{\mu}]^T\mathbf{a}_k = \rho_k\text{diag}(\pi_1, \ldots, \pi_J)\mathbf{b}_k$$

since $\mathbf{K}_Y\mathbf{b}_k = \text{diag}(\pi_1, \ldots, \pi_J)\mathbf{b}_k$ which follows from $\boldsymbol{\pi}^T\mathbf{b}_k = 0$. So, we have

$$[\boldsymbol{\mu}_1 - \boldsymbol{\mu}, \ldots, \boldsymbol{\mu}_J - \boldsymbol{\mu}]^T\mathbf{a}_k = \rho_k\mathbf{b}_k$$

and hence $\mathbf{a}_k^T(\boldsymbol{\mu}_j - \boldsymbol{\mu}) = \rho_k b_{kj}$ for $j = 1, \ldots, J$. Thus, the distance in (3.36) becomes (3.44).

□

The distances in (3.31), (3.42) and (3.43) are known to be equivalent in case that $\mathbf{K}_X$ and $\mathbf{K}_Y$ are invertible (Hastie et al., 1995). We now look for the relationship between the distances in (3.42) and (3.43) in the case of singularity for $\mathbf{K}_X$. Note that $\mathbf{K}_Y$ is always singular in our setting.

COROLLARY **III.1**. *The distance in (3.42) is equivalent to the distance measure in (3.43) in the sense of classification.*

Proof. As in the proof of Theorem III.3, we begin with

$$\begin{aligned} Dist_j^s(\mathbf{x}) &= \sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\mathbf{a}_k^T\mathbf{x} - \mathbf{a}_k^T\boldsymbol{\mu}_j)^2 \\ &= \sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\mathbf{a}_k^T\mathbf{x} - \mathbf{a}_k^T\boldsymbol{\mu})^2 - 2\sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\mathbf{a}_k^T\mathbf{x} - \mathbf{a}_k^T\boldsymbol{\mu})(\mathbf{a}_k^T\boldsymbol{\mu}_j - \mathbf{a}_k^T\boldsymbol{\mu}) \\ &\quad + \sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\mathbf{a}_k^T\boldsymbol{\mu}_j - \mathbf{a}_k^T\boldsymbol{\mu})^2. \end{aligned}$$

We have seen that $\mathbf{a}_k^T(\boldsymbol{\mu}_j - \boldsymbol{\mu}) = \rho_k b_{kj}$ for $j = 1, \ldots, J$. We then observe that

$$\mathbf{a}_k^T\mathbf{x} - \mathbf{a}_k^T\boldsymbol{\mu} = \rho_k^{-1}\tilde{\xi}_k(\mathbf{x}).$$

Thus, the distance above becomes

$$\begin{aligned} &\sum_{k=1}^{s} \frac{1}{\rho_k^2(1-\rho_k^2)}\tilde{\xi}_k(\mathbf{x})^2 - 2\sum_{k=1}^{s} \frac{1}{1-\rho_k^2}\tilde{\xi}_k(\mathbf{x})b_{kj} + \sum_{k=1}^{s} \frac{\rho_k^2}{1-\rho_k^2}b_{kj}^2 \\ &= \sum_{k=1}^{s} \left(\frac{1}{\rho_k^2(1-\rho_k^2)} - \frac{1}{1-\rho_k^2}\right)\tilde{\xi}_k(\mathbf{x})^2 + \sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\tilde{\xi}_k(\mathbf{x}) - b_{kj})^2 \\ &\quad + \sum_{k=1}^{s} \left(\frac{\rho_k^2}{1-\rho_k^2} - \frac{1}{1-\rho_k^2}\right)b_{kj}^2 \\ &= \sum_{k=1}^{s} \rho_k^{-2}\tilde{\xi}_k(\mathbf{x})^2 + \sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\tilde{\xi}_k(\mathbf{x}) - b_{kj})^2 - \sum_{k=1}^{s} b_{kj}^2. \end{aligned}$$

Since the term $\sum_{k=1}^{s} \rho_k^{-2}\tilde{\xi}_k(\mathbf{x})^2$ does not depend on the class membership, the class that minimizes (3.42) is identical to the class that minimizes (3.43).

□

Suppose that $s = r = \min(p, J-1) = J-1$. Then, the distance measure in (3.43) is equivalent to the distance measure

$$\sum_{k=1}^{J-1} \frac{1}{1-\rho_k^2}(\tilde{\xi}_k(\mathbf{x}) - b_{kj})^2 - \pi_j^{-1}. \tag{3.45}$$

This distance measure cannot be used in a dimension-reduction mode since it counts on the presence of $J-1$ discriminant coordinates. To establish equivalence, let $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_{J-1}]$

and $\widetilde{\mathbf{B}} = [\mathbf{B}, \mathbf{1}]$. Then $\widetilde{\mathbf{B}}$ is square and nonsingular with

$$\widetilde{\mathbf{B}}^T \text{diag}(\pi_1, \ldots, \pi_J)\widetilde{\mathbf{B}} = \begin{pmatrix} \mathbf{B}^T \text{diag}(\pi_1, \ldots, \pi_J)\mathbf{B} & \mathbf{B}^T \text{diag}(\pi_1, \ldots, \pi_J)\mathbf{1} \\ \mathbf{1}^T \text{diag}(\pi_1, \ldots, \pi_J)\mathbf{B} & \mathbf{1}^T \text{diag}(\pi_1, \ldots, \pi_J)\mathbf{1} \end{pmatrix} = \mathbf{I}_J$$

since $\mathbf{B}^T \text{diag}(\pi_1, \ldots, \pi_J)\mathbf{B} = \mathbf{B}^T\mathbf{K}_Y\mathbf{B} = \mathbf{I}_{J-1}$, $\mathbf{B}^T \text{diag}(\pi_1, \ldots, \pi_J)\mathbf{1} = \mathbf{B}^T\boldsymbol{\pi} = \mathbf{0}$ and $\mathbf{1}^T \text{diag}(\pi_1, \ldots, \pi_J)\mathbf{1} = \mathbf{1}^T\boldsymbol{\pi} = 1$. which follow from $\mathbf{b}_i^T\mathbf{K}_Y\mathbf{b}_i = 1, \mathbf{b}_i^T\mathbf{K}_Y\mathbf{b}_k = 0$ and $\boldsymbol{\pi}^T\mathbf{b}_i = 0$ for $i, k = 1, \ldots, J-1, i \neq k$. Since $\widetilde{\mathbf{B}}$ is nonsingular, we have $\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T = \text{diag}(\pi_1^{-1}, \ldots, \pi_J^{-1})$, or $\sum_{k=1}^{J-1} b_{kj}^2 + 1 = \pi_j^{-1}$ for $j = 1, \ldots, J$.

3.2.7.1 Example: Fisher's Irises Data

In this section we exemplify some of the previous discussions using Fisher's classic Iris data set. The iris data published by Fisher (1936) have been widely used for examples in discriminant analysis and cluster analysis. For this data, four measurements (sepal length and width, and petal length and width) were taken on each of fifty specimens of three different Iris types: namely, setosa, versicolor, and verginica.

The estimated canonical correlations are

$$\hat{\rho}_1 = 0.985, \quad \hat{\rho}_2 = 0.471.$$

The corresponding canonical variables of the $\mathbf{X}$ space are

$$-.145X_1 - .269X_2 + .386X_3 + .493X_4$$

and

$$-.021X_1 - 1.928X_2 + .83X_3 - 2.529X_3$$

for X_1 the sepal length, X_2 the sepal width, X_3 the petal length and X_4 the petal width. The estimated coefficient vectors of the first two canonical variables of the $\mathbf{Y}$ space are given by

$$\hat{\mathbf{b}}_1 = (-1.354, 0.324, 1.029) \quad \text{and} \quad \hat{\mathbf{b}}_2 = (-0.407, 1.376, -0.969).$$

(a) (b)

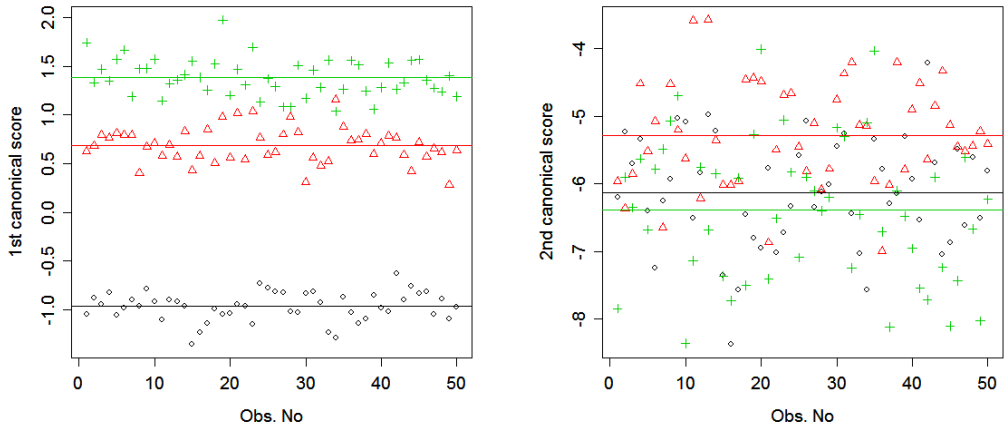

Figure 1: Plots of (a) the first canonical $\mathbf{X}$ scores (b) the second canonical $\mathbf{X}$ scores for 150 irises and the predicted canonical scores (horizontal lines) superimposed: black points for Sertosa, red points for Versicolor and green points for Verginica. Each point represents a score for an iris.

Table 1: Confusion matrix of classification of the Iris data

	Sertosa	Versicolor	Verginica	
Sertosa	50	0	0	50
Versicolor	0	48	2	50
Verginica	0	0	50	50

In accordance with our discussion of the role of the coefficient vector for the canonical $\mathbf{Y}$ variables in Section 3.2.6, we might expect the first discriminator or the first canonical variable of the $\mathbf{X}$ space to be able to distinguish Sertosa from the other species and the second discriminator to be able to distinguish Versicolor from the others. Figure 1 reveals that this is, indeed, the case. However, we also see that the discrimination power of the second discriminator is quite limited relative to the first. So, we will only use the first discriminators which results in a misclassification rate is 1.33%. Table 1 shows the result of classification using the CCA approach using the classification rule based on the distance in (3.44) with $s = 1$.

CHAPTER IV

MATHEMATICAL PRELIMINARIES

In this chapter, we collect a number of results and definitions that provide the mathematical prerequisite for the developments in subsequent chapters. We begin with a discussion of inner product spaces.

4.1 Hilbert Spaces

The concept of a Hilbert space occupies a fundamental role throughout this dissertation. In this section we lay out some of the basic facts about Hilbert spaces that will be used in the sequel.

Hilbert spaces are normed vector spaces whose norms stem from a bilinear function referred to as an inner product. The concepts of norms and inner products can be developed formally as follows.

DEFINITION **IV.1**. Let V be a vector space over $\mathbb{R}$. A norm on V is a function $\|\cdot\| : V \to \mathbb{R}$ such that for all $u, v \in V$ and $\alpha \in \mathbb{R}$,

(a) $\|u\| > 0$ if and only if $u \neq 0$,

(b) $\|\alpha u\| = |\alpha|\|u\|$,

(c) $\|u + v\| \leq \|u\| + \|v\|$.

A vector space V with a norm is called a normed (vector) space.

DEFINITION **IV.2**. Let V be a vector space over $\mathbb{R}$. An inner product on V is a function $\langle \cdot, \cdot \rangle$ on $V \times V \to \mathbb{R}$ such that for all $u, v, w \in V$ and $\alpha, \beta \in \mathbb{R}$,

(a) $\langle u, u \rangle > 0$ if and only if $u \neq 0$,

(b) $\langle u, v\rangle = \langle v, u\rangle$,

(c) $\langle \alpha u + \beta v, w\rangle = \alpha\langle u, w\rangle + \beta\langle v, w\rangle$.

A real vector space V with an inner product is called an inner product space. The function $\|\cdot\| : V \to \mathbb{R}$ defined by $\|u\| = \langle u, u\rangle^{1/2}, u \in V$ is a norm on V and hence an inner product space is a normed space.

DEFINITION **IV.3**. Let V be an inner product space and let A be a subset of V. The orthogonal complement of A is the set

$$A^{\perp} = \{u \in V : \langle u, a\rangle = 0 \text{ for all } a \in A\}.$$

The triangle inequality (i.e., the relation $\|u-v\| \leq \|u-w\|+\|w-v\|$ for all $u, v, w \in V$) immediately implies that if a sequence $\{u_n\}$ in V converges, then it is necessarily a Cauchy sequence. But the converse of this statement is not true. So, to avoid questions concerning the existence of the limit of a sequence in V, our interest is in a complete space.

DEFINITION **IV.4**. V is complete if for any Cauchy sequence $\{u_n\}$ with $u_n \in V$ there exists $u \in V$ such that $\|u_n - u\| \to 0$ as $n \to \infty$ for all n.

DEFINITION **IV.5**. An inner product space which is complete under the norm induced by the inner product is called a Hilbert space.

Although every inner product space does not have the completeness property, any inner product space can be completed to create a Hilbert space.

A matrix is a linear transformation in a finite-dimensional vector space and it has played an important role in the developments of our theory. Matrices are actually linear transformation which preserves the linear structure of the vector spaces. The matrices treated in the previous chapter were linear transformation between finite-dimensional real vector spaces and, as such, they are automatically bounded and compact. However, when the spaces being transformed are infinite-dimensional, conditions of boundedness

and compactness do not hold automatically. So, we now summarize some important concepts involving the properties of linear transformations or linear operators between linear spaces.

DEFINITION **IV.6**. Let V, W be real vector spaces. A mapping $T : V \to W$ is said to be a linear transformation if for all $\alpha \in \mathbb{R}$ and $u, v \in V$,

(a) $T(u + v) = T(u) + T(v)$,

(b) $T(\alpha u) = \alpha T(u)$.

If $W = \mathbb{R}$ then T is said to be a linear functional.

DEFINITION **IV.7**. If T is a linear transformation from V to W, the range and null space of T are defined by

$$\text{Im}(T) = \{w \in W : w = Tu \text{ for some } u \in V\}$$

and

$$\text{Ker}(T) = \{u \in V : Tu = 0\},$$

respectively. Also, the rank of T denoted by $r(T)$ is the dimension of $\text{Im}(T)$.

DEFINITION **IV.8**. Suppose that V and W are normed spaces with norms $\|\cdot\|_V$ and $\|\cdot\|_W$, respectively. Let T be a linear transformation from V to W.

(a) T is said to be bounded if there exists a finite M such that $\|Tu\|_W \leq M\|u\|_V$ for all $u \in V$.

(b) T is compact if for any bounded sequence $\{u_n\}$ in V the sequence $\{Tu_n\}$ in W contains a convergent subsequence.

(c) T is called an isometry if $\|Tu\|_W = \|u\|_V$ for $u \in V$.

(d) A one-to-one linear transformation T from V onto W is said to be an isomorphism.

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be real Hilbert spaces. The set of all bounded linear transformations from $\mathcal{H}_1$ to $\mathcal{H}_2$ is denoted by $B(\mathcal{H}_1, \mathcal{H}_2)$. Elements of $B(\mathcal{H}_1, \mathcal{H}_2)$ are also called bounded linear operators.

DEFINITION **IV.9**. Let $T \in B(\mathcal{H}_1, \mathcal{H}_2)$. A transformation $T^* \in B(\mathcal{H}_2, \mathcal{H}_1)$ such that $\langle Tu, v\rangle_{\mathcal{H}_2} = \langle u, T^*v\rangle_{\mathcal{H}_1}$ for $u \in \mathcal{H}_1, v \in \mathcal{H}_2$ is said to be the adjoint of the operator T. Now let $\mathcal{H}$ be a real Hilbert space and $T \in B(\mathcal{H}, \mathcal{H}) := B(\mathcal{H})$.

(a) T is said to be self-adjoint if $T^* = T$.

(b) T is positive if it is self-adjoint and $\langle Tu, u\rangle_{\mathcal{H}} \geq 0$ for $u \in \mathcal{H}$.

(c) T is said to be a projection if $T^2 = T$.

(d) T is normal if $TT^* = T^*T$.

A Banach space is a complete normed vector space.

THEOREM **IV.1**. (Open mapping theorem) *Let V and W be Banach spaces and $T \in B(V, W)$ map V onto W. If T is one-to-one then there exist $S \in B(W, V)$ such that $S \circ T = I_V$ and $T \circ S = I_W$.*

To provide solutions to the optimization problems posed in Chapters V and VI, we will need the concepts of eigenvalue and eigenvector of the linear operator that arises from the spectral decomposition of a bounded linear self-adjoint operator and also the concept of polar representation of a bounded linear operator. We collect some essential information about these notions in the remainder of this section.

DEFINITION **IV.10**. The spectrum $\sigma(T)$ of an operator $T \in B(\mathcal{H})$ is the set of all scalars λ for which $T - \lambda I$ is not invertible.

DEFINITION **IV.11**. Let V be a vector space and T be a linear transformation from V to V. A scalar λ is an eigenvalue of T if $Tv = \lambda v$ has a non-zero solution $v \in V$, and any

such non-zero solution is an eigenvector. The subspace Ker$(T - \lambda I)$ in V is called the eigenspace corresponding to λ and the multiplicity of λ is the dimension of Ker$(T - \lambda I)$.

Now for any eigenvalue of T we may find elements $v_1 \neq v_2$ such that $(T - \lambda I)v_1 = (T - \lambda I)v_2$. Thus, $(T - \lambda I)(v_1 - v_2) = 0$ which $T - \lambda I$ is not one-to-one. But, if $T - \lambda I$ is not one-to-one it is not invertible and consequently, any eigenvalue of T must be in $\sigma(T)$.

The spectrum $\sigma(T)$ of $T \in B(\mathcal{H})$ can be divided into three disjoint subsets. The subset of $\sigma(T)$ consisting of all eigenvalues of T is called the point spectrum of T. The set of λ's for which $T - \lambda I$ is a one-to-one mapping of $\mathcal{H}$ onto a dense proper subspace of $\mathcal{H}$ is called the continuous spectrum for T. Finally, the set consisting of all other $\lambda \in \sigma(T)$ is called the residual spectrum for T.

THEOREM **IV.2**. *Let $\mathcal{H}$ be a real Hilbert space and $T \in B(\mathcal{H})$.*

(a) *$\sigma(T)$ is a closed set.*

(b) *A normal operator has empty residual spectrum.*

Since a self-adjoint operator is normal, we observe from (b) that the spectrum $\sigma(T)$ of a self-adjoint operator can be decomposed into the point spectrum and the continuous spectrum.

DEFINITION **IV.12**. Let $\mathcal{A}$ be a σ-field in a set Ω and let $\mathcal{H}$ be a real Hilbert space. In this setting, a resolution of the identity on $\mathcal{A}$ is a mapping

$$E : \mathcal{A} \to B(\mathcal{H})$$

with the following properties:

(a) $E(\emptyset) = 0, E(\Omega) = 1$.

(b) Each $E(\omega)$ is a self-adjoint projection for $\omega \in \mathcal{A}$.

(c) $E(\omega_1 \cap \omega_2) = E(\omega_1)E(\omega_2)$.

(d) $E(\omega_1 \cup \omega_2) = E(\omega_1) + E(\omega_2)$ for $\omega_1 \cap \omega_2 = \emptyset$.

(e) For every $x \in \mathcal{H}$ and $y \in \mathcal{H}$, the set function $E_{x,y}$ defined by

$$E_{x,y}(\omega) = \langle E(\omega)x, y\rangle_{\mathcal{H}}$$

is a measure on $\mathcal{A}$.

THEOREM **IV.3**. *Let T be a normal operator on a real Hilbert space $\mathcal{H}$. Then there exists a unique resolution of the identity E on the Borel subsets of $\sigma(T)$ which satisfies*

$$T = \int_{\sigma(T)} \gamma dE(\gamma). \tag{4.1}$$

Since if $T \in B(\mathcal{H})$ is self-adjoint then it is normal, Theorem IV.3 is true for self-adjoint operator.

If $\mathcal{H}$ is a finite-dimensional Hilbert space and $T \in B(\mathcal{H})$ then the spectrum of T consists solely of eigenvalues of T. However, there are operators on infinite-dimensional spaces which have no eigenvalues at all. If $T \in B(\mathcal{H})$ is compact then the zero eigenvalue belongs to the spectrum $\sigma(T)$ and the set of non-zero eigenvalues of T consists of countable set of eigenvalues with finite multiplicity.

Suppose that T is compact. Let us define the operator $f \otimes g$ from $\mathcal{H}_1$ to $\mathcal{H}_2$ as

$$(f \otimes g)h = \langle g, h\rangle_{\mathcal{H}_1} f$$

for $f \in \mathcal{H}_2,\ g, h \in \mathcal{H}_1$. Then, $r(T)$ represents the cardinality of $\sigma(T)$ and (4.1) becomes

$$T = \sum_{j=1}^{r(T)} \gamma_j e_j \otimes e_j, \tag{4.2}$$

where $\gamma_1, \gamma_2, \ldots, \gamma_{r(T)}$ are non-zero distinct eigenvalues of T with associated eigenvectors e_j.

The polar representation for a bounded but non-self adjoint linear operator can be combined with our discussion of the spectral decomposition of a positive and self-adjoint linear operator (Naimark, 1960 and Rudin, 1973) to obtain a decomposition for operators between two Hilbert spaces. The specific result is that if $T \in B(\mathcal{H}_1, \mathcal{H}_2)$ then

$$T = W(T^*T)^{1/2} = W \int_{\sigma(T^*T)} \lambda^{1/2} dE(\lambda), \tag{4.3}$$

where W is a unique partial isometry (i.e., a norm preserving mapping from Ker$(T)^{\perp}$ to $\overline{\text{Im}(T)}$), $\sigma(T^*T) = \{\lambda \in \mathbb{R} : T^*T - \lambda I \text{ is not invertible}\}$ is a closed subset of $[0, \infty)$ and $\{E(\lambda) : \lambda \in \sigma(T^*T)\}$ is the unique resolution of the identity corresponding to T^*T.

Thus, if $T^*T \in B(\mathcal{H}_1)$ is compact we have

$$(T^*T)^{1/2} = \sum_{j=1}^{r(T)} \lambda_j^{1/2} \beta_j \otimes \beta_j,$$

since $r(T) = r(T^*T)$, and (4.3) becomes

$$T = \sum_{j=1}^{r(T)} \lambda_j^{1/2} \alpha_j \otimes \beta_j, \tag{4.4}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{r(T)} > 0$ are the eigenvalues of T^*T with associated eigenvectors $\beta_j, j = 1, \ldots, r(T)$ and

$$\alpha_j = W\beta_j = \lambda_j^{-1/2} W(T^*T)^{1/2} \beta_j = \lambda_j^{-1/2} T\beta_j$$

which follows from $(T^*T)^{1/2}\beta_j = \lambda_j^{1/2}\beta_j$.

4.2 Reproducing Kernel Hilbert Spaces and Stochastic Processes

Reproducing kernel Hilbert spaces (RKHS's) provide a fundamental tool for inference concerning second order stochastic process. This stems from the congruence between the Hilbert space spanned by a stochastic process and the RKHS generated by its covariance

kernel. The link between reproducing kernels and stochastic processes was initially established by Loève (1948) and was developed fully by Parzen in a series of articles (e.g., see Parzen, 1967).

Now we will review some basic facts about RKHS's. More details can be found in Aronszajn (1950), Parzen (1961) and Weinert (1982). We begin with the definition of positive definite functions.

DEFINITION **IV.13**. A symmetric, real-valued bivariate function K on $\mathcal{T} \times \mathcal{T}$ is said to be positive definite if, for any real $a_1, \ldots, a_n$, and $t_1, \ldots, t_n \in \mathcal{T}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K(t_i, t_j) \geq 0,$$

and strictly positive definite if "$>$" holds.

DEFINITION **IV.14**. Let $\mathcal{H}$ be a Hilbert space of functions on some set $\mathcal{T}$ and denote by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ the inner product in $\mathcal{H}$. A bivariate function on $\mathcal{T} \times \mathcal{T}$ is said to be a reproducing kernel (r.k.) for $\mathcal{H}$ if for every $t \in \mathcal{T}$ and $f \in \mathcal{H}$,

(a) $K(\cdot, t) \in \mathcal{H}$,

(b) $f(t) = \langle f, K(\cdot, t) \rangle_{\mathcal{H}}$.

When (a) and (b) hold, $\mathcal{H}$ is said to be a reproducing kernel Hilbert space with r.k. K.

The property (b) is termed the reproducing property of K. It can easily be shown that K is the unique r.k. and K is a symmetric and positive definite function. The reproducing property leads us to the following theorem.

THEOREM **IV.4**. (Moore-Aronszajn-Loève) *Given a positive definite functions K on $\mathcal{T} \times \mathcal{T}$, one can construct a unique RKHS $\mathcal{H}(K)$ of real-valued functions on E with K as its r.k.. The space $\mathcal{H}(K)$ is given by the closure of the linear span of $\{K(\cdot, t), t \in \mathcal{T}\}$, i.e.,*

$$\mathcal{H}(K) = \overline{\text{span}}\{K(\cdot, t), t \in \mathcal{T}\}. \tag{4.5}$$

Let H be the linear manifold spanned by $\{K(\cdot,t), t \in \mathcal{T}\}$: i.e., the set of all finite linear combinations of the form

$$\sum_{i=1}^{n} a_i K(\cdot, t_i)$$

for $a_1, \ldots, a_n \in \mathbb{R}$, $t_1, \ldots, t_n \in \mathcal{T}$ and $n = 1, 2, \ldots$ with the inner product

$$\langle \sum_{i=1}^{n} a_i K(\cdot, t_i), \sum_{j=1}^{m} b_j K(\cdot, s_j) \rangle_H = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j K(t_i, s_j)$$

for arbitrary points $t_1, \ldots, t_n, s_1, \ldots, s_m$ in $\mathcal{T}$. Then, H is an incomplete inner product space with r.k. K. But it can be completed by adjoining all limits of Cauchy sequences of the functions in $\mathcal{H}$. Let $\mathcal{H}(K)$ be the completion of H and define a norm on $\mathcal{H}(K)$ by

$$\|f\|^2_{\mathcal{H}(K)} = \lim_{n \to \infty} \|f_n\|^2_H$$

with a Cauchy sequence $\{f_n\}$ in H converging pointwise to f. Then, H is dense in $\mathcal{H}(K)$.

We now review Parzen's representation theory concerning various concrete function spaces that are congruent to the Hilbert space spanned by a second order stochastic process (Parzen, 1961). A fundamental tool in this development is the following result.

THEOREM **IV.5**. (Basic Congruence Theorem) *Let H_1 and H_2 be two abstract Hilbert spaces equipped with the inner products $\langle \cdot, \cdot \rangle_{H_1}$ and $\langle \cdot, \cdot \rangle_{H_2}$. Let $\{u(t), t \in \mathcal{T}\}$ be a family of vectors which spans H_1 and $\{v(t), t \in \mathcal{T}\}$ be a family of vectors which spans H_2. If for every s and t in $\mathcal{T}$*

$$\langle u(s), u(t) \rangle_{H_1} = \langle v(s), v(t) \rangle_{H_2}$$

then the spaces H_1 and H_2 are congruent and there exists an isometric isomorphism (one-to-one and onto inner product preserving linear mapping) ψ from H_1 to H_2 satisfying

$$\psi(u(t)) = v(t), \ t \in \mathcal{T}.$$

Let $(Q, \mathcal{B}, \nu)$ be a measure space and let $L^2(\nu)$ be the Hilbert space of all $\mathcal{B}$-measurable real valued functions defined on Q that are square integrable with respect to ν with inner product

$$\langle f_1, f_2 \rangle_{L^2(\nu)} = \int_Q f_1(q) f_2(q) d\nu(q)$$

for $f_1, f_2 \in L^2(\nu)$. The next theorem provides an explicit formula that can frequently be used to obtain the inner product for a RKHS $\mathcal{H}(K)$ generated by K.

THEOREM **IV.6**. *Suppose that there is a set of functions $\{\phi(t,\cdot), t \in \mathcal{T}\}$ in $L^2(\nu)$ such that*

$$K(s,t) = \langle \phi(s,\cdot), \phi(t,\cdot) \rangle_{L^2(\nu)} \tag{4.6}$$

for all $s, t \in \mathcal{T}$. Then the RKHS $\mathcal{H}(K)$ corresponding to K consists of all functions of the form

$$f(t) = \langle g(\cdot), \phi(t,\cdot) \rangle_{L^2(\nu)} \tag{4.7}$$

for some unique function g in $\overline{\text{span}}\{\phi(t,\cdot), t \in \mathcal{T}\} \cap L^2(\nu)$, with inner product given by

$$\langle f_1, f_2 \rangle_{\mathcal{H}(K)} = \langle g_1, g_2 \rangle_{L^2(\nu)} \tag{4.8}$$

for $f_1, f_2 \in \mathcal{H}(K)$ corresponding to $g_1, g_2 \in \overline{\text{span}}\{\phi(t,\cdot), t \in \mathcal{T}\}$.

We finish out this section with discussion of i) the basic congruence relation between the Hilbert space of random variables spanned by a second-order stochastic process and the RKHS determined by its second moment function, ii) the one-to-one correspondence between the Hilbert space of random variables spanned by a second-order stochastic process and the RKHS determined by its covariance function and iii) some examples of RKHS's.

Let $\{X(t), t \in \mathcal{T}\}$ be a second order stochastic process with the mean function

$$\mu(t) = \mathrm{E}[X(t)]$$

and covariance function

$$K(s,t) = \mathrm{Cov}(X(s), X(t))$$

for $s, t \in \mathcal{T}$. We denote by R the second moment function

$$R(s,t) = \mathrm{E}[X(s)X(t)].$$

Note that

$$R(s,t) = K(s,t) + \mu(s)\mu(t).$$

Now let $(\Omega, \mathcal{B}, P)$ be the probability space corresponding to the stochastic process $X(\cdot)$ (e.g., see Doob, 1953). If $L^2(P)$ denotes the set of all square integrable functions on $(\Omega, \mathcal{B}, P)$, we are interested in the subset of $L^2(P)$ obtained as the completion (in $L^2(P)$) of the set of all random variables of the form

$$\sum_{i=1}^{n} a_i X(t_i)$$

for some integer n, some constants $a_1, \ldots, a_n \in \mathbb{R}$, and some points $t_1, \ldots, t_n \in \mathcal{T}$. We denote this space by L_X^2 and observe that it is a Hilbert space with inner product

$$\langle U, V \rangle_{L_X^2} = \mathrm{E}[UV] \text{ for } U, V \in L_X^2.$$

Since R is symmetric and positive definite, it generates a RKHS $\mathcal{H}(R)$ as in (4.5) from Theorem IV.4. Then, by the reproducing property,

$$\langle R(\cdot, s), R(\cdot, t) \rangle_{\mathcal{H}(R)} = R(s,t) = \mathrm{E}[X(s)X(t)].$$

Hence, by Theorem IV.5, there is an isometry ψ from $\mathcal{H}(R)$ onto L_X^2 satisfying

$$\psi(R(\cdot, t)) = X(t)$$

and $\mathcal{H}(R)$ and L_X^2 are congruent. So, every random variable U in L_X^2 can be written

$$U = \psi(f)$$

for some unique f in $\mathcal{H}(R)$. Also, ψ satisfies

$$\mathrm{E}[\psi(f)\psi(g)] = \langle f, g \rangle_{\mathcal{H}(R)}$$

for any $f, g \in \mathcal{H}(R)$. Additional properties are

$$\mathrm{E}[\psi(f)] = \langle f, \mu \rangle_{\mathcal{H}(R)}$$

and

$$\mathrm{E}[\psi(f)X(t)] = f(t)$$

for any $f \in \mathcal{H}(R)$.

A case where a complete characterization of ψ is possible corresponds to processes of the form

$$X(t) = \int_Q \phi(t, q) dZ(q), \ t \in \mathcal{T}, \tag{4.9}$$

where $\{Z(B), B \in \mathcal{B}\}$ is a family of random variables on Q with uncorrelated increments and $\phi(t, \cdot) \in L^2(\nu)$ for $d\nu(q) = \mathrm{E}|dZ(q)|^2$. In this instance (4.6) and (4.7) hold and we have

$$\psi(f) = \int_Q g(q) dZ(q). \tag{4.10}$$

The covariance function K of the X process also generates a RKHS $\mathcal{H}(K)$ as in (4.5) since K is symmetric and positive. We may want to use $\mathcal{H}(K)$ to build a representation for a random function X. The Hilbert space L^2_X may not be the same for all values of μ since its inner product depends on μ. However with the additional assumption that μ belongs to a subset M of $\mathcal{H}(K)$, then according to Parzen (1961), the Hilbert space L^2_X is the same for all μ and the set of elements in $\mathcal{H}(K)$ is equal to the set of elements in $\mathcal{H}(R)$ although the two spaces are equipped with different norms.

PROPOSITION **IV.1**. *Assume that $\mu \in M$ with M a subset of $\mathcal{H}(K)$. Then there exists an isomorphism Ψ from $\mathcal{H}(K)$ to L^2_X defined by*

$$\Psi(K(\cdot, t)) = X(t)$$

for every t in $\mathcal{T}$ with the properties

$$\mathrm{E}[\Psi(f)] = \langle f, \mu \rangle_{\mathcal{H}(K)} \tag{4.11}$$

and

$$\text{Cov}(\Psi(f), \Psi(g)) = \langle f, g \rangle_{\mathcal{H}(K)} \tag{4.12}$$

for f, g *in* $\mathcal{H}(K)$.

Proof. Recall our definition for the linear span of $\{K(\cdot, t), t \in \mathcal{T}\}$ which we denoted by H. Then, any function f in H is of the form

$$f(\cdot) = \sum_{i=1}^{n} a_i K(\cdot, t_i)$$

for some integer n, real constant $a_1, \ldots, a_n$ and points $t_1, \ldots, t_n$ in $\mathcal{T}$. For a function f in H, define

$$\Psi(f) = \sum_{i=1}^{n} a_i X(t_i).$$

Then, we observe that for any functions f, g in H,

$$\text{E}[\Psi(f)] = \sum_{i=1}^{n} a_i \mu(t_i) = \langle \mu(\cdot), \sum_{i=1}^{n} a_i K(\cdot, t_i) \rangle_{\mathcal{H}(K)} = \langle \mu, f \rangle_{\mathcal{H}(K)},$$

and

$$\begin{aligned} \text{Cov}(\Psi(f), \Psi(g)) &= \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j K(t_i, s_j) = \langle \sum_{i=1}^{n} a_i K(\cdot, t_i), \sum_{j=1}^{m} b_j K(\cdot, s_j) \rangle_{\mathcal{H}(K)} \\ &= \langle f, g \rangle_{\mathcal{H}(K)} \end{aligned}$$

by the reproducing property of K.

To prove that the mapping Ψ is well defined, it suffices to show that

$$\Psi(f) = \sum_{i=1}^{n} a_i X(t_i) = 0 \quad \text{if and only if} \quad f(\cdot) = \sum_{i=1}^{n} a_i K(\cdot, t_i) = 0$$

which follows from the fact that

$$\text{E}|\Psi(f)|^2 = \text{Var}(\Psi(f)) + \{\text{E}[\Psi(f)]\}^2 = \|f\|^2_{\mathcal{H}(K)} + |\ \langle \mu, f \rangle_{\mathcal{H}(K)}|^2.$$

So, we see that Ψ is a one-to-one linear mapping from H onto the linear manifold spanned by the random function $\{X(t), t \in \mathcal{T}\}$.

From the above relations observe that for any sequence $\{f_n\}$ in H

$$\mathrm{E}|\Psi(f_n) - \Psi(f_m)|^2 = \|f_n - f_m\|^2_{\mathcal{H}(K)} + |\ \langle \mu, f_n - f_m\rangle_{\mathcal{H}(K)}|^2$$

for some n, m. Consequently, for any sequence $\{f_n\}$ in H, that $\{f_n\}$ will be a Cauchy sequence in $H(K)$ if and only if $\{\Psi(f_n)\}$ is a Cauchy sequence in L^2_X. Now any function f in $\mathcal{H}(K)$ may be represented as the limit of a sequence $\{f_n\}$ in H. For a converging sequence $\{f_n\}$ in H, the corresponding random variables $\{\Psi(f_n)\}$ are a Cauchy sequence and have a limit denoted by $\Psi(f)$. Thus, the linear transformation Ψ from $\mathcal{H}(K)$ to L^2_X is one-to-one, onto and satisfies (4.11) and (4.12).

□

The following property is a consequence of (4.12). By replacing g in (4.12) by $K(\cdot, t)$, we have

$$\mathrm{Cov}(\Psi(f), X(t)) = f(t). \tag{4.13}$$

Define the process

$$\widetilde{X}(t) = X(t) - \mu(t),\ t \in \mathcal{T},$$

which is a stochastic process with zero mean and covariance function K. Since $\mathrm{E}[\widetilde{X}(s)\widetilde{X}(t)] = K(s,t) = \langle K(\cdot, s), K(\cdot, t)\rangle_{\mathcal{H}(K)}$ for every $s, t \in \mathcal{T}$, there is an isometric isomorphism $\psi_{\widetilde{X}}$ between the Hilbert space spanned by the $\widetilde{X}$ process, $L^2_{\widetilde{X}}$, and $\mathcal{H}(K)$ satisfying

$$\psi_{\widetilde{X}}(K(\cdot, t)) = \widetilde{X}(t).$$

The isomorphism Ψ and the isometric isomorphism $\psi_{\widetilde{X}}$ are related in the following way

$$\begin{aligned}\Psi(K(\cdot, t)) &= X(t)\\ &= \widetilde{X}(t) + \mu(t)\\ &= \psi_{\widetilde{X}}(K(\cdot, t)) + \langle \mu, K(\cdot, t)\rangle_{\mathcal{H}(K)}.\end{aligned}$$

So, we have

$$\Psi(f) = \psi_{\widetilde{X}}(f) + \langle \mu, f\rangle_{\mathcal{H}(K)} \tag{4.14}$$

for $f \in \mathcal{H}(K)$ (which also belongs to $\mathcal{H}(R)$). Consequently, we see that every random variable in L^2_X has the form

$$U = \langle \mu, f \rangle_{\mathcal{H}(K)} + \psi_{\tilde{X}}(f)$$

for $f = \psi^{-1}_{\tilde{X}}(V)$ with $V \in L^2_{\tilde{X}}$.

EXAMPLE 1. Let $\{X(t), t \in \mathcal{T}\}$ be a second-order stochastic process with covariance function K. Let the index set $\mathcal{T}$ be finite dimensional, say $\mathcal{T} = \{t_1, \ldots, t_p\}$. Then $\mathbf{X} = (X(t_1), \ldots, X(t_p))^T$ with $\boldsymbol{\mu} = \mathrm{E}[\mathbf{X}]$ and

$$\mathrm{Var}(\mathbf{X}) = \{K(t_i, t_j)\}_{i,j=1}^p = \mathbf{K}.$$

Let $\mathcal{H}(\mathbf{K})$ be the linear manifold of all vectors of the form

$$\mathbf{f} = \mathbf{K}\mathbf{a} \text{ for } \mathbf{a} \in \mathrm{Ker}(\mathbf{K})^\perp$$

with inner product

$$\langle \mathbf{f}_1, \mathbf{f}_2 \rangle_{\mathcal{H}(\mathbf{K})} = \mathbf{f}_1^T \mathbf{K}^- \mathbf{f}_2, \tag{4.15}$$

for $\mathbf{f}_k = (f_k(t_1), \ldots, f_k(t_p))^T, k = 1, 2$, with $f_k(\cdot) = \sum_{i=1}^p a_{ki} K(\cdot, t_i)$. Note that the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}(\mathbf{K})}$ is well-defined.

First observe that if $\mathbf{f} = \mathbf{K}\mathbf{a}$ then $\mathbf{a} = \mathbf{K}^- \mathbf{f}$ since $\mathbf{a} \in \mathrm{Ker}(\mathbf{K})^\perp$ and hence $\mathbf{f}_1^T \mathbf{K}^- \mathbf{f}_2 = \mathbf{a}_1^T \mathbf{K} \mathbf{a}_2 = \sum_{i=1}^p \sum_{j=1}^p a_{1i} a_{2j} K(t_i, t_j)$. So it is obvious that $\langle \mathbf{f}, \mathbf{f} \rangle_{\mathcal{H}(\mathbf{K})} = \mathbf{f}^T \mathbf{K}^- \mathbf{f} = \mathbf{a}^T \mathbf{K} \mathbf{a} \geq 0$ for $\mathbf{f} \in \mathcal{H}(\mathbf{K})$ since K is positive definite. Also, we can easily show symmetry and linearity. So we now focus on the property that if $\langle \mathbf{f}, \mathbf{f} \rangle_{\mathcal{H}(K_X)} = 0$ then $\mathbf{f} = \mathbf{0}$.

First observe that $\mathbf{K}(\cdot, t_i) = (K(t_1, t_i), \ldots, K(t_p, t_i))^T = \mathbf{K}\mathbf{e}_i \in \mathcal{H}(\mathbf{K})$ for any $t_i \in \mathcal{T}$ with $\mathbf{e}_i$ an elementary vector of all zeros except for 1 in its ith component. Then,

$$\langle \mathbf{f}, \mathbf{K}(\cdot, t_j) \rangle_{\mathcal{H}(\mathbf{K})} = \mathbf{f}^T \mathbf{K}^- \mathbf{K}(\cdot, t_j) = \mathbf{a}^T \mathbf{K} \mathbf{K}^- \mathbf{K} \mathbf{e}_j = \mathbf{a}^T \mathbf{K}(\cdot, t_j) = f(t_j) \tag{4.16}$$

for any $f \in \mathcal{H}(\mathbf{K})$ and $t_j \in \mathcal{T}$. Now, observe from the Cauchy-Schwarz inequality that

$$f^2(t_j) = \langle \mathbf{f}, \mathbf{K}(\cdot, t_j) \rangle^2_{\mathcal{H}(\mathbf{K})} \leq \langle \mathbf{f}, \mathbf{f} \rangle_{\mathcal{H}(\mathbf{K})} K(t_j, t_j)$$

which implies that if $\langle \mathbf{f}, \mathbf{f} \rangle_{\mathcal{H}(\mathbf{K})} = 0$ then $f(t_j) = 0$ for all j or $\mathbf{f} = \mathbf{0}$. Result (4.16) has the consequence that $\mathcal{H}(\mathbf{K})$ is an inner product space with r.k. K. Since $\mathcal{H}(\mathbf{K})$ is finite dimensional it is also a Hilbert space. Thus, $\mathcal{H}(\mathbf{K})$ is an RKHS with r.k. K.

Let L^2_X be the set of all random variables of the form

$$\sum_{i=1}^{p} a_i X(t_i)$$

for $\mathbf{a} = (a_1, \ldots, a_p)^T \in \text{Ker}(\mathbf{K})^\perp$ with the inner product

$$\text{E}\left[\left(\mathbf{a}_1^T \mathbf{X}\right)\left(\mathbf{a}_2^T \mathbf{X}\right)\right] = \mathbf{a}_1^T \mathbf{K} \mathbf{a}_2 + \left(\mathbf{a}_1^T \boldsymbol{\mu}\right)\left(\mathbf{a}_2^T \boldsymbol{\mu}\right).$$

Then,

$$\Psi(\mathbf{f}) = \mathbf{f}^T \mathbf{K}^- \mathbf{X}$$

is an isomorphism from $\mathcal{H}(\mathbf{K})$ to L^2_X and it satisfies

$$\text{Var}(\Psi(\mathbf{f})) = \mathbf{f}^T \mathbf{K}^- \mathbf{K} \mathbf{K}^- \mathbf{f} = \|\mathbf{f}\|^2_{\mathcal{H}(\mathbf{K})}.$$

So, if we start with $\mathcal{H}(\mathbf{K})$ and translate back via the isomorphism Ψ then $\mathbf{K}^- \mathbf{f} = \mathbf{a}$ is always in $\text{Ker}(\mathbf{K})^\perp$. Thus, by working in the RKHS we automatically avoid the annoying condition that we need $\mathbf{a} \in \text{Ker}(\mathbf{K}_X)^\perp$ and $\mathbf{b} \in \text{Ker}(\mathbf{K}_Y)^\perp$ that were imposed in Chapter III.

EXAMPLE 2. Let $\{X(t), t \in \mathcal{T}\}$ be a second-order stochastic process with mean function $\text{E}[X(t)] = \mu(t)$ and covariance function K. Let $\mathcal{T} = [0, 1]$ and assume that K is continuous on $\mathcal{T} \times \mathcal{T}$. Then, Mercer's theorem (e.g., see Riesz and Sz.-Nagy, 1955) insures that

$$K(s, t) = \sum_{q=1}^{\infty} \lambda_q \phi_q(s) \phi_q(t) \tag{4.17}$$

with $\lambda_1, \lambda_2, \ldots$ nonnegative eigenvalues and $\phi_1, \phi_2, \ldots$ in $L^2[0,1]$ continuous eigenfunctions of the integral operator

$$\int_0^1 K(s,t)\phi(t)dt = \lambda\phi(s).$$

Theorem IV.6 is now seen to be applicable with $Q = \{1, 2, \ldots\}, \nu(B) = \sum_{q\in B} \lambda_q$ for $B \in \mathcal{B}$ and $\phi(t,q) = \phi_q(t)$. So, the RKHS corresponding to K is

$$\mathcal{H}(K) = \{f(\cdot) = \sum_{q=1}^{\infty} \lambda_q g_q \phi_q(\cdot) : \sum_{q=1}^{\infty} \lambda_q g_q^2 < \infty\}.$$

For $f_i(\cdot) = \sum_{q=1}^{\infty} \lambda_q g_{iq}\phi_q(\cdot), i = 1, 2$, in $\mathcal{H}(K)$ the inner product is given by

$$\langle f_1, f_2\rangle_{\mathcal{H}(K)} = \sum_{q=1}^{\infty} \lambda_q g_{1q} g_{2q} = \sum_{q=1}^{\infty} \lambda_q^{-1} \langle f_1, \phi_q\rangle_{L^2[0,1]} \langle f_2, \phi_q\rangle_{L^2[0,1]}.$$

Now define a linear mapping Γ from $\overline{\text{span}}\{\phi_q\}_{q=1}^{\infty}$ in $L^2[0,1]$ to $\mathcal{H}(K)$ by

$$\Gamma(f) = \sum_{q=1}^{\infty} \lambda_q^{1/2} f_q \phi_q$$

for $f = \sum_{q=1}^{\infty} f_q\phi_q$ in $\overline{\text{span}}\{\phi_q\}_{q=1}^{\infty}$. Since $\Gamma(f) = \sum_{q=1}^{\infty} \lambda_q(\lambda_q^{-1/2} f_q)\phi_q$,

$$\|\Gamma(f)\|^2_{\mathcal{H}(K)} = \sum_{q=1}^{\infty} f_q^2 = \|f\|^2_{L^2[0,1]}.$$

Consequently, Γ is an isometric isomorphism, and $\overline{\text{span}}\{\phi_q\}_{q=1}^{\infty}$ and $\mathcal{H}(K)$ are congruent.

From the Karhunen-Loève representation, for $\tilde{X}(t) = X(t) - \mu(t)$, we have

$$\tilde{X}(t) = \sum_{q=1}^{\infty} \langle \tilde{X}, \phi_q\rangle_{L^2[0,1]} \phi_q(t),\ t \in [0,1].$$

Then, we have $dZ(q) = \langle \tilde{X}, \phi_q\rangle_{L^2[0,1]}$, which are uncorrelated and $\lambda_q = \mathrm{E}[dZ(q)]^2 = d\nu(q)$. Thus, (4.10) and (4.14) have the consequence that

$$\Psi(f) = \sum_{q=1}^{\infty} g_q \langle \tilde{X}, \phi_q\rangle_{L^2[0,1]} + \langle \mu, f\rangle_{\mathcal{H}(K)} = \sum_{q=1}^{\infty} g_q \langle X, \phi_q\rangle_{L^2[0,1]}$$

for any function $f(\cdot) = \sum_{q=1}^{\infty} \lambda_q g_q \phi_q(\cdot)$ with $\sum_{q=1}^{\infty} \lambda_q g_q^2 < \infty$ because we observe $\langle \mu, f \rangle_{\mathcal{H}(K)} = \sum_{q=1}^{\infty} g_q \langle \mu, \phi_q \rangle_{L^2[0,1]}$. In the special case that $\sum_{q=1}^{\infty} g_q^2 < \infty$ the function $\sum_{q=1}^{\infty} g_q \phi_q$ is a member of $L^2[0,1]$ and this produces

$$\Psi(f) = \langle X, \sum_{q=1}^{\infty} g_q \phi_q \rangle_{L^2[0,1]}. \tag{4.18}$$

Since $\{f \in \mathcal{H}(K) : \sum_{q=1}^{\infty} g_q^2 < \infty\}$ is not dense in $\mathcal{H}(K)$, (4.18) is generally only a partial characterization of Ψ.

CHAPTER V

CANONICAL CORRELATIONS FOR STOCHASTIC PROCESSES

In this chapter, we introduce a general formulation of canonical correlation analysis developed by Eubank and Hsing (2005). Let $\{X(t), t \in \mathcal{T}\}$ and $\{Y(s), s \in \mathcal{S}\}$ be second order stochastic processes with

$$\mathrm{E}[X(t)] = \mathrm{E}[Y(s)] = 0$$

for all $t \in \mathcal{T}, s \in \mathcal{S}$ and auto and cross covariance functions

$$K_X(t_1, t_2) = \mathrm{E}[X(t_1)X(t_2)],\ t_1, t_2 \in \mathcal{T},$$

$$K_Y(s_1, s_2) = \mathrm{E}[Y(s_1)Y(s_2)],\ s_1, s_2 \in \mathcal{S},$$

and

$$K_{XY}(t, s) = \mathrm{E}[X(t)Y(s)],\ t \in \mathcal{T}, s \in \mathcal{S}.$$

We are interested in developing a technique for decomposition of the covariance structure of the processes X and Y that is similar in spirit to the canonical correlation approach described in Chapter II.

5.1 Canonical Correlation Analysis

First, recall the classical canonical correlation problem in Chapter II. Let $\langle \cdot, \cdot \rangle_{\mathbb{R}^p}$ be the standard Euclidean inner product on $\mathbb{R}^p$. Our interest was in finding the random variables $\eta = \langle \mathbf{a}, \mathbf{X} \rangle_{\mathbb{R}^p}$ and $\xi = \langle \mathbf{b}, \mathbf{Y} \rangle_{\mathbb{R}^q}$ with $\mathbf{a} \in \mathrm{Ker}(\mathbf{K}_X)^{\perp}$ and $\mathbf{b} \in \mathrm{Ker}(\mathbf{K}_Y)^{\perp}$ having the largest possible correlation with each other. The goal is to extend this idea to canonical correlation problems in infinite dimensional spaces.

Let L_X^2 and L_Y^2 be the Hilbert spaces spanned by the processes X and Y, respectively, as defined in Section 4.2. The associated inner products are

$$\langle U_1, U_2 \rangle_{L_X^2} = \mathrm{E}[U_1 U_2], \quad \text{for } U_1, U_2 \in L_X^2$$

and

$$\langle V_1, V_2 \rangle_{L_Y^2} = \mathrm{E}[V_1 V_2], \quad \text{for } V_1, V_2 \in L_Y^2,$$

respectively.

In general, the goal of canonical correlation analysis is to find random variables $\eta \in L_X^2, \xi \in L_Y^2$ such that η and ξ are most strongly correlated with each other. In other words, we wish to find random variables $\eta \in L_X^2$ and $\xi \in L_Y^2$ maximizing

$$\rho^2(\eta, \xi) = \frac{\mathrm{Cov}^2(\eta, \xi)}{\mathrm{Var}(\eta)\mathrm{Var}(\xi)}. \tag{5.1}$$

Provided the above optimization problem can be solved, we define the first canonical correlation ρ_1 and the associated canonical variables η_1, ξ_1 by

$$\rho_1^2 = \mathrm{Cov}^2(\eta_1, \xi_1) = \sup_{\eta \in L_X^2, \xi \in L_Y^2} \mathrm{Cov}^2(\eta, \xi), \tag{5.2}$$

where η, ξ are subject to

$$\mathrm{Var}(\eta) = \mathrm{Var}(\xi) = 1. \tag{5.3}$$

For $i > 1$, the ith canonical correlation ρ_i and the associated canonical variables η_i, ξ_i are defined by

$$\rho_i^2 = \mathrm{Cov}^2(\eta_i, \xi_i) = \sup_{\eta \in L_X^2, \xi \in L_Y^2} \mathrm{Cov}^2(\eta, \xi), \tag{5.4}$$

where η, ξ are subject to (5.3) and

$$\mathrm{Cov}(\eta, \eta_j) = \mathrm{Cov}(\xi, \xi_j) = 0, \quad j < i. \tag{5.5}$$

If η_1 and ξ_1 are well defined in (5.2), then there are sequences $\eta_{1m} = \sum_{i=1}^m a_{im} X(t_{im})$ and $\xi_{1n} = \sum_{i=1}^n b_{in} Y(s_{in})$ such that $\rho_1^2 = \lim_{n,m \to \infty} \mathrm{Corr}^2(\eta_{1m}, \xi_{1n})$ since $\eta_1 \in L_X^2$ and

$\xi_1 \in L_Y^2$. Consequently, the infinite dimensional definition of canonical variables is actually built up from the finite dimensional multivariate case.

To see whether the canonical correlations are well defined, we will show that the optimization problems in (5.2)–(5.5) can be solved. For this purpose, we will use the fact that the Hilbert spaces L_X^2 and L_Y^2 and the reproducing kernel Hilbert spaces (RKHS) corresponding to the X and Y auto-covariance functions are congruent (or isometrically isomorphic). (e.g., see Parzen, 1961)

Before doing this in general we will first work with the case where both $\mathcal{T}$ and $\mathcal{S}$ are finite dimensional. This serves two purposes: it provides a motivational framework for understanding the general case and it provides a useful setting for the development of data analytic tools. Thus, first suppose that $\mathcal{T} = \{t_1, \ldots, t_p\}$, $\mathcal{S} = \{s_1, \ldots, s_q\}$, $\mathbf{X} = (X(t_1), \ldots, X(t_p))^T$ and $\mathbf{Y} = (Y(s_1), \ldots, Y(s_q))^T$ with $\mathbf{X}$ and $\mathbf{Y}$ the p-dimensional and q-dimensional random vectors that represent the X and Y processes in this case. Define

$$\text{Var}(\mathbf{X}) = \{K_X(t_i, t_j)\}_{i,j=1}^p = \mathbf{K}_X, \ \text{Var}(\mathbf{Y}) = \{K_Y(s_i, s_j)\}_{i,j=1}^q = \mathbf{K}_Y,$$

and

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = \{K_{XY}(t_i, s_j)\}_{i,j=1}^{p,q} = \mathbf{K}_{XY}.$$

As in Chapter III, we allow $\mathbf{K}_X$ and $\mathbf{K}_Y$ to have less than full rank: i.e., $r_X = r(\mathbf{K}_X) \leq p$ and $r_Y = r(\mathbf{K}_Y) \leq q$. The resulting Hilbert spaces spanned by the processes X and Y are then given by

$$L_X^2 = \left\{\mathbf{a}^T\mathbf{X} : \mathbf{a} \in \text{Ker}(\mathbf{K}_X)^\perp\right\}$$

and

$$L_Y^2 = \left\{\mathbf{b}^T\mathbf{Y} : \mathbf{b} \in \text{Ker}(\mathbf{K}_Y)^\perp\right\}$$

with associated squared norms $\|U\|_{L_X^2}^2 = \mathbf{a}^T\mathbf{K}_X\mathbf{a}$ and $\|V\|_{L_Y^2}^2 = \mathbf{b}^T\mathbf{K}_Y\mathbf{b}$ for $U = \mathbf{a}^T\mathbf{X} \in L_X^2$ and $V = \mathbf{b}^T\mathbf{Y} \in L_Y^2$.

As defined in Section 4.2, $\mathcal{H}(K_X)$ is the linear manifold of all vectors of the form

$$\mathbf{f} = \mathbf{K}_X \mathbf{a}$$

with inner product

$$\langle \mathbf{f}_1, \mathbf{f}_2 \rangle_{\mathcal{H}(K_X)} = \mathbf{f}_1^T \mathbf{K}_X^- \mathbf{f}_2, \tag{5.6}$$

for $\mathbf{f}_k = (f_k(t_1), \ldots, f_k(t_p))^T, k = 1, 2$, with $f_k(\cdot) = \sum_{i=1}^p a_{ki} K_X(\cdot, t_i)$ and $\mathbf{K}_X^-$ the Moore-Penrose generalized inverse of $\mathbf{K}_X$. Also, $\mathcal{H}(K_Y)$ is the linear manifold of all vectors of the form

$$\mathbf{g} = \mathbf{K}_Y \mathbf{b}$$

with inner product

$$\langle \mathbf{g}_1, \mathbf{g}_2 \rangle_{\mathcal{H}(K_Y)} = \mathbf{g}_1^T \mathbf{K}_Y^- \mathbf{g}_2,$$

for $\mathbf{g}_k = (g_k(s_1), \ldots, g_k(s_q))^T, k = 1, 2$, with $g_k(\cdot) = \sum_{i=1}^q b_{ki} K_Y(\cdot, s_i), k = 1, 2$, and $\mathbf{K}_Y^-$ the Moore-Penrose generalized inverse of $\mathbf{K}_Y$. Then, as explained in Example 1 in Section 4.2, $\mathcal{H}(K_X)$ and $\mathcal{H}(K_Y)$ are the RKHS's with r.k.'s K_X and K_Y, respectively. Note also that $\mathcal{H}(K_X) = \text{Ker}(\mathbf{K}_X)^\perp$ and $\mathcal{H}(K_Y) = \text{Ker}(\mathbf{K}_Y)^\perp$.

We now provide some results which allow us to relate the problem of maximizing (5.1) to an equivalent optimization problem in the RKHS. The mapping ψ_X from $\mathcal{H}(K_X)$ to L_X^2 and the mapping ψ_Y from $\mathcal{H}(K_Y)$ to L_Y^2 defined by

$$\psi_X(\mathbf{f}) = \mathbf{f}^T \mathbf{K}_X^- \mathbf{X} \text{ for } \mathbf{f} \in \mathcal{H}(K_X),$$

and

$$\psi_Y(\mathbf{g}) = \mathbf{g}^T \mathbf{K}_Y^- \mathbf{Y} \text{ for } \mathbf{g} \in \mathcal{H}(K_Y)$$

are the isometric isomorphisms from $\mathcal{H}(K_X)$ to L_X^2 and from $\mathcal{H}(K_Y)$ to L_Y^2, respectively. So, $\mathcal{H}(K_X)$ and L_X^2 are congruent and $\mathcal{H}(K_Y)$ and L_Y^2 are congruent.

As a result of the isometries we can write

$$\text{Corr}^2(\eta,\xi) = \frac{(\mathbf{a}^T\mathbf{K}_{XY}\mathbf{b})^2}{(\mathbf{a}^T\mathbf{K}_X\mathbf{a})(\mathbf{b}^T\mathbf{K}_Y\mathbf{b})} = \frac{(\mathbf{f}^T\mathbf{K}_X^-\mathbf{K}_{XY}\mathbf{K}_Y^-\mathbf{g})^2}{(\mathbf{f}^T\mathbf{K}_X^-\mathbf{f})(\mathbf{g}^T\mathbf{K}_Y^-\mathbf{g})} = \frac{\text{Cov}^2(\psi_X(\mathbf{f}),\psi_Y(\mathbf{g}))}{\text{Var}(\psi_X(\mathbf{f}))\text{Var}(\psi_Y(\mathbf{g}))}$$

for $\eta = \mathbf{a}^T\mathbf{X}$ and $\xi = \mathbf{b}^T\mathbf{Y}$ with $\mathbf{a} \in \text{Ker}(\mathbf{K}_X)^\perp$ and $\mathbf{b} \in \text{Ker}(\mathbf{K}_Y)^\perp$. Moreover, observe that $\mathbf{f}^T\mathbf{K}_X^-\mathbf{K}_{XY}\mathbf{K}_Y^-\mathbf{g} = \langle \mathbf{f}, \mathbf{K}_{XY}\mathbf{K}_Y^-\mathbf{g}\rangle_{\mathcal{H}(K_X)}$. Hence

$$\text{Corr}^2(\eta,\xi) = \frac{\langle \mathbf{f}, \mathbf{T}\mathbf{g}\rangle^2_{\mathcal{H}(K_X)}}{\|\mathbf{f}\|^2_{\mathcal{H}(K_X)}\|\mathbf{g}\|^2_{\mathcal{H}(K_Y)}}, \tag{5.7}$$

where

$$(\mathbf{T}\mathbf{g})(t) = \mathbf{K}_{XY}(t,\cdot)\mathbf{K}_Y^-\mathbf{g} = \langle \mathbf{K}_{XY}(t,\cdot), \mathbf{g}\rangle_{\mathcal{H}(K_Y)},\ t \in \mathcal{T}$$

with $\mathbf{K}_{XY}(t_i,\cdot)$ the ith row vector of $\mathbf{K}_{XY}$. Also,

$$\mathbf{f}^T\mathbf{K}_X^-\mathbf{K}_{XY}\mathbf{K}_Y^-\mathbf{g} = \langle \mathbf{T}^*\mathbf{f}, \mathbf{g}\rangle_{\mathcal{H}(K_Y)},$$

where $(\mathbf{T}^*\mathbf{f})(s) = \mathbf{K}_{YX}(s,\cdot)\mathbf{K}_X^-\mathbf{f} = \langle \mathbf{K}_{XY}(\cdot,s), \mathbf{f}\rangle_{\mathcal{H}(K_X)}, s \in \mathcal{S}$ with $\mathbf{K}_{XY}(\cdot,s_j)$ the jth column vector of $\mathbf{K}_{XY}$. So, $\mathbf{T}$ is a linear operator from $\mathcal{H}(K_Y)$ into $\mathcal{H}(K_X)$ with adjoint $\mathbf{T}^*$.

Now, the CCA problem in the finite dimensional case becomes

$$\sup_{\substack{\mathbf{f}\in\mathcal{H}(K_X),\,\mathbf{g}\in\mathcal{H}(K_Y)\\ \|\mathbf{f}\|_{\mathcal{H}(K_X)}=\|\mathbf{g}\|_{\mathcal{H}(K_Y)}=1}} \langle \mathbf{f}, \mathbf{T}\mathbf{g}\rangle^2_{\mathcal{H}(K_X)}.$$

Thus, CCA development in the $\mathcal{H}(K_X)$ and $\mathcal{H}(K_Y)$ setting proceeds via the singular value decomposition of the operator $\mathbf{T}$. To do this we find the eigenvalues and eigenvectors of $\mathbf{T}^*\mathbf{T}$ and $\mathbf{T}\mathbf{T}^*$. That is, the eigenvectors and eigenvalues are obtained from

$$\mathbf{T}^*\mathbf{T}\mathbf{g} = \rho^2\mathbf{g}$$

and

$$\mathbf{T}\mathbf{T}^*\mathbf{f} = \rho^2\mathbf{f}$$

which are

$$\mathbf{K}_{YX}\mathbf{K}_X^-\mathbf{K}_{XY}\mathbf{K}_Y^-\mathbf{g} = \rho^2\mathbf{g}$$

and

$$\mathbf{K}_{XY}\mathbf{K}_Y^-\mathbf{K}_{YX}\mathbf{K}_X^-\mathbf{f} = \rho^2\mathbf{f}.$$

Premultiplying by $\mathbf{K}_Y^-$ and $\mathbf{K}_X^-$ and employing the isometric isomorphisms ψ_X and ψ_Y then returns us to the original CCA solutions detailed in Chapter III.

It has been seen that finding the canonical correlation and variables for the X and Y processes on finite dimensional index sets $\mathcal{T}$ and $\mathcal{S}$ are equivalent to optimization in RKHS's generated by the X and Y auto covariance matrices. The next step is to extend this idea directly to the infinite dimensional case. For this purpose, we will define the notion of canonical correlation in this setting by directly generalizing the notion of canonical correlations in the finite dimensional case.

First let $\mathcal{H}(K_X)$ and $\mathcal{H}(K_Y)$ be the RKHS's with r.k.'s K_X and K_Y as defined in (4.5) in Theorem IV.4 with associated norms and inner products $\|\cdot\|_{\mathcal{H}(K_X)}, \langle\cdot,\cdot\rangle_{\mathcal{H}(K_X)}$ and $\|\cdot\|_{\mathcal{H}(K_Y)}, \langle\cdot,\cdot\rangle_{\mathcal{H}(K_Y)}$. As explained in Section 4.2, $\mathcal{H}(K_X)$ and L_X^2 are congruent and $\mathcal{H}(K_Y)$ and L_Y^2 are congruent. So, let ψ_X and ψ_Y be the isometric isomorphisms ψ_X from $\mathcal{H}(K_X)$ to L_X^2 and from $\mathcal{H}(K_Y)$ to L_Y^2, respectively, that satisfy

$$\psi_X : \sum_i a_i K_X(\cdot, t_i) \to \sum_i a_i X(t_i)$$

and

$$\psi_Y : \sum_j b_j K_Y(\cdot, s_j) \to \sum_j b_j Y(s_j).$$

Now every random variables $\eta \in L_X^2$ and $\xi \in L_Y^2$ can be written as

$$\eta = \psi_X(f) \quad \text{and} \quad \xi = \psi_Y(g)$$

for some unique functions f in $H(K_X)$ and g in $\mathcal{H}(K_Y)$.

Since two spaces that are isometrically isomorphic are algebraically and topologically identical, solving the optimization problems in L_X^2 and L_Y^2 is equivalent to solving the optimization problems which are formulated in the RKHS's $\mathcal{H}(K_X)$ and $\mathcal{H}(K_Y)$. We now can restate (5.2)-(5.5) in terms of optimization in $\mathcal{H}(K_X)$ and $\mathcal{H}(K_Y)$ as follows: Define the first canonical correlation ρ_1 and the associated RKHS vectors f_1, g_1 by

$$\rho_1^2 = \text{Cov}^2(\psi_X(f_1), \psi_Y(g_1)) = \sup_{f\in\mathcal{H}(K_X), g\in\mathcal{H}(K_Y)} \text{Cov}^2(\psi_X(f), \psi_Y(g)), \tag{5.8}$$

where f and g are subject to

$$\|f\|^2_{\mathcal{H}(K_X)} = \text{Var}(\psi_X(f)) = 1 = \text{Var}(\psi_Y(g)) = \|g\|^2_{\mathcal{H}(K_Y)}. \tag{5.9}$$

For $i > 1$, define the ith canonical correlations ρ_i and the associated RKHS vectors f_i, g_i by

$$\rho_i^2 = \text{Cov}^2(\psi_X(f_i), \psi_Y(g_i)) = \sup_{f\in\mathcal{H}(K_X), g\in\mathcal{H}(K_Y)} \text{Cov}^2(\psi_X(f), \psi_Y(g)), \tag{5.10}$$

where f and g are subject to (5.9) and

$$\text{Cov}(\psi_X(f), \psi_X(f_j)) = \text{Cov}(\psi_Y(g), \psi_Y(g_j)) = 0, \quad j < i. \tag{5.11}$$

For $\eta \in L_X^2, \xi \in L_Y^2$, there exist sequences $\eta_m = \sum_{i=1}^m a_{im}X(t_{im})$ and $\xi_n = \sum_{j=1}^n b_{jn}Y(s_{jn})$ such that $\text{Cov}(\psi_X(f), \psi_Y(g)) = \text{Cov}(\eta, \xi) = \lim_{n,m\to\infty} \text{Cov}(\eta_m, \xi_n)$. Hence

$$\begin{aligned}
\text{Cov}(\eta, \xi) &= \lim_{m,n\to\infty} \sum_{i=1}^m \sum_{j=1}^n a_{im}b_{jn}K_{XY}(t_{im}, s_{jn}) \\
&= \lim_{m,n\to\infty} \sum_{i=1}^m \sum_{j=1}^n a_{im}b_{jn}\langle K_{XY}(t_{im}, \cdot), K_Y(\cdot, s_{jn})\rangle_{\mathcal{H}(K_Y)} \\
&= \lim_{m,n\to\infty} \sum_{i=1}^m \sum_{j=1}^n a_{im}b_{jn}\langle K_X(t_{im}, *), \langle K_{XY}(*, \cdot), K_Y(\cdot, s_{jn})\rangle_{\mathcal{H}(K_Y)}\rangle_{\mathcal{H}(K_X)} \\
&= \lim_{m,n\to\infty} \langle \sum_{i=1}^m a_{im}K_X(t_{im}, *), \langle K_{XY}(*, \cdot), \sum_{j=1}^n b_{jn}K_Y(\cdot, s_{jn})\rangle_{\mathcal{H}(K_Y)}\rangle_{\mathcal{H}(K_X)}
\end{aligned}$$

by the reproducing properties of K_X and K_Y. Then, for $f_m(\cdot) = \sum_{i=1}^m a_{im} K_X(\cdot, t_{im})$ and $g_n(\cdot) = \sum_{j=1}^n b_{jn} K_Y(\cdot, s_{jn})$, we have

$$\begin{aligned}\text{Cov}(\psi_X(f), \psi_Y(g)) &= \lim_{m,n\to\infty} \langle f_m(*), \langle K_{XY}(*,\cdot), g_n(\cdot)\rangle_{\mathcal{H}(K_Y)}\rangle_{\mathcal{H}(K_X)} \\ &= \langle f(*), \langle K_{XY}(*,\cdot), g(\cdot)\rangle_{\mathcal{H}(K_Y)}\rangle_{\mathcal{H}(K_X)}\end{aligned}$$

with $f = \psi_X^{-1}(\eta) \in \mathcal{H}(K_X), g = \psi_Y^{-1}(\xi) \in \mathcal{H}(K_Y)$ the limits of the sequences f_m and g_n.

Now define the operator T from $\mathcal{H}(K_Y)$ to $\mathcal{H}(K_X)$ by

$$(Tg)(t) = \langle K_{XY}(t,\cdot), g(\cdot)\rangle_{\mathcal{H}(K_Y)}. \tag{5.12}$$

As a result of the above arguments

$$\text{Cov}(\psi_X(f), \psi_Y(g)) = \langle f, Tg\rangle_{\mathcal{H}(K_X)}$$

for any $f \in \mathcal{H}(K_X)$ and $g \in \mathcal{H}(K_Y)$. We then observe that

$$\begin{aligned}\langle f, Tg\rangle_{\mathcal{H}(K_X)} &= \text{Cov}(\psi_X(f), \psi_Y(g)) \\ &\leq \text{Var}(\psi_X(f))^{1/2}\text{Var}(\psi_Y(g))^{1/2} \\ &= \|f\|_{\mathcal{H}(K_X)}\|g\|_{\mathcal{H}(K_Y)}\end{aligned}$$

from the Cauchy-Schwarz inequality. Thus, when $f = Tg$ we have $\|Tg\|_{\mathcal{H}(K_X)} \leq \|g\|_{\mathcal{H}(K_Y)}$ and it follows that T is a bounded linear operator with operator norm at most 1. Also, from our previous development

$$\langle f_m, \langle K_{XY}(\cdot,*), g_n\rangle_{H(K_Y)}\rangle_{\mathcal{H}(K_X)} = \langle\langle f_m, K_{XY}(\cdot,*)\rangle_{\mathcal{H}(K_X)}, g_n\rangle_{H(K_Y)}$$

by the reproducing property. Taking limits as $n, m \to \infty$ then shows that

$$\langle f, Tg\rangle_{\mathcal{H}(K_X)} = \langle\langle f, K_{XY}(\cdot,*)\rangle_{\mathcal{H}(K_X)}, g\rangle_{\mathcal{H}(K_Y)};$$

i.e., the adjoint of $T \in B(\mathcal{H}(K_Y), \mathcal{H}(K_X))$ is given by

$$(T^* f)(s) = \langle f, K_{XY}(\cdot, s)\rangle_{\mathcal{H}(K_X)} \tag{5.13}$$

for $f \in \mathcal{H}(K_X)$.

We have now seen that $\text{Cov}(\psi_X(f), \psi_Y(g)) = \langle f, Tg \rangle_{\mathcal{H}(K_X)}$. So, analogous to the finite dimensional case, the polar representation of the bounded linear operator T in (4.3) should provide the solutions for the canonical problems (5.8)-(5.11) in the RKHS setting.

Suppose that the the largest value in the spectrum of T^*T, λ_1, is an eigenvalue of finite multiplicity with an associated eigenfunction g_1. That is,

$$\lambda_1 = \sup_{\|g\|_{\mathcal{H}(K_Y)}=1} \langle T^*Tg, g \rangle_{\mathcal{H}(K_Y)} = \sup_{\|g\|_{\mathcal{H}(K_Y)}=1} \int_{\sigma(T^*T)} \lambda dE_{g,g}(\lambda)$$

with $\sigma(T^*T)$ necessarily being a closed subset of $[0, 1]$. Then, $f_1 = Wg_1$ and $\eta_1 = \psi_X(f_1), \xi_1 = \psi_Y(g_1), \rho_1 = \lambda_1^{1/2}$. Continuing in this manner, if the second largest point in the spectrum is an eigenvalue of finite multiplicity, we have $f_2 = Wg_2$ and $\eta_2 = \Psi_X(f_2), \xi_2 = \psi_Y(g_2), \rho_2 = \lambda_2^{1/2}$, etc. However, in general, T^*T may not have any point spectra. In that case the canonical correlations and variables apparently cannot be defined.

An important special case of the previous development is the case where T is compact. As explained in Section 4.1, the spectrum $\sigma(T^*T)$ is known to consist of a countable set of non-zero eigenvalues with finite multiplicities and the polar representation of T is given by

$$T = \sum_{j=1}^{r(T)} \lambda_j^{1/2} \alpha_j \otimes \beta_j,$$

where $1 \geq \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{r(T)} > 0$ are the eigenvalues of T^*T with associated eigenvectors $\beta_j, j = 1, \ldots, r(T)$, and $\alpha_j = W\beta_j = T\beta_j/\lambda_j^{1/2}$. Then, the Cauchy-Schwarz and Bessel's inequalities ensure that ρ_i, f_i, g_i in (5.8) and (5.11) are given by $\rho_i = \lambda_i^{1/2}, f_i = \alpha_i, g_i = \beta_i$. Consequently, the canonical variables of the X space and Y space are

$$\eta_i = \psi_X(f_i) \quad \text{and} \quad \xi_i = \psi_Y(g_i),$$

where f_i, g_i are the eigenfunctions of TT^* and T^*T corresponding to their eigenvalues ρ_i^2, respectively, and f_i, g_i satisfy

$$\|f_i\|_{\mathcal{H}(K_X)} = \|g_i\|_{\mathcal{H}(K_Y)} = 1.$$

Finally, let us mention the relationship between the RKHS vectors f_i and g_i. For the polar representation of the compact operator T above, we have seen that

$$f_i = Tg_i/\rho_i,$$

or

$$Tg_i = \rho_i f_i. \tag{5.14}$$

Applying the operator T^* to both sides of (5.14) gives

$$T^* f_i = \rho_i g_i \tag{5.15}$$

since $T^*Tg_i = \rho_i^2 g_i$.

Now, suppose that the X and Y processes have non-zero mean functions $\mu_X(t) = \mathrm{E}[X(t)]$ and $\mu_Y(s) = \mathrm{E}[Y(s)]$ for all $t \in \mathcal{T}, s \in \mathcal{S}$. We see from Proposition IV.1 that if $\mu_X \in \mathcal{H}(K_X)$ and $\mu_Y \in \mathcal{H}(K_Y)$ then there exist linear mappings Ψ_X from $\mathcal{H}(K_X)$ to L_X^2 and Ψ_Y from $\mathcal{H}(K_Y)$ to L_Y^2. The linear mapping Ψ_X satisfies

$$\Psi_X(K_X(\cdot, t)) = X(t),\ t \in \mathcal{T},$$

$$\mathrm{E}[\Psi_X(f)] = \langle f, \mu_X \rangle_{\mathcal{H}(K_X)},$$

and

$$\mathrm{Cov}(\Psi_X(f^{(1)}), \Psi_Y(f^{(2)})) = \langle f^{(1)}, f^{(2)} \rangle_{\mathcal{H}(K_X)}.$$

The linear mapping Ψ_Y has similar properties. Thus, in this instance, (5.2)-(5.5) can be formulated exactly as before provided we use the linear mapping ψ_X, ψ_Y in lieu of isometries between L_X^2, L_Y^2 and $\mathcal{H}(R_X)$, $\mathcal{H}(R_Y)$.

5.2 Canonical Correlation Analysis and Regression

As we have shown in Section 3.1.2, linear regression can be viewed as a special case of CCA. In this section, we will demonstrate this remains true in the infinite dimensional setting.

Let Y be a random variable with zero mean and finite second moment. Let $\{X(t), t \in \mathcal{T}\}$ be a zero-mean, second order stochastic process with covariance kernel $K_X(s,t)$ for $s,t \in \mathcal{T}$. We observe the predictor function $\{X(t), t \in \mathcal{T}\}$ and the response variable Y. Assume without lose of generality that $\text{Var}(Y) = 1$.

For a linear regression problem we seek that random variable in L_X^2 whose mean square distance from Y is smallest. That is, we want to find a functional m satisfying

$$\inf_{m \in L_X^2} \text{E}|Y - m|^2. \tag{5.16}$$

The solution to this optimization problem was provided by Parzen (1961).

Recall now the RKHS $\mathcal{H}(K_X)$ determined by the covariance function K_X and the isometric isomorphism ψ_X between L_X^2 and $\mathcal{H}(K_X)$. Let $v(t) = \text{E}[YX(t)] = K_{XY}(t)$. The resulting best least-squares linear approximation of Y is then

$$m^*(\omega) = \psi_X(v)$$

with mean square error of prediction given by

$$\text{E}|Y - m^*|^2 = \text{E}|Y|^2 - \|v\|^2_{\mathcal{H}(K_X)} = 1 - \|v\|^2_{\mathcal{H}(K_X)}.$$

Now, the canonical correlation problem involving a zero-mean, second-order stochastic process X and a random variable Y with finite second moment can be defined as finding $\eta \in L_X^2$ to maximizes

$$\text{Corr}(\eta, Y) = \frac{\text{Cov}(\eta, Y)}{[\text{Var}(\eta)]^{1/2}}.$$

As in Section 5.1, the correlation between η and Y is written as

$$\frac{\langle K_{XY}(\cdot), f(\cdot)\rangle_{\mathcal{H}(K_X)}}{\|f\|_{\mathcal{H}(K_X)}}.$$

Hence the canonical variable of the X space is

$$\eta = \psi_X(f),$$

where f is the eigenvector of $TT^*f = \rho^2 f$ with the operator T^* from $\mathcal{H}(K_X)$ to $\mathbb{R}$ defined as

$$T^*f = \langle K_{XY}(\cdot), f(\cdot)\rangle_{\mathcal{H}(K_X)}, \quad f \in \mathcal{H}(K_X).$$

Also, from the fact that $\langle T^*f, g\rangle_{\mathbb{R}} = \langle f, Tg\rangle_{\mathcal{H}(K_X)}$, the operator T from $\mathbb{R}$ to $\mathcal{H}(K_X)$ is defined as

$$(Tg)(\cdot) = K_{XY}(\cdot)g, \quad g \in \mathbb{R}.$$

To demonstrate the connection between regression and CCA as in Section 3.1.2, we first observe that f can be obtained from

$$K_{XY}(\cdot)\langle K_{XY}, f\rangle_{\mathcal{H}(K_X)} = \rho^2 f(\cdot).$$

So, $\rho^2 = \|K_{XY}\|^2_{\mathcal{H}(K_X)}$ and we have only one canonical X variable $\eta = \psi_X(f)$ satisfying

$$\langle K_{XY}, f\rangle_{\mathcal{H}(K_X)} = \rho.$$

Thus,

$$f(\cdot) = \frac{K_{XY}(\cdot)}{\|K_{XY}\|_{\mathcal{H}(K_X)}}.$$

Since $v(\cdot) = K_{XY}(\cdot)$, the relationship between $v \in \mathcal{H}(K_X)$ and $f \in \mathcal{H}(K_X)$ are obtained by

$$v(\cdot) = \rho f(\cdot)$$

which is an exact parallel of what transpires for the finite dimensional setting.

CHAPTER VI

DISCRIMINANT ANALYSIS FOR STOCHASTIC PROCESSES

In this chapter we plan to extend the results from Chapter III concerning discriminant analysis to encompass stochastic processes. For this purpose, let $\{X(t), t \in \mathcal{T}\}$ be a second order stochastic process with mean function

$$\mu(t) = \mathrm{E}[X(t)]$$

and covariance function $K_X(s,t) = \mathrm{Cov}(X(s), X(t))$ for $s, t \in \mathcal{T}$. Also let G represent the class membership of the process from the populations numbered 1 to J. We define

$$\pi_j = P(G = j)$$

and

$$\mu_j(\cdot) = \mathrm{E}[X(\cdot)|G = j].$$

We assume that

$$K_j(s,t) = \mathrm{E}[(X(s) - \mu_j(s))(X(t) - \mu_j(t))|G = j],\ s, t \in \mathcal{T}$$

for $j = 1, \ldots, J$ have a common form that we denote by K_W. That is, $K_1 = \cdots = K_J = K_W$.

6.1 Discriminant Analysis

Let L_X^2 be the Hilbert space spanned by the X process with inner product

$$\langle U_1, U_2 \rangle_{L_X^2} = \mathrm{E}[U_1 U_2] \text{ for } U_1, U_2 \in L_X^2.$$

6.1.1 Fisher's Linear Discriminant Analysis

Let us begin by developing an infinite dimensional extension of Fisher's method. In that respect, we are interested in finding a random variable $\ell \in L_X^2$ maximizing

$$\text{Var}_G(\text{E}[\ell|G])/\text{E}_G[\text{Var}(\ell|G)] \tag{6.1}$$

which represents the ratio of between-class variability to within-class variability as in the finite dimensional case.

Define the kernel function

$$K_B(s,t) = \sum_{j=1}^{J} \pi_j(\mu_j(s) - \mu(s))(\mu_j(t) - \mu(t)),\ s,t \in \mathcal{T} \tag{6.2}$$

with

$$\mu(\cdot) = \sum_{j=1}^{J} \pi_j \mu_j(\cdot)$$

and let $\mathcal{H}(K_X)$ be the RKHS with the r.k. K_X as in Chapter IV. Also, denote the RKHS's generated by K_W and K_B by $\mathcal{H}(K_W)$ and $\mathcal{H}(K_B)$, respectively; i.e., let $\mathcal{H}(K_W) = \overline{\text{span}}\{K_W(\cdot,t), t \in \mathcal{T}\}$ and $\mathcal{H}(K_B) = \overline{\text{span}}\{K_B(\cdot,t), t \in \mathcal{T}\}$. Then, observe that

$$\sum_{i=1}^{n} c_i K_B(s,t_i) = \sum_{j=1}^{J} \pi_j b_j(\mu_j(s) - \mu(s)) = \sum_{j=1}^{J} \pi_j b_j \mu_j(s)$$

with $b_j = \sum_{i=1}^{n} c_i(\mu_j(t_i) - \mu(t_i))$ since $\sum_{j=1}^{J} \pi_j b_j = \sum_{i=1}^{n} c_i \sum_{j=1}^{J} \pi_j(\mu_j(t_i) - \mu(t_i)) = 0$. Consequently, we have shown

$$\mathcal{H}(K_B) = \Big\{ \sum_{j=1}^{J} c_j^* \mu_j(\cdot) : \sum_{j=1}^{J} c_j^* = 0 \Big\}$$

which consists of contrast among the class mean functions.

Now assume that $\mu_j \in \mathcal{H}(K_W)$. Clearly, $K_B(\cdot,t) \in \mathcal{H}(K_W)$. We then see that the set of elements in $\mathcal{H}(K_X)$ is equal to the set of elements for $\mathcal{H}(K_W)$; but, the two spaces are equipped with different norms. To see that $\mathcal{H}(K_X) = \mathcal{H}(K_W)$, for positive definite

functions $K_1(s,t)$ and $K(s,t)$, let us first write $K_1 \ll K$ if $K(s,t) - K_1(s,t)$ is also a positive definite function. Then, we know that $\mathcal{H}(K_W) \subset \mathcal{H}(K_X)$ and $\|h\|_{\mathcal{H}(K_X)} \leq \|h\|_{\mathcal{H}(K_W)}$ for $h \in \mathcal{H}(K_W)$ since $K_W \ll K_X$ as a result of Theorem I in Aronszajn (1950).

Now define a linear operator L from $\mathcal{H}(K_X)$ to $\mathcal{H}(K_W)$ satisfying

$$L(K_X(\cdot,t)) = K_W(\cdot,t),\ t \in \mathcal{T}. \tag{6.3}$$

Then, L is a one-to-one and onto linear mapping since

$$\begin{aligned}
\|\sum_{i=1}^{n} a_i K_X(\cdot,t_i)\|^2_{\mathcal{H}(K_X)} &= \sum_{i=1}^{n}\sum_{k=1}^{n} a_i a_k K_X(t_i,t_k) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{n} a_i a_k K_W(t_i,t_k) + \sum_{i=1}^{n}\sum_{k=1}^{n} a_i a_k K_B(t_i,t_k) \\
&= \|\sum_{i=1}^{n} a_i K_W(\cdot,t_i)\|^2_{\mathcal{H}(K_W)} \\
&\quad + \sum_{j=1}^{J} \pi_j \big|\langle \mu_j - \mu, \sum_{i=1}^{n} a_i K_W(\cdot,t_i)\rangle_{\mathcal{H}(K_W)}\big|^2.
\end{aligned}$$

Also, we observe that, for $h \in \mathcal{H}(K_W)$ and $f \in \mathcal{H}(K_X)$,

$$\langle h, f\rangle_{\mathcal{H}(K_X)} = \langle h, Lf\rangle_{\mathcal{H}(K_W)} \tag{6.4}$$

which follows from the fact that

$$\langle h, K_X(\cdot,t)\rangle_{\mathcal{H}(K_X)} = h(t) = \langle h, K_W(\cdot,t)\rangle_{\mathcal{H}(K_W)} = \langle h, L(K_X(\cdot,t))\rangle_{\mathcal{H}(K_W)}.$$

The operator L is bounded with operator norm at most one because

$$\langle Lf, f\rangle_{\mathcal{H}(K_X)} = \langle Lf, Lf\rangle_{\mathcal{H}(K_W)} = \|Lf\|^2_{\mathcal{H}(K_W)} \geq \|Lf\|^2_{\mathcal{H}(K_X)}.$$

Hence, $\|Lf\|^2_{\mathcal{H}(K_X)} \leq \langle Lf, f\rangle_{\mathcal{H}(K_X)} \leq \|Lf\|_{\mathcal{H}(K_X)}\|f\|_{\mathcal{H}(K_X)}$ and $\|Lf\|_{\mathcal{H}(K_X)} \leq \|f\|_{\mathcal{H}(K_X)}$. Further, $L \in B(\mathcal{H}(K_X), \mathcal{H}(K_W))$ is positive because $\langle Lf, f\rangle_{\mathcal{H}(K_X)} = \|Lf\|^2_{\mathcal{H}(K_W)} \geq 0$.

For $f \in \mathcal{H}(K_X)$, observe that

$$\begin{aligned} f(t) &= \langle K_X(\cdot, t), f \rangle_{\mathcal{H}(K_X)} \\ &= \langle K_B(\cdot, t), f \rangle_{\mathcal{H}(K_X)} + \langle K_W(\cdot, t), f \rangle_{\mathcal{H}(K_X)} \\ &= \langle K_B(\cdot, t), Lf \rangle_{\mathcal{H}(K_W)} + \langle K_W(\cdot, t), Lf \rangle_{\mathcal{H}(K_W)}. \end{aligned}$$

Thus, let us now define the operator T_B from $\mathcal{H}(K_W)$ to $\mathcal{H}(K_W)$ by

$$(T_B h)(t) = \langle K_B(t, \cdot), h(\cdot) \rangle_{\mathcal{H}(K_W)} \tag{6.5}$$

for $h \in \mathcal{H}(K_W)$. Then, $f(t) = (T_B Lf)(t) + (Lf)(t) \in \mathcal{H}(K_W)$ and hence $\mathcal{H}(K_X) \subset \mathcal{H}(K_W)$. Therefore, $\mathcal{H}(K_X) = \mathcal{H}(K_W)$.

As in Section 5.1, let us first consider the problem of maximizing (6.1) in the finite dimensional case. Suppose that $\mathcal{T} = \{t_1, \ldots, t_p\}$ and let $\mathbf{X} = (X(t_1), \ldots, X(t_p))^T$ with $\mathbf{K}_X = \{K_X(t_k, t_l)\}_{k,l=1}^p$. Set $\mathbf{K}_W = \{K_W(t_k, t_l)\}_{k,l=1}^p$ and $\mathbf{K}_B = \{K_B(t_k, t_l)\}_{k,l=1}^p$. Then, the linear discriminant functions in the finite dimensional case are obtained from the classical multivariate setting as in Section 3.2.3.1. To see this first note that in this instance we had $L_X^2 = \{\mathbf{a}^T\mathbf{X} : \mathbf{a} \in \text{Ker}(\mathbf{K}_X)^\perp\}$ with squared norm

$$\text{E}[(\mathbf{a}^T\mathbf{X})^2] = \mathbf{a}^T\mathbf{K}_X\mathbf{a} + (\mathbf{a}^T\boldsymbol{\mu})^2$$

for $\mathbf{a} = (a_1, \ldots, a_p)^T$. The corresponding RKHS is

$$\mathcal{H}(K_X) = \{\mathbf{f} = \mathbf{K}_X\mathbf{a} : \mathbf{a} \in \text{Ker}(\mathbf{K}_X)^\perp\}$$

with associated inner product

$$\langle \mathbf{f}_1, \mathbf{f}_2 \rangle_{\mathcal{H}(K_X)} = \mathbf{f}_1^T\mathbf{K}_X^-\mathbf{f}_2.$$

Assume that $\boldsymbol{\mu}_j \in \text{Ker}(\mathbf{K}_W)^\perp$ for all j. Then, $\text{Ker}(\mathbf{K}_X)^\perp = \text{Ker}(\mathbf{K}_W)^\perp$ and

$$L_X^2 = \{\mathbf{a}^T\mathbf{X} : \mathbf{a} \in \text{Ker}(\mathbf{K}_W)^\perp\}.$$

Also, the RKHS determined by K_W, which is equal set-wise to $\mathcal{H}(K_X)$, is

$$\mathcal{H}(K_W) = \{\mathbf{h} = \mathbf{K}_W\mathbf{a} : \mathbf{a} \in \text{Ker}(\mathbf{K}_W)^\perp\}$$

with associated inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}(K_W)}$ producing the squared norm

$$\|\mathbf{h}\|^2_{\mathcal{H}(K_W)} = \mathbf{h}^T\mathbf{K}_W^-\mathbf{h}.$$

Just as in Section 5.1, with finite dimensions, an isomorphism (one-to-one and onto linear mapping) Ψ_W from $\mathcal{H}(K_W)$ to L^2_X is given by

$$\Psi_W(\mathbf{h}) = \mathbf{h}^T\mathbf{K}_W^-\mathbf{X},\ \mathbf{h} \in \mathcal{H}(K_W).$$

Then, we observe that

$$\text{E}_G[\text{Var}(\Psi_W(\mathbf{h})|G)] = \mathbf{h}^T\mathbf{K}_W^-\mathbf{K}_W\mathbf{K}_W^-\mathbf{h} = \|\mathbf{h}\|^2_{\mathcal{H}(K_W)}.$$

COROLLARY **VI.1**. *Let L^2_X be the Hilbert space spanned by the process $\{X(t), t \in \mathcal{T}\}$ with $\mathcal{T} = \{t_1, \ldots, t_p\}$ and let $\mathcal{H}(K_W)$ be the RKHS generated by the within-class covariance function K_W. Then, maximizing (6.1) over $\ell \in L^2_X$ is equivalent to maximizing*

$$\langle\mathbf{h}, \mathbf{T}_B\mathbf{h}\rangle_{\mathcal{H}(K_W)} \tag{6.6}$$

over $\mathbf{h} \in \mathcal{H}(K_W)$ subject to $\|\mathbf{h}\|^2_{\mathcal{H}(K_W)} = 1$, where

$$(\mathbf{T}_B\mathbf{h})(t) = \mathbf{K}_B(t,\cdot)\mathbf{K}_W^-\mathbf{h} = \langle\mathbf{K}_B(t,\cdot), \mathbf{h}\rangle_{\mathcal{H}(K_W)}$$

with $\mathbf{K}_B(t_i,\cdot)$ the ith row vector of $\mathbf{K}_B$.

Proof. In the finite dimensional case, $\ell = \mathbf{a}^T\mathbf{X}$ and (6.1) becomes

$$\frac{\mathbf{a}^T\mathbf{K}_B\mathbf{a}}{\mathbf{a}^T\mathbf{K}_W\mathbf{a}} = \frac{\mathbf{h}^T\mathbf{K}_W^-\mathbf{K}_B\mathbf{K}_W^-\mathbf{h}}{\mathbf{h}^T\mathbf{K}_W^-\mathbf{h}} \tag{6.7}$$

since $\mathbf{h} = \mathbf{K}_W\mathbf{a}$ with $\mathbf{a} = (a_1, \ldots, a_p)^T$. Now observe that

$$\mathbf{K}_B(t,\cdot)\mathbf{K}_W^-\mathbf{h} = \langle\mathbf{K}_B(t,\cdot), \mathbf{h}\rangle_{\mathcal{H}(K_W)}.$$

So, (6.7) becomes

$$\frac{\langle \mathbf{h}, \mathbf{K}_B \mathbf{K}_W^- \mathbf{h} \rangle_{\mathcal{H}(K_W)}}{\|\mathbf{h}\|^2_{\mathcal{H}(K_W)}} = \frac{\langle \mathbf{h}, \mathbf{T}_B \mathbf{h} \rangle_{\mathcal{H}(K_W)}}{\|\mathbf{h}\|^2_{\mathcal{H}(K_W)}} \tag{6.8}$$

and the result follows.

□

The solution to the optimization problem in (6.6) can be obtained from the eigenvalue decomposition of the finite dimensional operator $\mathbf{T}_B$. If we start with the eigenvalue decomposition of the operator $\mathbf{T}_B$ on $B(\mathcal{H}(K_W))$ then the eigenvalues and eigenvectors are obtained from

$$\mathbf{T}_B \mathbf{h} = \gamma \mathbf{h}$$

which is

$$\mathbf{K}_B \mathbf{K}_W^- \mathbf{h} = \gamma \mathbf{h}.$$

Premultiplying by $\mathbf{K}_W^-$ and using the isomorphism Ψ_W returns us to the matrix case in (3.29).

Now we wish to extend this idea to the problem of finding linear discriminant functions in the infinite dimensional setting. To do this, we first establish the following result.

PROPOSITION **VI.1**. *Assume that $\mu_j \in \mathcal{H}(K_W)$ for all $j = 1, \ldots, J$. Then, there exists a one-to-one linear mapping Ψ_W between $\mathcal{H}(K_W)$ and L^2_X defined by*

$$\Psi_W(K_W(\cdot, t)) = X(t)$$

for every t in $\mathcal{T}$ with the properties

$$\mathrm{E}[\Psi_W(h)] = \langle h, \mu \rangle_{\mathcal{H}(K_W)}, \tag{6.9}$$

$$\mathrm{E}[\Psi_W(h)|G = j] = \langle h, \mu_j \rangle_{\mathcal{H}(K_W)}, \tag{6.10}$$

$$\mathrm{E}_G[\mathrm{Var}(\Psi_W(h)|G)] = \|h\|^2_{\mathcal{H}(K_W)} \tag{6.11}$$

for $h \in \mathcal{H}(K_W)$ *and*

$$\mathrm{Cov}(\Psi_W(h^{(1)}), \Psi_W(h^{(2)})) = \langle h^{(1)}, h^{(2)} \rangle_{\mathcal{H}(K_W)} + \langle h^{(1)}, T_B h^{(2)} \rangle_{\mathcal{H}(K_W)} \tag{6.12}$$

for $h^{(1)}, h^{(2)} \in \mathcal{H}(K_W)$ *and* T_B *defined in (6.5).*

Proof. For any function of the form $h(\cdot) = \sum_{k=1}^{n} a_k K_W(\cdot, t_k)$ define

$$\Psi_W(h) = \sum_{k=1}^{n} a_k X(t_k).$$

Then, $\Psi_W(h)$ is well defined as a member of L_X^2 since

$$\begin{aligned}
\mathrm{E}\Big|\sum_{k=1}^{n} a_k X(t_k)\Big|^2 &= \sum_{k=1}^{n}\sum_{l=1}^{n} a_k a_l [K_X(t_k, t_l) + \mu(t_k)\mu(t_l)] \\
&= \sum_{k=1}^{n}\sum_{l=1}^{n} a_k a_l \Big[K_W(t_k, t_l) \\
&\qquad + \sum_{j=1}^{J} \pi_j (\mu_j(t_k) - \mu(t_k))(\mu_j(t_l) - \mu(t_j)) + \mu(t_k)\mu(t_l) \Big] \\
&= \sum_{k=1}^{n}\sum_{l=1}^{n} a_k a_l K_W(t_k, t_l) + \sum_{j=1}^{J} \pi_j \sum_{k=1}^{n}\sum_{l=1}^{n} a_k a_l \mu_j(t_k)\mu_j(t_l) \\
&= \Big\| \sum_{k=1}^{n} a_k K_W(\cdot, t_k) \Big\|_{\mathcal{H}(K_W)}^2 + \sum_{j=1}^{J} \pi_j \Big| \langle \mu_j, \sum_{k=1}^{n} a_k K_W(\cdot, t_k) \rangle_{\mathcal{H}(K_W)} \Big|^2
\end{aligned}$$

by the reproducing property of K_W. So,

$$\Psi_W(h) = \sum_{k=1}^{n} a_k X(t_k) = 0 \quad \text{if and only if} \quad h(\cdot) = \sum_{k=1}^{n} a_k K_W(\cdot, t_k) = 0.$$

It is now clear that Ψ_W defines a one-to-one linear mapping from the linear manifold spanned by $\{K_W(\cdot, t), t \in \mathcal{T}\}$ onto the linear manifold spanned by the X process with the properties

$$\mathrm{E}[\Psi_W(h)] = \sum_{k=1}^{n} a_k \mu(t_k) = \langle \sum_{k=1}^{n} a_k K_W(\cdot, t_k), \mu(\cdot) \rangle_{\mathcal{H}(K_W)} = \langle h, \mu \rangle_{\mathcal{H}(K_W)},$$

$$\mathrm{E}[\Psi_W(h) | G = j] = \sum_{k=1}^{n} a_k \mu_j(t_k) = \langle \sum_{k=1}^{n} a_k K_W(\cdot, t_k), \mu_j(\cdot) \rangle_{\mathcal{H}(K_W)} = \langle h, \mu_j \rangle_{\mathcal{H}(K_W)},$$

$$\begin{aligned}
\mathrm{E}_G[\mathrm{Var}(\Psi_W(h)|G)] &= \sum_{k=1}^{n}\sum_{l=1}^{n} a_k a_l K_W(t_k, t_l) \\
&= \langle \sum_{k=1}^{n} a_k K_W(\cdot, t_k), \sum_{l=1}^{n} a_l K_W(\cdot, t_l) \rangle_{\mathcal{H}(K_W)} \\
&= \| \sum_{k=1}^{n} a_k K_W(\cdot, t_k) \rangle^2_{\mathcal{H}(K_W)} = \|h\|^2_{\mathcal{H}(K_W)},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Cov}(\Psi_W(h^{(1)}), \Psi_W(h^{(2)})) &= \sum_{k=1}^{n}\sum_{l=1}^{m} a_k b_l K_X(t_k, s_l) \\
&= \sum_{k=1}^{n}\sum_{l=1}^{m} a_k b_l K_W(t_k, s_l) + \sum_{k=1}^{n}\sum_{l=1}^{m} a_k b_l K_B(t_k, s_l) \\
&= \langle \sum_{k=1}^{n} a_k K_W(\cdot, t_k), \sum_{l=1}^{m} b_l K_W(\cdot, s_l) \rangle_{\mathcal{H}(K_W)} \\
&\quad + \langle \sum_{k=1}^{n} a_k K_W(\cdot, t_k), \langle K_B(\cdot, *), \sum_{l=1}^{n} b_l K_W(*, s_l) \rangle_{\mathcal{H}(K_W)} \\
&= \langle h^{(1)}, h^{(2)} \rangle_{\mathcal{H}(K_W)} + \langle h^{(1)}(\cdot), \langle K_B(\cdot, *), h^{(2)}(*) \rangle_{\mathcal{H}(K_W)} \rangle_{\mathcal{H}(K_W)}.
\end{aligned}$$

Moreover, Cauchy sequences in L^2_X correspond to Cauchy sequences in $\mathcal{H}(K_W)$ and conversely as a result of the identity

$$\mathrm{E}|\Psi_W(h_n) - \Psi_W(h_m)|^2 = \|h_n - h_m\|^2_{\mathcal{H}(K_W)} + \sum_{j=1}^{J} \pi_j |\langle \mu_j, h_n - h_m \rangle_{\mathcal{H}(K_W)}|^2.$$

Thus, the result follows.

□

For any $\ell \in L^2_X$, there exists a sequence $\ell_n = \sum_{t \in \mathcal{T}_n} a_t X(t)$ with $\mathcal{T}_n$ being n dimensional subsets of $\mathcal{T}$ such that

$$\lim_{n \to \infty} \mathrm{E}[\ell_n - \ell]^2 = 0.$$

Then, $\mathrm{Var}_G(\mathrm{E}[\ell_n - \ell|G]) \leq \mathrm{Var}(\ell_n - \ell) \leq \mathrm{E}[\ell_n - \ell]^2$ which has the consequence that $\lim_{n\to\infty} \mathrm{Cov}_G(\mathrm{E}[\ell_n|G], \mathrm{E}[\ell|G]) = \mathrm{Var}_G(\mathrm{E}[\ell|G])$ as a result of $\lim_{n\to\infty} \mathrm{Cov}_G(\mathrm{E}[\ell_n|G] - \mathrm{E}[\ell|G], \mathrm{E}[\ell|G]) = 0$ and the Cauchy-Schwarz inequality. So, we see that $\mathrm{Var}_G(\mathrm{E}[\ell|G]) =$

$\lim_{n\to\infty} \mathrm{Var}_G(\mathrm{E}[\ell_n|G])$ or

$$\begin{aligned}
\mathrm{Var}_G(\mathrm{E}[\ell|G]) &= \lim_{n\to\infty} \sum_{s\in\mathcal{T}_n}\sum_{t\in\mathcal{T}_n} a_s a_t K_B(s,t) \\
&= \lim_{n\to\infty} \sum_{s\in\mathcal{T}_n}\sum_{t\in\mathcal{T}_n} a_s a_t \langle K_B(s,\cdot), K_W(\cdot,t)\rangle_{\mathcal{H}(K_W)} \\
&= \lim_{n\to\infty} \sum_{s\in\mathcal{T}_n}\sum_{t\in\mathcal{T}_n} a_s a_t \langle K_W(s,*), \langle K_B(*,\cdot), K_W(\cdot,t)\rangle_{\mathcal{H}(K_W)}\rangle_{\mathcal{H}(K_W)} \\
&= \lim_{n\to\infty} \langle \sum_{s\in\mathcal{T}_n} a_s K_W(s,*), \langle K_B(*,\cdot), \sum_{t\in\mathcal{T}_n} a_t K_W(\cdot,t)\rangle_{\mathcal{H}(K_W)}\rangle_{\mathcal{H}(K_W)}
\end{aligned}$$

by the reproducing property of K_W. Hence, for any sequence $h_n = \sum_{t\in\mathcal{T}_n} a_t K_W(\cdot,t)$ converging to h in the norm of $\mathcal{H}(K_W)$, we have

$$\begin{aligned}
\mathrm{Var}_G(\mathrm{E}[\ell|G]) &= \lim_{n\to\infty} \langle h_n(*), \langle K_B(*,\cdot), h_n(\cdot)\rangle_{\mathcal{H}(K_W)}\rangle_{\mathcal{H}(K_W)} \\
&= \langle h(*), \langle K_B(*,\cdot), h(\cdot)\rangle_{\mathcal{H}(K_W)}\rangle_{\mathcal{H}(K_W)}.
\end{aligned}$$

Then, using the isomorphism Ψ_W reveals that

$$\mathrm{Var}_G[\mathrm{E}(\ell|G)] = \mathrm{Var}_G[\mathrm{E}(\Psi_X(h)|G)] = \langle h, T_B h\rangle_{\mathcal{H}(K_W)}.$$

THEOREM **VI.1**. *The operator T_B from $\mathcal{H}(K_W)$ to $\mathcal{H}(K_W)$ in (6.5) is bounded, self-adjoint, positive and compact.*

Proof. We observe from (6.12) that

$$\begin{aligned}
\left|\langle h^{(1)}, T_B h^{(2)}\rangle_{\mathcal{H}(K_W)}\right| &= \sum_{j=1}^{J} \pi_j \langle h^{(1)}, \mu_j - \mu\rangle_{\mathcal{H}(K_W)} \langle \mu_j - \mu, h^{(2)}\rangle_{\mathcal{H}(K_W)} \\
&\le \|h^{(1)}\|_{\mathcal{H}(K_W)} \|h^{(2)}\|_{\mathcal{H}(K_W)} \sum_{j=1}^{J} \pi_j \|\mu_j - \mu\|^2_{\mathcal{H}(K_W)} \\
&= M \|h^{(1)}\|_{\mathcal{H}(K_W)} \|h^{(2)}\|_{\mathcal{H}(K_W)}
\end{aligned}$$

with $M = \sum_{j=1}^{J} \pi_j \|\mu_j - \mu\|^2_{\mathcal{H}(K_W)} < \infty$ since $\mu_j - \mu \in \mathcal{H}(K_W)$ for all j. Replacing $h^{(1)}$ by $T_B h^{(2)}$ entails that $\|T_B h^{(2)}\|_{\mathcal{H}(K_W)} \le M \|h^{(2)}\|_{\mathcal{H}(K_W)}$ and so T_B is a bounded linear operator.

The operator T_B in $B(\mathcal{H}(K_W))$ clearly has finite rank since

$$(T_B h)(t) = \sum_{j=1}^{J} \pi_j(\mu_j(t) - \mu(t))\langle \mu_j - \mu, h\rangle_{\mathcal{H}(K_W)}$$

and so $\text{Im}(T_B) = \text{span}\{\mu_j - \mu, j = 1, \ldots, J : \sum_{j=1}^{J} \pi_j(\mu_j(\cdot) - \mu(\cdot)) = 0\}$. Thus, $r(T_B) \leq J - 1$, which means that T_B is compact. Also, observe that

$$\langle h^{(1)}, T_B h^{(2)}\rangle_{\mathcal{H}(K_W)} = \sum_{j=1}^{J} \pi_j \langle h^{(1)}, \mu_j - \mu\rangle_{\mathcal{H}(K_W)} \langle \mu_j - \mu, h^{(2)}\rangle_{\mathcal{H}(K_W)} = \langle T_B h^{(1)}, h^{(2)}\rangle_{\mathcal{H}(K_W)}$$

and that

$$\langle T_B f, f\rangle_{\mathcal{H}(K_W)} = \sum_{j=1}^{J} \pi_j \langle \mu_j - \mu, f\rangle^2_{\mathcal{H}(K_W)} \geq 0.$$

So, $T_B \in B(\mathcal{H}(K_W))$ is a self-adjoint, compact and positive operator.

□

We now can restate the discrimination problem (6.1) in the RKHS $\mathcal{H}(K_X)$ as follows: The RKHS variate f can be obtained by solving

$$\sup_{h \in \mathcal{H}(K_W)} \text{Var}_G(\text{E}[\Psi_W(h)|G]) \tag{6.13}$$

subject to

$$\text{E}_G[\text{Var}(\Psi_W(h)|G)] = \|h\|^2_{\mathcal{H}(K_W)} = 1. \tag{6.14}$$

It is seen that $\text{Var}_G(\text{E}[\Psi_W(h)|G]) = \langle h, T_B h\rangle_{\mathcal{H}(K_W)}$ and hence characterization of the solutions to problem (6.13) can be achieved by the study of the operator T_B.

Analogous to the finite dimensional case, the spectral decomposition of T_B will provide the solutions to the optimization in (6.1). Thus, as in (4.1), write T_B as

$$T_B = \sum_{i=1}^{J-1} \gamma_i \alpha_i \otimes \alpha_i, \tag{6.15}$$

where $\gamma_1 \geq \ldots \geq \gamma_{J-1} > 0$ are eigenvalues of T_B and $\alpha_i, i = 1, \ldots, J-1$, are the associated eigenfunctions. Note that $\{\alpha_i, i = 1, \ldots, J-1\}$ in $\mathcal{H}(K_W)$ are an orthonormal basis for $\overline{\text{Im}(T_B)} = \text{Im}(T_B)$.

THEOREM **VI.2**. *Suppose that T_B has the spectral decomposition in (6.15). Then, the h_i satisfying (6.13) and (6.14) are given by $h_i = \alpha_i$ and the corresponding linear discriminant variables of the X space are $\ell_i = \Psi_W(f_i)$.*

Proof. We have

$$\text{Var}_G(\text{E}[\Psi_W(h)|G]) = \langle h, T_B h\rangle_{\mathcal{H}(K_W)} = \sum_{i=1}^{J-1} \gamma_i \langle \alpha_i, h\rangle^2_{\mathcal{H}(K_W)}.$$

Since the $\{\alpha_j\}$ are orthonormal in $\mathcal{H}(K_X)$,

$$\text{Var}_G(\text{E}[\Psi_W(h)|G]) \leq \gamma_1 \sum_{i=1}^{J-1} \langle \alpha_i, h\rangle^2_{\mathcal{H}(K_W)} \leq \gamma_1 \|h\|^2_{\mathcal{H}(K_W)}$$

by Bessel's inequality. Then, the equality holds if and only if $h = \alpha_1$. For the general case we have $h \perp \alpha_i, 1 \leq i \leq k-1$ and $\text{Var}_G(\text{E}[\Psi_W(h)|G]) \leq \gamma_k \|h\|^2_{\mathcal{H}(K_W)}$, with inequality if and only if $h = \alpha_k$.

□

We now see that the linear discriminant functions are given by

$$\ell_i = \Psi_W(h_i),\ i = 1, \ldots, J-1, \tag{6.16}$$

where the h_i are the eigenvectors of T_B corresponding to its positive eigenvalue γ_i. The h_i satisfy the constraints

$$\|h_i\|_{\mathcal{H}(K_W)} = 1$$

and

$$\langle h_i, h_k\rangle_{\mathcal{H}(K_W)} = 0,\ k \neq i.$$

With $x(\cdot) = X(\cdot, \omega_0)$ for $\omega_0 \in \Omega$

$$\ell(\omega_0) = \Psi_W(h)(\omega_0) := \Psi_{W,x}(h).$$

Let us now adopt the notation $\ell(x) = \Psi_{W,x}(h)$ instead of ℓ to explicitly emphasize the dependency on x. The classification rule is then to classify a new curve x to class i if

$$Dist_i^s(x) = \min_j Dist_j^s(x),$$

where the squared Mahalanobis distance $Dist_j^s$ confined to the subspace defined by the first s ($\leq J-1$) linear discriminant functions is given by

$$Dist_j^s(x) = \sum_{k=1}^{s} \big(\ell_k(x) - \mathrm{E}[\ell_k|G=j]\big)^2 = \sum_{k=1}^{s} \big(\Psi_{W,x}(h_k) - \langle h_k, \mu_j\rangle_{\mathcal{H}(K_W)}\big)^2. \quad (6.17)$$

6.1.2 Generalized Fisher's Linear Discriminant Analysis

We now wish to formulate the general version of Fisher's discrimination method in this section with respect to optimization over $\mathcal{H}(K_X)$. In this regard, maximizing (6.1) over $\ell \in L_X^2$ is equivalent to maximizing

$$\mathrm{Var}_G(\mathrm{E}[\ell|G])/\mathrm{Var}(\ell) \quad (6.18)$$

over $\ell \in L_X^2$ since $\mathrm{Var}(\ell) = \mathrm{Var}_G(\mathrm{E}[\ell|G]) + \mathrm{E}_G[\mathrm{Var}(\ell|G)]$ implies that (6.18) equals

$$\frac{\mathrm{Var}_G(\mathrm{E}[\ell|G])/\mathrm{E}_G[\mathrm{Var}(\ell|G)]}{1+\mathrm{Var}_G(\mathrm{E}[\ell|G])/\mathrm{E}_G[\mathrm{Var}(\ell|G)]}$$

and $\frac{x}{1+x}$ is an increasing function in $x \geq 0$.

We have seen from Proposition IV.1 that given $\mu_j \in \mathcal{H}(K_W)$ a linear mapping Ψ_X between $\mathcal{H}(K_X)$ and L_X^2 defined by

$$\Psi_X(K_X(\cdot,t)) = X(t),\ t \in \mathcal{T}$$

is an isomorphism with the properties

$$\mathrm{E}[\Psi_X(f)] = \langle f, \mu\rangle_{\mathcal{H}(K_X)} \quad (6.19)$$

and

$$\mathrm{Cov}(\Psi_X(f^{(1)}), \Psi_X(f^{(2)})) = \langle f^{(1)}, f^{(2)}\rangle_{\mathcal{H}(K_X)}. \quad (6.20)$$

In addition, Ψ_X has the following property

$$\mathrm{E}[\Psi_X(f)|G=j] = \langle f, \mu_j\rangle_{\mathcal{H}(K_X)}. \quad (6.21)$$

As in Section 6.1.1, observe that

$$\text{Var}_G(\text{E}[\ell|G]) = \text{Var}_G(\text{E}[\Psi_X(f)|G]) = \langle f(*), \langle K_B(*,\cdot), f(\cdot)\rangle_{\mathcal{H}(K_X)}\rangle_{\mathcal{H}(K_X)}$$

and

$$\text{Var}(\ell) = \text{Var}(\Psi_X(f)) = \|f\|^2_{\mathcal{H}(K_X)}.$$

We now define the operator C_B from $\mathcal{H}(K_X)$ to $\mathcal{H}(K_X)$ by

$$(C_B f)(t) = \langle K_B(t,\cdot), f(\cdot)\rangle_{\mathcal{H}(K_X)} \tag{6.22}$$

for $f \in \mathcal{H}(K_X)$. Consequently, we now can restate the general discrimination problem (6.18) in the RKHS $\mathcal{H}(K_X)$ as finding $f^* \in \mathcal{H}(K_X)$ such that

$$\langle f^*, C_B f^*\rangle_{\mathcal{H}(K_X)} = \sup_{f\in\mathcal{H}(K_X)} \langle f, C_B f\rangle_{\mathcal{H}(K_X)} \tag{6.23}$$

subject to

$$\|f\|^2_{\mathcal{H}(K_X)} = 1.$$

Thus, characterization of the solutions to problem (6.23) is achieved through study of the operator C_B.

THEOREM **VI.3**. *The operator C_B in (6.22) is a bounded linear operator from $\mathcal{H}(K_X)$ to $\mathcal{H}(K_X)$ with operator norm at most 1.*

Proof. We see that

$$\text{Cov}_G\left(\text{E}[\Psi_X(f^{(1)})|G], \text{E}[\Psi_X(f^{(2)})|G]\right) = \langle f^{(1)}, C_B f^{(2)}\rangle_{\mathcal{H}(K_X)}$$

for any functions $f^{(1)}$ and $f^{(2)}$ in $\mathcal{H}(K_X)$. Now observe from the Cauchy-Schwarz inequality that

$$\begin{aligned}
\left|\langle f^{(1)}, C_B f^{(2)}\rangle_{\mathcal{H}(K_X)}\right| &= \left|\text{Cov}_G\left(\text{E}[\Psi_X(f^{(1)})|G], \text{E}[\Psi_X(f^{(2)})|G]\right)\right| \\
&\leq \text{Var}_G(\text{E}[\Psi_X(f^{(1)})|G])^{1/2}\text{Var}_G(\text{E}[\Psi_X(f^{(2)})|G])^{1/2} \\
&\leq \text{Var}(\Psi_X(f^{(1)}))^{1/2}\text{Var}(\Psi_X(f^{(2)}))^{1/2} = \|f^{(1)}\|_{\mathcal{H}(K_X)}\|f^{(2)}\|_{\mathcal{H}(K_X)}.
\end{aligned}$$

Replacing $f^{(1)}$ by $C_B f^{(2)}$ entails that $\|C_B f^{(2)}\|_{\mathcal{H}(K_X)} \leq \|f^{(2)}\|_{\mathcal{H}(K_X)}$ and completes the proof.

□

We can easily see that C_B is self-adjoint, compact and positive in a similar way to the operator T_B. So, C_B has the spectral decomposition as in (6.15). Let λ_i, β_i be the eigenvalues and the corresponding eigenvectors of the operator C_B, respectively. Note that $1 \geq \lambda_1 \geq \cdots \geq \lambda_{r(C_B)} > 0$. Then, the solutions to problem (6.23) subject to $\|f\|^2_{\mathcal{H}(K_X)} = 1$ are given by $f_i = \beta_i$ with β_i the eigenvectors of the operator C_B and the corresponding linear discriminant variables of the X space are $\ell_i = \Psi_X(f_i)$.

We may be interested in the relationship between the operators $T_B \in B(\mathcal{H}(K_W))$ and $C_B \in B(\mathcal{H}(K_X))$ and the relationship between the isomorphisms Ψ_W and Ψ_X. These links are addressed by the following results.

LEMMA **VI.1**. *Let T_B and C_B be the operators defined as in (6.5) and (6.22), respectively. Also, let L be the linear transformation defined in (6.3). Then, the operators C_B is related to T_B in the following way:*

$$C_B = T_B \circ L.$$

Also, $\Psi_X(f) = \Psi_W(Lf) = \Psi_W(f) - \Psi_W(C_B f)$ for $f \in \mathcal{H}(K_X)$.

Proof. We observe from (6.4) that

$$(T_B L f)(t) = \langle K_B(\cdot, t), Lf \rangle_{\mathcal{H}(K_W)} = \langle K_B(\cdot), f \rangle_{\mathcal{H}(K_X)} = (C_B f)(t).$$

for $f \in \mathcal{H}(K_X)$. Also, the isomorphisms Ψ_X from $\mathcal{H}(K_X)$ to L^2_X and Ψ_W from $\mathcal{H}(K_W)$ to L^2_X are related in that

$$\Psi_X(f) = \Psi_W(Lf), \ f \in \mathcal{H}(K_X)$$

which follows from

$$\Psi_X(K_X(\cdot, t)) = X(t) = \Psi_W(K_W(\cdot, t)) = \Psi_W(L(K_X(\cdot, t)).$$

From $f(t) = (Lf)(t) + (T_B Lf)(t) = (Lf)(t) + (C_B f)(t)$, we then have

$$(C_B f)(t) = ((I - L)f)(t).$$

That is, $L = I - C_B$.

□

Since the linear transformation $L \in B(\mathcal{H}(K_X), \mathcal{H}(K_W))$ is one-to-one and onto, we observe from the open mapping theorem and the relationship between T_B and C_B that

$$T_B = C_B \circ L^{-1}$$

with $L^{-1} \in B(\mathcal{H}(K_W), \mathcal{H}(K_X))$ satisfying $L \circ L^{-1} = I_{\mathcal{H}(K_W)}$ and $L^{-1} \circ L = I_{\mathcal{H}(K_X)}$. So, the compactness of T_B is a consequence of the compactness of C_B and the boundedness of L^{-1}.

Now Fisher's linear discriminant function was originally obtained by solving

$$\sup_{\ell \in L_X^2} \text{Var}_G(\text{E}[\ell|G]) = \sup_{f \in \mathcal{H}(K_W)} \langle f, T_B f \rangle_{\mathcal{H}(K_W)}$$

subject to

$$\text{E}_G[\text{Var}(\ell|G)] = 1.$$

If we let $\ell_{\text{Fisher},i}$ be the solutions to problem in (6.1), we then get

$$\ell_{\text{Fisher},i} = \frac{\ell_i}{\left(\text{E}_G[\text{Var}(\ell_i|G)]\right)^{1/2}}$$

with ℓ_i the solutions to the problem (6.18). But we can observe that

$$\begin{aligned} \text{E}_G[\text{Var}(\ell_i|G)] &= \langle f_i(\cdot), \langle K_W(\cdot, *), f_i(*) \rangle_{\mathcal{H}(K_X)} \rangle_{\mathcal{H}(K_X)} \\ &= \|f_i\|^2_{\mathcal{H}(K_X)} - \langle f_i, C_B f_i \rangle_{\mathcal{H}(K_X)} = 1 - \lambda_i. \end{aligned}$$

Thus, we have

$$\ell_{\text{Fisher},i} = (1 - \lambda_i)^{-1/2} \Psi_X(f_i)$$

with f_i the eigenvectors of the operator C_B. So, the squared Mahalanobis distance $Dist_j^s$ in (6.17) becomes

$$Dist_j^s(x) = \sum_{k=1}^{s} \frac{1}{1-\lambda_k}\big(\ell_k(x) - \mathrm{E}[\ell_k|G=j]\big)^2 = \sum_{k=1}^{s} \frac{1}{1-\lambda_k}\big(\Psi_{X,x}(f_k) - \langle f_k, \mu_j\rangle_{\mathcal{H}(K_X)}\big)^2. \tag{6.24}$$

6.1.3 Bayes Procedure: Linear Discriminant Analysis

In this section, we wish to consider the classification of a Gaussian process $\{X(t), t \in \mathcal{T}\}$ under the assumption of a common within-class covariance function. This problem dates back to Parzen (1962, 1963) who developed a unified approach to the extraction of signal in noise problems based on RKHS theory.

Let us consider the stochastic model

$$X(t) = \sum_{j=1}^{J} \mu_j(t)Y(j) + e(t) \tag{6.25}$$

with $\mathrm{E}[e(t)] = 0$ and $\mathrm{Cov}(e(s), e(t)) = K_W(s,t)$. Then, our interest is in prediction of the membership of X corresponding to the population indexed by $Y(\cdot)$.

Let Ω be the space of all real-valued functions on $\mathcal{T}$. For $j = 1, \ldots, J$, let P_0 and P_j be the probability measures defined on the measurable subsets B of Ω by

$$P_0(B) = P[\{e(t), t \in \mathcal{T}\} \in B]$$

and

$$P_j(B) = P\Big[\Big\{\sum_{j=1}^{J} \mu_j(t)Y(j) + e(t), t \in \mathcal{T}\Big\} \in B \Big| G = j\Big].$$

Let $p_j(X(t), t \in \mathcal{T})$ denote the probability density of the process $\{\mu_j(t) + e(t), t \in \mathcal{T}\}$ with respect to the process $\{e(t), t \in \mathcal{T}\}$. Recalling our definitions of the RKHS $\mathcal{H}(K_W)$ and the isomorphism Ψ_W between $\mathcal{H}(K_W)$ and L_X^2, we see that if $\mu_j \in \mathcal{H}(K_W)$ and $e(t)$

is a normal process, then the probability density functional of P_j with respect to P_0 is

$$p_j(X(t), t \in \mathcal{T}) = \frac{dP_j}{dP_0} = \exp\left\{\Psi_W(\mu_j) - \frac{1}{2}\|\mu_j\|^2_{\mathcal{H}(K_W)}\right\}$$
$$= \exp\left\{\Psi_W(\mu_j) - \frac{1}{2}\Big(\mathrm{E}[\Psi_W(\mu_j)|G = j]\Big)^2\right\}$$

from (6.10).

The Bayes classifier classifies a new observation to the class which maximizes the posterior probability

$$P(G = j|X(t), t \in \mathcal{T}) = \frac{\pi_j p_j(X(t), t \in \mathcal{T})}{\sum_{k=1}^{J} \pi_k p_k(X(t), t \in \mathcal{T})}.$$

However,

$$P(G = j|X(t), t \in \mathcal{T}) \propto \exp\left\{\Psi_W(\mu_j) - \frac{1}{2}\|\mu_j\|^2_{\mathcal{H}(K_W)} + \log \pi_j\right\}$$

since $\mathrm{E}[\Psi_W(\mu_j)|G = j] = \|\mu_j\|^2_{\mathcal{H}(K_W)}$. So, we can define the discriminant function for class j to be

$$d_j(x) = \Psi_{W,x}(\mu_j) - \frac{1}{2}\|\mu_j\|^2_{\mathcal{H}(K_W)} + \log \pi_j$$

and we classify x to the class for which $d_j(x)$ is largest.

6.2 Fisher's Linear Discrimination and Bayes Procedure

Suppose $J = 2$ and $\pi_1 = \pi_2$. Assume that μ_1 and μ_2 belong to $\mathcal{H}(K_W)$. In this instance, the Bayes classification becomes

classify x to class 1 if $d_1(x) - d_2(x) = \Psi_{W,x}(\mu_1 - \mu_2) - \langle \mu_1 - \mu_2, \mu \rangle_{\mathcal{H}(K_W)} > 0$

with $\mu = (\mu_1 + \mu_2)/2$.

In contrast, Fisher's linear discriminant function is obtained by maximizing

$$\frac{\big|\langle \mu_1 - \mu_2, h\rangle_{\mathcal{H}(K_W)}\big|^2}{\|h\|^2_{\mathcal{H}(K_W)}}$$

over $h \in \mathcal{H}(K_W)$. This ratio has maximum $\|\mu_1 - \mu_2\|^2_{\mathcal{H}(K_W)}$ which is attained when $h = \mu_1 - \mu_2$. Hence, Fisher's linear discriminant function is $\ell = \Psi_W(\mu_1 - \mu_2)$. The corresponding classification rule to (6.17) is then to classifying x to class 1 if

$$\left|\Psi_{W,x}(\mu_1 - \mu_2) - \langle \mu_1 - \mu_2, \mu_1 \rangle_{\mathcal{H}(K_W)}\right| < \left|\Psi_{W,x}(\mu_1 - \mu_2) - \langle \mu_1 - \mu_2, \mu_2 \rangle_{\mathcal{H}(K_W)}\right|,$$

which provides exactly the same rule as in the Bayes procedure.

6.3 Fisher's Linear Discrimination and Canonical Correlation Analysis

It was seen that Fisher's linear discriminant functions can be derived from canonical correlation analysis in the finite dimensional case in Section 3.2.6. Our goal is now to generalize that result to the infinite dimensional setting.

Let $\{Y(j), j = 1, \ldots, J\}$ be a family of indicator variables for a collection of mutually exclusive and exhaustive populations numbered 1 to J. We define $\pi_j = P(G = j) = P(Y(j) = 1)$. Then auto and cross covariance functions for the X and Y processes are given by

$$K_X(s,t) = \text{Cov}(X(s), X(t)),\ K_Y(i,j) = \text{Cov}(Y(i), Y(j))$$

and

$$K_{XY}(s,j) = \text{Cov}(X(s), Y(j))$$

for $s, t \in \mathcal{T}$ and $i, j \in \{1, \ldots, J\}$ and recall that

$$\mathbf{K}_Y = \{K_Y(i,j)\}_{i,j=1}^J = \text{diag}(\pi_1, \ldots, \pi_J) - \boldsymbol{\pi}\boldsymbol{\pi}^T.$$

Now let $\mathcal{H}(K_X), \mathcal{H}(K_Y)$ be the RKHS's with r.k.'s K_X, K_Y, respectively. In particular, $\mathcal{H}(K_Y)$ is the linear manifold of functions on $\{1, \ldots, J\}$ of the form

$$\sum_{j=1}^J b_j K_Y(\cdot, j)$$

for $\mathbf{b} = (b_1, \ldots, b_J)^T \in \text{Ker}(\mathbf{K}_Y)^\perp$. The associated inner product is

$$\langle g^{(1)}, g^{(2)} \rangle_{\mathcal{H}(K_Y)} = \mathbf{b}_1^T \mathbf{K}_Y \mathbf{b}_2 \tag{6.26}$$

for $g^{(1)}, g^{(2)} \in \mathcal{H}(K_Y)$ and $\mathbf{b}_1, \mathbf{b}_2 \in \text{Ker}(\mathbf{K}_Y)^\perp$. Since $\boldsymbol{\pi}^T \mathbf{b}_1 = \boldsymbol{\pi}^T \mathbf{b}_2 = 0$ as in Section 3.2.6, (6.26) becomes

$$\langle g^{(1)}, g^{(2)} \rangle_{\mathcal{H}(K_Y)} = \mathbf{b}_1^T \text{diag}(\pi_1, \ldots, \pi_J) \mathbf{b}_2.$$

Let $\mathbf{g} = (g(1), \ldots, g(J))^T = \mathbf{K}_Y \mathbf{b}$. Then $\mathbf{1} \in \text{Ker}(\mathbf{K}_Y)$ and premultiplying by $\mathbf{1}^T$ produces

$$\sum_{j=1}^{J} g(j) = 0.$$

Further, from $\mathbf{g} = \mathbf{K}_Y \mathbf{b} = \text{diag}(\pi_1, \ldots, \pi_J)\mathbf{b}$ it follows that

$$\mathbf{b} = \text{diag}(\pi_1^{-1}, \ldots, \pi_J^{-1})\mathbf{g} = \left(\frac{g(1)}{\pi_1}, \ldots, \frac{g(J)}{\pi_J} \right).$$

Thus, the associated inner product in $\mathcal{H}(K_Y)$ is

$$\langle g^{(1)}, g^{(2)} \rangle_{\mathcal{H}(K_Y)} = \sum_{j=1}^{J} \frac{g^{(1)}(j) g^{(2)}(j)}{\pi_j}. \tag{6.27}$$

As explained in Section 5.1, the canonical variables of the X space and Y space in this setting are

$$\eta_i = \Psi_X(f_i) \quad \text{and} \quad \mathbf{b}_i^T \mathbf{Y} = \Psi_Y(g_i) = \sum_{j=1}^{J} \frac{g_i(j) Y(j)}{\pi_j},$$

where $\mathbf{Y} = (Y(1), \ldots, Y(J))^T$ and f_i, g_i are the singular functions of the operator T given by

$$(Tg)(t) = \langle K_{XY}(t, \cdot), g \rangle_{\mathcal{H}(K_Y)} = \sum_{j=1}^{J} \frac{K_{XY}(t, j) g(j)}{\pi_j}$$

for $t \in \mathcal{T}$ and $g \in \mathcal{H}(K_Y)$ and f_i, g_i satisfying

$$\|f_i\|^2_{\mathcal{H}(K_X)} = 1 \quad \text{and} \quad \langle f_i, f_l \rangle_{\mathcal{H}(K_X)} = 0, \tag{6.28}$$

and

$$\sum_{j=1}^{J} g_i(j) = 0, \quad \sum_{j=1}^{J} \frac{g_i^2(j)}{\pi_j} = 1 \quad \text{and} \quad \sum_{j=1}^{J} \frac{g_i(j)g_l(j)}{\pi_j} = 0 \tag{6.29}$$

for $i \neq l$ and $i, l = 1, \ldots, J$. Note that the operator T from $\mathcal{H}(K_Y)$ to $\mathcal{H}(K_X)$ is clearly bounded and compact since $\dim(\mathcal{H}(K_Y))$ is finite.

We now provide a general result that links Fisher's discriminant functions and canonical correlation analysis.

THEOREM **VI.4**. *For $i = 1, \ldots, J-1$, the canonical variables of the X space, η_i, are identical to the linear discriminant functions, ℓ_i apart from scaling factors and the canonical correlations ρ_i are precisely square roots of the eigenvalues obtained from the spectral decomposition of the operator C_B.*

Proof. Let us first observe that

$$\Psi_X(f) = \Psi_W(f) - \Psi_W(C_B f), \ f \in \mathcal{H}(K_X) \tag{6.30}$$

and

$$\begin{aligned}\langle Lf^{(1)}, Lf^{(2)} \rangle_{\mathcal{H}(K_W)} &= \langle Lf^{(1)}, f^{(2)} \rangle_{\mathcal{H}(K_X)} = \langle (I - C_B)f^{(1)}, f^{(2)} \rangle_{\mathcal{H}(K_X)} \\ &= \langle f^{(1)}, f^{(2)} \rangle_{\mathcal{H}(K_X)} - \langle C_B f^{(1)}, f^{(2)} \rangle_{\mathcal{H}(K_X)}\end{aligned}$$

for $f^{(1)}, f^{(2)}$ in $\mathcal{H}(K_X)$. The canonical variables for X are then given by $\eta_i = \Psi_X(f_i), i = 1, \ldots, J-1$, where $f_i \in \mathcal{H}(K_X)$ are obtained from

$$TT^* f_i = \rho_i^2 f_i \tag{6.31}$$

and Fisher's discriminant functions are given by $\ell_i = \Psi_W(h_i), i = 1, \ldots, J-1$, where $h_i \in \mathcal{H}(K_W)$ are obtained from

$$T_B h_i = \gamma_i h_i. \tag{6.32}$$

We can see that, for $f \in \mathcal{H}(K_X)$,

$$\begin{aligned}(TT^* f)(t) &= \langle K_{XY}(t, \cdot), (T^* f)(\cdot) \rangle_{\mathcal{H}(K_Y)} \\ &= \langle TK_{XY}(t, \cdot), f(\cdot) \rangle_{\mathcal{H}(K_X)}.\end{aligned}$$

However,

$$TK_{XY}(t,\cdot) = \langle K_{XY}(\cdot,*), K_{XY}(t,*)\rangle_{\mathcal{H}(K_Y)} = \sum_{j=1}^{J} \frac{K_{XY}(\cdot,j)K_{XY}(t,j)}{\pi_j}$$

and, for $i = 1, \ldots, J$,

$$K_{XY}(\cdot,i) = \text{Cov}(X(\cdot),Y(i)) = \text{E}[X(\cdot)Y(i)] - \text{E}[X(\cdot)]\text{E}[Y(i)] = \pi_i(\mu_i(\cdot) - \mu(\cdot)) \quad (6.33)$$

since

$$\text{E}[X(\cdot)Y(i)] = \text{E}_G[\text{E}(X(\cdot)Y(i)|G)] = \sum_{j=1}^{J} \pi_j \text{E}[X(\cdot)|G=j]\delta_{ij} = \pi_i\mu_i(\cdot)$$

with $\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ otherwise, $\text{E}[X(\cdot)] = \mu(\cdot)$, and $\text{E}[Y(i)] = \pi_i$. So,

$$TK_{XY}(t,\cdot) = \sum_{j=1}^{J} \pi_j(\mu_j(\cdot) - \mu(\cdot))(\mu_j(t) - \mu(t)) = K_B(t,\cdot),$$

$$(TT^*f)(t) = \langle TK_{XY}(t,\cdot), f(\cdot)\rangle_{\mathcal{H}(K_X)} = \langle K_B(t,\cdot), f(\cdot)\rangle_{\mathcal{H}(K_X)} \quad (6.34)$$

and (6.34) becomes

$$(TT^*f)(t) = (C_Bf)(t).$$

Now use the fact that $C_Bf = T_BLf$ and $f = Lf + T_BLf$ for $f \in \mathcal{H}(K_X)$ to rewrite (6.31) as

$$(T_BLf_i)(t) = \rho_i^2[(Lf_i)(t) + (T_BLf_i)(t)];$$

i.e.,

$$(T_BLf_i)(t) = \frac{\rho_i^2}{1-\rho_i^2}(Lf_i)(t).$$

Since the f_i satisfy $\|f_i\|^2_{\mathcal{H}(K_X)} = 1$,

$$\|Lf_i\|^2_{\mathcal{H}(K_W)} = \|f_i\|^2_{\mathcal{H}(K_X)} - \langle C_Bf_i, f_i\rangle_{\mathcal{H}(K_X)} = 1 - \rho_i^2.$$

Moreover, the h_i in $\mathcal{H}(K_W)$ corresponding to Fisher's discriminant functions $\ell_i = \Psi_W(h_i)$ satisfy $\|h_i\|^2_{\mathcal{H}(K_W)} = 1$ and $Lf_i = (1-\rho_i^2)f_i$ from $f_i = Lf_i + C_Bf_i = Lf_i + \rho_i^2 f_i$. Thus,

the h_i are related to the f_i via

$$h_i = \frac{Lf_i}{\|Lf_i\|_{\mathcal{H}(K_W)}} = \frac{Lf_i}{(1-\rho_i^2)^{1/2}} = (1-\rho_i^2)^{1/2} f_i. \tag{6.35}$$

Also, from the relationship between the isomorphisms Ψ_X from $\mathcal{H}(K_X)$ to L_X^2 and Ψ_W from $\mathcal{H}(K_W)$ to L_X^2

$$\eta_i = \Psi_X(f_i) = \Psi_W(Lf_i) = (1-\rho_i^2)\Psi_W(f_i) = (1-\rho_i^2)^{1/2}\Psi_W(h_i) = (1-\rho_i^2)^{1/2}\ell_i.$$

Therefore, Fisher's discriminant functions ℓ_i are related to the canonical X variables η_i in

$$\ell_i = \frac{\eta_i}{(1-\rho_i^2)^{1/2}}. \tag{6.36}$$

This result is the exact parallels of what transpires in the finite dimensional setting.

□

Note that the eigenvalues of T_B and TT^* are related as

$$\gamma_i = \frac{\rho_i^2}{1-\rho_i^2}. \tag{6.37}$$

Also, the canonical X variables and the generalized Fisher's discriminant functions in Section 6.1.2 are identical since $TT^* f = C_B f$.

Now we wish to interpret the canonical variables of the Y space from canonical correlation analysis in Chapter IV. The canonical variables of the Y space are obtained from

$$T^*Tg = \lambda g.$$

Then we have

$$(T^*Tg)(l) = \langle K_{XY}(\cdot, l), (Tg)(\cdot)\rangle_{\mathcal{H}(K_X)}.$$

Now observe that

$$(Tg)(\cdot) = \sum_{j=1}^{J} \frac{K_{XY}(\cdot, j)g(j)}{\pi_j} = \sum_{j=1}^{J} (\mu_j(\cdot) - \mu(\cdot))g(j) = \sum_{j=1}^{J} g(j)\mu_j(\cdot)$$

because $\sum_{j=1}^{J} g(j) = 0$. So, the operator T from $\mathcal{H}(K_Y)$ to $\mathcal{H}(K_X)$ provides a contrast among the population mean functions. Hence

$$(T^*Tg)(l) = \pi_l \sum_{j=1}^{J} g(j)\langle \mu_l(\cdot) - \mu(\cdot), \mu_j(\cdot) - \mu(\cdot)\rangle_{\mathcal{H}(K_X)}.$$

Also,

$$\langle f, Tg\rangle_{\mathcal{H}(K_X)} = \sum_{j=1}^{J} g(j)\langle f, \mu_j\rangle_{H(K_X)},$$

which is the contrast among the transformed mean functions $m_j = \langle f, \mu_j\rangle_{\mathcal{H}(K_X)}$.

Let $\Psi_X(f_1)$ and $\Psi_Y(g_1)$ be the first canonical variables of the X and Y processes. Then, f_1 and g_1 are obtained by maximizing

$$\Big|\sum_{j=1}^{J} g(j)\langle f, \mu_j\rangle_{H(K_X)}\Big|,$$

subject to $\|f\|_{\mathcal{H}(K_X)} = 1, \sum_{j=1}^{J} g(j) = 0$ and $\sum_{j=1}^{J} \frac{g^2(j)}{\pi_j} = 1$. Thus, we have exactly the same interpretation as in the finite dimensions. The functions g provide the coefficient of a contrast in transformed means and so it measures the importance of the transformed means $m_j = \langle f, \mu_j\rangle_{\mathcal{H}(K_X)}$ in the contrast. Also, it plays an important role in classification analogous to the finite dimensions.

From (5.14), we have

$$f_i \propto Tg_i$$

and we have seen that $Tg_i, i = 1, \ldots, J-1$ are the orthogonal contrasts among class means. Thus, $Tg_i, i = 1, \ldots, J-1$ are exactly the same as the RKHS vectors f_i apart from a constant of proportionality.

6.4 Classification

A goal of discriminant analysis is in construction of classification rule. In this section, the classification rule based on the canonical variables of the X and Y processes will be formulated as in Section 3.2.7.

Let $\eta = \Psi_X(f)$ and $\xi = \mathbf{b}^T\mathbf{Y} = \sum_{j=1}^{J} \frac{g(j)Y(j)}{\pi_j}$ be a pair of canonical variables for the X and Y processes corresponding to the canonical correlation ρ. Since η is the best linear predictors of ξ, ξ can be predicted from η. The predicted score is given by

$$\mathrm{E}[\xi] + \frac{\mathrm{Cov}(\xi,\eta)}{\mathrm{Var}(\eta)}(\eta - \mathrm{E}[\eta]) = \rho(\eta - \mathrm{E}[\eta])$$

because $\mathrm{E}[\xi] = \sum_{j=1}^{J} g(j) = 0$ and $\mathrm{Var}(\eta) = 1$.

Now we provide the classification rule in the subspace defined by the predicted scores of the first s ($\leq J-1$) canonical variables of the X space. Let $\tilde{\xi}(\omega_0) := \tilde{\xi}_i(x) = \rho_i(\eta_i(x) - \mathrm{E}[\eta_i]), i = 1, \ldots, J-1$, with $x(\cdot) = X(\cdot, \omega_0)$. Then, the squared Mahalanobis distance is

$$\sum_{k=1}^{s} \frac{1}{\rho_k^2(1-\rho_k^2)}(\tilde{\xi}_k(x) - \bar{\tilde{\xi}}_{kj})^2 \tag{6.38}$$

with $\bar{\tilde{\xi}}_{kj} = \mathrm{E}[\tilde{\xi}_k | G = j] = \rho_k \langle f_k, \mu_j - \mu \rangle_{\mathcal{H}(K_X)}$. We can easily see from this that the distances in (6.17), (6.24) and (6.38) are the same. However, these distances are expressed in terms of either $\langle f_k, \mu_j \rangle_{\mathcal{H}(K_W)}$ or $\langle f_k, \mu_j \rangle_{\mathcal{H}(K_X)}$ which pose practical problems for estimation from data.

Our goal is now to find new classification rule which is free of inner products and is equivalent to the distances (6.17), (6.24) and (6.38). Now our goal is to find the equivalent classification rule to the distances (6.17), (6.24) and (6.38) through CCA. As in Section 3.2.7, we can introduce distance measures constructed from the CCA approach as follows: for a sample path x,

$$\sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\tilde{\xi}_k(x) - b_{kj})^2 - \sum_{k=1}^{s} b_{kj}^2 \tag{6.39}$$

and

$$\sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\Psi_{X,x}(f_k) - \tilde{\eta}_{kj})^2 \tag{6.40}$$

with $\tilde{\eta}_{kj} = \langle f_k, \mu \rangle_{\mathcal{H}(K_X)} + \rho_k b_{kj}$ the predicted score of η_k via ξ_k for the class j. The proposed classification rule is to classifying a sample path x to the class whose index minimizes (6.40).

THEOREM **VI.5**. *The distances in (6.17), (6.24),(6.38) and (6.40) are the same.*

Proof. We can easily see that (6.17) and (6.24) are identical from the fact that

$$\Psi_W(h_i) = (1-\rho_i^2)^{-1/2}\Psi_X(f_i)$$

and

$$\langle f_i, \mu_j\rangle_{\mathcal{H}(K_X)} = \langle Lf_i, \mu_j\rangle_{\mathcal{H}(K_W)} = (1-\rho_i^2)^{1/2}\langle h_i, \mu_j\rangle_{\mathcal{H}(K_W)},$$

where h_i are the eigenvectors of T_B and f_i are the eigenvectors of TT^* associated with its eigenvalues ρ_i^2.

We start with (6.24). We see from Theorem 14 that (6.24) becomes

$$Dist_j^s(x) = \sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\eta_k(x) - \langle f_k, \mu_j\rangle_{\mathcal{H}(K_X)})^2.$$

Then, observe that

$$Dist_j^s(x) = \sum_{k=1}^{s} \frac{1}{1-\rho_k^2}\left(\eta_k(x) - \mathrm{E}[\eta_k] + \mathrm{E}[\eta_k] - \langle f_k, \mu_j\rangle_{\mathcal{H}(K_X)}\right)^2.$$

From (5.15) and (6.33), we have

$$(T^* f_k)(j) = \langle f_k, K_{XY}(\cdot, j)\rangle_{\mathcal{H}(K_X)} = \pi_j\langle f_k, \mu_j - \mu\rangle_{\mathcal{H}(K_X)} = \rho_k g_k(j).$$

Hence we have $\langle f_k, \mu_j - \mu\rangle_{\mathcal{H}(K_X)} = \rho_k\pi_j^{-1}g_k(j) = \rho_k b_{kj}$ for $j = 1, \ldots, J$. So,

$$\langle f_k, \mu_j\rangle_{\mathcal{H}(K_X)} - \mathrm{E}[\eta_k] = \langle f_k, \mu_j - \mu\rangle_{\mathcal{H}(K_X)}.$$

Thus, the result follows.

□

COROLLARY **VI.2**. *The distance measure in (6.39) is equivalent to the distances in (6.17), (6.24),(6.38) and (6.40) in the sense of classification.*

Proof. We begin with (6.24) since (6.24) and (6.38) are identical. Then, observe that

$$\begin{aligned} Dist_j^s(x) = &\sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\eta_k(X) - \mathrm{E}[\eta_k])^2 + \sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\langle f_k, \mu_j\rangle_{\mathcal{H}(K_X)} - \mathrm{E}[\eta_k])^2 \\ &- 2\sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\eta_k(x) - \mathrm{E}[\eta_k])(\langle f_k, \mu_j\rangle_{\mathcal{H}(K_X)} - \mathrm{E}[\eta_k]). \end{aligned}$$

We first see that

$$\eta_k(x) - \mathrm{E}[\eta_k] = \rho_k^{-1}\tilde{\xi}_k(x).$$

Also, we have seen that

$$\langle f_k, \mu_j \rangle_{\mathcal{H}(K_X)} - \mathrm{E}[\eta_k] = \rho_k b_{kj}$$

for $j = 1, \ldots, J$. Thus, $Dist_j^s$ can be simplified to

$$\begin{aligned}
&\sum_{k=1}^{s} \frac{1}{\rho_k^2(1-\rho_k^2)}\tilde{\xi}_k(x)^2 + \sum_{k=1}^{s} \frac{\rho_k^2}{1-\rho_k^2} b_{kj}^2 - 2\sum_{k=1}^{s} \frac{1}{1-\rho_k^2}\tilde{\xi}_k(x) b_{kj} \\
&= \sum_{k=1}^{s} \rho_k^{-2}\tilde{\xi}_k(x)^2 + \sum_{k=1}^{s} \frac{1}{1-\rho_k^2}(\tilde{\xi}_k(x) - b_{kj})^2 - \sum_{k=1}^{s} b_{kj}^2
\end{aligned}$$

and the desired result is obtained.

□

6.5 Computation

Let $X_1, \ldots, X_N$ be iid copies of a random continuous curve X. Let X_{ij} be the ith curve randomly drawn from the jth class. Also let μ_j be the true mean curve of an individual from the jth class and e_{ij} be the random noise processes with mean zero and covariance kernel K_W. We will focus on the case of $\mathcal{T} = [0, 1]$ and smooth covariance function K_X of the X process. Then,

$$X_{ij}(t) = \mu_j(t) + e_{ij}(t),\ i = 1, \ldots, N_j,\ j = 1, \ldots, J,\ t \in [0, 1].$$

In practice X_{ij} is observed at a discrete set of finitely many points $t_1, \ldots, t_m$. Let X_{ijk} be the value for the ith curve at t_k from the jth class. Observe that

$$X_{ijk} = X_{ij}(t_k) + \upsilon_{ijk},\ k = 1, \ldots, m,$$

where υ_{ijk}'s are the uncorrelated measurement errors with zero mean and constant variance σ^2. We now have

$$X_{ijk} = \mu_j(t_k) + \epsilon_{ijk},\ k = 1, \ldots, m,$$

where $\epsilon_{ijk} = e_{ij}(t_k) + \upsilon_{ijk}$ satisfy

$$\begin{aligned}\text{Cov}(\epsilon_{ijk}, \epsilon_{i'jk'}) &= K_W(t_k, t_k) + \sigma^2, \quad i = i' \text{ and } k = k' \\ &= K_W(t_k, t_{k'}), \qquad i = i' \text{ and } k \neq k' \\ &= 0, \qquad i \neq i'.\end{aligned}$$

Defining $\bar{X}_{jk} = \frac{1}{N_j}\sum_{i=1}^{N_j} X_{ijk}$, we have

$$\bar{X}_{jk} = \mu_j(t_k) + \bar{\epsilon}_{jk}, \tag{6.41}$$

where

$$\bar{\epsilon}_{jk} = \frac{1}{N_j}\sum_{i=1}^{N_j} \epsilon_{ijk}$$

and

$$\begin{aligned}\text{Cov}(\bar{\epsilon}_{jk}, \bar{\epsilon}_{jk'}) &= \frac{1}{N_j}\left\{K_W(t_k, t_k) + \sigma^2\right\}, \quad k = k' \\ &= \frac{1}{N_j}K_W(t_k, t_{k'}), \qquad k \neq k'.\end{aligned}$$

We first propose to estimate the between-class covariance kernel $K_B(\cdot, \cdot)$ defined in (6.2). For this purpose, we will estimate μ_j and μ. One natural approach is to use nonparametric function estimation. Then, in general, the estimate of μ_j has the following form

$$\hat{\mu}_j(t) = \sum_{k=1}^{m} w_k(t, \lambda)\bar{X}_{jk},$$

where $w_k(t, \lambda)$ is a weight function at t depending on a smoothing parameter, λ. Now let us assume that the μ_j's are smooth and use a smoothing spline to estimate μ_j and μ. (e.g., see Eubank, 1999). Specifically, cubic spline smoothing will be used in where we estimate μ_j by the minimizer $\hat{\mu}_j$ of

$$\frac{1}{mN_j}(\bar{\mathbf{X}}_{jk} - \boldsymbol{\mu}_j)^T\mathbf{W}^{-1}(\bar{\mathbf{X}}_{jk} - \boldsymbol{\mu}_j) + \lambda\int_0^1 \left\{\mu_j^{(2)}(t)\right\}^2 dt,$$

where $\bar{\mathbf{X}}_{jk} = (\bar{X}_{j1}, \ldots, \bar{X}_{jm})^T$, $\boldsymbol{\mu}_j = (\mu_j(t_1), \ldots, \mu_j(t_m))^T$ and $\mathbf{W} = \mathbf{K}_W + \sigma^2\mathbf{I}$ with $\mathbf{K}_W = \{K_W(t_k, t_{k'})\}_{k,k'=1}^m$. Then, μ can be estimated by $\hat{\mu}(t) = \sum_{k=1}^m p_j\hat{\mu}_j(t)$ with $p_j =$

$\frac{N_j}{N}$. Combining these estimators produces

$$\widehat{K}_B(s,t) = \sum_{j=1}^{J} p_j(\hat{\mu}_j(s) - \hat{\mu}(s))(\hat{\mu}_j(t) - \hat{\mu}(t)),\ s,t \in \mathcal{T}$$

and

$$\widehat{\mathbf{K}}_B = \left\{\widehat{K}_B(t_k, t_{k'})\right\}_{k,k'=1}^{m},$$

where $\{t_1, \ldots, t_m\}$ is a finite dimensional subset of $\mathcal{T}$.

Now to estimate $K_W(\cdot,\cdot)$ let $\widetilde{\mathbf{K}}_W(m) = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_j)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_j)^T$, where $\mathbf{X}_{ij} = (X_{ij}(t_1), \ldots, X_{ij}(t_m))^T$ and $\bar{\mathbf{X}}_j = \frac{1}{N_j}\sum_{i=1}^{N_j}\mathbf{X}_{ij}$. We now adopt the approach discussed in Silverman (1996). Compute the eigenvalues and eigenvectors of the generalized eigen equation

$$\widetilde{\mathbf{K}}_W(m)\mathbf{e} = \lambda(I + \vartheta\mathbf{\Omega})\mathbf{e}, \tag{6.42}$$

where ϑ is a smoothing parameter and $\mathbf{\Omega}$ is such that $\mathbf{e}^T\mathbf{\Omega}\mathbf{e} = \int (e'')^2$ for the cubic smoothing spline.

Let M_ϑ be the number of the nonzero eigenvalues of the eigen equation

$$\widetilde{\mathbf{K}}_W(m)^{[-i]}\mathbf{e} = \lambda(I + \vartheta\mathbf{\Omega})\mathbf{e},$$

where $\widetilde{\mathbf{K}}_W(m)^{[-i]}$ is the sample pooled covariance matrix computed with the ith observation $\mathbf{X}_i = (X_i(t_1), \ldots, X_i(t_m))^T$ left out. Also, let $\mathbf{e}_l^{[-i]}(\vartheta),\ l = 1, \ldots, M_\vartheta$ be the eigenvectors corresponding to the nonzero eigenvalues of the above eigen equation. For $l = 1, \ldots, M_\vartheta$, let $\mathbf{\Pi}_l^{[-i]}(\vartheta)$ be the projection onto the linear space spanned by $\mathbf{e}_1^{[-i]}(\vartheta), \ldots, \mathbf{e}_{M_\vartheta}^{[-i]}(\vartheta)$. Then, the smoothing parameter ϑ is chosen by minimizing

$$CV(\vartheta) = \sum_{l=1}^{M_\vartheta}\sum_{i=1}^{N}\left\|(\mathbf{I} - \mathbf{\Pi}_l^{[-i]}(\vartheta))\mathbf{X}_i\right\|_{\mathbb{R}^2}^2.$$

From the linear system (6.42) retain $q \leq m$ smoothed principal component vectors for use in subsequent analysis. If λ_i, e_i denote the resulting eigenvalues and smoothed principal components we then estimate $K_W(s,t)$ on $[0,1]\times[0,1]$ by $\widehat{K}_W(s,t) = \sum_{i=1}^{q}\lambda_i e_i(s)e_i(t)$.

Define $\widehat{\mathbf{K}}_W = \{\widehat{K}_W(t_k, t_{k'})\}_{k,k'=1}^m$ and perform an eigenvalue decomposition on

$$\widehat{\mathbf{K}}_W^{-1/2}\widehat{\mathbf{K}}_B\widehat{\mathbf{K}}_W^{-1/2}$$

to obtain eigenvalues $\hat{\gamma}_i$ with the associated eigenvectors $\hat{\mathbf{u}}_i$. Let $\hat{\boldsymbol{l}}_i = \widehat{\mathbf{K}}_W^{-1/2}\hat{\mathbf{u}}_i$. Then, the squared correlations are

$$\hat{\rho}_i^2 = \frac{\hat{\gamma}_i}{1+\hat{\gamma}_i}$$

and the canonical vectors $\hat{\mathbf{a}}_i$ are

$$\hat{\mathbf{a}}_i = (1-\hat{\rho}_i^2)^{1/2}\hat{\boldsymbol{l}}_i$$

which produce the estimated RKHS function

$$\hat{f}_i = (1-\hat{\rho}_i^2)^{-1/2}\hat{h}_i$$

with $\hat{h}_i(\cdot) = \sum_{k=1}^m \hat{l}_{ik}\widehat{K}_W(\cdot, t_k)$ and corresponding estimated canonical variate

$$\hat{\eta}_i = \sum_{k=1}^m \hat{a}_{ik}X(t_k).$$

Now compute

$$\hat{\mathbf{b}}_i = \hat{\rho}_i^{-1}[\hat{\boldsymbol{\mu}}_1 - \hat{\bar{\boldsymbol{\mu}}}, \ldots, \hat{\boldsymbol{\mu}}_J - \hat{\bar{\boldsymbol{\mu}}}]^T\hat{\mathbf{a}}_i$$

with $\hat{\boldsymbol{\mu}}_j = \{\hat{\mu}_j(t_k)\}_{k=1}^m$ and $\hat{\bar{\boldsymbol{\mu}}} = \{\hat{\bar{\mu}}(t_k)\}_{k=1}^m$. Then, we have

$$\hat{\eta}_{ir} = \sum_{k=1}^m \hat{a}_{ik}X_r(t_k) \quad \text{and} \quad \hat{\xi}_{ir} = \sum_{k=1}^J \hat{b}_{ik}Y_r(k)$$

for $r = 1, \ldots, N$. So, for any fixed i our transformed data is

$$(\hat{\eta}_{ir}, \hat{\xi}_{ir}),\ r = 1, \ldots, N.$$

Now, regress the $\hat{\xi}_{ir}$'s on $\hat{\eta}_{ir}$'s to get the predicted canonical X scores

$$\hat{\hat{\eta}}_i = b_{i0} + b_{1i}\hat{\xi}_i$$

with

$$b_{1i} = \frac{\sum_{r=1}^{N}(\hat{\xi}_{ir} - \bar{\hat{\xi}}_i)(\hat{\eta}_{ir} - \bar{\hat{\eta}}_i)}{\sum_{r=1}^{N}(\hat{\xi}_{ir} - \bar{\hat{\xi}}_i)^2},$$

$$b_{0i} = \bar{\hat{\eta}}_i - b_{1i}\bar{\hat{\xi}}_i$$

and $\bar{\hat{\eta}}_i = \frac{1}{N}\sum_{r=1}^{N}\hat{\eta}_{ir}, \bar{\hat{\xi}}_i = \frac{1}{N}\sum_{r=1}^{N}\hat{\xi}_{ir}$. Thus, given a sample path x, we assign x to the class whose index minimizes

$$\sum_{i=1}^{s}\frac{1}{1-\hat{\rho}_i^2}(\hat{\eta}_i(x) - \hat{\bar{\eta}}_{ij})^2$$

with $\hat{\eta}_i(x) = \sum_{k=1}^{m}\hat{a}_{ik}x(t_k)$.

EXAMPLE 1. To illustrate the use of our estimation method, take $\mathcal{T} = [0, 1]$ and consider the case where

$$\begin{pmatrix} Y(1) \\ Y(2) \end{pmatrix} \sim Multinomial(1; \pi_1, \pi_2)$$

with $\pi_1 = \pi_2 = .5$. Let

$$X(t) = \mu_1(t)Y(1) + \mu_2(t)Y(2) + \sum_{i=1}^{20} i^{-1/2}U_i\sqrt{2}\cos(i\pi t) \text{ for } t \in \mathcal{T}, \tag{6.43}$$

and the U_i being i.i.d. standard normal random variables and

$$\mu_1(t) = 3\sqrt{2}\cos(\pi t) + \sqrt{2}\cos(2\pi t),$$

$$\mu_2(t) = \sqrt{2}\cos(2\pi t).$$

A typical data set consisting of 50 sample paths of process (6.43) is shown in Figure 2 and the true mean functions of two different classes is shown in Figure 3. In this instance,

$$K_B(s,t) = \pi_1\pi_2(\mu_1(s) - \mu_2(s))(\mu_1(t) - \mu_2(t)) = \frac{18}{4}\cos(\pi s)\cos(\pi t),$$

$$K_W(s,t) = \sum_{i=1}^{20}\frac{2}{i}\cos(i\pi s)\cos(i\pi t).$$

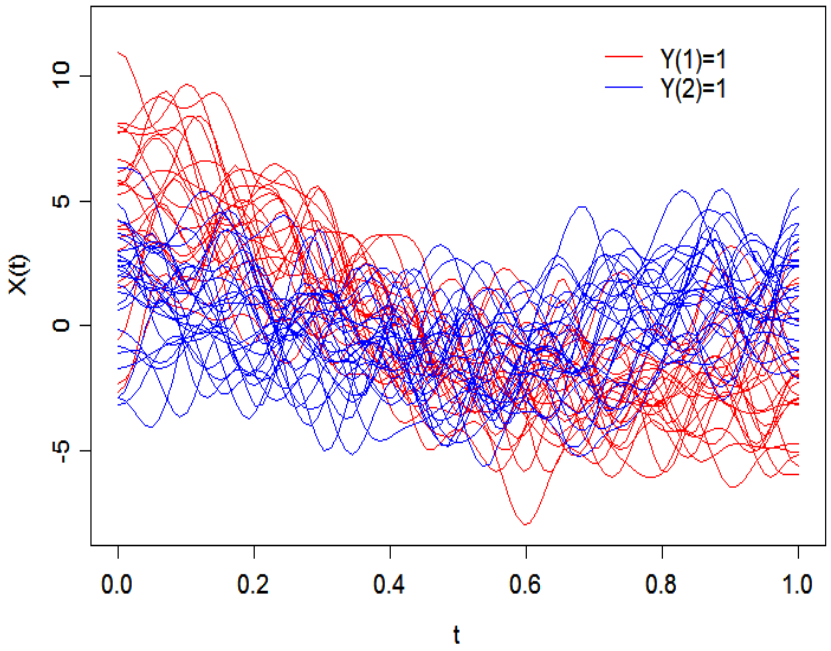

Figure 2: Sample paths of 50 curves from 2 different classes: 23 for class 1 and 27 for class 2. The red curves are from class 1 and the blue curves are from class 2.

We see that μ_1 and μ_2 belong to $\mathcal{H}(K_W)$. The integral representation Theorem then has the consequence that $\mathcal{H}(K_W)$ consists of functions of the form

$$h(t) = \sum_{i=1}^{20} \nu_i \kappa_i \sqrt{2} \cos(i\pi t)$$

for real coefficients $\kappa_i = \nu_i^{-1} \langle h(\cdot), \sqrt{2}\cos(i\pi \cdot)\rangle_{L^2[0,1]}$ and $\nu_i = i^{-1},\ i = 1, \ldots, 20$. The associated inner product is

$$\langle h_1, h_2 \rangle_{\mathcal{H}(K_W)} = \sum_{i=1}^{20} \nu_i \kappa_{1i} \kappa_{2i} = \sum_{i=1}^{20} \nu_i^{-1} \langle h_1(\cdot), \sqrt{2}\cos(i\pi \cdot)\rangle_{L^2[0,1]} \langle h_2(\cdot), \sqrt{2}\cos(i\pi \cdot)\rangle_{L^2[0,1]}.$$

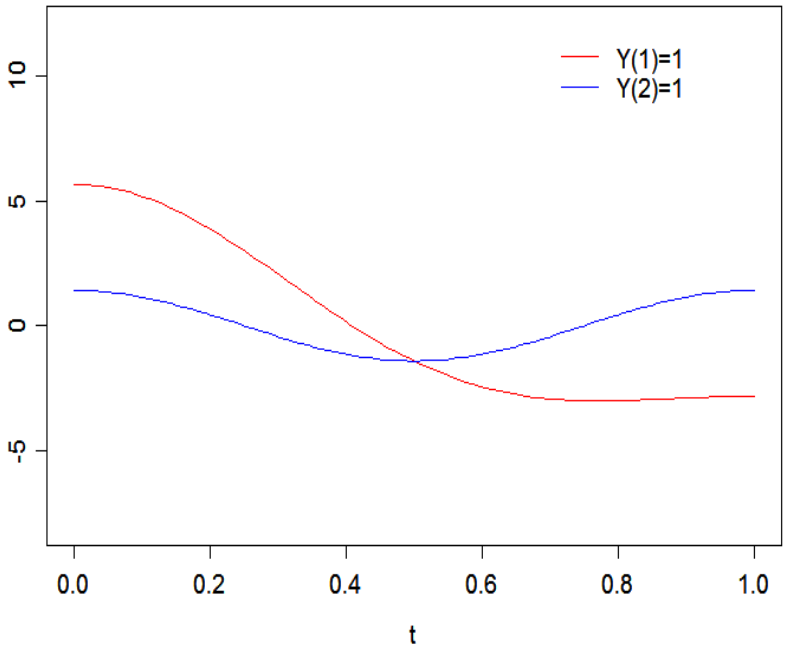

Figure 3: True mean functions μ_1 and μ_2.

Direct calculations then lead us to

$$\begin{aligned}(T_B h)(t) &= \frac{9}{4}\sqrt{2}\cos(\pi t)\langle\sqrt{2}\cos(\pi\cdot), h(\cdot)\rangle_{\mathcal{H}(K_W)} \\ &= \frac{9}{4}\sqrt{2}\cos(\pi t) \\ &\quad \times \sum_{i=1}^{20} \nu_i^{-1}\langle\sqrt{2}\cos(\pi\cdot), \sqrt{2}\cos(i\pi\cdot)\rangle_{L^2[0,1]}\langle\sum_{k=1}^{20}\nu_k\kappa_k\sqrt{2}\cos(k\pi\cdot), \sqrt{2}\cos(i\pi\cdot)\rangle_{L^2[0,1]} \\ &= \frac{9}{4}\kappa_1\sqrt{2}\cos(\pi t).\end{aligned}$$

Thus, $(T_B h)(t) = \gamma h(t)$ entails that there is only one nonzero eigenvalue $\gamma_1 = 9/4$.

Now observe that $h_1(t) = \sqrt{2}\cos(\pi t)/\kappa_1$ which follows from $K_B(s,t) = (T_B h_1)(s)h_1(t)$.

Moreover,

$$\|h_1\|^2_{\mathcal{H}(K_W)} = \kappa_1^{-2}\sum_{i=1}^{20}\nu_i^{-1}\langle\sqrt{2}\cos(\pi\cdot), \sqrt{2}\cos(i\pi\cdot)\rangle^2_{L^2[0,1]} = \kappa_1^{-2}\nu_1^{-1} = \kappa_1^{-2}.$$

So, the corresponding eigenfunction is

$$h_1(t) = \sqrt{2}\cos(\pi t).$$

For $e(t) = X(t) - \mu_1(t)Y(1) - \mu_2(t)Y(2) = \sum_{i=1}^{20} i^{-1/2}U_i\sqrt{2}\cos(i\pi t)$, we then have

$$\psi_e(h_1) = \kappa_1 U_1 = U_1$$

from (4.10). Therefore, from (4.14),

$$\begin{aligned}\ell_1 &= \Psi_W(h_1) = \psi_e(h_1) + \langle h_1, \mu_1 Y(1) + \mu_2 Y(2)\rangle_{\mathcal{H}(K_W)} \\ &= U_1 + \langle\sqrt{2}\cos(\pi\cdot), 3Y(1)\sqrt{2}\cos(\pi\cdot) + \sqrt{2}\cos(2\pi\cdot)\rangle_{\mathcal{H}(K_W)} = 3Y(1) + U_1.\end{aligned}$$

From (6.35), (6.37), and (6.36), we find that the first canonical correlation, RKHS variate, and canonical variable of the X process are

$$\rho_1^2 = \frac{\gamma_1}{1+\gamma_1} = \frac{9}{13}$$

(i.e., $\rho_1 = 3/\sqrt{13} \approx .83$),

$$f_1(t) = (1-\rho_1^2)^{1/2}h_1(t) = \sqrt{13}\cos(\pi t)/\sqrt{2},$$

and

$$\eta_1 = (1-\rho_1^2)^{1/2}\ell_1 = \frac{2}{\sqrt{13}}(3Y(1) + U_1).$$

Consequently,

$$(b_{11}, b_{12})^T = (1, -1)^T$$

and

$$(\tilde{\eta}_{11}, \tilde{\eta}_{12})^T = (1.664, 0)^T.$$

The data in Figure 2 were analyzed via our estimation algorithm. We initially took $m = 100$ equally spaced points on $[0, 1]$ and $q = 20$. The smoothing parameter for cubic spline smoothing was chosen by generalized cross validation (GCV) for estimation of the

between-class covariance kernel K_B while the smoothing parameter at $\vartheta = .0008$ was used for estimation of the within-class covariance kernel K_W. The true and estimated between and within class covariance functions are shown in Figure 5. The estimated first canonical correlation in this case was found to be $\hat{\rho}_1 = .831$ with

$$(\hat{b}_{11}, \hat{b}_{12})^T = (1.083, -.923)^T$$

and

$$(\hat{\tilde{\eta}}_{11}, \hat{\tilde{\eta}}_{12})^T = (1.264, -.407)^T.$$

Figure 4 (a) and (b) provide the plots of the estimated and true eigenfunctions of T_B and T^*T corresponding to $\hat{\gamma}_1$ and $\hat{\rho}_1$, respectively, and Figure 6 shows the estimated versus true first canonical scores of the X space. Figure 7 is a plot of canonical X scores superimposed on the predicted canonical X scores assigned to the classes. From Table 2, the misclassification rate was 1 out of 50 or 2%.

(a) (b)

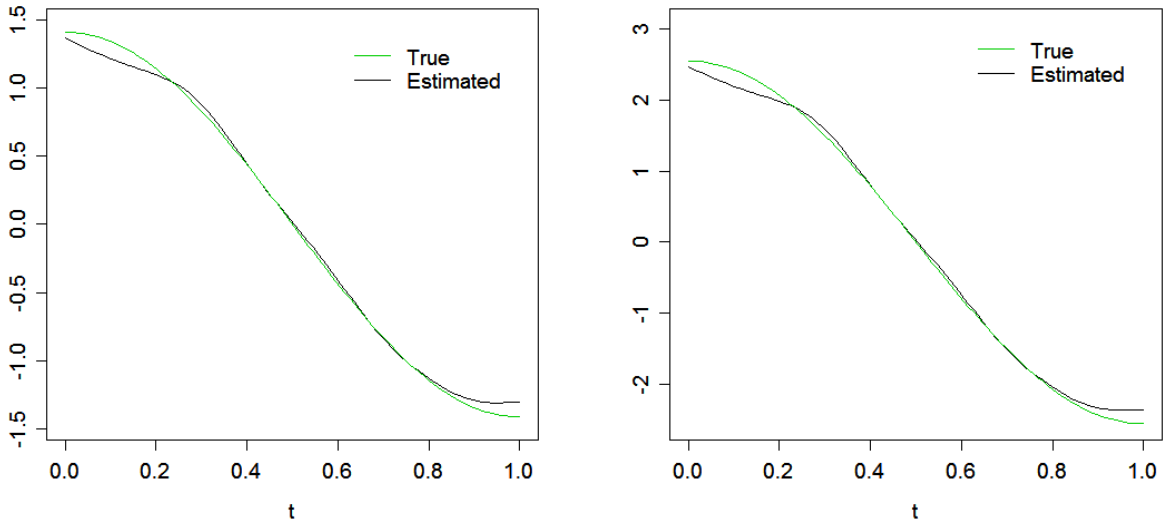

Figure 4: Estimated and true RKHS functions in $\mathcal{H}(K_W) = \mathcal{H}(K_X)$: (a) h_1 (green curve) and $\hat{h}_1$ (black curve); (a) f_1 (green curve) and $\hat{f}_1$ (black curve).

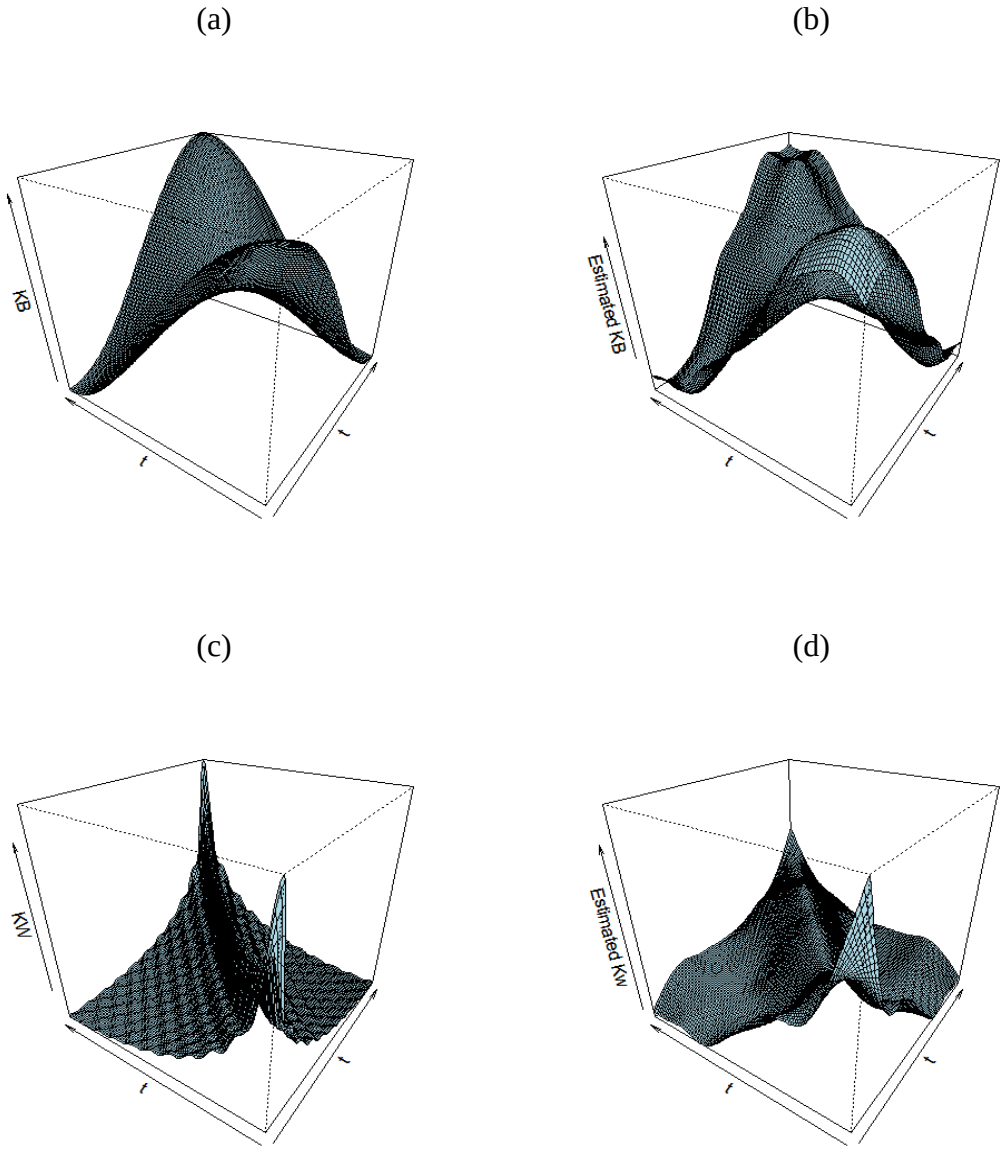

Figure 5: True and estimated between class covariance functions: (a) $K_B(\cdot,\cdot)$ and (b) $\widehat{K}_B(\cdot,\cdot)$; True and estimated within class covariance functions: (c) $K_W(\cdot,\cdot)$ and (d) $\widehat{K}_W(\cdot,\cdot)$.

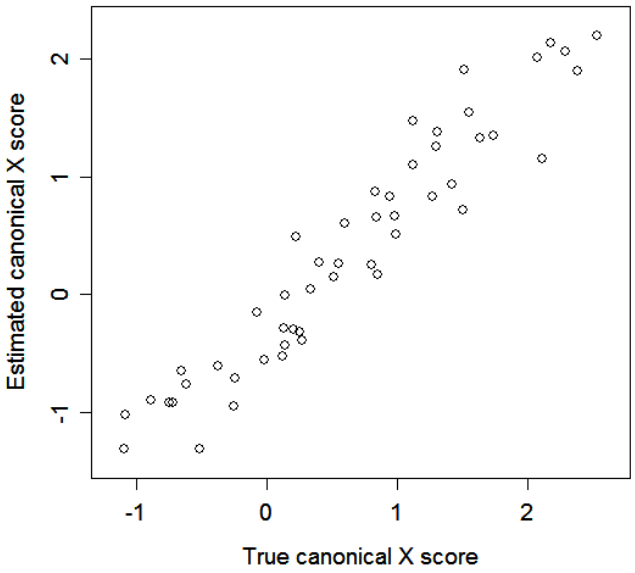

Figure 6: Estimated versus true canonical X scores: $\hat{\eta}_1$ versus η_1.

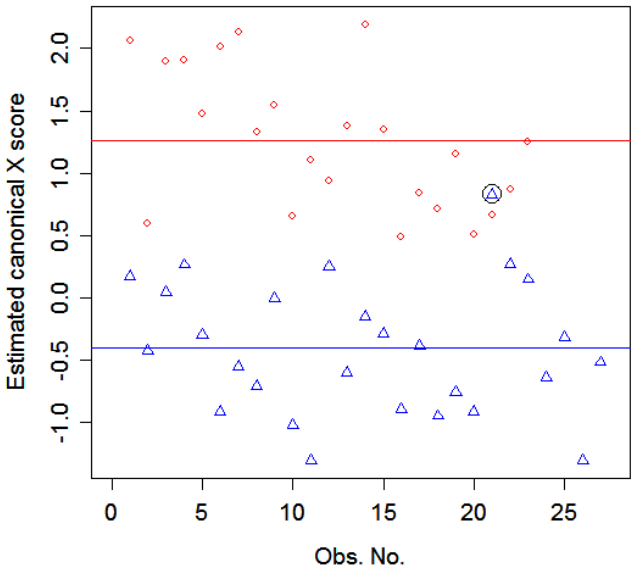

Figure 7: Each point represents the canonical X score for a sample path and the horizontal lines provide the values of $\tilde{\eta}_{1j}$. The sample curve corresponding to the point marked with black circle was misclassified.

Table 2: Confusion matrix of classification for the simulated data set

	Class 1	Class 2	
Class 1	23	0	23
Class 2	1	26	27

EXAMPLE 2. (Canadian Weather Data) Monthly temperatures for 35 weather stations distributed across Canada were measured. Canada can be divided into Atlantic, Continental, Pacific and Arctic meteorological zones and 14 stations are in the Atlantic zone, 5 stations in the Pacific, 13 stations in the Continental and 3 stations in the Arctic zones. Ramsay and Silverman (1997) used these data to conduct functional principal components analysis and functional analysis of variance. Figure 8 (a) and (b) show the monthly temperatures of 35 weather stations and mean monthly temperatures for the Canadian weather stations, respectively.

(a) (b)

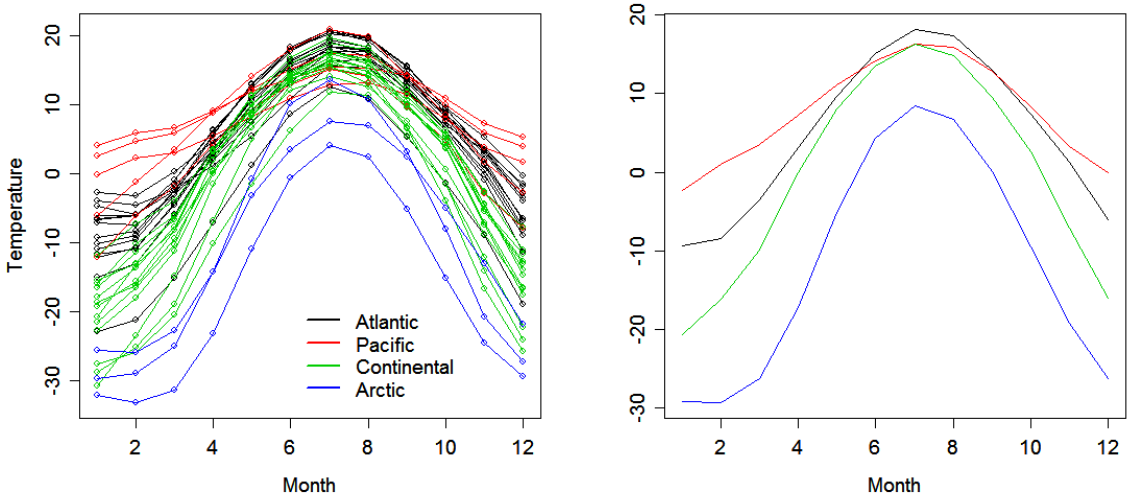

Figure 8: (a) Monthly temperatures for Canadian weather stations; (b) Mean monthly temperatures for the Canadian weather stations.

Let us analyze these data by our estimation algorithm. The estimated canonical corre-

lations are

$$\hat{\rho}_1 = .931,\ \hat{\rho}_2 = .910,\ \hat{\rho}_3 = .823$$

and the estimated coefficient vectors of the canonical variables of the Y space are

$$\hat{\mathbf{b}}_1 = (-.040, -.938, -.316, 3.123)^T,$$

$$\hat{\mathbf{b}}_2 = (-1.148, .067, 1.106, .450)^T,$$

$$\hat{\mathbf{b}}_3 = (.425, -2.261, .606, -.841)^T.$$

Figure 9 shows the estimated eigenfunctions of T^*T corresponding to $\hat{\rho}_1, \hat{\rho}_2$ and $\hat{\rho}_3$.

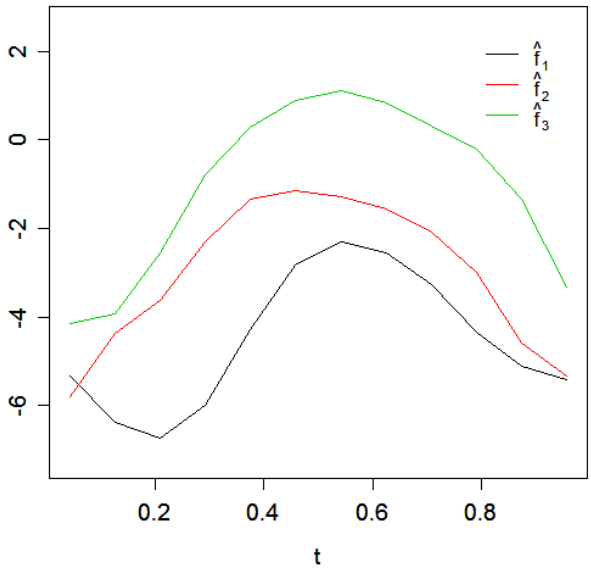

Figure 9: Estimated RKHS functions: $\hat{f}_1$ (black curve), $\hat{f}_2$ (red curve) and $\hat{f}_3$ (green curve)

As we investigated in Section 3.2.6, we can expect the first discriminator or canonical X variable to distinguish the Arctic zone from the others by looking at $\hat{\mathbf{b}}_1$. Similarly, we can expect the second discriminator to distinguish the Atlantic zone from the other zones and expect the third discriminator to distinguish the Pacific zone from the Atlantic and Continental zones. Since the three discriminators play different roles, they all contribute

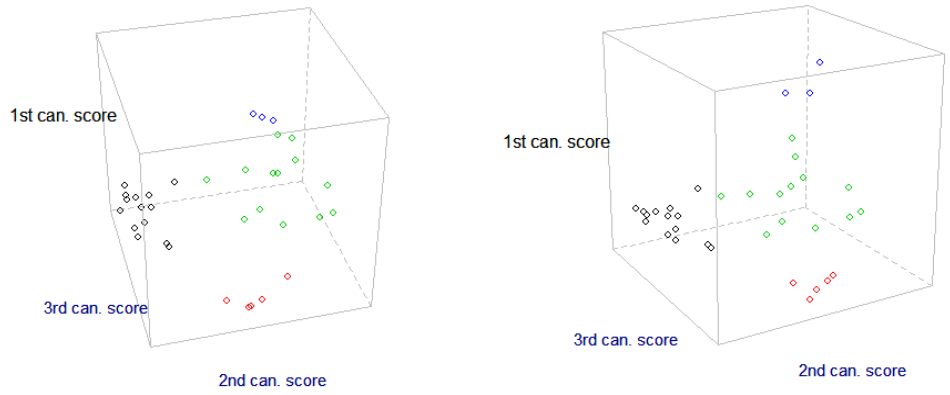

Figure 10: Plot of the canonical X scores of 35 weather stations. Each point represents score for a sample path.

to discrimination and the estimated canonical correlations are all large. As a result, we use all three discriminators for discrimination purposes. Table 3 is a confusion matrix of classification for the Canadian weather data. Figure 10 provides the canonical X scores for 35 temperature profiles with different angles. Separation is very clear and there is no misclassification.

Table 3: Confusion matrix of classification for Canadian monthly temperature data

	Atlantic	Pacific	Continental	Arctic	
Atlantic	14	0	0	0	14
Pacific	0	5	0	0	5
Continental	0	0	13	0	13
Arctic	0	0	0	3	3

CHAPTER VII

SUMMARY AND FUTURE RESEARCH

7.1 Summary

Multivariate analysis under the less than full rank scenario plays an important role as a beginning step for the development of infinite dimensional statistical methods. We have investigated multivariate canonical correlation analysis and discriminant analysis including Bayes' classifier and Fisher's discriminant method under the less than full rank scenario in Chapter III. Under this condition, we have shown the well-known connection between canonical correlation analysis and Fisher's discriminant method. Also, we have introduced some distance measures for classification and have shown the equivalence of those distance measures in a sense that parallels work by Hastie et al. (1995).

In this dissertation, discrimination and classification in infinite dimensional settings is motivated by the connection between Fisher's discriminant analysis method and canonical correlation analysis that is well known for the finite dimensional case. We have shown that this connection extends to infinite dimensions using the abstract canonical correlation concept developed by Eubank and Hsing (2005). A key part of this dissertation involved using this approach to develop a theoretical framework for discrimination and classification of sample paths from stochastic processes through use of the Loève-Parzen isomorphism that connects a second order process to the reproducing kernel Hilbert space generated by its covariance kernel. This paradigm provides a seamless transition between finite and infinite dimensional settings and lends itself well to computation via smoothing and regularization. In addition, we have developed and illustrated a new computational procedure with simulated data and Canadian weather data.

7.2 Future Research

One of the goals of this dissertation work was to develop a general methodological paradigm that simultaneously includes classical multivariate analysis, functional data analysis, etc. Statistical methods for analyzing functional data will ultimately parallel those for multivariate analysis. Among many possible extensions, a future research area concerns the infinite-dimensional extensions of Bayes' classifier method from multivariate analysis. Further, discriminant analysis and multivariate analysis of variance are closely related concepts that, in a sense, represent different sides of the same coin. As a result, this dissertation work also provides a theoretical structure from which one can extend ANOVA and MANOVA to the infinite dimensional setting. So, the next research area to consider is the development of high dimensional ANOVA techniques that can be applied to, e.g. the FDA context. Among other applications, the methodology developed in this dissertation can be applied to discriminant analysis for FDA bioinformatics data. Subsequent studies will pursue the development of large sample theory for the tests and estimators.

We conclude by mentioning a few other remaining problems that will be focused of future investigations. First, we have roughly shown that MANOVA under the less than full rank scenario parallels to the classical developments. This should be proved more precisely and connected to the general theory of multivariate linear models. Secondly, the computation algorithm in Section 6.5 needs to be refined for more complex data structure such as the data with noise, surfaces, etc. Finally, one can generalize the case of

$$K_B(s,t) = \sum_{j=1}^{\infty} \pi_j(\mu_j(s) - \bar{\mu}(s))(\mu_j(t) - \bar{\mu}(t)) \tag{7.1}$$

to situations

$$K_B(s,t) = \int_Q (\mu(s,q) - \bar{\mu}(s))(\mu(t,q) - \bar{\mu}(t))dP(q) \tag{7.2}$$

with P a Stieltjes measure on Q. This provides a collection of useful extensions of previous

developments that includes framework for the development of abstract regression concepts. We plan to explore this topic in some detail.

REFERENCES

Aronszajn, N. (1950). Theory of reproducing kernel. *Transaction of the American Mathematical Society* **68**, 337–404.

Doob, J. L. (1953). *Stochastic Processes*. New York: Wiley.

Eubank, R. (1999). *Nonparametric Regression and Spline Smoothing*. New York: Dekker.

Eubank, R. and Hsing, T. (2005). Canonical correlation for stochastic processes. Preprint.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

Hall, P., Poskitt, D. S., and Presnell, B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics* **43**, 1–9.

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *Annals of Statistics* **23**, 73–102.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer-Verlag.

He, G., Müller, H.-G., and Wang, J.-L. (2002). Functional canonical analysis for square integrable stochatic processes. *Journal of Multivariate Analysis* **85**, 54–77.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **20**, 321–377.

James, G. and Hastie, T. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B* **63**, 533–550.

Jewell, N. P. and Bloomfield, P. (1983). Canonical correlations of past and future for time series: Definitions and theory. *Annals of Statistics* **11**, 837–847.

Kettenring, J. R. (1971). Canonical analysis of several sets of variates. *Biometrika* **58**, 433–451.

Leurgans, S. E., Moyeed, R. A., and Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society, Series B* **55**, 725–740.

Loève, M. (1948). Fonctions aléatoires du second ordre. In P. Lévy (ed.), *Processus Stochastiques et Mouvement Brownien*. Paris: Gauthier-Villars.

Naimark, M. (1960). *Normed Rings, translated from the first Russian edition by Leo F. Boron*. Groningen: Noordhoff.

Parzen, E. (1961). An approach to time series analysis. *Annals of Mathematical Statistics* **32**, 951–989.

Parzen, E. (1962). Extraction and detection problems and reproducing kernel hilbert spaces. *SIAM J. ser A* **1**, 492–519.

Parzen, E. (1963). Probability density functionals and reproducing kernel hilbert spaces. In M. Rosenblatt (ed.), *Proceedings of the Symposium on Time Series Analysis*, pp. 155–169. New York: Wiley.

Parzen, E. (1967). *Time Series Analysis Papers*. San Fransico: Holden-Day.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. New York: Springer.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–259.

Riesz, F. and Sz.-Nagy, B. (1955). *Functional Analysis*. New York: Dover Publications.

Rudin, W. (1973). *Functional Analysis*. New York: McGraw-Hill.

Silverman, B. W. (1996). Smoothed functional principal component analysis by choice of norm. *Annals of Statistics* **24**, 1–24.

Tiao, G. and Tsay, R. (1989). Model specification in multivariate time series. *Journal of the Royal Statistical Society, Series B* **51**, 157–213.

Tsay, R. and Tiao, G. (1985). Use of canonical analysis in time series identification. *Biometrika* **72**, 299–315.

Weinert, H. L. (1982). *Reproducing Kernel Hilbert Spaces: Applications in Statistical Signal Processing*. Stroudsburg, PA: Hutchinson Ross.

VITA

Hyejin Shin received a Bachelor of Science degree in Statistics from Chonnam National University, Korea in 1999. She received a Masters of Science degree in Statistics from Seoul National University, Korea, under the direction of Dr. Byeong Uk Park in 2001. She was admitted to the Ph.D. program in the Department of Statistics at Texas A&M University in September 2001 and she received her Ph.D. degree in May 2006. Her permanent address is 34-5 Jangcheon-dong, Sucheon, Chonnam, 540-190, Republic of Korea.