

An RKHS Framework for Functional Data Analysis

Ana Kupresanin

*Department of Mathematics and Statistics, Arizona State University, Tempe, AZ
85287-1804*

Hyejin Shin

Mathematics and Statistics, Auburn University, Auburn, AL 36849-5310

David King

*Department of Mathematics and Statistics, Arizona State University, Tempe, AZ
85287-1804*

R. L. Eubank*

*Department of Mathematics and Statistics, Arizona State University, Tempe, AZ
85287-1804*

Abstract

Linear combinations of random variables play a crucial role in multivariate analysis. Two extension of this concept are considered for functional data and shown to coincide using the Loève-Parzen reproducing kernel Hilbert space representation of a stochastic process. This theory is then used to provide an extension of the multivariate concept of canonical correlation. A solution to the regression problem of best linear unbiased prediction is obtained from this abstract canonical correlation formulation. The classical identities of Lawley and Rao that lead to canonical factor analysis are also generalized to the functional data setting. Finally, the relationship between Fisher's linear discriminant analysis and canonical correlation analysis for random vectors is extended to include situations with function-valued random elements. This allows for classification using the canonical Y scores and related distance measures.

Key words: best linear prediction, classification, discriminant analysis, factor analysis, H-valued random variable

AMS 2000 Subject Classification: Primary 62H30, 62M99

1 Introduction

Functional data analysis (FDA) is a rapidly developing area in statistics due, in large part, to the pioneering work of Ramsay and Silverman (2005). The basic FDA premise is that one has infinite dimensional observations in the form of curves and wishes to analyze the data using techniques that parallel those from multivariate analysis.

In this article we describe a theoretical framework that can be used to formulate FDA methodology. Our approach relies on the Loève-Parzen congruence that links a second order stochastic process with the reproducing kernel Hilbert space (RKHS) generated by its covariance kernel. This congruence provides the vehicle for developing a rigorous formulation of functional canonical correlation (CCA) as detailed in Section 3. Functional CCA is then used to provide a generalization of key results for multivariate regression, factor analysis, MANOVA and discriminant analysis. In all cases these extensions are backward compatible in that they reduce to their parallels from multivariate analysis when the dimensionality is finite.

The paper is organized as follows. In the next section we extend the concept of linear combinations of random variables to the FDA setting. Then, in Section 3 we use this idea to describe the functional canonical correlation concept of Eubank and Hsing (2008). Section 4 generalizes a formula that connects multivariate regression and canonical correlation while Section 5 provides a similar extension of the Rao (1955) canonical factor analysis identity. Finally, Section 6 details how CCA can be applied to situations with multiple populations to produce formulations of functional analysis of variance and discriminant analysis. In the process we extend the equivalence between Fisher's linear discriminant analysis and canonical correlation analysis to an abstract data setting.

2 H-valued random variables

In this section we examine two ways of modeling the random structure that produces functional data. Both approaches have appeared in the FDA literature. Their unifying theme is that the data are, in some sense, realizations of "random functions." This intuitive view can be made rigorous through con-

* Corresponding author

Email addresses: amk@math.1a.asu.edu (Ana Kupresanin),
hjshin@auburn.edu (Hyejin Shin), dbking@asu.edu (David King),
eubank@math.asu.edu (R. L. Eubank).

sideration of Hilbert space valued random variables as we now explain.

Let $\{\Omega, \mathcal{A}, P\}$ be a probability space with \mathcal{H} representing a real, separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$. The σ -field generated by the class of all open subsets of \mathcal{H} is denoted by \mathcal{B} . A mapping $X : \Omega \rightarrow \mathcal{H}$ is called an \mathcal{H} -valued random variable if X is \mathcal{B} -measurable. A prototypical setting for FDA derives from this perspective by assuming that data are realization of an \mathcal{H} -valued random variable with \mathcal{H} a Hilbert function space.

The finite dimensional multivariate paradigm relies on linear combinations of vector random variables for the purpose of dimensionality reduction. A parallel of this approach for \mathcal{H} -valued random variables employs linear functionals of X (Remark 7.1.2 of Laha and Rohatgi 1979). Specifically, we can obtain a real valued, Hilbert space indexed, stochastic process $\{U(f) : f \in \mathcal{H}\}$ by taking

$$U(f) = \langle f, X \rangle_{\mathcal{H}}.$$

Assume $\int_{\mathcal{H}} \|f\|^2 dP_X(f) < \infty$ for all $f \in \mathcal{H}$ with P_X the probability measure that X induces on \mathcal{H} . Then (see, e.g., Laha and Rohatgi 1979), there is an element of \mathcal{H} that represents the mean for $U(\cdot)$ and a linear operator that provides the covariances for these random variables. The mean is determined by the (unique) member μ of \mathcal{H} that satisfies

$$\mathbb{E}[U(f)] := \int_{\mathcal{H}} \langle h, f \rangle_{\mathcal{H}} dP_X(h) = \langle \mu, f \rangle_{\mathcal{H}}$$

for all $f \in \mathcal{H}$. Since μ plays no role in our development until Section 5, we will assume that $\mu = 0$ for the present. In that event, the covariance operator for X is determined (uniquely) by the linear mapping $S_X : \mathcal{H} \rightarrow \mathcal{H}$ that satisfies

$$\mathbb{E}[U(f)U(g)] := \int_{\mathcal{H}} \langle h, f \rangle_{\mathcal{H}} \langle h, g \rangle_{\mathcal{H}} dP_X(h) = \langle f, S_X g \rangle_{\mathcal{H}}$$

for all $f, g \in \mathcal{H}$.

The covariance operator is Hilbert-Schmidt (Proposition 7.5.2 of Laha and Rohatgi 1979) and therefore admits the decomposition

$$S_X = \sum_{j=1}^{\infty} \lambda_j \phi_j \otimes_{\mathcal{H}} \phi_j,$$

where $\lambda_1 > \lambda_2 > \dots$ are the eigenvalues of the operator, $\{\phi_j\}_{j=1}^{\infty}$ are the associated eigenfunctions and $\otimes_{\mathcal{H}}$ is the operator defined by

$$(g \otimes_{\mathcal{H}} f)h = \langle g, h \rangle_{\mathcal{H}} f.$$

We can now create the pre-Hilbert space

$$\left\{ a : a = \sum_{j=1}^n f_j U(\phi_j), f_j \in \mathbb{R}, n \in \mathbb{Z}^+ \right\}$$

equipped with the inner product $\langle a_1, a_2 \rangle_{L_U^2} = E[a_1 a_2]$. The completion of this space will be denoted by L_U^2 and can be viewed as the set of all linear combinations, in an extended sense, of the members of $\{U(\phi_j) : j \in \mathbb{Z}^+\}$.

The covariance kernel for the $U(\cdot)$ process is

$$\text{Cov}(U(f_1), U(f_2)) = \langle f_1, S_X f_2 \rangle_{\mathcal{H}} := K_U(f_1, f_2)$$

for $f_1, f_2 \in \mathcal{H}$. The Moore-Aronszajn theorem (Section 2 of Aronszajn 1950) ensures that there is a unique reproducing kernel Hilbert space associated with K_U . Following Parzen (1970, Section 9) we can characterize this RKHS as

$$\mathcal{H}(K_U) = \left\{ \ell : \ell(g) = \sum_{j=1}^{\infty} \lambda_j f_j \langle g, \phi_j \rangle_{\mathcal{H}}, \sum_{j=1}^{\infty} \lambda_j f_j^2 < \infty \right\}.$$

This Hilbert space is congruent to L_U^2 ; that is, there is a 1-1, norm-preserving, linear map Ψ_U that maps $\mathcal{H}(K_U)$ onto L_U^2 . This congruence is an example of the Loève-Parzen RKHS representation for a second order process (e.g., Loève 1948 and Parzen 1961a). For the $U(\cdot)$ process it is possible to characterize the congruence as we will now describe.

Using the eigensystem for S_X we may express X in terms of a Karhunen–Loève type expansion as

$$X = \sum_{j=1}^{\infty} \langle X, \phi_j \rangle_{\mathcal{H}} \phi_j.$$

Thus, Theorem 4D of Parzen (1961a) has the consequence that

$$\Psi_U(\ell) = \sum_{j=1}^{\infty} f_j U(\phi_j) = \sum_{j=1}^{\infty} f_j \langle X, \phi_j \rangle_{\mathcal{H}}.$$

Combining this expression with the form of the RKHS $\mathcal{H}(K_U)$ we see that L_U^2 consists of random variables that are linear combination of the $U(\phi_i)$ with coefficients that satisfy $\sum_{j=1}^{\infty} \lambda_j f_j^2 < \infty$. In particular, a conclusion that may be drawn from this is that not all random variables in L_U^2 can be expressed as $\langle X, f \rangle_{\mathcal{H}}$ for some $f \in \mathcal{H}$.

A somewhat more specific scenario that has received attention in the FDA literature has $\mathcal{H} = L^2(T)$, where T is an interval subset of the line, the inner product is

$$\langle f_1, f_2 \rangle_{L^2(T)} = \int_T f_1(t) f_2(t) d\nu(t)$$

for $f_1, f_2 \in L^2(T)$ with ν some sigma-finite measure and the associated squared norm is $\|f_1\|_{L^2(T)}^2 = \langle f_1, f_1 \rangle_{L^2(T)}$. In this case we can think of an \mathcal{H} -valued random variable as corresponding to the stochastic process $\{X(t) : t \in T\}$ with covariance kernel

$$K_X(s, t) = \text{Cov}(X(s), X(t)).$$

This point-wise view can be made rigorous if, for example, K_X is continuous on T or, equivalently, $X(\cdot)$ is mean-squared continuous. We will presume that to be the case in what follows.

To work with linear combinations of $X(\cdot)$ we begin with the space

$$\left\{ a : a = \sum_{j=1}^n a_j X(t_j), t_j \in T, a_j \in \mathbb{R}, n \in \mathbb{Z}^+ \right\}$$

under the inner product $\langle a_1, a_2 \rangle_{L_X^2} = \text{E}[a_1 a_2]$. The completion of this pre-Hilbert space will be denoted by L_X^2 . Its elements represent a natural extension of the finite dimensional linear combinations of random variables that are a staple of multivariate analysis.

As was the case for K_U the covariance kernel K_X generates a reproducing kernel Hilbert space which we denote by $\mathcal{H}(K_X)$. The form of $\mathcal{H}(K_X)$ is determined from the integral operator defined by

$$(\mathcal{K}_X f)(t) = \int_T f(s) K_X(t, s) d\nu(s)$$

that represents a positive, compact operator on $L^2(T)$. Let $\{(\tilde{\lambda}_j, \tilde{\phi}_j)\}_{j=1}^\infty$ be the eigenvalue-eigenvector sequence for \mathcal{K}_X . Then, the RKHS corresponding to K_X is

$$\mathcal{H}(K_X) = \left\{ f : f = \sum_{j=1}^\infty \tilde{\lambda}_j f_j \tilde{\phi}_j, \sum_{j=1}^\infty \tilde{\lambda}_j f_j^2 < \infty \right\}. \quad (1)$$

The congruence mapping between L_X^2 and $\mathcal{H}(K_X)$ is given explicitly as

$$\Psi_X(f) = \sum_{j=1}^\infty f_j \langle X, \tilde{\phi}_j \rangle_{L^2(T)}.$$

Thus, as was true for elements of L_U^2 , we need only have $\sum_{j=1}^\infty \tilde{\lambda}_j f_j^2 < \infty$ meaning that not every element of L_X^2 can be expressed as $\langle f, X \rangle_{L^2(T)}$ for some $f \in L^2(T)$.

We now have two seemingly different ways of viewing information that may arise in an FDA context. However, it turns out that the two perspectives are equivalent as we shall demonstrate. First, observe that the eigensystem for

\mathcal{K}_X allows us to write

$$X(t) = \sum_{k=1}^{\infty} \langle X, \tilde{\phi}_k \rangle_{L^2(T)} \tilde{\phi}_k(t)$$

with the series converging in mean square. Thus,

$$\begin{aligned} K_U(f_1, f_2) &= \mathbb{E}[\langle f_1, X \rangle_{L^2(T)} \langle f_2, X \rangle_{L^2(T)}] \\ &= \int_T \int_T f_1(t) f_2(s) K_X(t, s) d\nu(t) d\nu(s) \\ &= \langle f_1, \mathcal{K}_X f_2 \rangle_{L^2(T)} \end{aligned}$$

from which we may conclude that $S_X = \mathcal{K}_X$ and, hence, that $\lambda_j = \tilde{\lambda}_j, \phi_j = \tilde{\phi}_j, j = 1, 2, \dots$. The Hilbert spaces L_X^2 and L_U^2 are therefore identical while $\mathcal{H}(K_X)$ and $\mathcal{H}(K_U)$ are congruent. In particular, this entails that there is no advantage to be gained from inferential formulations that deal with realizations of random variables of the form $\langle f, X \rangle_{L^2(T)}$ versus those relying on realizations of $X(\cdot)$ (and conversely) provided that all elements in L_X^2 and L_U^2 are allowed to participate in methodological developments.

There has, however, been some interest shown in basing inferential procedures on the elements of

$$L_{X0}^2 = \left\{ a : a = \sum_{j=1}^{\infty} f_j U(\phi_j), \sum_{j=1}^{\infty} f_j^2 < \infty \right\}.$$

The members of L_{X0}^2 can all be represented as $U(f) = \langle X, f \rangle_{L^2(T)}$ for some $f \in L^2(T)$ which makes them directly computable. But, L_{X0}^2 is a proper subset of the space $L_U^2 = L_X^2$ that contains all the random variables that can be constructed from the X process through linear operations. The relation between these two collections of random variables is illustrated by our next result.

Proposition 2.1 *Let $X(\cdot)$ be continuous in quadratic mean with ν Lebesgue measure and $T = [0, 1]$. Then, $L_X^2 = \bar{L}_{X0}^2$.*

Proof. The first step is to show that $\bar{L}_{X0}^2 \subset L_X^2$. In this regard consider a random variable $\int_0^1 f(t)X(t)dt \in L_{X0}^2$ and suppose without loss that $f \cdot X$ is nonnegative. Then,

$$\int_0^1 f(t)X(t)dt = \sup \left(\sum_i [\inf_{t \in A_i} X(t)f(t)] \nu(A_i) \right)$$

with the supremum extending over all finite decompositions $\{A_i\}$ of $[0, 1]$. Thus, for each n there exists a decomposition of $[0, 1]$ into finitely many in-

tervals $(A_1^n, \dots, A_{k_n}^n)$ such that

$$\int_0^1 f(t)X(t)dt = \lim_{n \rightarrow \infty} \sum_{i=1}^{k_n} [\inf_{t \in A_i^n} X(t)f(t)]\nu(A_i^n).$$

Let $\alpha_i^n = \nu(A_i^n)f(t_i^n)$, where $t_i^n \in A_i^n$ is such that $f(t_i^n)X(t_i^n) = \inf_{t \in A_i^n} f(t)X(t)$. Then,

$$\int_0^1 f(t)X(t)dt = \lim_{n \rightarrow \infty} \sum_{i=1}^{k_n} \alpha_i^n \nu(A_i^n) \in L_X^2.$$

To establish the converse result, fix $t_0 \in [0, 1]$ and consider the random variable $X(t_0)$. For each $n = 1, 2, \dots$, define $f_n(t) = \frac{n}{2}\chi_{B_n}(t)$ with χ_{B_n} the indicator function for $B_n = [t_0 - \frac{1}{n}, t_0 + \frac{1}{n}]$. The random variables $\int_0^1 f_n(t)X(t)dt$ are all in $L_{X_0}^2$ and

$$\lim_{n \rightarrow \infty} \int_0^1 f_n(t)X(t)dt = \lim_{n \rightarrow \infty} \frac{n}{2} \int_{t_0 - \frac{1}{n}}^{t_0 + \frac{1}{n}} X(t)dt = X(t_0)$$

which completes the proof. \square

From Proposition 2.1 we see that if $\mathcal{H} = L^2[0, 1]$ every random variable from L_X^2 can be approximated by a sequence of random variables in $L_{X_0}^2$. This, of course, does not mean that for each random variable $Z \in L_X^2$ there is a square integrable function f such that $Z = \langle X, f \rangle_{L^2[0,1]}$. Further, when one wants to approximate a random variable $Z \in L_X^2$ by a sequence of random variables in $L_{X_0}^2$ it is not enough to consider only convergent sequences of square integrable functions. This can be seen from the second part of the proof in which a divergent sequence of square integrable functions generates a sequence of random variables that converges to an L_X^2 random variable. Inference methods that work only on $L_{X_0}^2$ can be expected to produce suboptimal results in general and, as suggested by the proof, may incur numerical instabilities depending on how they handle the boundary of $L_{X_0}^2$.

We conclude this section by mentioning a characterization of $\mathcal{H}(K_X)$ due to Nashed and Wahba (1974) that provides insight into the mechanics that underlie the canonical correlation concept discussed in the next section. For this purpose observe that since \mathcal{K}_X is a positive operator we may write

$$K_X(s, t) = \int_T Q(s, u)Q(t, u)d\nu(u)$$

for a positive definite function Q that produces the square-root operator $\mathcal{K}_X^{1/2}$ defined by

$$(\mathcal{K}_X^{1/2}f)(t) = \int_T f(s)Q(t, s)d\nu(s).$$

Then, Proposition 2.2 of Nashed and Wahba (1974) gives

$$\mathcal{H}(K_X) = \left\{ f \in L^2(T) : \sum_{j=1}^{\infty} \langle f, \phi_j \rangle_{L^2(T)} / \lambda_j < \infty \right\}$$

with inner product

$$\langle f_1, f_2 \rangle_{\mathcal{H}(K_X)} = \sum_{j=1}^{\infty} \frac{\langle f_1, \phi_j \rangle_{L^2(T)} \langle f_2, \phi_j \rangle_{L^2(T)}}{\lambda_j}. \quad (2)$$

The fact that the $\mathcal{H}(K_X)$ norm must be finite is referred to as Picard's condition. In this respect it is known (e.g., Theorem 2.8 of Engl et al. 2000) to be a necessary and sufficient condition for existence of a solution of (in this case) $\mathcal{K}_X^{1/2} g = f$. If f satisfies Picard's condition, then

$$g = \mathcal{K}_X^{-1/2} f = \sum_{j=1}^{\infty} \frac{\langle f, \phi_j \rangle_{L^2(T)}}{\lambda_j} \phi_j$$

with $\mathcal{K}_X^{-1/2}$ the Moore-Penrose inverse of $\mathcal{K}_X^{1/2}$. From (1) we see that $\mathcal{H}(K_X)$ is the range of $\mathcal{K}_X^{1/2}$ which is a subset of $L^2(T)$. Although the range of $\mathcal{K}_X^{1/2}$ is not closed in $L^2(T)$ it becomes closed under the RKHS norm obtained from (2). Among other things this allows us to write

$$\langle f_1, f_2 \rangle_{\mathcal{H}(K_X)} = \langle \mathcal{K}_X^{-1/2} f_1, \mathcal{K}_X^{-1/2} f_2 \rangle_{L^2(T)} \quad (3)$$

with no ambiguity as to the meaning of $\mathcal{K}_X^{-1/2}$.

3 Canonical Correlation

Canonical correlation analysis (CCA) was developed by Hotelling (1936) to study the relationship between two vector random variables. However, CCA is intimately related to a variety of other multivariate concepts including multivariate regression, factor analysis, MANOVA and discriminant analysis. In subsequent sections we will see that the same basic relationships extend to situations where the random elements in question are infinite dimensional. As a first step in this development we will lay out a rigorous formulation of the infinite-dimensional canonical correlation problem for \mathcal{H} -valued processes with $\mathcal{H} = L^2(T)$.

Let $\{X(t), t \in T\}$ and $\{Y(s), s \in T\}$ be two zero mean stochastic processes with covariance functions K_X and K_Y and cross-covariance function K_{XY} defined by

$$K_{XY}(t, s) = \mathbb{E}[X(t)Y(s)], \quad s, t \in T. \quad (4)$$

As in the previous section, we use L_X^2 and L_Y^2 to denote the Hilbert spaces spanned by X and Y , respectively, and let $\mathcal{H}(K_X)$ and $\mathcal{H}(K_Y)$ be the corresponding RKHSs with reproducing kernels K_X and K_Y . Similarly, let Ψ_X and Ψ_Y be the congruence mappings between the Hilbert spaces spanned by the processes and the RKHSs spanned by their covariance functions.

Following Eubank and Hsing (2008) we define the first canonical correlation ρ_1 and the associated canonical variables $\zeta_1 = \Psi_X(f_1)$ and $\eta_1 = \Psi_Y(g_1)$ by

$$\begin{aligned} \rho_1^2 &= \text{Cov}^2(\zeta_1, \eta_1) = \sup_{\substack{\zeta \in L_X^2, \eta \in L_Y^2 \\ \text{Var}(\zeta) = \text{Var}(\eta) = 1}} \text{Cov}^2(\zeta, \eta) \\ &= \sup_{\substack{f \in \mathcal{H}(K_X), g \in \mathcal{H}(K_Y) \\ \|f\|_{\mathcal{H}(K_X)}^2 = \|g\|_{\mathcal{H}(K_Y)}^2 = 1}} \text{Cov}^2(\Psi_X(f), \Psi_Y(g)) = \text{Cov}^2(\Psi_X(f_1), \Psi_Y(g_1)). \end{aligned}$$

For $i > 1$, the canonical correlations ρ_i and canonical variables ζ_i and η_i are defined sequentially using the same criterion with the additional restriction of being uncorrelated with the previous canonical variables in the sequence. Using the reproducing property of K_X and K_Y one finds that

$$\text{Cov}(\Psi_X(f), \Psi_Y(g)) = \langle \langle K_{XY}(\star, \cdot), g(\cdot) \rangle_{\mathcal{H}(K_Y)}, f(\star) \rangle_{\mathcal{H}(K_X)} = \langle f, Rg \rangle_{\mathcal{H}(K_X)},$$

where the operator R from $\mathcal{H}(K_Y)$ to $\mathcal{H}(K_X)$ is defined by

$$(Rg)(t) = \langle K_{XY}(t, \cdot), g(\cdot) \rangle_{\mathcal{H}(K_Y)}, \quad t \in T. \quad (5)$$

An application of (3) reveals that

$$\langle Rg, f \rangle_{\mathcal{H}(K_X)} = \langle \mathcal{K}_X^{-1/2} \langle (\mathcal{K}_Y^{-1/2} K_{XY})(\star, \cdot), (\mathcal{K}_Y^{-1/2} g)(\cdot) \rangle_{L^2(T)}, \mathcal{K}_X^{-1/2} f(\star) \rangle_{L^2(T)}$$

which means that R has a representation as $\mathcal{K}_X^{-1/2} \mathcal{K}_{XY} \mathcal{K}_Y^{-1/2}$ with \mathcal{K}_{XY} the integral operator corresponding to the cross-covariance kernel (4).

The operator R is compact with adjoint $(R^*f)(t) = \langle K_{XY}(\cdot, t), f(\cdot) \rangle_{\mathcal{H}(K_X)}$ for all $t \in T$. Therefore, it follows from Eubank and Hsing (2008) that canonical variables and correlations are provided by the singular value decomposition (e.g., Engl et al. 2000)

$$R = \sum_{j=1}^{\infty} \rho_j g_j \otimes_{\mathcal{H}(K_Y)} f_j \quad (6)$$

where the g_j and f_j are the eigenfunctions of R^*R and RR^* , respectively, and $\rho_1^2 \geq \rho_2^2 \geq \dots \geq 0$ are the associated eigenvalues. As in the multivariate case, the ρ_i give the requisite canonical correlations while the canonical variables of the X and Y spaces are $\Psi_X(f_i)$ and $\Psi_Y(g_i)$, $i = 1, 2, \dots$. Eubank and Hsing (2008) show that these definition of canonical correlations and variables agree with those from multivariate analysis when the index set T is finite dimensional.

Other approaches to functional (and more general) forms of canonical correlation include the works of Dauxois and Pousse (1975), Dauxois, Nkiet and Romain (2004) and He et al. (2003). In the context of functional data all these references can be viewed as working with the Hilbert space indexed processes $\langle X, f \rangle_{L^2(T)}, \langle Y, g \rangle_{L^2(T)}, f, g \in L^2(T)$ which leads to a singular value decomposition of the operator $\mathcal{K}_X^{-1/2} \mathcal{K}_{XY} \mathcal{K}_Y^{-1/2}$ defined on the range of $\mathcal{K}_Y^{1/2}$ with the $L^2(T)$ norm. The Dauxois and Pousse (1975) formulation is valid when this range and the range of $\mathcal{K}_X^{1/2}$ are closed (e.g., Dauxois, Nkiet and Romain 2004 pg. 129) which is only true when the two operators are finite dimensional. Thus, their approach agrees with standard multivariate canonical correlation and, in that sense, provides a generalization of the classical concept. On the other hand it fails to provide a satisfactory extension for the case that is of interest here because it does not return a solution to the actual problem; that is, the Dauxois and Pousse (1975) construction does not in general provide the “linear combinations” of the X and Y processes that are maximally correlated. This is because they deal only with random variables that are obtained from $L^2(T)$ inner products and, as demonstrated in Section 2, not all “linear combinations” will be of this form. The optimal solution can, in fact, lie on the $L^2(T)$ boundary of the range of $\mathcal{K}_X^{1/2}, \mathcal{K}_Y^{1/2}$ as shown in Cupidon et al. (2007a). It would therefore seem that the Dauxois and Pousse (1975) claim of providing “a most general possible definition and formulation of the notion of canonical correlation” was, at the least, overly optimistic.

He, et al. (2003) recognize the difference between returning solutions from $L^2(T)$ versus those from L_X^2, L_Y^2 and give conditions on the singular functions which assure that an optimal set of canonical variables will be obtained from their canonical correlation analysis. However, their conditions will not hold in general and, as noted in Eubank and Hsing (2008), their solution will not in general agree with the one described here. We believe that a rigorous, backward compatible, solution to the functional canonical correlation problem cannot be formulated without the use of the RKHS congruences or their equivalent.

With the Dauxois and Pousse (1975), Dauxois, Nkiet and Romain (2004) and He et al. (2003) formulations providing cases in point, we can say that the RKHS approach outlined in this section provides a viable setting for rigorous development of functional CCA methodology for two reasons. First, it produces optimal solutions in L_X^2, L_Y^2 that provide true extension of the finite dimensional multivariate approach via inclusion of elements from $L^2(T)$ as well as finite dimensional linear combinations of the processes. Secondly, the operative domains for optimization are RKHSs and therefore closed under their norm induced topology. This latter factor is what insures the existence of canonical correlations and variables.

Estimation methods for the canonical correlations and variables determined by

(6) have been developed in Eubank and Hsing (2008) and Kupresanin (2008). The latter approach is based on a sample version of the algorithm for evaluating RKHS inner products and congruence mappings in Parzen (1963, Section 5). Large sample theory for regularized estimators of functional canonical correlations and variables are given in Cupidon, et al. (2007b).

4 Best linear prediction

As in the previous section we assume that we have zero mean processes X and Y with common index set T and consider the problem of predicting $Y(\cdot)$ using only information about $X(\cdot)$. Specifically, we want to find a best linear predictor (BLP) $\hat{Y}(t)$ of $Y(t)$ in the sense that

$$E[|Y(t) - \hat{Y}(t)|^2] = \inf_{a \in L_X^2} E[|Y(t) - a|^2].$$

The existence and uniqueness of the best linear predictor are provided by the standard Hilbert space projection theorem. But, in this case it is possible to provide a useful characterization of $\hat{Y}(t)$ as demonstrated by Parzen (1963, Theorem 3.1). An extended version of Parzen's result can be stated as follows.

Theorem 4.1 *The best linear predictor of $Y(t)$ given $\{X(s), s \in T\}$ is $\hat{Y}(t) = \Psi_X(K_{XY}(\cdot, t))$.*

An application of this result to our setting provides a direct connection between the BLP and CCA that can be formalized as follows.

Corollary 4.1 $\Psi_X(K_{XY}(\cdot, t)) = \sum_{j=1}^{\infty} \rho_j g_j(t) \Psi_X(f_j)$.

Proof. The result stems from the fact that $K_{XY}(t, s) = \sum_{j=1}^{\infty} \rho_j f_j(t) g_j(s)$. To see that this is true observe that since $K_{XY}(t, \cdot)$ is a function in $\mathcal{H}(K_Y)$ for each $t \in T$ (e.g., Proposition A.3 of Eubank and Hsing 2008), the reproducing property of K_Y in conjunction with (6) gives

$$\begin{aligned} K_{XY}(t, s) &= \langle K_{XY}(t, \cdot), K_Y(\cdot, s) \rangle_{\mathcal{H}(K_Y)} = (RK_Y(\cdot, s))(t) \\ &= \sum_{j=1}^{\infty} \rho_j \langle g_j, K_Y(\cdot, s) \rangle_{\mathcal{H}(K_Y)} f_j(t) = \sum_{j=1}^{\infty} \rho_j g_j(s) f_j(t). \end{aligned}$$

The linearity of Ψ_X can now be used to complete the proof. □

As a result of the theorem we see that the BLP of $Y(t)$ is a linear combination of the canonical variable $\zeta_j = \Psi_X(f_j)$ of the X space with weights obtained from i) the canonical correlations between $Y(\cdot)$ and $X(\cdot)$ and ii) the values of

$\{g_j(t)\}_{j=1}^{\infty}$ derived from the right-hand singular vectors of R . The implication is that CCA provides all the information required for the regression of Y on X in exactly the same manner that it does in multivariate analysis.

5 Factor analysis

In this section we apply results from Parzen (1970) to develop essential identities that lay a theoretical foundation for functional factor analysis. We begin by establishing a parallel of the canonical factor analysis formula of Rao (1955) from multivariate analysis.

The functional extension of the standard factor analysis model (e.g., Section 6.2 of Basilevsky 1994) takes the form of a signal-plus-noise model with

$$Y(t) = X(t) + N(t), \quad t \in T, \quad (7)$$

where Y, X have covariance kernels K_X, K_Y as before, the noise process N has covariance kernel K and the signal X is uncorrelated with the noise process. All processes are assumed to be $L^2(T)$ valued and have zero means.

If we now apply the canonical correlation approach of the previous section to model (7) we find that $(R^*Rg)(t) = (Rg)(t) = \langle K_X(t, \cdot), g(\cdot) \rangle_{\mathcal{H}(K_Y)}$ because $K_{XY} = K_X$. Since $K_Y = K_X + K$ this translates to squared canonical correlations and singular functions for the Y space that are solutions to

$$\langle K(t, \cdot), g(\cdot) \rangle_{\mathcal{H}(K_Y)} = (1 - \rho^2)g(t), \quad t \in T. \quad (8)$$

For the finite dimensional case K_Y and K can be represented by matrices \mathbf{K}_Y and \mathbf{K} while a function g corresponds to a vector \mathbf{g} . The $\mathcal{H}(K_Y)$ inner product then takes the form $\langle g, g \rangle_{\mathcal{H}(K_Y)} = \mathbf{g}^T \mathbf{K}_Y^{-1} \mathbf{g}$ with $\Psi_Y(g) = \mathbf{g}^T \mathbf{K}_Y^{-1} \mathbf{Y} := \mathbf{m}^T \mathbf{Y}$ for \mathbf{Y} a vector with elements $Y(t), t \in T$. Thus, (8) is equivalent to

$$\left[\mathbf{K}_Y - \frac{1}{1 - \rho^2} \mathbf{K} \right] \mathbf{m} = \mathbf{0} \quad (9)$$

which is identical to equation (3.2.3) of Rao (1955).

To complete the factor model we need to add structure to the X process. Specifically, we will assume that X is a second order process with the representation

$$X(t) = \sum_{j=1}^{\infty} Z_j \phi_j(t), \quad t \in T,$$

where the Z_j are zero mean, uncorrelated random variables with $E[Z_j^2] = \lambda_j$ and $\{\phi_j\}$ is an orthogonal sequence of functions in $\mathcal{H}(K)$ with $\|\phi_j\|_{\mathcal{H}(K)}^2 =$

$\gamma_j, j = 1, \dots$. In factor analysis terminology, the Z_j are the *factors* and we can refer to the ϕ_j as the *loading functions*.

Let \mathcal{H}_o be the Hilbert space of sequences of the form $f = \{f_j\}_{j=1}^\infty$ with squared norm $\sum_{j=1}^\infty \lambda_j f_j^2$ and define the operator $\Phi : \mathcal{H}_o \rightarrow \mathcal{H}(K)$ by

$$(\Phi f)(t) = \sum_{j=1}^{\infty} \lambda_j f_j \phi_j(t), \quad t \in T.$$

Then, one may show that $(\Phi\Phi^*g)(t) = \langle K_X(t, \cdot), g \rangle_{\mathcal{H}(K)}$.

Now, $\Phi\Phi^*\phi_j = \lambda_j\gamma_j\phi_j$ which characterizes the loading functions as being the eigenvectors of the operator $\Phi\Phi^*$ on $\mathcal{H}(K)$. In the finite dimensional case we can write the X process as the vector $\mathbf{X} = \Phi\mathbf{Z}$ with \mathbf{Z} having $\mathbb{E}[\mathbf{Z}\mathbf{Z}^T] = \mathbf{\Lambda}$ for a diagonal matrix $\mathbf{\Lambda}$ and Φ a matrix satisfying $\Phi^T\mathbf{K}^{-1}\Phi = \mathbf{\Gamma}$ with $\mathbf{\Gamma}$ diagonal. Since $\mathbf{K}_X = \Phi\mathbf{\Lambda}\Phi^T$, the eigen-equations for $\Phi\Phi^*$ can now be expressed as

$$\left[\mathbf{K}^{-1/2}\mathbf{K}_Y\mathbf{K}^{-1/2} - (1 + \lambda_j\gamma_j)\mathbf{I} \right] \mathbf{K}^{-1/2}\phi_j = \mathbf{0} \quad (10)$$

with \mathbf{I} the identity matrix and ϕ_j the j th column of Φ . Formula (10) agrees with the Lawley factor analysis “normal equations” as set out in Theorem 6.5 of Basilevsky (1994).

The reason that we cannot immediately connect $\Phi\Phi^*$ to the operator R is because the former involves the $\mathcal{H}(K)$ inner product while the latter is phrased in terms of the inner product for $\mathcal{H}(K_Y)$. To make the connection between the two inner products we will use two results: namely,

$$\langle f, g \rangle_{\mathcal{H}(K_Y)} = \langle f, (I + \Phi\Phi^*)^{-1}g \rangle_{\mathcal{H}(K)}$$

and

$$K(t, \cdot) = (I + \Phi\Phi^*)^{-1}K_Y(t, \cdot), \quad t \in T.$$

The first result is a consequence of developments in Section 6 of Parzen (1970) while the latter follows from

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

An application of these two identities reveals that $Rg = \rho^2g$ is equivalent to $\Phi\Phi^*g = (\rho^2/(1 - \rho^2))g$. Consequently, the singular functions from the Y space are, apart from a normalization, the loading functions from the factor analysis model. We note in passing that the present formulation provides a meaningful interpretation for the standard factor analysis “identifiability” constraint that $\Phi^T\mathbf{K}^{-1}\Phi$ be diagonal. From the RKHS perspective this is just an orthogonality condition for the loading vectors in the Hilbert space $\mathcal{H}(K)$ that represents their home.

6 ANOVA and discriminant analysis

It is well known that CCA, discriminant analysis and MANOVA are closely linked in the case of ordinary multivariate analysis. In this section we show that this connection remains true in the functional data analysis setting.

One formulation of functional ANOVA/discrimination is to assume that an \mathcal{H} -valued random variable is observed from one of J possible populations. If π_j is the possibility that the reading comes from population j , we can view the problem as seeing the pair (X, G) , where X takes values in $L^2(T)$ and $G \in \{1, \dots, J\}$ is a categorical response variable representing the class membership with $P(G = j) = \pi_j$. It is assumed that the populations have mean functions $E[X(\cdot)|G = j] = \mu_j(\cdot)$ that do not all coincide.

To proceed in an environment such as that of Sections 2–3 we must first extend the basic congruence mappings to allow for non-zero mean functions. In that regard if, e.g., the X process has mean $\mu_X(t) = E[X(t)] \neq 0$ it follows from Parzen (1961b, Section 5) that if $\mu_X \in \mathcal{H}(K_X)$ there exist a one-to-one linear mapping from $\mathcal{H}(K_X)$ onto L^2_X that we will continue to denote by Ψ_X . This mapping satisfies i) $\Psi_X(K_X(\cdot, t)) = X(t)$, $t \in T$, ii) $E[\Psi_X(f)] = \langle f, \mu_X \rangle_{\mathcal{H}(K_X)}$ and iii) $\text{Cov}(\Psi_X(f_1), \Psi_X(f_2)) = \langle f_1, f_2 \rangle_{\mathcal{H}(K_X)}$. The analogous mapping for the Y process will be denoted as Ψ_Y .

Rather than being $L^2(T)$ -valued the Y process in this setting is $\{Y(j) : j = 1, \dots, J\}$ with $Y(j)$ an indicator variable that is 1 when X derives from the j th population. The covariance function for the Y process is $K_Y(i, j) = \text{Cov}(Y(i), Y(j))$ with matrix form $\mathbf{K}_Y = \text{diag}(\pi_1, \dots, \pi_J) - \boldsymbol{\pi}\boldsymbol{\pi}^T$ with $\boldsymbol{\pi} = (\pi_1, \dots, \pi_J)^T$. Now let $\mathcal{H}(K_X)$, $\mathcal{H}(K_Y)$ be the RKHSs corresponding to K_X , K_Y , respectively. In particular, $\mathcal{H}(K_Y)$ consists of functions on $\{1, \dots, J\}$ of the form $\sum_{j=1}^J b_j K_Y(\cdot, j)$ for $\mathbf{b} = (b_1, \dots, b_J)^T \in \text{Ker}(\mathbf{K}_Y)^\perp$ with $\text{Ker}(\mathbf{K}_Y)^\perp$ denoting the orthogonal complement of the null space of \mathbf{K}_Y . The associated inner product is $\langle g_1, g_2 \rangle_{\mathcal{H}(K_Y)} = \mathbf{g}_1^T \mathbf{K}_Y^- \mathbf{g}_2$, where $\mathbf{g}_i = (g_i(1), \dots, g_i(J))^T$, $i = 1, 2$, for $g_1, g_2 \in \mathcal{H}(K_Y)$ and \mathbf{K}_Y^- the Moore-Penrose generalized inverse of \mathbf{K}_Y . Since $\mathbf{K}_Y^G = \text{diag}(\pi_1^{-1}, \dots, \pi_J^{-1})$ is a generalized inverse of \mathbf{K}_Y , the inner product can also be expressed as

$$\langle g_1, g_2 \rangle_{\mathcal{H}(K_Y)} = \sum_{j=1}^J \frac{g_1(j)g_2(j)}{\pi_j}.$$

The canonical variables of the X and Y spaces are

$$\zeta_i = \Psi_X(f_i) \quad \text{and} \quad \eta_i = \Psi_Y(g_i) = \mathbf{g}_i^T \mathbf{K}_Y^- \mathbf{Y},$$

where $\mathbf{Y} = (Y(1), \dots, Y(J))^T$ with f_i, g_i the singular functions of the operator

R in (5). One finds that

$$(Rg)(t) = \sum_{j=1}^J \frac{K_{XY}(t, j)g(j)}{\pi_j} = \sum_{j=1}^J g(j)\mu_j(t)$$

with $K_{XY}(t, j) = \text{Cov}(X(t), Y(j)), t \in T, j = 1, \dots, J - 1$. Thus, R maps to contrasts among the population mean functions with the consequence that all the mean functions coincide if and only if the R singular values $\rho_1, \dots, \rho_{J-1}$ vanish. Testing the functional ANOVA hypothesis that $\mu_1(\cdot) = \dots = \mu_J(\cdot)$ is therefore equivalent to the hypothesis that all the canonical correlations are zero.

The canonical X and Y variables may be used for classification purposes. Let (ζ_k, η_k) be the canonical variable pair corresponding to ρ_k and define $\zeta_k^c = \zeta_k - \text{E}[\zeta_k], \eta_k^c = \eta_k - \text{E}[\eta_k]$. Then, for example, to predict η_k^c from ζ_k^c we can use $\rho_k \zeta_k^c$. Considerations of this nature suggest classification via minimization of distance measures such as

$$\sum_{k=1}^s \frac{1}{1 - \rho_k^2} (\zeta_k^c - \rho_k \pi_j^{-1} g_k(j))^2, \quad (11)$$

$$\sum_{k=1}^s \frac{1}{\rho_k^2 (1 - \rho_k^2)} (\eta_k - \rho_k \text{E}[\zeta_k | G = j])^2 \quad (12)$$

and

$$\sum_{k=1}^s \frac{1}{1 - \rho_k^2} (\eta_k - \pi_j^{-1} g_k(j))^2 - \sum_{k=1}^s (\pi_j^{-1} g_k(j))^2 \quad (13)$$

for some $s \leq J - 1$.

Shin (2008) has used an RKHS formulation to generalize the classical Fisher's method for linear discriminant analysis (LDA) to the FDA setting. Specifically, she defines the discriminant functions to be random variables $\ell \in L_X^2 \setminus \{0\}$ that maximize $\text{Var}(\text{E}[\ell | G]) / \text{E}[\text{Var}(\ell | G)]$. Thus, the first linear discriminant function ℓ_1 is defined by

$$\text{Var}(\text{E}[\ell_1 | G]) = \sup_{\ell \in L_X^2} \text{Var}(\text{E}[\ell | G]),$$

where ℓ is subject to $\text{E}[\text{Var}(\ell | G)] = 1$. The i th linear discriminant function for $i > 1$ is defined similarly subject to the additional restriction $\text{E}[\text{Cov}(\ell_i, \ell_k | G)] = 0, k < i$.

Assume that the within group covariance kernel K_W is the same across the J populations: i.e., $K_W(s, t) = \text{E}[(X(s) - \mu_j(s))(X(t) - \mu_j(t)) | G = j], s, t \in T$ for $j = 1, \dots, J$. Next, define the kernel function $K_B(s, t) = \sum_{j=1}^J \pi_j (\mu_j(s) - \mu(s))(\mu_j(t) - \mu(t))$ for $s, t \in T$ with $\mu(\cdot) = \sum_{j=1}^J \pi_j \mu_j(\cdot)$ and denote the RKHS generated by K_W as $\mathcal{H}(K_W)$. Then, Shin (2008) shows that if $\mu_j \in \mathcal{H}(K_W)$ for all $j = 1, \dots, J$, there exists a one-to-one linear mapping Ψ_W from $\mathcal{H}(K_W)$

onto L_X^2 defined by $\Psi_W : K_W(\cdot, t) \rightarrow X(t)$ for every t in T with the properties i) $E[\Psi_W(h)] = \langle h, \mu \rangle_{\mathcal{H}(K_W)}$, ii) $E[\Psi_W(h)|G = j] = \langle h, \mu_j \rangle_{\mathcal{H}(K_W)}$ and iii) $E[\text{Var}(\Psi_W(h)|G)] = \|h\|_{\mathcal{H}(K_W)}^2$ for $h \in \mathcal{H}(K_W)$. Using Ψ_W we see that

$$\text{Var}(E[\Psi_W(h)|G]) = \langle h, R_B h \rangle_{\mathcal{H}(K_W)}$$

with

$$(R_B h)(t) = \langle K_B(t, \cdot), h(\cdot) \rangle_{\mathcal{H}(K_W)} \quad (14)$$

for $h \in \mathcal{H}(K_W)$. The operator R_B has spectral decomposition

$$R_B = \sum_{i=1}^{J-1} \gamma_i h_i \otimes_{\mathcal{H}(K_W)} h_i,$$

with $\gamma_1 \geq \dots \geq \gamma_{J-1} \geq 0$ the eigenvalues and $h_i, i = 1, \dots, J-1$, the associated eigenfunctions for the operator. The linear discriminant functions are

$$\ell_i = \Psi_W(h_i), \quad i = 1, \dots, J-1,$$

with classification obtained from squared Mahalanobis distance based on the first s ($\leq J-1$) linear discriminant functions. That is, for the j th population we employ the distance measure

$$\sum_{k=1}^s \left(\Psi_W(h_k) - \langle h_k, \mu_j \rangle_{\mathcal{H}(K_W)} \right)^2 \quad (15)$$

that compares the process ‘‘score’’ vector $(\Psi_W(h_1), \dots, \Psi_W(h_s))$ to its conditional mean $(\langle h_1, \mu_j \rangle_{\mathcal{H}(K_W)}, \dots, \langle h_s, \mu_j \rangle_{\mathcal{H}(K_W)})$ for the j th population.

Now CCA and Fisher’s LDA are well known to be equivalent in the case of finite dimensional multivariate analysis. We now show that this equivalence continues to hold for FDA situations.

Theorem 6.1 *Assume that $\mu_j \in \mathcal{H}(K_W), j = 1, \dots, J$. Then,*

$$\gamma_i = \frac{\rho_i^2}{1 - \rho_i^2}, \quad h_i = (1 - \rho_i^2)^{1/2} f_i \text{ and } \ell_i = \zeta_i / (1 - \rho_i^2)^{1/2} \quad (16)$$

for $i = 1, \dots, J-1$.

Proof. For the theorem to be true it must be that both R and R_B are defined on the same space: i.e., we must have $\mathcal{H}(K_W) = \mathcal{H}(K_X)$. To see that this is so first observe that $\mathcal{H}(K_W) \subset \mathcal{H}(K_X)$ due to of Theorem I in Aronszajn (1950). To establish the converse define $L : \mathcal{H}(K_X) \rightarrow \mathcal{H}(K_W)$ by

$$L(K_X(\cdot, t)) = K_W(\cdot, t), \quad t \in T. \quad (17)$$

Then, L is a one-to-one, onto and, for $h \in \mathcal{H}(K_W)$ and $f \in \mathcal{H}(K_X)$,

$$\langle h, f \rangle_{\mathcal{H}(K_X)} = \langle h, Lf \rangle_{\mathcal{H}(K_W)}$$

because $\langle h, K_X(\cdot, t) \rangle_{\mathcal{H}(K_X)} = h(t) = \langle h, K_W(\cdot, t) \rangle_{\mathcal{H}(K_W)}$. Since $K_X = K_B + K_W$, for $f \in \mathcal{H}(K_X)$

$$\begin{aligned} f(t) &= \langle K_B(\cdot, t), f \rangle_{\mathcal{H}(K_X)} + \langle K_W(\cdot, t), f \rangle_{\mathcal{H}(K_X)} \\ &= \langle K_B(\cdot, t), Lf \rangle_{\mathcal{H}(K_W)} + \langle K_W(\cdot, t), Lf \rangle_{\mathcal{H}(K_W)}. \end{aligned}$$

Thus, $f = (R_B Lf) + (Lf) \in \mathcal{H}(K_W)$.

Now R_B induces an operator C_B on $\mathcal{H}(K_X)$ determined from $(C_B f)(t) = \langle K_B(t, \cdot), f(\cdot) \rangle_{\mathcal{H}(K_X)}$. One finds that $C_B = R_B L$ for L in (17) because

$$(R_B Lf)(t) = \langle K_B(\cdot, t), Lf \rangle_{\mathcal{H}(K_W)} = \langle K_B(\cdot, t), f \rangle_{\mathcal{H}(K_X)} = (C_B f)(t).$$

Since $f = (Lf) + (R_B Lf)$, we have $L = I - C_B$ and, for $f \in \mathcal{H}(K_X)$, $\|Lf\|_{\mathcal{H}(K_W)}^2 = \langle Lf, f \rangle_{\mathcal{H}(K_X)} = \langle (I - C_B)f, f \rangle_{\mathcal{H}(K_X)}$. The linear mappings Ψ_X from $\mathcal{H}(K_X)$ to L_X^2 and Ψ_W from $\mathcal{H}(K_W)$ to L_X^2 are similarly related in that $\Psi_X(f) = \Psi_W(Lf)$ for $f \in \mathcal{H}(K_X)$ which follows from $\Psi_X(K_X(\cdot, t)) = X(t) = \Psi_W(K_W(\cdot, t)) = \Psi_W(L(K_X(\cdot, t)))$.

For $f \in \mathcal{H}(K_X)$ we have

$$(RR^* f)(t) = \langle K_{XY}(t, \cdot), (R^* f)(\cdot) \rangle_{\mathcal{H}(K_Y)} = \langle RK_{XY}(t, \cdot), f(\cdot) \rangle_{\mathcal{H}(K_X)}.$$

But, $K_{XY}(\cdot, i) = \text{Cov}(X(\cdot), Y(i)) = \pi_i(\mu_i(\cdot) - \mu(\cdot))$ for $i = 1, \dots, J$ and

$$\begin{aligned} RK_{XY}(t, \cdot) &= \langle K_{XY}(\cdot, *), K_{XY}(t, *) \rangle_{\mathcal{H}(K_Y)} \\ &= \sum_{j=1}^J \frac{K_{XY}(\cdot, j)K_{XY}(t, j)}{\pi_j} = K_B(t, \cdot). \end{aligned}$$

Thus, $RR^* = C_B$.

We can now use the fact that $RR^* f = C_B f = R_B Lf$ and $f = Lf + R_B Lf$ for $f \in \mathcal{H}(K_X)$ to write

$$R_B Lf_i = \frac{\rho_i^2}{1 - \rho_i^2} Lf_i.$$

Also, $Lf_i = (1 - \rho_i^2)f_i$ from $Lf_i = (I - C_B)f_i = (I - RR^*)f_i$ and $\|Lf_i\|_{\mathcal{H}(K_W)}^2 = \|f_i\|_{\mathcal{H}(K_X)}^2 - \langle C_B f_i, f_i \rangle_{\mathcal{H}(K_X)} = 1 - \rho_i^2$ since $\|f_i\|_{\mathcal{H}(K_X)}^2 = 1$ while the h_i in $\mathcal{H}(K_W)$ corresponding to $\ell_i = \Psi_W(h_i)$, satisfy $\|h_i\|_{\mathcal{H}(K_W)}^2 = 1$. Thus,

$$h_i = \frac{Lf_i}{\|Lf_i\|_{\mathcal{H}(K_W)}} = \frac{Lf_i}{(1 - \rho_i^2)^{1/2}} = (1 - \rho_i^2)^{1/2} f_i.$$

The proof is completed using the relationship between Ψ_X and Ψ_W . \square

Theorem 6.1 has the implication that the RKHS vectors h_i are the optimum orthogonal contrasts among the class means and the Fisher's discriminant functions are given by $\ell_i = \gamma_i^{-1/2} \Psi_W(Rg_i) = \Psi_W\left(\frac{Rg_i}{\|Rg_i\|_{\mathcal{H}(K_W)}}\right)$ which is an exact parallel of what transpires for the finite dimensional setting. (See, e.g., Kshirsagar 1972.)

We can also interpret the canonical variables of the Y space from a discrimination perspective. To see this first observe that

$$\langle f, Rg \rangle_{\mathcal{H}(K_X)} = \sum_{j=1}^J g(j) \langle f, \mu_j \rangle_{\mathcal{H}(K_X)}$$

is a contrast among the transformed mean functions $\langle f, \mu_j \rangle_{\mathcal{H}(K_X)}$, $j = 1, \dots, J$. In this respect consider, for example, $\Psi_X(f_1)$ and $\Psi_Y(g_1)$. Then, f_1 and g_1 are obtained by maximizing $\left| \sum_{j=1}^J g(j) \langle f, \mu_j \rangle_{\mathcal{H}(K_X)} \right|$ subject to $\|f\|_{\mathcal{H}(K_X)} = 1$, $\sum_{j=1}^J g(j) = 0$ and $\sum_{j=1}^J g^2(j)/\pi_j = 1$. The function g_1 provides the coefficients of a contrast in the transformed means $\langle f, \mu_j \rangle_{\mathcal{H}(K_X)}$ and therefore measures their importance in the contrast. Thus, just as in the finite dimensional case, the canonical variables of the Y space given by $\Psi_Y(g_i)$ provide optimum scores to represent the relative location of the different populations or groups.

We conclude with a result that connects up the various classification metrics.

Theorem 6.2 *The distance measures (11)–(13) and (15) are all equivalent.*

Proof. Let $\theta_{kj} = \pi_j^{-1} g_k(j)$ and $\bar{\eta}_{kj} = \rho_k \mathbb{E}[\zeta_k | G = j]$. The relations $\Psi_W(h_k) = (1 - \rho_k^2)^{-1/2} \Psi_X(f_k)$ and

$$\langle f_k, \mu_j \rangle_{\mathcal{H}(K_X)} = \langle Lf_k, \mu_j \rangle_{\mathcal{H}(K_W)} = (1 - \rho_k^2)^{1/2} \langle h_k, \mu_j \rangle_{\mathcal{H}(K_W)}$$

ensure that (15) becomes

$$\sum_{k=1}^s \frac{1}{1 - \rho_k^2} (\zeta_k - \langle f_k, \mu_j \rangle_{\mathcal{H}(K_X)})^2 = \sum_{k=1}^s \frac{1}{1 - \rho_k^2} (\zeta_k^c - \langle f_k, \mu_j - \mu \rangle_{\mathcal{H}(K_X)})^2.$$

Since $(R^* f_k)(j) = \langle f_k(\cdot), K_{XY}(\cdot, j) \rangle_{\mathcal{H}(K_X)} = \pi_j \langle f_k, \mu_j - \mu \rangle_{\mathcal{H}(K_X)}$ and $g_k(j) = \sum_{i=1}^J b_{ki} K_Y(i, j) = \pi_j \theta_{kj}$, we have $\langle f_k, \mu_j - \mu \rangle_{\mathcal{H}(K_X)} = \rho_k \pi_j^{-1} g_k(j) = \rho_k \theta_{kj}$. Moreover, $\bar{\eta}_{kj} = \mathbb{E}[\eta_k | G = j] = \rho_k \langle f_k, \mu_j - \mu \rangle_{\mathcal{H}(K_X)}$. Thus, (15) is equivalent to (11) and (12). Using the fact that $\bar{\eta}_{kj} = \rho_k^2 \theta_{kj}$, (12) simplifies to

$$\sum_{k=1}^s \rho_k^{-2} \eta_k^2 + \sum_{k=1}^s \frac{1}{1 - \rho_k^2} (\eta_k - \theta_{kj})^2 - \sum_{k=1}^s \theta_{kj}^2$$

and (13) is obtained. \square

Hastie, et al. (1995) consider penalized forms of discriminant and canonical correlation analysis where they establish relationships between scoring methods that parallel some of those in Theorem 6.2. In the context of FDA the population entities targeted by their proposed estimation methods involve inverses of integral operators. Since integral operators are not invertible it is difficult to assess the statistical relevance of the methodology they propose. The framework for FDA set out in this paper could provide a means to rigorously link the Hastie, et al. penalized estimation methods to well defined parameters and thereby clarify the implications of their work.

Acknowledgments. This paper would not exist without the genius and guidance of our advisor/grand-advisor and friend Emanuel Parzen.

References

- [1] Aronszajn, N., 1950. Theory of reproducing kernels, *Amer. Math. Soc. Trans.* **68**, 337–404.
- [2] Basilevsky, A., 1994. *Statistical Factor Analysis and Related Methods: Theory and Applications*, Wiley, New York.
- [4] Cupidon, J., Eubank, R., Gilliam, D. and Ruymgaart, F., 2007a. Some properties of canonical correlations and variates in infinite dimensions. *J. Multivariate Anal.* **99**, 1083–1104.
- [4] Cupidon, J., Eubank, R., Gilliam, D. and Ruymgaart, F., 2007b. The delta-method for analytic functions of random operators with application to functional data. *Bernoulli* **13**, 1179–1194.
- [5] Dauxois, J. and Nkiet, G. and Romain, Y., 2004a. Canonical analysis relative to a closed subspace, *Linear Algebra Appl.* **388**, 119–145.
- [6] Dauxois, J. and Pousse, A., 1976. Une extension de l’analyse canonique. quelques applications, *Ann. Inst. H. Poincaré Probab. Statist.* **11**, 355–379.
- [7] Engl, H., Hanke, M. and Neubauer, A., 2000. *Regularization of Inverse Problems*. Kluwer Academic Publishers, Boston.
- [8] Eubank, R. and Hsing, T., 2008. Canonical correlation for stochastic processes, *Stochastic Process. Appl.* **118**, 1634–1661.
- [9] He, G., Müller, H.-G. and Wang, J.-L., 2003. Functional canonical correlation analysis for square integrable stochastic processes, *J. Multivariate Anal.* **85**, 54–77.
- [10] Hotelling, H., 1936. Relations between two sets of variates, *Biometrika* **20**, 321–377.
- [11] Kshirsagar, A., 1972. *Multivariate Analysis*. Marcel Dekker, New York.
- [12] Kupresanin, A., 2008. *Topics in Functional Canonical Correlation and*

- Regression*. Ph.D. Thesis, Dept. Math. and Stat., Arizona State University.
- [13] Laha, R. and Rohatgi, V., 1979. *Probability Theory*. Wiley, New York.
 - [14] Loève, M., (1948). Fonctions aléatoires du second ordre, Supplement to P. Lévy *Processus Stochastiques et Mouvement Brownien*, Gauthier-Villars, Paris.
 - [15] Nashed, M. and Wahba, G., 1974. Convergence rates of approximate least squares solutions of linear integral operator equations of the first kind. *Math. Comp* **28**, 69–80.
 - [16] Parzen, E., 1961a. An approach to time series analysis, *Ann. Math. Statist.* **32**, 951–989.
 - [17] Parzen, E., 1961b. Regression analysis of continuous parameter time series, in *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, 469–489, University of California Press, Berkeley.
 - [18] Parzen, E., 1963. A new approach to the synthesis of optimal smoothing and prediction problems. In *Mathematical Optimization Techniques*, 75–108, University of California Berkeley Press, Berkeley.
 - [19] Parzen, E., 1970. Statistical inference on time series by RKHS methods, In *12th Biennial Seminar Canadian Mathematical Congress Proc.*, R. Pyke, ed., Canadian Mathematical Congress, Montreal, 1-37.
 - [20] Ramsay, J. O. and Silverman, B. W., 2005. *Functional Data Analysis, 2nd Edition*. Springer-Verlag, New York.
 - [21] Rao, C. R., 1955. Estimation and tests of significance in factor analysis. *Psychometrika* **20**, 93–111.
 - [22] Shin, H., 2008. An extension of Fisher’s discriminant analysis for stochastic processes, *J. Multivariate Anal.* **99**, 1191–1216.