Journal of Statistical Planning and Inference 140 (2010) 3627–3637

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

An RKHS framework for functional data analysis

Ana Kupresanin [a], Hyejin Shin [b], David King [a], R.L. Eubank [a,*]

[a] Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287-1804, USA
[b] Mathematics and Statistics, Auburn University, Auburn, AL 36849-5310, USA

ARTICLE INFO

Available online 20 May 2010

MSC:
primary, 62H30;
62M99

Keywords:
Best linear prediction
Classification
Discriminant analysis
Factor analysis
H-valued random variable

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.

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 Loéve–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 analysis (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.

* Corresponding author. Tel.: +1 480 965 3724; fax: +1 480 965 8119.
E-mail addresses: amk@math.la.asu.edu (A. Kupresanin), hjshin@auburn.edu (H. Shin), dbking@asu.edu (D. King), eubank@math.asu.edu (R.L. Eubank).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2010.04.030

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 consideration 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 $X:\Omega\to\mathcal{H}$ 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 (Laha and Rohatgi, 1979, Remark 7.1.2). 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

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

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 $S_X:\mathcal{H}\to\mathcal{H}$ that satisfies

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

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

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

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

where $\lambda_1>\lambda_2>\cdots$ 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

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

equipped with the inner product $\langle a_1,a_2\rangle_{L_U^2}=\mathrm{E}[a_1a_2]$. 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

$$\mathrm{Cov}(U(f_1),U(f_2))=\langle f_1,S_Xf_2\rangle_{\mathcal{H}}:=K_U(f_1,f_2)$$

for $f_1,f_2\in\mathcal{H}$. The Moore–Aronszajn theorem (Aronszajn, 1950, Section 2) 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

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

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; 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

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