An RKHS Framework for Functional Data Analysis

Ana Kupresanin Hyejin Shin David King Randy Eubank

Arizona State University

Auburn University

August 1, 2009

As such the data for the ith sample can be interpreted as a discretized realization $h_{i,k} = h_i(s_k)$ of a stochastic predictor process $h(s)$ evaluated at discrete values s_k of a (continuous) index domain. We assume (a) that the jth class is observed with relative frequency $\pi_j = P[G = j]$ and that (b) its predictor process $h(s)$ has a class-specific mean function $\mu_j(s) = E[h(s)|G = j]$ and (c) a covariance function $\Sigma(s,t) = E[\{h(s) - \mu_j(s)\}\{h(t) - \mu_j(t)\}|G = j]$ that is shared among the classes.

The functional canonical variate problem in discriminant analysis consists of finding normalized functions $\beta_k(s)$ such that the associated functionals $\eta_k = \int \beta_k(s)h(s)\,ds$ have means that are optimally separated among the classes, that is, $\Sigma_{j=1}^{J} \pi_j E(\eta_k|G = j)^2$ is maximized. The normalization has the form $\int\int \beta_k(s)\Sigma(s,t)\beta_l(t)\,ds\,dt = \delta_{kl}$.

For the functional classification problem in discriminant analysis, one assumes that the predictor processes are Gaussian. The Bayes optimal classifies assigns a predictor function $h(s)$ to the class j that minimizes $d(h, \mu_j) - 2\log \pi_j$, where the Mahalanobis distance d is defined as

$$d(h, \mu_j) = \int\int \big[h(s) - \mu_j(s)\big]\Sigma^{-1}(s,t)\big[h(t) - \mu_j(t)\big]\,ds\,dt.$$

For this to be meaningful, one has to assume that an inverse Σ^{-1} of the covariance operator $\beta(\cdot) \mapsto \int \Sigma(\cdot, t)\beta(t)\,dt$ exists, but even if it exists, it generally cannot be represented by a kernel $\Sigma^{-1}(s,t)$. The double integral in the definition of $D(h, \mu_j)$ therefore has to be taken with a grain of salt.

Hastie, Buja, Tibshirani 1995, Penalized Discriminant Analysis, *Ann. Statist.*

Outline

- $\{\Omega, \mathcal{A}, P\}$ a probability space
- $\mathcal{H}$ – a real, separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $|| \cdot ||_{\mathcal{H}}$
- $\mathcal{B}$ – the σ-field generated by the class of all open subsets of $\mathcal{H}$
- $X : \Omega \to \mathcal{H}$ is called an *$\mathcal{H}$-valued random variable* if X is $\mathcal{B}$-measurable
- In particular $\{U(f) : f \in \mathcal{H}\}$ with

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

is a real valued, Hilbert space indexed, stochastic process

- P_X – probability measure X induces on $\mathcal{H}$
- assume $\int_{\mathcal{H}} \|f\|^2 dP_X(f) < \infty$
- assume $\mathrm{E}[U(f)] = 0$ for all $f \in \mathcal{H}$

Then, the covariance operator for X is the unique linear operator $S_X : \mathcal{H} \to \mathcal{H}$ that satisfies

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

for all $f, g \in \mathcal{H}$. S_X is compact with representation

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

- $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ – eigenvalues
- $\{\phi_j\}_{j=1}^{\infty}$ – eigenfunctions
- $(g \otimes_{\mathcal{H}} f)h = \langle g, h\rangle_{\mathcal{H}} f$

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

$$\langle a_1, a_2 \rangle_{L^2_U} = \mathrm{E}[a_1 a_2]$$

The completion of this space, L^2_U, can be viewed as the set of all linear combinations of the members of $\{U(\phi_j) : j \in \mathbb{Z}^+\}$.

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

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

for $f_1, f_2 \in \mathcal{H}$.

There is a unique reproducing kernel Hilbert space (RKHS) associated with K_U that (Parzen 1970) can be characterize as

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

There is a 1-1, norm-preserving, linear (congruence) map Ψ_U that maps $\mathcal{H}(K_U)$ onto L_U^2 (e.g., Loève 1948 and Parzen 1961a). The Karhunen–Loève expansion and Theorem 4D of Parzen (1961a) give

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

Theorem

L^2_U consists of random variables that are linear combination of the $U(\phi_i)$ with coefficients that satisfy $\sum_{j=1}^{\infty} \lambda_j f_j^2 < \infty$.

Corollary

Not all random variables in L^2_U can be expressed as $\langle X, f \rangle_{\mathcal{H}}$ for some $f \in \mathcal{H}$.

Outline

- $\mathcal{H} = L^2(T)$, where T is an interval subset of the line,

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

and $||f_1||^2_{L^2(T)} = \langle f_1, f_1 \rangle_{L^2(T)}$

- $\mathcal{H}$-valued random variable now correspond to the stochastic process $\{X(t) : t \in T\}$ with covariance kernel

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

- $(S_X f)(t) = \int_T f(s) K_X(t,s) d\nu(s)$

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 $\langle a_1, a_2 \rangle_{L_X^2} = \mathrm{E}[a_1 a_2]$. The Hilbert space spanned by $X(\cdot)$, L_X^2, is the completion of this pre-Hilbert space. It represents the (most?) natural extension of the finite dimensional linear combinations of random variables from multivariate analysis.

The covariance kernel K_X generates the RKHS $\mathcal{H}(K_X)$

$$\mathcal{H}(K_X) = \left\{ f : f = \sum_{j=1}^{\infty} \lambda_j f_j \phi_j, \sum_{j=1}^{\infty} \lambda_j f_j^2 < \infty \right\}$$

which is congruent to $\mathcal{H}(K_U)$.

The congruence mapping between L^2_X and $\mathcal{H}(K_X)$ is

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

with $\sum_{j=1}^{\infty} \lambda_j f_j^2 < \infty$.

Theorem

$L^2_U = L^2_X$

Corollary

When $\mathcal{H} = L^2(T)$ not all random variables in L^2_U can be expressed as $\int_T X(t) f(t) d\nu(t)$ for some $f \in \mathcal{H}$.

$\mathcal{H}(K_X)$ and $Im(S_X)$

Nashed and Wahba (1974) show that $\mathcal{H}(K_X)$ is the range of $S_X^{1/2}$ which is a subset of $L^2(T)$. Although the range of $S_X^{1/2}$ is not closed in $L^2(T)$ it becomes closed under the RKHS norm.

Finiteness of the $\mathcal{H}(K_X)$ norm is equivalent to Picard's condition which is necessary and sufficient for existence of a solution $S_X^{1/2}g = f$. Among other things this allows us to write

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

with no ambiguity as to the meaning of $S_X^{-1/2}$.

Outline

- $\{X(t), t \in T\}$, $\{Y(s), s \in T\}$ – zero mean stochastic processes with covariance functions K_X and K_Y and cross-covariance function

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

- L_X^2 , L_Y^2 – the Hilbert spaces spanned by X and Y
- $\mathcal{H}(K_X)$, $\mathcal{H}(K_Y)$ – the RKHSs
- Ψ_X and Ψ_Y – the congruence mappings

The first canonical correlation ρ_1 is

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

But,

$$\mathrm{Cov}(\Psi_X(f), \Psi_Y(g)) = \langle f, Rg \rangle_{\mathcal{H}(K_X)},$$

where $R : \mathcal{H}(K_Y) \mapsto \mathcal{H}(K_X)$ is defined by

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

Since

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

R has a representation as $S_X^{-1/2} S_{XY} S_Y^{-1/2}$ with S_{XY} the integral operator corresponding to the cross-covariance kernel.

The operator R is Hilbert-Schmidt. As in the finite dimensional case (Kshirsagar 1972) the canonical variables and correlations are obtained from the singular value decomposition

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

where

- g_j, f_j are the eigenfunctions of R^*R and RR^*
- $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq 0$ are the eigenvalues of R^*R and RR^*

The ρ_i are the canonical correlations while the canonical variables are $\Psi_X(f_i)$ and $\Psi_Y(g_i), i = 1, 2, \ldots$

Related approaches:

- He, Müller and Wang (2003) – works under restrictions on the singular values.
- Dauxois and Pousse (1976) contend that they provide

  > *a most general possible definition and formulation of the notion of canonical correlation*

  Their approach, along with others from the "French school" (e.g., Dauxois, Nkiet and Romain 2004), works only when $Im(S_X^{1/2}), Im(S_Y^{1/2})$ are closed in $L_2(T)$, which is basically Hotelling's original result. In general, the canonical variables cannot be obtained from this approach (Cupidon, et al. 2007).

Outline

Best Linear Prediction

Theorem

(Parzen 1963)The best linear predictor of $Y(t)$ *given* $\{X(s), s \in T\}$ *is* $\hat{Y}(t) = \Psi_X(K_{XY}(\cdot, t))$.

Corollary

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

Factor Analysis

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

- Y, X have covariance kernels K_X, K_Y
- the noise process N has covariance kernel K
- the signal X is uncorrelated with the noise
- all processes are $L^2(T)$ valued and have zero means

We find that $K_{XY} = K_X, K_Y = K_X + K$ and the squared canonical correlations and singular functions for the Y space are solutions to

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

This is equivalent to a result of Rao (1955) that produces his canonical correlation factor model.

Discriminant Analysis/Manova

- J populations
- The Y process is $\{Y(j) : j = 1, \ldots, J\}$ with $Y(j)$ an indicator variable that is 1 when X derives from the jth population
- $\mathrm{E}[X(\cdot)|$ population $j] = \mu_j(\cdot), j = 1, \ldots, J$ do not all coincide

Theorem

(Parzen 1961b) If $E[X(\cdot)] = \mu_X(\cdot) \in \mathcal{H}(K_X)$ there exists a one-to-one linear map Ψ_X mapping from $\mathcal{H}(K_X)$ onto L^2_X that satisfies

- $\Psi_X(K_X(\cdot, t)) = X(t), \;\; t \in T$
- $E[\Psi_X(f)] = \langle f, \mu_X \rangle_{\mathcal{H}(K_X)}$
- $Cov(\Psi_X(f_1), \Psi_X(f_2)) = \langle f_1, f_2 \rangle_{\mathcal{H}(K_X)}$.

The operator R can be expressed as

$$(Rg)(t) = \sum_{j=1}^{J} g(j)\mu_j(t)$$

with $\sum_{j=1}^{J} g(j) = 0$. Thus, all the mean functions coincide if and only if the R singular values $\rho_1, \ldots, \rho_{J-1}$ vanish. Testing the functional ANOVA hypothesis that $\mu_1(\cdot) = \cdots = \mu_J(\cdot)$ is equivalent to the hypothesis that all the canonical correlations are zero. The canonical X and Y variables may be used for classification purposes as detailed in Shin (2008).

Outline

References I

- Cupidon, J., Eubank, R., Gilliam, D. and Ruymgaart, F. (2007). Some properties of canonical correlations and variates in infinite dimensions. *J. Multivar. Anal.* **99**, 1083–1104.
- Dauxois, J., Nkiet, G. and Romain, Y. (2004). Canonical analysis relative to a closed subspace, *Linear Algebra Appl.* **388**, 119–145.
- Dauxois, J. and Pousse, A. (1976). Une extension de l'analyse canonique. quelques applications, *Ann. Inst. H. Poincaré Probab. Statist.* **11**, 355–379.
- Eubank, R. and Hsing, T. (2008). Canonical correlation for stochastic processes. *Stochastic Process. Appl.* **118**, 1634–1661.
- 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.
- Kshirsagar, A. (1972). *Multivariate Analysis*. Marcel-Dekker, New York.

References II

- Loève, M. (1948). Fonctions aléatoires du second ordre, Supplement to P. Lévy *Processes Stochastiques et Mouvement Brownien*, Gauthier-Villars, Paris.
- Parzen, E. (1961a). An approach to time series analysis, *Ann. Math. Statist.* **32**, 951–989.
- 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.
- 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.
- Parzen, E. (1970). Statistical inference on time series by RKHS methods. In *12th Biennial Seminar Canadian Mathematical Congress Proc.*, pp. 1–37, R. Pyke, ed. Canadian Mathematical Congress, Montreal.

References III

- Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika* **20**, 93–111.
- Shin, H. (2008). An extension of Fisher's discriminant analysis for stochastic processes, *J. Multivariate Anal.* **99**, 1191–1216.