Estimating the Mean and Covariance Structure Nonparametrically When the Data are Curves

By John A. Rice and B. W. Silverman

Data Feature.

The basic observed responses are curves.

eg. Growth curve

Studies of walking : the motion of a particular joint

Mouse wheel-running, food consumption and body mass curves

General Set-up.

Suppose we have independent data:

$$\mathbf{X}_i = (\mathbf{X}_i(t_1), \cdots, \mathbf{X}_i(t_p))^T, i = 1, \cdots, n$$

where

$$\mathbf{X}_i(t_j) = \mathbf{X}_{ij}, \qquad j = 1, \cdots, p$$

with $E(X_i(t)) = \mu(t)$ and $Cov(X_i(s), X_i(t)) = \gamma(s,t)$
i.e. the sample curves are modelled as being independent realizations of a stochastic process $X(t)$

Goal.

Nonparametrically estimate the mean curve $\mu(t)$ and covariance structure.

Assume that there is an orthogonal expression in the L^2 sense of γ in terms of eigenvalues and eigenfunctions

$$\gamma(s,t) = \sum_{\nu} \gamma_\nu \ \phi_\nu(s) \ \phi_\nu(t)$$

Then a random curve from the population may be expressed as

$$X(t) = \mu(t) + \sum_{\nu} \xi_\nu \ \phi_\nu(t)$$

where ξ_ν are uncorrelated random variables with

$$E(\xi_\nu) = 0, \qquad Var(\xi_\nu) = \gamma_\nu .$$

Estimate the mean curve $\mu(t)$.

First ignore the time sequence t_j and consider the estimation of the vector $\mu = (\mu(t_j))$

— A standard problem in Multivariate analysis

$$\mathbf{X}_i = \mu + \epsilon_i, \qquad i = 1, \cdots, n$$

where ϵ_i are independent and

$$E(\epsilon_i) = \mathbf{0}, \qquad Var(\epsilon_i) = \Gamma = (\gamma(t_j, t_k))_{p \times p}$$

Gauss-Markov Theory.

Regardless of Γ, the generalize least squares estimate of μ

$$\hat{\mu} = \bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$$

Assume that μ is a smooth curve

One natural smoothing approach: use penalized least squares

Estimate $\mu(t)$ by the curve $\hat{\mu}$ that minimize

$$\frac{1}{n} \sum_i \left|\left|\mathbf{X}_i - \mu\right|\right|^2 + \alpha \int (\mu'')^2 dt$$

$$\Longleftrightarrow \quad \hat{\mu} \text{ minimizes}$$

$$\left|\left|\bar{\mathbf{X}} - \mu\right|\right|^2 + \alpha \int (\mu'')^2 dt$$

where a smoothing parameter $\alpha > 0$.

i.e. $\hat{\mu}$ is a natural cubic spline with knots $t_1, \cdots, t_p$.

Choice of α.

An popular automatic method of choosing α — **CV**

It is much more natural to leave out an entire observed curve at a time.

For each k,

$$\bar{\mathbf{X}}^{-k} = (n-1)^{-1} \sum_{i \neq k} \mathbf{X}_i$$

Let $\hat{\mu}^{-k}$ be the estimate of μ such that

$$\begin{aligned}\hat{\mu}^{-k} &= \arg\min_{\mu} \left|\left|\bar{\mathbf{X}}^{-k} - \mu\right|\right|^2 + \alpha \int (\mu'')^2 dt \\ &= \arg\min_{\mu} \sum_{j=1}^{p} \{\bar{\mathbf{X}}_j^{-k} - \mu(t_j)\}^2 + \alpha \int (\mu'')^2 dt\end{aligned}$$

Then the cross-validation score

$$
\begin{aligned}
S(\alpha) &= \sum_{k=1}^{n} \left|\left|\mathbf{X}_k - \hat{\mu}^{-k}\right|\right|^2 \\
&= \sum_{k=1}^{n} \sum_{j=1}^{p} \{\mathbf{X}_{kj} - \hat{\mu}^{-k}(t_j)\}^2
\end{aligned}
$$

Choose α to minimize $S(\alpha)$.

Advantage of leaving out whole curves rather than individual data points when constructing $S(\alpha)$.

Define the summed mean square error

$$
M_n(\alpha) = \mathbf{E} \sum_j \{\hat{\mu}(t_j) - \mu(t_j)\}^2
$$

where $\hat{\mu}$ is obtained from a sample of n indept curves $\mathbf{X}_i$.

$$
\begin{aligned}
\mathbf{E} n^{-1} S(\alpha) \quad &= \mathbf{E}\, n^{-1} \sum_{k=1}^{n} \sum_{j=1}^{p} \{\mathbf{X}_{kj} - \hat{\mu}^{-k}(t_j)\}^2 \\
&= \sum_{j=1}^{p} \mathbf{E}\{\mathbf{X}_{nj} - \hat{\mu}^{-n}(t_j)\}^2 \\
&= \sum_{j=1}^{p} \mathbf{E}\{\mathbf{X}_{nj} - \mu(t_j)\}^2 + \sum_{j=1}^{p} \mathbf{E}\{\hat{\mu}^{-n}(t_j) - \mu(t_j)\}^2 \\
&\quad -2 \sum_{j=1}^{p} \mathbf{E}\{\mathbf{X}_{nj} - \mu(t_j)\}\{\hat{\mu}^{-n}(t_j) - \mu(t_j)\} \\
&= tr(\Gamma) + M_{n-1}(\alpha)
\end{aligned}
$$

Remarks.

- Regardless of the covariance function γ, the α that minimizes $n^{-1}\ E\ S(\alpha)$ will be the minimizer of $M_{n-1}(\alpha)$.
  For moderate n, the α minimizing $M_{n-1}(\alpha)$ will be close to achieving a minimum of $M_n(\alpha)$.

- The approach does not depend on the covariance.
  In contrast, Hart and Wehrly (1986) construct an estimate of $M_n(\alpha)$ based on a prior estimate of the covariance.
  Diggle and Hutchison (1989) assume a parametric form for the covarinace.

Analysis of Variability: Univariate Case.

Estimate the eigenfunctions of the covariance kernel γ.

Resons for interest in the eigenfunctions and eigenvalues:

- they can often be interpreted physically, giving insight into, and a simple description of, the large-scale features of variability

- The first few eigenfunctions and eigenvalues can be used for data reduction.

- The scores of individual curves on the leading eigenfunctions can be used for description, clustering, classification and prediction.

Some Assumptions.

- The mean curve has been subtracted from the individual curves.

- WLOG, the eigenvectors are normalized to have unit length.

Nonparametric procedure.

Estimation of the leading eigenfunction $\mathbf{u}_1$ of γ.

$\mathbf{G}$ — Sample covariance matrix

Bounding the roughness of $\mathbf{u}_1$, a class of estimate is obtained by

$$\max \ \mathbf{u}^T \, \mathbf{G} \, \mathbf{u}$$
$$\text{subject to } \ \|\mathbf{u}\| = 1 \ \text{ and } \ \mathbf{u}^T \, \mathbf{D} \, \mathbf{u} \leq \beta$$

where β : a smoothing parameter

$\mathbf{D}$: roughness matrix,

eg. $\mathbf{F}$: a second-differencing operator and $\mathbf{D} = \mathbf{F}^T \mathbf{F}$.

An equivalent formulation:

$$\hat{\mathbf{u}}_1 = \arg\max_{\|\mathbf{u}\|=1} \; \mathbf{u}^T \left(\mathbf{G} - \lambda\,\mathbf{D}\right) \mathbf{u}$$

where λ is a smoothing parameter.
i.e. $\hat{\mathbf{u}}_1$ is the leading eigenvector of $\mathbf{G} - \lambda\,\mathbf{D}$.

Understanding λ.
The smallest eigenvalues of $\mathbf{D}$ have associated eigenfunctions which are smooth. As λ increases, the solution $\mathbf{u}_1(\lambda)$ is pulled towards smooth functions, because it will tend to lie in the part of the space spanned by eigenfunctions of $\mathbf{D}$ with small eigenvalues.

An alternative characterization of the leading eigenfunction $\mathbf{u}_1(0)$ of $\mathbf{G}$

$$\mathbf{u}_1(0) = \arg\min_{\mathbf{u}} \sum_{i=1}^{n} \|\mathbf{X}_i - (\mathbf{X}_i^T\ \mathbf{u})\ \mathbf{u}\|^2$$

i.e. $\mathbf{u}_1(0)$ can be regarded as a best predictor, in that the individual realization $\mathbf{x}$ obtained from the population of curves can each be well modelled as a multiple of $\mathbf{u}_1(0)$.

Choice of λ: CV method

$\mathbf{G}^{(-i)}$: the covariance matrix of the data leaving the ith curve out.

$\mathbf{u}_1^{(-i)}(\lambda)$: the leading eigenvector of $\mathbf{G}^{(-i)} - \lambda\, D$. Then

$$\mathbf{CV}(\lambda) = \sum_{i=1}^{n} \|\mathbf{X}_i - \{\mathbf{X}_i^T\ \mathbf{u}_1^{(-i)}(\lambda)\}\ \mathbf{u}_1^{(-i)}(\lambda)\|^2$$

Choose λ^* to minimize $\mathbf{CV}(\lambda)$.

Understanding λ^*.

Sample $\mathbf{X}_1, \cdots, \mathbf{X}_n$ i.i.d $\sim (\mathbf{0}, \Gamma)$.

$\hat{\mathbf{u}}$: estimate of the leading eigenvector of Γ based on the sample of size n.

$\mathbf{X}$ indept of $\mathbf{X}_1, \cdots, \mathbf{X}_n$.

Define the mean-square prediction error

$$L_n(\hat{\mathbf{u}}) = \mathbf{E}_{\mathbf{X}} \|\mathbf{X} - (\mathbf{X}^T\ \hat{\mathbf{u}})\hat{\mathbf{u}}\|^2 = tr(\Gamma) - \hat{\mathbf{u}}^T\ \Gamma\ \hat{\mathbf{u}}$$

and

$$R_n(\lambda) = \mathbf{E}[L_n\{\hat{\mathbf{u}}_1(\lambda)\}]$$

Then it is straightforward to show that

$$\mathbf{E}n^{-1}\mathbf{CV}(\lambda) = R_{n-1}(\lambda)$$

i.e. for moderate n, λ^* is hopefully close to the minimizer of $R_n(\lambda)$.

Estimate the higher order eigenfunctions.

Suppose that smooth orthogonal eigenfunctions $\mathbf{u}_1, \cdots, \mathbf{u}_k$ have been found. Let $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$ be orthogonal matrix with $\mathbf{Q}_1 = (\mathbf{u}_1, \cdots, \mathbf{u}_k)$. Then

$$\mathbf{u}_{k+1}(\lambda) = \underset{\|\mathbf{u}\|=1,\ \mathbf{Q}_1^T \mathbf{u}=0}{\arg\max}\ \mathbf{u}^T\ (\mathbf{G} - \lambda\ \mathbf{D})\ \mathbf{u}$$

Let $\hat{\mathbf{v}}(\lambda)$ be the leading eigenvector of $\mathbf{Q}_2^T\ (\mathbf{G} - \lambda\ \mathbf{D})\ \mathbf{Q}_2$, i.e.

$$\mathbf{v}(\lambda) = \underset{\|\mathbf{v}\|=1}{\arg\max}\ \mathbf{v}^T\ \mathbf{Q}_2^T\ (\mathbf{G} - \lambda\ \mathbf{D})\ \mathbf{Q}_2\ \mathbf{v}$$

Then $\mathbf{u}_{k+1}(\lambda) = \mathbf{Q}_2\ \mathbf{v}(\lambda)$.

Choice of λ.

Exactly the same arguements as before yield the **CV** score

$$\mathbf{CV}_{k+1}(\lambda) = \sum_{i=1}^{n} \|\mathbf{Q}_2^T\, \mathbf{X}_i - \{(\mathbf{Q}_2^T\, \mathbf{X}_i)^T\, \mathbf{v}^{(-i)}(\lambda)\}\, \mathbf{v}^{(-i)}(\lambda)\|^2$$

where $\mathbf{v}^{(-i)}(\lambda)$ is the leading eigenvector of $\mathbf{Q}_2^T\, (\mathbf{G} - \lambda\, \mathbf{D})\, \mathbf{Q_2}$ by leaving the ith curve out.

Substituting $\mathbf{u}_{k+1}^{(-i)}(\lambda) = \mathbf{Q}_2\, \mathbf{v}^{(-i)}(\lambda)$, we obtain

$$\mathbf{CV}_{k+1}(\lambda) = \sum_{i=1}^{n} \|\mathbf{X}_i - \{\mathbf{X}_i^T\, \mathbf{u}_{k+1}^{(-i)}(\lambda)\}\, \mathbf{u}_{k+1}^{(-i)}(\lambda)\|^2$$

Analysis of Variability : Multivariate Case.

Motivation. There are several curves per subject in some data sets.

eg. In the data set, hip and knee curves per child

Objective. Explore methods to describing and analyzing the relationships between the curves.

eg. How the knee and hip move jointly.

Methodology.

Taking into account the dependentce between them, we consider the closed curve $(h(t), k(t))^T = (h(t_1), \cdots, h(t_p), k(t_1), \cdots, k(t_p))^T$ in the plane with

$$\mathbf{E}((h(t), k(t))^T) = (\mu_H(t), \mu_K(t))^T, \qquad Var((h(t), k(t))^T) = \Sigma$$

Estimation.

- Seperately estimate the mean hip and knee curves.

- Carry out eigenfunction analysis on the n vectors $(h(t_1), \cdots, h(t_p), k(t_1), \cdots, k(t_p))^T$. The penalty matrix is of block diagonal form so that elements of h and k are not linked, but the covariance matrix allows correlation between hip and knee angles.