The functional singular value decomposition for bivariate stochastic processes

Daniel Gervini[1]
University of Wisconsin–Milwaukee

July 23, 2009

[1] Department of Mathematical Sciences, University of Wisconsin–Milwaukee, P.O. Box 413, Milwaukee, WI 53201 (email: gervini@uwm.edu).

Abstract

In this article we present some statistical applications of the functional singular value decomposition (FSVD). This tool allows us to decompose the sample mean of a bivariate stochastic process into components that are functions of separate variables. These components are sometimes interpretable functions that summarize salient features of the data. The FSVD can be used to visually detect outliers, to estimate the mean of a stochastic process or to obtain individual smoothers of the sample surfaces. As estimators of the mean, we show by simulation that FSVD estimators are competitive with tensor-product splines in some situations.

Key Words: Eigenvalues and eigenfunctions; Functional data analysis; Outlier detection; Principal component analysis; Spectral decomposition; Spline smoothing.

1 Introduction

The analysis of samples of curves has become more common in statistical applications in recent years. In many applications, the data consists of discrete realizations of a univariate process, say $X(t)$, where t can be time (e.g. growth curves in Gasser et al., 2004), distance (e.g. biomarker expression curves in Morris and Carroll, 2006) or age (e.g. income distribution densities in Kneip and Utikal, 2001), among other possibilities. More examples and statistical methodology can be found in Ramsay and Silverman (2002, 2005) or Ferraty and Vieu (2006).

Multivariate stochastic processes, on the other hand, have received less attention. By multivariate process we mean a real-valued process $X(\mathbf{s})$ that is a function of a multidimensional variable s. They are also known as random fields (Adler and Taylor, 2007). Although they are less common in statistics than univariate processes, they play an important role in fMRI studies and spatial statistics (Taylor and Worseley, 2007; Nychka, 2000). In these applications s is a point in $\mathbb{R}^2$ or $\mathbb{R}^3$. However, in other situations the variables do not belong to a single natural space. For example, $X(s,t)$ may be the mortality rate for individuals of age s during year t in a given country, or the outcome of a multichannel electroencephalography study where t is time and s is the location of the electrode on the scalp. It is clear that the variables s and t belong to different spaces; although the product space could be regarded as a single space, this would be more a mathematical formalization than a natural structure implied by the data.

To understand more clearly the problems involved, in Fig. 1 we have plotted the sample mean of log-mortality rates for ten European countries. The raw mean shows some irregularities due to random noise. To regularize a bivariate estimator like this, one would normally employ a smoothing method based on splines (Gu, 2000) or kernels (Härdle and Müller, 2000). However, those global smoothers will most likely level off important features of the data, like the increased mortality rates during the Second World War, which are sharp but localized features.

In this paper we present a different approach, based on a generalization of the singular value decomposition. The basic idea is to approximate a bivariate function $\mu(s,t)$ with a sum of functions of separate variables, $\mu^{(p)}(s,t) = \sum_{k=1}^{p} \lambda_k^{1/2} \phi_k(s) \psi_k(t)$, where ϕ_k and ψ_k are univariate functional principal components (Silverman, 1996; Yao and Lee, 2006; Gervini, 2006). The components are sometimes interpretable

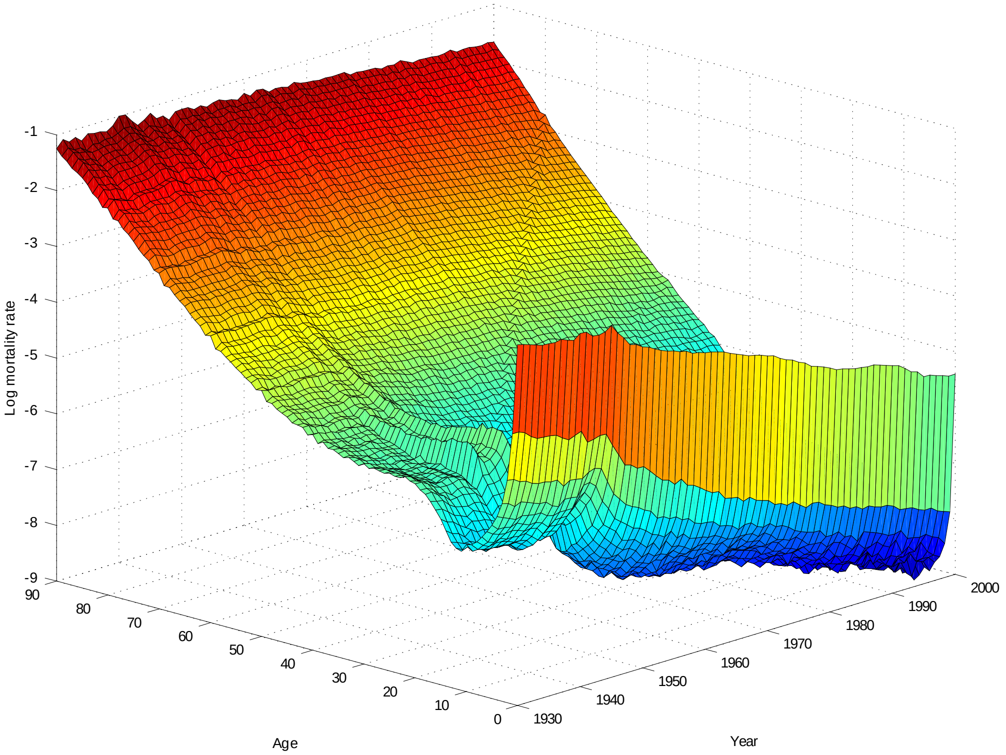

Figure 1: Human Mortality Data. Mean of log-mortality rates for ten European countries.

functions that summarize important features of the data, and can be used, for example, to detect atypical observations. Individual smoothers of the sample surfaces can also be obtained as by-products. The bivariate singular value decomposition has been used in image analysis and physics (Dente et al., 1996; Aubry et al., 1991), under the name of "biorthogonal decomposition". However, these articles disregard smoothing issues, using raw principal components for estimation. In most statistical applications, that would lead to extremely noisy and uninformative estimates. In contrast, the method we present here produces smooth and regular estimators.

This article is organized as follows. The functional singular value decomposition (FSVD) is presented in Section 2, and smooth estimators of the components are introduced in Section 3. An application to a real dataset in Section 4 illustrates the potential of the FSVD as a graphical tool. In Section 5 we compare by simulation the behavior of the FSVD with tensor-product splines as estimators of the mean. Abbreviated proofs of the theorems are given in the Appendix; more detailed proofs and additional material is available on a Technical Report that will be posted on the author's website.

2 The functional singular value decomposition

Let $X(s,t)$ be a real-valued stochastic process in $L^2(\mathcal{S}\times\mathcal{T})$ with finite expectation $\mu(s,t)$ and finite covariance function $\rho\{(s_1,t_1),(s_2,t_2)\}$. We assume that $\mathcal{S}$ and $\mathcal{T}$ are closed intervals in $\mathbb{R}$. Let us define the kernel functions

$$k_1(s_1,s_2)=\int_{\mathsf{T}}\mu(s_1,t)\mu(s_2,t)\,dt$$

and

$$k_2(t_1,t_2)=\int_{\mathsf{S}}\mu(s,t_1)\mu(s,t_2)\,ds.$$

We say that $\phi\in L^2(\mathcal{S})$ is an eigenfunction of k_1 with eigenvalue λ if $\int_{\mathsf{S}}k_1(s,u)\phi(u)du=\lambda\phi(s)$ for almost every $s\in\mathcal{S}$. The eigenfunctions of k_2 are defined in a similar way, only that they belong to $L^2(\mathcal{T})$. The next theorem establishes the existence of a decomposition of k_1, k_2 and μ in terms of these eigenfunctions.

Theorem 1 *There exist a non-increasing sequence of positive eigenvalues $\{\lambda_k\}$ of k_1 and k_2, an orthonormal sequence $\{\phi_k\}$ of eigenfunctions of k_1 and an orthonormal sequence $\{\psi_k\}$ of eigenfunctions of k_2 such that*

$$k_1(s_1, s_2) = \sum_{k\geq 1} \lambda_k \phi_k(s_1)\phi_k(s_2), \tag{1}$$

$$k_2(t_1, t_2) = \sum_{k\geq 1} \lambda_k \psi_k(t_1)\psi_k(t_2) \tag{2}$$

and

$$\mu(s, t) = \sum_{k\geq 1} \lambda_k^{1/2} \phi_k(s)\psi_k(t). \tag{3}$$

The series (1), (2) and (3) converge in the sense of the L^2 norm. If in addition $\mu(s, t)$ is continuous, then $\{\phi_k\}$ and $\{\psi_k\}$ are continuous functions and the convergence of (1) and (2) is absolute and uniform in both variables, with the identities holding for each (s_1, s_2) and each (t_1, t_2). If the right-hand side of (3) converges uniformly and absolutely, then the identity also holds for every (s, t).

Theorem 1 implies that the truncated series

$$\mu^{(p)}(s, t) = \sum_{k=1}^{p} \lambda_k^{1/2} \phi_k(s)\psi_k(t) \tag{4}$$

converges to $\mu(s, t)$ in the sense of $L^2(\mathcal{S} \times \mathcal{T})$ as p increases, and that the convergence is pointwise for every (s, t) if the right-hand side of (3) converges uniformly and absolutely. The latter occurs if, for instance, the ϕ_ks and the ψ_ks are uniformly bounded and $\sum_{k\geq 1} \lambda_k^{1/2}$ is finite.

In analogy with the multivariate singular value decomposition, the truncated series $\mu^{(p)}$ given by (4) provides the best possible approximation of μ among linear combinations of functions of separate variables, in the sense of the $L^2(\mathcal{S} \times \mathcal{T})$ norm.

Theorem 2 *Let $\mathcal{H}_p$ be the class of functions $h(s, t) = \sum_{k=1}^{p} a_k f_k(t) g_k(s)$ with $\{f_k\}$ and $\{g_k\}$ orthonormal in $L^2(\mathcal{T})$ and $L^2(\mathcal{S})$, respectively. Then*

$$\min_{h \in \mathsf{H}_p} \|\mu - h\|^2 = \|\mu - \mu^{(p)}\|^2,$$

with $\mu^{(p)}$ as in (4).

The function $\mu^{(p)}(s,t)$ is the sum of p functions of separate variables, $d_k(s,t) = \lambda_k^{1/2}\phi_k(s)\psi_k(t)$, that we will call "detail functions". The detail functions are orthogonal in both variables, and $\|d_k\| = \lambda_k^{1/2}$, so they provide finer levels of detail as k increases. An appealing feature of the detail functions is that they are often interpretable functions, giving us information about the most relevant characteristics of the process under investigation.

Of course, all this would be of little practical use if the computation of the ϕ_ks and ψ_ks required a good preliminary estimator of μ. But we show below that good estimators of the eigenfunctions can be obtained from the raw data, and these estimators are then used to construct a smooth estimator of μ.

3 Smooth estimation of the eigenfunctions

Let $X_1, \ldots, X_n$ be an i.i.d. sample of the process X. In most cases, the X_is are observed on a discrete grid $\{s_j\} \times \{t_k\} \subset \mathcal{S} \times \mathcal{T}$ with random error, so the data follows the model

$$x_{ijk} = X_i(s_j, t_k) + \varepsilon_{ijk},\ i = 1, \ldots, n,\ j = 1, \ldots, m,\ k = 1, \ldots, r. \tag{5}$$

We will assume that $\mathrm{E}(\varepsilon_{ijk}) = 0$, ε_{ijk} is independent of X_i, ε_{ijk} and $\varepsilon_{i'j'k'}$ are independent if $i \neq i'$, and $\mathrm{E}(\varepsilon_{ijk}\varepsilon_{ij'k'}) = \sigma^2\delta_{jj'}\delta_{kk'}$ (where δ is Kronecker's delta).

The simplest estimator of μ at the grid points is the cross sectional mean, $\hat{\mu}(s_j, t_k) = \sum_{i=1}^n x_{ijk}/n$. The corresponding estimators of the kernel functions k_1 and k_2, using the trapezoid rule for numerical integration, are

$$\hat{k}_1(s_j, s_{j'}) = \sum_{k=1}^{r} u_k \hat{\mu}(s_j, t_k)\hat{\mu}(s_{j'}, t_k)$$

and

$$\hat{k}_2(t_k, t_{k'}) = \sum_{j=1}^{m} v_j \hat{\mu}(s_j, t_k)\hat{\mu}(s_j, t_{k'}),$$

where $u_1 = (t_2 - t_1)/2$, $u_k = (t_{k+1} - t_{k-1})/2$, $k = 2, \ldots, r-1$, $u_r = (t_r - t_{r-1})/2$, and $v_1 = (s_2 - s_1)/2$, $v_j = (s_{j+1} - s_{j-1})/2$, $j = 2, \ldots, m-1$, $v_m = (s_m - s_{m-1})/2$.

From $\hat{k}_1$ and $\hat{k}_2$ we can compute smooth estimators of the eigenfunctions $\{\phi_k\}$ and $\{\psi_k\}$ using spline models (such as B-splines; de Boor, 2001) as follows. We know that

$$\phi_1 = \mathrm{argmax}_{\|g\|=1} \textstyle\iint k_1(s_1,s_2)g(s_1)g(s_2)\mathrm{d}s_1\mathrm{d}s_2.$$

Then, given a spline basis $\{\beta_1,\ldots,\beta_q\}$ in $L^2(\mathcal{S})$, we write $g(s)=\sum_{j=1}^q b_j\beta_j(s)$ and define

$$\hat{\mathbf{b}}_1 = \mathrm{argmax}\{\mathbf{b}^T\hat{\mathbf{\Omega}}\mathbf{b} : \mathbf{b}^T\mathbf{\Gamma}\mathbf{b}=1\},$$

where $\hat{\Omega}_{ij} = \iint \hat{k}_1(s_1,s_2)\beta_i(s_1)\beta_j(s_2)\mathrm{d}s_1\mathrm{d}s_2$ and $\Gamma_{ij}=\int \beta_i(s)\beta_j(s)\mathrm{d}s$. Then $\hat{\phi}_1(s) = \sum_{j=1}^q \hat{b}_{1j}\beta_j(s)$ is a spline estimator of the first eigenfunction of k_1.

For the rest of the eigenfunctions we proceed sequentially: since

$$\phi_k = \mathrm{argmax}\left\{\textstyle\iint k_1(s_1,s_2)g(s_1)g(s_2)\mathrm{d}s_1\mathrm{d}s_2 : \|g\|=1 \text{ and } \langle g,\phi_j\rangle = 0 \text{ for } j<k\right\},$$

we define

$$\hat{\mathbf{b}}_k = \mathrm{argmax}\{\mathbf{b}^T\hat{\mathbf{\Omega}}\mathbf{b} : \mathbf{b}^T\mathbf{\Gamma}\mathbf{b}=1, \mathbf{b}^T\mathbf{\Gamma}\hat{\mathbf{b}}_j=0, j<k\} \qquad (6)$$

and set $\hat{\phi}_k(s)=\sum_{j=1}^q \hat{b}_{kj}\beta_j(s)$. The corresponding eigenvalues can be estimated by $\hat{\lambda}_k = \hat{\mathbf{b}}_k^T\hat{\mathbf{\Omega}}\hat{\mathbf{b}}_k$.

Computationally, (6) is a very simple problem. Let $\mathbf{V}=\mathrm{diag}(v_1,\ldots,v_m)$, $\mathbf{B}\in\mathbb{R}^{q\times m}$ with $B_{ij}=\beta_i(s_j)$, and $\mathbf{K}_1\in\mathbb{R}^{m\times m}$ with $K_{1ij}=\hat{k}_1(s_i,s_j)$. Then, using the trapezoid rule for numerical integration, $\hat{\mathbf{\Omega}}=\mathbf{B}^T\mathbf{V}\mathbf{K}_1\mathbf{V}\mathbf{B}$ and $\mathbf{\Gamma}=\mathbf{B}^T\mathbf{V}\mathbf{B}$. If $\mathbf{\Gamma}^{1/2}$ denotes the symmetric square root of $\mathbf{\Gamma}$ and $\hat{\mathbf{c}}_k$ the kth unit-norm eigenvector of $\mathbf{\Gamma}^{-1/2}\hat{\mathbf{\Omega}}\mathbf{\Gamma}^{-1/2}$, then $\hat{\mathbf{b}}_k = \mathbf{\Gamma}^{-1/2}\hat{\mathbf{c}}_k$.

If the true eigenfunctions belong to the space generated by the specified spline basis, and the eigenvalues of $\mathbf{\Gamma}^{-1/2}\mathbf{\Omega}\mathbf{\Gamma}^{-1/2}$ (with $\mathbf{\Omega}$ given below) have multiplicity one, then the above estimators are consistent. This is a consequence of the next theorem together with the results of Tyler (1981).

Theorem 3 *Let* $\mathbf{\Omega}\in\mathbb{R}^{q\times q}$ *be given by* $\Omega_{ij}=\iint k_1(s_1,s_2)\beta_i(s_1)\beta_j(s_2)\mathrm{d}s_1\mathrm{d}s_2$. *If* $\max v_j\to 0$ *as* $m\to\infty$ *and* $\max u_k\to 0$ *as* $r\to\infty$, *then* $\hat{\mathbf{\Omega}}\to\mathbf{\Omega}$ *in probability as* n, m *and* r *go to infinity.*

In practice, though, the eigenfunctions may not belong to a spline space. But the asymptotic bias will be negligible if the spline basis is appropriately chosen. For that reason, in this paper we use adaptive free-knot splines as in Gervini (2006). Another possibility is to use a large number of basis functions with global regularization, as in Silverman (1996), but we prefer the free-knot approach because it provides better fits for the local features of the eigenfunctions.

Concretely, the algorithm we implemented aggregates knots by maximizing (6) over a grid of candidates (usually the grid $\{s_j\}$ itself) until there is no significant improvement on the objective function (6). Repeated knots are allowed, since they provide better resolution of the local features of the components (at the expense of fewer degrees of differentiability). The optimal number of knots can be chosen either subjectively or by cross-validation. This procedure must be repeated for each component because the optimal placement and number of knots changes with each component.

The eigenfunctions $\{\psi_k\}$ of k_2 are estimated in a similar way, using a spline basis in $L^2(\mathcal{T})$. Since the choice of sign of the eigenfunctions is always arbitrary, care must be taken so that $\hat{\lambda}_k^{1/2} = \iint \hat{\mu}(s,t)\hat{\phi}_k(s)\hat{\psi}_k(t)\mathrm{d}s\mathrm{d}t$ is positive. As before, we use the trapezoid rule for numerical integration, so $\hat{\lambda}_k^{1/2} = \hat{\phi}_k(\mathbf{s})^T\mathbf{V}\bar{\mathbf{X}}\mathbf{U}\hat{\psi}_k(\mathbf{t})$, where $\hat{\phi}_k(\mathbf{s})$ is the vector with elements $\hat{\phi}_k(s_j)$ and $\hat{\psi}_k(\mathbf{t})$ is the vector with elements $\hat{\psi}_k(t_j)$; $\bar{\mathbf{X}}$ is the average of the matrices $\mathbf{X}_i$ with elements $(X_i)_{jk} = x_{ijk}$ and $\mathbf{U} = \mathrm{diag}(u_1, \ldots, u_r)$.

The eigenfunctions are estimated sequentially until a given order p, and then we define

$$\hat{\mu}^{(p)}(s,t) = \sum_{k=1}^{p} \hat{\lambda}_k^{1/2}\hat{\phi}_k(s)\hat{\psi}_k(t).$$

The order p must be chosen with care, to reduce bias as much as possible. For reasons that will become clearer in Sections 4 and 5, we recommend to use a large p as long as the estimators of the eigenfunctions are not overwhelmed by noise, even if the corresponding $\hat{\lambda}_k$s seem to be negligibly small.

Interestingly, $\hat{\mu}^{(p)}$ can be further decomposed into terms that represent the individual contributions of the X_is, since $\hat{\lambda}_k^{1/2} = \sum_{i=1}^{n} \hat{w}_{ik}/n$ with $\hat{w}_{ik} = \hat{\phi}_k(\mathbf{s})^T\mathbf{V}\mathbf{X}_i\mathbf{U}\hat{\psi}_k(\mathbf{t})$. Note that $\hat{w}_{ik}$ is an estimator of $w_{ik} = \iint X_i(s,t)\phi_k(s)\psi_k(t)\mathrm{d}s\mathrm{d}t$. Then we can define individual predictors of the unobserved sample paths $X_i(s,t)$,

$$\hat{X}_i^{(p)}(s,t) = \sum_{k=1}^{p} \hat{w}_{ik}\hat{\phi}_k(s)\hat{\psi}_k(t).$$

The score vectors $\hat{\mathbf{w}}_i$ are useful for exploratory data analysis; for example, they may reveal outliers or unusual groupings in the data, as we show by example in

Section 4. The predictors $\hat{X}_i^{(p)}$ can also be used to select the best order p by cross-validation.

4 Example: evolution of human mortality in the 20th century

The socioeconomic progress experienced by western European countries after the Second World War is very graphically exemplified by the evolution of human mortality curves. Mortality rates, which are the percentages of people of certain age who die in a given year, can be seen as longitudinal of functional data in two senses: for a given year, mortality rates are a function of age; and for each age, the evolution of mortality rates over the years are a time series. But a thorough statistical analysis must take into account the interplay between these two variables; that is, the data must be seen as realizations of a bivariate stochastic process.

In this section we analyze mortality rates between the years of 1930 and 2000, for people ranging from 0 to 90 years of age. The data was downloaded from the Human Mortality Database website, www.mortality.org. We only included countries of western Europe for which complete data was available: Belgium, Denmark, England, Finland, France, Italy, the Netherlands, Norway, Spain and Sweden. For country i we defined $X_i(s,t)$ as the logarithm of the mortality rate for age s at year t; the data was observed on the grid $\{0,1,\ldots,90\} \times \{1930,1931,\ldots,2000\}$.

We computed three pairs of eigenfunctions, which are shown in Fig. 2. The corresponding root-eigenvalues were $\hat{\lambda}_1^{1/2} = 435.85$, $\hat{\lambda}_2^{1/2} = 11.09$ and $\hat{\lambda}_3^{1/2} = 6.71$. Clearly, the first eigenvalue is dominant. However, the second and third detail functions do improve the fit in ways that are visually noticeable (the fact that obvious visual improvements may be associated with very small eigenvalues was observed by Dente et al., 1996).

We see that $\hat{\phi}_1(s)$ (Fig. 2(a)) can be interpreted as the basic shape of a human mortality curve: high infant mortality is followed by a sharp decrease until adolescence, then a sharp increase occurs that levels off at ages 20 to 30, followed by a steady increase from then on. The companion eigenfunction $\hat{\psi}_1(t)$ (Fig. 2(b)) is the overall mortality trend over this 71-year period: a modest decrease in the early 30's was punctuated by the Second World War, followed by a remarkably fast

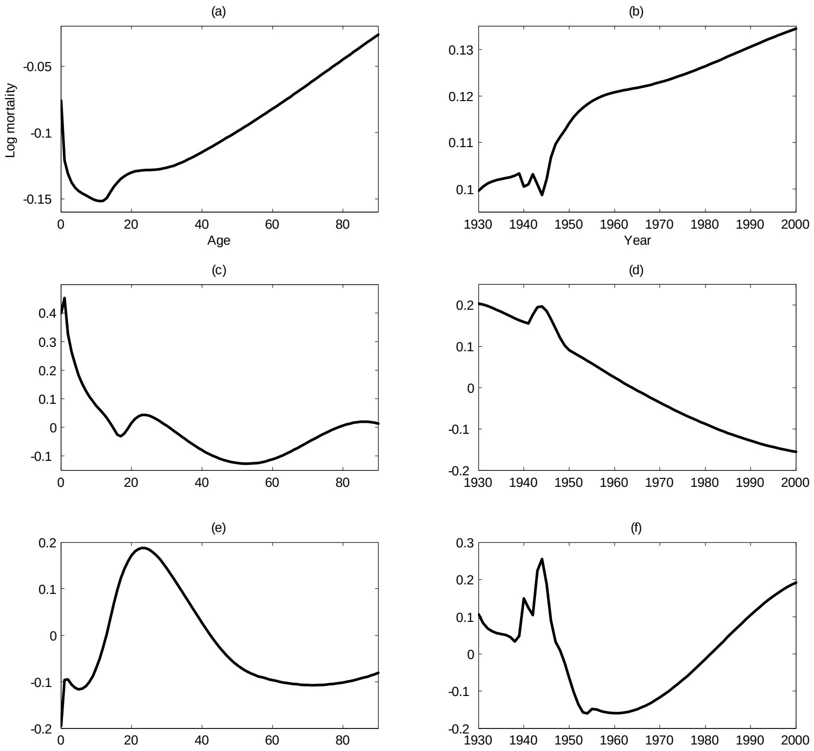

Figure 2: Human Mortality Data. Free-knot spline estimators of the eigenfunctions: (a) $\hat{\phi}_1(s)$, (b) $\hat{\psi}_1(t)$, (c) $\hat{\phi}_2(s)$, (d) $\hat{\psi}_2(t)$, (e) $\hat{\phi}_3(s)$ and (f) $\hat{\psi}_3(t)$.

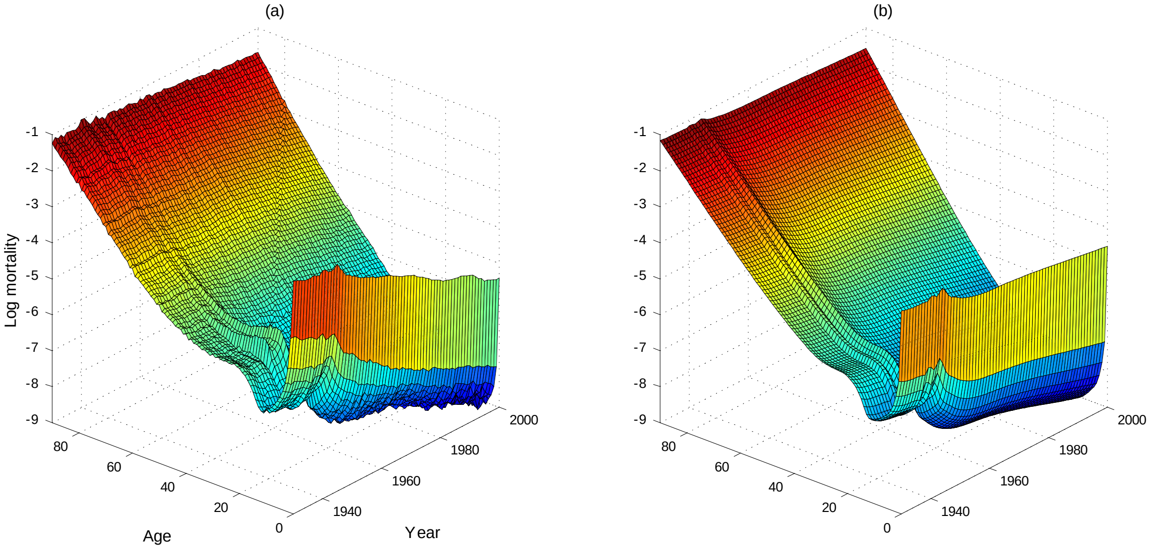

Figure 3: Human Mortality Data. (a) Raw mean and (b) first-order singular value approximation.

decrease in mortality that has continued until these days. The first-order approximation $\hat{\mu}^{(1)}$ is depicted in Fig. 3, together with the raw mean. We see that the approximation is very good, but some flaws are obvious. For example, newborn mortality ($s = 0$) remains constant over the years in Fig. 3(b) while it is obviously decreasing in Fig. 3(a).

The second component $\hat{\phi}_2(s)$ (Fig. 2(c)) is mostly related to infant mortality, with $\hat{\psi}_2(t)$ (Fig. 2(d)) showing a steady decrease over the years except for the war period. Clearly, $\hat{\mu}^{(2)}$ (Fig. 4(b)) provides a better fit for infant mortality than $\hat{\mu}^{(1)}$. The third-order approximation $\hat{\mu}^{(3)}$ (Fig. 5(b)) improves the fit for the war years. Note that for this period, $\hat{\mu}^{(2)}$ underestimates mortality for ages 20 to 30 and overestimates it for ages 60 and over. Higher levels of detail could be added, but it is hard to see any features of the raw mean that have not been accounted for by $\hat{\mu}^{(3)}$.

An analysis of individual countries also reveals interesting facts. The scatter plot of the component scores (Fig. 6) shows three points that stand apart from the rest. The most extreme case, having the smallest first-component score and the largest third-component score, is Finland. This is an unexpected result for someone

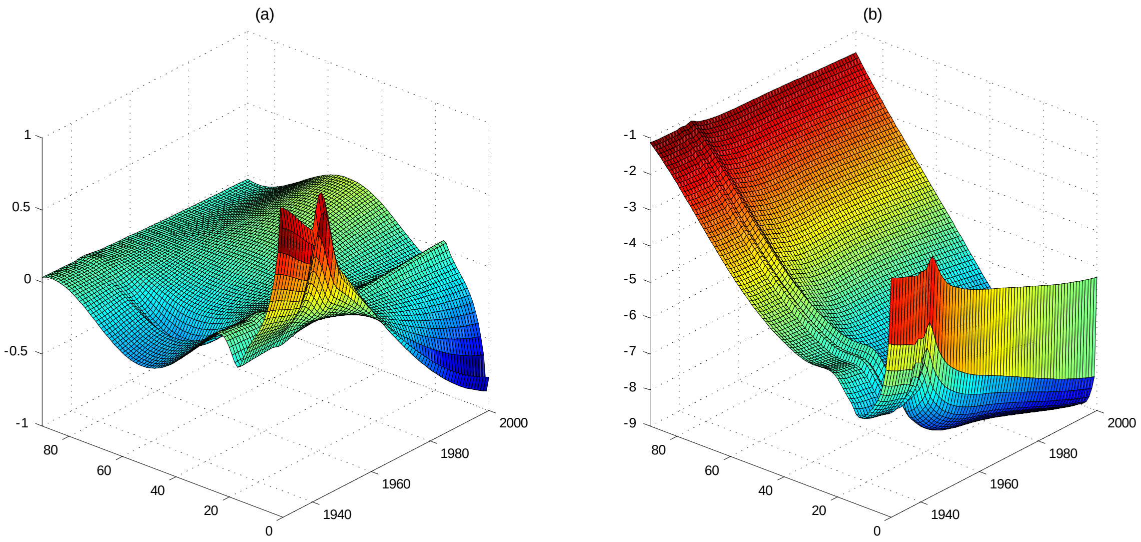

Figure 4: Human Mortality Data. (a) Second-order detail function and (b) second-order singular value approximation of the mean.

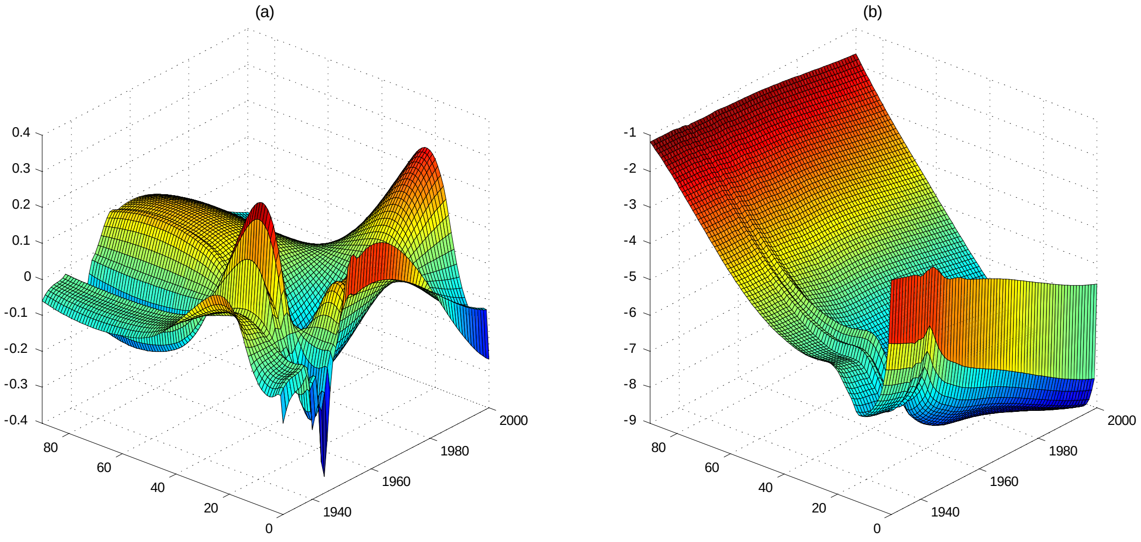

Figure 5: Human Mortality Data. (a) Third-order detail function and (b) third-order singular value approximation of the mean.

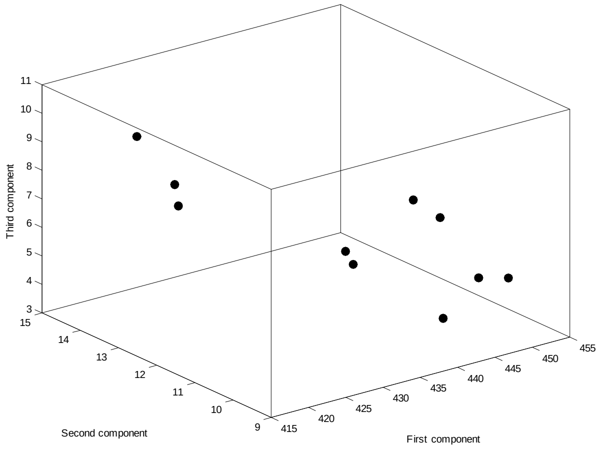

Figure 6: Human Mortality Data. Individual component scores of the ten countries.

unfamiliar with Finnish history, but it turns out that Finland was fighting on two different fronts during the war years. A quick comparison of the individual mortality plots (shown in the Technical Report) reveals that Finland, indeed, experienced the largest increase in mortality rate for the 20-40 age bracket during the war years among the countries in this sample (this is precisely what a small first-component score accompanied by a large third-component score indicates, according to our interpretation of the components).

The other two atypical points are Spain and Italy. Spain did not participate in the Second World War but went through a civil war in the 1930s, showing a different mortality pattern from the rest of the countries; in particular, the decrease in child mortality after 1945 was not as fast as for the other countries. Italy, by contrast, has the largest second-component score and is the country with the fastest post-war decrease in infant mortality.

This example illustrates the kind of insight that can be gained from the functional singular value decomposition. While other methods (like tensor-product splines) can provide estimators of the mean function, the FSVD also offers an interpretable decomposition of the mean that can reveal interesting aspects of the data.

5 Simulations

As mentioned before, we see the FSVD mainly as a tool for graphical and exploratory data analysis, but since (4) can be used as an estimator of μ, we ran a Monte Carlo study to compare its performance with that of tensor-product spline estimators. Specifically, we wanted to assess the ability of our free-knot component estimators to adapt to local features of μ, and the potential dangers of underestimating the approximation order p.

We generated data from a mean-plus-error model $x_{ijk} = \mu(s_j, t_k) + \varepsilon_{ijk}$. Two different means were considered, $\mu_1(s,t) = \sum_{k=1}^{2} \lambda_k^{1/2}\phi_k(s)\psi_k(t)$ and $\mu_2(s,t) = \sum_{k=1}^{3} \lambda_k^{1/2}\phi_k(s)\psi_k(t)$, with $\phi_k(s) = \sqrt{2}\sin(2k\pi s)$, $\psi_k(t) = \sqrt{2}\cos(2k\pi t)$, $\lambda_1 = 1$, $\lambda_2 = 1/2$ and $\lambda_3 = 1/32$. The grids $\{s_j\}$ and $\{t_k\}$ consisted of $m = r$ equispaced points in $[0,1]$, and the errors ε_{ijk} were independent $N(0,\sigma^2)$. We considered two grid sizes, $m = 20$ and $m = 30$, two sample sizes, $n = 10$ and $n = 50$, and two error variances, $\sigma^2 = 1$ and $\sigma^2 = 4$. Each model was replicated 200 times (although not all combinations of factors were considered; see Table 1).

For the tensor-product spline estimator, we took two bases of cubic B-splines with knots placed at the grid points. The estimator was regularized by penalizing the integrated squared partial derivatives, as explained in Hastie et al. (2001, ch. 5). The choice of a good smoothing parameter is crucial for the behavior of these estimators. To be as fair as possible with tensor-product splines, we chose the optimal smoothing parameter: the minimizer of $\|\hat{\mu} - \mu\|$. In practice this cannot be done because μ is unknown, so the estimation errors reported in Table 1 (under "TPS") will be lower than those attainable in practice.

As FSVD estimator of μ we took a two-component decomposition, $\hat{\mu}^{(2)}$, with $\hat{\phi}_k$s and $\hat{\psi}_k$s estimated by free-knot cubic splines, as explained in Section 3. Here the number of knots plays the role of smoothing parameter, so we considered two possibilities: a fixed number of knots (3 for ϕ_1, 5 for ϕ_2, 2 for ψ_1 and 4 for ψ_2), and an optimal number of knots (the number that minimizes $\|\hat{\phi}_k - \phi_k\|$ or $\|\hat{\psi}_k - \psi_k\|$, up to a maximum of 10 knots). The estimation errors are reported in Table 1 as "SVf" and "SVo", respectively. These two are extreme cases, so the actual estimation error of $\hat{\mu}^{(2)}$ when the number of knots is selected by the user will fall somewhere between these two.

Table 1 shows the root integrated squared errors, $E^{1/2}(\|\hat{\mu} - \mu\|^2)$. Standard

Model parameters				Root ISE		
Mean	σ	m	n	TPS	SVf	SVo
μ_1	1	20	10	.159	.111	.097
			50	.085	.075	.047
		30	10	.114	.090	.069
			50	.063	.070	.034
μ_1	2	20	10	.277	.196	.184
			50	.147	.103	.089
		30	10	.197	.140	.124
			50	.104	.086	.062
μ_2	2	20	10	.285	.264	.255
			50	.160	.205	.197
		30	10	.212	.225	.217
			50	.110	.196	.187

Table 1: Simulation Results. Root mean integrated squared errors for tensor-product spline estimator (TPS) and FSVD estimators with fixed number of knots (SVf) and optimal number of knots (SVo).

errors are not given, to avoid overcrowding the table, but all the differences are significant (the Technical Report shows boxplots of the simulated squared errors). We see that for μ_1, for which the order p of $\hat{\mu}$ is correctly specified, the FSVD estimator with a fixed number of knots outperforms the tensor-product spline estimator in all situations but one ($\sigma = 1$, $m = 30$, $n = 50$), while the FSVD estimator with optimal number of knots outperforms the tensor-product spline estimator in *all* situations (usually by a considerable margin).

For μ_2 the situation reverses, as expected, since the order p is now underspecified and then the bias does not vanish, even as m or n increase. Of course, it can be argued that p in practice is also chosen in a data-driven way: for large m and n, the estimators $\hat{\phi}_3$ and $\hat{\psi}_3$ will be regular enough to call for a three-component estimator, which will make the FSVD estimator competitive again. The conclusion of this Monte Carlo study, then, is that FSVD estimators are competitive and even better than tensor-product splines as long as the number of components is not severely underspecified. Even if the estimated eigenvalues are small, for estimation purposes it is safer to include as many eigenfunctions as possible, as long as they are not overwhelmed by noise.

Acknowledgment

This research was supported by the National Science Foundation, award number DMS 0604396.

A Appendix

The following proofs use functional analysis results that can be found, for instance, in Gohberg et al. (2003). Given $\mu \in L^2(\mathcal{S} \times \mathcal{T})$, define the operator $\mathfrak{M} : L^2(\mathcal{T}) \to L^2(\mathcal{S})$ as $(\mathfrak{M}f)(s) = \int_{\mathsf{T}} \mu(s,t)f(t)\mathrm{d}t$. The adjoint of $\mathfrak{M}$ is the operator $\mathfrak{M}^* : L^2(\mathcal{S}) \to L^2(\mathcal{T})$ given by $(\mathfrak{M}^*g)(t) = \int_{\mathsf{S}} \mu(s,t)g(s)\mathrm{d}s$. Let $\mathfrak{K}_1 = \mathfrak{M}\mathfrak{M}^*$ and $\mathfrak{K}_2 = \mathfrak{M}^*\mathfrak{M}$. They are self-adjoint operators, $\mathfrak{K}_1 : L^2(\mathcal{S}) \to L^2(\mathcal{S})$ and $\mathfrak{K}_2 : L^2(\mathcal{T}) \to L^2(\mathcal{T})$, with kernels $k_1(s_1,s_2) = \int \mu(s_1,t)\mu(s_2,t)\mathrm{d}t$ and $k_2(t_1,t_2) = \int \mu(s,t_1)\mu(s,t_2)\mathrm{d}s$, respectively.

Remember that for $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$, the tensor-product operator $g \otimes f : \mathcal{H}_1 \to \mathcal{H}_2$ is defined as $(g \otimes f)(h) = \langle f, h \rangle g$.

A.1 Proof of Theorem 1

Since $\mathfrak{K}_2$ is a self-adjoint integral operator, the spectral decomposition implies that $\mathfrak{K}_2 = \sum \lambda_k \psi_k \otimes \psi_k$, where $\lambda_k > 0$ and $\{\psi_k\}$ is an orthonormal system of eigenfunctions of $\mathfrak{K}_2$, which can be completed to a basis of $L^2(\mathcal{T})$ by adding an orthonormal basis of $\ker(\mathfrak{K}_2)$, say $\{\tilde{\psi}_k\}$ (Gohberg et al., 2003, p. 180). This proves (2) of Theorem 1. Note that $\ker(\mathfrak{K}_2) = \ker(\mathfrak{M})$: clearly $\ker(\mathfrak{M}) \subseteq \ker(\mathfrak{K}_2)$ because $\mathfrak{K}_2 = \mathfrak{M}^*\mathfrak{M}$; but for any $f \in \ker(\mathfrak{K}_2)$, $0 = \langle f, \mathfrak{K}_2 f \rangle = \|\mathfrak{M}f\|^2$, which implies $f \in \ker(\mathfrak{M})$ and then $\ker(\mathfrak{K}_2) \subseteq \ker(\mathfrak{M})$.

Now define $\phi_k = \lambda_k^{-1/2}\mathfrak{M}\psi_k$.The ϕ_ks are orthonormal in $L^2(\mathcal{S})$, since

$$\begin{aligned}\langle \phi_j, \phi_k \rangle &= \lambda_j^{-1/2}\lambda_k^{-1/2}\langle \mathfrak{M}\psi_j, \mathfrak{M}\psi_k \rangle \\ &= \lambda_j^{-1/2}\lambda_k^{-1/2}\langle \psi_j, \mathfrak{K}_2\psi_k \rangle = \lambda_j^{-1/2}\lambda_k^{-1/2}\ \lambda_k\delta_{jk}.\end{aligned}$$

To prove (3) of Theorem 1, define the operator $\mathfrak{L} = \sum \lambda_k^{1/2}\phi_k \otimes \psi_k$. This operator is well defined, since for any $f \in L^2(\mathcal{T})$, we have $\mathfrak{L}f = \sum \lambda_k^{1/2}\langle\psi_k, f\rangle\phi_k$ and

$$\|\mathfrak{L}f\|^2 = \sum \lambda_k|\langle\psi_k, f\rangle|^2 \leq \|f\|^2 \sum \lambda_k < \infty.$$

Direct calculation shows that $\mathfrak{L}\psi_k = \mathfrak{M}\psi_k$, and $\mathfrak{L}\tilde{\psi}_k = \mathfrak{M}\tilde{\psi}_k = 0$ because $\ker(\mathfrak{K}_2) = \ker(\mathfrak{M})$. Since $\{\psi_k\} \cup \{\tilde{\psi}_k\}$ is a basis of $L^2(\mathcal{T})$, it follows that $\mathfrak{L} = \mathfrak{M}$, which is (3) of Theorem 1 in different words.

The identity (1) of Theorem 1 follows from (3), since $\mathfrak{K}_1 = \mathfrak{M}\mathfrak{M}^*$. In particular, this shows that the positive eigenvalues of $\mathfrak{K}_1$ are the same as those of $\mathfrak{K}_2$, and the ϕ_ks can be taken as the corresponding eigenfunctions.

If the mean function $\mu(s,t)$ is continuous, Mercer's Theorem (Gohberg et al., 2003, p. 198) implies that the ψ_ks are continuous and k_2 satisfies (2) in Theorem 1 in a pointwise manner, with the series converging absolutely and uniformly.

The ϕ_ks are continuous by definition when μ is continuous. To prove that the identity (1) in Theorem 1 holds pointwise and that the series converges absolutely and uniformly, we essentially mimic the proof of Mercer's Theorem. See the Technical Report for details.

Finally, to show that expression (3) in Theorem 1 holds pointwise when the series on the right-hand side converges absolutely and uniformly, note that both sides of expression (3) define the same operator from $L^2(\mathcal{T})$ to $L^2(\mathcal{S})$, so the identity must hold almost everywhere, and by continuity, it must actually hold everywhere.■

Remark. As by-products of the proof of Theorem 1 we obtain the identities

$$\phi_k(s) = \frac{1}{\lambda_k^{1/2}}(\mathfrak{M}\psi_k)(s) = \frac{1}{\lambda_k^{1/2}} \int \mu(s,t)\psi_k(t)\mathrm{d}t,$$

and

$$\psi_k(t) = \frac{1}{\lambda_k^{1/2}}(\mathfrak{M}^*\phi_k)(t) = \frac{1}{\lambda_k^{1/2}} \int \mu(s,t)\phi_k(s)\mathrm{d}s.$$

A.2 Proof of Theorem 2

Since $\{f_k\}$ and $\{g_k\}$ are orthonormal,

$$\|\mu - h\|^2 = \|\mu\|^2 - 2\sum_{k=1}^{p} a_k\langle g_k, \mathfrak{M}f_k\rangle + \sum_{k=1}^{p} a_k^2,$$

which is minimized by $a_k = \langle g_k, \mathfrak{M} f_k \rangle,\ k = 1, \dots, p$. Then, minimizing $\|\mu - h\|^2$ is equivalent to maximizing $\sum_{k=1}^{p} |\langle g_k, \mathfrak{M} f_k \rangle|^2$. By Cauchy-Schwartz inequality,

$$\begin{aligned} \sum_{k=1}^{p} |\langle g_k, \mathfrak{M} f_k \rangle|^2 &\leq \sum_{k=1}^{p} \|g_k\|^2 \, \|\mathfrak{M} f_k\|^2 \\ &= \sum_{k=1}^{p} |\langle \mathfrak{M} f_k, \mathfrak{M} f_k \rangle|^2 = \sum_{k=1}^{p} |\langle f_k, \mathfrak{K}_2 f_k \rangle|^2 . \qquad (7) \end{aligned}$$

It is well known (or see Gohberg et al., 2003, Section 4.9) that (7) is maximized by the leading p eigenfunctions of $\mathfrak{K}_2$, and the maximum value is $\sum_{k=1}^{p} \lambda_k$. Therefore $\sum_{k=1}^{p} |\langle g_k, \mathfrak{M} f_k \rangle|^2 \leq \sum_{k=1}^{p} \lambda_k$ and equality holds for $f_k = \psi_k$ and $g_k = \phi_k$, which completes the proof. ■

A.3 Proof of Theorem 3

Let $z_{ijk} = x_{ijk} - \mu(s_j, t_k)$, and define $\mathbf{M}_0 = [\mu(s_j, t_k)]_{(j,k)}$, $\mathbf{X}_i = [x_{ijk}]_{(j,k)}$ and $\mathbf{Z}_i = [z_{ijk}]_{(j,k)}$. Since $\hat{\mathbf{\Omega}} = \mathbf{B}^\top \mathbf{V} \mathbf{K}_1 \mathbf{V} \mathbf{B}$ and $\mathbf{K}_1 = \bar{\mathbf{X}} \mathbf{U} \bar{\mathbf{X}}^\top$, we can write

$$\begin{aligned} \hat{\Omega}_{hh'} &= \beta_h(\mathbf{s})^\top \mathbf{V} \bar{\mathbf{X}} \mathbf{U} \bar{\mathbf{X}}^\top \mathbf{V} \beta_{h'}(\mathbf{s}) \\ &= \beta_h(\mathbf{s})^\top \mathbf{V} \mathbf{M}_0 \mathbf{U} \mathbf{M}_0^\top \mathbf{V} \beta_{h'}(\mathbf{s}) \qquad (8) \\ &\quad + 2\beta_h(\mathbf{s})^\top \mathbf{V} \bar{\mathbf{Z}} \mathbf{U} \mathbf{M}_0^\top \mathbf{V} \beta_{h'}(\mathbf{s}) \qquad (9) \\ &\quad + \beta_h(\mathbf{s})^\top \mathbf{V} \bar{\mathbf{Z}} \mathbf{U} \bar{\mathbf{Z}}^\top \mathbf{V} \beta_{h'}(\mathbf{s}). \qquad (10) \end{aligned}$$

We will show that (8) goes to $\Omega_{hh'}$ as m and r go to infinity, and that (9) and (10) go to zero in probability as n goes to infinity, uniformly in m and r.

Since

$$\beta_h(\mathbf{s})^\top \mathbf{V} \bar{\mathbf{X}} \mathbf{U} \bar{\mathbf{X}}^\top \mathbf{V} \beta_{h'}(\mathbf{s}) =$$
$$\sum_{j=1}^{m} \sum_{j'=1}^{m} \beta_h(s_j) v_j \left\{ \sum_{k=1}^{r} u_k \mu(s_j, t_k) \mu(s_{j'}, t_k) \right\} v_{j'} \beta_{h'}(s_{j'}),$$

it is clear that (8) goes to $\Omega_{hh'}$ as m and r go to infinity, because both $\max v_j$ and $\max u_k$ go to zero as m and r go to infinity.

With respect to (9), note that we can write it as $2\bar{y}$, with

$$y_i = \beta_h(\mathbf{s})^\top \mathbf{V} \mathbf{Z}_i \mathbf{U} \mathbf{M}_0^\top \mathbf{V} \beta_{h'}(\mathbf{s}).$$

The y_is are i.i.d. with $\mathrm{E}(y_i) = 0$ and $\mathrm{V}(y_i) = \mathrm{V}\left\{\sum_{j=1}^{m}\sum_{k=1}^{r}\beta_h(s_j)v_j z_{ijk}u_k a_{kh'}\right\}$, with $a_{kh'} = \sum_{j'=1}^{m}\mu(s_{j'}, t_k)v_{j'}\beta_{h'}(s_{j'})$. It can be proved that

$$\lim_{\substack{m\to\infty\\ r\to\infty}} \mathrm{V}(y_i) = \iiiint \beta_h(s_1)\alpha_{h'}(t_1)\beta_h(s_2)\alpha_{h'}(t_2)\rho\{(s_1,t_1),(s_2,t_2)\}\mathrm{d}s_1\mathrm{d}s_2\mathrm{d}t_1\mathrm{d}t_2,$$

where $\alpha_{h'}(t_k) = \int \mu(s,t_k)\beta_{h'}(s)\mathrm{d}s$ as $m \to \infty$ (see Technical Report). Then $\mathrm{V}(y_i)$ is bounded for any m and r, and a simple application of Tchebyshev's Inequality implies that (9) goes to zero in probability as n goes to infinity, uniformly in m and r.

Regarding (10), note that

$$\beta_h(\mathbf{s})^\top \mathbf{V}\bar{\mathbf{Z}}\mathbf{U}\bar{\mathbf{Z}}^\top \mathbf{V}\beta_{h'}(\mathbf{s}) \leq \|\mathbf{U}^{1/2}\bar{\mathbf{Z}}^\top\mathbf{V}\beta_h(\mathbf{s})\|\|\mathbf{U}^{1/2}\bar{\mathbf{Z}}^\top\mathbf{V}\beta_{h'}(\mathbf{s})\|.$$

For a given index h, we can write $\mathbf{U}^{1/2}\bar{\mathbf{Z}}^\top\mathbf{V}\beta_h(\mathbf{s}) = \bar{\mathbf{w}}$, with $\mathbf{w}_i = \mathbf{U}^{1/2}\mathbf{Z}_i^\top\mathbf{V}\beta_h(\mathbf{s})$. The $\mathbf{w}_i$s are i.i.d. with $\mathrm{E}(\mathbf{w}_i) = 0$ and

$$\lim_{\substack{m\to\infty\\ r\to\infty}} \sum_{k=1}^{r} \mathrm{V}(w_{ik}) = \iiint \beta_h(s_1)\beta_h(s_2)\rho\{(s_1,t),(s_2,t)\}\mathrm{d}s_1\mathrm{d}s_2\mathrm{d}t$$

(again, see Technical Report). Since $\mathrm{E}(\|\bar{\mathbf{w}}\|^2) = n^{-1}\sum_{k=1}^{r}\mathrm{V}(w_{ik})$, a straightforward application of Markov's Inequality implies that $\|\bar{\mathbf{w}}\|$ goes to zero in probability as n goes to infinity, uniformly in m and r, and consequently the same is true for (10). ■

References

Adler, R. J., and Taylor, J. E. (2007). *Random Fields and Geometry*. New York: Springer-Verlag.

Aubry, N., Guyonnet, R., and Lima, R. (1991). Spatio-temporal analysis of complex signals: theory and applications. *Journal of Statistical Physics*, 64, 683–739.

De Boor, C. (2001). *A Practical Guide to Splines*. New York: Springer-Verlag.

Dente, J. A., Vilela Mendes, R., Lambert, A., and Lima, R. (1996). The bi-orthogonal decomposition in image processing: signal analysis and texture segmentation. *Signal Processing: Image Communication*, 8, 131–148.

Ferraty, F., and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer-Verlag.

Gasser, T., Gervini, D., and Molinari, L. (2004). Kernel estimation, shape-invariant modeling and structural analysis. In *Methods in Human Growth Research*, eds. R. Hauspie, N. Cameron & L. Molinari. Cambridge: Cambridge University Press, pp.179–204.

Gervini, D. (2006). Free-knot spline smoothing for functional data. *Journal of the Royal Statistical Society*, Ser. B, 68, 671–687.

Gohberg, I., Goldberg, S., and Kaashoek, M. A. (2003). *Basic Classes of Linear Operators*. Basel: Birkhäuser Verlag.

Gu, C. (2000). Multivariate spline regression. In *Smoothing and Regression: Approaches, Computation, and Application*, ed. M. Schimek. New York: Wiley, pp. 329–355.

Härdle, W., and Müller, M. (2000). Multivariate and semiparametric kernel regression. In *Smoothing and Regression: Approaches, Computation, and Application*, ed. M. Schimek. New York: Wiley, pp. 357–391.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* New York: Springer.

Kneip, A., and Utikal, K. J. (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association*, 96, 519–532.

Morris, J. S., and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society*, Ser. B, 68, 179–199.

Nychka, D. (2000). Spatial-process estimates as smoothers. In *Smoothing and Regression: Approaches, Computation, and Application*, ed. M. Schimek. New York: Wiley, pp. 393–424.

Ramsay, J. O., and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer-Verlag.

Ramsay, J. O., and Silverman, B. W. (2005). *Functional Data Analysis* (2nd edition). New York: Springer-Verlag.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Annals of Statistics*, 24, 1–24.

Taylor, J. E., and Worsley, K. J. (2007). Detecting sparse signals in random fields, with an application to brain mapping. *Journal of the American Statistical Association*, 102, 913–928.

Tyler, D. E. (1981). Asymptotic inference for eigenvectors. *The Annals of Statistics*, 9, 725–736.

Yao, F., and Lee, T. C. M. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society*, Ser. B, 68, 3–25.