ELSEVIER Computational Statistics & Data Analysis 24 (1997) 255–270

COMPUTATIONAL STATISTICS & DATA ANALYSIS

Simultaneous non-parametric regressions of unbalanced longitudinal data

Philippe C. Besse, Hervé Cardot*, Fréderic Ferraty

Laboratoire de Statistique et Probabilités, URA CNRS 745, Université Paul Sabatier, 31064 Toulouse Cedex, France

Received November 1995; revised June 1996

Abstract

The aim of this paper is to simultaneously estimate n curves corrupted by noise, this means several observations of a random process. The non-parametric estimation of the sampled paths leads to a new kind of functional principal components analysis which simultaneously takes into account a dimensionality and a smoothness constraint. Furthermore, the use of B-spline approximation to estimate the curves allows the study of unbalanced longitudinal data. The relationship between the choice of the smoothing parameter and that of dimensionality is discussed. A simulation study shows good behaviors of this proposed estimate compared to n independent smoothing splines under generalized cross-validation. Finally, the methodology of this paper is illustrated by its application to a real world data set.

Keywords: Non-parametric regression; Functional principal component analysis; Hybrid splines; B-splines; Rainfall data

1. Introduction

The aim of this paper is the functional estimation of several curves, this means simultaneous non-parametric regressions of smooth sampled curves corrupted by noise. Furthermore we consider the case of missing or unbalanced data. This means that curves are not necessarily observed at the same times. The regressions could be achieved by computing n classical non-parametric (kernel or spline) regressions

* Corresponding author. E-mail: besse@cict.fr.

PII S0167-9473(96)00067-9

where each smoothing parameter is independently optimized by generalized cross-validation (GCV). But this approach does not take into account the fact that the data are independent observations of the same process and simulations show that it induces a loss of fit.

This kind of data has been previously studied by Besse and Ramsay (1986), Ramsay and Dalzell (1991), Besse and Pousse (1992) and Kneip (1995) which deal with principal components analysis (PCA) of curves. These are closely related to other papers which aim at approximating a covariance function by smooth eigenfunctions (Rice and Silverman, 1991; Jones and Rice, 1992; Pezzuli and Silverman, 1993). In another context, Boularan et al. (1993) propose a two-step non-parametric model well adapted to unbalanced data, for example, growth curves. In that case, each curve contributes to the functional estimation of a common effect as well as being separately estimated.

In some other cases, for instance when curves do not roughly share the same common shape, it is more interesting to take the covariance function into account. This is achieved by both considering dimensional and smoothness constraints in a simultaneous non-parametric estimation. This leads to the definition of a new kind of functional PCA. The solution is similar to those found in parallel works by Denby and Mallows (1993), Besse (1994) and Silverman (1995) but takes the unbalanced measurement into account.

The organization of the paper is as follows. In Section 2 we establish notations and define the framework of our study. Section 3 is devoted to the estimation problem under rank and smoothness constraints. The unbalanced sample paths are, at first, estimated by means of B-splines. Afterwards the smooth functional PCA is applied to the coordinates of the trajectories in the space spanned by the B-splines. Section 4 discusses the choice of the smoothing parameter value jointly with the choice of dimensionality. In Section 5 we compare estimates of simulated noisy functions which are obtained, on the one hand, by spline smoothing of each curve with a GCV smoothing parameter and, on the other hand, by this new functional PCA. All programs are written in S+ (Becker et al., 1988) and are available on request. Finally, the methodology of the paper is illustrated by its application to a set of rainfall data.

2. Functional framework

Let (Ω, A) be a measurable space with the probability measure P and z a vector random function (r.f.) mapping from (Ω, A, P) into (H, B_H) where H is a separable Hilbert space and B_H its Borel field. We assume that z is a second-order r.f., $\mathbb{E}\|z\|_H^2 < \infty$, where $\|.\|_H$ denotes the norm of H, and define $\mu = \mathbb{E}(z)$.

Now let us assume z is a "smooth" stochastic process whose sample paths belong to a q-dimensional subspace H_q of H. This means that the r.f. z can be decomposed as a linear combination of q smooth elements of H. One way to control the smoothness and the regularity of the trajectories of the r.f. z is to let H be a Sobolev space $W^m[0,1]$, that is to say the collection of functions on $[0,1]$ (without loss of generality) which obey

$$\{f\colon f, f', \dots, f^{(m-1)} \text{ absolutely continuous, } f^{(m)} \in L_2[0,1]\}.$$

The smoothness of a function f in $W^m[0,1]$ can be controlled by the semi-norm

$$\|f\|_m^2 = \int_0^1 (f^{(m)})^2(t)\,\mathrm{d}t \leq c \quad (c > 0). \tag{1}$$

One could generalize to any equivalent semi-norm (see Wahba, 1990) leading to Tchebycheffian splines by replacing the differential operator D^m in the expression

$$\|f\|_m^2 = \|D^m f\|_{L_2}^2$$

by some more general differential operators.

Let us assume, for almost all ω in Ω, the function $z(\omega,.)$ belongs to H_q which is a q-dimensional subspace of $W^m[0,1]$ and the r.f. satisfies the smoothness constraint

$$\mathbb{E}\|z\|_m^2 < c, \quad c > 0.$$

In practice, each sample path $\{z(\omega_i,.);\ i = 1,\dots,n\}$ of the r.f. z is observed at a finite number, p_i, of points in the interval $[0,1]$, $0 \leq t_{i1} < \cdots < t_{p_i} \leq 1$. Usually the measuring apparatus introduces measurement errors which can be supposed to be independent and identically distributed with finite variance σ^2. This leads to the observation of the random vector[1] of $\mathbb{R}^{p_i}$:

$$\boldsymbol{y}_i = \boldsymbol{z}_i + \boldsymbol{\varepsilon}_i, \tag{2}$$

with the following assumptions and notation for $i,j = 1,\dots,n$:

$$\mathbb{E}\boldsymbol{\varepsilon}_i = 0,$$

$$\mathbb{E}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i' = \sigma^2\boldsymbol{I},$$

$$\mathbb{E}\boldsymbol{\varepsilon}_j\boldsymbol{z}_i' = 0,$$

$$\boldsymbol{z}_i = (z(\omega_i, t_{i1}), \dots, z(\omega_i, t_{p_i}))'.$$

As a first step, to achieve our functional data analysis, we have to estimate the trajectories from the noisy and discrete observations. It will be done non-parametrically to “let the data a chance to speak”. The use of B-splines (De Boor, 1978) seems to be appropriate since we do not need to assume the discretization design is the same for all the curves and these functions have good approximation properties. Let us denote by

$$\mathcal{B}_{k,m} = \{B_l;\ l = 1,\dots,r = k + m + 1\}$$

a basis of B-splines of degree m defined on $[0,1]$ with k equispaced knots and consider $\mathcal{S}_{k,m}$ the space generated by $\mathcal{B}_{k,m}$. It can be noticed that $\mathcal{S}_{k,m}$ is a subspace of $W^m[0,1]$.

[1] In the succeeding discussion, we will use bold italic letters to denote vectors or matrices.

Let $\boldsymbol{B}_k(t)$ be the vector of $B_l(t)$, $l = 1, \ldots, r$ and

$$\boldsymbol{A}_{ki} = \frac{1}{p_i} \sum_{j=1}^{p_i} \boldsymbol{B}_k(t_{ij}) \boldsymbol{B}_k'(t_{ij}).$$

Then, the least squares estimate $\hat{z}_i$ of z_i in $\mathscr{S}_{k,m}$ is

$$\hat{z}_i = \sum_{l=1}^{r} s_{il} B_l,$$

where

$$\boldsymbol{s}_i = \boldsymbol{A}_{ki}^{-1} \boldsymbol{b}_i$$

is the vector of coordinates of the estimated trajectory in the basis of B-splines and

$$\boldsymbol{b}_i = 1/p_i \sum_{j=1}^{p_i} y_{ij} \boldsymbol{B}_k(t_{ij}).$$

Let us denote by $\boldsymbol{C}$ the scalar product matrix of the B-splines in $L^2(T)$ and by $\boldsymbol{G}$ the matrix associated to the semi-norm which measures the smoothness of the functions of $\mathscr{S}_{k,m}$:

$$[\boldsymbol{C}]_{kl} = \int B_k(t) B_l(t)\,\mathrm{d}t, \quad k, l = 1, \ldots, r,$$

$$[\boldsymbol{G}]_{kl} = \int B_k^{(m)}(t) B_l^{(m)}(t)\,\mathrm{d}t, \quad k, l = 1, \ldots, r,$$

with these notations,

$$\|\hat{z}_i\|_{L_2}^2 = \boldsymbol{s}_i' \boldsymbol{C} \boldsymbol{s}_i = \|\boldsymbol{s}_i\|_{\boldsymbol{C}}^2,$$

$$\|\hat{z}_i\|_m^2 = \boldsymbol{s}_i' \boldsymbol{G} \boldsymbol{s}_i = \|\boldsymbol{s}_i\|_{\boldsymbol{G}}^2.$$

3. Model and estimation

In practice we only observe a finite number of independent realizations of the smooth stochastic process z. The previous discussion leads us to consider the following model:

$$y_i(t_{ij}) = z_i(t_{ij}) + \varepsilon_{ij}, \ t_{i1} < \cdots < t_{ip_i};$$

$$i = 1, \ldots, n \quad \text{and} \quad j = 1, \ldots, p_i;$$

$$\text{with} \begin{cases} \mathbb{E}(\varepsilon_{ij}) = 0 \text{ and } \mathbb{E}(\varepsilon_{ij}\varepsilon_{ij'}) = \sigma^2 \delta_{jj'}, \\ \sigma \text{ unknown } (\sigma > 0), \\ z_i \text{ independent of } \varepsilon_{i'}, \quad i' = 1, \ldots, n, \\ z_i \in H_q \text{ a.s.}, \\ \|z_i\|_m^2 \le c \text{ a.s.}, \end{cases} \tag{3}$$

The data specificity suggests two kinds of constraints in the estimation procedure. On the one hand, a dimensionality constraint assumes that curves only span a subspace H_q of W_T^m. On the other hand, each real curve is assumed to be sufficiently smooth to satisfy constraint (1) with a common constant c.

A least squares estimation of the sample paths and the subspace leads us to consider the following optimization problem in which the smoothness constraint is taken into account by introducing a Lagrange multiplier ρ. This is the common smoothing parameter for all the curves.

$$\min_{\hat{z}_i, H_q^r} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\|\hat{z}_i - \tilde{z}_i\|_{L^2}^2 + \rho\|\tilde{z}_i\|_m^2);\ \tilde{z}_i \in H_q^r,\ \dim H_q^r = q \right\}, \tag{4}$$

where H_q^r is a q dimensional affine subspace of $\mathscr{S}_{k,m}$ to be estimated.

With the above notations, the optimization problem (4) is equivalent to

$$\min_{u_i, A_q} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\|\boldsymbol{s}_i - \boldsymbol{u}_i\|_C^2 + \rho\|\boldsymbol{u}_i\|_G^2);\ \boldsymbol{u} \in A_q,\ \dim A_q = q \right\}, \tag{5}$$

where A_q is a q-dimensional affine subspace of $\mathbb{R}^r$.

Let us denote by $\bar{\boldsymbol{s}} = (1/n)\sum_{i=1}^{n} \boldsymbol{s}_i$ the mean coordinates, by $\boldsymbol{x}_i$ the vector $(\boldsymbol{s}_i - \bar{\boldsymbol{s}})$ and by $\boldsymbol{\Gamma}$ the empirical covariance matrix in the B-splines basis:

$$\boldsymbol{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'.$$

Finally, let us define $\boldsymbol{H}_\rho = (\boldsymbol{C} + \rho \boldsymbol{G})^{-1}$ which can be interpreted as a smoothing matrix.

Proposition 1. *The solution of problem* (4) *is given by*

$$\hat{\boldsymbol{u}}_i = \boldsymbol{H}_\rho^{1/2} \widehat{\boldsymbol{P}}_q \boldsymbol{H}_\rho^{1/2} \boldsymbol{C}\boldsymbol{x}_i + \boldsymbol{H}_\rho \boldsymbol{C}\bar{\boldsymbol{s}}, \quad i = 1, \ldots, n.$$

The matrix $\widehat{\boldsymbol{P}}_q = \boldsymbol{V}_q \boldsymbol{V}_q'$ is the orthogonal projector on the subspace generated by the first q eigenvectors of the matrix

$$\boldsymbol{H}_\rho^{1/2} \boldsymbol{C}\boldsymbol{\Gamma}\boldsymbol{C}\boldsymbol{H}_\rho^{1/2}.$$

The proof of Proposition 1 is to be found in the Appendix.

Afterwards, we can construct the estimated sample paths:

$$\hat{z}_i(t) = \sum_{l=1}^{r} \hat{u}_{il} B_l(t), \quad t \in T,\ i = 1, \ldots, n. \tag{6}$$

One can notice that the smooth estimated sample paths are obtained by the generalized singular value decomposition (SVD) of $\boldsymbol{X}\boldsymbol{C}\boldsymbol{H}_\rho$ with respect to the metrics $(1/n)\boldsymbol{I}_n$ and $\boldsymbol{H}_\rho^{-1}$,$\boldsymbol{X}$ being the matrix whose rows are the $\boldsymbol{x}_i$'s. It can be interpreted as a kind of smooth approximation of the Karhunen–Loeve expansion of the observed process. Convergence properties are on study.

This estimation procedure is very similar to those proposed by Denby and Mallows (1993) and Besse (1994). They lead to the spectral analysis of matrices of the form $A(\lambda)^{1/2}\boldsymbol{S}A(\lambda)^{1/2}$ where $A(\lambda)$ is a Hat-matrix associated to the smoothing parameter λ and $\boldsymbol{S}$ is an empirical covariance matrix. The main interest of our estimator is that it can be applied to any type of discretization since we work on the coordinates of the approximated sample paths. That is not the case for the methods presented before because they work directly on the discretized sample paths and therefore have to assume the discretization design is the same for each curve.

Our estimator of the sample paths can be viewed as a B-splines approximation of the smoothing splines. They are called hybrid splines by Kelly and Rice (1990) and Champely (1994) in the context of a non-parametric estimator of a regression function.

They have mainly two advantages compared with the usual smoothing splines and the B-splines. On the one hand, if k is sufficiently large, the tradeoff between the smoothness and the fidelity to the data can be considered as a function of only one parameter, the smoothing parameter ρ as in the case of smoothing splines. On the other hand, they are easily computable like B-splines. Computations only deal with spectral analyses or inverses of $(r \times r)$ well-conditioned matrices.

4. Smoothing and dimension choice

Both hybrid splines and SVD act as smoothing tools. The first is determined by a smoothing parameter ρ whose optimal value depends on data regularity. In contrast, SVD of longitudinal data gives approximations of trajectories whose smoothness usually depends on the number q of selected components; neglected components are often essentially noisy. These parameter values must be jointly optimized in order to obtain better fits of the original functions. GCV is commonly used to optimize the smoothing parameter of the spline. Unfortunately, its application to the dimension choice in PCA (Krzanowski, 1987) is not convincing (Besse and Ferré, 1993) and never used in practice.

Besse and Pousse (1992) suggest another strategy based on a stability criterion: PCA results are assumed to be reliable if the estimated subspace $\widehat{E_q}$ is stable with respect to data perturbations. It is then proposed to find a dimension and a smoothing parameter which produce stability in this sense.

We consider the following loss function which measures the subspace estimation quality; it is based on the usual norm of matrices when it is applied to measure distances between projectors, and thus to measure distances between the associated subspaces:

$$\mathscr{L}_q = \tfrac{1}{2}\|\boldsymbol{P}_q - \widehat{\boldsymbol{P}}_q\|^2 = q - \operatorname{tr}\boldsymbol{P}_q\widehat{\boldsymbol{P}}_q, \tag{7}$$

where $\widehat{\boldsymbol{P}}_q$ (resp. $\boldsymbol{P}_q$) is the orthogonal projection onto $\widehat{E}_q$ (resp. E_q).

In that context, $\operatorname{tr}\boldsymbol{P}_q\widehat{\boldsymbol{P}}_q$ is also the sum of the squared canonical correlation coefficients between the component sets which respectively span E_q and $\widehat{E}_q$.

A risk function is then defined by taking the expectation:

$$R_q = \mathbb{E}(\mathscr{L}_q). \tag{8}$$

The criterion R_q is symmetrically defined since its value is invariant under any permutation of the observations y_i, resampling methods such as the bootstrap and the jackknife are natural candidates to compute estimates for this risk function. Besse and Falguerolles (1993) compared these methods which are all computationally expensive. Furthermore Besse (1992) proposed an approximation of the jackknife which does not require much computation.

If n is large enough, it is quite acceptable to consider that any row elimination introduces only a small perturbation in further computations. Perturbation theory leads to Taylor's series expansions of the eigenelements of $\boldsymbol{XCH}_\rho^{1/2}$ which lead to a Taylor's series expansion of the jackknife estimate and then to an analytic approximation:

$$\widehat{R_{JKq}} = \widehat{R_{Pq}} + \mathrm{O}(n^{-2}).$$

An analytic approximation of the jackknife estimate is given by

$$\widehat{R_{Pq}} = \frac{1}{n-1} \sum_{k=1}^{q} \sum_{j=k+1}^{p} \frac{(1/n)\sum_{i=1}^{n} c_{ik}^2 c_{ij}^2}{(\lambda_j - \lambda_k)^2}, \tag{9}$$

where c_{ij} denotes the general entry of the matrix $\boldsymbol{XCH}_\rho^{1/2} \boldsymbol{V}_q$.

This shows the importance of the gap between successive eigenvalues. The leading term in $\widehat{R_{Pq}}$ depends on the difference between the eigenvalues associated with the last selected dimension and the first neglected one. This is consistent with intuitive considerations: if the difference $(\lambda_q - \lambda_{q+1})$ is large enough, data perturbations cannot lead to the swapping of the associated eigenvectors v_q and v_{q+1}. Optimality is achieved by minimizing $\widehat{R_{Pq}}$ with respect to both ρ and q. Simulations show that this is not easy, but this heuristic approach leads to fairly workable choices of ρ and q values. An asymptotic study still needs to be developed.

5. Applications

5.1. A simulated data example

Artificial data sets were generated to illustrate the above strategy in the usual context of cubic spline smoothing ($m = 2$). The four data sets of unbalanced observed values $y_i(t_{ij})$ are obtained by adding to the same "common effects" $z_i(t_{ij})$ a simulated white noise of controlled standard deviation:

(i) The true curves have been constructed as follows:

$$z_i(t_{ij}) = f_i(t_{ij}) = a_i t_j + b_i \cos(2\pi t_j) + c_i \cos(4\pi t_j) \quad \text{for} \quad \begin{cases} i = 1,\ldots,50, \\ j = 0,\ldots,p_i, \end{cases}$$

where a_i, b_i, c_i are pseudo-random numbers uniformly and independently distributed in the range $[0,1]$ and $t_{ij} = j/p_i$. As regards the length of each discretization p_i, we draw a uniformly distributed number u_i in the range [25, 30]; we allocate the closest integer of u_i to p_i. Clearly, the curves are observed at different subdivisions of [0, 1]. Moreover, H_q is generated by $\{t, \cos(2\pi t), \cos(4\pi t)\}$; the known rank q of H_q is equal to 3.

(ii) The noise was simulated as follows: let $N = \sum_{i=1}^{50}(p_i + 1); N$ observations of independently and identically normally distributed pseudo-random variables with mean 0 and standard deviation $\sigma_1 = 0.9$, $\sigma_2 = 0.7$, $\sigma_3 = 0.5$, $\sigma_4 = 0.3$ were generated.

However, it is interesting to know the signal-to-noise ratio. To this end, we take the same discretization for each curve $t_{ij} = \tau_j = j/30$, $j = 1,\ldots,30$. Let $\boldsymbol{Z}$ be the (50×30) matrix such that $[\boldsymbol{Z}]_{ij} = z_i(\tau_j)$ and let $\boldsymbol{D}$ be the (50×50) diagonal matrix such that $[\boldsymbol{D}]_{ii} = p_i/N$. The positive eigenvalues of the "common effect" pseudo-covariance matrix $\boldsymbol{Z}'\boldsymbol{D}\boldsymbol{Z}$ were

$$\lambda_1 = 1.70, \quad \lambda_2 = 1.27, \quad \lambda_3 = 0.82$$

and the noise variances satisfied

$$\lambda_2 > \sigma_1^2 \simeq \lambda_3 > \sigma_2^2 > \sigma_3^2 > \sigma_4^2$$

in order to test different levels of signal-to-noise ratios.

Since the data were simulated, the true quadratic risk

$$Q_r = \sum_{i=1}^{n} \frac{p_i}{N} \|z_i - \hat{z}_i\|^2$$

could be computed. Initially, $\hat{z}_i$ is the spline regression estimate for smoothing parameter value computed by the `smooth.spline` function of Splus. Each curve was separately estimated. In a second step, $\hat{z}_i$ is given by (6) for different values of q.

The plots (Figs. 1–3) are computed on the data set with a middle standard deviation: $\sigma_2 = 0.7$. Fig. 1 displays $Q_r(q)$ for different smoothing parameter values. The minimum is reached for $q = 3$ and $\rho \simeq 2 \times 10^{-5}$. These plots are to be compared with the value of $Q_r = 3.5$ when each curve is separately estimated. Fig. 2, which compares a "true" curve with its estimates, shows that GCV oversmooths, but even a smaller value of the smoothing parameter does not improve the fit. Only the truncation of the singular value decomposition, which, in this case, takes the covariance structure into account, makes visible the third data component $(\cos(4\pi t))$ in the estimates.

This estimation does not use the best smoothing parameter value which could be found in Fig. 1 by minimizing Q_r. It is deduced from Fig. 3 that displays $\widehat{R_{Pq}}$ versus $\log(\rho)$. Such a graph is not always easy to interpret because of the highly nonlinear behavior of eigenelements. Very sharp peaks occur in the case of almost equal eigenvalues. As $\log(\rho)$ grows, $\widehat{E}_3$ is successively reliable, then suddenly becomes very unstable and, finally, when the third component vanishes by oversmoothing, $\widehat{E}_3$ becomes as stable as $\widehat{E}_2$. A competitive ρ value lies between e^{-9} and e^{-7}. It is not exactly the value which minimizes Q_r, but it seems accurate enough to lead to a fairly good fit.

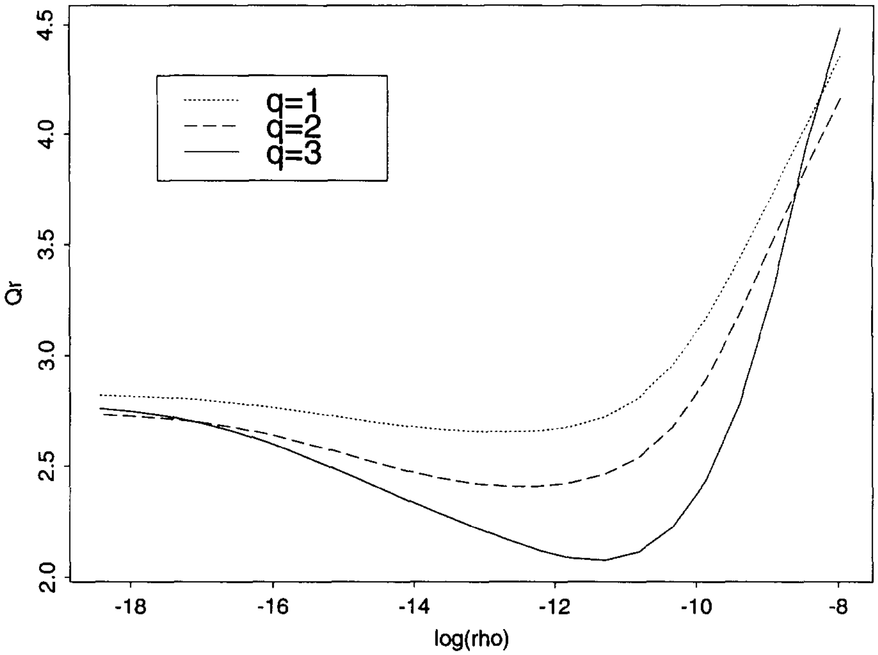

Fig. 1. The quadratic error $Q_r(q)$ versus the smoothing parameter $\log(\rho)$.

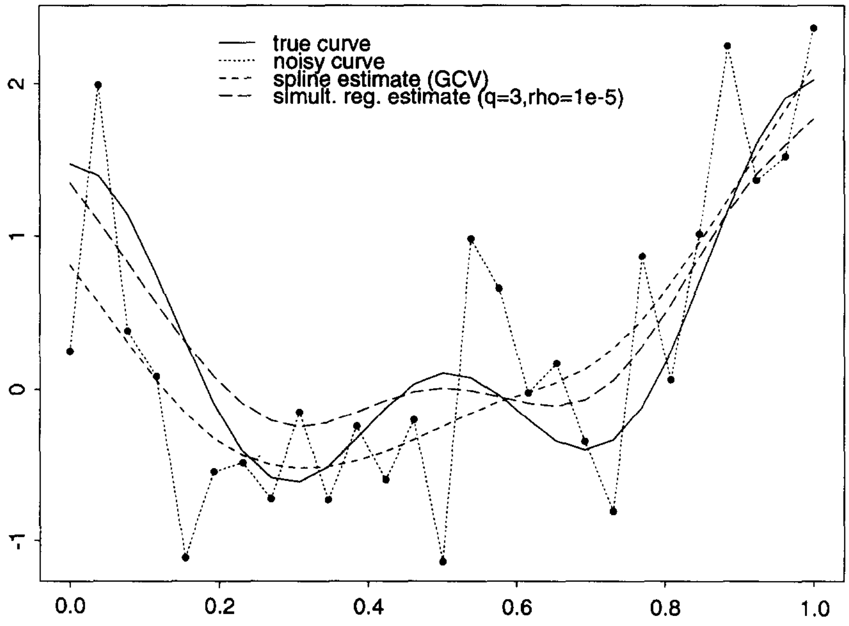

Fig. 2. A true curve $f(t_j)$, compared with its different estimates.

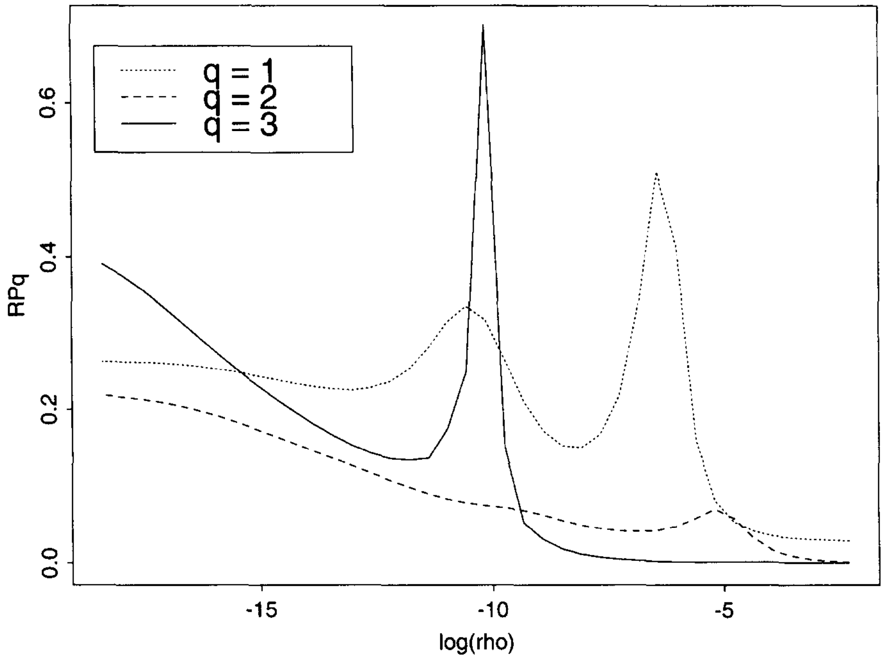

Fig. 3. The stability estimate $\widehat{R_{Pq}}$ versus $\log(\rho)$ for different dimensions.

Table 1
Q_r values of the two kinds of estimates for different noise standard deviations; optimal ρ and q values are deduced from $\widehat{R_{Pq}}$ plots

Std. dev. of the noise	Spline + GCV	Simultaneous regressions		
	Q_r	ρ	q	Q_r
0.3	0.8	2×10^{-6}	3	0.5
0.5	2.1	6×10^{-6}	3	1.2
0.7	3.5	2×10^{-5}	3	2.1
0.9	6.2	1×10^{-4}	1	4.8

Table 1 gives the quadratic errors for each value of the noise standard deviation and for each kind of estimate; Q_r naturally increases with σ but the functional PCA estimate is always competitive. As the noise variance becomes greater and greater, the quality of the estimate is improved by increasing the smoothing parameter (ρ) and also by reducing the dimension (q).

This study, only based on simulated data, gives an insight into the good behavior of these simultaneous non-parametric estimations resulting from that new kind of functional PCA. Nevertheless, only an asymptotic study could definitively justify such a practice.

5.2. Rainfall data

In this section, the methodology described above is illustrated by an application to rainfall data. The data set (ECOSTAT, 1991) consists of monthly observation, during 10 years, of rainfalls in 26 French towns. We considered these data as 26×5 sampled curves each observed during two years. This cutoff has been chosen to deal with more complicated curves. A study of yearly curves led to the same kind of results as those displayed below, but fewer components were required.

Furthermore, unbalanced data were obtained by randomly removing 15% of the observed values. Some of these raw curves are displayed in Fig. 4. All the curves are strongly unsmoothed and exhibit a very large variability. This is also emphasized by considering the first three eigenfunctions (Fig. 5) of a classical PCA computed on the balanced raw data. Finally, a classical transformation, the square root, was performed to stabilize the variance as it can be done in the case of a Poisson distribution.

The smoothed functional PCA of these unbalanced data was then computed by considering the hybrid spline approximation of each curve under the dimensionality constraint. This methodology requires to choose jointly the smoothing parameter value and the dimensionality. This was achieved by considering the graphs of the stability estimate (Fig. 6). It displays the estimate of the stability score $\widehat{R_{Pq}}$ for different dimensions q calculated at a grid (on a logarithm scale) of values of the smoothing parameter. For small values of $\rho(\log(\rho) < -5)$, only the first component, which mainly takes a trend into account, is reliable. For larger values of $\rho(\log(\rho) > 6)$, the transformed data are oversmoothed and many components vanish. As suggested

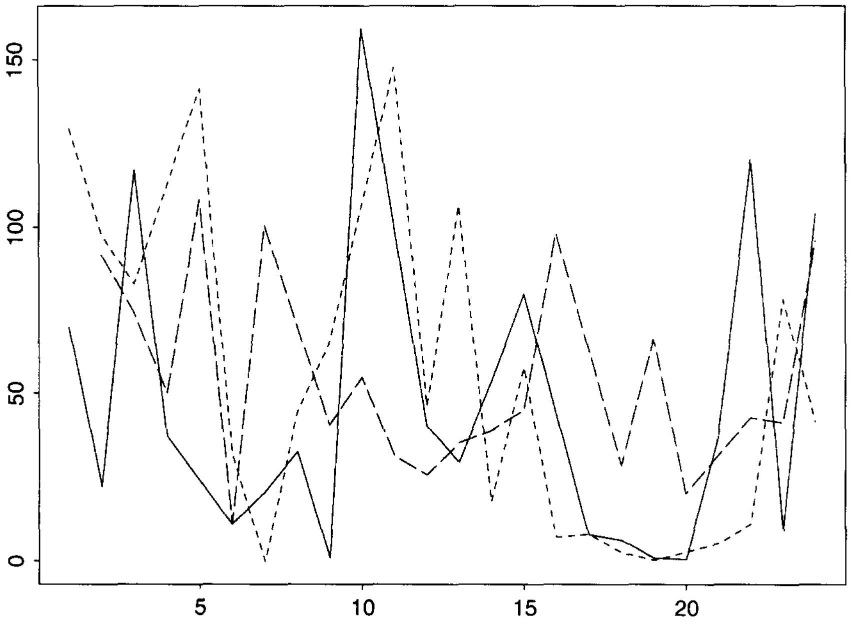

Fig. 4. Three curves coming from pluviometrical data set.

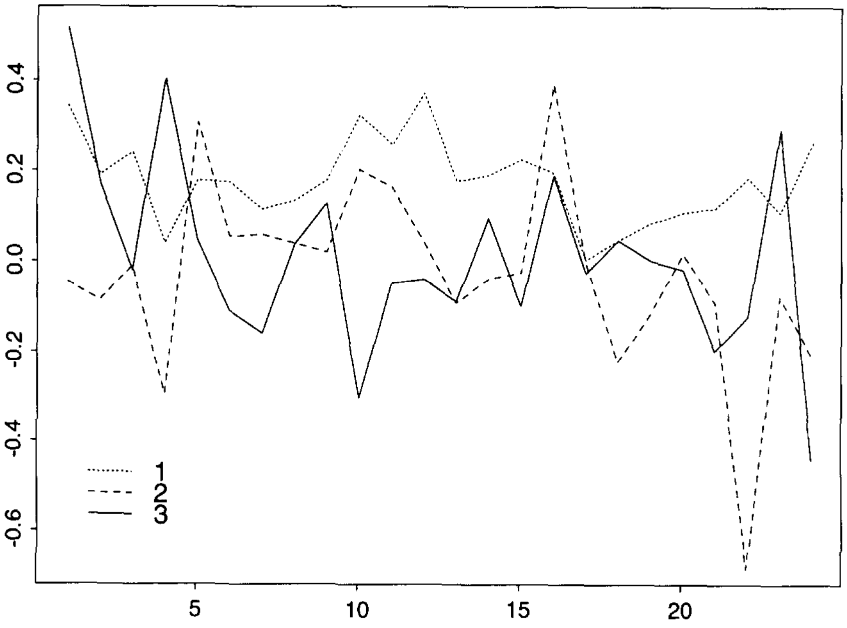

Fig. 5. The first three eigenfunctions computed by the classical PCA on the original data set.

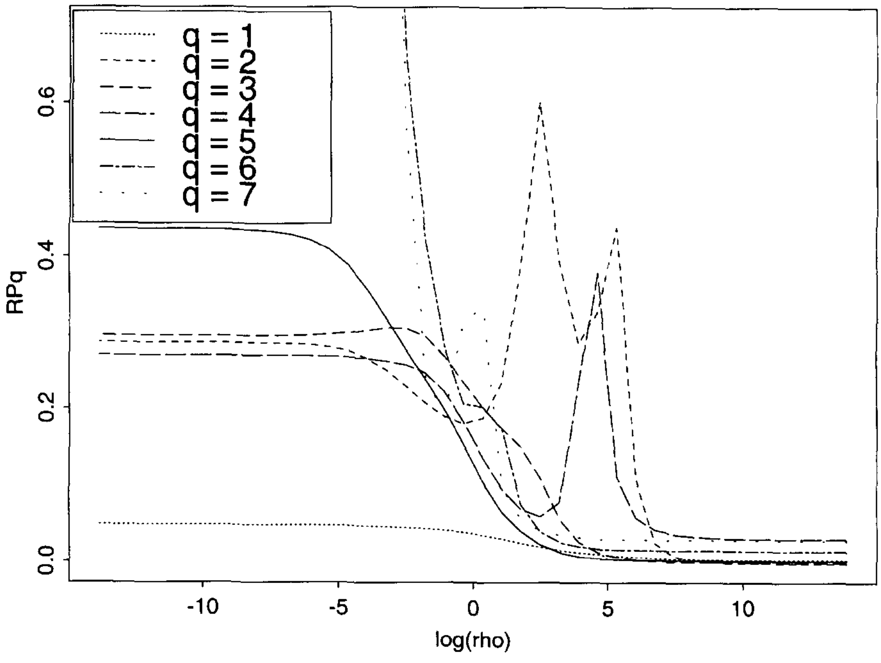

Fig. 6. The stability estimate $\widehat{R_{P_q}}$ versus $\log(\rho)$ for different dimensions.

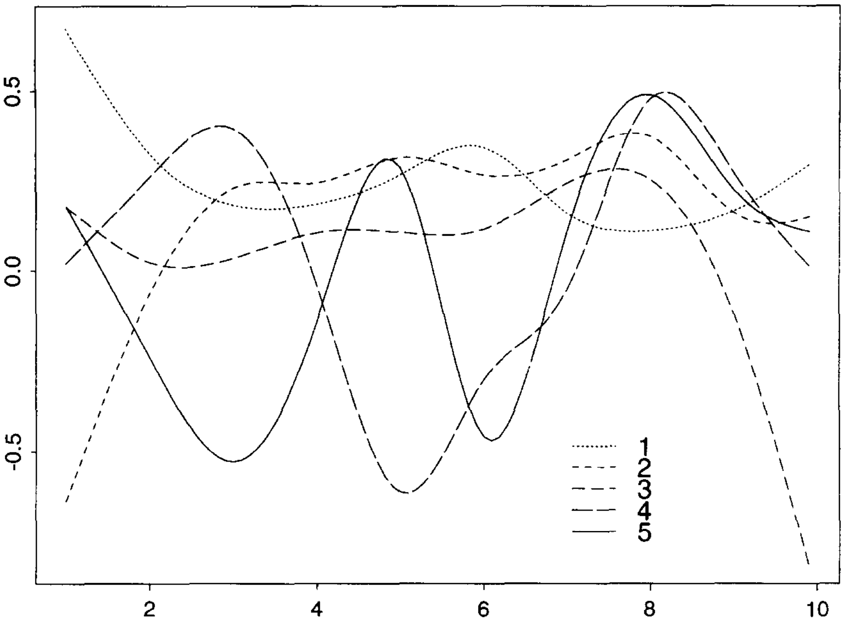

Fig. 7. The first five eigenfunctions providing $\widehat{P}_q$.

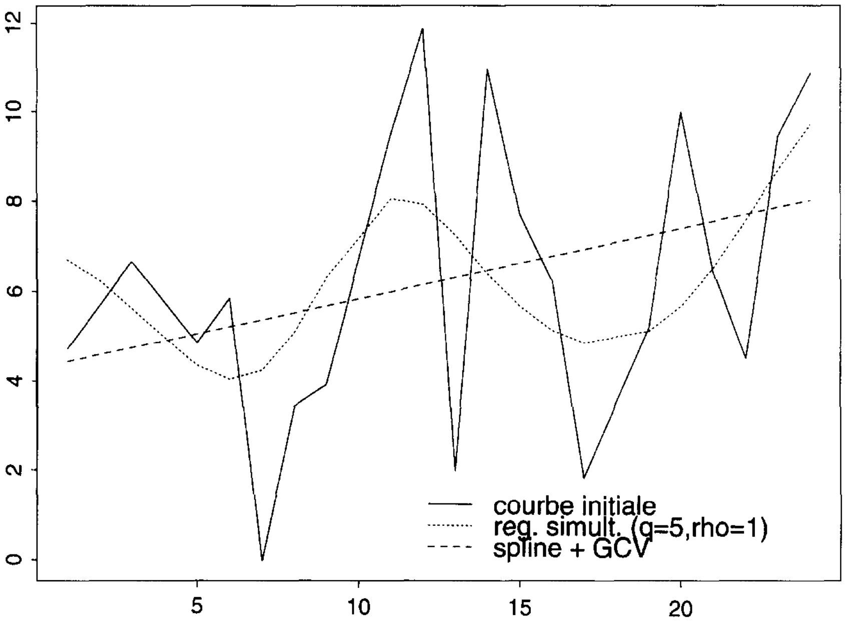

Fig. 8. Comparison between the simultaneous regression estimate ($q = 5$, $\rho = 1$) and the spline estimate (gcv).

by the drop in R_{P5}, we retained $q = 5$ and $\rho \approx 1$ corresponding to the minimum reached by R_{P5}.

In Fig. 7 the first five smooth eigenfunctions are given. Finally, in Fig. 8, a raw curve, its classical spline estimate and the proposed smoothed functional PCA estimate are compared. It clearly appears that the latter only is able to emphasis natural quasi-periodic components in spite of both missing data and such a large variability of the noise.

Appendix

Proof of Proposition 1. The criterion to minimize is decomposed as follows:

$$\frac{1}{n}\sum_{i=1}^{n}(\|s_i - u_i\|_C^2 + \rho\|u_i\|_G^2) = \frac{1}{n}\sum_{i=1}^{n}(\|x_i - (u_i - \bar{u})\|_C^2 + \rho\|u_i - \bar{u}\|_G^2)$$
$$+\|\bar{s} - \bar{u}\|_C^2 + \rho\|\bar{u}\|_G^2.$$

This leads us to estimate $\bar{u}$ by smoothing the empirical mean:

$$\hat{\bar{u}} = H_\rho C\bar{s}$$

and also to solve:

$$\min_{u_i \in A_q}\left\{\frac{1}{n}\sum_{i=1}^{n}(\|x_i - (u_i - \hat{\bar{u}})\|_C^2 + \rho\|u_i - \hat{\bar{u}}\|_G^2)\right\}, \tag{A.1}$$

where $A_q = \bar{s} + E_q$ such that E_q is a q-dimensional subspace of $\mathbb{R}^r$.

Let us define the transformation

$$\tilde{x}_i = H_\rho C x_i,$$

which can be interpreted as smoothing the centered coordinates. Now, let Q_i be the following quantity:

$$Q_i = \|x_i - (u_i - \hat{\bar{u}})\|_C^2 + \rho\|u_i - \hat{\bar{u}}\|_G^2.$$

Since for all v in $\mathbb{R}^r, v'Cv = \operatorname{tr} v'Cv = \operatorname{tr} vv'C$, we have

$$\|x_i - (u_i - \hat{\bar{u}})\|_C^2 = \operatorname{tr} x_i x_i' C - 2\operatorname{tr} x_i(u_i - \hat{\bar{u}})'C$$
$$+\operatorname{tr}(u_i - \hat{\bar{u}})(u_i - \hat{\bar{u}})'C;$$

then, we can write

$$Q_i = \operatorname{tr} H_\rho^{-1}\tilde{x}_i\tilde{x}_i'H_\rho^{-1}C^{-1} - 2\operatorname{tr}\tilde{x}_i(u_i - \hat{\bar{u}})'H_\rho^{-1} + \operatorname{tr}(u_i - \hat{\bar{u}})(u_i - \hat{\bar{u}})'H_\rho^{-1},$$

which can be decomposed as

$$\begin{aligned}Q_i &= \operatorname{tr}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i'\boldsymbol{H}_\rho^{-1} + \operatorname{tr}(\boldsymbol{u}_i - \hat{\boldsymbol{u}})(\boldsymbol{u}_i - \hat{\boldsymbol{u}})'\boldsymbol{H}_\rho^{-1}\\ &\quad -2\operatorname{tr}\tilde{\boldsymbol{x}}_i(\boldsymbol{u}_i - \hat{\boldsymbol{u}})'\boldsymbol{H}_\rho^{-1} + \operatorname{tr}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i'(\rho\boldsymbol{G} + \rho^2\boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G})\\ &= \|\tilde{\boldsymbol{x}}_i - (\boldsymbol{u}_i - \hat{\boldsymbol{u}})\|^2_{\boldsymbol{H}_\rho^{-1}} + \operatorname{tr}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i'(\rho\boldsymbol{G} + \rho^2\boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}).\end{aligned}$$

Since only the first term depends on $\boldsymbol{u}_i$, the problem (A.1) becomes

$$\min_{\boldsymbol{u}_i\in A_q}\left\{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\boldsymbol{x}}_i - (\boldsymbol{u}_i - \hat{\boldsymbol{u}})\|^2_{\boldsymbol{H}_\rho^{-1}}\right\}. \tag{A.2}$$

The solution is achieved by the spectral analysis of the matrix

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i' = \boldsymbol{H}_\rho\boldsymbol{C}\boldsymbol{\Gamma}\boldsymbol{C}\boldsymbol{H}_\rho$$

with respect to the metric $\boldsymbol{H}_\rho^{-1}$:

$$\boldsymbol{H}_\rho\boldsymbol{C}\boldsymbol{\Gamma}\boldsymbol{C}\tilde{\boldsymbol{V}} = \tilde{\boldsymbol{V}}\tilde{\boldsymbol{L}} \quad\text{and}\quad \tilde{\boldsymbol{V}}'\boldsymbol{H}_\rho^{-1}\tilde{\boldsymbol{V}} = \boldsymbol{I}, \tag{A.3}$$

where $\tilde{\boldsymbol{V}}$ is the matrix whose columns are the eigenvectors of $\boldsymbol{H}_\rho\boldsymbol{C}\boldsymbol{\Gamma}\boldsymbol{C}$, $\tilde{\boldsymbol{L}}$ is the diagonal matrix containing the eigenvalues and $\boldsymbol{I}$ is the identity matrix in $\mathbb{R}^r$.

Moreover, it is easy to infer from the above eigenequation (8) that

$$\boldsymbol{H}_\rho^{1/2}\boldsymbol{C}\boldsymbol{\Gamma}\boldsymbol{C}\boldsymbol{H}_\rho^{1/2}\boldsymbol{H}_\rho^{-1/2}\tilde{\boldsymbol{V}} = \boldsymbol{H}_\rho^{-1/2}\tilde{\boldsymbol{V}}\tilde{\boldsymbol{L}}.$$

Thus, setting $\boldsymbol{V} = \boldsymbol{H}_\rho^{-1/2}\tilde{\boldsymbol{V}}$, we have

$$\boldsymbol{H}_\rho^{1/2}\boldsymbol{C}\boldsymbol{\Gamma}\boldsymbol{C}\boldsymbol{H}_\rho^{1/2}\boldsymbol{V} = \boldsymbol{V}\tilde{\boldsymbol{L}} \quad\text{and}\quad \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}, \tag{A.4}$$

where $\boldsymbol{V}$ and $\tilde{\boldsymbol{L}}$ are obtained from the eigenanalysis of the real symmetric matrix

$$\boldsymbol{H}_\rho^{1/2}\boldsymbol{C}\boldsymbol{\Gamma}\boldsymbol{C}\boldsymbol{H}_\rho^{1/2}.$$

The projector $\tilde{\boldsymbol{P}}_q$ is then derived from $\tilde{\boldsymbol{V}}_q$,

$$\tilde{\boldsymbol{P}}_q = \tilde{\boldsymbol{V}}_q\tilde{\boldsymbol{V}}_q'\boldsymbol{H}_\rho^{-1},$$

and applied to $\boldsymbol{H}_\rho\boldsymbol{C}\boldsymbol{x}_i$. □

References

Becker, R., J. Chambers and A. Wilks, *The new S language, a programming environment for data analysis and graphics* (Wadsworth and Brooks/Cole, 1988).

Besse, P., Pca stability and choice of dimensionality, *Statist. Probab. Lett.*, **13** (1992) 405–410.

Besse, P., Models for multivariate analysis, in: R. Dutter and W. Grossman (Eds.), *Compstat 94* (Physica-Verlag, Wurzburg, 1994) 271–285.

Besse, P. and A.d. Falguerolles, Application of resampling methods to the choice of dimension in principal component analysis, in: W. Härdle and L. Simar (Eds.), *Computer intensive methods in statistics* (Physica-Verlag, Wurzburg, 1993) 167–176.

Besse, P. and L. Ferré, Sur l'usage de la validation croisée en analyse en composantes principales, *Rev. Statist. Appl.*, **XLI** (1993) 71–76.

Besse, P. and A. Pousse, Extension des analyses factorielles, in: J. Droesbeke et al. (Eds.), *Modèles pour l'analyse des données multidimensionnelles* (Economica, 1992) 129–158.

Besse, P. and J. Ramsay, Principal component analysis of sampled curves, *Psychometrika*, **51** (1986) 285–311.

Boularan, J., L. Ferré and P. Vieu, Growth curves: a two stage nonparametric approach, *J. Statist. Plann. Inference*, **38** (1993) 327–350.

Champely, S., Analyse de données fonctionnelles, aproximation par les splines de regression, Ph.D. Thesis (Université Lyon-1, France, 1994).

De Boor, C., *A practical guide to splines* (Springer, Berlin, 1978).

Denby, L. and C. Mallows, Smooth reduced-rank approximations, in: *I.S.I., 49th session, contributed papers*, Vol. 1 (1993) 355–356.

ECOSTAT, Banque de données statistiques pour l'enseignement, C.R.D.P. de Montpellier, Unité de la documentation Statistique (1991).

Jolliffe, I., *Principal component analysis* (Springer, Berlin, 1986).

Jones, M. and J. Rice, Displaying the important features of large collections of similar curves, *Amer. Statist.*, **46** (1992) 140–145.

Kato, T., *Perturbation theory for linear operator* (Springer, Berlin, 1966).

Kelly, C. and J. Rice, Monotone smoothing with application to dose-response curves and the assessment of synergism, *Biometrics*, **46** (1990) 1071–1085.

Kneip, A., Nonparametric estimation of common regressors for similar curve data, *Ann. Statist.*, **22** (1995) 1386–1427.

Krzanowski, W., Cross validation choice in principal components analysis, *Biometrics*, **43** (1987) 575–584.

Pezzulli, S. and B. Șilverman, On smoothed principal components analysis, *Comput. Statist.*, **8** (1993) 1–16.

Ramsay, J. and C. Dalzell, Some tools for functional data analysis, *J. Roy. Statist. Soc. Ser. B*, **53** (1991) 539–572; with discussion.

Rice, J. and B. Silverman, Estimating the mean and covariance structure nonparametrically when the data are curves, *J. Roy. Statist. Soc. Ser. B*, **53** (1991) 233–243.

Silverman, B., Smoothed functional principal components analysis by choice of norm, to be published.

Wahba, G., *Spline models for observational data* (SIAM, Philadelphia, PA, 1990).