The Annals of Statistics
1996, Vol. 24, No. 1, 1–24

SMOOTHED FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS BY CHOICE OF NORM[1]

BY BERNARD W. SILVERMAN

University of Bristol

The principal components analysis of functional data is often enhanced by the use of smoothing. It is shown that an attractive method of incorporating smoothing is to replace the usual L^2-orthonormality constraint on the principal components by orthonormality with respect to an inner product that takes account of the roughness of the functions. The method is easily implemented in practice by making use of appropriate function transforms (Fourier transforms for periodic data) and standard principal components analysis programs. Several alternative possible interpretations of the smoothed principal components as obtained by the method are presented. Some theoretical properties of the method are discussed: the estimates are shown to be consistent under appropriate conditions, and asymptotic expansion techniques are used to investigate their bias and variance properties. These indicate that the form of smoothing proposed is advantageous under mild conditions, indeed milder than those for existing methods of smoothed functional principal components analysis. The choice of smoothing parameter by cross-validation is discussed. The methodology of the paper is illustrated by an application to a biomechanical data set obtained in the study of the behaviour of the human thumb–forefinger system.

1. Introduction. Suppose we have data $X_1(t), \ldots, X_n(t)$ that are assumed to be drawn from a stochastic process X on a bounded interval $\mathcal{I}$, say. Data that are (or may be considered as being) of this kind, rather than the vectors of standard multivariate analysis, arise in an increasing number of fields of application. Rice and Silverman (1991) discussed a method for smoothed principal components analysis of such *functional* data, and some properties of this method were described by Pezzulli and Silverman (1993). For an important perspective on the analysis of functional data, see Ramsay and Dalzell (1991) and its published discussion.

In this paper, an alternative approach to functional principal components analysis (PCA) will be proposed and investigated. This approach is simpler both conceptually and computationally than the Rice–Silverman approach. We present both theoretical and empirical results which indicate that the

Received September 1994; revised March 1995.

[1]Supported by the British Science and Engineering Research Council (now EPSRC); part of the research was carried out during a visit to Stanford University, with support form NSF Grant DMS-92-09130.

AMS 1991 *subject classifications*. Primary 62G07, 62H25; secondary 65D10, 73P20.

Key words and phrases. Biomechanics, consistency, cross-validation, functional data analysis, mean integrated square error, PCA, roughness penalty, smoothing.

method allows the principal components of interest all to be estimated simultaneously with a single choice of smoothing parameter; in the Rice–Silverman method a sequence of estimation steps with decreasing smoothing parameters was necessary. We investigate conditions under which applying smoothing in the new framework will allow an improvement in accuracy of the principal component weight functions. These conditions are even milder than the corresponding conditions for the Rice–Silverman approach [Pezzulli and Silverman (1993)].

The method proposed has in common with the Rice–Silverman method the property of being a "nonparametric" method, in that the estimated principal component functions are not constrained a priori to lie in any particular finite-dimensional space; the data are allowed to speak for themselves to a greater extent in the estimation process than would be the case if, for example, we projected onto a finite-dimensional basis and then performed a standard multivariate analysis. The way in which the roughness penalty is included in the procedure is in some ways analogous to the approach of Leurgans, Moyeed and Silverman (1993) to the canonical correlation analysis of functional data.

The paper is set out as follows. In Section 2, we first of all establish notation and review the existing approach. In Section 3, the new method is set out, and the details of its implementation in practice are described. The estimated principal components produced by the method have various possible interpretations, and these are discussed in Section 4. In Section 5, the estimates are shown under suitable conditions to be consistent. The theoretical accuracy of the smoothing method is investigated in Section 6, where an argument based on asymptotic expansions is used to obtain approximations for the bias and variance of the estimators. These expressions are used to determine conditions under which smoothing is advantageous, and to find the ideal values of the smoothing parameter. In Section 7, a practical cross-validation method for the automatic choice of the smoothing parameter is set out, and in Section 8 the methodology of the paper is illustrated by its application to a set of biomechanical data collected in a study of the way the human thumb and forefinger squeeze an object.

2. Notation and more detailed background. We first of all set out some notation that will be useful throughout the paper. For simplicity, we shall assume that the mean function of the process X is known and has been subtracted off, so without loss of generality we assume that $EX(t) = 0$. Define Γ to be the covariance function

$$\Gamma(s,t) = EX(s)X(t)$$

and $\hat{\Gamma}$ to be the sample covariance

$$\hat{\Gamma}(s,t) = n^{-1} \sum_{i=1}^{n} X_i(s) X_i(t).$$

Given functions f and g, let (f, g) be the usual L^2 inner product

$$(f, g) = \int_{\mathcal{J}} f(t)g(t)\,dt.$$

We shall assume that Γ has an orthonormal expansion (in the L^2 sense) in terms of eigenfunctions γ_j, so that

$$\Gamma(s, t) = \sum_{j=1}^{\infty} \lambda_j \gamma_j(s)\gamma_j(t),$$

with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. The (γ_j, X) are the principal components of the stochastic process X. Dauxois, Pousse and Romain (1982) showed that the eigenfunctions and the eigenvalues of the sample covariance $\hat{\Gamma}$ are consistent estimators of the γ_j and λ_j, respectively, and obtained some asymptotic results for them. However, in many cases the principal component weight functions γ_j estimated purely by a decomposition of $\hat{\Gamma}$ are excessively variable or rough.

The approach adopted in this paper will be based on a roughness penalty. In order to quantify the "roughness" of a function f on $\mathcal{J}$, a roughness penalty such as $\int f''^2$ is used; for a detailed discussion of the use of roughness penalties in statistical problems, see, for example, Green and Silverman (1994). To define the roughness penalty, we shall assume that $\mathcal{S}$ is a space of suitably smooth functions on $\mathcal{J}$ and that R is a linear differential operator defined on $\mathcal{S}$. For functions f and g in $\mathcal{S}$, define the bilinear form $[f, g] = (Rf, Rg)$. The roughness penalty will be assumed to be $[f, f]$ for all f in $\mathcal{S}$.

Define the operator $Q = R^*R$, where R^* is the adjoint of R. Let $\mathcal{V}$ be the space of "very smooth" functions f for which R^*Rf is defined and falls in L^2. Then, provided $f \in \mathcal{S}$ and $g \in \mathcal{V}$, we will have $[f, g] = (Rf, Rg) = (f, Qg)$.

In our case, the space $\mathcal{S}$ will usually be the space of functions with square-integrable second derivative on $\mathcal{J}$, subject to periodic boundary conditions if appropriate. We will have $Rf = f''$ and

$$[f, g] = \int_{\mathcal{J}} f''(t)g''(t)\,dt.$$

The roughness penalty will be $\int_{\mathcal{J}} f''(t)^2\,dt$. In this case, subject to the periodic boundary conditions, the operator R is self-adjoint. The space $\mathcal{V}$ will be the space of functions g such that g and its first three derivatives are absolutely continuous on the periodic extension of $\mathcal{J}$, and $\int_{\mathcal{J}} g''''^2$ is finite. Integrating by parts twice, it can be seen at once that $\int_{\mathcal{J}} f''g'' = \int_{\mathcal{J}} fg''''$ for sufficiently regular functions f and g. If we are not assuming periodicity, then $\mathcal{V}$ will be the space of functions g for which g has square-integrable fourth derivative on $\mathcal{S}$ and g'' and g''' are zero at the boundaries. In either case Q is the fourth-derivative operator, and it is easy to check that $\int_{\mathcal{J}} f''g'' = \int_{\mathcal{J}} fg''''$, integrating by parts twice.

The approach of Rice and Silverman (1991) to smoothed functional principal components analysis is also based on a roughness penalty idea. Their

method depends on a sequence of smoothing parameters α_j; the estimate $\hat{\gamma}_j$ of the jth principal component weight function is found by maximizing

$$\int\int \gamma(s)\hat{\Gamma}(s,t)\gamma(t)\,ds\,dt - \alpha_j[\gamma,\gamma], \tag{1}$$

subject to the constraints $\int \gamma^2 = 1$ and $(\gamma, \hat{\gamma}_k) = 0$ for $k < j$. It has been found in practice to be appropriate to use smaller values of the smoothing parameter for higher-order principal components.

The basic idea of this approach, as with roughness penalty methods generally, is that the form (1) quantifies the trade-off between "fidelity" to the data (in this case the sample variance of the projection in the direction of γ) and roughness as measured by the roughness penalty. The smoothing parameter controls the relative importance of the two terms. In the next section, a different method for making this trade-off will be described.

3. Modifying the norm.

3.1. *Rearranging the roughness penalty*. The alternative approach investigated in this paper is obtained by a simple rearrangement of the maximization in (1). Rather than penalizing the variance of a principal component for roughness, we build the roughness penalty into the orthonormality constraint instead. For a given smoothing parameter α, define the inner product

$$(f,g)_\alpha = (f,g) + \alpha[f,g], \tag{2}$$

with corresponding squared norm $\|f\|_\alpha^2 = (f,f)_\alpha$. [These are of course slight generalizations of standard Sobolev inner products and norms; see Adams (1975).]

Consider, now, the effect of performing a principal component analysis imposing orthonormality with respect to the inner product $(\cdot,\cdot)_\alpha$. In other words we find a series of functions $\tilde{\gamma}_j$ such that $\tilde{\gamma}_j$ maximizes $\int\int \gamma(s)\hat{\Gamma}(s,t)\gamma(t)\,ds\,dt$ subject to $\|\gamma\|_\alpha^2 = 1$ and $(\gamma, \tilde{\gamma}_k)_\alpha = 0$ for $k < j$. We then estimate the jth eigenvalue of the variance operator by

$$\tilde{\lambda}_j = \int\int \tilde{\gamma}_j(s)\hat{\Gamma}(s,t)\tilde{\gamma}_j(t)\,ds\,dt. \tag{3}$$

In order to understand the motivation for this procedure, consider the leading eigenfunction first of all. In the Rice–Silverman procedure the idea was, subject to the eigenfunction being of fixed L^2-norm, to maximize "variance" minus "roughness" (with a suitable weighting). In the current procedure, we maximize "variance" subject to "L^2-norm" plus "roughness" being fixed. To make a fair comparison, let us consider scale-invariant versions of the procedure, where only the direction of the function matters. The Rice–Silverman approach will maximize

$$\frac{\mathrm{var}(\gamma, X) - \alpha_1[\gamma,\gamma]}{(\gamma,\gamma)},$$

while the new approach will maximize

$$\frac{\operatorname{var}(\gamma, X)}{(\gamma, \gamma) + \alpha[\gamma, \gamma]}.$$

In both cases roughness will be penalized, but in a somewhat different way.

3.2. *Practicalities.* In this subsection we set out an algorithm for performing the PCA in the way described above. Let us assume (as is the case in many applications) that the interval $\mathcal{J}$ may be considered as being periodic. We concentrate on the case where R is the second-derivative operator.

In the context of periodic boundary conditions the algorithm now works in terms of real Fourier transforms. The formulas obtained using complex Fourier transforms are slightly simpler, but if we wish to use a standard SPLUS routine such as `prcomp` for the principal component decomposition step below, then complex data will present problems.

Let ϕ_ν be a series of Fourier functions

$$\phi_\nu(t) = \begin{cases} 2^{1/2}|\mathcal{J}|^{-1/2} \sin\left(\dfrac{2\pi\nu t}{|\mathcal{J}|}\right), & \text{for } \nu > 0, \\ 2^{1/2}|\mathcal{J}|^{-1/2} \cos\left(\dfrac{2\pi\nu t}{|\mathcal{J}|}\right), & \text{for } \nu < 0, \end{cases}$$

$$\phi_0(t) = |\mathcal{J}|^{-1/2}.$$

Define ρ_ν to be the eigenvalues of the differential operator R, so that (since the Fourier functions are the eigenfunctions of R) $R\phi_\nu = \rho_\nu \phi_\nu$; since R is the second-derivative operator, we have

$$\rho_\nu = -\frac{4\pi^2\nu^2}{|\mathcal{J}|^2}.$$

Define the operator Q to be the fourth-derivative operator, so that Q has eigenvalues ρ_ν^2. Define an operator S by

$$S = (I + \alpha Q)^{-1/2}.$$

This simple operator notation means that S is an operator such that if f is any function in $\mathcal{J}$, we will have $(I + \alpha Q)S^2 f = f$, so that $S^2 f$ is a solution in $\mathcal{J}$ of the differential equation

$$g + \alpha g'''' = f.$$

In the periodic case the easiest way to write down S explicitly is in the Fourier domain: if $f = \sum f_\nu \phi_\nu$, then $Sf = \sum_\nu s_\nu f_\nu \phi_\nu$, where $s_\nu = (1 + \alpha\rho_\nu^2)^{-1/2}$.

It is also interesting to note that S^2 has the property that, given any function f, the minimum over g of $\int (f - g)^2 + \alpha[g, g]$ will be given by setting $g = S^2 f$. Thus S^2 corresponds in a certain sense to the usual spline smoothing operator [see Green and Silverman (1994)]. In spline smoothing one is given a sequence of values $f(t_i)$ (possibly subject to error) and the

smoother g is defined to be the minimizer of the penalized mean square error $n^{-1}\Sigma(f(t_i) - g(t_i))^2 + \alpha[g, g]$; the expression $\int (f - g)^2 + \alpha[g, g]$ minimized by $S^2 f$ is simply the continuous analog of the penalized mean square error, and in this sense $S^2 f$ can be regarded as the "spline smoother" of a continuously observed function f.

We can now develop an algorithm for PCA with respect to the inner product $(\cdot, \cdot)_\alpha$ in the periodic case. In the succeeding discussion, we use boldface letters to denote vectors of Fourier coefficients, so that $\mathbf{f}$ is the vector of coefficients f_ν of a function f. We use letters Q and S to denote the corresponding operators in either domain; in the Fourier domain they are, of course, diagonal matrices, with diagonal entries ρ_ν^2 and s_ν, respectively. It will always be entirely clear from the context whether Q and S represent matrices or linear operators on functions.

For any functions f and g, we have

$$(f, g)_\alpha = \mathbf{f}^T\mathbf{g} + \alpha\mathbf{f}^T Q\mathbf{g} = \mathbf{f}^T S^{-2}\mathbf{g}.$$

Let $\tilde{\Gamma}$ denote the sample covariance matrix of the Fourier transformed data,

$$\tilde{\Gamma} = n^{-1}\sum_i \mathbf{X}_i\mathbf{X}_i^T.$$

Suppose that the required estimated eigenfunctions are $\tilde{\gamma}_j$ and that $\tilde{\gamma}_j = Sa_j$. In the Fourier domain we will then have the corresponding vectors of coefficients $\tilde{\gamma}_j$ successively maximizing $\gamma^T\tilde{\Gamma}\gamma$ subject to orthonormality requirements of the form $\tilde{\gamma}_j^T S^{-2}\tilde{\gamma}_k = \delta_{jk}$. This implies that the vectors of Fourier coefficients $\mathbf{a}_j$ successively maximize $\mathbf{a}^T S\tilde{\Gamma}S\mathbf{a}$ subject to standard orthonormality $\mathbf{a}_j^T\mathbf{a}_k = \delta_{jk}$. The matrix $S\tilde{\Gamma}S$ is the covariance matrix of the "half-spline-smoothed" Fourier transforms $S\mathbf{X}_i$. Hence this yields the following algorithm, which is easily implemented in SPLUS.

1. Fourier-transform the data.
2. Operate by S (analogous to a "half-spline-smooth").
3. Perform a standard PCA on the resulting sample Fourier coefficients, and get eigenfunctions $\mathbf{a}_j$, say.
4. Let $\tilde{\boldsymbol{\gamma}}_j = S\mathbf{a}_j$.
5. Apply an inverse Fourier transform to $\tilde{\boldsymbol{\gamma}}_j$ to get the required estimated eigenfunctions $\tilde{\gamma}_j$.

In practice the algorithm is implemented by truncating the Fourier series at some point. Typically, but by no means necessarily, the data will be obtained by sampling the continuous curves at some regular rate, and the Fourier transforms will then be obtained by fast Fourier transformation of that data. An SPLUS program for carrying out the analysis is available by anonymous FTP from `ftp.statistics.bristol.ac.uk` in the directory `pub / reports / FDA`.

The principal components analysis of the half-smoothed data yields standard deviations σ_j, say, each of which is the sample standard deviation of the set $\{(a_j, SX_i): i = 1, \ldots, n\}$. Since $(a_j, SX_i) = (\tilde{\gamma}_j, X_i)$, we have $\sigma_j^2 = \tilde{\lambda}_j$, where

$\tilde{\lambda}_j$ are as defined in (3). We shall refer further to these standard deviations in Section 4.

In this section we have concentrated exclusively on the roughness penalty $\int f''^2$ with periodic boundary conditions on the functions in $\mathscr{S}$. To deal with more general roughness penalties and spaces of smooth functions, one possible approach is to replace the Fourier functions ϕ_ν by eigenfunctions of the operator $Q = R^*R$, and to let ρ_ν^2 be the eigenvalues of Q. At somewhat greater cost in linear algebra, one could expand, if it is more convenient, in a basis other than eigenfunctions of Q. The operator S would of course then not be diagonal in the basis ϕ_ν.

4. Some interpretations of the estimated PCA. It turns out that there are a number of interesting alternative interpretations of the smoothed PCA as we have developed it. We shall consider these in turn.

4.1. *The half-smoothed data.* By construction, the $\mathbf{a}_j$ define principal components of the half-spline-smoothed data SX_i. Other authors [e.g., Ramsay and Dalzell (1991)] have suggested carrying out functional principal components by first of all smoothing the data in some way and then carrying out a PCA. Our construction illustrates an interesting connection between the type of smoothing applied to the data and that implicitly used in the estimation of the principal components in that the following two procedures are equivalent:

1. Smooth the data by S, then perform a PCA, and then smooth the principal components by operating by S.
2. Estimate the principal components by the smoothed PCA procedure using the roughness penalty $[\cdot,\cdot]$.

4.2. *PCA of the smoothed data.* The $\tilde{\gamma}_j$ are principal components of the *smoothed* original data S^2X with respect to the inner product $(\cdot,\cdot)_\alpha$, in the following sense:

1. The $\tilde{\gamma}_j$ are orthonormal with respect to $(\cdot,\cdot)_\alpha$.
2. The sample correlation of $\{(\tilde{\gamma}_j, S^2X_i)_\alpha\}$ and $\{(\tilde{\gamma}_k, S^2X_i)_\alpha\}$ is zero for $j \neq k$.

Thus the smoothed data can be decomposed as a sum of uncorrelated terms orthogonal with respect to $(\cdot,\cdot)_\alpha$ as

$$S^2X_i = \sum_j \left(\tilde{\gamma}_j, S^2X_i\right)_\alpha \tilde{\gamma}_j. \tag{4}$$

The first property is immediate from the construction. To demonstrate the second, we know that the sample correlations of the $\{\tilde{\boldsymbol{\gamma}}_j^T\mathbf{X}_i\}$ are zero for varying j; we then have $\tilde{\boldsymbol{\gamma}}_j^T\mathbf{X}_i = \tilde{\boldsymbol{\gamma}}_j^TS^{-2}(S^2\mathbf{X}_i) = (\tilde{\gamma}_j, S^2X_i)_\alpha$.

The variances σ_j^2 can be interpreted as contributions to the variability of the *smoothed* data S^2X_i, measuring variability in the $\|\cdot\|_\alpha^2$-norm, which of course incorporates information about variability of derivatives. However, it

can be seen that the total variability in this sense is the sample variance of $\|S^2X_i\|_\alpha^2 = (SX_i, SX_i) = \|SX_i\|^2$, and so the smoothing will have more effect in the calculation of the total variance than the implicit roughening involved in the norm $\|\cdot\|_\alpha^2$.

It can also easily be seen from the above discussion that the $\tilde{\gamma}_j$ can be found successively by maximizing the sample variance of $\{(\gamma, S^2X_i)_\alpha\}$ subject to $\|\gamma\|_\alpha^2 = 1$ and $(\gamma, \tilde{\gamma}_k)_\alpha = 0$ for $k < j$.

Note that this is the usual sense in which PCA with respect to a particular inner product is understood. The PCA as we have carried it out in Section 3 is a hybrid procedure, in that we consider the variances of (γ, X) but we impose orthonormality with respect to $(\cdot,\cdot)_\alpha$.

4.3. *A biorthogonal expansion of the original data.* One of the most instructive ways of viewing the PCA we have derived is in terms of a "biorthogonal" expansion of the original data. Define functions $b_j = S^{-1}a_j = S^{-2}\tilde{\gamma}_j$, in practice most easily found in the Fourier domain. Then it follows from (4) that

$$X_i = \sum_j \left\{ \left(\tilde{\gamma}_j, S^2X_i\right)_\alpha b_j \right\} = \sum_j (\tilde{\gamma}_j, X_i) b_j. \tag{5}$$

Although the functions b_j are not orthogonal in the usual sense, this expansion does indeed give a decomposition of the observations into effects that are uncorrelated with one another. Thus the usual interpretation of PCA as yielding "modes of variability" of the data remains. The price that is paid for the nonorthogonality of the b_j is that the principal component scores are obtained by taking inner products with the $\tilde{\gamma}_j$, which are of course smoothed versions of the b_j. The variances σ_j^2 are the sample variances of the coefficients in the expansions in (5).

4.4. *PCA of the data with respect to a dual inner product.* Another interpretation of the roughened principal components b_j can be given by considering an inner product dual to $(\cdot,\cdot)_\alpha$. Given functions f and g, define

$$\langle\langle f, g \rangle\rangle = \mathbf{f}^T S^2 \mathbf{g} = (Sf, Sg) = \left(S^2 f, S^2 g\right)_\alpha.$$

It then follows from the discussion of Section 4.2 that the b_j are the principal components of the original data with respect to the norm generated by $\langle\langle \cdot,\cdot \rangle\rangle$. Since $b_j = S^{-2}\tilde{\gamma}_j$, the b_j successively maximize the sample variance of $\{\langle\langle b, X_i\rangle\rangle\}$ subject to $\langle\langle b, b\rangle\rangle = 1$ and $\langle\langle b, b_k\rangle\rangle = 0$ for $k < j$, as required. The expansion (5) can be rewritten as

$$X_i = \sum_j \langle\langle b_j, X_i\rangle\rangle b_j.$$

Relative to this norm, the variances σ_j^2 do indeed quantify the contribution of the jth principal component to the overall variability of the original data.

5. Consistency results.

5.1. *Preliminary remarks.* In this section, the consistency of the proposed method is proved, under suitable conditions.

Before embarking on the proof, some remarks about the role of consistency proofs and the framework within which we are working may be helpful. In the author's view, asymptotic results, such as consistency proofs and also results on rate of convergence, should be seen not as *limiting* results (which are of course of no immediate practical use, because all real samples are finite) but as large-sample *approximation* results, which are of great use as an aid to our general intuition about the method and are often of direct use when we have a reasonably large amount of data. Of course, if [e.g., following Tukey (1977)] one views smoothing methods as data-analytic procedures without an underlying probability model, then there is little meaning in any finite-sample or asymptotic results based on models. However, in the author's view it is ultimately more fruitful to consider smoothing methods as *model-based methods for data exploration and summary*, in which case theoretical results are clearly relevant. For further discussion of the distinction between model-based and non-model-based approaches see, for example, Green and Silverman (1994).

In the case of functional data analysis, consistency results are particularly important because in some contexts certain "obvious" procedures are not consistent and do not give meaningful information about the data under consideration. See Leurgans, Moyeed and Silverman (1993) for an example where a consistency proof is valuable in distinguishing between useful and misleading approaches.

Finally, we remark that the asymptotic framework in which we shall work is to assume we have an increasing number of observations, each of which is a continuously observed function. Of course in practice functional observations can only ever be made discretely, but the whole aim of functional data analysis is to gain additional intuition by considering functions as single observations in function space rather than as high-dimensional vectors. With modern data collection techniques it is in any case very often the case that the data are observed at extremely rapid sampling rates and so are most naturally considered as being continuously observed.

5.2. *Statement of assumptions and the consistency theorem.* Suppose that we have independent identically distributed observations X_i drawn from a finite-variance stochastic process X defined on a compact set $\mathcal{I}$. Except where otherwise stated, our notation is as defined in Sections 2 and 3. Our proof applies to a more general roughness penalty than $\int f''^2$; we can allow $[f,g]$ to be any nonnegative-definite, symmetric bilinear form defined on a subspace $\mathcal{S}$ of $L^2(\mathcal{I})$. We then pursue all the definitions of Section 3, substituting this more general roughness penalty. The choice of such matters as the boundary conditions imposed on "smooth" functions is then governed by the specification of the subspace $\mathcal{S}$.

We shall make the following assumptions:

1. The covariance function Γ is strictly positive-definite, and the trace of Γ, $\int \Gamma(s, s)\, ds$, is finite. It then follows [see, e.g., Taylor and Lay (1980)] that there is a complete sequence of eigenfunctions γ_j of Γ, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots > 0$.
2. Each of the eigenfunctions γ_j falls in $\mathscr{S}$ and hence has finite roughness $[\gamma_j, \gamma_j]$.
3. All the eigenvalues λ_j have multiplicity 1, so that $\lambda_1 > \lambda_2 > \cdots > 0$. The method we describe can be extended to deal with the case of multiple eigenvalues but for simplicity we shall not do this.

THEOREM 1. *Define the functions $\tilde{\gamma}_j$ as in Section* 3, *and set*

$$\gamma_j^* = \tilde{\gamma}_j / \|\tilde{\gamma}_j\| \quad \textit{for each } j. \tag{6}$$

Assume that $\alpha \to 0$ as $n \to \infty$ and that the assumptions set out above hold. Then, for each j, with probability 1,

$$\tilde{\lambda}_j \to \lambda_j \quad \textit{as } n \to \infty$$

and

$$(\gamma_j^*, \gamma_j)^2 \to 1 \quad \textit{as } n \to \infty. \tag{7}$$

REMARKS. Throughout this section, convergence of any random quantity will be taken to be convergence with probability 1 (in other words, convergence almost surely) and will always be as $n \to \infty$. Note that the conclusion (7) is equivalent to either of the following conclusions:

(i) The sign of γ_j^* can be chosen to ensure that $\|\gamma_j^* - \gamma_j\| \to 0$.
(ii) Defining P_γ as the projection onto the subspace generated by γ,

$$P_\gamma x = (\gamma, \gamma)^{-1}(\gamma, x)\gamma, \tag{8}$$

we have

$$\|P_{\tilde{\gamma}_j} - P_{\gamma_j}\| \to 0. \tag{9}$$

5.3. *Proof of Theorem* 1.

General structure. The proof of Theorem 1 is carried out by induction. For any positive integer k, define the statement $\mathscr{H}_k$ to be the following three convergences as $n \to \infty$:

$$\tilde{\lambda}_k \to \lambda_k; \tag{10}$$

$$\alpha[\tilde{\gamma}_k, \tilde{\gamma}_k] \to 0; \quad \text{and} \tag{11}$$

$$(\gamma_k^*, \gamma_k)^2 \to 1. \tag{12}$$

Because $\|\tilde{\gamma}_k\|_\alpha^2 = 1$, the limit (11) is equivalent to

$$\|\tilde{\gamma}_k\|^2 \to 1. \tag{13}$$

The limit (12) is, as noted above, equivalent to the limit

$$\|P_{\tilde{\gamma}_k} - P_{\gamma_k}\| \to 0. \tag{14}$$

We shall prove the following: for any $j \geq 1$, if $\mathscr{H}_k$ is true for all $k < j$, then $\mathscr{H}_j$ is true also. In the case $j = 1$ our proof will demonstrate the unconditional truth of $\mathscr{H}_1$. In the case of any $j > 1$ we will have shown that if $\mathscr{H}_k$ is true for $k = 1, \ldots, j-1$, then $\mathscr{H}_k$ is true for $k = 1, \ldots, j-1, j$. By induction this will demonstrate the truth of $\mathscr{H}_j$ for all j, completing the proof of Theorem 1. [This form of the principle of mathematical induction is sometimes called the *principle of complete induction*; see, e.g., Spivak (1967), page 23.]

Details of the proof. Now consider any fixed $j \geq 1$, and assume inductively that $\mathscr{H}_k$ is true for all $k < j$. (If $j = 1$ there is nothing to assume.)

Let P be the projection perpendicular to $\gamma_1, \ldots, \gamma_{j-1}$,

$$Px = x - \sum_{k=1}^{j-1} (x, \gamma_k)\gamma_k,$$

and let $\tilde{P}$ be the projection in the space $\mathscr{S}$ in the $(\cdot,\cdot)_\alpha$ inner product perpendicular to $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_{j-1}$,

$$\tilde{P}x = x - \sum_{k=1}^{j-1} (x, \tilde{\gamma}_k)_\alpha \tilde{\gamma}_k \quad \text{for } x \text{ in } \mathscr{S}.$$

For $j = 1$, the sums are empty and we define P and $\tilde{P}$ each to be the identity.

We can now demonstrate the closeness of P and $\tilde{P}$ in a useful sense.

LEMMA 1. *For any $j \geq 1$, suppose that $\mathscr{H}_k$ holds for all $k < j$. Then, defining P and $\tilde{P}$ as above,*

$$\sup_{\|x\|_\alpha \leq 1} |(Px, \Gamma Px) - (\tilde{P}x, \hat{\Gamma}\tilde{P}x)| \to 0. \tag{15}$$

PROOF. Consider $j \geq 2$. Confining attention throughout to x in $\mathscr{S}$, for each $k < j$, using the triangle and Cauchy–Schwarz inequalities and the fact that $\|x\| \leq \|x\|_\alpha$, we have

$$\begin{aligned}
&\sup_{\|x\|_\alpha \leq 1} \|(x, \gamma_k)\gamma_k - (x, \tilde{\gamma}_k)_\alpha \tilde{\gamma}_k\| \\
&\quad \leq \sup_{\|x\|_\alpha \leq 1} \|(x, \gamma_k)\gamma_k - (x, \tilde{\gamma}_k)\tilde{\gamma}_k\| + \sup_{\|x\|_\alpha \leq 1} \alpha[x, \tilde{\gamma}_k]\|\tilde{\gamma}_k\| \\
&\quad \leq \sup_{\|x\| \leq 1} \|(x, \gamma_k)\gamma_k - (x, \gamma_k^*)\gamma_k^*\| + \sup_{\|x\| \leq 1} \|(x, \gamma_k^*)\gamma_k^* - (x, \tilde{\gamma}_k)\tilde{\gamma}_k\| \\
&\qquad + \sup_{\|x\|_\alpha \leq 1} (\alpha[x, x])^{1/2} \big(\alpha[\tilde{\gamma}_k, \tilde{\gamma}_k]\big)^{1/2} \|\tilde{\gamma}_k\| \qquad (16) \\
&\quad = \|P_{\gamma_k} - P_{\tilde{\gamma}_k}\| + \big(1 - \|\tilde{\gamma}_k\|^2\big) + o(1). \qquad (17)
\end{aligned}$$

The quantity in (16) is $o(1)$ because of the inductive assumption (11) and the facts that $\alpha[x, x] \leq \|x\|_\alpha^2$ and $\|\tilde{\gamma}_k\| \leq 1$. Substituting the alternative forms (14) and (13) of the inductive hypotheses into (17) now gives

$$\sup_{\|x\|_\alpha \leq 1} \|(x, \gamma_k)\gamma_k - (x, \tilde{\gamma}_k)_\alpha \tilde{\gamma}_k\| \to 0 \quad \text{for all } k < j. \tag{18}$$

It follows by summing over $k < j$ that

$$\sup_{x \in \mathscr{S}} \frac{\|(\tilde{P} - P)x\|}{\|x\|_\alpha} \to 0. \tag{19}$$

In the case $j = 1$, (19) of course holds trivially.

Dauxois, Pousse and Romain (1982) showed, by appealing to the strong law of the large numbers in Hilbert space, that $\|\hat{\Gamma} - \Gamma\| \to 0$. By combining this result with the uniform convergence result (19), we have

$$\sup_{\|x\|_\alpha \leq 1} |(Px, \Gamma Px) - (\tilde{P}x, \hat{\Gamma}\tilde{P}x)| \to 0,$$

completing the proof of the lemma. □

We can now prove the three parts of $\mathscr{H}_j$ successively.

PROOF OF (10) FOR $k = j$. The maximum value over $\|\gamma\| \leq 1$ of $(P\gamma, \Gamma P\gamma)$ is λ_j and is attained at γ_j, and the maximum of $(\tilde{P}\gamma, \hat{\Gamma}\tilde{P}\gamma)$ over $\|\gamma\|_\alpha \leq 1$ is $\tilde{\lambda}_j$, and is attained at $\tilde{\gamma}_j$. We therefore have

$$\lambda_j = (P\gamma_j, \Gamma P\gamma_j) \geq (P\gamma_j^*, \Gamma P\gamma_j^*) \geq (P\tilde{\gamma}_j, \Gamma P\tilde{\gamma}_j) \tag{20}$$

$$= (\tilde{P}\tilde{\gamma}_j, \hat{\Gamma}\tilde{P}\tilde{\gamma}_j) + o(1) = \tilde{\lambda}_j + o(1), \tag{21}$$

using property (15). On the other hand, we have, again using (15),

$$\begin{aligned} \tilde{\lambda}_j = (\tilde{P}\tilde{\gamma}_j, \hat{\Gamma}\tilde{P}\tilde{\gamma}_j) &\geq \left(\frac{\tilde{P}\gamma_j}{\|\gamma_j\|_\alpha}, \frac{\hat{\Gamma}\tilde{P}\gamma_j}{\|\gamma_j\|_\alpha}\right) \\ &\geq \left(\frac{P\gamma_j}{\|\gamma_j\|_\alpha}, \frac{\Gamma P\gamma_j}{\|\gamma_j\|_\alpha}\right) + o(1) \\ &= \frac{\lambda_j}{\|\gamma_j\|_\alpha^2} + o(1) = \lambda_j + o(1); \end{aligned} \tag{22}$$

the fact that $\|\gamma_j\|_\alpha \to 1$ holds since $\alpha \to 0$ and $[\gamma_j, \gamma_j]$ remains fixed. Combining (21) and (22) now completes the proof of (10) for $k = j$. □

PROOF OF (11) FOR $k = j$. Since all the inequalities in (20) tend to equalities, it also follows that

$$\alpha\big[\tilde{\gamma}_j, \tilde{\gamma}_j\big] = 1 - \|\tilde{\gamma}_j\|^2 = 1 - \frac{\big(P\tilde{\gamma}_j, \Gamma P\tilde{\gamma}_j\big)}{\big(P\gamma_j^*, \Gamma P\gamma_j^*\big)} \to 0,$$

completing the proof of (11), the second part of $\mathscr{H}_k$, for $k = j$. □

PROOF OF (12) FOR $k = j$. We first consider $j \geq 2$ and set $x = \tilde{\gamma}_j$ in (18). Since $\|\tilde{\gamma}_j\|_\alpha = 1$ and, for each $k < j$,

$$\|\big(\tilde{\gamma}_j, \gamma_k\big)\gamma_k - \big(\tilde{\gamma}_j, \tilde{\gamma}_k\big)_\alpha \tilde{\gamma}_k\| = \|\big(\tilde{\gamma}_j, \gamma_k\big)\gamma_k\| = |\big(\tilde{\gamma}_j, \gamma_k\big)|,$$

it follows from (13) for $k = j$ and from (18) that, for each $k < j$,

$$\lim(\gamma_k, \gamma_j^*) = \lim\big(\gamma_k, \tilde{\gamma}_j\big) = 0,$$

so that

$$\sum_{k<j} (\gamma_k, \gamma_j^*)^2 \to 0. \tag{23}$$

For $j = 1$, (23) is trivially true since the sum is empty.

We now consider the expansion of γ_j^* in terms of the complete orthonormal sequence γ_i. We have

$$P\gamma_j^* = P\sum_{i=1}^{\infty} (\gamma_j^*, \gamma_i)\gamma_i = \sum_{i=j}^{\infty} (\gamma_j^*, \gamma_i)\gamma_i \tag{24}$$

since $P\gamma_i = 0$ for $i < j$ and 1 for $i \geq j$. Now using the fact that $\Gamma\gamma_i = \lambda_i\gamma_i$ yields

$$\Gamma P\gamma_j^* = \sum_{i=j}^{\infty} \lambda_i(\gamma_j^*, \gamma_i)\gamma_i. \tag{25}$$

Putting (24) and (25) together, and using the orthonormality of the γ_i, now gives

$$\big(P\gamma_j^*, \Gamma P\gamma_j^*\big) = \sum_{i\geq j} \lambda_i(\gamma_i, \gamma_j^*)^2. \tag{26}$$

The fact that P is a projection implies that $\|P\gamma_j^*\|^2 \leq \|\gamma_j^*\|^2 = 1$; combining this with (26) now gives

$$\begin{aligned} \lambda_j - \big(P\gamma_j^*, \Gamma\gamma_j^*\big) &\geq \lambda_j\|P\gamma_j^*\|^2 - \sum_{i\geq j} \lambda_i(\gamma_i, \gamma_j^*)^2 \\ &= \sum_{i>j} \big(\lambda_j - \lambda_i\big)(\gamma_i, \gamma_j^*)^2 \\ &\geq \big(\lambda_j - \lambda_{j+1}\big)\sum_{i>j} (\gamma_i, \gamma_j^*)^2 \geq 0 \end{aligned} \tag{27}$$

since (λ_i) is a decreasing sequence.

Since all the inequalities in (20) tend to equalities, $\lambda_j - (P\gamma_j^*, \Gamma\gamma_j^*) \to 0$ and so all the inequalities in (27) tend to equalities. Since $\lambda_j \neq \lambda_{j+1}$, it follows that

$$\sum_{i>j} (\gamma_i, \gamma_j^*)^2 \to 0. \tag{28}$$

Combining (23) and (28) with the property that $\Sigma_i(\gamma_i, \gamma_j^*)^2 = \|\gamma_j^*\|^2 = 1$ demonstrates that $(\gamma_j, \gamma_j^*)^2 \to 1$, completing the proof of (12) for $k = j$. By the inductive argument laid out above, the proof of Theorem 1 is now complete. □

6. The effect of smoothing. In this section, some heuristic calculations are carried out to investigate the effect that smoothing, in the sense we have described, has on the estimation of the eigenfunctions and eigenvalues of Γ. The calculations parallel those given in Pezzulli and Silverman (1993) for the Rice–Silverman method of smoothing. They will demonstrate that the method of smoothing proposed in this paper is appropriate under even milder conditions than the Rice–Silverman method.

6.1. *Asymptotic expansions*. We shall concentrate on the estimation of any particular eigenfunction γ with eigenvalue λ, and we shall let $\tilde{\gamma}$ and $\tilde{\lambda}$ be the estimates as defined in Section 3. The basic idea of our heuristic calculations is to consider an asymptotic expansion to find the leading bias and variance terms in both $\tilde{\lambda}$ and $\tilde{\gamma}$. Because we have already shown that the method is consistent, it is reasonable to assume that an asymptotic expansion will have good approximation properties.

It will be assumed that γ is the kth eigenfunction γ_k of Γ and that it corresponds to an eigenvalue λ_k of multiplicity 1. The subscript k will be omitted almost throughout. We shall measure accuracy in the estimation of γ in terms of integrated square error. We shall assume that the eigenfunctions γ_j all fall in the space of "very smooth" functions $\mathscr{V}$.

The asymptotic expansions are carried out as follows. It can be seen from the definition of $\tilde{\gamma}_j$ that the $\tilde{\gamma}_j$ are solutions of the generalized eigenproblem

$$\hat{\Gamma}\tilde{\gamma} = \tilde{\lambda}(I + \alpha Q)\tilde{\gamma}. \tag{29}$$

We set $\varepsilon = n^{-1/2}$, because in various senses the difference between $\hat{\Gamma}$ and Γ is exactly of order ε; specifically, the covariance structure of the process $n^{1/2}(\hat{\Gamma} - \Gamma)$ does not vary with n.

The eigenproblem (29) is a perturbation of the eigenproblem $\Gamma\gamma = \lambda\gamma$ in two ways, in that the matrices on both sides of the equation are subject to small perturbations. With this in mind, we expand (29) by setting

$$\begin{aligned}
\hat{\Gamma} &= \Gamma + \varepsilon\Delta, \\
\tilde{\gamma} &= \gamma + \varepsilon\gamma^{(1)} + \alpha\gamma^{(2)} + \varepsilon^2\gamma^{(11)} + \varepsilon\alpha\gamma^{(12)} + \cdots, \\
\tilde{\lambda} &= \lambda + \varepsilon\lambda^{(1)} + \alpha\lambda^{(2)} + \varepsilon^2\lambda^{(11)} + \varepsilon\alpha\lambda^{(12)} + \cdots .
\end{aligned}$$

Set

$$\rho = [\gamma, \gamma] = (\gamma, Q\gamma).$$

Write P for the projection P_γ onto the space generated by γ. Define Π to be the mapping onto the space perpendicular to γ given by

$$\Pi = \sum_{j \neq k} (\lambda - \lambda_j)^{-1} P_{\gamma_j} \tag{30}$$

so that, for any x, $\Pi(\lambda - \Gamma)x$ is the projection of x onto the space perpendicular to γ, namely, $(I - P)x$. Note also that $\Pi\gamma = 0$, so that $\Pi(I - P) = \Pi$.

We will need to substitute the expansions of $\tilde{\gamma}$ and $\tilde{\lambda}$ in two equations. The first is the eigenfunction condition (29), which gives

$$\begin{aligned}(\Gamma + \varepsilon\Delta)(\gamma + \varepsilon\gamma^{(1)} + \alpha\gamma^{(2)} + \cdots) \\ = (\lambda + \varepsilon\lambda^{(1)} + \alpha\lambda^{(2)} + \cdots)(I + \alpha Q) \\ \times(\gamma + \varepsilon\gamma^{(1)} + \alpha\gamma^{(2)} + \cdots),\end{aligned} \tag{31}$$

and the second is the normalization condition $(\tilde{\gamma}, (I + \alpha Q)\tilde{\gamma}) = 1$, which becomes

$$\big((\gamma + \varepsilon\gamma^{(1)} + \alpha\gamma^{(2)} + \cdots), (I + \alpha Q)(\gamma + \varepsilon\gamma^{(1)} + \alpha\gamma^{(2)} + \cdots)\big) = 1. \tag{32}$$

Our strategy will now be to match the coefficients of powers of α and ε in (31) and (32). These will give various expressions for the terms in the expansions of $\tilde{\gamma}$ and $\tilde{\lambda}$, which will then be used to investigate the mean integrated squared error properties of $\tilde{\gamma}$ as an estimator of γ. We shall not give explicit expressions for all the terms considered but will confine ourselves to obtaining properties that will be needed in the summing-up discussion in Section 6.2.

Terms in ε. Matching terms in ε in (31) and (32), respectively, gives

$$\Gamma\gamma^{(1)} + \Delta\gamma = \lambda\gamma^{(1)} + \lambda^{(1)}\gamma \tag{33}$$

and

$$(\gamma, \gamma^{(1)}) = 0. \tag{34}$$

The second equality follows from the property $(\gamma, \gamma) = 1$ and can be written as $P\gamma^{(1)} = 0$. We now take the inner product of γ with (33). Using the fact that $(\gamma, \Gamma\gamma^{(1)}) = (\Gamma\gamma, \gamma^{(1)}) = (\lambda\gamma, \gamma^{(1)})$, this yields

$$\lambda^{(1)} = (\gamma, \Delta\gamma).$$

We also have

$$(\lambda - \Gamma)\gamma^{(1)} = -\lambda^{(1)}\gamma + \Delta\gamma.$$

Operating by Π and using the fact that $(I - P)\gamma^{(1)} = \gamma^{(1)}$ then yields

$$\gamma^{(1)} = \Pi\Delta\gamma. \tag{35}$$

Terms in α. We now match the terms in α in the two expansions (31) and (32). This yields

$$\Gamma\gamma^{(2)} = \lambda\gamma^{(2)} + \lambda^{(2)}\gamma + \lambda Q\gamma \tag{36}$$

and

$$2(\gamma, \gamma^{(2)}) + (\gamma, Q\gamma) = 0. \tag{37}$$

As before, taking the inner product of γ with (36) gives

$$\lambda^{(2)} = -\lambda(\gamma, Q\gamma) = -\lambda\rho; \tag{38}$$

rearranging (36) to give an expression for $(\lambda - \Gamma)\gamma^{(2)}$ and then operating by Π gives

$$(I - P)\gamma^{(2)} = -\lambda\Pi Q\gamma.$$

The component of $\gamma^{(2)}$ parallel to γ is given from (37); we have

$$P\gamma^{(2)} = (\gamma, \gamma^{(2)})\gamma = -\tfrac{1}{2}(\gamma, Q\gamma)\gamma = -\tfrac{1}{2}\rho\gamma,$$

so that $\gamma^{(2)} = -\frac{1}{2}\rho\gamma - \lambda\Pi Q\gamma$ and

$$\|\gamma^{(2)}\|^2 = \tfrac{1}{4}\rho^2 + \lambda^2\|\Pi Q\gamma\|^2. \tag{39}$$

Terms in $\alpha\varepsilon$. We now consider terms in $\alpha\varepsilon$. We will only need an expression for $(I - P)\gamma^{(12)}$, which we shall find by following the same steps as previously. From (31) we have

$$\begin{aligned}\Gamma\gamma^{(12)} + \Delta\gamma^{(2)} = {} & \lambda\gamma^{(12)} + \lambda Q\gamma^{(1)} + \lambda^{(1)}\gamma^{(2)} + \lambda^{(1)}Q\gamma \\ & + \lambda^{(2)}\gamma^{(1)} + \lambda^{(12)}\gamma.\end{aligned} \tag{40}$$

Rearranging (40) and operating by Π gives

$$\begin{aligned}(I - P)\gamma^{(12)} &= \Pi(\lambda - \Gamma)\gamma^{(12)} \\ &= \Pi\,\Delta\gamma^{(2)} - \lambda\Pi Q\gamma^{(1)} - \lambda^{(1)}\Pi\gamma^{(2)} - \lambda^{(1)}\Pi Q\gamma - \lambda^{(2)}\Pi\gamma^{(1)}.\end{aligned}$$

We now use the property that $\Pi = \Pi(I - P)$ and the values we have already derived for $\lambda^{(1)}$, $\lambda^{(2)}$, $\gamma^{(1)}$ and $\gamma^{(2)}$ to yield

$$\begin{aligned}(I - P)\gamma^{(12)} = {} & -\lambda\Pi\,\Delta\Pi Q\gamma - \tfrac{1}{2}\rho\Pi\,\Delta\gamma - \lambda\Pi Q\Pi\,\Delta\gamma + \lambda(\gamma, \Delta\gamma)\Pi^2 Q\gamma \\ & -(\gamma, \Delta\gamma)\Pi Q\gamma + \lambda\rho\Pi^2\,\Delta\gamma \\ = {} & -\lambda\Pi\,\Delta\Pi Q\gamma - \tfrac{1}{2}\rho\Pi\,\Delta\gamma - \lambda\Pi Q\Pi\,\Delta\gamma + \lambda\Pi^2 QP\,\Delta\gamma \\ & - \Pi QP\,\Delta\gamma + \lambda\rho\Pi^2\,\Delta\gamma.\end{aligned} \tag{41}$$

Terms in ε^2. Matching terms in ε^2 and going through the familiar manipulations gives

$$\Gamma\gamma^{(11)} + \Delta\gamma^{(1)} = \lambda\gamma^{(11)} + \lambda^{(1)}\gamma^{(1)} + \lambda^{(11)}\gamma \tag{42}$$

and

$$(\gamma, \gamma^{(11)}) = -\tfrac{1}{2}(\gamma^{(1)}, \gamma^{(1)}) = -\tfrac{1}{2}\|\Pi\,\Delta\gamma\|^2.$$

It follows from (42) that

$$(I - P)\gamma^{(11)} = \Pi(\Delta - \lambda^{(1)})\gamma^{(1)} = \Pi\,\Delta\Pi\,\Delta\gamma - \Pi^2\,\Delta P\,\Delta\gamma.$$

6.2. *Mean integrated square error calculations.* In order to proceed we need some moment properties of Δ. It is clear that $E\Delta = 0$. In order to deal with quadratic forms, suppose that A is any self-adjoint operator. Then, using the fact that

$$\operatorname{cov}(X(s)X(t), X(u)X(v)) = \Gamma(s,u)\Gamma(t,v) + \Gamma(s,v)\Gamma(t,u),$$

we can conclude that

$$E(\Delta A\Delta) = \Gamma A\Gamma + \operatorname{tr}(A\Gamma)\Gamma. \tag{43}$$

For further details, see Pezzulli and Silverman (1993).

Applying (43) to find $E(I - P)\gamma^{(11)}$ yields the value zero, since the relation $\Pi\gamma = 0$ eliminates all four terms one has to consider. Therefore we have

$$E\gamma^{(11)} = -\tfrac{1}{2}\gamma E(\Pi\,\Delta\gamma, \Pi\,\Delta\gamma) = -\tfrac{1}{2}\gamma\big(\gamma, E\,\Delta\Pi^2\,\Delta\gamma\big) = -\tfrac{1}{2}\lambda\operatorname{tr}(\Pi^2\Gamma)\gamma,$$

and hence, since $\|\gamma\|^2 = 1$, we have

$$\big(\gamma^{(2)}, E\gamma^{(11)}\big) = \tfrac{1}{4}\lambda\rho\operatorname{tr}(\Pi^2\Gamma). \tag{44}$$

Bias terms. Using the fact that $E\Delta = 0$, we have from (35) that $E\gamma^{(1)} = 0$. Therefore the leading terms in the bias of $\tilde{\gamma}$ will be $\alpha\gamma^{(2)} + \varepsilon^2 E\gamma^{(11)}$, and so we will have

$$\begin{aligned}\|E\tilde{\gamma} - \gamma\|^2 &\approx \alpha^2\|\gamma^{(2)}\|^2 + 2\alpha\varepsilon^2\big(\gamma^{(2)}, E\gamma^{(11)}\big) + O(\varepsilon^4, \alpha\varepsilon^3, \alpha^2\varepsilon^2, \alpha^3)\\ &\approx \alpha^2\big(\tfrac{1}{4}\rho^2 + \lambda^2\|\Pi Q\gamma\|^2\big) + \tfrac{1}{2}\alpha\varepsilon^2\lambda\rho\operatorname{tr}(\Pi^2\Gamma),\end{aligned} \tag{45}$$

substituting (39) and (44).

The corresponding approximation for the norm of the leading bias term in the Rice–Silverman procedure with smoothing parameter $\tilde{\alpha}$ is [from Pezzulli and Silverman (1993)] $\tilde{\alpha}^2\lambda^2\|\Pi Q\gamma\|^2$; this is not on its own directly comparable with (45) because of the different roles that the smoothing parameters play in the two procedures.

Variance terms. The variance part of the mean integrated square error is $E\|\tilde{\gamma} - E\tilde{\gamma}\|^2$, which can be expanded as a power series in α and ε by substituting the expansion of $\tilde{\gamma}$. Since $\gamma^{(2)}$ is purely deterministic, the leading variance term will be $n^{-1}E(\gamma^{(1)}, \gamma^{(1)}) = n^{-1}\lambda\operatorname{tr}(\Pi^2\Gamma)$. This does not depend on the amount of smoothing applied, and the effect of the smoothing is on the next term in the variance. Define $V_n(\alpha)$ to be the variance term with smoothing parameter α, and define $V_n(0)$ to be the variance term when the principal components analysis is carried out without any smoothing. Then we have

$$V_n(\alpha) - V_n(0) \approx 2\alpha n^{-1}E\big(\gamma^{(1)}, \gamma^{(12)}\big) + O(\alpha\varepsilon^3, \alpha^2\varepsilon^2). \tag{46}$$

There are six terms in the expression (41) for $(I - P)\gamma^{(12)}$. Take the inner product of each of them with $\gamma^{(1)} = \Pi\,\Delta\gamma$ and then apply (43); this gives 12 terms in all, 9 of which are zero. For example,

$$E(\Pi\,\Delta\gamma, \Pi\,\Delta\Pi Q\gamma) = E(\gamma, \Delta\Pi^2\,\Delta\Pi Q\gamma)$$
$$= (\gamma, \Gamma\Pi^2\Gamma\Pi Q\gamma) + \operatorname{tr}(\Gamma\Pi^2)(\gamma, \Gamma\Pi Q\gamma),$$

which is equal to zero, because, for any x, $(\gamma, \Gamma\Pi x) = (\Gamma\gamma, \Pi x) = \lambda(\gamma, \Pi x) = 0$.

Pursuing manipulations of this type, we obtain

$$E(\gamma^{(1)}, \gamma^{(12)}) = -\tfrac{1}{2}\lambda\rho\operatorname{tr}(\Pi^2\Gamma) - \lambda^2\operatorname{tr}(\Pi^2 Q\Pi\Gamma) + \lambda^2\rho\operatorname{tr}(\Pi^3\Gamma). \tag{47}$$

Pezzulli and Silverman (1993) show that for the Rice–Silverman procedure with smoothing parameter $\tilde{\alpha}$ the approximation corresponding to (46) is $V_n(\tilde{\alpha}) - V_n(0) \approx 2\tilde{\alpha}n^{-1}\{-\lambda\operatorname{tr}(\Pi^2 Q\Pi\Gamma) + \lambda\rho\operatorname{tr}(\Pi^3\Gamma)\}$. Just as in the case of the leading bias term, this cannot in isolation be compared directly with (47), but it is interesting to note that the expression depends linearly on the eigenvalue λ rather than quadratically on λ and the roughness ρ of the eigenfunction.

6.3. *Is smoothing advantageous*? In this section we put together the results obtained above to determine under what conditions the estimation of γ_k will be improved by smoothing to some degree. We shall investigate this question by considering whether the mean integrated square error of $\tilde{\gamma}_k$ increases or decreases as α moves away from zero. If the derivative of the mean square error, considered as a function of α, is negative at $\alpha = 0$, then we can conclude that some degree of smoothing will give better estimation of γ_k in the L^2 sense. We shall make the dependence on k explicit from here onward. We shall write ρ_j for the roughness of the jth eigenfunction,

$$\rho_j = [\gamma_j, \gamma_j] = (\gamma_j, Q\gamma_j).$$

Let $M_n(\alpha)$ be the mean integrated square error of $\tilde{\gamma}_k$ for sample size n and smoothing parameter α. We can then see, by combining (46) and (45), that

$$nM_n'(0) = 2(\gamma^{(2)}, E\gamma^{(11)}) + 2E(\gamma^{(1)}, \gamma^{(12)}) + o(1) \quad \text{as } n \to \infty. \tag{48}$$

Substituting for the terms in (48) then gives

$$\begin{aligned} nM_n'(0) &\approx -\frac{1}{2}\lambda_k\rho_k\operatorname{tr}(\Pi^2\Gamma) - 2\lambda_k^2\operatorname{tr}(\Pi^2 Q\Pi\Gamma) + 2\lambda_k^2\rho_k\operatorname{tr}(\Pi^3\Gamma) \\ &= -\frac{1}{2}\lambda_k\rho_k\sum_{j\neq k}\frac{\lambda_j}{(\lambda_k - \lambda_j)^2} - 2\lambda_k^2\sum_{j\neq k}\frac{\lambda_j\rho_j}{(\lambda_k - \lambda_j)^3} \\ &\quad + 2\lambda_k^2\rho_k\sum_{j\neq k}\frac{\lambda_j}{(\lambda_k - \lambda_j)^3} \\ &= -2\left(\lambda_k^2\beta_k^{(1)} + \frac{1}{4}\rho_k\lambda_k\beta_k^{(2)}\right) = -2\beta_k, \end{aligned} \tag{49}$$

say, where

$$\beta_k^{(1)} = \sum_{j\neq k}\lambda_j(\lambda_k - \lambda_j)^{-3}(\rho_j - \rho_k)$$

and

$$\beta_k^{(2)} = \sum_{j \neq k} \lambda_j \big(\lambda_k - \lambda_j \big)^{-2}.$$

In order for $M_n'(0)$ to be negative, we require $\beta_k > 0$. This is a mild condition, indeed (since $\beta_k^{(2)}$ is necessarily positive) milder than the corresponding criterion $\beta_k^{(1)} > 0$ obtained by Pezzulli and Silverman (1993) for the Rice–Silverman method. A sufficient, but by no means necessary, condition for both β_k and $\beta_k^{(1)}$ to be positive is for $\rho_j < \rho_k$ if and only if $j < k$, so that the eigenfunctions of lower index than k are smoother than γ_k and those of higher index are rougher than γ_k.

6.4. *The ideal amount of smoothing.* Define

$$\rho_{jk} = \big(\gamma_j, Q\gamma_k \big) = \big[\gamma_j, \gamma_k \big].$$

For small α, the results (45) and (49) give the approximation

$$M_n(\alpha) \approx M_n(0) + \eta_k \alpha^2 - 2\beta_k n^{-1} \alpha, \tag{50}$$

where β_k is as defined in (49) and

$$\begin{aligned} \eta_k &= \lambda_k^2 \| \Pi Q \gamma_k \|^2 + \tfrac{1}{4} \rho_k^2 \\ &= \tfrac{1}{4} \rho_k^2 + \lambda_k^2 \sum_l \sum_{j \neq l} \big(\lambda_j - \lambda_l \big)^{-2} \rho_{lk}^2. \end{aligned} \tag{51}$$

Up to the degree of approximation in (50), the optimal α for the estimation of γ_k will be

$$\alpha_k^* = n^{-1} \eta_k^{-1} \max(\beta_k, 0). \tag{52}$$

It can be seen from the definitions of η_k and β_k that these quantities are both weighted sums depending on the quantities ρ_{ij}; the weights depend on the ratios between the various λ_j's. Although the way in which α_k depends on the quantities λ_j and ρ_{ij} is not transparent, this provides theoretical support for the empirical observation that the appropriate amount of smoothing does not depend dramatically on the index k.

It may be interesting to compare the results of Pezzulli and Silverman (1993). From equation (25) of that paper, but using our notation, the asymptotically optimal value of the smoothing parameter for the estimation of γ_k by the Rice–Silverman method will satisfy

$$\begin{aligned} \tilde{\alpha}_k^* &= n^{-1} \| \Pi Q \gamma_k \|^{-2} \max\big(\lambda_k \beta_k^{(1)}, 0 \big) \\ &= \lambda_k \frac{n^{-1} \max(\text{first term of } (49), 0)}{\text{first term of } (51)}. \end{aligned} \tag{53}$$

Leaving aside any differences caused by the fact that it is only the first terms of (49) and (51) that are involved, it is striking that $\tilde{\alpha}_k^*$ is λ_k multiplied by an expression that depend on the λ_j's only through the ratios between them. In most practical cases, the λ_k decay rapidly, and this illustrates why in practice

it is found appropriate in the Rice–Silverman method to choose much smaller smoothing parameters for the estimation of higher-order eigenvalues.

7. Choosing the smoothing parameter. Rice and Silverman (1991) discussed a cross-validation method for choosing the smoothing parameter in their procedure. Of course, in many problems a subjective choice of smoothing parameter is satisfactory or even preferable; for general remarks on this matter, see, for example, Green and Silverman [(1994), Section 3.1]. Nevertheless there are many contexts where an automatic choice of smoothing parameter may be helpful, and cross-validation provides a natural approach.

Because the estimation procedure works essentially by estimating all the eigenfunctions simultaneously, it is most natural to construct a cross-validation score that takes all the eigenfunctions into account. In order to consider how such a score could be calculated, suppose that X is an observation from the population. Then for each m the principal components $\gamma_1, \ldots, \gamma_m$ have the property that they explain more of the variation in X than any other collection of m vectors. Suppose $\tilde{\gamma}_1, \ldots$ are estimates of the principal component functions. Let G_m be the $m \times m$ matrix whose (i, j) element is the inner product $(\tilde{\gamma}_i, \tilde{\gamma}_j)$. Then the component of X orthogonal to the subspace spanned by $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_m$ is of course

$$\xi_m = X - \sum_{i=1}^{m} \sum_{j=1}^{m} \left(G_m^{-1}\right)_{ij} (\tilde{\gamma}_i, X)\, \tilde{\gamma}_j.$$

If we wished to consider the efficacy of the first m components, then a natural measure to consider would be $E\|\xi_m\|^2$; in order not to be tied to a particular m, one could, for example, seek to minimize $\sum_m E\|\xi_m\|^2$. In both cases, of course, we do not have new observations X to work with, and the usual cross-validation paradigm has to be used, as follows:

1. Subtract off the overall mean from the observed data X_i.
2. For a given smoothing parameter α, let $\tilde{\gamma}_j^{[i]}(\alpha)$ be the estimate of γ_j obtained from all the data except X_i.
3. Define $\xi_m^{[i]}(\alpha)$ to be the component of X_i orthogonal to the subspace spanned by $\{\tilde{\gamma}_j^{[i]}(\alpha)\colon j = 1, \ldots, m\}$.
4. Combine the "deleted remainders" $\xi_m^{[i]}(\alpha)$ to obtain the cross-validation scores

$$\mathrm{CV}_m(\alpha) = \sum_{i=1}^{n} \|\xi_m^{[i]}(\alpha)\|^2 \tag{54}$$

and possibly

$$\mathrm{CV}(\alpha) = \sum_{m=1}^{\infty} \mathrm{CV}_m(\alpha). \tag{55}$$

In practice one would of course truncate the sum in (55) at some convenient point. Indeed, given n data curves one can estimate at most $n - 1$

principal components and so the sum must be truncated at $m = n - 1$ if not at a smaller value.

5. Minimize $\mathrm{CV}_m(\alpha)$ or $\mathrm{CV}(\alpha)$ to provide the choice of smoothing parameter.

Clearly there are other possible ways of combining the $\mathrm{CV}_m(\alpha)$ to produce a cross-validation score to take more than one value of m into account; this is a matter for future investigation.

There are various computational tricks that can be used to compute the cross-validation score efficiently. For example, the covariance matrix (in the Fourier transform domain) of the population with the ith observation deleted can be calculated economically from the full covariance operator $\hat{\Gamma}$, and furthermore its eigenvalues can be related to those of $\hat{\Gamma}$. We shall not pursue these in detail here.

8. An example. In this section, the methodology described above is applied to an example. The records consist of the force exerted by the thumb and forefinger during each of 20 brief squeezes. The task required of the subject during each squeeze was to maintain a background force on a force meter and then to give a force impulse aimed at peaking at a predetermined maximum value, returning to the baseline afterward. The interest in the experiment is to study the behaviour of the muscle group controlling the thumb–forefinger muscle group and the way in which the brain controls this system. The data were collected at the MRC Applied Psychology Unit, Cambridge, by R. Flanagan, and were kindly supplied to the author by J. O. Ramsay. For a detailed description of the data and of another approach to its analysis see Ramsay, Flanagan and Wang (1995) and Ramsay (1995).

In Figure 1, the raw data curves are presented. It can be seen that these exhibit considerable local variability, and we shall see this reflected in the principal components. Ramsay, Flanagan and Wang (1995) fitted a parametric curve to each of these curves and then smoothed the residual curves

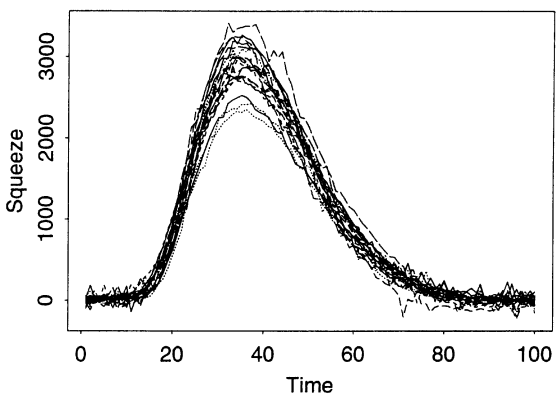

FIG. 1. *Grip force data.*

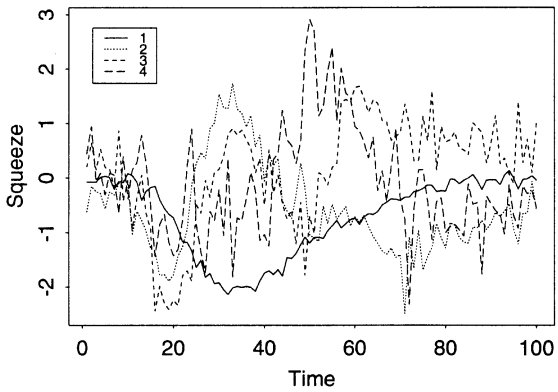

FIG. 2. *First four raw principal components for grip force data.*

before applying their own approach to functional PCA. We shall work with the raw data and investigate the consequences of our method of smoothed PCA.

In Figures 2 and 3 the effect of applying principal components analysis without smoothing and of using our method for smoothed PCA are compared. In both cases, the first four principal component curves are plotted. It can be seen that the raw principal component curves are very noisy. The smoothing parameter in Figure 3 is chosen by minimizing the score $\mathrm{CV}(\alpha)$ defined in Section 7. It was found satisfactory to calculate the cross-validation score at a grid (on a logarithmic scale) of values of the smoothing parameter α and pick out the minimum. The grid can be quite coarse since small changes in the numerical value of α do not make very much difference to the smoothed principal components. For this example the cross-validation scores were

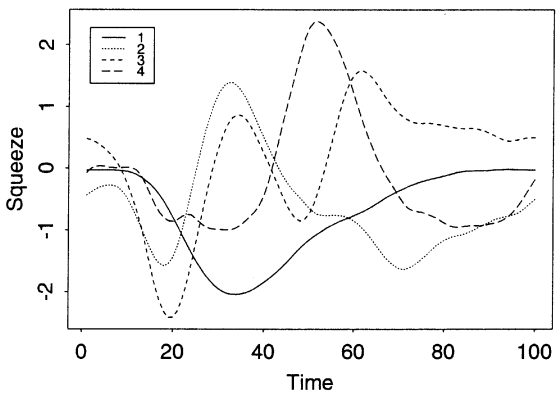

FIG. 3. *First four smoothed principal components for grip force data, smoothing parameter chosen by cross-validation.*

calculated for $\alpha = 0$ and $\alpha = 1.5^{i-1}$, for $i = 1, \ldots, 30$, and the minimum of $\mathrm{CV}(\alpha)$ was attained by setting $\alpha = 37$.

In Figure 4, the effect on the mean curve of adding and subtracting a multiple of each of the first four smoothed principal components is given. It can be seen that the first component corresponds to an effect whereby the shape of the impulse is not substantially changed, but its overall scale is increased. The second component (with appropriate sign) corresponds roughly to a compression in the overall time scale on which the "squeeze" takes place. Both of these effects were removed in the analysis of Ramsay, Flanagan and Wang (1995) before any detailed analysis was carried out. It is, however, interesting to note that they occur as separate components and therefore are essentially uncorrelated with one another, and with the effects found subsequently. The third component corresponds to an effect whereby the main part takes place more quickly but the tail after the main part is extended to the right. This corresponds to an effect detected by Ramsay, Flanagan and Wang (1995) expressed in rather different terms. The fourth component corresponds to a higher peak correlated with a tail-off that is faster initially, but subsequently slower than the mean. The first and second effects are transparent in their interest, and the third and fourth are of biomechanical interest in indicating ways in which the system compensates for departure from the (remarkably reproducible) overall mean. The smoothing we have described makes the effects very much clearer than they are in the raw principal component plot.

The estimated variances σ^2 indicate that the four components displayed respectively explain 86.2, 6.7, 3.5 and 1.7% of the variability in the original data, with 1.9% accounted for by the remaining components. Examination of the individual principal component scores indicates that there is one curve

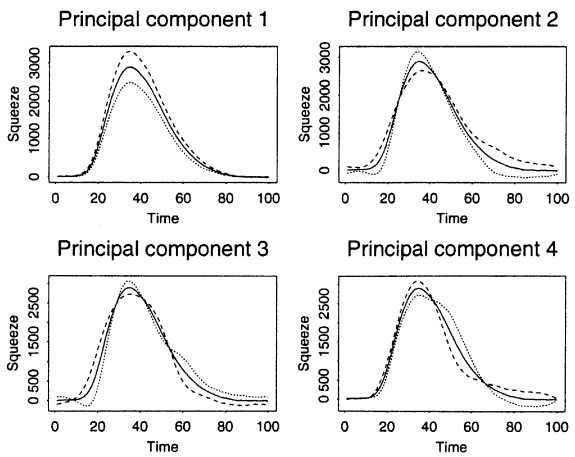

FIG. 4. *Effect on the overall mean curve of adding and subtracting a suitable multiple of each of the first four smoothed principal components.*

with a fairly extreme value of principal component 2 (corresponding to moving more quickly than average through the cycle), but this curve is not unusual in other respects.

Acknowledgments. I am extremely grateful to Trevor Hastie, Richard Olshen, Guy Nason, Jim Ramsay and John Rice for their help in various ways.

REFERENCES

ADAMS, R. A. (1975). *Sobolev Spaces*. Academic Press, New York.

DALZELL, C. J. and RAMSAY, J. O. (1993). Computing reproducing kernels with arbitrary boundary constraints. *SIAM J. Sci. Comput.* **14** 511–518.

DAUXOIS, J., POUSSE, A. and ROMAIN, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Multivariate Anal.* **12** 136–154.

GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, London.

LEURGANS, S. E., MOYEED, R. A. and SILVERMAN, B. W. (1993). Canonical correlation analysis when the data are curves. *J. Roy. Statist. Soc. Ser. B* **55** 725–740.

PEZZULLI, S. D. and SILVERMAN, B. W. (1993). Some properties of smoothed principal components analysis for functional data. *Comput. Statist. Data Anal.* **8** 1–16.

RAMSAY, J. O. (1995). Some tools for the multivariate analysis of functional data. In *Recent Advances in Descriptive Multivariate Analysis* (W. Krzanowski, ed.) 269–282. Clarendon, Oxford.

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

RAMSAY, J. O., FLANAGAN, R. and WANG, X. (1995). The functional data analysis of the pinch force of human fingers. *J. Roy. Statist. Soc. Ser. C* **44** 17–30.

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

SPIVAK, M. (1967). *Calculus*. Addison-Wesley, Reading, MA.

TAYLOR, A. E. and LAY, D. C. (1980). *Introduction to Functional Analysis*. Wiley, New York.

TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

SCHOOL OF MATHEMATICS
UNIVERSITY WALK
BRISTOL BS8 1TW
UNITED KINGDOM