Submitted to the Annals of Statistics

METHODOLOGY AND THEORY FOR PARTIAL LEAST SQUARES APPLIED TO FUNCTIONAL DATA

By Aurore Delaigle*, and Peter Hall*

University of Melbourne

The partial least squares procedure was originally developed to estimate the slope parameter in multivariate parametric models. More recently it has gained popularity in the functional data literature. There, the partial least squares estimator of slope is either used to construct linear predictive models, or as a tool to project the data onto a one dimensional quantity that is employed for further statistical analysis. Although the partial least squares approach is often viewed as an attractive alternative to projections onto the principal component basis, its properties are less well known than those of the latter, mainly because of its iterative nature. We develop an explicit formulation of partial least squares for functional data, which leads to insightful results and motivates new theory, demonstrating consistency and establishing convergence rates.

1. Introduction. Partial least squares (PLS) is an iterative procedure for estimating the slope of linear models. The technique was originally developed in high dimensional and collinear multivariate settings and is especially popular in chemometrics. See Wold (1975), Martens and Naes (1989), Helland (1990), Frank and Friedman (1993), Garthwaite (1994), Goutis and Fearn (1996), Durand and Sabatier (1997) and Nguyen and Rocke (2004).

The iterative nature of PLS can make it difficult to uncover properties in a clear and explicit way, and for a long time PLS was regarded as a technique that worked well, but whose properties were relatively obscure. Early theoretical developments of multivariate PLS can be found in Lorber, Wangen and Kowalski (1987) and Höskuldsson (1988), and further developments include those of Phatak, Rilley and Penlidis (2002), Phatak and de Hoog (2003), Bro and Eldén (2009) and Krämer and Sugiyama (2011).

More recently, the method has been applied in the functional data context by Preda and Saporta (2005a), who suggest using PLS for estimating slope in functional linear models; see also Reiss and Ogden (2007). Also in the

*Research supported by grants and fellowships from the Australian Research Council.

AMS 2000 subject classifications: Primary 62G08

Keywords and phrases: Central limit theorem, computational algorithm, consistency, convergence rates, functional linear models, generalised Fourier basis, principal components, projection, stochastic expansion

functional setting, the intrinsic iterative nature of PLS has made it difficult to develop intuition, and derive clear and explicit theoretical properties. In this paper we provide a transparent account of theoretical issues that underpin PLS methods in linear models for prediction from functional data, and show that they motivate an alternative formulation of PLS in that setting. This "alternative PLS," which we refer to here as APLS, has the advantage that it is expressed only in terms of functions that are explicitly computable. These attributes make APLS particularly attractive, relative to the conventional PLS formulation, and permit detailed theoretical development.

We give concise stochastic expansions for the difference between estimators derived using APLS, and the quantities to which these estimators converge in the large-sample limit. These expansions are valid uniformly in estimators based on the first $O(n^{1/2})$ APLS basis functions, where n denotes sample size. The expansions also lead easily and directly to a variety of results about our estimators, including convergence rates and central limit theorems.

Besides functional linear models, PLS is employed in a variety of other data functional problems. For example, Ferraty and Vieu (2006) use it to define a semi-metric for nonparametric functional predictors or classifiers; Escabias, Aguilera and Valderrama (2007) employ PLS with logit regression; Preda, Saporta and Lévéder (2007), and Delaigle and Hall (2012), use it for functional data classification. See also Preda and Saporta (2005b), Krämer et al. (2008) and Aguilear et al. (2010).

2. Functional linear models.

2.1. *General bases for inference in functional linear models.* Let $\mathcal{X} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ denote a sample of independent data pairs, all distributed as (X, Y), where X is a random function defined on the nondegenerate, compact interval $\mathcal{I}$ and satisfying $\int_{\mathcal{I}} E(X^2) < \infty$, and Y is a scalar random variable generated by the linear model

$$Y = a + \int_{\mathcal{I}} b\, X + \epsilon\,. \tag{2.1}$$

Here, a denotes a scalar parameter, ϵ is a scalar random variable with finite mean square and satisfying $E(\epsilon \,|\, X) = 0$, and b, a function-valued parameter, is a square-integrable function on $\mathcal{I}$.

Predicting the value of Y, given X, amounts to estimating the function

$$g(x) = E(Y \,|\, X = x) = a + \int_{\mathcal{I}} b\, x\,, \tag{2.2}$$

which itself requires to estimate the scalar a and the function b from the data. A standard approach is to express X and b in terms of an orthonormal basis $\psi_1, \psi_2, \ldots$ defined on $\mathcal{I}$. Expansions for X and b in this basis can be written as $X = \sum_j (\int_{\mathcal{I}} X\,\psi_j)\,\psi_j$ and $b = \sum_j v_j\,\psi_j$, where $v_j = \int_{\mathcal{I}} b\,\psi_j$. Since in practice we can calculate only a finite number of terms, the infinite-dimensional expansion for b is approximated by a sum of p terms, where $p \geq 1$ is an integer, and each term of this sum is then estimated from the data. Note that $\int_{\mathcal{I}} b\,X = \sum_j v_j \int_{\mathcal{I}} X\,\psi_j$, which motivates us to take $a = E(Y) - \int_{\mathcal{I}} b\,E(X)$ and define $\beta_1, \ldots, \beta_p$ to be the sequence $v_1, \ldots, v_p$ that minimises

$$s_p(v_1, \ldots, v_p) = E\Big\{ \int_{\mathcal{I}} b\,(X - EX) - \sum_{j=1}^{p} v_j \int_{\mathcal{I}} (X - EX)\,\psi_j \Big\}^2 . \tag{2.3}$$

The functions

$$b_p = \sum_{j=1}^{p} \beta_j \psi_j,\; g_p(x) = E(Y) + \int_{\mathcal{I}} b_p\,(x - EX) = E(Y) + \sum_{j=1}^{p} \beta_j \int_{\mathcal{I}} (x - EX)\psi_j \tag{2.4}$$

are approximations to b and to $g(x)$, respectively. Their accuracy, as p increases, depends on the choice of the sequence $\psi_1, \psi_2, \ldots$.

Sometimes the basis is chosen independently of the data (e.g. sine-cosine basis, spline basis, etc). Then the functions ψ_j are known, and an empirical version of (2.4) is obtained by replacing the scalars $\beta_1, \ldots, \beta_p$ by the sequence $v_1, \ldots, v_p$ that minimises

$$n^{-1} \sum_{i=1}^{n} \Big\{ Y_i - \bar{Y} - \sum_{j=1}^{p} v_j \int_{\mathcal{I}} (X_i - \bar{X})\,\psi_j \Big\}^2 . \tag{2.5}$$

A drawback of such bases is that there is no reason why their first p elements should capture the most important information about the regression function g, available from the data. It seems more attractive to use bases that adapt to the properties of the population represented by the data. We discuss two such adaptive bases in sections 2.2 and 2.3, respectively.

2.2. *Principal component basis.* One of the most popular adaptive bases is the so-called principal component basis, constructed from the covariance function $K(s,t) = \mathrm{cov}\{X(s), X(t)\}$ of the random process X. As is common in mathematical analysis, we shall use the notation K also for the linear transformation (a functional) that takes a square-integrable function ψ to $K(\psi)$ given by $K(\psi)(t) = \int_{\mathcal{I}} \psi(s)\,K(s,t)\,ds$.

Since $\int_{\mathcal{I}} E(X^2) < \infty$ then $\int_{\mathcal{I}} K(t,t)\,dt < \infty$, and we can write the spectral decomposition of K as

$$K(s,t) = \sum_{k=1}^{\infty} \theta_k\, \phi_k(s)\, \phi_k(t)\,, \tag{2.6}$$

where the principal component basis $\phi_1, \phi_2, \ldots$ is a complete orthonormal sequence of eigenvectors (i.e. eigenfunctions) of the transformation K, with respective nonnegative eigenvalues $\theta_1, \theta_2, \ldots$. That is, $K(\phi_k) = \theta_k\, \phi_k$ for $k \geq 1$. Positive definiteness of K implies that the eigenvalues are nonnegative, and the condition $\int_{\mathcal{I}} E(X^2) < \infty$ entails $\sum_k \theta_k < \infty$. Therefore we can, and do, order the terms in the series in (2.6) so that

$$\theta_1 \geq \theta_2 \geq \ldots \geq 0\,. \tag{2.7}$$

In practice the scalars θ_j and the functions ϕ_j are unknown and are estimated from the data, as follows. First, the covariance function is estimated by

$$\widehat{K}(s,t) = \frac{1}{n} \sum_{i=1}^{n} \{X_i(s) - \bar{X}(s)\}\, \{X_i(t) - \bar{X}(t)\}\,, \tag{2.8}$$

where $\bar{X}(t) = n^{-1} \sum_{i=1}^{n} X_i(t)$. Then, $\theta_1, \ldots, \theta_n$ and $\phi_1, \ldots, \phi_n$ are estimated by the eigenvalues $\widehat{\theta}_1 \geq \widehat{\theta}_2 \geq \ldots \widehat{\theta}_n \geq 0$ and the eigenfunctions $\widehat{\phi}_1, \ldots, \widehat{\phi}_n$ of the transformation represented by $\widehat{K}$, which can have at most n nonzero eigenvalues. Finally, an empirical version of $\beta_1, \ldots, \beta_p$ is defined to be the sequence $v_1, \ldots, v_p$ that minimises (2.5), where each ψ_j there is replaced by $\widehat{\phi}_j$. Then, g_p at (2.4) is replaced by its corresponding empirical version. In the rest of this paper, to avoid confusion with projections of b onto other bases, we shall add a superscript $^{\rm PC}$ to coefficients obtained from projection of b onto one of the functions ϕ_j; that is, we shall use the notation $\beta_j^{\rm PC} = \int_{\mathcal{I}} b\phi_j$.

The literature on functional linear models based on principal component analysis (PCA) is large. It includes, for example, work by Cai and Hall (2006), Reiss and Ogden (2007), Apanasovich and Goldstein (2008), Cardot and Sarda (2008), Baillo (2009), Müller and Yao (2010), Wu, Fan and Müller (2010) and Yao and Müller (2010).

2.3. *The orthonormal PLS basis.* The principal component basis introduced in section 2.2 is defined in terms of the population, but only through X. In particular, while its first p elements $\phi_1, \ldots, \phi_p$ usually contain most of the information related to the covariance of X, these are not necessarily important for representing b, and all or some of the most important terms

accounting for the interaction between b and X might come from later principal components. In prediction, to capture the main effects of interaction using only a few terms, one could construct the basis in a way that takes this interaction into account.

Motivated by such considerations, the standard PLS basis, adapted to the functional context, is defined iteratively by choosing ψ_p in a sequential manner, to maximise the covariance functional

$$f_p(\psi_p) = \mathrm{cov}\left\{Y - g_{p-1}(X), \int_{\mathcal{I}} X\,\psi_p\right\}, \tag{2.9}$$

subject to

$$\int_{\mathcal{I}}\int_{\mathcal{I}} \psi_j(s)K(s,t)\,\psi_p(t)\,ds\,dt = 0 \text{ for } 1 \le j \le p-1\,, \text{ and } \|\psi_p\| = 1\,, \tag{2.10}$$

where $\|\cdot\|$ is a norm (see section 3.1), and given that $\psi_1, \ldots, \psi_{p-1}$ have already been chosen. (Recall that g_p was defined at (2.4).) In practice, the covariances in (2.9) are replaced by estimates, and empirical versions of the ψ_js are constructed by an iterative algorithm described in appendix A.2.

Partial least squares can also be used for prediction in nonlinear models, where the basis that it produces is sometimes, but not always, effective for prediction. Specifically, although the PLS basis enables a consistent approximation to g in such cases, a large number of terms may be required to get a good approximation.

3. Properties of theoretical functional partial least squares. For prediction and estimation of b, the PLS basis is sometimes preferred to the PCA basis, partly because it can often capture the relevant information with fewer terms; see our data illustrations in section 6. Detailed theoretical properties for inference in functional linear models based on the PCA basis have been studied in a number of papers, but few results exist about their functional PLS counterpart. In this section we provide new insight into the theoretical PLS basis defined at (2.9)–(2.10), and give an explicit description of the space generated by the first p PLS basis functions $\psi_1, \ldots, \psi_p$. These properties motivate an alternative formulation of functional PLS, which we call APLS. It permits us to define the functional PLS basis very simply, and to construct an explicitly defined algorithm to implement empirical PLS; see section 4. The explicit nature of the algorithm will allow us to derive detailed theoretical properties of empirical functional PLS, including convergence rates; see section 5.

3.1. *Explicit form of the orthonormal PLS basis.* Our first result, Theorem 3.1, below, gives an explicit account of the constrained optimisation described in section 2.3. We use the following notation. Given α_1 and α_2 in the class $\mathcal{C}(\mathcal{I})$ of all square-integrable functions on $\mathcal{I}$, write $\int_{\mathcal{I}}\int_{\mathcal{I}} \alpha_1\, \alpha_2\, K$ to denote $\int_{\mathcal{I}}\int_{\mathcal{I}} \alpha_1(s)\, \alpha_2(t)\, K(s,t)\, ds\, dt$. For any $x \in \mathcal{C}(\mathcal{I})$, define $\|x\|^2 = \int_{\mathcal{I}}\int_{\mathcal{I}} x\, x\, K$. (Some implementations of PLS, for example the one in appendix A.2, take $\|x\|^2 = \int_{\mathcal{I}} x^2$, but this affects only the scale, not the main properties of the functions ψ_j).

THEOREM 3.1. *If $\int_{\mathcal{I}} E(X^2) < \infty$ then the function ψ_p that maximises f at (2.9), given $\psi_1, \ldots, \psi_{p-1}$ and subject to (2.10), is determined by*

$$\psi_p = c_0 \left[K\left\{ b - \sum_{j=1}^{p-1} \left(\int_{\mathcal{I}} b\, \psi_j \right) \psi_j \right\} + \sum_{k=1}^{p-1} c_k\, \psi_k \right], \tag{3.1}$$

where, for $1 \le k \le p-1$, the constants c_k are obtained by solving the linear system of $p-1$ equations

$$\int_{\mathcal{I}}\int_{\mathcal{I}} \psi_j\, \psi_p\, K = 0, \quad j = 1, \ldots, p-1, \tag{3.2}$$

and where c_0 is defined uniquely, up to a sign change, by the property

$$\|\psi_p\| = 1\,. \tag{3.3}$$

One of the interesting implications of the theorem is that for each j, the jth basis function determined by PLS can be expressed as a linear combination of j explicitly defined functions. More precisely, the theorem implies that $\psi_1 = d_1\, K(b)$, where, by (3.3) with $p = 1$, $d_1 = \|K(b)\|^{-1}$, and more generally, the following properties follow from the representation (3.1); the first property implies the second:

(a) For each $p \ge 1$, and given $\psi_1, \ldots, \psi_{p-1}$, the function ψ_p is the linear combination of $K(b), \ldots, K^p(b)$ for which (2.10) holds, and is unique up to a sign change. (b) For each $p \ge 1$, representing a function as a linear form in $\psi_1, \ldots, \psi_p$ is equivalent to representing it as a linear combination of $K(b), \ldots, K^p(b)$. (3.4)

These properties motivate the APLS formulation and underpin the rest of the paper. Interestingly, (3.4) continues to hold if equations (3.2) are replaced by $\int_{\mathcal{I}} \psi_j \psi_p = 0$ for $j = 1, \ldots, p-1$. In particular, although the functions $\psi_2, \ldots, \psi_p$ will change in this case, the spaced spanned by $\psi_1, \ldots, \psi_p$ will not alter.

Result (3.4) is a deterministic functional version of a known result for empirical PLS in the multivariate context. More specifically, suppose we have n observations of a q-variate (non functional) predictor of a variable Y, let $\boldsymbol{X}$ be the $n \times q$ matrix of observations on the predictor, and let $\boldsymbol{y}$ be the $n \times 1$ vector containing the observations on Y. Then it has been established that the space spanned by the first p empirical PLS components is equal to the space generated by $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, \boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, \ldots, \boldsymbol{D}^{p-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, where $\boldsymbol{D} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$. See for example Bro and Eldén (2009), and compare the empirical algorithm in section 4.1. This is itself a particular case of results that are available more generally in Krylov spaces, although again in the multivariate rather than functional setting that is the subject of this paper.

3.2. *Expansions in a non-orthogonal PLS basis.* The properties at (3.4) give a clear and explicit account of the form taken by the PLS basis functions. For example, they show that for each p, the space generated by $\psi_1, \ldots, \psi_p$ is the same as the space generated (i.e. spanned) by $K(b), \ldots, K^p(b)$. Note that the functions $K^j(b)$ are explicitly defined, since we have $K^j(b) = \sum_k \theta_k^j \beta_k^{\mathrm{PC}} \phi_k$, where ϕ_k is the kth PCA basis function.

Next, if we note that $a = E(Y) - \int_{\mathcal{I}} b\, E(X)$ and define $\gamma_1, \ldots, \gamma_p$ to be the sequence $w_1, \ldots, w_p$ that minimises

$$t_p(w_1, \ldots, w_p) = E\Big\{ \int_{\mathcal{I}} (X - EX)\, b - \sum_{j=1}^{p} w_j \int_{\mathcal{I}} (X - EX)\, K^j(b) \Big\}^2 \quad (3.5)$$

(compare (2.3)), then the slope function approximation b_p at (2.4) has two equivalent expressions:

$$b_p = \sum_{j=1}^{p} \gamma_j\, K^j(b) = \sum_{j=1}^{p} \beta_j\, \psi_j\,, \quad (3.6)$$

where $\beta_1, \ldots, \beta_p$ are as defined in section 2.1 if we take the general $\psi_1, \ldots, \psi_p$ introduced there to be the specific functions given by Theorem 3.1.

In matrix notation,

$$\gamma \equiv (\gamma_1, \ldots, \gamma_p)^{\mathrm{T}} = H^{-1}(\alpha_1, \ldots, \alpha_p)^{\mathrm{T}}\,, \quad (3.7)$$

where $H = (h_{jk})_{1 \le j,k \le p}$ denotes a $p \times p$ matrix,

$$h_{jk} = \int_{\mathcal{I}} K^{j+1}(b)\, K^k(b) = \sum_{r=1}^{\infty} (\beta_r^{\mathrm{PC}})^2\, \theta_r^{j+k+1}\,, \quad (3.8)$$

$$\alpha_j = \int_{\mathcal{I}} K(b)\, K^j(b) = \sum_{r=1}^{\infty} (\beta_r^{\rm PC})^2\, \theta_r^{j+1} = h_{0j}\,. \tag{3.9}$$

Here we have used the fact that, for p fixed, the matrix H is nonsingular because, for finite p, the equivalence of the expansion in the orthogonal basis $\psi_1, \ldots, \psi_p$ and in the basis $K(b), \ldots, K^p(b)$ implies that the sequence $\gamma_1, \ldots, \gamma_p$ that minimises (3.5) is unique. See also our discussion on Hankel matrices in section 5.3.

The pth order approximation $g_p(x)$ to $g(x) = E(Y \,|\, X = x)$, resulting from the pth order approximation of b by either of the identities at (3.6), is given equivalently by the second formula at (2.4) or by the expression

$$g_p(x) = a + \int_{\mathcal{I}} b_p\, x = E(Y) + \sum_{j=1}^{p} \gamma_j \int_{\mathcal{I}} (x - EX)\, K^j(b)\,. \tag{3.10}$$

We denote by APLS the formulation of PLS based on the sequence $K(b), \ldots, K^p(b)$.

For the approximation at (3.6) to converge to b, that function should be expressible as a linear form in $K(b), K^2(b), \ldots$:

$$b = \sum_{j=1}^{\infty} w_j\, K^j(b)\,, \tag{3.11}$$

where the w_js are constants and the series converges in L^2. The next theorem gives conditions under which, for a general b in $\mathcal{C}(\mathcal{I})$, there exist $w_1, w_2, \ldots$ such that (3.11) holds.

THEOREM 3.2. *If $\int_{\mathcal{I}} E(X^2) < \infty$, and the eigenvalues of K are all nonzero, then each $b \in \mathcal{C}(\mathcal{I})$ can be written as at* (3.11)*, where the series converges in L^2.*

Under the side condition $\int_{\mathcal{I}} E(X^2) < \infty$ the assumption in Theorem 3.2 that all eigenvalues of K be nonzero is both necessary and sufficient for (3.11) to hold for all $b \in \mathcal{C}(\mathcal{I})$. However, if some eigenvalues θ_j, corresponding to respective eigenvectors ϕ_j, vanish, then the respective values of $\int_{\mathcal{I}} (X - EX)\, \phi_j$ vanish with probability 1, and so those indices make zero contribution to $\int_{\mathcal{I}} (X - EX)\, b = \sum_j \int_{\mathcal{I}} (X - EX)\, \phi_j \cdot \int_{\mathcal{I}} b\, \phi_j$. Therefore we can delete the components of $b = \sum_j \phi_j \int_{\mathcal{I}} b\, \phi_j$ that correspond to indices j for which $\theta_j = 0$, without affecting the value of $\int_{\mathcal{I}} b\, X$; and it is only through the latter integral that b influences prediction. Therefore the theorem can be stated in a form which asserts that even if some of the eigenvalues of K

vanish, the representation at (3.11) is sufficiently accurate for all purposes of prediction based on (2.1). The only reason we have not taken this course is to make our arguments relatively simple and transparent.

Note that the w_js in (3.11) are not determined uniquely. In particular, (3.11) implies that $K(b) = \sum_j w_j K^{j+1}(b)$, and so the following expansion, among many others, is alternative to (3.11): $b = \sum_{j=1}^{\infty} (w_j + w_{j+1}) K^{j+1}(b)$. This lack of uniqueness does not violate the equivalence noted in (3.4)(b), since that property is asserted only for a finite sequence $\psi_1, \ldots, \psi_p$. However, it makes it impossible to treat usefully the relationship between the infinite expansion of a function b in terms of the sequence $K(b), K^2(b), \ldots$, and its infinite expansion in terms of the PCA basis, $\phi_1, \phi_2, \ldots$, introduced in section 2.2. Nevertheless we can discuss the pth order PLS projection $b_p = \sum_{j=1}^{p} \gamma_j K^j(b)$ of b onto the finite dimensional space spanned by $K(b), \ldots, K^p(b)$, for an arbitrary but fixed $p \geq 1$.

To this end, recall that $\beta_1^{\rm PC}, \beta_2^{\rm PC}, \ldots$ denote the Fourier coefficients of b with respect to the PCA basis $\phi_1, \phi_2, \ldots$. Then,

$$b_p = \sum_{j=1}^{p} \gamma_j K^j(b) = \sum_{j=1}^{p} \gamma_j \sum_{k=1}^{\infty} \beta_k^{\rm PC} \theta_k^j \phi_k = \sum_{k=1}^{\infty} \beta_k^{\rm PC} \left(\sum_{j=1}^{p} \gamma_j \theta_k^j \right) \phi_k \,, \tag{3.12}$$

and the last series expresses b_p in terms of the components of the PCA basis.

4. Empirical implementation of APLS.

4.1. *Algorithm for empirical APLS.* A standard algorithm for empirical implementation of PLS based on the sequence $\psi_1, \ldots, \psi_p$ is given in appendix A.2. In this section we describe a simple empirical algorithm for implementing APLS based on the non-orthogonal sequence $K(b), \ldots, K^p(b)$. As we shall see, this algorithm will permit simple derivation of theoretical properties of PLS. In section 4.2 we shall deduce two algorithms that are numerically more stable.

To estimate $K(b), \ldots, K^p(b)$, first note that we can estimate $K(b)$ by

$$\widehat{K(b)} = \frac{1}{n} \sum_{i=1}^{n} X_i^{\rm cent} Y_i^{\rm cent} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) (Y_i - \bar{Y}) \,,$$

where $X_i^{\rm cent} = X_i - \bar{X}$ and $Y_i^{\rm cent} = Y_i - \bar{Y}$. Then, given an estimator $\widehat{K^j(b)}$ of $K^j(b)$, we can estimate $K^{j+1}(b)(t)$ by $\widehat{K^{j+1}}(b)(t) = \int_{\mathcal{I}} \widehat{K^j(b)}(s) \, \widehat{K}(s,t) \, ds$, where $\widehat{K}$ is the conventional estimator of the covariance function,

$\widehat{K}(s,t) = n^{-1} \sum_{i=1}^n \{X_i(s) - \bar{X}(s)\} \{X_i(t) - \bar{X}(t)\}$. Having calculated $\widehat{K^j(b)}$ for $1 \le j \le p$ we take $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_p$ to minimise

$$U_p(w_1, \ldots, w_p) = \frac{1}{n} \sum_{i=1}^n \left\{ Y_i^{\rm cent} - \sum_{j=1}^p w_j \int_{\mathcal{I}} X_i^{\rm cent} \, \widehat{K^j(b)} \right\}^2 \tag{4.1}$$

with respect to $w_1, \ldots, w_p$ (compare (3.5)). In matrix notation,

$$\widehat{\gamma} \equiv (\widehat{\gamma}_1, \ldots, \widehat{\gamma}_p)^{\rm T} = \widehat{H}^{-1} (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_p)^{\rm T} , \tag{4.2}$$

where $\widehat{H} = (\widehat{h}_{jk})_{1 \le j,k \le p}$ denotes a $p \times p$ matrix,

$$\widehat{h}_{jk} = \int_{\mathcal{I}} \int_{\mathcal{I}} \widehat{K}(s,t) \, \widehat{K^j(b)}(s) \, \widehat{K^k(b)}(t) \, ds \, dt = \int_{\mathcal{I}} \widehat{K^{j+1}}(b) \, \widehat{K^k(b)} , \tag{4.3}$$

$$\widehat{\alpha}_j = \int_{\mathcal{I}} \widehat{K(b)} \, \widehat{K^j(b)} . \tag{4.4}$$

Finally we construct an estimator of g based on (3.10):

$$\widehat{g}_p(x) = \bar{Y} + \sum_{j=1}^p \widehat{\gamma}_j \int_{\mathcal{I}} (x - \bar{X}) \, \widehat{K^j(b)} . \tag{4.5}$$

REMARK 1. Formula (3.8) demonstrates that the theoretical version H of $\widehat{H}$ is a symmetric matrix. Our estimator $\widehat{H}$ does not necessarily enjoy that property, but an alternative estimator of h_{jk} can be defined to satisfy it. More precisely we can take $\tilde{h}_{jk} = \int_{\mathcal{I}} \widehat{K^{j+k}}(b) \, \widehat{K(b)}$, which produces a symmetric estimator $\widetilde{H} = (\tilde{h}_{jk})$ of H. We could use $\widetilde{H}$ in place of $\widehat{H}$, but computing $\tilde{h}_{jk}$ requires $\widehat{K}$ to be iterated $j + k$ times, whereas $\widehat{h}_{jk}$ needs iteration at most $\max(j + 1, k)$ times. Therefore we prefer the version $\widehat{H}$.

4.2. *Stabilised algorithm for empirical APLS.* The algorithm described in section 4.1 would provide a good solution if we were able to work in exact arithmetic, but it can be unstable in finite precision arithmetic. This is because, due to the non-unicity of the expression for b in terms of the infinite series $K(b), K^2(b), \ldots$, as p increases the linear system of equations given by the empirical version of (3.5) (see (4.1)) becomes closer to singular. Therefore, in finite precision arithmetic, as p increases it becomes more difficult to numerically identify one or more of the valid expressions arising from a large number of terms in the sequence $\widehat{K}(b), \widehat{K}^2(b), \ldots$.

There exist a number of numerical methods for overcoming this numerical difficulty. A simple approach is to transform the linear system of equations

by Gram-Schmidt orthogonalisation; see section 7.7 of Lange (1999). There, the columns of the $n \times p$ matrix with (i,j)th element equal to $\int_{\mathcal{I}} X_i^{\text{cent}} \widehat{K^j(b)}$ are transformed into p orthonormal vectors $u_1, \ldots, u_p$ by the modified Gram-Schimdt algorithm (a numerically stabilised version of Gram-Schmidt algorithm, see appendix A.3). Instead of using $\widehat{\gamma}$ in (4.2), the sequence that minimises (4.1) can then be computed by solving, with respect to $w_1, \ldots, w_p$, the equivalent equation $\boldsymbol{R}(w_1, \ldots, w_p)^{\mathrm{T}} = \boldsymbol{U}^{\mathrm{T}}(Y_1^{\text{cent}}, \ldots, Y_n^{\text{cent}})^{\mathrm{T}}$, where $\boldsymbol{U}$ is a matrix with columns $u_1, \ldots, u_p$ and $\boldsymbol{R}$ is an upper $p \times p$ triangular matrix. Let $\widehat{\gamma}^* = (\widehat{\gamma}_1^*, \ldots, \widehat{\gamma}_p^*)^{\mathrm{T}}$ be the solution of this equation. We can estimate g by $\widehat{g}_p^*(x) = \bar{Y} + \sum_{j=1}^p \widehat{\gamma}_j^* \int_{\mathcal{I}} (x - \bar{X}) \widehat{K^j(b)}$.

Alternatively, having constructed $\widehat{K^j(b)}$ for $1 \leq j \leq p$ as in section 4.1, we can also transform them into an orthonormal sequence $\widehat{\psi}_1, \ldots, \widehat{\psi}_p$ satisfying the standard PLS constraints, $\int_{\mathcal{I}}\int_{\mathcal{I}} \widehat{\psi}_j \widehat{\psi}_k \widehat{K} = 0$ for $j \neq k$ (compare (2.10)), using for example the modified Gram-Schmidt algorithm. Then we can calculate an empirical version $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ of $\beta_1, \ldots, \beta_p$, the latter defined in section 2.1 (taking there the ψ_js to be the empirical PLS basis functions), by finding the sequence $v_1, \ldots, v_p$ that minimises (2.5). Finally, we can estimate g by

$$\tilde{g}_p(x) = \bar{Y} + \sum_{j=1}^{p} \widehat{\beta}_j \int_{\mathcal{I}} (x - \bar{X}) \widehat{\psi}_j \,. \tag{4.6}$$

In exact arithmetic, $\widehat{g}_p^*$ and $\widehat{\gamma}^*$ would be equal to, respectively, $\widehat{g}_p$ and $\widehat{\gamma}$ defined in (4.5) and (4.2). Likewise, $\tilde{g}_p$, would be equal to $\widehat{g}_p$. In practice, these approximations differ because we can only work in finite precision arithmetic, and the algorithms leading to $\widehat{g}_p^*$ and $\tilde{g}_p$ are much more numerically stable than the one leading to $\widehat{g}_p$. In general, for prediction we found the algorithm leading to $\tilde{g}_p$ to be preferable. However, the algorithm of section 4.1 is important for developing intuition and assembling theoretical arguments. On the theoretical side, the simple, explicit formulae in section 4.1 permit us to establish consistency and derive rates of convergence. Of course, the equivalence between $\tilde{g}_p$, $\widehat{g}_p$ and $\widehat{g}_p^*$ implies that, in order to derive the theoretical properties of $\tilde{g}_p$ and $\widehat{g}_p^*$, it suffices to derive them for $\widehat{g}_p$ (all three have the same theoretical properties). On the intuitive side we note that the explicit formulation of the quantities involved in our empirical algorithms for APLS gives a much clearer account of what partial least-squares does, than the standard empirical iterative PLS algorithm in appendix A.2.

5. Asymptotic properties of empirical APLS.

5.1. *Introduction.* To our knowledge, the only existing theoretical results for functional PLS are those of Preda and Saporta (2005a), who state generalisations to the functional data context of some results of Höskuldsson (1988). Although they are of interest, the theoretical arguments there are iterative and not explicit, and consistency of the PLS approximation is mentioned without a proof and without regularity conditions or convergence rates. This is because those results are based on the iterative empirical approximation of PLS, and the inexplicit form of the algorithm (see appendix A.2) apparently makes it very difficult to derive explicit theoretical results.

Our alternative formulation, APLS, of the functional partial least-squares problem permits us to derive many properties. As already explained in section 4.2, the theoretical properties of the empirical approximations $\widehat{g}_p^*$ and $\tilde{g}_p$ in section 4.2 are identical to those of $\widehat{g}_p$ in section 4.1.

5.2. *Main results.* Define $\mu = E(X)$, a function, and observe that we can write:

$$\widehat{K} = K + n^{-1/2}\,\xi + n^{-1}\,\eta\,, \quad \widehat{K(b)} = K(b) + n^{-1/2}\,\xi_0 + n^{-1}\,\eta_0\,, \tag{5.1}$$

where ξ and η are functions of two variables, ξ_0 and η_0 are functions of a single variable, each equals $O_P(1)$. More specifically,

$$\begin{aligned}
\xi(s,t) &= \frac{1}{n^{1/2}} \sum_{i=1}^{n} (1-E)\,\{X_i(s)-\mu(s)\}\,\{X_i(t)-\mu(t)\}\,,\\
\xi_0(t) &= \frac{1}{n^{1/2}} \sum_{i=1}^{n} (1-E)\,\{X_i(t)-\mu(t)\}\,\{Y_i-E(Y_i)\}\,,\\
\eta(s,t) &= -n\,\{\bar{X}(s)-\mu(s)\}\,\{\bar{X}(t)-\mu(t)\}\,,\\
\eta_0(t) &= -n\,\{\bar{X}(t)-\mu(t)\}\,(\bar{Y}-E\bar{Y})\,.
\end{aligned}$$

For any square-integrable function L of two variables, define $|||L|||^2 = \int_{\mathcal{I}}\int_{\mathcal{I}} L^2$ and put $R_1 = |||K||| + n^{-1/2}\,|||\xi||| + n^{-1}\,|||\eta|||$, $R_2 = |||\xi||| + |||\eta|||$. Define too

$$\zeta_j(t) = \int_{\mathcal{I}} K^j(b)(s)\,\xi(s,t)\,ds\,, \tag{5.2}$$

and

$$\xi_j = K^{j-1}(\xi_0) + \sum_{k=0}^{j-2} K^k(\zeta_{j-k-1})\,, \tag{5.3}$$

$$\|\eta_j\| \le R_1^{j-1}\,\|\eta_0\| + R_2 \sum_{k=1}^{j-1} R_1^{j-k-1}\,\left(\left\|K^k(b)\right\| + \left|||K^{k-1}|||\,\|\xi_0\|\right)$$

$$+ R_2 \, |||\xi||| \sum_{k=1}^{j-1} R_1^{j-k-1} \sum_{\ell=0}^{k-2} |||K^\ell||| \, \big\|K^{k-\ell-1}(b)\big\| \, . \tag{5.4}$$

Theorem 5.1 below requires no assumptions beyond the model at (2.1), and the condition that

$$\int_{\mathcal{I}} b^2 < \infty \, , \quad E\|X\|^4 < \infty \, , \quad E\big(\epsilon^2\big) < \infty \, . \tag{5.5}$$

(Recall that ϵ, satisfying $E(\epsilon \,|\, X) = 0$, is the error in the model at (2.1).) Note that, under (5.5), it follows from (5.3) and (5.4) that $\|\xi_j\| + \|\eta_j\| = O_P(n^{-1/2})$. Theorem 5.1 shows that the empirical approximations $\widehat{K^j(b)}$ to the basis functions used by APLS, converge in probability to their theoretical values $K^j(b)$ at a rate $n^{-1/2}$.

THEOREM 5.1. *If* (5.5) *holds then, for each* $j \geq 1$,

$$\widehat{K^j(b)} = K^j(b) + n^{-1/2} \, \xi_j + n^{-1} \, \eta_j \, , \tag{5.6}$$

where ξ_j *is defined at* (5.3) *and* η_j *satisfies* (5.4).

The next theorem shows that the matrix entries $\widehat{h}_{jk}$, defined at (4.3), converge in probability to their theoretical counterparts h_{jk}, at (3.8), at a rate $n^{-1/2}$. This theorem will be used to establish consistency of the empirical coefficients $\widehat{\gamma}_j$ used in the empirical APLS expansion at (4.5). Note that, since $|||K|||^2 = \sum_j \theta_j^2$, the condition $0 < \theta_1 < |||K|||$ imposed in Theorem 5.2 is equivalent to asserting that at least two values of θ_j are nonzero. The condition $|||K||| < 1$ can be ensured by simply changing the scale on which X is measured, and so is imposed without loss of generality.

THEOREM 5.2. *Assume* (5.5), *that* $\theta_1, \theta_2, \ldots$ *is the eigenvalue sequence in the representation* (2.6), *ordered such that* (2.7) *holds, and that* $0 < \theta_1 < |||K||| < 1$. *Then* $\|\eta_j\| = O_p(|||K|||^j)$ *uniformly in* $1 \leq j \leq C \, n^{1/2}$, *and*

$$\begin{aligned} \widehat{h}_{jk} = h_{jk} + n^{-1/2} \int_{\mathcal{I}} \big\{ \xi_{j+1} \, K^k(b) + K^{j+1}(b) \, \xi_k \big\} \\ + O_p\big(n^{-1} \, \theta_1^j \, |||K|||^k + n^{-2} \, |||K|||^{j+k}\big), \end{aligned} \tag{5.7}$$

uniformly in $1 \leq j \leq k \leq C \, n^{1/2}$ *as* $n \to \infty$, *for each* $C > 0$.

Our next result, Theorem 5.3, applies Theorems 5.1 and 5.2 to derive a stochastic expansion for the difference between the theoretical approximant

$g_p(x)$, at (3.10), and its estimator $\widehat{g}_p(x)$, at (4.5). Let $\Delta_{1jk} = \int_{\mathcal{I}}\{\xi_{j+1}\, K^k(b) + K^{j+1}(b)\, \xi_k\}$, denoting the coefficient of $n^{-1/2}$ in the expansion (5.7), and put $\Delta_1 = (\Delta_{1jk})$, a $p \times p$ matrix, and $\delta = (\Delta_{101}, \ldots, \Delta_{10p})^{\mathrm{T}}$, a p-vector. Also, let $\lambda = \lambda(p)$ be the smallest eigenvalue of the $p \times p$ matrix $H = (h_{jk})$, introduced in section 3.2.

THEOREM 5.3. *Under the conditions of Theorem 5.2, and if each $\theta_j > 0$,*

$$\left\| \widehat{\gamma} - \left\{ \gamma + n^{-1/2}\, H^{-1}\, (\delta - \Delta_1\, \gamma) \right\} \right\| = O_p\big(n^{-1}\, \lambda^{-3}\big)\,, \tag{5.8}$$

$$\begin{aligned} \widehat{g}_p(x) - g_p(x) = \bar{Y} - EY + n^{-1/2} \sum_{j=1}^{p} \Bigg[\big\{ H^{-1}(\delta - \Delta_1\, \gamma) \big\}_j \int_{\mathcal{I}} (x - EX) K^j(b) \\ + \gamma_j \int_{\mathcal{I}} \Big\{ (x - EX)\, \xi_j - n^{1/2}\, (\bar{X} - EX)\, K^j(b) \Big\} \Bigg] \\ + O_p\Big(n^{-1}\, \lambda^{-1}\, \|\gamma\| + n^{-1}\, \lambda^{-3} \Big)\,, \end{aligned} \tag{5.9}$$

uniformly in functions x and integers p for which $\|x\| \le C$, $1 \le p \le C\, n^{1/2}$ and $n^{1/2}\lambda \to \infty$, where $C > 0$ is fixed but arbitrary.

Note that, by (3.8), $|h_{jk}| \le \theta_1^{j+k+1}\, \|b\|^2$, and therefore $\|Hv\| \le C_1\, \|v\|$ for all p-vectors v, where the constant C_1 does not depend on p. (Here we have used the condition $\theta_1 < 1$, which we introduced in Theorem 5.2 and also imposed in Theorem 5.3.) Hence $\lambda \le C_1$ for all p. Note too that since, for finite p, H is nonsingular (see section 3.2), then its smallest eigenvalue $\lambda = \lambda(p)$ is positive. On the other hand, when $p = \infty$ the sequence $\gamma_1, \gamma_2 \ldots$, that minimises (3.5) is not unique (see section 3.2), and so we can have $\lambda \to 0$ as $p \to \infty$. The condition $n^{1/2}\lambda \to \infty$ imposed in Theorem 5.3 reflects this property, and essentially puts an upper bound to the speed at which p can tend to infinity as a function of n.

5.3. *Implications of the main theorems and additional results.*

5.3.1. *Consistency and rates of convergence.* Let X_0 have the same distribution as $X_1 \ldots, X_n$ but be independent of those random functions, and let $\|\cdot\|_{\mathrm{pred}}$ denote the predictive L_2 norm, conditional on $X_1 \ldots, X_n$: if W is a random variable then $\|W\|_{\mathrm{pred}} = \big\{E\big(W^2 \,\big|\, X_1, \ldots, X_n\big)\big\}^{1/2}$. For example, taking $W = \widehat{g}_p(X_0) - g(X_0)$ we obtain a measure of the accuracy with which $\widehat{g}_p(X_0)$ predicts $g(X_0)$. We shall show in section 7.6 that if $p = p(n)$ is chosen to diverge no faster than $n^{1/2}$, and sufficiently slowly to ensure that

$$n^{-1/2}\, \lambda^{-1} \|\gamma\| + n^{-1}\, \lambda^{-3} \to 0 \tag{5.10}$$

as $n \to \infty$, then

$$\|\widehat{g}_p(X_0) - g(X_0)\|_{\rm pred} = O_p\Big\{n^{-1/2}\,\lambda^{-1}(1+\|\gamma\|) + n^{-1}\,\lambda^{-3} + t_p(\gamma_1,\ldots,\gamma_p)^{1/2}\Big\}\,, \tag{5.11}$$

where t_p is as at (3.5). It follows from Theorem 3.2 that if all of the eigenvalues θ_j are nonzero then $t_p(\gamma_1,\ldots,\gamma_p) \to 0$ as $p \to \infty$ (As remarked in the paragraph immediately below that theorem, the condition that each θ_j is nonzero can be dropped.) Therefore, (5.10) implies that $\widehat{g}_p(X_0)$ is consistent for $g(X_0)$.

Additionally, Theorems 5.1–5.3 make it clear that, provided p does not diverge too fast as a function of n, the quantities $\sup_{j\le p} \|\widehat{K^j(b)} - K^j(b)\|$, $\sup_{1\le j,k\le p} |\widehat{h}_{jk} - h_{jk}|$ and $\sup_{j\le p} |\widehat{\gamma}_j - \gamma_j|$ (see (5.12) below) converge in probability to zero as n diverges.

5.3.2. *Results in supremum metrics.* For our expansions of the function $\widehat{K^j(b)}$ at (5.6), and of the vector $\widehat{\gamma}$ at (5.8), our bounds on remainder terms are given in L_2 metrics. In either case they can be extended to the supremum metric. For example, (5.8) itself implies that:

$$\sup_{1\le j\le p} \Big|\widehat{\gamma}_j - \gamma_j - n^{-1/2}\,\big\{H^{-1}\,(\delta - \Delta_1\,\gamma)\big\}_j\Big| = O_p\big(n^{-1}\,\lambda^{-3}\big)\,. \tag{5.12}$$

Theorem 5.4, below, states a version of (5.6) in the L_∞ metric. It makes use of the following regularity conditions:

$$\text{for both } D_i \equiv 1 \text{ and } D_i \equiv Y_i\,,$$
$$\sup_{t\in\mathcal{I}} \left|\frac{1}{n^{1/2}} \sum_{i=1}^n \{X_i(t)\,D_i - EX_i(t)\,D_i\}\right| = O_p(1)\,, \tag{5.13}$$

$$\sup_{t\in\mathcal{I}} \int_{\mathcal{I}} \left|\frac{1}{n^{1/2}} \sum_{i=1}^n (1-E)\,\{X_i(s) - EX_i(s)\}\,\{X_i(t) - EX_i(t)\}\right|^2 ds = O_p(1)\,. \tag{5.14}$$

Conditions (5.13) and (5.14) will be discussed in appendix A.1.

THEOREM 5.4. *If* (5.5), (5.13) *and* (5.14) *hold then* $\sup_{t\in\mathcal{I}} |\xi_j(t)| = O_p(1)$ *for each* j, *and*

$$\sup_{t\in\mathcal{I}} \Big|\widehat{K^j(b)}(t) - \big\{K^j(b)(t) + n^{-1/2}\,\xi_j(t)\big\}\Big| = O_p\big(n^{-1}\big)\,.$$

5.3.3. *Interpreting stochastic expansions.* The coefficients of $n^{-1/2}$ in the expansions of $\widehat{K^j(b)}(t) - K^j(b)(t)$, $\widehat{h}_{jk} - h_{jk}$, $\widehat{\gamma}_j - \gamma_j$ and $\widehat{\gamma}_p(x) - \gamma_p(x)$ in (5.6), (5.7), (5.8) (see also (5.12)) and (5.9), respectively, are each equal to n^{-1} multiplied by a sum of n independent and identically distributed random variables with zero mean, plus a term that equals $O_p(n^{-1})$. In these cases, for fixed (j,t), (j,k), j and (p,x), respectively, the independent random variables do not depend on n. Therefore, their variances can be computed easily.

For example, in the case of $\widehat{h}_{jk} - h_{jk}$, using (5.7) and the definitions of ξ and ξ_0, we have, under the conditions of Theorem 5.2 and for each fixed j and k,

$$\widehat{h}_{jk} = h_{jk} + n^{-1} \sum_{i=1}^{n} Z_{ijk} + O_p\big(n^{-1}\big)\,, \tag{5.15}$$

where the independent and identically distributed random variables $Z_{1jk}, \ldots,$ Z_{njk} are given by

$$\begin{aligned}
Z_{ijk} = (1-E) \int_{\mathcal{I}} \Bigg(& K^k(b)(u) \Bigg[\{Y_i - E(Y_i)\} \int_{\mathcal{I}} K^j(u,t)\, \{X_i(t) - \mu(t)\}\, dt \\
& + \sum_{\ell=0}^{j-1} \int_{\mathcal{I}} \{X_i(t) - \mu(t)\}\, K^\ell(t,u)\, dt \int_{\mathcal{I}} K^{j-\ell}(b)(s)\, \{X_i(s) - \mu(s)\}\, ds \Bigg] \\
& + K^{j+1}(b)(u) \Bigg[\{Y_i - E(Y_i)\} \int_{\mathcal{I}} K^{k-1}(u,t)\, \{X_i(t) - \mu(t)\}\, dt \\
& + \sum_{\ell=0}^{k-2} \int_{\mathcal{I}} \{X_i(t) - \mu(t)\}\, K^\ell(t,u)\, dt \int_{\mathcal{I}} K^{k-\ell-1}(b)(s)\, \{X_i(s) - \mu(s)\}\, ds \Bigg] \Bigg)\, du\,.
\end{aligned}$$

The distribution of Z_{ijk} does not depend on n, and, under the assumption of finite fourth moment of X and finite second moment of ϵ (see (5.5)), Z_{ijk} has finite variance σ^2_{jk}, say. Hence, for each fixed j and k it follows from (5.15) that $n^{1/2}\,(\widehat{h}_{jk} - h_{jk})$ is asymptotically normal $\mathrm{N}(0, \sigma^2_{jk})$.

5.3.4. *Hankel matrix properties.* In section 3.2 we demonstrated that $\alpha_j = \int x^j\, m(dx)$, where m is the measure that places mass $(\beta_r^{\mathrm{PC}})^2\, \theta_r$ at the point θ_r for $r \geq 1$; m has no mass anywhere else. Therefore the $p \times p$ matrix $H = (\alpha_{j+k})$ is a Hankel matrix for which the associated nonnegative measure, m, is discrete and compactly supported. The latter property implies that m is completely determined by its moments α_j, and hence that the Hankel matrix is "determinate"; see, for example, Berg and Szwarc (2010). In such cases the smallest eigenvalue of H can converge to zero arbitrarily

fast as p diverges (Berg and Szwarc, 2010, Theorem 2.5), although more is known about the case where m is a continuous than that of a discrete measure, and it is particularly challenging to develop general theory describing properties of H^{-1} in the context of our measures m. (See Lascoux (1990) and Hou et al. (2003) for access to the literature on inverses of Hankel matrices and their determinants.) Nevertheless, as we noted in section 3.2, H is generally nonsingular for all p.

6. Numerical illustrations. In this section we illustrate, numerically, in a few examples, the fact that the algorithms in section 4.2 and appendix A.2 do indeed solve the same problem. We also illustrate the main difference between the PLS basis and the PCA basis, namely that PLS can capture the interaction between X and Y using a smaller number of terms than PCA.

In our first illustration, we take the X_is from a real data study, and generate the Y_is according to the linear model at (2.1). By choosing the population in this way we can represent, in simulations, the vagaries of real data, but can still compare the performance of our methodology with the "truth." We take the X_i curves from a benchmark Phoneme dataset, which can be downloaded from `www-stat.stanford.edu/ElemStatLearn`. In these data, $X_i(t)$ represents log-periodograms constructed from recordings of different phonemes. The periodograms are available at 256 equispaced frequencies t, which for simplicity we denote by $t = 1, 2, \ldots, 256$. Hence, in this example $\mathcal{I} = [1, 256]$. See Hastie et al. (2009) for more information about this dataset. We used the $N = 1717$ data curves $X_i(t)$ that correspond to the phonemes "aa" as in "dark" and "ao" as in "water".

We computed the first $J = 20$ empirical PCA basis functions $\widehat{\phi}_1(t), \ldots, \widehat{\phi}_{20}(t)$, and considered four different curves b, which we constructed by taking $b(t) = \sum_{j=1}^{J} a_j \widehat{\phi}_j(t)$ for four different sequences of a_js: (i) $a_j = (-1)^j \cdot 1\{j \leq 5\}$; (ii) $a_j = (-1)^j \cdot 1\{6 \leq j \leq 10\}$; (iii) $a_j = (-1)^j \cdot 1\{11 \leq j \leq 15\}$; (iv) $a_j = (-1)^j \cdot 1\{16 \leq j \leq 20\}$. These four models were chosen to illustrate clearly the advantages of the PLS basis over the PCA basis. Example (i) illustrates a situation particularly favourable to PCA, where the interaction between X and Y can be represented by the first few PCA basis functions. There we do not expect that PCA will need many more terms than PLS to achieve a small prediction error. On going from example (i) to example (iv), the function b is represented by five consecutively indexed PCA basis functions in each case, but with their indices successively larger. However, as we shall see below, in those cases too, PLS manages to construct a basis that captures the interaction between X and Y using only the first few terms.

In the four cases, for $i = 1, \ldots, N$ we generated the Y_is by taking $Y_i =$

$\int_{\mathcal{I}} X_i b + \epsilon_i$, where $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, and where $5\sigma^2$ was equal to the empirical variance of the $\int_{\mathcal{I}} bX_i$s calculated from the N observations. Then, in each case, we randomly split these N observations in two parts: a training sample of size n, and a test sample of size $N - n$. We did this 200 times for each of $n = 30$, $n = 50$ and $n = 100$, so for each setting we generated 200 test and training samples.

For each set of test and training samples generated in this way, we constructed our predictor using only the test sample, and then we applied it to predict $\int_{\mathcal{I}} bX_i$ for each X_i in the associated training sample. In other words, we constructed $\bar{X}$, $\bar{Y}$ and $\widehat{b}$ from the training sample only, where $\widehat{b}$ was the empirical version of b_p calculated either via the first p terms of the PLS basis (calculated from the algorithm in appendix A.2 or the second algorithm of section 4.2), or via the first p terms of the PCA basis, for each of $p = 1, \ldots, 10$. Then, for each observation X_i in the test sample, we calculated the predictor $\widehat{Y}_i = \bar{Y} + \int_{\mathcal{I}} \widehat{b}(X_i - \bar{X})$ of $\int_{\mathcal{I}} bX_i$. Note that this predictor includes the estimator $\bar{Y} - \int_{\mathcal{I}} b\bar{X}_i$ of the intercept because, although our data were generated from a model with no intercept, in practice we are not supposed to know this.

To quantify the quality of prediction, we calculated the prediction error $\mathrm{PE} = (N-n)^{-1} \sum_{i=1}^{N-n} (\widehat{Y}_i - \int_{\mathcal{I}} bX_i)^2$ in each case, for each method, and for each test sample. In figure 1 we show boxplots of these prediction errors calculated in each case from the 200 test samples. Note that here the two PLS algorithms gave exactly the same estimators, and so the boxplots only show the results for the standard PLS algorithm and for the PCA method. These boxplots show that as the information about the interaction between X and Y moves further away in the sequence of $\widehat{\phi}_j$s (that is, going from case (i) to case (iv)), PLS can capture the interaction using fewer terms than PCA. For example, in case (i), PLS took $p = 3$ components to reach the prediction error that PCA reached with $p = 5$, but in case (iv), the prediction error was already very small for PLS with $p = 10$, and was still very large for PCA with $p = 10$. We also calculated the integrated squared error $\mathrm{ISE} = \int_{\mathcal{I}} (\widehat{b} - b)^2$ for each method and test sample. In figure 2 we show boxplots of these ISEs calculated from the 200 test samples, for models (i), (iii) and (iv). We can see that the PLS estimator of b needs less components than the PCA estimator to reach small ISE values.

In our second example we took the orange juice data which comprise $N = 216$ observations $\big(X_i(t), Y_i\big)$, $i = 1, \ldots, N$, where each Y_i is the saccharose content of a sample of orange juice and X_i is a curve representing the first derivative of near-infrared spectra of the juice at 700 equispaced points t. We take $t = 1, \ldots, 700$ (hence $\mathcal{I} = [1, 700]$). The data can be found at

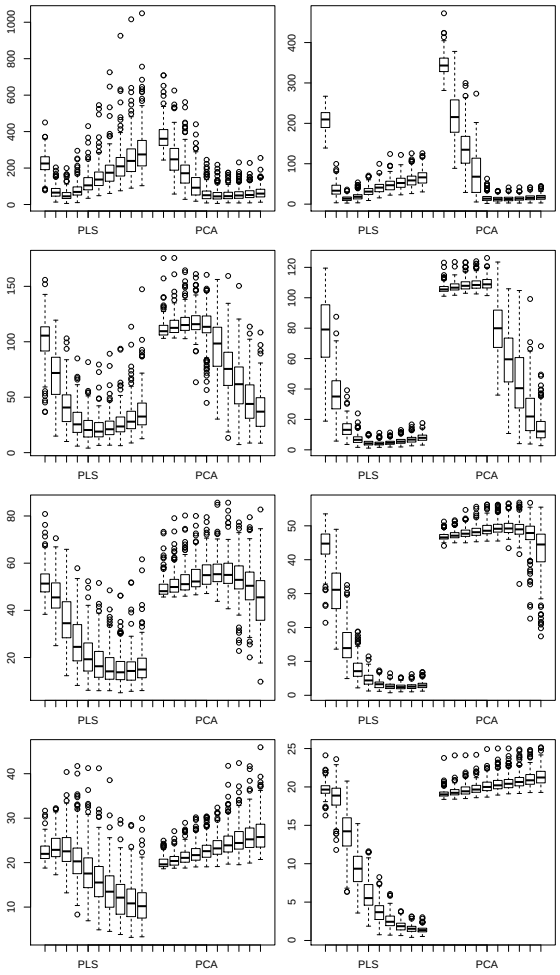

FIG 1. *Boxplots of the prediction error using the first p PLS components (first group of 10 boxes) or the first p PCA components (last group of 10 boxes), calculated from 200 samples of sizes* $n = 30$ *(first column) or* $n = 100$ *(second column) generated from the phoneme data. The curve b is that in case (i), (ii), (iii) and (iv), in respectively row 1, 2, 3 and 4. From left to right, each group of 10 boxplots addresses the settings indexed by* $p = 1$ *to* $p = 10$.

`www.ucl.ac.be/mlg/index.php?page=DataBases`. As with our simulated data above, we split the observations randomly into a training sample of size n and a test sample of size $N - n$, for each of $n = 30$, 50 and 100. We did this 200 times for each n. Then in each case we calculated our predictor, as above, from the training sample, and applied it for predicting $\int_{\mathcal{I}} X_i b$ for the corresponding test sample. Here we did not know the true model, so we calculated an estimator of the prediction error as $\widehat{\mathrm{PE}} = (N -$

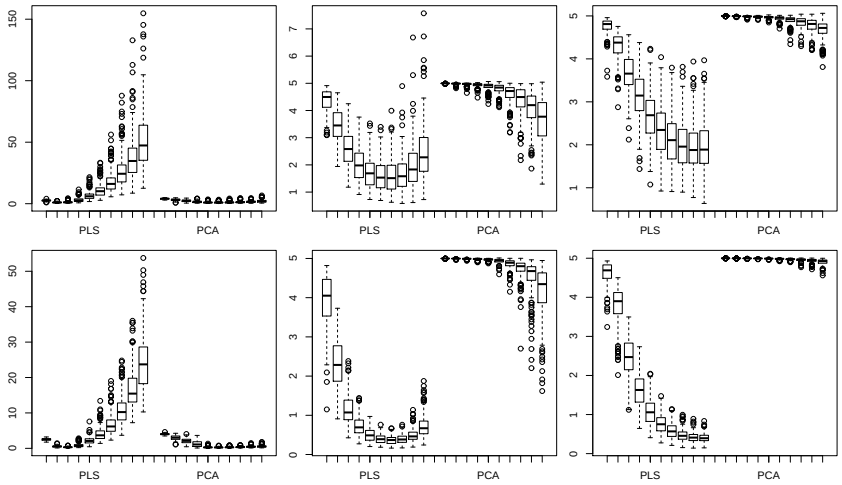

FIG 2. *Boxplots of the ISE of* $\widehat{b}$ *using the first* p *PLS components (first group of 10 boxes) or the first* p *PCA components (last group of 10 boxes), calculated from 200 samples of sizes* $n = 30$ *(first row) or* $n = 100$ *(second row) generated from the phoneme data. The curve* b *is that in case (i), (iii) and (iv), in respectively column 1, 2 and 3. From left to right, each group of 10 boxplots addresses the settings indexed by* $p = 1$ *to* $p = 10$.

$n)^{-1} \sum_{i=1}^{N-n} (\widehat{Y}_i - Y_i)^2$, for each (X_i, Y_i) in the test sample. In this way we obtained 200 values of $\widehat{\text{PE}}$ for each n. Figure 3 shows, for each n, boxplots of these 200 $\widehat{\text{PE}}$s, for $p = 1$ to 8. As above, the two PLS algorithms (the algorithm in appendix A.2 and the second algorithm of section 4.2) gave exactly the same results, except for $p = 8$ where the numerical roundings of both methods differed somewhat. Therefore we show the boxplots for both algorithms. In this example too we can see that the two PLS algorithms clearly solve the same problem, and that PLS needs fewer terms (i.e. p is smaller) to capture the same interactions as PCA. This can be advantageous when computing time is an issue, for example when a linear prediction is associated with a complex nonparametric procedure. For example, in Ferraty and Vieu (2006), the linear fit is used in combination with nonparametric estimators of $E(Y|X)$.

7. Technical arguments.

7.1. *Proof of Theorem 3.1.* Defining $\sigma^2 = \text{var}(\epsilon)$ we see that the right-hand side of (2.9) can be expressed as

$$\text{cov}\Big\{ \Big(\int_{\mathcal{I}} b\, X \Big) - \sum_{j=1}^{p-1} \Big(\int_{\mathcal{I}} b\, \psi_j \Big) \Big(\int_{\mathcal{I}} X\, \psi_j \Big), \int_{\mathcal{I}} X\, \psi_p \Big\}$$

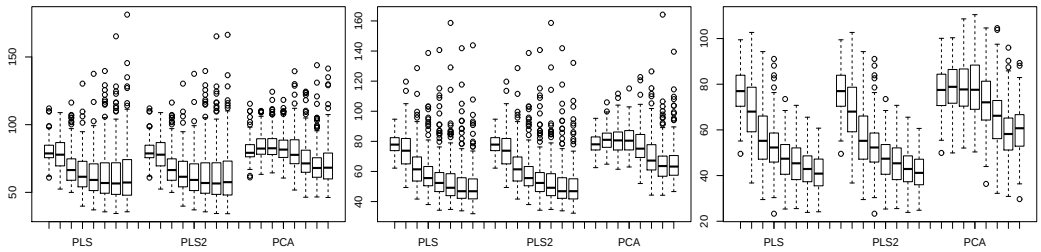

FIG 3. *Boxplots of the estimated prediction error using the first p PLS components calculated by the algorithm of appendix A.2 (first group of 8 boxes) or the second algorithm of section 4.2 (second group of 8 boxes, denoted by PLS2), or the first p PCA components (last group of 8 boxplots). Each box was calculated from 200 samples of sizes* $n = 30$ *(first column),* $n = 50$ *(second column), or* $n = 100$ *(third column) drawn randomly from the orange data. From left to right, each group of 8 boxplots is for* $p = 1$ *to* $p = 8$.

$$= \int_{\mathcal{I}}\int_{\mathcal{I}} b\,\psi_p\,K - \sum_{j=1}^{p-1}\left(\int_{\mathcal{I}} b\,\psi_j\right)\left(\int_{\mathcal{I}}\int_{\mathcal{I}} \psi_j\,\psi_p\,K\right).$$

The partial derivative of the right-hand side here, with respect to ψ_p, equals

$$K\left\{b - \sum_{j=1}^{p-1}\left(\int_{\mathcal{I}} b\,\psi_j\right)\psi_j\right\}. \tag{7.1}$$

The equation in c_k at (3.2) is the result of adjoining Lagrange multipliers on the right-hand side so as to accommodate the first $p-1$ constraints in (2.10). The factor c_0 on the right-hand side of (3.1) accommodates the last constraint in (2.10).

7.2. *Proof of Theorem 3.2.* Recall that $\mathcal{C}(\mathcal{I})$ is the space of all square-integrable functions on $\mathcal{I}$, and suppose $b = \sum_j \beta_j^{\mathrm{PC}}\,\phi_j \in \mathcal{C}(\mathcal{I})$. Write $\mathcal{C}_p(\mathcal{I})$ for the p-dimensional subspace of $\mathcal{C}(\mathcal{I})$ generated by the PCA basis functions $\phi_1, \ldots, \phi_p$, and let K_p denote the transformation that takes $b_p \equiv \sum_{1\le j\le p} \beta_j^{\mathrm{PC}}\,\phi_j \in \mathcal{C}_p(\mathcal{I})$ to $\sum_{1\le j\le p} \theta_j \beta_j^{\mathrm{PC}}\,\phi_j$. Now,

$$(\theta_1\,I - K_p)\ldots(\theta_p\,I - K_p)\,b_p = 0$$

for all $b_p \in \mathcal{C}_p(\mathcal{I})$. Therefore,

$$a_0\,b_p + a_1\,K_p(b_p) + \ldots + a_p\,K_p^p(b_p) = 0 \tag{7.2}$$

for all $b_p \in \mathcal{C}_p(\mathcal{I})$, where $a_0, \ldots, a_p$ are constants and $a_0 = \theta_1 \ldots \theta_p$. In particular, a_0 is nonzero, and so (7.2) implies that, for constants $c_1, \ldots, c_p$,

$$b_p = c_1\,K(b_p) + \ldots + c_p\,K^p(b_p)\,. \tag{7.3}$$

Let P_p denote the projection operator that takes $b = \sum_j \beta_j^{\rm PC}\,\phi_j \in \mathcal{C}(\mathcal{I})$ to $P_p(b) = b_p \in \mathcal{C}_p(\mathcal{I})$. Since P_p and K commute then $K^j(b_p) = K^j\,P_p(b) = P_p\,K^j(b)$. Therefore (7.3) implies that $b_p = P_p\{c_1\,K(b) + \ldots + c_p\,K^p(b)\}$, or equivalently,

$$P_p\big[b - \big\{c_1\,K(b) + \ldots + c_p\,K^p(b)\big\}\big] = 0\,. \tag{7.4}$$

In view of (7.4), if we let $\mathcal{D}(\mathcal{I})$ denote the vector space generated by $K(b)$, $K^2(b), \ldots$, and if we define $P_p\{\mathcal{D}(\mathcal{I})\} = \{P_p(z) : z \in \mathcal{D}(\mathcal{I})\}$, then $P_p(b) \in P_p\{\mathcal{D}(\mathcal{I})\}$ for all p. Now, $P_p\{\mathcal{D}(\mathcal{I})\} \subseteq \mathcal{D}(\mathcal{I})$, which is closed under limit operations in L_2. Therefore, the limit as $p \to \infty$ of $P_p(b)$, i.e. b, must be in $\mathcal{D}(\mathcal{I})$.

7.3. *Proof of Theorem 5.1.* Assume it can be proved that (5.6) holds, with ξ_j and η_j satisfying (5.3) and (5.4), for a particular $j \geq 1$; in view of (5.1), (5.6) is valid for $j = 1$. Then,

$$\begin{aligned}
\widehat{K^{j+1}}(b)(t) &= \int_{\mathcal{I}} \widehat{K^j(b)}(s)\,\widehat{K}(s,t)\,ds \\
&= \int_{\mathcal{I}} \big\{K^j(b) + n^{-1/2}\xi_j + n^{-1}\eta_j\big\}(s)\big(K + n^{-1/2}\xi + n^{-1}\eta\big)(s,t)\,ds \\
&= K^{j+1}(b)(t) + n^{-1/2}\int_{\mathcal{I}} \big\{K^j(b)(s)\,\xi(s,t) + \xi_j(s)\,K(s,t)\big\}\,ds \\
&\quad + n^{-1}\int_{\mathcal{I}} \big\{K^j(b)(s)\,\eta(s,t) + \eta_j(s)\,K(s,t) + \xi_j(s)\,\xi(s,t)\big\}\,ds \\
&\quad + n^{-3/2}\int_{\mathcal{I}} \{\xi_j(s)\,\eta(s,t) + \eta_j(s)\,\xi(s,t)\}\,ds \\
&\quad + n^{-2}\int_{\mathcal{I}} \eta_j(s)\,\eta(s,t)\,ds\,.
\end{aligned} \tag{7.5}$$

Therefore, taking ξ_{j+1} to be given by the coefficient of $n^{-1/2}$ in (7.5), and recalling the definition of ζ_j at (5.2), we have:

$$\begin{aligned}
\xi_{j+1}(t) &= \int_{\mathcal{I}} \big\{K^j(b)(s)\,\xi(s,t) + \xi_j(s)\,K(s,t)\big\}\,ds \\
&= K(\xi_j)(t) + \zeta_j(t) = K^2(\xi_{j-1})(t) + K(\zeta_{j-1})(t) + \zeta_j(t)\,,
\end{aligned}$$

which, on iteration, gives (5.3).

Finally we derive the bound at (5.4) on the remainder, again arguing by induction; assuming that (5.4) holds for j we establish it for $j+1$. Taking η_{j+1} to equal n times the sum of the terms in n^{-1}, $n^{-3/2}$ and n^{-2} in (7.5), we deduce that

$$\eta_{j+1}(t) = \int_{\mathcal{I}} \big\{K^j(b)(s)\,\eta(s,t) + \eta_j(s)\,K(s,t) + \xi_j(s)\,\xi(s,t)\big\}\,ds$$

$$+ n^{-1/2} \int_{\mathcal{I}} \{\xi_j(s)\, \eta(s,t) + \eta_j(s)\, \xi(s,t)\}\, ds + n^{-1} \int_{\mathcal{I}} \eta_j(s)\, \eta(s,t)\, ds\,.$$

Therefore,

$$\begin{aligned}
\|\eta_{j+1}\| &\le \|K^j(b)\|\, \|\!|\eta|\!\| + \|\eta_j\|\, \|\!|K|\!\| + \|\xi_j\|\, \|\!|\xi|\!\| + n^{-1/2} \left(\|\xi_j\|\, \|\!|\eta|\!\| + \|\eta_j\|\, \|\!|\xi|\!\|\right) \\
&\qquad + n^{-1}\, \|\eta_j\|\, \|\!|\eta|\!\| \\
&\le \left(\left\|K^j(b)\right\| + \|\xi_j\|\right) R_2 + \|\eta_j\|\, R_1 \\
&\le \left(\left\|K^j(b)\right\| + \|\xi_j\|\right) R_2 + \left\{\left(\left\|K^{j-1}(b)\right\| + \|\xi_{j-1}\|\right) R_2 + \|\eta_{j-1}\| R_1\right\} R_1 \\
&= \left\{\left(\left\|K^j(b)\right\| + \|\xi_j\|\right) + \left(\left\|K^{j-1}(b)\right\| + \|\xi_{j-1}\|\right) R_1\right\} R_2 + \|\eta_{j-1}\|\, R_1^2 \\
&\le \left\{\sum_{k=1}^{j} \left(\left\|K^k(b)\right\| + \|\xi_k\|\right) R_1^{j-k}\right\} R_2 + \|\eta_1\|\, R_1^j\,,
\end{aligned}$$

where the last identity follows on iteration. Observe too that, by (5.1), $\eta_1 = \eta_0$. Therefore,

$$\|\eta_{j+1}\| \le R_1^j\, \|\eta_0\| + R_2 \sum_{k=1}^{j} R_1^{j-k} \left(\left\|K^k(b)\right\| + \|\xi_k\|\right). \tag{7.6}$$

Note too that, by (5.2), $\|\zeta_j\| \le \|K^j(b)\|\, \|\!|\xi|\!\|$, and so, by (5.3),

$$\|\xi_k\| \le \left\|\!\left|K^{k-1}\right|\!\right\|\, \|\xi_0\| + \|\!|\xi|\!\| \sum_{\ell=0}^{k-2} \left\|\!\left|K^{\ell}\right|\!\right\|\, \left\|K^{k-\ell-1}(b)\right\|.$$

Hence, by (7.6),

$$\begin{aligned}
\|\eta_{j+1}\| - R_1^j\, \|\eta_0\| &\le R_2 \sum_{k=1}^{j} R_1^{j-k} \left\{\left\|K^k(b)\right\| + \left\|\!\left|K^{k-1}\right|\!\right\|\, \|\xi_0\| \right. \\
&\qquad \left. + \|\!|\xi|\!\| \sum_{\ell=0}^{k-2} \left\|\!\left|K^{\ell}\right|\!\right\|\, \left\|K^{k-\ell-1}(b)\right\| \right\} \\
&= R_2 \sum_{k=1}^{j} R_1^{j-k} \left(\left\|K^k(b)\right\| + \left\|\!\left|K^{k-1}\right|\!\right\|\, \|\xi_0\|\right) \\
&\qquad + R_2\, \|\!|\xi|\!\| \sum_{k=1}^{j} R_1^{j-k} \sum_{\ell=0}^{k-2} \left\|\!\left|K^{\ell}\right|\!\right\|\, \left\|K^{k-\ell-1}(b)\right\|. \qquad (7.7)
\end{aligned}$$

Result (5.4) for $j+1$ follows from (7.7).

7.4. *Proof of Theorem 5.2.* The representation (5.6) implies that

$$\widehat{h}_{jk} = h_{jk} + n^{-1/2} \int_{\mathcal{I}} \{\xi_{j+1}\, K^k(b) + K^{j+1}(b)\, \xi_k\} + n^{-1}\, R_{jk}\,, \tag{7.8}$$

where

$$\begin{aligned}
\widehat{h}_{jk} &= \int_{\mathcal{I}} \widehat{K^{j+1}}(b)\, \widehat{K^k(b)}\,, \quad h_{jk} = \int_{\mathcal{I}} K^{j+1}(b)\, K^k(b)\,, \\
|R_{jk}| &\le \Big| \int_{\mathcal{I}} \{\xi_{j+1}\, \xi_k + K^{j+1}(b)\, \eta_k + K^k(b)\, \eta_{j+1}\} \\
&\qquad + n^{-1/2} \int_{\mathcal{I}} (\eta_{j+1}\, \xi_k + \xi_{j+1}\, \eta_k) + n^{-1} \int_{\mathcal{I}} \eta_{j+1}\, \eta_k \Big| \\
&\le \|\xi_{j+1}\|\, \|\xi_k\| + \big\|K^{j+1}(b)\big\|\, \|\eta_k\| + \big\|K^k(b)\big\|\, \|\eta_{j+1}\| \\
&\qquad + n^{-1/2}\, (\|\eta_{j+1}\|\, \|\xi_k\| + \|\xi_{j+1}\|\, \|\eta_k\|) + n^{-1}\, \|\eta_{j+1}\|\, \|\eta_k\|\,.
\end{aligned} \tag{7.9}$$

Next we bound $|R_{jk}|$. Observe that $|\!|\!| K^k |\!|\!|^2 = \sum_j \theta_j^{2k} = O(\theta_1^{2k})$, $\|K^k(b)\|^2 = \sum_j \theta_j^{2k}\, (\int_{\mathcal{I}} b\phi_j)^2 = O(\theta_1^{2k})$ and $|\!|\!|\eta|\!|\!| + |\!|\!|\xi|\!|\!| + \|\eta_0\| + \|\xi_0\| = O_p(1)$ as $n \to \infty$. Hence, by (5.2), $\|\zeta_j\| \le \|K^j(b)\|\, |\!|\!|\xi|\!|\!| = O_p(\theta_1^j)$, uniformly in $j \ge 1$, and therefore by (5.3),

$$\|\xi_j\| = O_p\Big(\theta_1^j + \sum_{k=0}^{j-2} \theta_1^k\, \theta_1^{j-k-1}\Big) = O_p\big(j\, \theta_1^j\big)\,, \tag{7.10}$$

uniformly in $j \ge 1$. Note too that

$$R_1^j = \Big\{ \big(|\!|\!| K |\!|\!| + n^{-1/2}\, |\!|\!|\xi|\!|\!| + n^{-1}\, |\!|\!|\eta|\!|\!|\big)^j \Big\} = O_p\big(|\!|\!| K |\!|\!|^j\big)\,,$$

uniformly in $1 \le j \le C\, n^{1/2}$, for any $C > 0$. More simply, $R_2 = O_p(1)$. Hence, by (5.4),

$$\|\eta_j\| = O_p\Big(|\!|\!| K |\!|\!|^j + \sum_{k=1}^{j-1} |\!|\!| K |\!|\!|^{j-k-1}\, \theta_1^k + \sum_{k=1}^{j-1} |\!|\!| K |\!|\!|^{j-k-1}\, k\, \theta_1^k \Big) = O_p\big(|\!|\!| K |\!|\!|^j\big)\,, \tag{7.11}$$

uniformly in $1 \le j \le C\, n^{1/2}$. (Here we have used the property $0 < \theta_1 < |\!|\!| K |\!|\!| < 1$.) Combining (7.9)–(7.11) we find that:

$$R_{jk} = O_p\Big\{ jk\, \theta_1^{j+k} + \theta_1^j\, |\!|\!| K |\!|\!|^k + \theta_1^k\, |\!|\!| K |\!|\!|^j \tag{7.12}$$

$$+ n^{-1/2}\big(|||K|||^j\,\theta_1^k + |||K|||^k\,\theta_1^j\big) + n^{-1}\,|||K|||^{j+k}\Big\}$$
$$= O_p\big(\theta_1^j\,|||K|||^k + n^{-1}\,|||K|||^{j+k}\big)\,, \tag{7.13}$$

uniformly in $1 \le j \le k \le C\,n^{1/2}$. Theorem 5.2 follows from (7.8) and (7.13).

7.5. *Proof of Theorem 5.3.* From (3.10), (3.11) and (4.5) we deduce that:

$$\begin{aligned}
&\widehat{g}_p(x) - g_p(x) - (\bar{Y} - EY)\\
&= \sum_{j=1}^{p} \left\{ \widehat{\gamma}_j \int_{\mathcal{I}} (x - \bar{X})\,\widehat{K^j(b)} - \gamma_j \int_{\mathcal{I}} (x - EX)\,K^j(b) \right\}\\
&= \sum_{j=1}^{p} \left[\big(\widehat{\gamma}_j - \gamma_j\big) \int_{\mathcal{I}} (x - EX)\,K^j(b) + \gamma_j \int_{\mathcal{I}} (x - EX)\,\big\{\widehat{K^j(b)} - K^j(b)\big\} \right.\\
&\qquad - \gamma_j \int_{\mathcal{I}} (\bar{X} - EX)\,K^j(b) + \big(\widehat{\gamma}_j - \gamma_j\big) \int_{\mathcal{I}} (x - EX)\,\big\{\widehat{K^j(b)} - K^j(b)\big\}\\
&\qquad \left. - (\widehat{\gamma}_j - \gamma_j) \int_{\mathcal{I}} (\bar{X} - EX)\,K^j(b) - \widehat{\gamma}_j \int_{\mathcal{I}} (\bar{X} - EX)\,\big\{\widehat{K^j(b)} - K^j(b)\big\} \right].
\end{aligned} \tag{7.14}$$

Combining (5.6), (7.10) and the bound $\|\eta_j\| = O_p(|||K|||^j)$, valid uniformly in $1 \le j \le C\,n^{1/2}$ for each $C > 0$ and given in Theorem 5.2, we deduce that

$$\big\|\widehat{K^j(b)} - K^j(b)\big\| \le n^{-1/2}\,\|\xi_j\| + n^{-1}\,\|\eta_j\| = O_p\big(n^{-1/2}\,j\,\theta_1^j + n^{-1}\,|||K|||^j\big)\,, \tag{7.15}$$

uniformly in $1 \le j \le C\,n^{1/2}$ for each $C > 0$.

More simply, $\|\bar{X} - EX\| = O_p(n^{-1/2})$. Combining this bound with (7.10), (7.14) and the properties $\|x\| \le C$ and $\|K^j(b)\| = O(\theta_1^j)$, we deduce that

$$\begin{aligned}
&\widehat{g}_p(x) - g_p(x) - (\bar{Y} - EY)\\
&\quad = \sum_{j=1}^{p} \left[\big(\widehat{\gamma}_j - \gamma_j\big) \int_{\mathcal{I}} (x - EX)\,K^j(b) \right.\\
&\qquad \left. + \gamma_j \int_{\mathcal{I}} (x - EX)\,\big\{\widehat{K^j(b)} - K^j(b)\big\} - \gamma_j \int_{\mathcal{I}} (\bar{X} - EX)\,K^j(b) \right]\\
&\qquad + O_p\left\{ n^{-1/2} \sum_{j=1}^{p} \big(|\widehat{\gamma}_j - \gamma_j| + n^{-1/2}\,|\widehat{\gamma}_j|\big)\,\big(j\,\theta_1^j + n^{-1/2}\,|||K|||^j\big) \right\},
\end{aligned} \tag{7.16}$$

uniformly in $1 \le p \le C\,n^{1/2}$ and $\|x\| \le C$, for each $C > 0$. Using (5.6) and the bound $\|\eta_j\| = O_p(|\!|\!|K|\!|\!|^j)$, we deduce from (7.16) that

$$\begin{aligned}\widehat{g}_p(x) - g_p(x) - (\bar{Y} - EY) \\ &= \sum_{j=1}^{p} \Bigg[\big(\widehat{\gamma}_j - \gamma_j\big) \int_{\mathcal{I}} (x - EX)\, K^j(b) \\ &\qquad + \gamma_j \int_{\mathcal{I}} \Big\{ n^{-1/2}\,(x - EX)\,\xi_j - (\bar{X} - EX)\,K^j(b) \Big\} \Bigg] \\ &\qquad + O_p\Bigg[n^{-1/2} \sum_{j=1}^{p} \Big\{ |\widehat{\gamma}_j - \gamma_j| \,\big(j\,\theta_1^j + n^{-1/2}\,|\!|\!|K|\!|\!|^j\big) \\ &\qquad\qquad + n^{-1/2}\,(|\widehat{\gamma}_j| + |\gamma_j|)\,|\!|\!|K|\!|\!|^j \Big\} \Bigg], \end{aligned} \tag{7.17}$$

uniformly in $1 \le p \le C\,n^{1/2}$ and $\|x\| \le C$, for each $C > 0$.

Given any $p \times p$ matrix M, define its norm by $\|M\| = \sup_{v\,:\,\|v\|=1} \|Mv\|$. Writing Δ for a particular $p \times p$ matrix, and recalling that $\lambda = \lambda(p)$ denotes the smallest eigenvalue of H, we have $\|\Delta\,H^{-1}\| \le \|\Delta\|/\lambda$. Therefore, if $\widehat{H} = (\widehat{h}_{jk})$ is the $p \times p$ matrix obtained when $\widehat{h}_{jk}$ is defined as at (4.3), and we put $\Delta = \widehat{H} - H$, then, provided that $\|\Delta\|/\lambda \le \rho$ where $\rho \in (0,1)$ is fixed, we have:

$$\widehat{H}^{-1} = \big(I + H^{-1}\,\Delta\big)^{-1}\,H^{-1} = \big[I - H^{-1}\,\Delta + O_p\big\{(\|\Delta\|/\lambda)^2\big\}\big]\,H^{-1}\,. \tag{7.18}$$

Here the matrix M represented by $O_p\{(\|\Delta\|/\lambda)^2\}$ is interpreted as having the property $\|Mv\| \le (1-\rho)^{-1}\,(\|\Delta\|/\lambda)^2\,\|v\|$ for all p-vectors v (provided that $\|\Delta\|/\lambda \le \rho$), where on this occasion $\|Mv\|$ and $\|v\|$ denote vector norms of the indicated quantities, and $\|\Delta\|$ is the matrix norm of Δ.

We know from (5.7) that $\widehat{h}_{jk} = h_{jk} + n^{-1/2}\,\Delta_{1jk} + n^{-1}\,\Delta_{2jk}$, where

$$\Delta_{1jk} = \int_{\mathcal{I}} \big\{\xi_{j+1}\,K^k(b) + K^{j+1}(b)\,\xi_k\big\}\,,\ |\Delta_{2jk}| = O_p\big(\theta_1^j\,|\!|\!|K|\!|\!|^k + n^{-1}\,|\!|\!|K|\!|\!|^{j+k}\big)\,, \tag{7.19}$$

the latter property holding uniformly in $1 \le j \le k \le C\,n^{1/2}$. Note too that, by (7.10), $\|\xi_j\| = O_p(j\,\theta_1^j)$, uniformly in $j \ge 1$, and that $\|K^j(b)\| = O(\theta_1^j)$, so $|\Delta_{1jk}| = O_p\{\max(j,k)\,\theta_1^{j+k}\}$. Therefore, if we define $\Delta_{jk} = \widehat{h}_{jk} - h_{jk}$ then, since $\theta_1 < |\!|\!|K|\!|\!| < 1$, we have: $n\,\sum\sum_{j,k \le p} \Delta_{jk}^2 = O_p(1)$, uniformly in $p \le C\,n^{1/2}$. Hence, $\|\Delta\| = O_p(n^{-1/2})$, uniformly in $p \le C\,n^{1/2}$, where $\Delta = (\Delta_{jk})$ is a $p \times p$ matrix. Therefore, if p is chosen to diverge so slowly that $p = O(n^{1/2})$ and $\lambda = \lambda(p)$ satisfies $n^{1/2}\,\lambda \to \infty$ then, by (7.18),

$$\widehat{H}^{-1} = \big\{I - H^{-1}\,\Delta + O_p\big(n^{-1}\,\lambda^{-2}\big)\big\}\,H^{-1}\,, \tag{7.20}$$

uniformly in $p \le C\,n^{1/2}$. (Here $O_p(n^{-1}\,\lambda^{-2})$ denotes a $p \times p$ matrix, M say, for which $\|Mv\|/\|v\| = O_p(n^{-1}\,\lambda^{-2})$ uniformly in nonzero p-vectors v.) Note too that, if we define Δ_ℓ to be the $p \times p$ matrix with (j,k)th element $\Delta_{\ell jk}$, for $\ell = 1, 2$, then, in view of the second formula at (7.19), $\sum\sum_{j,k \le p} \Delta_{2jk}^2 = O_p(1)$, and so $\|\Delta_2\| = O_p(1)$ uniformly in $p \le C\,n^{1/2}$. Therefore (7.20) and the property $\Delta = \widehat{H} - H = n^{-1/2}\,\Delta_1 + n^{-1}\,\Delta_2$ imply that

$$\widehat{H}^{-1} = \big\{I - n^{-1/2}\,H^{-1}\,\Delta_1 + O_p\big(n^{-1}\,\lambda^{-2}\big)\big\}\,H^{-1}\,. \tag{7.21}$$

(here we used the fact that $\lambda \le h_{1,1} = O(1)$).

Recalling the definitions of $\widehat{h}_{jk}$, $\widehat{\alpha}_j$ and α_j at (4.3), (4.4) and (3.9), we deduce that $\widehat{\alpha}_j = \widehat{h}_{0j}$. Noting that result (5.7) can be extended to $\widehat{h}_{0j}$, we have that $\widehat{\alpha}_j = \alpha_j + n^{-1/2}\,\Delta_{10j} + n^{-1}\,\Delta_{20j}$, where Δ_{10j} and Δ_{20j} are given by (7.19). Note too that, by (4.2) and (3.7), $\widehat{\gamma}_j = (\widehat{H}^{-1}\widehat{\alpha})_j$ and $\gamma_j = (H^{-1}\alpha)_j$, where $\alpha = (\alpha_1, \ldots, \alpha_p)^{\mathrm{T}}$ and $\widehat{\alpha} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_p)^{\mathrm{T}}$. Since $K^j(b) = O(\theta_1^j)$ uniformly in $j \ge 1$, $\|\eta_j\| = O_p(|\!|\!|K|\!|\!|^j)$ uniformly in $1 \le j \le C\,n^{1/2}$ (see Theorem 5.2), and $\|\xi_j\| = O_p(j\,\theta_1^j)$ uniformly in $j \ge 1$ (see (7.10)), then, by (5.6), $\|\widehat{K^j(b)}\| = O_p(\theta_1^j + n^{-1/2}\,j\,\theta_1^j + n^{-1}\,|\!|\!|K|\!|\!|^j)$ uniformly in $1 \le j \le C\,n^{1/2}$. Using formula (4.4) for $\widehat{\alpha}_j$, and the fact that $0 < \theta_1 < |\!|\!|K|\!|\!| < 1$, we deduce that:

$$\|\widehat{\alpha}\| \le \Big\{\sum_{j=1}^{p} \big\|\widehat{K(b)}\big\|^2\,\big\|\widehat{K^j(b)}\big\|^2\Big\}^{1/2} = O_p(1)\,, \tag{7.22}$$

uniformly in $1 \le p \le C\,n^{1/2}$.

Therefore, defining $\delta = (\Delta_{101}, \ldots, \Delta_{10p})^{\mathrm{T}}$, we have, by (7.21),

$$\begin{aligned}\widehat{\gamma} = \widehat{H}^{-1}\,\widehat{\alpha} &= H^{-1}\,\big(\alpha + n^{-1/2}\,\delta\big) - n^{-1/2}\,H^{-1}\,\Delta_1\,H^{-1}\,\alpha + O_p\big(n^{-1}\,\lambda^{-3}\big) \\ &= \gamma + n^{-1/2}\,H^{-1}\,(\delta - \Delta_1\,\gamma) + O_p\big(n^{-1}\,\lambda^{-3}\big)\,,\end{aligned} \tag{7.23}$$

uniformly in $1 \le p \le C\,n^{1/2}$, where the two vectors denoted by $O_p(n^{-1}\,\lambda^{-3})$ have the property that their norms equal $O_p(n^{-1}\,\lambda^{-3})$ uniformly in $1 \le p \le C\,n^{1/2}$.

Next we combine (7.17) and (7.23), obtaining:

$$\begin{aligned}&\widehat{g}_p(x) - g_p(x) - (\bar{Y} - EY) \\ &\qquad = \sum_{j=1}^{p} \Big[n^{-1/2}\,\big\{H^{-1}\,(\delta - \Delta_1\,\gamma)\big\}_j\,\int_{\mathcal{I}} (x - EX)\,K^j(b) \\ &\qquad\qquad + \gamma_j \int_{\mathcal{I}} \Big\{n^{-1/2}\,(x - EX)\,\xi_j - (\bar{X} - EX)\,K^j(b)\Big\}\Big]\end{aligned}$$

$$+ O_p\Big[n^{-1}\,\lambda^{-3} + n^{-1/2}\sum_{j=1}^{p}\Big\{|\widehat{\gamma}_j - \gamma_j|\,\big(j\,\theta_1^j + n^{-1/2}\,|\!|\!|K|\!|\!|^j\big) + n^{-1/2}\,(|\widehat{\gamma}_j| + |\gamma_j|)\,|\!|\!|K|\!|\!|^j\Big\}\Big], \tag{7.24}$$

uniformly in $1 \le p \le C\,n^{1/2}$ and $\|x\| \le C$ for each $C > 0$. Here we have used the fact that, if $V = (V_1, \ldots, V_p)^{\mathrm{T}}$ is the vector denoted by $O_p(n^{-1}\,\lambda^{-3})$ on the far right-hand side of (7.23), then

$$\sum_{j=1}^{p}\left|V_j\int_{\mathcal{I}}(x - EX)\,K^j(b)\right| \le \|V\|\left\{\sum_{j=1}^{p}\left|\int_{\mathcal{I}}(x - EX)\,K^j(b)\right|^2\right\}^2 = O_p\big(n^{-1}\,\lambda^{-3}\big)\,,$$

uniformly in $1 \le p \le C\,n^{1/2}$ and $\|x\| \le C$, since $\sum_{j\ge1}\|K^j(b)\|^2 < \infty$.

Note too that, since $\|K^j(b)\| = O(\theta_1^j)$ and $\|\xi_j\| = O_p(j\,\theta_1^j)$, uniformly in $1 \le j \le C\,n^{1/2}$, then by (7.19), $|\Delta_{1jk}| = O_p(jk\,\theta_1^{j+k})$, uniformly in $1 \le j, k \le C\,n^{1/2}$, and therefore,

$$\|\Delta_1\|^2 \le \sum_{j=1}^{p}\sum_{k=1}^{p}\Delta_{1jk}^2 = O_p(1)\,, \quad \|\delta\|^2 = \sum_{j=1}^{p}\Delta_{10j}^2 = O_p(1)\,,$$

uniformly in $1 \le p \le C\,n^{1/2}$. Hence, by (7.23) and (7.22),

$$\begin{aligned}\|\widehat{\gamma} - \gamma\| &= O_p\Big\{n^{-1/2}\,\lambda^{-1}\,(\|\delta\| + \|\Delta_1\|\,\|\gamma\|) + n^{-1}\,\lambda^{-3}\Big\}\\ &= O_p\Big\{n^{-1/2}\,\lambda^{-1}\,(1 + \|\gamma\|) + n^{-1}\,\lambda^{-3}\Big\}\,.\end{aligned}$$

Therefore,

$$\begin{aligned}&\sum_{j=1}^{p}\Big\{|\widehat{\gamma}_j - \gamma_j|\,\big(j\,\theta_1^j + n^{-1/2}\,|\!|\!|K|\!|\!|^j\big) + n^{-1/2}\,(|\widehat{\gamma}_j| + |\gamma_j|)\,|\!|\!|K|\!|\!|^j\Big\}\\ &\quad = O_p\big(\|\widehat{\gamma} - \gamma\| + n^{-1/2}\,\|\gamma\|\big) = O_p\Big\{n^{-1/2}\,\lambda^{-1}\,(1 + \|\gamma\|) + n^{-1}\,\lambda^{-3}\Big\}\,.\end{aligned} \tag{7.25}$$

Result (5.8) is a consequence of (7.23), and (5.9) follows from (7.24) and (7.25).

7.6. *Proof of* (5.11). To derive (5.11), note that minor modifications of the argument used to derive (5.9) can be employed to show that, under the conditions of Theorem 5.3,

$\|\widehat{g}_p(X_0) - g_p(X_0)\|_{\mathrm{pred}}$

$$= \Big\| \bar{Y} - EY + n^{-1/2} \sum_{j=1}^{p} \Big[\big\{ H^{-1} (\delta - \Delta_1 \gamma) \big\}_j \int_{\mathcal{I}} (X_0 - EX)\, K^j(b) + \gamma_j \int_{\mathcal{I}} \Big\{ (X_0 - EX)\, \xi_j - n^{1/2} (\bar{X} - EX)\, K^j(b) \Big\} \Big] \Big\|_{\rm pred} + O_p\Big(n^{-1}\, \lambda^{-1}\, \|\gamma\| + n^{-1}\, \lambda^{-3} \Big)\,, \tag{7.26}$$

uniformly in p satisfying $1 \le p \le C\, n^{1/2}$, for each $C > 0$. The predictive norm on the right-hand side of (7.26) can be shown to equal $O_p\{n^{-1/2}\, \lambda^{-1}\, (1 + \|\gamma\|)\}$, and so if (5.10) holds then

$$\|\widehat{g}_p(X_0) - g_p(X_0)\|_{\rm pred} = O_p\Big\{ n^{-1/2}\, \lambda^{-1} (1 + \|\gamma\|) + n^{-1}\, \lambda^{-3} \Big\}\,. \tag{7.27}$$

Since $\|g_p(X_0) - g(X_0)\|_{\rm pred} = t_p(\gamma_1, \ldots, \gamma_p)^{1/2}$ then (7.27) implies (5.11).

Acknowledgements. We are grateful to Peter Forrester and Alan McIntosh for helpful discussion.

References

Aguilera, M., Escabiasa, M., Preda, C. and Saporta, G. (2010). Using basis expansions for estimating functional PLS regression: Applications with chemometric data. *Chemom. Intell. Lab.* **104**, 289–305.

Apanasovich, T.V. and Goldstein, E. (2008). On prediction error in functional linear regression. *Statist. Probab. Lett.* **78**, 1807–1810.

Baillo, A. (2009). A note on functional linear regression. *J. Amer. Statist. Assoc.* **79**, 657–669.

Berg, C. and Szwarc, R. (2010). The smallest eigenvalue of Hankel matrices. Manuscript available at `http://arxiv.org/abs/0906.4506`.

Bro, R. and Eldén, L. (2009). PLS works. *J. Chemom.* **23**, 69–71.

Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34**, 2159–2179.

Cardot, H. and Sarda, P. (2008). Varying-coefficient functional linear regression models. *Commun. Statist. Theor. Meth.* **37**, 3186–3203.

Delaigle, A. and Hall, P. (2012). Achieving near-perfect classification for functional data. *J. Roy. Stat. Soc. Ser.* B, doi 10.1111/j.1467-9868.2011.01003.x.

Durand, J. and Sabatier, R. (1997). Additive splines for partial least squares regression. *J. Amer. Statist. Assoc.* **92**, 1546–1554.

Escabias, M., Aguilera, A.M. and Valderrama, M.J. (2007). Functional PLS logit regression model. *Comput. Statist. Data Anal.* **51**, 4891–4902.

Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109–148 .

Garthwaite, P.H. (1994). An interpretation of partial least squares. *J. Amer. Statist. Assoc.* **89**, 122–127.

Goutis, C. and Fearn, T. (1996). Partial least squares regression on smooth factors. *J. Amer. Statist. Assoc.* **91**, 627–632.

Hastie, T. Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd Ed. *Springer-Verlag.*

Helland, I. (1990). Partial least squares regression and statistical models, *Scand. J. Statist.* **17**, 97–114.

Höskuldsson, A. (1988). PLS regression methods. *J. Chemom.* **2**, 211–228.

Hou, Q.H., Lascoux, A. and Mu, Y.P. (2003). Evaluation of some Hankel determinants. *Adv. Appl. Math.* **34**, 845–852.

Krämer, N., Boulesteix, A.-L. and Tutz, G. (2008). Penalized partial least squares with applications to B-spline transformations and functional data. *Chemom. Intell. Lab.* **94**, 60–69.

Krämer, N. and Sugiyama, M. (2011). The degrees of freedom of partial least squares regression. *J. Amer. Statist. Assoc.* **106**, 697–705.

Lange, K. (1999). *Numerical Analysis for Statisticians.* Springer-Verlag.

Lascoux, A. (1990). Inversion des matrices de Hankel. *Lin. Algebra Appl.* **129**, 77–102.

Lorber, A., Wangen, L. E., and Kowalski, B. R. (1987). A theoretical foundation for the PLS algorithm. *J. Chemom.* **1**, 19–31.

Martens, H. and Naes, T. (1989). *Multivariate Calibration.* Wiley, New York.

Müller, H.-G. and Yao, F. (2010). Additive modelling of functional gradients. *Biometrika* **97**, 791–805.

Nguyen, D.V. and Rocke, D.M. (2004). On partial least squares dimension reduction for microarray-based classification: a simulation study. *Comput. Statist. Data Anal.* **46**, 407–425.

Phatak, A. and de Hoog, F. (2003). Exploiting the connection between PLS, Lanczos, and conjugate gradients: Alternative proofs of some properties of PLS. *J. Chemom.* **16**, 361–367.

Phatak, A., Rilley, P., and Penlidis, A. (2002). The asymptotic variance of the univariate PLS estimator. *Lin. Algebra Appl.* **354**, 245–253.

Preda, C. and Saporta, G. (2005a). PLS regression on a stochastic process. *Comput. Statist. Data Anal.* **48**, 149–158.

Preda, C. and Saporta, G. (2005b). Clusterwise PLS regression on a stochastic process. *Comput. Statist. Data Anal.* **49**, 99–108.

Preda, C., Saporta, G. and Lévéder, C. (2007). PLS classification of functional data. *Comput. Statist.* **22**, 223–235.

Reiss, P.T. and Ogden, R.T. (2007). Functional principal component regression and functional partial least squares. *J. Amer. Statist. Assoc.* **102**, 984–996.

Wold, H. (1975). Soft Modeling by Latent Variables; the Nonlinear Iterative Partial Least Squares Approach. In *Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett*, Ed. J. Gani. Academic Press, London.

Wu, Y., Fan, J. and Müller, H.-J. (2010). Varying-coefficient functional linear regression. *Bernoulli* **16**, 730–758.

Yao, F. and Müller, H.-G. (2010). Functional quadratic regression. *Biometrika* **97**, 49–64.

APPENDIX A: APPENDIX

A.1. Conditions (5.13) and (5.14). Here we give examples where (5.13) and (5.14) hold. Assume that $E(X) = 0$. Then the Karhunen-Loève expansion of X_i, founded on the principal component basis introduced in section 2.2, is given by $X_i = \sum_j \theta_j^{1/2}\, \xi_{ij}\, \phi_j$, where the random variables ξ_{ij}, for $j \geq 1$, are uncorrelated and have zero mean and unit variance. For simplicity we suppose that they have identical distributions with bounded fourth moments, that $E(\epsilon^4) < \infty$, and that the eigenvalues θ_j and eigenvectors ϕ_j satisfy the condition: $\sum_{j=1}^{\infty} \theta_j^{1/2} \sup_{t\in\mathcal{I}} |\phi_j(t)| < \infty$. Then,

$$\begin{aligned} E\biggl[\sup_{t\in\mathcal{I}} \biggl|\frac{1}{n^{1/2}} \sum_{i=1}^{n} \{X_i(t)\, D_i - EX_i(t)\, D_i\}\biggr|\biggr] \\ \leq \sum_{j=1}^{\infty} \theta_j^{1/2} \Bigl\{\sup_{t\in\mathcal{I}} |\phi_j(t)|\Bigr\}\, E\biggl|\frac{1}{n^{1/2}} \sum_{i=1}^{n} (1-E)\, \xi_{ij}\, D_i\biggr| \\ \leq \bigl(E\xi_{11}^4 \cdot ED_1^4\bigr)^{1/4} \sum_{j=1}^{\infty} \theta_j^{1/2} \Bigl\{\sup_{t\in\mathcal{I}} |\phi_j(t)|\Bigr\} < \infty\,, \end{aligned} \tag{A.1}$$

$$\begin{aligned} E\biggl[\sup_{t\in\mathcal{I}} \biggl|\frac{1}{n^{1/2}} \sum_{i=1}^{n} (1-E)\, \{X_i(s) - EX_i(s)\}\, \{X_i(t) - EX_i(t)\}\biggr|^2\biggr] \\ = E\biggl[\sup_{t\in\mathcal{I}} \biggl|\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} (\theta_j\, \theta_k)^{1/2}\, \phi_j(s)\, \phi_k(t) \biggl\{n^{-1/2} \sum_{i=1}^{n} (1-E)\, \xi_{ij}\, \xi_{ik}\biggr\}\biggr|^2\biggr] \\ \leq E\bigl(\xi_{11}^4\bigr) \biggl[\sum_{j=1}^{\infty} \theta_j^{1/2} \Bigl\{\sup_{t\in\mathcal{I}} |\phi_j(t)|\Bigr\}\biggr]^4\,, \end{aligned} \tag{A.2}$$

where we have used the properties

$$\begin{aligned} \biggl\{E\biggl|\frac{1}{n^{1/2}} \sum_{i=1}^{n} (1-E)\, \xi_{ij}\, D_i\biggr|\biggr\}^2 \leq E\bigl\{(\xi_{11}\, D_1)^2\bigr\} \leq \bigl(E\xi_{11}^4 \cdot ED_1^4\bigr)^{1/2}\,, \\ \biggl\{E\biggl|\frac{1}{n^{1/2}} \sum_{i=1}^{n} (1-E)\, \xi_{ij}\, \xi_{ik}\biggr|\biggr\}^2 \leq E\bigl\{(\xi_{1j}\, \xi_{1k})^2\bigr\} \leq E\bigl(\xi_{11}^4\bigr)\,. \end{aligned}$$

Properties (5.13) and (5.14) follow from (A.1) and (A.2), respectively.

A.2. Conventional implementation via the PLS basis. Inference is based on a dataset $\mathcal{X} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of independent data pairs distributed as (X, Y). We first introduce the centred data $X_i^{[1]} = X_i - \bar{X}$ and $Y_i^{[1]} = Y_i - \bar{Y}$, for $1 \le i \le n$. Here and below, a superscript in square brackets denotes the number, or index, of a step in our algorithm. The algorithm goes as follows. For $j = 1, \ldots, p$:

(1) Estimate ψ_j by the empirical covariance of $X_i^{[j]}$ and $Y_i^{[j]}$: $\widehat{\psi}_j = \sum_{i=1}^n X_i^{[j]} Y_i^{[j]} / \big\| \sum_{i=1}^n X_i^{[j]} Y_i^{[j]} \big\|$.

(2) Fit the models $Y_i^{[j]} = \beta_j \int_{\mathcal{I}} X_i^{[j]} \widehat{\psi}_j + \epsilon_i^{[j]}$ and $X_i^{[j]}(t) = \delta_j(t) \int_{\mathcal{I}} X_i^{[j]} \widehat{\psi}_j + \eta_i^{[j]}(t)$ by least-squares, i.e. take

$$\widehat{\beta}_j = \sum_{i=1}^n Y_i^{[j]} \int_{\mathcal{I}} X_i^{[j]} \widehat{\psi}_j \Big/ \sum_{i=1}^n \Big\{ \int_{\mathcal{I}} X_i^{[j]} \widehat{\psi}_j \Big\}^2,$$

$$\widehat{\delta}_j(t) = \sum_{i=1}^n X_i^{[j]}(t) \int_{\mathcal{I}} X_i^{[j]} \widehat{\psi}_j \Big/ \sum_{i=1}^n \Big\{ \int_{\mathcal{I}} X_i^{[j]} \widehat{\psi}_j \Big\}^2.$$

(3) Calculate $X_i^{[j+1]}(t) = X_i^{[j]}(t) - \widehat{\delta}_j(t) \int_{\mathcal{I}} X_i^{[j]} \widehat{\psi}_j$ and $Y_i^{[j+1]} = Y_i^{[j]} - \widehat{\beta}_j \int_{\mathcal{I}} X_i^{[j]} \widehat{\psi}_j$.

After completion of steps (1) to (3) for all j, define $M = \big(M_{j,k}\big)_{1 \le j,k \le p}$ by $M^{-1} = \Big(\int_{\mathcal{I}} \widehat{\delta}_j \widehat{\psi}_k \Big)_{1 \le j,k \le p}$. Then $\widehat{b}_p(t) = \sum_{j,k=1}^p \widehat{\beta}_k M_{j,k} \widehat{\psi}_j(t)$ and $\tilde{g}_p(x) = \bar{Y} + \int_{\mathcal{I}} \widehat{b}_p \, (x - \bar{X})$.

A.3. Modified Gram-Schmidt algorithm. This algorithm turns a set of linearly independent functions $v_1, \ldots, v_p$ into a set of orthogonal functions $u_1, \ldots, u_p$, where orthogonality is defined with respect to a scalar product $\langle \cdot, \cdot \rangle$. For example, for the second algorithm in section 4.2, the scalar product between two functions f_1 and f_2 is defined by $\langle f_1, f_2 \rangle = \int_{\mathcal{I}} \int_{\mathcal{I}} f_1(s) f_2(t) \widehat{K}(s,t) \, ds \, dt$. The modified Gram-Schmidt algorithm is described in Lange (1999), section 7.7. It works as follows:

for $j = 1, \ldots, p$
  $u_j^{[1]} = v_j$
  for $i = 1, \ldots, j-1$
    $u_j^{[i+1]} = u_j^{[i]} - \langle u_j^{[i]}, u_i \rangle \, u_i$
  end loop i
  $u_j = u_j^{[j]} / \| u_j^{[j]} \|$
end loop j

Department of Mathematics and Statistics, University of Melbourne, Parkville, VIC, 3010, Australia.
E-mail: A.Delaigle@ms.unimelb.edu.au
halpstat@ms.unimelb.edu.au