Electronic Journal of Statistics
Vol. 0 (0000)
ISSN: 1935-7524
DOI: 0

Weighted least squares methods for prediction in the functional data linear model

Aurore Delaigle*

Department of Mathematics and Statistics, University of Melbourne, VIC, 3010, Australia and Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.
e-mail: A.Delaigle@ms.unimelb.edu.au

Peter Hall*

Department of Mathematics and Statistics, University of Melbourne, VIC, 3010, Australia and Department of Statistics, University of California at Davis, Davis, CA 95616, USA.
e-mail: halpstat@ms.unimelb.edu.au

Tatiyana V.Apanasovich†

Department of Pharmacology and Experimental Therapeutics, Thomas Jefferson University, Philadelphia, PA 1910, USA.
e-mail: Tatiyana.Apanasovich@jefferson.edu

Abstract: The problem of prediction in functional linear regression is conventionally addressed by reducing dimension via the standard principal component basis. In this paper we show that weighted least-squares methods can be more effective when the experimental errors are heteroscedastic. We give a concise theoretical result which demonstrates the effectiveness of this approach, even when the model for the variance is inaccurate, and we explore the numerical properties of the method. We show too that the advantages of the suggested adaptive techniques are not found only in low-dimensional aspects of the problem; rather, they accrue almost equally among all dimensions.

AMS 2000 subject classifications: Primary 62J05, 62G05.
Keywords and phrases: Cross-validation, eigenfunction, eigenvector, functional data analysis, functional linear regression, mean squared error, orthogonal series, principal component analysis, rate of convergence, weighted linear regression.

Contents

*Supported by a grant from the Australian Research Council
†Supported by an NSF grant #0707106

1. Introduction

The functional linear model has the appearance of being rather conventional. It involves representing a scalar response, Y, as

$$Y = \alpha + \int_{\mathcal{I}} \beta\, X + \text{error}\,, \tag{1.1}$$

where X denotes the function-valued explanatory variable, α is a scalar, β is the function-valued slope parameter, and $\mathcal{I}$ is a known compact interval. However, estimation of β is generally a nonparametric problem, and the level of complexity implicit in that property can carry over to the problem of prediction, in which we wish to estimate $\alpha + \int_{\mathcal{I}} \beta\, x$ for a given function x. Sometimes $\alpha + \int_{\mathcal{I}} \beta\, x$ can be estimated root-n consistently, where n denotes sample size, but more commonly, estimators converge at strictly slower rates. Cai and Hall (2006) discuss these issues, and Faraway (1997), Ferraty and Vieu (2000, 2006), Cuevas *et al.* (2002), Ramsay and Silverman (2005, Chapter 12), Cardot *et al.* (2006) and Cardot and Sarda (2006) address functional linear regression in more general terms.

A standard approach to estimating α and β is to first estimate the principal component basis from a sample of observations of (X, Y), and then construct an estimator of $\mu(x) = \alpha + \int_{\mathcal{I}} \beta\, x$ in terms of that basis, using least squares. However, in practice the distribution of the error in (1.1) can be a functional of the distribution of X, and the optimal choice of basis can depend significantly on x. To address these challenges we could construct the basis so that it gave greater emphasis to observations of X that were relatively close to x. For example, we could restrict attention to X for which $\|X - x\| \le \delta$, where $\|\cdot\|$ was a suitable distance measure and δ played the role of bandwidth, although δ would not necessarily be chosen to converge to zero as n increased. More subtly, the basis could be constructed by applying kernel weights to each observation. See Mas (2008) for theoretical results addressing problems of this type.

Although this approach is attractive, practical difficulties can arise from the implicit reduction in sample size that is involved. An alternative method is to estimate the variance, $\sigma(X)^2$ say, of the distribution of the error in (1.1) conditional on X, and adapt prediction to the level of variability there. We suggest solving this problem by modelling $\sigma(x)^2$ as a function of $\alpha + \int_{\mathcal{I}} \beta\, x$, and using its

inverse, with x replaced by a data value X, as a weight in the basic least-squares problem. We then show that calculations can be simplified by computing a new principal component basis, adapted to heteroscedasticity. While our approach has some similarities with the weighted least squares method used for finite dimensional data, it differs significantly due to the intrinsic nonparametric, and infinite dimensional, characters of functional linear regression; we quantify these issues in theoretical terms.

In summary, this paper makes three main contributions. First, we show in section 2 that adaptive modification of the standard principal component basis, or a nearly-equivalent method based on weighted least-squares, can be advantageous when undertaking functional linear prediction, i.e. when estimating $\mu(x)$. Secondly, we suggest approximations to the value of $\sigma(x)^2$, and we employ them to construct a second basis, this time adapted to heteroscedasticity. Then, in sections 3 and 4 we show that this approach can give real and effective reductions in mean squared error, even when the model we use to estimate variance is not completely correct. Alternatively, a nonparametric approach can be used to estimate variance. These methodologies all have analogues in cases where Y is a multivariate response, although for simplicity and transparency we focus only on the univariate case.

The main theoretical result in section 3 gives a concise account of the way adaptive methods can improve the performance of estimators in functional linear regression. In particular, we show that the advantages accrue almost equally among all dimensions; they are not principally to be found in low-dimensional aspects of the problem.

Previous developments of principal components analysis for functional data play a central role in our work. Early contributions include Besse and Ramsay (1986), Ramsay and Dalzell (1991) and Rice and Silverman (1991). From that point a very substantial literature has developed, including but by no means limited to the work of Silverman (1995, 1996), Brumback and Rice (1998), Cardot *et al.* (1999, 2000, 2003), Cardot (2000), Girard (2000), James *et al.* (2000), Boente and Fraiman (2000), He *et al.* (2003), Ramsay and Silverman (2005, Chap. 8–10), Yao *et al.* (2005), Hall and Hosseini-Nasab (2006), and the work of Jank and Shmueli (2006), Ocaña *et al.* (2007), Reiss and Ogden (2007), Huang *et al.* (2008).

2. Methodology

2.1. Orthogonal series approach to inference in the linear model

The functional linear model argues that independent data pairs $(X_{[1]}, Y_1), \ldots, (X_{[n]}, Y_n)$, distributed as (X, Y), are generated as

$$Y = \alpha + \int_{\mathcal{I}} \beta\, X + \epsilon\,, \tag{2.1}$$

where α is a scalar, β and X are functions defined on the compact interval $\mathcal{I}$, and $E(\epsilon \,|\, X) = 0$. Square-bracketed subscripts here distinguish the ith ob-

servation of X, $X_{[i]}$, from the ith principal component score, which is conventionally represented by X_i. The prediction problem is that of estimating $\mu(x) = E(Y \,|\, X = x) = \alpha + \int_{\mathcal{I}} \beta\, x$ with (α, β) at (2.1), where x denotes a particular value of X and μ is a scalar functional.

A standard approach to estimating $\mu(x)$ is to introduce an orthonormal basis, say $\psi_1, \psi_2, \ldots$, and argue that β and x admit convergent expansions with respect to this sequence, i.e.

$$\beta = \sum_{j=1}^{\infty} b_j\, \psi_j\,, \quad x = \sum_{j=1}^{\infty} x_j\, \psi_j\,, \quad \mu(x) = \alpha + \sum_{j=1}^{\infty} b_j\, x_j\,, \tag{2.2}$$

where $b_j = \int_{\mathcal{I}} \beta\, \psi_j$ and $x_j = \int_{\mathcal{I}} x\, \psi_j$. Estimators $\hat\alpha$ of α and $\hat b_j$ of b_j, for $j \ge 1$, are then constructed from the data by minimising

$$S_r(\alpha, b_1, \ldots, b_r) = \sum_{i=1}^{n} \left(Y_i - \alpha - \sum_{j=1}^{r} b_j\, X_{ij} \right)^2, \tag{2.3}$$

where $X_{ij} = \int X_{[i]}\, \psi_j$ and r denotes the frequency cut-off, a smoothing parameter. These definitions of $\hat\alpha$ and $\hat b_1, \ldots, \hat b_r$ reflect the definitions of α and β at (2.1) and, for appropriate choice of r, ensure consistency. The resulting estimator of μ is

$$\widehat\mu(x) = \hat\alpha + \sum_{j=1}^{r} \hat b_j\, x_j\,. \tag{2.4}$$

A thresholding method could also be used instead of "cut-off smoothing," but the difficulty of estimating the variance of $\hat b_j$ makes that approach unattractive.

2.2. Principal component basis

It is common to take $\psi_1, \psi_2, \ldots$ to be the principal component basis, ordered so that the corresponding eigenvalues form a decreasing sequence. Specifically, define $K(s,t) = \mathrm{cov}\{X(s), X(t)\}$ to be the covariance function of X, and construct the spectral decomposition of K,

$$K(s,t) = \sum_{j=1}^{\infty} \theta_j\, \psi_j(s)\, \psi_j(t)\,, \tag{2.5}$$

where $\theta_1 \ge \theta_2 \ge \ldots \ge 0$ and (θ_j, ψ_j) are the (eigenvalue, eigenfunction) pairs of the transformation that takes ψ to $K\psi$, defined by $(K\psi)(t) = \int_{\mathcal{I}} K(s,t)\, \psi(s)\, ds$. Then the orthonormal functions ψ_j make up the principal component basis. The jth uncentred principal component score of X is $X_j = \int_{\mathcal{I}} X\, \psi_j$.

In practice the principal component basis is unknown, and needs to be estimated from data. To this end we define

$$\widehat K(s,t) = n^{-1} \sum_{i=1}^{n} \{X_{[i]}(s) - \bar X(s)\}\, \{X_{[i]}(t) - \bar X(t)\} = \sum_{j=1}^{\infty} \hat\theta_j\, \hat\psi_j(s)\, \hat\psi_j(t)\,,$$

where $\bar{X} = n^{-1} \sum_i X_{[i]}$, $\widehat{K}(s,t)$ is an estimator of $K(s,t)$, $(\hat{\theta}_j, \psi_j)$ are (eigenvalue, eigenfunction) pairs for the transformation represented by $\widehat{K}$, and the order of the indices j is chosen to ensure that $\hat{\theta}_1 \geq \hat{\theta}_2 \geq \ldots$ Then $\hat{\theta}_j$ and $\hat{\psi}_j$ are our estimators of θ_j and ψ_j, respectively, and we would replace (2.3) by

$$\hat{S}_r(\alpha, b_1, \ldots, b_r) = \sum_{i=1}^{n} \left(Y_i - \alpha - \sum_{j=1}^{r} b_j \, \widehat{X}_{ij} \right)^2 , \tag{2.6}$$

where $\widehat{X}_{ij} = \int_{\mathcal{I}} X_{[i]} \, \hat{\psi}_j$, giving the obvious estimator $\widehat{\mu}(x)$ of $\mu(x)$. Equivalently, since $\hat{\alpha} = \bar{Y} - \int_{\mathcal{I}} \hat{\beta} \, \bar{X}$ then, writing $\overline{\widehat{X}}_j = n^{-1} \sum_i \widehat{X}_{ij}$, we can minimise

$$\hat{S}_r^{\text{equiv}}(b_1, \ldots, b_r) = \sum_{i=1}^{n} \left\{ Y_i - \bar{Y} - \sum_{j=1}^{r} b_j \left(\widehat{X}_{ij} - \overline{\widehat{X}}_j \right) \right\}^2 \tag{2.7}$$

over $b_1, \ldots, b_r$, obtaining the same numerical values $\hat{b}_1, \ldots, \hat{b}_r$ as we do when minimising (2.6). Then, defining $\hat{x}_j = \int_{\mathcal{I}} x \, \hat{\psi}_j$, we take

$$\widehat{\mu}(x) = \bar{Y} + \sum_{j=1}^{r} \hat{b}_j \, (\hat{x}_j - \overline{\widehat{X}}_j) \tag{2.8}$$

to be our estimator of $\mu(x)$. In a slight abuse of notation, when discussing practical implementation we shall write $\hat{\alpha}$ and $\hat{b}_j$ for the quantities that minimise (2.6) rather than (2.3).

2.3. *Adapting to the variance of Y*

The estimator $\widehat{\mu}$ at (2.8) is conventional, but does not take into account the fact that the errors at (2.1) are often heteroscedastic. When a significant amount of variability is explained by that aspect of the problem, we should replace $\hat{S}_r(\alpha, b_1, \ldots, b_r)$ at (2.6) by its form where a weight, equal to an approximation to the inverse of the variance of $Y_i - \alpha - \int_{\mathcal{I}} \beta \, X_{[i]}$ conditional on $X_{[i]}$, is incorporated into the series at (2.6).

In conventional parametric regression, the conditional variance of the regression errors is often modelled as a function of the assumed parametric form of $E(Y|X)$. See for example Carroll and Ruppert (1988). In the functional data context we propose modeling $\text{var}(\epsilon \,|\, X)$ by

$$\sigma(X)^2 = g\left(\alpha + \sum_{j=1}^{r} b_j \, X_j \right) , \tag{2.9}$$

with g a univariate function and α, $b_1, \ldots, b_r$ and r as in (2.6) and where X_j is the jth principal component score. The appropriate choice of g depends on the data at hand, but an adaptive choice that is often suitable is the "power of

the mean" model, $g(u) = |c_1 u|^{c_2}$, where c_1 and c_2 are constants; or the version of the model which includes an intercept term. See Carroll and Ruppert (1988, pp. 5, 65).

Note that it is not necessary to have consistent estimators of the variances in order to enjoy improved statistical performance, even in the asymptotic limit. In particular, approximate parametric variance models can bring significant improvement. We shall take this point up again in section 3; see point (ii) below the Theorem there. When a reasonable parametric model cannot be formulated, the alternative is to use a nonparametric estimator of g. When sample sizes are small, it is not always possible to construct an accurate nonparametric variance estimator. However, we shall show in section 4 that such techniques can be useful.

To estimate $\sigma^2(X)$ we interpret the unweighted estimators $\hat\alpha$ and $\hat\beta = \sum_{j\le r} \hat b_j \hat\psi_j$ as pilot estimators of α and $\beta = \sum_j b_j \psi_j$, respectively, and use them to calculate residuals $\hat\epsilon_i = Y_i - \hat\alpha - \sum_j \hat b_j \widehat X_{ij}$. Since these quantities are already centred then, in a parametric context where $g \equiv g(\cdot;\boldsymbol\theta)$ and $\boldsymbol\theta$ denotes a vector of parameters, we define

$$\widehat T(\boldsymbol\theta) = \sum_{i=1}^n \left\{ \hat\epsilon_i^2 - g\Big(\hat\alpha + \sum_{j=1}^r \hat b_j \widehat X_{ij}; \boldsymbol\theta\Big)\right\}^2 \tag{2.10}$$

and choose $\hat{\boldsymbol\theta}$ to minimise $\widehat T(\boldsymbol\theta)$. In this notation our estimator of $\mathrm{var}(Y - \alpha - \int_{\mathcal I} \beta\, X \,|\, X = x)$ is, when $x = X_{[i]}$,

$$\widehat w(X_{[i]})^{-1} = g\Big(\hat\alpha + \sum_{j=1}^r \hat b_j \widehat X_{ij}; \boldsymbol\theta\Big)\,.$$

We proceed similarly in the nonparametric context, but there we estimate g by a nonparametric regression estimator applied to the data $(\hat\alpha + \sum_{j=1}^r \hat b_j \widehat X_{ij}\,, \hat\epsilon_i^2)$, for $i = 1, \ldots, n$.

Next we incorporate these weights into the objective function at (2.7), obtaining:

$$\widehat U_t(b_1, \ldots, b_t) = \sum_{i=1}^n \left\{ Y_i - \bar Y_w - \sum_{j=1}^t b_j \left(\widehat X_{ij} - \overline{\widehat X}_{j,w}\right)\right\}^2 \widehat w(X_{[i]})\,, \tag{2.11}$$

where $\bar Y_w = \{\sum_i \widehat w(X_{[i]})\}^{-1} \sum_i \widehat w(X_{[i]})\, Y_i$ and $\overline{\widehat X}_{j,w} = \{\sum_i \widehat w(X_{[i]})\}^{-1} \sum_i \widehat w(X_{[i]})\, \hat X_{ij}$; and we choose $\tilde b_{w1}, \ldots, \tilde b_{wt}$ to minimise $\widehat U_t(b_1, \ldots, b_t)$. A new estimator of $\mu(x)$ is given by the following analogue of (2.4), based on the new coefficient estimators:

$$\widetilde\mu_w(x) = \bar Y_w + \sum_{j=1}^t \tilde b_{wj} \left(\hat x_j - \overline{\widehat X}_{j,w}\right). \tag{2.12}$$

A computational advantage of defining estimators by minimising $\hat{S}_r(\alpha, b_1, \ldots, b_r)$ at (2.3), rather than $\widehat{U}_t(b_1, \ldots, b_t)$ at (2.11), is that the "ex ex transpose" matrix in the former case is simple to invert. Indeed, by definition of $\widehat{X}_{ij}$ in terms of the orthogonal functions $\hat{\psi}_j$, the matrix with (j,k)th term $n^{-1}\sum_i (\widehat{X}_{ij} - \overline{\widehat{X}}_j)(\widehat{X}_{ik} - \overline{\widehat{X}}_k)$ is diagonal. The fact that this does not hold in the case of the objective function $\widehat{U}_t(b_1, \ldots, b_t)$ reflects the fact that the orthonormal basis functions $\hat{\psi}_j$ are not necessarily, in this case, the natural ones. Instead we could replace $\widehat{U}_t(b_1, \ldots, b_t)$ by

$$\widehat{V}_s(b_1, \ldots, b_s) = \sum_{i=1}^n \left\{ Y_i - \bar{Y}_w - \sum_{j=1}^s b_j \left(\check{X}_{ij} - \overline{\check{X}}_{j,w} \right) \right\}^2 \widehat{w}(X_{[i]})\,, \tag{2.13}$$

where we define $\check{X}_{ij} = \int_{\mathcal{I}} X_{[i]}\,\hat{\phi}_j$ and $\overline{\check{X}}_{j,w} = \{\sum_i \widehat{w}(X_{[i]})\}^{-1} \sum_i \widehat{w}(X_{[i]})\check{X}_{ij}$, and where the orthonormal functions $\hat{\phi}_1, \hat{\phi}_2, \ldots$, with corresponding eigenvalues $\widehat{\omega}_1 \geq \widehat{\omega}_2 \geq \ldots$, are defined by the following spectral decomposition

$$\left\{ \sum_{i=1}^n \widehat{w}(X_{[i]}) \right\}^{-1} \sum_{i=1}^n \{X_{[i]}(s) - \bar{X}_w(s)\}\,\{X_{[i]}(t) - \bar{X}_w(t)\}\,\widehat{w}(X_{[i]}) = \sum_{j=1}^\infty \widehat{\omega}_j\,\hat{\phi}_j(s)\,\hat{\phi}_j(t).$$

Taking $\check{b}_{w1}, \ldots, \check{b}_{ws}$ to minimise $\widehat{V}_s(b_1, \ldots, b_s)$, a competitor with $\widetilde{\mu}_w(x)$ at (2.12) is given by

$$\check{\mu}_w(x) = \bar{Y}_w + \sum_{j=1}^s \check{b}_{wj} \left(\int_{\mathcal{I}} x\,\hat{\phi}_j - \overline{\check{X}}_{j,w} \right). \tag{2.14}$$

The numerical differences between $\widetilde{\mu}_w$ and $\check{\mu}_w$ are generally very small.

2.4. Practical choice of smoothing parameters

The methodology outlined in sections 2.2 and 2.3 involves two smoothing parameters: r, in the equivalent objective functions $\hat{S}_r$ and $\hat{S}_r^{\text{equiv}}$ at (2.6) and (2.7), and t, in $\widehat{U}_t$ at (2.11), or s, in $\widehat{V}_s$ at (2.13). We propose selecting these parameters by cross-validation, as follows. Omit the data pair $(X_{[i]}, Y_i)$ from the sample, and, using the remaining $n-1$ pairs, construct the predictor $\check{\mu}_w(x)$ at (2.14) for a general r and s; denote it by $\check{\mu}_{w,-i}(x \,|\, r, s)$. Put $W(r,s) = \sum_i \{Y_i - \check{\mu}_{w,-i}(X_{[i]} \,|\, r, s)\}^2$, and choose (r,s) to minimise $W(r,s)$. The same approach is used to select r and t for the predictor $\widetilde{\mu}_w(x)$ at (2.12).

3. Theoretical properties

Recall that the pair (X, Y) is generated by the model at (2.1), where the error, ϵ, has zero mean, and we wish to estimate $\mu(x) = \alpha + \int_{\mathcal{I}} \beta\, x$ for a particular

function x. Our estimator, which is equivalent to that given at (2.12) with $w = \tau(X_{[i]})^{-2}$, is defined by $\bar\mu_w(x) = \bar Y_w + \sum_{j\le r} \hat b_j\,(x_j - \bar X_{j,w})$, where $\bar Y_w = \{\sum_i \tau(X_{[i]})^{-2}\}^{-1} \sum_i \tau(X_{[i]})^{-2}\, Y_i$ and $\bar X_{j,w} = \{\sum_i \tau(X_{[i]})^{-2}\}^{-1} \sum_i \tau(X_{[i]})^{-2}\, X_{ij}$ and $\hat b_1, \hat b_2, \ldots$ are chosen to minimise

$$\sum_{i=1}^n \left\{ Y_i - \bar Y_w - \sum_{j=1}^r b_j\,(X_{ij} - \bar X_{j,w}) \right\}^2 \frac{1}{\tau(X_{[i]})^2}\,. \tag{3.1}$$

Since we centre the principal component scores X_{ij} at their respective means $\bar X_{j,w}$, which are consistent estimators of the respective jth components of $E\{X\,\tau(X)^{-2}\}$, we may, and do, assume without loss of generality that $E\{X\,\tau(X)^{-2}\} = 0$.

The eigenfunctions ψ_j and eigenvalues θ_j are defined by (2.5), and we assume of them that:

$$\text{the principal components } \textstyle\int_{\mathcal I} X\,\psi_j \text{ are independent}\,, \tag{3.2}$$

$$r = r(n) \to \infty \text{ as } n \to \infty \text{ and } r = O\big(n^{-\eta+(1/2)}\big) \text{ for some } \eta \in (0, \tfrac12)\,, \tag{3.3}$$

$$E\|X\|^k < \infty\,, \sup_{j\ge1}\, \theta_j^{-k}\, E\big(\textstyle\int_{\mathcal I} X\,\psi_j\big)^{2k} < \infty \text{ for each integer } k \ge 1\,. \tag{3.4}$$

Assumption (3.2) can be relaxed to a mixing condition.

We suppose too that we model the variance $\mathrm{var}(\epsilon\,|\,X) = \sigma(X)^2$ as $\tau(X)^2$, where the function τ is known but might not equal σ. That is, our model may not actually be correct. We make simplifying assumptions that relate to this model:

$\epsilon = \sigma(X)\,\delta$, where δ is stochastically independent of X, $E(\delta) = 0$, $E(\delta^2) = 1$, the functional σ is bounded, τ is bounded above zero, and $\sigma(X)$ and $\tau(X)$ depend on only a finite number of the principal component scores $X_j = \int_{\mathcal I} X\,\psi_j$; that is, for some $t \ge 1$ and positive integers $j_1, \ldots, j_t$ we can write $\sigma(X)^2 = \mathrm{var}(\epsilon\,|\,X) = h(X_{j_1}, \ldots, X_{j_t})$, where h is a positive, t-variate function which is bounded away from zero and infinity, and $\tau(X)^2$ can be represented in the same way; (3.5)

and either:

the empirical basis $\hat\psi_1, \hat\psi_2, \ldots$ is used to construct the predictor $\bar\mu_w$, and there exist positive constants γ_b, γ_x, γ_θ and C such that, for all j, $|b_j| \le C\,j^{-\gamma_b}$, $|x_j| \le C\,j^{-\gamma_x}$, $C^{-1}\,j^{-\gamma_\theta} \le \theta_j \le C\,j^{-\gamma_\theta}$, $\theta_j - \theta_{j+1} \ge C^{-1}\,j^{-\gamma_\theta-1}$, $\gamma_b > \gamma_\theta + 2$, $\gamma_x > \frac12$, $\gamma_\theta > 1$ and $\gamma_\theta + 1 > 2\,\gamma_x$, (3.6)

or:

the principal component basis $\psi_1, \psi_2, \ldots$ is known and is used in place of the empirical basis to construct $\bar\mu_w$, and moreover, $\sum_j \theta_j^{-1}\,x_j^2 = \infty$ and $\int_{\mathcal I} \beta^2 < \infty$. (3.7)

Condition (3.5) can be relaxed by noting that if $\sigma(\cdot)$ is sufficiently regular, and if the scores X_j are independent, then $\sigma(X)$ can be approximated by a sequence of functions $\sigma_t(X_1, \ldots, X_t)$, for $t \geq 1$, where $\sigma(X) - \sigma_t(X_1, \ldots, X_t)$ converges to zero as $t \to \infty$, with a similar constraint imposed on $\tau(X)$. If we strengthen the condition $|x_j| \leq C\, j^{-\gamma_x}$ by asking that $C^{-1}\, j^{-\gamma_x} \leq |x_j| \leq C\, j^{-\gamma_x}$ then the assumption $\gamma_\theta + 1 > 2\,\gamma_x$ in (3.6) implies that $\sum_j \theta_j^{-1}\, x_j^2 = \infty$, which is stated explicitly in (3.7). There is a sense in which (3.7) is unnecessary since it addresses a case that is of only technical interest, but it permits milder assumptions on the eigenvalue sequence θ, and the functions b and x, than does our derivation under (3.6), and moreover our proof in the case of (3.7) is so much more transparent that it sheds considerably more light on the argument leading to Theorem 3.1 than does the proof when (3.6) obtains. Indeed, the length of the proof in the case of (3.6) prevents us from giving it here; it is similar to that of Theorem 2.2 of Hall and Hosseini-Nasab (2009). A proof of Theorem 3.1 under (3.7) is given in section 5.

Write $\mathrm{AMSE}\{\bar\mu_w(x) - \mu(x)\}$ for the asymptotic mean squared error of the estimator $\bar\mu_w(x)$. The following result describes asymptotic properties of this quantity.

Theorem 3.1. *If (3.2)–(3.5), and either (3.6) or (3.7), hold then as n and r diverge together,*

$$\mathrm{AMSE}\{\bar\mu_w(x) - \mu(x)\} = n^{-1}\, \frac{E\{\sigma(X)^2\, \tau(X)^{-4}\}}{[E\{\tau(X)^{-2}\}]^2} \sum_{j=1}^{r} \theta_j^{-1}\, x_j^2 + \Bigg(\sum_{j=r+1}^{\infty} b_j\, x_j \Bigg)^2 . \tag{3.8}$$

Among the implications that can be drawn from (3.8) are the following:

(i) If the model, τ^2, for the variance, σ^2, is essentially correct, i.e. if τ equals a constant multiple of σ, then the factor, $\rho^2 \equiv E\{\sigma(Z)^2\, \tau(Z)^{-4}\}\, [E\{\tau(Z)^{-2}\}]^{-2}$, outside the first term in (3.8), which represents the variance contribution to asymptotic mean squared error, reduces to simply $[E\{\sigma(X)^{-2}\}]^{-1}$; whereas that factor would be simply $E\{\sigma(X)^2\}$ if we were to use unweighted least-squares, i.e. if we were to take $\tau(X)$ to be constant. The fact that, by Jensen's inequality, $[E\{\sigma(X)^{-2}\}]^{-1} \leq E\{\sigma(X)^2\}$, demonstrates the effectiveness of the adaptive approach.

(ii) If the model is essentially incorrect, i.e. if τ does not equal a constant multiple of σ, then the estimator remains consistent and enjoys the same convergence rate as before, but with an inflated constant multiplier. More generally, if the variance functional σ^2 is not constant, and if the model is wrong but approximately correct (in particular, if $\tau(X)$ is sufficiently close to $\sigma(X)$ for sufficiently many values of X), then ρ^2 is reduced relative to the value it would have if we were to simply take $\tau \equiv 1$.

(iii) The factor ρ^2, defined in (i) above, is applied to each and every term in the series $\sum_{j \leq r} \theta_j^{-1}\, x_j^2$ in (3.8); it does not reduce in size as j increases. Therefore

the advantages of correcting for heteroscedasticity are valid with equal force for arbitrarily large dimension; they do not relate just to low-dimensional aspects of the problem.

(iv) As is to be expected, the effect of weighting has an impact only on the variance contribution to asymptotic mean squared error, not on the bias component. However, even if the problem is finite-dimensional the impact of the variance component persists even in the asymptotic limit, and so there is always something to be gained, in asymptotic terms, by adapting the estimator appropriately to heteroscedasticity.

(v) The result $\sum_j \theta_j^{-1} x_j^2 = \infty$ in (3.7) implies that the estimator $\bar{\mu}_w(x)$ has nonparametric, rather than parametric, convergence rates. It holds if we treat x as a realisation of X, and average over all such realisations. In particular, if x is distributed like X then the random variables $\theta_j^{-1} X_j^2$ all have unit mean, and so $\sum_{j\ge1} E(\theta_j^{-1} X_j^2) = \sum_{j\ge1} 1 = \infty$, of which a modest refinement is the assumption that $\sum_{j\ge1} E\{\min(\theta_j^{-1} X_j^2, c)\} = \infty$ for some $c > 0$. (See Appendix (i) for details.) The latter property implies that the assumption $\sum_{j=1}^{\infty} \theta_j^{-1} x_j^2 = \infty$ in (3.7) holds "on average."

4. Numerical illustrations

4.1. Real data example

We applied our method to Australian rainfall data analysed by Lavery *et al.* (1992), and available at `http://dss.ucar.edu/datasets/ds482.1`. The data consist of daily rainfall measurements, observed over the years from 1840 to 1990 inclusive, at 191 Australian weather stations. Our goal is to predict the rainfall during the first week of June from the rainfall curve over the other weeks in the year.

We considered two versions of this prediction problem. In the first, for any given station we took Y to equal the rainfall during the first week of June (rainfall in June is particularly variable), averaged over the years where the station had been observed. Our predictor $X(t)$ was measured from the second week of June to the end of May, rainfall was averaged over the years where the station was observed, and $X(t)$ was computed by passing a local polynomial smoother through discrete observations. We removed one weather station which appeared to be an outlier. Then we applied our method to the $n = 190$ remaining stations.

In the second version of the prediction problem, for any given station we took Y to equal the total rainfall during the first week of June in a given year. (We used the year 1987 as not all stations were operative after 1987.) Our predictor $X(t)$ was measured from the second week of June in the previous year (i.e. 1986) to the end of May in the same year (i.e. 1987), and again $X(t)$ was computed by passing a local polynomial smoother through discrete observations. For each

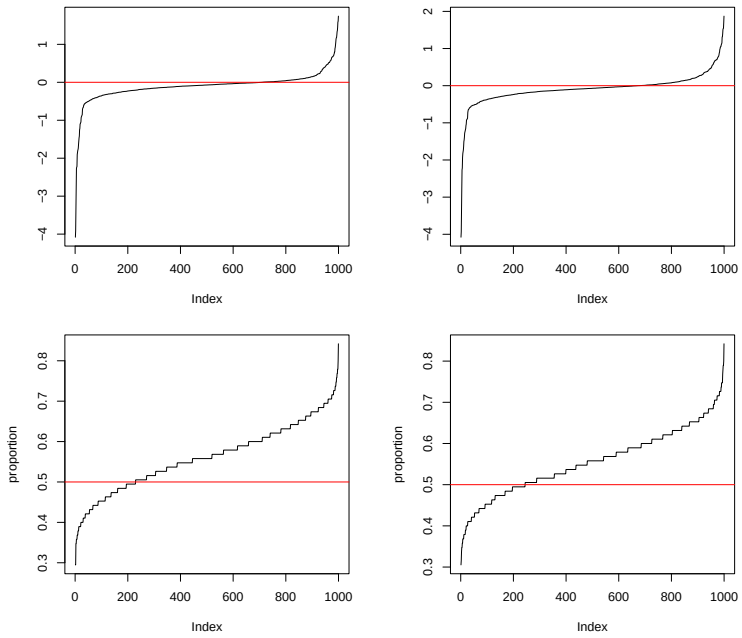

FIG 1. *Plot of the 1000 ordered values of* $\log(\widetilde{MSE}_{w,b}/MSE_b)$*, in the top left panel;* $\log(\check{MSE}_{w,b}/MSE_b)$*, top right panel;* $\tilde{p}_{w,b}$*, bottom left panel; and* $\check{p}_{w,b}$*, bottom right panel; for the first prediction problem. Horizontal lines are for reference only.*

year, some of the stations had missing values and we kept only the $n = 149$ stations with no missing observation from June 1986 to June 1987.

The majority of weather stations fall into one of two classes, which respectively comprise most stations in southern parts of the continent (these tend to follow a "European" rainfall pattern, where the majority of rain comes in cooler months and summer is relatively dry), and most stations in northern regions (these exhibit a "tropical" pattern where most rain falls in mid to late summer and the cooler months are relatively dry). Only a small number of weather stations have more complex rainfall patterns that are not of one of these two types, although some northern stations reflect southern rainfall patterns, and vice versa. These features suggest that most of the data might reasonably be assumed to come from a mixture of two populations. Those populations could produce different error variances in the linear model, leading to heteroscedasticity.

In the first prediction problem, to test our method we generated $B = 1000$ samples, each of size $n = 95$, by randomly removing half of the 190 stations. For each of the B samples we then applied our method to predict the value of Y for each of the 95 removed stations. To construct the weights we used the nonparametric variance estimator described in section 2. Note that, since these were real data, we did not know the true value of the target μ, and so we compared the predicted value with the true value of Y. For each of the B

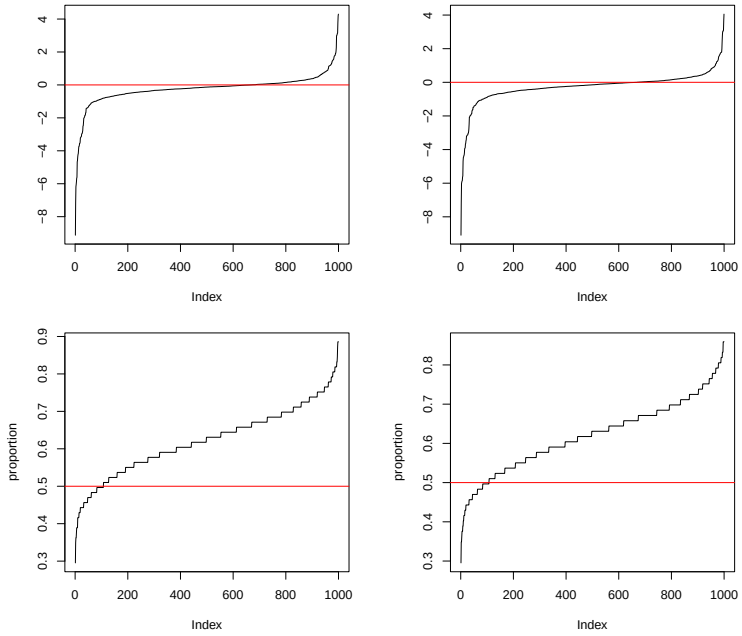

FIG 2. *Plot of the 1000 ordered values of* $\log(\widetilde{MSE}_{w,b}/MSE_b)$, *in the top left panel;* $\log(\check{MSE}_{w,b}/MSE_b)$, *top right panel;* $\tilde{p}_{w,b}$, *bottom left panel; and* $\check{p}_{w,b}$, *bottom right panel; for the second prediction problem. Horizontal lines are for reference only.*

samples we calculated the mean squared errors for the 95 predicted stations. That is, for $b = 1, \ldots, B$ we calculated

$$\widetilde{\text{MSE}}_{w,b} = \frac{1}{95}\sum_{i=1}^{95}\{\tilde{\mu}_w(X_{[i]}) - Y_i\}^2, \quad \check{\text{MSE}}_{w,b} = \frac{1}{95}\sum_{i=1}^{95}\{\check{\mu}_w(X_{[i]}) - Y_i\}^2$$

$$\text{and} \qquad \text{MSE}_b = \frac{1}{95}\sum_{i=1}^{95}\{\hat{\mu}(X_{[i]}) - Y_i\}^2.$$

For each of the B samples we also calculated the proportion of the 95 predicted stations that were better predicted by the weighted methods. In other words, for $b = 1, \ldots, B$ we computed $\tilde{p}_{w,b} = \#\{[\tilde{\mu}_w(X_{[i]}) - Y_i]^2 < [\hat{\mu}(X_{[i]}) - Y_i]^2\}/95$ and $\check{p}_{w,b} = \#\{[\check{\mu}_w(X_{[i]}) - Y_i]^2 < [\hat{\mu}(X_{[i]}) - Y_i]^2\}/95$.

In Figure 1 we present graphs of the resulting $B = 1000$ ordered values of $\log(\widetilde{\text{MSE}}_{w,b}/\text{MSE}_b)$, of $\log(\check{\text{MSE}}_{w,b}/\text{MSE}_b)$, of $\tilde{p}_{w,b}$ and of $\check{p}_{w,b}$ for the first prediction problem. We see that both weighted methods gave very similar results, and that both strongly bettered the unweighted predictor $\hat{\mu}$: for about 70% of the 1000 samples, the MSE of the weighted methods was less than that for the unweighted method. Moreover, in about 75% of cases, the proportions $\tilde{p}_{w,b}$ and $\check{p}_{w,b}$ were higher than 0.5, meaning that for a large number of the $B = 1000$ samples, more than half of the 95 predicted values were closer to the true Y_i when

using the weighted method than when the unweighted method was employed.

We proceeded similarly in the second prediction problem, each time randomly splitting the sample in two parts (of respective sizes 74 and 75). The results are shown in Figure 2. They are similar to, but favour more strongly the weighted approach, than the results in the first prediction problem: for about 70% of the 1000 samples, the MSE of the weighted methods was less than that for the unweighted method, and in about 95% of cases, the proportions $\tilde{p}_{w,b}$ and $\check{p}_{w,b}$ were higher than 0.5.

4.2. Simulations

We also tested the weighted methods on some generated data (the advantage here is that we know the target $\mu(x) = E(Y|X = x)$, and thus it is easier to assess the quality of the predictors). For $t \in [0, 365]$ we took $X(t)$ to be a smoothed version of the rainfall, averaged over the years for which the station was in operation, at time t, at each of the 190 Australian weather stations used in the first prediction problem of section 4.1. We generated 190 Y-values according to the model at (2.1), where we took $\alpha = 0$, $\beta(t) = 0.02 \cdot \sin\{8 - (\pi/20t)\}$ and $\epsilon = f(X)\,U$ where $U \sim U[-3/4, 3/4]$ and $f(X)^2 = 0.1 \cdot \{\int \beta(t)X(t)\,dt\}^2$.

We proceeded as in section 4.1 and, by randomly splitting the data $(X_{[i]}, Y_i)$ into two parts, constructed $B = 1000$ samples of size $n = 95$, and each time applied the method to the 95 remaining data points. We compared the results obtained when estimating the function g at (2.9) nonparametrically, or parametrically using a correct model or a wrong model. More precisely we tried the following three parametric models for g:

(i) $g(u) = |c_1 u|^{c_2}$ (the true g);

(ii) $g(u) = |c_1|/(3 + |u|^{c_2})$ (an approximation of the true g);

(iii) $g(u) = |c_1|\{\cos(c_2 u) + 1\}$ (a model that is very different from the true g).

In this case, since we knew the target μ we replaced Y_i by $\mu(X_{[i]})$ in the definitions of $\widetilde{\mathrm{MSE}}_{w,b}$, MSE_b, $\check{\mathrm{MSE}}_{w,b}$, $\tilde{p}_{w,b}$ and $\check{p}_{w,b}$. Figure 3 shows the results obtained by estimating g nonparametrically or using the parametric models (i), (ii) and (iii). The figure illustrates the improvement that can be gained by using a weighted version of the predictor when the parametric variance model is correct or approximately correct, or when the variance is estimated nonparametrically. When we are able to formulate the correct parametric model for the variance, the parametric variance estimator can give better results than a nonparametric estimator of g, but when we assume a parametric model for g that is very far from the true one (case (iii)), the weighted estimator can perform poorly, and is strongly outperformed by a nonparametric estimator of g.

5. Proof of Theorem

We give a proof when (3.7) holds. We consider first the homoscedastic case, where both σ and τ equal constants, and then we generalise our argument

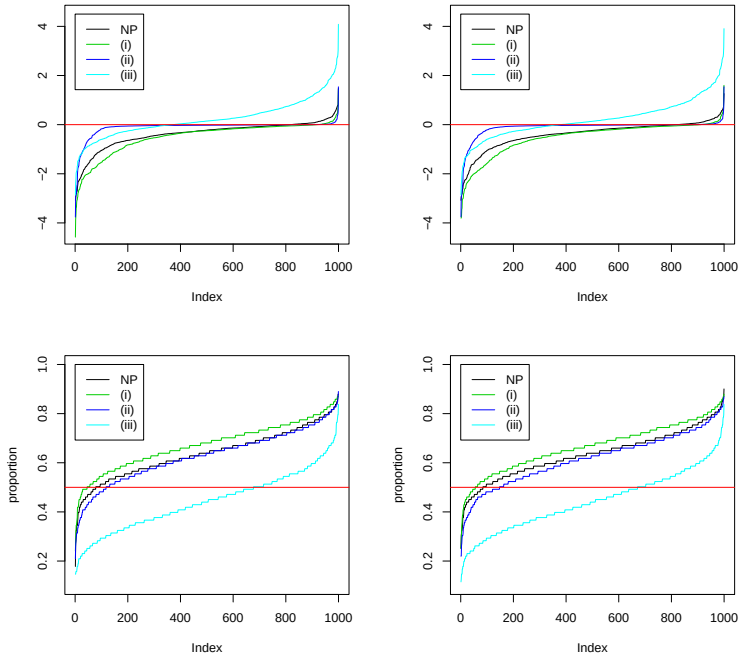

FIG 3. *Plot of the 1000 ordered values of* $\log(\widetilde{MSE}_{w,b}/MSE_b)$, *in the top left panel;* $\log(\check{MSE}_{w,b}/MSE_b)$, *top right panel;* $\tilde{p}_{w,b}$, *bottom left panel; and* $\check{p}_{w,b}$, *bottom right panel; for the generated data, estimating* g *with a nonparametric estimator (NP) or assuming the parametric models* (i), (ii) *or* (iii). *Horizontal lines are for reference only.*

to the heteroscedastic setting. The model (2.1) can be written equivalently as $Y = \alpha + \int_{\mathcal{I}} \beta\,(X - EX) + \epsilon$, for the same function β but for a different scalar α, which now equals $E(Y)$. We shall work with this model below. The least-squares estimator of $\mu(x)$ is the same as before, but the corresponding estimator of α is now simply $\hat{\alpha} = \bar{Y}$. In particular, using the new model and making assumptions (3.7) and (3.4), $\hat{\alpha}$ is root-n consistent for α:

$$\hat{\alpha} - \alpha = O_p\big(n^{-1/2}\big)\,. \tag{5.1}$$

The least-squares estimators $\hat{b}_1, \ldots, \hat{b}_r$ are the solutions of

$$\hat{S}\,(\hat{b}_1, \ldots, \hat{b}_r)^{\mathrm{T}} = \hat{s}\,, \tag{5.2}$$

where $\hat{S} = (\hat{s}_{j_1 j_2})$ is an $r \times r$ matrix, $\hat{s} = (\hat{s}_1, \ldots, \hat{s}_r)^{\mathrm{T}}$ is an r-vector,

$$\hat{s}_{j_1 j_2} = \frac{1}{n} \sum_{i=1}^{n} (X_{ij_1} - \bar{X}_{j_1})\,(X_{ij_2} - \bar{X}_{j_2})\,, \quad \hat{s}_j = \frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)\,(Y_i - \bar{Y})\,, \tag{5.3}$$

$X_{ij} = \int X_{[i]}\,\psi_j$, $\bar{X}_j = n^{-1} \sum_i X_{ij}$ and $\bar{Y} = n^{-1} \sum_i Y_i$. Without loss of generality, each $E(X_{ij}) = 0$. Put $Z_{ij} = X_{ij}\,\theta_j^{-1/2}$. Then the variables Z_{ij} have

zero mean and unit variance, and $Z_{i_1 j_1}$ and $Z_{i_2 j_2}$ are independent for arbitrary i_1, i_2 and for $j_1 \neq j_2$ (see (3.2)). In this notation, $\hat{s}_{j_1 j_2} = (\theta_{j_1}\,\theta_{j_2})^{1/2}\,\hat{t}_{j_1 j_2}$ and $\hat{s}_j = \theta_j^{1/2}\,\hat{t}_j$, where

$$\begin{aligned}
\hat{t}_{j_1 j_2} &= \frac{1}{n}\sum_{i=1}^{n} (Z_{ij_1} - \bar{Z}_{j_1})\,(Z_{ij_2} - \bar{Z}_{j_2})\,, \\
\hat{t}_j &= \frac{1}{n}\sum_{i=1}^{n} (Z_{ij} - \bar{Z}_j)\,(Y_i - \bar{Y}) \\
&= \frac{1}{n}\sum_{i=1}^{n} (Z_{ij} - \bar{Z}_j)\left\{\int_{\mathcal{I}} \beta\,(X_{[i]} - \bar{X}) + \epsilon_i - \bar{\epsilon}\right\} \\
&= \sum_{k=1}^{\infty} b_k\,\theta_k^{1/2}\,\hat{t}_{jk} + \hat{u}_j\,, \\
\hat{u}_j &= \frac{1}{n}\sum_{i=1}^{n} (Z_{ij} - \bar{Z}_j)\,(\epsilon_i - \bar{\epsilon})\,,
\end{aligned}$$

$\bar{Z}_j = n^{-1}\sum_i Z_{ij}$ and $\bar{\epsilon} = n^{-1}\sum_i \epsilon_i$. Define too $\hat{v}_j = \sum_{k\geq r+1} b_k\,\theta_k^{1/2}\,\hat{t}_{jk}$, and put $\widehat{T} = (\hat{t}_{j_1 j_2})$ and $D = \mathrm{diag}(\theta_1^{1/2}, \ldots, \theta_r^{1/2})$, denoting $r \times r$ matrices, and $\hat{u} = (\hat{u}_1, \ldots, \hat{u}_r)^{\mathrm{T}}$, $\hat{v} = (\hat{v}_1, \ldots, \hat{v}_r)^{\mathrm{T}}$, $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_r)^{\mathrm{T}}$ and $b = (b_1, \ldots, b_r)^{\mathrm{T}}$, representing $r \times 1$ vectors. In this notation, (5.2) is equivalent to $\widehat{T}\,D\,(\hat{b} - b) = \hat{u} + \hat{v}$; see Appendix (ii). Define $\|A\|^2 = \sum_{j_1}\sum_{j_2} a_{j_1 j_2}^2$. We shall show shortly that $\|A\| \to 0$ in probability as $n \to \infty$; see the paragraph containing (5.7). Therefore the probability that $\widehat{T} = I + A$ is invertible converges to 1. When $\widehat{T}$ is invertible,

$$\bar{\mu}_w(x) - \mu(x) - (\hat{\alpha} - \alpha) + \sum_{j=r+1}^{\infty} b_j\,x_j = \sum_{j=1}^{r} (\hat{b}_j - b_j)\,x_j = \sum_{j=1}^{r} \left\{(\widehat{T}\,D)^{-1}(\hat{u} + \hat{v})\right\}_j x_j\,. \tag{5.4}$$

Write $\hat{t}_{j_1 j_2} = \delta_{j_1 j_2} + a_{j_1 j_2}$, where $\delta_{j_1 j_2}$ denotes the Kronecker delta and $A = (a_{j_1 j_2})$ is an $r \times r$ random matrix. Assuming that $\nu \equiv \|A\| \to 0$ in probability,

$$\begin{aligned}
\widehat{T}^{-1} &= (I + A)^{-1} \\
&= I - A + \ldots + (-1)^{k-1}\,A^{k-1} + (-1)^k\,A^k\,\left(I - A + A^2 - \ldots\right) \\
&= I - A + \ldots + (-1)^{k-1}\,A^{k-1} + A^k\,A_k\,,
\end{aligned}$$

where $A_k = (a_{k, j_1 j_2})$ denotes a random matrix and $\|A_k\| \leq (1 - \nu)^{-1}$. Therefore, writing $A^k = (a_{j_1 j_2}^{(k)})$, and defining $c_r = \sum_{j\leq r} x_j^2\,\theta_j^{-1}$, we have,

$$\begin{aligned}
\left(\sum_{j=1}^{r}\Big[\;\; D^{-1} \quad \left\{\widehat{T}^{-1} - I + A - \ldots + (-1)^k\,A^{k-1}\right\}(\hat{u} + \hat{v})\Big]_j\,x_j\right)^2 \\
= \quad \left[\sum_{j=1}^{r}\left\{D^{-1}\,A^k\,A_k\,(\hat{u} + \hat{v})\right\}_j x_j\right]^2
\end{aligned}$$

$$\leq \quad c_r \sum_{j=1}^{r} \left[\left\{ A^k \, A_k \, (\hat{u} + \hat{v}) \right\}_j \right]^2 \tag{5.5}$$

$$= \quad c_r \sum_{j=1}^{r} \left\{ \sum_{j_1=1}^{r} \sum_{j_2=1}^{r} a_{jj_1}^{(k)} \, a_{k,j_1 j_2} \, (\hat{u} + \hat{v})_{j_2} \right\}^2$$

$$\leq \quad c_r \left\{ \sum_{j=1}^{r} \sum_{j_1=1}^{r} \left(a_{jj_1}^{(k)} \right)^2 \right\} \sum_{j_1=1}^{r} \left\{ \sum_{j_2=1}^{r} a_{k,j_1 j_2} \, (\hat{u} + \hat{v})_{j_2} \right\}^2$$

$$\leq \quad c_r \left\| A^k \right\|^2 \|A_k\|^2 \sum_{j=1}^{r} \{ (\hat{u} + \hat{v})_j \}^2 . \tag{5.6}$$

Assumptions (3.2) and (3.4) imply that, for each integer $\ell \geq 1$,

$$\sup_{1 \leq j_1, j_2 \leq r} E\big(a_{j_1 j_2}^{2\ell} \big) = O\big(n^{-\ell} \big) . \tag{5.7}$$

Therefore, $E(\|A^k\|^2) = O\{(r^2/n)^k\}$, and so, since (3.3) implies that $r/n^{1/2} \to 0$, we have $\nu \to 0$ in probability. The property $r/n^{1/2} \to 0$ also entails $\|A_k\| = O_p(1)$. Furthermore, $E(\hat{u}_j^2) = O(n^{-1})$ uniformly in $j \geq 1$ (see Appendix (iii)), and if $1 \leq j \leq r$ then $|\hat{v}_j| = |\sum_{k \geq r+1} b_k \, \theta_k^{1/2} \, a_{jk}|$. The latter result, (3.4), (5.7) and the fact that $(\sum_j |b_j| \, \theta_j^{1/2})^2 \leq (\sum_j b_j^2) \, (\sum_j \theta_j) < \infty$ imply that $E(\hat{v}_j^2) = O(n^{-1})$, uniformly in $1 \leq j \leq r$. (Note that, in view of the second part of (3.7), $\sum_j b_j^2 < \infty$, and by (3.4), $E\|X\|^2 = \sum_j \theta_j < \infty$.)

Combining these results we deduce that the right hand side of (5.6) equals

$$O_p\Big\{ c_r \left(r^2/n \right)^k r \, n^{-1} \Big\} = O_p\big(c_r \, r^{2k+1} \, n^{-(k+1)} \big) .$$

Hence, by (5.4) and (5.6),

$$\bar{\mu}_w(x) - \mu(x) - (\hat{\alpha} - \alpha) \tag{5.8}$$

$$= \sum_{j=1}^{r} \theta_j^{-1/2} \, x_j \left[\left\{ I - A + \ldots + (-1)^{k-1} \, A^{k-1} \right\} (\hat{u} + \hat{v}) \right]_j$$

$$- \sum_{j=r+1}^{\infty} b_j \, x_j + O_p\big(c_r^{1/2} \, r^{k+(1/2)} \, n^{-(k+1)/2} \big) . \tag{5.9}$$

Using the fact that $r^2/n \to 0$ it can be shown by direct calculation that, for each integer $k \geq 1$,

$$E\left[\sum_{j=1}^{r} \theta_j^{-1/2} \, x_j \left\{ A^k \, (\hat{u} + \hat{v}) \right\}_j \right]^2 = o\big(c_r \, n^{-1} \big) . \tag{5.10}$$

Taking k arbitrarily large, and using (5.9), (5.10) and the fact that $r = O(n^{-\eta + (1/2)})$ for some $\eta > 0$ (see (3.3)), we deduce that,

$$\bar{\mu}_w(x) - \mu(x) - (\hat{\alpha} - \alpha) = V - \sum_{j=r+1}^{\infty} b_j \, x_j + o\big(c_r^{1/2} \, n^{-1/2} \big) , \tag{5.11}$$

where $V = \sum_{j \le r} \theta_j^{-1/2}\, x_j\, (\hat{u} + \hat{v})_j$.

Note too that, since we are addressing the homoscedastic case,

$$E(V^2) = \sum_{j_1=1}^{r} \sum_{j_2=1}^{r} (\theta_{j_1}\, \theta_{j_2})^{-1/2}\, x_{j_1}\, x_{j_2}\, \{\sigma^2\, E(\hat{t}_{j_1 j_2}) + E(\hat{v}_{j_1}\, \hat{v}_{j_2})\}\,. \tag{5.12}$$

Now, $E(\hat{t}_{j_1 j_2}) = n^{-1}\,(1 - n^{-1})\,\delta_{j_1 j_2}$ and, recalling that each $E(Z_{ij}) = 0$,

$$\begin{aligned} n\, E(\hat{t}_{j_1 k_1}\, \hat{t}_{j_2 k_2}) &= E\Big\{(Z_{1 j_1} - \bar{Z}_{j_1})\,(Z_{1 k_1} - \bar{Z}_{k_1})\,(Z_{1 j_2} - \bar{Z}_{j_2})\,(Z_{1 k_2} - \bar{Z}_{k_2})\Big\} \\ &= \big(1 - n^{-1}\big)^2\, E(Z_{1 j_1}\, Z_{1 k_1}\, Z_{1 j_2}\, Z_{1 k_2}) \\ &= \big(1 - n^{-1}\big)^2\, \delta_{j_1 j_2}\, \delta_{k_1 k_2}\,, \end{aligned}$$

using the properties $j_1, j_2 \le r$ and $k_1, k_2 \ge r + 1$, and the fact that the Z_{ij}s are independent. Hence,

$$\begin{aligned} E(\hat{v}_{j_1}\, \hat{v}_{j_2}) &= \sum_{k_1=r+1}^{\infty} \sum_{k_2=r+1}^{\infty} b_{k_1}\, b_{k_2}\, (\theta_{k_1}\, \theta_{k_1})^{1/2}\, E(\hat{t}_{j_1 k_1}\, \hat{t}_{j_2 k_2}) \\ &= n^{-1}\, \big(1 - n^{-1}\big)^2\, \delta_{j_1 j_2}\, d_r\,, \end{aligned} \tag{5.13}$$

where $d_r = \sum_{k \ge r+1} b_k^2\, \theta_k$. Using (5.12), (5.13) and the fact that $d_r \to 0$ as $r \to \infty$, we deduce that, as r and n diverge together,

$$E(V^2) = \frac{1}{n} \sum_{j=1}^{r} \theta_j^{-1}\, x_j^2\, \{(1 - n^{-1})\, \sigma^2 + (1 - n^{-1})^2\, d_r\} \sim \sigma^2\, c_r\, n^{-1}\,. \tag{5.14}$$

In view of the first part of (3.7), $c_r \to \infty$ as $r \to \infty$. Formula (3.8), but with $[E\{\sigma(X)^{-2}\}]^{-1}$ replaced by σ^2, follows from this property, (5.1), (5.11) and (5.14).

Next we outline the argument that extends this result to the heteroscedastic setting. First we discuss a version of the theorem in an artificial problem where the error variance is a function of Z, say, which is independent of (X, Y) but is observed along with that pair. That is, the model (2.1) now has the form $Y = \alpha + \int_{\mathcal{I}} \beta\, X + \sigma(Z)\, \delta$, where the perturbation δ is independent of X and Z and has zero mean and unit variance. The appropriately weighted criterion function is that at (3.1) but with $\tau(X_{[i]})$ replaced by $\tau(Z_i)$. In this case the proof above is easily re-worked, in particular with the factor $\tau(Z_i)^{-2}$ included in both series at (5.3) and in subsequent series, to show that the asymptotic mean squared error of $\bar{\mu}_w(x)$ continues to be given by (3.8) but with $E\{\sigma(X)^2\, \tau(X)^{-4}\}\, [E\{\tau(X)^{-2}\}]^{-2}$ replaced by $E\{\sigma(Z)^2\, \tau(Z)^{-4}\}\, [E\{\tau(Z)^{-2}\}]^{-2}$. To appreciate the origins of this result, note that in the simpler model where β vanishes and $Y = \alpha + \sigma(Z)\, \delta$, the variance of the weighted least-squares estimator of α is exactly $\rho(n)^2 \equiv E[\{\sum_i \sigma(Z_i)^2\, \tau(Z_i)^{-4}\}\, \{\sum_i \tau(Z_i)^{-2}\}^{-2}]$; and, under the assumption in (3.5) that $\sigma(z)$ is bounded and $\tau(z)$ is bounded away from zero,

$$\rho(n)^2 \sim n^{-1}\, E\{\sigma(Z)^2\, \tau(Z)^{-4}\}\, [E\{\tau(Z)^{-2}\}]^{-2}.$$

This result continues to hold when σ and τ are functions of X rather than Z, and depend on only a finite number of principal component scores. The proof proceeds by noting first that if $\mathrm{var}(\epsilon \,|\, X) = h(X_{j_1}, \ldots, X_{j_t})$, as in (3.5), and if we assume that the components with indices $j_1, \ldots, j_t$ are known and therefore do not need to be estimated, then we are in exactly the position addressed in the previous paragraph; we can take Z to be $(X_{j_1}, \ldots, X_{j_t})$ and replace X by the expansion $\sum_j' X_j \, \psi_j$ where the summation $\sum_j'$ is over only those indices j not included among $j_1, \ldots, j_t$. Moreover, the asymptotic mean squared error formula is unaffected if we eliminate the components corresponding to $j = j_1, \ldots, j_t$, or if we take those components to be known.

Appendix

Appendix (i)

Here we prove that if it holds that

$$\text{(i) } E\{\sum_j \min(\theta_j^{-1} \, X_j^2, c)\} = \infty \text{ for some } c > 0,$$

then

$$\text{(ii) } \sum_j \theta_j^{-1} \, X_j^2 = \infty \text{ with probability one.}$$

Suppose that (i) holds, and note that, by Kolmogorov's three-series theorem, and since we assumed (see (3.2)) that the principal components X_j are independent, then $\sum_j \theta_j^{-1} \, X_j^2 < \infty$ if and only if there exists $c > 0$ such that (a) $P(U_j > c) < \infty$, (b) $\sum_j E(V_j) < \infty$ and (c) $\sum_j \mathrm{var}(V_j) < \infty$, where $U_j = \theta_j^{-1} \, X_j^2$ and $V_j = U_j$ if $U_j \le c$ and $V_j = 0$ otherwise. From this result, and the fact that, by the zero-one law, $\sum_j U_j$ either converges almost everywhere or diverges almost everywhere, we see that if (ii) fails then (a), (b) and (c) must all be true for some $c > 0$. Now, (a) and (b) together imply that $\sum_j E\{\min(U_j, c)\} < \infty$, from which it follows directly that $\sum_j E\{\min(U_j, C)\} < \infty$ for all $C \in (0, c]$. Moreover, $\sum_j E\{\min(U_j, C)\} < \infty$ also holds for $C > c$, since in that case, $E\{\min(U_j, C)\} \le \max(C, 1) \, E\{\min(U_j, c)\}$.
Therefore, if (ii) fails then $E\{\sum_j \min(\theta_j^{-1} \, X_j^2, c)\} < \infty$ for all $c > 0$. Consequently, $E\{\sum_j \min(\theta_j^{-1} \, X_j^2, c)\} = \infty$ for some $c > 0$ (or equivalently, (i)) implies (ii).

Appendix (ii)

Here we prove that $\widehat{T} \, D \, (\hat{b} - b) = \hat{u} + \hat{v}$. Observe from the definitions of D, $\hat{S}$ and $\widehat{T}$ that $\widehat{T} \, D = D^{-1} \, D \, \widehat{T} \, D = D^{-1} \, \hat{S}$. Since, by (5.2), $\hat{S} \, \hat{b} = \hat{s}$, then $\widehat{T} \, D \, \hat{b} = D^{-1} \, \hat{S} \, \hat{b} = D^{-1} \, \hat{s}$, and therefore

$$\left(\widehat{T} \, D \, \hat{b}\right)_j = \theta_j^{-1/2} \, \hat{s}_j \, . \tag{A.1}$$

Using the relation $Z_{ij} = X_{ij}\,\theta_j^{-1/2}$ we deduce that

$$\theta_j^{-1/2}\,\hat{s}_j = \frac{1}{n}\sum_{i=1}^{n}(Z_{ij}-\bar{Z}_j)\,(Y_i-\bar{Y}) = \hat{t}_j = \sum_{k=1}^{\infty} b_k\,\theta_k^{1/2}\,\hat{t}_{jk} + \hat{u}_j\,. \tag{A.2}$$

(The final identity in this string of identities was derived in section 5.) Combining (A.1) and (A.2) we deduce that

$$\bigl(\widehat{T}\,D\,\hat{b}\bigr)_j = \sum_{k=1}^{\infty} b_k\,\theta_k^{1/2}\,\hat{t}_{jk} + \hat{u}_j\,. \tag{A.3}$$

Additionally, $(\widehat{T}\,D\,b)_j = \sum_{1\le k\le r}\hat{t}_{jk}\,\theta_k^{1/2}\,b_k$. Subtracting this formula from (A.3) we deduce that

$$\bigl(\widehat{T}\,D\,(\hat{b}-b)\bigr)_j = \sum_{k=r+1}^{\infty} b_k\,\theta_k^{1/2}\,\hat{t}_{jk} + \hat{u}_j = \hat{v}_j + \hat{u}_j\,,$$

which is the desired result.

Appendix (iii)

It can be proved from the definition of $\hat{u}_j$, and the fact that $E(\delta^2) = 1$, that

$$\tfrac{1}{2}\,n\,\theta_j\,E(\hat{u}_j^2) \le E\bigl\{(X_{1j}-\bar{X}_j)^2\,\sigma(X_{[1]})^2\bigr\} + \frac{1}{n^2}\sum_{i=1}^{n} E\bigl\{(X_{1j}-\bar{X}_j)^2\,\sigma(X_{[i]})^2\bigr\}\,. \tag{A.4}$$

The fact that the function h, in assumption (3.5), is bounded means that σ is also bounded, so we may assume that $(0<)\,\sigma \le B_1$, say. Therefore, using (A.4),

$$n\,E\bigl(\hat{u}_j^2\bigr) \le 4\,\theta_j^{-1}\,B_1^2\,E\bigl\{(X_{1j}-\bar{X}_j)^2\bigr\}\,. \tag{A.5}$$

Property (3.4) implies that, for some $B_2 > 0$, $E\{(X_{1j}-\bar{X}_j)^2\} \le B_2\,\theta_j$ for all j, and hence, by (A.5), $n\,E(\hat{u}_j^2) \le 4\,B_1^2\,B_2$, as had to be shown.

Acknowledgements

We thank a referee and an associate editor for their helpful comments which led to an improved version of the manuscript. The Australian weather data we used in the paper were assembled by the Australian Bureau of Meteorology. They are available from the Research Data Archive, maintained by the Computational and Information Systems Laboratory at the National Center for Atmospheric Research (NCAR). NCAR is sponsored by the National Science Foundation. Bob Dattore is greatly acknowledged for providing the data.

References

Besse, P. and Ramsay J.O. (1986). Principal components-analysis of sampled functions. *Psychometrika* **51**, 285–311.

Boente, G. and Fraiman, R. (2000). Kernel-based functional principal components. *Statist. Probab. Lett.* **48**, 335–345.

Brumback, B.A. and Rice, J.A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Amer. Statist. Assoc.* **93**, 961–976.

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

Cardot, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *J. Nonparam. Statist.* **12**, 503–538.

Cardot, H., Crambes, C. Kneip, A. and Sarda, P. (2006). Smoothing splines estimators in functional linear regression with errors-in-variables. *Comput. Statist. Data Anal.* **51**, 4832–4848.

Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Statist. Probab. Lett.* **45**, 11–22.

Cardot, H., Ferraty, F. and Sarda, P. (2000). Étude asymptotique d'un estimateur spline hybride pour le modèle linéaire fonctionnel. *C.R. Acad. Sci. Paris* Sér. I **330**, 501–504.

Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13**, 571–591.

Cardot, H. and Sarda, P. (2006). Linear regression models for functional data. In: *The Art of Semiparametrics* Eds. S. Sperlich, W. Härdle and G. Aydinli, pp. 49–66. Springer, Berlin.

Carroll, R. and Ruppert, D. (1988). *Transformation and weighting in regression.* Chapman and Hall, New York.

Cuevas, A., Febrero, M. and Fraiman, R. (2002). Linear functional regression: the case of fixed design and functional response. *Canad. J. Statist.* **30**, 285–300.

Faraway, J.J. (1997). Regression analysis for a functional response. *Technometrics* **39**, 254–261.

Ferraty, F. and Vieu, P. (2000). Fractal dimensionality and regression estimation in semi-normed vectorial spaces. *C.R. Acad. Sci. Paris* Sér. I **330**, 139–142.

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

Girard, S. (2000). A nonlinear PCA based on manifold approximation. *Comput. Statist.* **15**, 145–167.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc.* Ser. B **68**, 109–126.

Hall, P. and Hosseini-Nasab, M. (2009). Theory for high-order bounds in functional principal components analysis. *Math. Proc. Camb. Phil. Soc.* **146**, 225–256.

He, G.Z., Müller, H.-G. and Wang, J.-L. (2003). Functional canonical

analysis for square integrable stochastic processes. *J. Multivar. Anal.* **85**, 54–77.

Huang, J.Z., Shen, H. and Buja, A. (2008). Functional principal components analysis via penalized rank one approximation. *Electron. J. Stat.* **2**, 678–695.

James, G.M., Hastie, T.J. and Sugar, C.A. (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587–602.

Jank, W. and Shmueli, G. (2006). Functional data analysis in electronic commerce research. *Statist. Sci.* **21**, 155–166.

Lavery, B., Kariko, A. and Nicholls, N. (1992). A historical rainfall data set for Australia. *Aust. Met. Mag.*, **40**, 33–39.

Mas, A. (2008). Local functional principal component analysis. *Complex Anal. Oper. Theory* **2**, 135–167.

Ocaña, F.A., Aguilera, A.M. and Escabias, M. (2007). Computational considerations in functional principal component analysis. *Comput. Statist.* **22**, 449–465.

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. and Silverman, B.W. (2005). *Functional Data Analysis.*, 2nd Edn. Springer, New York.

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

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.

Silverman, B.W. (1995). Incorporating parametric effects into functional principal components analysis. *J. Roy. Statist. Soc.* Ser. B **57**, 673–689.

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

Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873–2903.