Biometrika (2016), **103**, 1, *pp.* 147–159
Printed in Great Britain
doi: 10.1093/biomet/asv062
Advance Access publication 19 January 2016

Partially functional linear regression in high dimensions

By DEHAN KONG

Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.
kongdehanstat@gmail.com

KAIJIE XUE, FANG YAO

Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5S 3G3, Canada
kaijie@utstat.toronto.edu fyao@utstat.toronto.edu

and HAO H. ZHANG

Department of Mathematics, University of Arizona, Tucson, Arizona 85721, U.S.A.
hzhang@math.arizona.edu

Summary

In modern experiments, functional and nonfunctional data are often encountered simultaneously when observations are sampled from random processes and high-dimensional scalar covariates. It is difficult to apply existing methods for model selection and estimation. We propose a new class of partially functional linear models to characterize the regression between a scalar response and covariates of both functional and scalar types. The new approach provides a unified and flexible framework that simultaneously takes into account multiple functional and ultrahigh-dimensional scalar predictors, enables us to identify important features, and offers improved interpretability of the estimators. The underlying processes of the functional predictors are considered to be infinite-dimensional, and one of our contributions is to characterize the effects of regularization on the resulting estimators. We establish the consistency and oracle properties of the proposed method under mild conditions, demonstrate its performance with simulation studies, and illustrate its application using air pollution data.

Some key words: Functional data; Functional linear regression; Model selection; Principal components; Regularization; Smoothly clipped absolute deviation.

1. Introduction

Functional linear regression is widely used to model the prediction of a functional predictor through a linear operator, often realized via an integral form of a regression parameter function; see Ramsay & Dalzell (1991), Cuevas et al. (2002), Cardot et al. (2003), Ramsay & Silverman (2005) and Yao et al. (2005a). To capture the regression relation between the response and a functional predictor, regularization is necessary. One common approach is functional principal component analysis, which has been studied by Rice & Silverman (1991), Yao et al. (2005b), Cai & Hall (2006), Hall et al. (2006), Hall & Horowitz (2007) and Zhang & Chen (2007), among others. Functional linear models have been extended to generalized functional linear models (Escabias et al., 2004; Cardot & Sarda, 2005; Müller & Stadtmüller, 2005), varying-coefficient

models (Fan & Zhang, 2000; Fan et al., 2003), wavelet-based functional models (Morris et al., 2003), functional additive models (Müller & Yao, 2008) and quadratic models (Yao & Müller, 2010).

Classical functional linear regression is designed to describe the relation between a real-valued response and one functional explanatory variable. However, in many real-world problems, it is common to also collect information on a large number of nonfunctional predictors. How to incorporate scalar predictors into functional linear regression and perform model selection or regularization is an important issue. For a standard linear regression with scalar covariates only, various penalization procedures have been proposed and studied, including the lasso (Tibshirani, 1996), the smoothly clipped absolute deviation (Fan & Li, 2001) and the adaptive lasso (Zou, 2006).

In this work, we develop a class of partially functional linear regression models to handle multiple functional and nonfunctional predictors and automatically identify important risk factors by suitable regularization. Shin (2009) and Lu et al. (2014) considered similar partially functional linear and quantile models, respectively, but did not deal with variable selection or with multiple functional predictors and high-dimensional scalar covariates. We propose a unified framework that combines the regularization of each functional predictor as a whole with a penalty on high-dimensional scalar covariates. Due to the differences between the functional and scalar predictors, we use two regularizing operations. Shrinkage penalties are imposed on the effects of both functional predictors and scalar covariates to achieve model selection and enhance interpretability, while a data-adaptive truncation that plays the role of a tuning parameter is applied to the functional predictors. We treat the functional predictors as infinite-dimensional processes; this distinguishes our approach from methods that fix the number of principal components (e.g., Li et al., 2010). A main contribution of our work is to quantify the theoretical impact of functional principal component estimation with diverging truncation, especially when the number of scalar covariates is permitted to diverge at an exponential order of the sample size.

2. REGULARIZED PARTIALLY FUNCTIONAL LINEAR REGRESSION

2·1. *Classical functional linear model via principal components*

Let $X(\cdot)$ be a square-integrable random function defined on a closed interval T of the real line with continuous mean and covariance functions, denoted by $E\{X(t)\} = \mu(t)$ and $\mathrm{cov}\{X(s), X(t)\} = K(s,t)$, respectively. The classical functional linear model is

$$Y = \mu_Y + \int_T \{X(t) - \mu(t)\}\beta(t)\,\mathrm{d}t + \epsilon, \tag{1}$$

where the regression parameter function $\beta(\cdot)$ is assumed to be square-integrable and ϵ is a random error independent of $X(t)$. Mercer's theorem implies that there exists a complete orthonormal basis $\{\phi_k\}$ in $L_2(T)$ and a nonincreasing sequence of nonnegative eigenvalues $\{w_k\}$ such that $K(s,t) = \sum_{k=1}^{\infty} w_k\phi_k(s)\phi_k(t)$ with $\sum_{k=1}^{\infty} w_k < \infty$. We further assume that $w_1 > w_2 > \cdots \geqslant 0$. Let $\{(y_i, x_i) : i = 1, \ldots, n\}$ be independent and identically distributed observations from (Y, X). The Karhunen–Loève expansion $x_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t)$ forms the foundation of functional principal component analysis, where the coefficients $\xi_{ik} = \int_T \{x_i(t) - \mu(t)\}\phi_k(t)\,\mathrm{d}t$ are uncorrelated random variables with mean zero and variances $E(\xi_{ik}^2) = w_k$, also called the functional principal component scores. Expanded on the orthonormal eigenbasis $\{\phi_k\}$, the regression function becomes $\beta(t) = \sum_{k=1}^{\infty} b_k\phi_k(t)$, and the functional linear model (1) can be written as $y_i = \mu_Y + \sum_{k=1}^{\infty} b_k\xi_{ik} + \epsilon_i$. The basis with respect to which the regression parameter b is expanded is determined by the covariance function K. This is not unnatural since $\{\phi_k\}$ is

the unique canonical basis leading to a generalized Fourier series which gives the most rapidly convergent representation of X in the L^2 sense.

2·2. *Partially functional linear regression with regularization*

We now consider functional linear regression with multiple functional and scalar predictors. Suppose that the data are $\{Y, X(\cdot), Z\}$, where Y is a scalar continuous response, $X(\cdot) = \{X_j(\cdot) : j = 1, \ldots, d\}$ are d functional predictors, and $Z = (Z_1, \ldots, Z_{p_n})^{\mathrm{T}}$ is a p_n-dimensional vector of scalar covariates. This is motivated by commonly encountered situations where both functional and nonfunctional predictors may affect the response. We assume that the number of functional predictors d is fixed, while the number of scalar covariates p_n may grow with the sample size. Specifically, we allow p_n to be ultrahigh-dimensional, such that $\log p_n = O(n^{\alpha})$ for some $\alpha > 0$. Without loss of generality, we assume that the response Y, the functional predictors $\{X_j : j = 1, \ldots, d\}$ and the scalar covariates $\{Z_l : l = 1, \ldots, p_n\}$ have been centred to have mean zero. We then model the linear relationship between Y and (X, Z) by

$$Y = \sum_{j=1}^{d} \int_T X_j(t)\beta_j(t)\,\mathrm{d}t + Z^{\mathrm{T}}\gamma + \epsilon, \tag{2}$$

where $\{\beta_j(\cdot) : j = 1, \ldots, d\}$ are square-integrable regression parameter functions, $\gamma = (\gamma_1, \ldots, \gamma_{p_n})^{\mathrm{T}}$ contains the regression coefficients of nonfunctional covariates, and ϵ is the random error, which is independent of $\{X_j(\cdot) : j = 1, \ldots, d\}$ and Z with $E(\epsilon) = 0$ and $\mathrm{var}(\epsilon) = \sigma^2$. For convenience, assume that the first q_n scalar covariates are significant while the rest are not. In other words, the true values of the regression coefficients, γ_0^{T}, are equal to $(\gamma_0^{(1)\mathrm{T}}, \gamma_0^{(2)\mathrm{T}})$, where $\gamma_0^{(1)}$ is a $q_n \times 1$ vector corresponding to significant effects and $\gamma_0^{(2)}$ is a $(p_n - q_n) \times 1$ vector of zeros. We also assume that only the first g functional predictors are significant or, equivalently, that the true values of the regression functions, β_{j0}, are such that $\beta_{j0}(t) \equiv 0$ for $j = g + 1, \ldots, d$. Each functional predictor $X_j(\cdot)$ is an infinite-dimensional process and requires regularization. Therefore the proposed model has a partially functional structure that combines the multiple functional and high-dimensional scalar components into a single linear framework.

Let $\{(y_i, x_i, z_i) : i = 1, \ldots, n\}$ denote independent and identically distributed realizations from the population (Y, X, Z). Let x_{ij} denote the jth component of x_i for $j = 1, \ldots, d$, and let z_{il} be the lth component of z_i for $l = 1, \ldots, p_n$. We further write $Y_M = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $Z_M = (z_1, \ldots, z_n)^{\mathrm{T}}$. To estimate the functions $\{\beta_j(\cdot) : j = 1, \ldots, d\}$ and the regression coefficients $\{\gamma_l : l = 1, \ldots, p_n\}$, we consider the least-squares loss, which couples $\beta_j(t) = \sum_k b_{jk}\phi_{jk}(t)$ with $x_{ij}(t) = \sum_k \xi_{ijk}\phi_{jk}(t)$ for each $j = 1, \ldots, d$ given the complete orthonormal basis series $\{\phi_{jk}\}_{k=1,2,\ldots}$,

$$\begin{aligned} L(b, \gamma \mid \mathcal{D}_n) &= \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{d} \int_T x_{ij}(t)\beta_j(t)\,\mathrm{d}t - z_i^{\mathrm{T}}\gamma \right\}^2 \\ &= \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \sum_{k=1}^{\infty} b_{jk}\xi_{ijk} - z_i^{\mathrm{T}}\gamma \right)^2, \end{aligned} \tag{3}$$

where $\mathcal{D}_n = \{(y_i, x_i, z_i) : i = 1, \ldots, n\}$ and $b = (b_1^{\mathrm{T}}, \ldots, b_d^{\mathrm{T}})^{\mathrm{T}}$ with $b_j = (b_{j1}, b_{j2}, \ldots)^{\mathrm{T}}$ for each j. It is evident that the loss function (3) should not be directly minimized due to the infinite

expansions of the functional predictors and high-dimensional scalar covariates, requiring suitable regularization for both X and Z.

A primary goal for (2) is to extract useful information from Z and X, whereas the classical functional linear model focuses only on a single functional predictor. It is therefore essential to select and estimate the nonzero coefficients in γ and nonzero functions in $b_1, \ldots, b_d$ to enhance model prediction and interpretability. To achieve simultaneous variable selection and estimation, we introduce a shrinkage penalty function $J_\lambda(\cdot)$ associated with a tuning parameter λ. Many penalty choices are available for variable selection. We use the smoothly clipped absolute deviation penalty of Fan & Li (2001), whose derivative is $J'_\lambda(|\gamma|) = \lambda[I(|\gamma| \leqslant \lambda) + I(|\gamma| > \lambda)(a\lambda - |\gamma|)_+/\{($

using $\lambda_{jn} = \lambda_n \big(\sum_{k=1}^{s_{nj}} \hat{w}_{jk} \big)^{1/2}$, which simplifies both the computation and the theoretical analysis. The estimated regression parameter functions are $\hat{\beta}_j(t) = \sum_{k=1}^{s_{nj}} \hat{b}_{jk} \hat{\phi}_{jk}(t)$.

2·3. *Algorithms and parameter tuning*

The optimization of (4) can be seen as a group smoothly clipped absolute deviation problem with different weights on the penalties, and the individual γ_l can be treated as a group of size 1. We propose two algorithms to solve the minimization problem (4), depending on the dimension p_n. Generally, when p_n is moderately large, say $p_n < n$, we modify the local linear approximation algorithm (Zou & Li, 2008), which inherits the computational efficiency and sparsity of lasso-type solutions. For ultrahigh p_n, especially $p_n \gg n$, the local linear approximation algorithm may not be applicable, and in that case we modify the concave-convex procedure used in Kim et al. (2008). The Appendix gives the details.

Two sets of tuning parameters play crucial roles in the penalized procedure (4). The parameter λ_n in the smoothly clipped absolute deviation directly controls the sparsity of both the functional and the nonfunctional predictors. Wang et al. (2007) showed that minimizing the BIC can identify the true model consistently, while generalized crossvalidation may lead to overfitting. The truncation parameters s_{nj} control the dimensions of the functional spaces used to approximate the true function parameters. In most previous work, each s_{nj} was chosen based on the functional principal component representation, such as leave-one-curve-out crossvalidation (Rice & Silverman, 1991) and the pseudo-AIC (Yao et al., 2005a). However, a sensible tuning criterion for s_{nj}

We minimize

$$\text{AIC}(s_{nj} : j \in D) = \log \text{RSS}(s_{nj} : j \in D) + 2n^{-1} \sum_{j \in D} s_{nj}$$

with respect to combinations of $\{s_{nj} : j \in D\}$, where

$$\text{RSS}(s_{nj} : j \in D) = \sum_{i=1}^{n} \left\{ y_i - \sum_{j \in D} \sum_{k=1}^{s_{nj}} \hat{\xi}_{ijk} \hat{b}^*_{jk}(s_{nj}) - \sum_{l \in S} z_{il} \hat{\gamma}^*_l(s_{nj}) \right\}^2,$$

with $\hat{b}^*_{jk}(s_{nj})$ and $\hat{\gamma}^*_l(s_{nj})$ being the refitted values using ordinary least squares.

3. Asymptotic properties

Denote the true values of $b^{(1)}$ and γ by $b_0^{(1)}$ and γ_0, respectively, and similarly for the remaining parameters. Recall that the boundedness of the covariance functions $K_j(s, t)$ and the regression operators implies that $\sum_{k=1}^{\infty} w_{jk} < \infty$ and $\sum_{k=1}^{\infty} b_{jk0}^2 < \infty$. We impose mild conditions on the decay rates of the eigenvalues $\{w_{jk}\}$ and regression coefficients $\{b_{jk0}\}$, similar to those adopted by Hall & Horowitz (2007) and Lei (2014). We assume:

Condition 1. $w_{jk} - w_{j(k+1)} \geqslant Ck^{-a-1}$ for $k \geqslant 1$, $j = 1, \ldots, d$.

This implies that $w_{jk} \geqslant Ck^{-a}$. As the covariance functions $K_1, \ldots, K_d$ are bounded, one has $a > 1$. Regarding the regression function $\beta_j(\cdot)$, in order to prevent the coefficients b_{jk0} from decreasing too slowly, we assume that:

Condition 2. $|b_{jk0}| \leqslant Ck^{-b}$ for $k > 1$, $j = 1, \ldots, d$.

These decay conditions are needed only to control the tail behaviour for large k, and so are not as restrictive as they appear. Without loss of generality, we use a common truncation parameter s_n in the theoretical analysis. It is important to control s_n appropriately. On the one hand, s_n cannot be too large due to increasingly unstable functional principal component estimates:

Condition 3. $(s_n^{2a+2} + s_n^{a+4})/n = o(1)$.

On the other hand, s_n cannot be too small, so that the covariances between Z and the unobservable $\{\xi_{jk} : k \geqslant s_n + 1\}$ are asymptotically negligible:

Condition 4. $s_n^{2b-1}/n \to \infty$ as $n \to \infty$.

Combining Conditions 3 and 4 entails that $b > \max(a + 3/2, a/2 + 5/2)$ is a sufficient condition for such an s_n to exist. This implies that the regression function is smoother than the lower bound on the smoothness of K_j. Regarding the dimension of scalar covariates, assume that the number of significant covariates satisfies:

Condition 5. $s_n^{a+2} q_n^2/n = o(1)$.

Such $q_n = o(n^{1/2} s_n^{-a/2-1})$ does exist and is allowed to diverge with the sample size, given Condition 3. The dimension of the candidate set, p_n, is allowed to be ultrahigh.

Condition 6. $p_n = O\{\exp(n^{\alpha})\}$ for some $\alpha \in (0, 1/2)$.

Finally, we require the following to hold for the tuning parameter λ_n and the sparsity of γ, in order to achieve consistent estimation:

Condition 7. $\lambda_n = o(1), \max\{n^{2\alpha-1},\ n^{-1}(q_n + s_n)\} = o(\lambda_n^2)$ and $\min_{l=1,\ldots,q_n} |\gamma_{l0}|/\lambda_n \to \infty$.

In the Supplementary Material we give standard conditions on the underlying processes x_{ij}, describe how the data are sampled and smoothed, and present auxiliary lemmas and proofs.

To facilitate the theoretical analysis, we reparameterize by writing $\tilde{b}_{jk} = w_{jk}^{1/2} b_{jk}$, so that the functional principal component scores serving as predictor variables are on a common scale of variability. This reparameterization is used only for technical derivations and does not appear in the estimation procedure. Let $\tilde{\eta} = (\tilde{b}^{(1)\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}}$, where $\tilde{b}^{(1)} = (\tilde{b}_1^{(1)\mathrm{T}}, \ldots, \tilde{b}_d^{(1)\mathrm{T}})^{\mathrm{T}}$ with $\tilde{b}_j^{(1)} = A_j b_j^{(1)}$; here A_j is the $s_n \times s_n$ diagonal matrix with $A_j(k,k) = w_{jk}^{1/2}$. Then, minimization of (4) is equivalent to minimizing

$$Q_n(\tilde{\eta}) = \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{d} \sum_{k=1}^{s_n} (\hat{\xi}_{ijk} w_{jk}^{-1/2}) \tilde{b}_{jk} - z_i^{\mathrm{T}} \gamma \right\}^2 + 2n \sum_{l=1}^{p_n} J_{\lambda_n}(|\gamma_l|) + 2n \sum_{j=1}^{d} J_{\lambda_{jn}}(\|b_j^{(1)}\|).$$

Theorem 1 establishes the estimation and selection consistency for both the functional and the scalar regression parameters. For a random variable ε with mean zero, ε is said to be a sub-Gaussian random variable if there exists some positive constant $C_1 > 0$ such that $\mathrm{pr}(|\varepsilon| > t) \leqslant \exp(-2^{-1} C_1 t^2)$ for $t \geqslant 0$. Let $\check{b}^{(1)}$ denote the estimate of $\tilde{b}^{(1)}$.

THEOREM 1. *If $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed sub-Gaussian random variables, then under Conditions 1–7 and* S1–S5 *in the Supplementary Material, there exists a local minimizer $\check{\eta} = (\check{b}^{(1)\mathrm{T}}, \hat{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ of $Q_n(\tilde{\eta})$ such that $\|\check{\eta} - \tilde{\eta}_0\| = O_{\mathrm{p}}[\{(q_n + s_n)/n\}^{1/2}]$ and $\mathrm{pr}(\hat{\gamma}_2 = 0,\ \check{b}^{(1)} = 0,\ j = g+1, \ldots, d) \to 1$.*

The estimation consistency result is expressed in terms of $\tilde{b}^{(1)}$, not the original parameter $b^{(1)} = (b_1^{(1)\mathrm{T}}, \ldots, b_d^{(1)\mathrm{T}})^{\mathrm{T}}$. For estimation, given $\hat{b}^{(1)} = A_j^{-1} \check{b}^{(1)}$, it is easy to deduce that $\|\hat{\beta}_j - \beta_{j0}\|_{L^2}^2 = O_{\mathrm{p}}\{s_n^a (q_n + s_n)/n\}$, where $\hat{\beta}_j = \sum_{k=1}^{s_n} \hat{b}_{jk} \hat{\phi}_{jk}$ and $\beta_{j0} = \sum_{k=1}^{\infty} b_{jk0} \phi_{jk}$. Theorem 2 establishes the asymptotic normality of the q_n-dimensional vector $\hat{\gamma}^{(1)}$. Write $\Sigma_1 = E(z_i^{(1)} z_i^{(1)\mathrm{T}})$ and $\hat{\Sigma}_1 = n^{-1} \sum_{i=1}^{n} z_i{}^{(1)} z_i{}^{(1)\mathrm{T}}$ with $z_i{}^{(1)} = (z_{i1}, \ldots, z_{iq_n})^{\mathrm{T}}$.

THEOREM 2. *If $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed sub-Gaussian random variables and $q_n = o(n^{1/3})$, then under Conditions 1–7 and* S1–S5 *in the Supplementary Material, for the local minimizer in Theorem 1, $n^{1/2} A_n \hat{\Sigma}_1 (\hat{\gamma}^{(1)} - \gamma_0^{(1)}) \to N(0,\ \sigma^2 H^* + B^*)$ in distribution, for any $r \times q_n$ matrix A_n such that $G = \lim_{n\to\infty} A_n A_n^{\mathrm{T}}$ is positive definite; here $\sigma^2 = \mathrm{var}(\epsilon)$, $H^* = \lim_{n\to\infty} A_n \Sigma_1 A_n{}^{\mathrm{T}}$ and $B^* = \lim_{n\to\infty} A_n B_n A_n^{\mathrm{T}}$ with*

$$B_n = \mathrm{cov}\left\{ \sum_{j=1}^{g} \sum_{k=1}^{s_n} \sum_{v \neq k} b_{jk0} (w_{jk} - w_{jv})^{-1} \langle \Xi_j, \phi_{jv} \rangle \int (x_{ij} \otimes x_{ij}) \phi_{jk} \phi_{jv} \right\},$$

where $\Xi_j = (\Xi_{j1}, \ldots, \Xi_{jq_n})^{\mathrm{T}}$, $E\{X_j(t) Z_l\} = \Xi_{jl}(t)$ and $(x_{ij} \otimes x_{ij})(s,t) = x_{ij}(s) x_{ij}(t)$.

The asymptotic covariance is inflated by estimating the unobservable functional principal component scores. The inflation is quantified by a convergent sequence B_n associated with the truncation size s_n.

4. SIMULATION STUDIES

The simulated data $\{y_i : i = 1, \ldots, n\}$ are generated from the model

$$y_i = \sum_{j=1}^{d} \int_0^1 \beta_j(t) x_{ij}(t)\,\mathrm{d}t + z_i^{\mathrm{T}}\gamma + \epsilon_i = \sum_{j=1}^{d} \sum_{k} b_{jk}\xi_{ijk} + z_i^{\mathrm{T}}\gamma + \epsilon_i,$$

with $d = 4$ functional predictors and p_n scalar covariates; the errors $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed from $N(0, \sigma^2)$, and γ is the vector of scalar coefficients. The functional predictors have mean zero and covariance function derived from the Fourier basis $\phi_{2\ell-1} = 2^{-1/2}\cos\{(2\ell-1)\pi t\}$ and $\phi_{2\ell} = 2^{-1/2}\sin\{(2\ell-1)\pi t\}$ $(\ell = 1, \ldots, 25; t \in T = [0, 1])$. The underlying regression function is $\beta_j(t) = \sum_{k=1}^{50} b_{jk}\phi_k(t)$, a linear combination of the eigenbasis. The scalar covariates $z_i = (z_{i1}, \ldots, z_{ip_n})^{\mathrm{T}}$ are jointly normal with zero mean, unit variance and AR(0·5) correlation structure. Next, we describe how to generate the $d = 4$ functional predictors $x_{ij}(t)$. For $j = 1, \ldots, 4$, define $V_{ij}(t) = \sum_{k=1}^{50} \tilde{\xi}_{ijk}\phi_k(t)$, where $\{\tilde{\xi}_{ijk} : i = 1, \ldots, n\}$ are independent and identically distributed as $N(0,\ 16k^{-2})$ for different i and j. The four functional predictors are then defined through the linear transformations

$$x_{i1} = V_{i1} + 0{\cdot}5(V_{i2} + V_{i3}), \quad x_{i2} = V_{i2} + 0{\cdot}5(V_{i1} + V_{i3}),$$
$$x_{i3} = V_{i3} + 0{\cdot}5(V_{i1} + V_{i2}), \quad x_{i4} = V_{i4}.$$

Here, the first three functional predictors are correlated with each other. To be more realistic, we allow moderate correlation between V_{i1} and z_i $(i = 1, \ldots, n)$ by giving $\tilde{\xi} = (\tilde{\xi}_{i11}, \tilde{\xi}_{i12}, \tilde{\xi}_{i13}, \tilde{\xi}_{i14})^{\mathrm{T}}$ and $z_i = (z_{i1}, \ldots, z_{ip_n})^{\mathrm{T}}$ a correlation structure specified by $\mathrm{corr}(\tilde{\xi}_{i1k}, z_{il}) = r^{|k-l|+1}$ $(k = 1, \ldots, 4; l = 1, \ldots, p_n)$ with $r = 0{\cdot}2$. For the actual observations, we assume them to be realizations of $\{x_{ij}(\cdot) : j = 1, 2, 3, 4\}$ at 100 equally spaced times $\{t_{ijl} \in T : l = 1, \ldots, 100\}$ with independent and identically distributed noise $\varepsilon_{ijl} \sim N(0, 1)$.

We use 200 Monte Carlo runs for model assessment. Since inferences on both the parametric component γ and the functional components β_j are of interest, we report the Monte Carlo averages for the numbers of false nonzero and false zero functional predictors, as well as the functional mean squared error, $\mathrm{MSE_f} = \sum_{j=1}^{d} E(\|\hat{\beta}_j - \beta_j\|_{L^2}^2)$. For the scalar covariates, we report the Monte Carlo averages of the numbers of false nonzero and false zero scalar covariates, along with the scalar mean squared error, $\mathrm{MSE_s} = E(\|\hat{\gamma} - \gamma\|^2)$. The prediction error is assessed using an independent test set of size $N = 1000$ for each Monte Carlo repetition, and is defined as $\mathrm{PE} = N^{-1}\sum_{i=1}^{N}(y_i^* - \hat{y}_i^*)^2 - \sigma^2$ where $\{x_{ij}^*, z_i^*, y_i^* : j = 1, \ldots, 4\}$ are the testing data generated from the same model; the predictions are $\hat{y}_i^* = \sum_j \sum_k \hat{\xi}_{ijk}^* \hat{b}_{jk} + z_i^{*\mathrm{T}}\hat{\gamma}$, obtained by inserting estimates from the corresponding training sample.

Design I is for a moderate number of scalar covariates, with sample size $n = 200$ and error variance $\sigma^2 = 1$. Specifically, for $j = 1$ or 2, $b_{j1} = 1$, $b_{j2} = 0{\cdot}8$, $b_{j3} = 0{\cdot}6$, $b_{j4} = 0{\cdot}5$, $b_{jk} = 8(k-2)^{-4}$ for $k = 5, \ldots, 50$, $\beta_3 = \beta_4 = 0$, and $\gamma = (1_5^{\mathrm{T}}, 0_{15}^{\mathrm{T}})^{\mathrm{T}}$. Hence $p_n = 20$ and $q_n = 5$. To illustrate the effect of the choice of s_n, in Table 1 we exhibit the results for s_n ranging from 1 to 16 with λ_n chosen by BIC. The selection of functional and scalar predictors is quite accurate and stable for a wide range of s_n, but with a very small number of false nonzero scalars. For functional predictors, the functional mean squared error improves until s_n reaches an optimal

Table 1. *Simulation results for sample size* $n = 200$ *based on* 200 *Monte Carlo replicates for Designs I and II; values reported are Monte Carlo averages with standard errors in parentheses. We first use* ABIC *to choose the tuning parameter* λ_n *and a common truncation* s_n*, and then tune* s_{nj} *jointly with* AIC *by refitting the selected model using ordinary least squares. In Design II, step* 1 *results are based on the original sample in each Monte Carlo run, while step* 2 *yields improved results by fitting the penalized procedure to the selected model in step* 1 *with an additional sample of* $n = 200$

Design	s_n	FZ_f	FN_f	MSE_f	FZ_s	FN_s	MSE_s	PE
I	1	0·95	0	4·1 (0·03)	0·39	7·1	4·7 (0·15)	27·9 (0·6)
($p_n = 20$)	2	0·35	0	2·2 (0·10)	0·05	3·2	1·1 (0·06)	7·9 (0·3)
	3	0	0	0·6 (0·01)	0	0·91	0·22 (0·013)	1·5 (0·03)
	4	0	0	0·12 (0·005)	0	0·36	0·067 (0·005)	0·21 (0·007)
	5	0	0	0·14 (0·005)	0	0·38	0·069 (0·004)	0·19 (0·006)
	6	0	0	0·19 (0·007)	0	0·44	0·072 (0·004)	0·19 (0·006)
	10	0	0	0·65 (0·03)	0	0·38	0·073 (0·004)	0·22 (0·006)
	16	0·03	0·08	3·1 (0·13)	0	0·11	0·074 (0·006)	0·54 (0·1)
					$\hat{s}_n = 4{\cdot}30$ (0·050)			
	ABIC	0	0	0·13 (0·007)	0	0·34	0·067 (0·004)	0·20 (0·006)
					$\hat{s}_{n1} = 4{\cdot}78$ (0·075), $\hat{s}_{n2} = 4{\cdot}87$ (0·071)			
	Tune s_{nj}	0	0	0·09 (0·003)	0	0·28	0·065 (0·004)	0·18 (0·006)
II					$\hat{s}_n = 4{\cdot}07$ (0·034)			
($p_n = 1000$)	Step 1	0	0·04	0·18 (0·004)	0	7·4	0·36 (0·018)	1·1 (0·032)
	Step 2	0	0	0·095 (0·005)	0	0·10	0·047 (0·003)	0·17 (0·004)
					$\hat{s}_{n1} = 4{\cdot}76$ (0·066), $\hat{s}_{n2} = 4{\cdot}62$ (0·055)			
	Tune s_{nj}	0	0	0·076 (0·003)	0	0·09	0·046 (0·003)	0·16 (0·004)

FZ_f, number of false zero functional predictors; FN_f, number of false nonzero functional predictors; MSE_f, functional mean squared error; FZ_s, number of false zero scalar covariates; FN_s, number of false nonzero scalar covariates; MSE_s, scalar mean squared error; PE, prediction error.

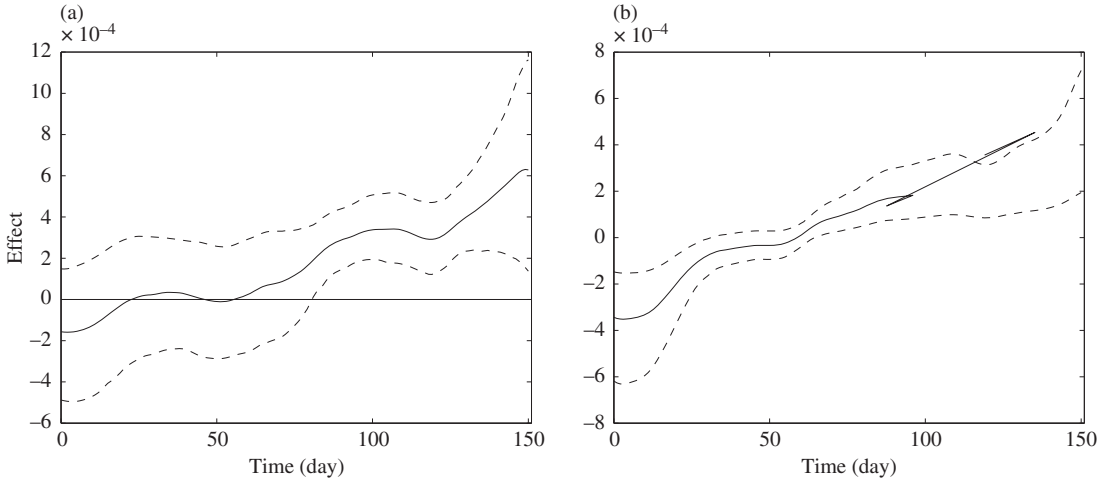
level, and then deteriorates as s_n continues to increase. For the scalar covariates, the mean squared error and prediction error appear more stable beyond the optimal level. We then use ABIC with a common s_n to select both s_n and λ_n. It yields similar results to those at the optimal mean squared and prediction errors, selecting an average $\hat{s}_n = 4{\cdot}30$ with a standard error of 0·050. Refitting the selected model using ordinary least squares with jointly tuned s_{nj} via AIC improves the estimation of the functional coefficients and the overall prediction.

Design II illustrates the situation where the scalar covariates are of ultrahigh dimension, $\gamma = (1_5^{\mathrm{T}}, 0_{995}^{\mathrm{T}})^{\mathrm{T}}$ with $p_n = 1000$, while other settings remain the same as in Design I. The ABIC yields results similar to those giving the optimal estimation and prediction. The number of false nonzero scalar covariates, the scalar mean squared error and the prediction error in step 1 become larger than those in Design I, mainly due to the ultrahigh number of insignificant scalar covariates. The functional mean squared error is also higher, as the correlation between functional predictors and scalar covariates becomes greater for larger p_n. To improve the estimation and prediction, for each Monte Carlo run, after obtaining the estimates based on ABIC in step 1, we generate an additional sample of size 200 and implement the penalized procedure using the significant variables and s_n selected in step 1. The results from step 2, summarized in Table 1, are dramatically improved and become comparable to those for Design I. This hints at a promising two-step procedure via sample splitting when p_n is ultrahigh, in a similar spirit to the approach of Fan et al. (2012). Further improvement can be achieved by refitting the selected model with jointly tuned s_{nj} using ordinary least squares. Additional simulations are presented in the Supplementary Material.

5. APPLICATION

We applied our method to a dataset from the National Mortality, Morbidity, and Air Pollution Study that contains air pollution measurements and mortality counts for U.S. cities collected during the census in year 2000. A main goal of the study was to investigate the impact of air pollution on the nonaccidental mortality rate across different cities in the U.S.A., whilst taking into account climate patterns and information from the U.S. census. In previous work, a two-stage analysis was conducted: first the short-term effect of certain air pollutants on the mortality count for each city was modelled; then the estimates across different cities were combined (Peng et al., 2005, 2006). By contrast, we apply the partially functional linear regression model to the data for different cities. In particular, we are interested in studying the effect on mortality of particulate matter with an aerodynamic diameter of less than $2{\cdot}5\,\mu\mathrm{m}$, abbreviated as PM 2·5 and measured in $\mu\mathrm{g\,m^{-3}}$, because its negative impact on health, as revealed by toxicological and epidemiological studies, has brought it to the public's attention in recent years. Other studies (Samoli et al., 2013; Pascal et al., 2014) have shown that PM 2·5 has a larger effect on mortality in warm weather, so we focus on daily concentration measurements of PM 2·5 from 1 April 2000 to 31 August 2000; these, along with daily observations of temperature and humidity, were treated as the functional predictors. After removing cities with more than ten consecutive missing measurements of PM 2·5, a total of 69 cities were included in our analysis. The response of interest is the log-transformed total nonaccidental mortality rate in the following month, September 2000, among individuals of age 65 and older, who account for the majority of nonaccidental deaths. The scalar covariates available from the U.S. census for each city are land area per individual, water area per individual, proportion of urban population, proportion of the population with at least a high school diploma, proportion of the population with at least a university degree, proportion of the population below the poverty line, and proportion of household owners.

The ABIC was used first to choose significant predictors with a common truncation, and this was followed by a least-squares refitting using AIC to tune s_{nj} jointly. Among scalar covariates, our analysis shows that only the proportion of household owners has a negative effect ($-1{\cdot}80$ with a standard error of 0·41), indicating that household owners tend to have a lower mortality rate. The standard error was based on 1000 bootstrap samples, obtained by fitting the selected model using ordinary least squares. Our method also selected two significant functional predictors, PM 2·5 and temperature. The least-squares refitting chose the truncation numbers $\hat{s}_{n1} = 2$ and $\hat{s}_{n2} = 2$. The estimated regression parameter functions together with their 95% bootstrap confidence bands are plotted in Fig. 1. We observe that higher PM 2·5 concentrations in the summer, especially in July and August, can lead to increased mortality in the period immediately afterwards. This coincides with the findings of Samoli et al. (2013) and Pascal et al. (2014), but needs to be interpreted with caution, as the effect could be partially explained by the proximity of the pollution period to the time of death. Higher temperatures in the summer, in contrast to lower temperatures in April, may also increase the mortality rate, agreeing with the results of Curriero et al. (2002). To better understand the effects of functional predictors, we fitted a linear regression using only the selected scalar covariate, obtaining $R^2 = 0{\cdot}15$. Including temperature leads to $R^2 = 0{\cdot}25$, and including both temperature and PM 2·5 yields $R^2 = 0{\cdot}38$. A heuristic F-test for the significance of two principal component scores of temperature gives a p-value of 0·01, and adding an extra two principal component scores of PM 2·5 gives a p-value of 0·0008. For comparison, we also fitted the marginal models containing only PM 2·5 or temperature using classical functional linear regression. The marginal F-tests for temperature and PM 2·5 gave p-values of 0·0001 and 0·004, respectively. The regression parameter functions show similar patterns and are omitted. We conclude that, after adjusting for temperature and household ownership, summer PM 2·5 concentrations have a significant impact on the near-future mortality rate of elder residents in U.S. cities.


(a)
× 10−4
12
10
8
6
4
2
0
−2
−4
−6
Effect
0
50
100
150
Time (day)
(b)
× 10−4
8
6
4
2
0
−2
−4
−6
−8
0
50
100
150
Time (day)

APPENDIX

Algorithm details

Recall that $Y_M = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $Z_M = (z_1, \ldots, z_n)^{\mathrm{T}}$, where $z_i = (z_{i1}, \ldots, z_{ip_n})^{\mathrm{T}}$. In addition, M_j is a $n \times s_n$ matrix with (i, k)th element ξ_{ijk}, $M = (M_1, \ldots, M_d)$, and $N = (M, Z_M) = (N_1, \ldots, N_n)^{\mathrm{T}}$ is a $n \times (ds_n + p_n)$ matrix. Further, $\eta = (b^{(1)\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}}$. The solution to (4) is equivalent to

$$\arg\min_{\eta} \left\{ (2n)^{-1} \|Y_M - N\eta\|^2 + \sum_{r=1}^{R} J_{\lambda_r}(\|\eta_r\|) \right\}$$

where $R = d + p_n$. The tuning parameter is $\lambda_r = \lambda_{rn}$ with group size $K_r = s_n$ if $r = 1, \ldots, d$, and $\lambda_r = \lambda_n$ with group size $K_r = s_n$ if $r = d + 1, \ldots, d + p_n$.

When p_n is moderately large, say $p_n < n$, one can modify the local linear approximation algorithm of Zou & Li (2008), which inherits the computational efficiency and sparsity of lasso-type solutions. Denote the initial estimate from the ordinary least-squares solution by $\hat{\eta}^{(0)}$, and solve $\hat{\eta}^{(1)} = \arg\min_{\eta}\{(2n)^{-1} \|Y_M - N\eta\|^2 + \sum_{r=1}^{R} J'_{\lambda_r}(\|\eta_r^{(0)}\|)\|\eta_r\|\}$. Since some of the $J'_{\lambda_r}(\|\eta_r^{(0)}\|)$ are zero, we use a similar algorithm proposed by Zou & Li (2008). Write $V = \{r : J'_{\lambda_r}(\|\eta_r^{(0)}\|) = 0\}$, $W = \{r : J'_{\lambda_r}(\|\eta_r^{(0)}\|) > 0\}$, $N = (N_V, N_W)$ and $\eta^{(1)} = (\eta_V^{(1)\mathrm{T}}, \eta_W^{(1)\mathrm{T}})^{\mathrm{T}}$. Our algorithm is as follows.

Algorithm 1.

(i) Reparameterize the response vector by $Y_M^* = N\eta^{(0)}$, and reparameterize the observed data matrix by $N_r^* = N_r K_r^{1/2} / J'_{\lambda_r}(\|\eta_r^{(0)}\|)$ for $r \in W$ and $N_r^* = N_r$ for $r \in V$.

(ii) Let P_V denote the projection matrix of the space $\{N_r^* : r \in V\}$, where $P_V = N_V(N_V^{\mathrm{T}} N_V)^{-1} N_V{}^{\mathrm{T}}$. Then, calculate $Y_M^{**} = Y_M^* - P_V Y_M^*$ and $N_W^{**} = N_W^* - P_V N_W^*$.

(iii) Find $\hat{\eta}_W^* = \arg\min_{\beta}\{(2n)^{-1} \|Y_M^{**} - N_W^{**}\eta\|^2 + \sum_{r \in W} K_r^{1/2} \|\eta_r\|\}$.

(iv) Compute $\hat{\eta}_V^* = (N_V^{*\mathrm{T}} N_V^*)^{-1} N_V^{*\mathrm{T}} (Y_M^* - N_W^* \hat{\eta}*$

Escabias, M., Aguilera, A. M. & Valderrama, M. J. (2004). Principal component estimation of functional logistic regression: Discussion of two different approaches. *J. Nonparam. Statist.* **16**, 365–84.
Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.
Fan, J. & Zhang, J.-T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J. R. Statist. Soc.* B **62**, 303–22.
Fan, J., Guo, S. & Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Statist. Soc.* B **74**, 37–65.
Fan, J., Yao, Q. & Cai, Z. (2003). Adaptive varying-coefficient linear models. *J. R. Statist. Soc.* B **65**, 57–80.
Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *J. Comp. Graph. Statist.* **7**, 397–416.
Hall, P. & Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35**, 70–91.
Hall, P., Müller, H. G. & Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34**, 1493–517.
Kim, Y., Choi, H. & Oh, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *J. Am. Statist. Assoc.* **103**, 1665–73.
Lei, J. (2014). Adaptive global testing for functional linear models. *J. Am. Statist. Assoc.* **109**, 624–34.
Li, Y., Wang, N. & Carroll, R. J. (2010). Generalized functional linear models with semiparametric single-index interactions. *J. Am. Statist. Assoc.* **105**, 621–33.
Lu, Y., Du, J. & Sun, Z. (2014). Functional partially linear quantile regression model. *Metrika* **77**, 317–32.
Morris, J. S., Vannucci, M., Brown, P. J. & Carroll, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *J. Am. Statist. Assoc.* **98**, 573–97.
Müller, H.-G. & Stadtmüller, U. (2005). Generalized functional linear models. *Ann. Statist.* **33**, 774–805.
Müller, H.-G. & Yao, F. (2008). Functional additive models. *J. Am. Statist. Assoc.* **103**, 1534–44.
Parikh, N. & Boyd, S. (2013). Proximal algorithms. *Foundat. Trends Optimiz.* **1**, 123–231.
Pascal, M., Falq, G., Wagner, V., Chatignoux, E., Corso, M., Blanchard, M., Host, S., Pascal, L. & Larrieu, S. (2014). Short-term impacts of particulate matter (PM10, PM10–2·5, PM2·5) on mortality in nine French cities. *Atmosph. Envir.* **95**, 175–84.
Peng, R. D., Dominici, F., Pastor-Barriuso, R., Zeger, S. L. & Samet, J. M. (2005). Seasonal analyses of air pollution and mortality in 100 US cities. *Am. J. Epidemiol.* **161**, 585–94.
Peng, R. D., Dominici, F. & Louis, T. A. (2006). Model choice in time series studies of air pollution and mortality. *J. R. Statist. Soc.* A **169**, 179–203.
Ramsay, J. O. & Dalzell, C. J. (1991). Some tools for functional data analysis (with Discussion). *J. R. Statist. Soc.* B **53**, 539–72.
Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*. New York: Springer, 2nd ed.
Rice, J. A. & Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* B **53**, 233–43.
Samoli, E., Stafoggia, M., Rodopoulou, S., Ostro, B., Declercq, C., Alessandrini, E., Díaz, J., Karanasiou, A., Kelessis, A. G., Le Tertre, A. et al.(2013). Associations between fine and coarse particles and mortality in Mediterranean cities: Results from the MED-PARTICLES project. *Envir. Health Perspect.* **121**, 932–8.
Shin, H. (2009). Partial functional linear regression. *J. Statist. Plan. Infer.* **139**, 3405–18.
Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.
Wang, H., Li, R. & Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **93**, 553–68.
Yao, F. & Müller, H.-G. (2010). Functional quadratic regression. *Biometrika* **97**, 49–64.
Yao, F., Müller, H.-G. & Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *J. Am. Statist. Assoc.* **100**, 577–90.
Yao, F., Müller, H.-G. & Wang, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873–903.
Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B **68**, 49–67.
Zhang, J.-T. & Chen, J. (2007). Statistical inferences for functional data. *Ann. Statist.* **35**, 1052–79.
Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.
Zou, H. & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36**, 1509–33.

[*Received October* 2014. *Revised October* 2015]