Biometrika (20XX), **XX**, 1, *pp*. 1–13
© 20XX Biometrika Trust
Printed in Great Britain

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 AND FANG YAO

Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5S 3G3, Canada

kaijie@utstat.toronto.edu fyao@utstat.toronto.edu

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 non-functional 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 those covariates, including both functional and scalar types. The new approach provides a unified and flexible framework to simultaneously take into account multiple functional and ultra-high dimensional scalar predictors, identify important features and improve 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 impact of regularization on the resulting estimators. We establish the consistency and oracle properties of the proposed method under mild conditions, illustrate its performance with simulation studies, and apply it to 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 by an integral form of a regression parameter function; see Ramsay & Dalzell (1991), Cardot et al. (2003), Cuevas et al. (2002), Yao et al. (2005a) and Ramsay & Silverman (2005). 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), Hall et al. (2006), Cai & Hall (2006), Zhang & Chen (2007), and Hall & Horowitz (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 problems, it is common to also collect a large number of non-functional predictors. How to incorporate scalar predictors in functional linear regression and perform model selection/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 non-functional 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 regularizes each functional predictor as a whole, combined 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 functional predictors. We treat the functional predictors as infinite-dimensional processes, which distinguishes our work from work that fixes the number of principal components (Li et al., 2010). A main contribution 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

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

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 non-increasing sequence of non-negative 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 $w_1 > w_2 > \cdots \geq 0$. 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 $\xi_{ik} = \int_T \{x_i(t) - \mu(t)\}\phi_k(t)dt$ 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 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 non-functional predictors affect the response. We assume the number of functional predictors d to be fixed, while the number of scalar covariates p_n may grow with the sample size. Specifically we allow p_n to be ultra-high 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 centered to have mean zero. We then model the linear relationship between Y and (X, Z) by

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

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 non-functional covariates, and ϵ is the random error 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 regression coefficients γ_0^{T} equals $(\gamma_0^{(1)\mathrm{T}}, \gamma_0^{(2)\mathrm{T}})$, where $\gamma_0^{(1)}$ is a $q_n \times 1$ vector corresponding to significant effects and $\gamma_0^{(2)}$ is a $(p_n - q_n) \times 1$ zero vector. We also assume that only the first g functional predictors are significant, equivalently, the true values of regression functions $\beta_{j0}(t) \equiv 0$ 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 one 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 $\{\phi_{jk}\}_{k=1,2,\ldots}$,

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

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.

One 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 thus essential to select and estimate the nonzero coefficients in γ and nonzero functions in $b_1, \ldots, b_d$ to en-

hance 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 $J'_\lambda(|\gamma|) = \lambda\{I(|\gamma| \le \lambda) + I(|\gamma| > \lambda)(a\lambda - |\gamma|)_+/\{(a-1)\lambda\}$ with $a = 3.7$ suggested by Fan & Li (2001) for implementation.

Due to the infinite dimensionality of the functional predictors, smoothing and regularization are necessary in estimation. It is sensible to control the complexity of $\beta_j(t)$ as a whole function, rather than treating its basis terms as separate predictors. We adopt the simple yet effective truncation approach in the spirit of controlling smoothness as in classical nonparametric regression. Denote the truncated form by $X_{s_j}(t) = \mu_j(t) + \sum_{k=1}^{s_j} \xi_{jk}\phi_{jk}(t)$ for $j = 1, \ldots, d$, where s_j is the truncation parameter. Correspondingly, for $\beta_j(t) = \sum_{k=1}^{\infty} b_{jk}\phi_{jk}(t)$, write $b_j = (b_j^{(1)\mathrm{T}}, b_j^{(2)\mathrm{T}})^{\mathrm{T}}$, where $b_j^{(1)} = (b_{j1}, ..., b_{js_j})^{\mathrm{T}}$ and $b_j^{(2)} = (b_{j,s_j+1}, \ldots)^{\mathrm{T}}$. Unlike with non-functional effects, there is no underlying separation between $b_j^{(1)}$ and $b_j^{(2)}$. For the sake of adaptivity, we allow s_j to vary with the sample size, $s_j \equiv s_{nj}$, such that it plays a role of a smoothing parameter that balances the trade-off between bias and variance. The coefficients in $b_j^{(2)}$ associated with higher-order basis functions are nonzero but decay rapidly. It is of interest to study the impact of s_{nj} on the convergence rates of the resulting estimators.

In practice we do not observe the entire trajectories x_{ij}, but only have intermittent noisy measurements $W_{ijl} = x_{ij}(t_{ijl}) + \varepsilon_{ijl}$, where $\{\varepsilon_{ijl}, i = 1 \ldots, n\}$ are independent and identically distributed measurement errors independent of x_{ij}, satisfying $E(\varepsilon_{ijl}) = 0$, $\mathrm{var}(\varepsilon_{ijl}) = \sigma_{xj}^2$ for $i = 1, \ldots, n$ and $l = 1, \ldots, m_{ij}$. When the repeated observations are sufficiently dense for each subject, a common practice is to run a smoother through $\{(t_{ijl}, W_{ijl}), l = 1, \ldots, m_{ij}\}$, and then the estimates $\{\hat{x}_{ij}, i = 1, \ldots, n, j = 1, \ldots, d\}$ are used to construct the covariance, eigenvalues/basis, and functional principal component scores; details are given in the Supplementary Material. A theoretical justification of the asymptotic equivalence between the estimators obtained from $\hat{x}_{ij}$ and those from the true x_{ij} is given in the Supplementary Material. The unobservable functional principal component scores $\{\xi_{ijk} : k = 1, \ldots, s_n; j = 1, \ldots, d; i = 1, \ldots, n\}$ are estimated by functional principal component analysis based on the observed data $\{(t_{ijl}, W_{ijl}) : l = 1, \ldots, m_{ij}; j = 1, \ldots, d; i = 1, \ldots, n\}$. Therefore we minimize

$$\min_{b^{(1)},\gamma} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{d}\sum_{k=1}^{s_{nj}} \hat{\xi}_{ijk} b_{jk} - z_i^{\mathrm{T}}\gamma)^2 + 2n\sum_{j=1}^{d} J_{\lambda_{jn}}\big(\|b_j^{(1)}\|\big) + 2n\sum_{l=1}^{p_n} J_{\lambda_n}(|\gamma_l|), \qquad (4)$$

given suitable choices of s_{nj}, λ_n and λ_{jn}, where $\|b_j^{(1)}\|$ is the Euclidean norm invoking a group penalty that shrinks the regression functions of unimportant functional predictors to zero. To regularize all predictors on a comparable scale, one often standardizes the predictors before imposing a penalty associated with a common tuning parameter (Fan & Li, 2001). Thus we standardize $(z_{1l}, \ldots, z_{nl})^{\mathrm{T}}$ to have unit variance. The variability of the jth functional predictor can be approximated by $\sum_{k=1}^{s_{nj}} \hat{w}_{jk}$, where $\hat{w}_{jk}$ is the kth estimated eigenvalue of the jth predictor. Since standardization is equivalent to adding weights to the penalty function, we suggest using $\lambda_{jn} = \lambda_n(\sum_{k=1}^{s_{nj}} \hat{w}_{jk})^{1/2}$, which simplifies both the computation and 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 penalties, and the individual γ_l can be treated as group of size one. We propose two algorithms to solve the minimization problem (4), which adapts to 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 ultra-high p_n, especially $p_n \gg n$, the local linear approximation algorithm may not be applicable, and then 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 non-functional predictors. Wang et al. (2007) showed that minimizing the BIC can identify the true model consistently, while generalized cross-validation might lead to over-fitting. The truncation parameters s_{nj} control the dimensions of the functional spaces to approximate the true function parameters. Previous work mostly chose s_{nj} based on the functional principal component representation, such as leave-one-curve-out cross-validation (Rice & Silverman, 1991) and the pseudo-AIC (Yao et al., 2005a). However, a sensible tuning criterion of s_{nj} in a regression setting should take into account its impact on the response. The process X_j is infinite-dimensional, and its coefficient function $\beta_j(t) = \sum_{k=1}^{\infty} b_{jk}\phi_k(t)$ does not have a finite cut-off. This is similar to the situation that the true model does not lie in the space formed by finite-dimensional candidate models. Therefore we propose a hybrid tuning procedure, which in principle combines BIC for tuning λ_n and AIC for choosing s_{nj} due to the infinite-dimensional parameter spaces of $\{\beta_j, j = 1, \ldots, \infty\}$. In practice it is computationally prohibitive to choose $\{s_{nj} : j = 1, \ldots, d\}$ simultaneously in the penalized procedure (4). Table 1 suggests that, when using a common truncation, the selection of both functional and scalar covariates is accurate and stable for a wide range of s_n. Thus we propose to use a common truncation parameter s_n when solving (4), then refit the selected model with the significant functional and scalar predictors using ordinary least squares, while different truncation parameters s_{nj} are tuned simultaneously by AIC for the retained functional predictors.

Specifically, for a fixed pair (s_n, λ_n), the ABIC criterion is defined as

$$\text{ABIC}(s_n, \lambda_n) = \log\left\{\text{RSS}(s_n, \lambda_n)\right\} + 2g(s_n, \lambda_n)s_n/n + n^{-1}\text{df}(s_n, \lambda_n)\log(n),$$

where

$$\text{RSS}(s_n, \lambda_n) = \sum_{i=1}^{n}\Big\{y_i - \sum_{j=1}^{d}\sum_{k=1}^{s_n}\hat{\xi}_{ijk}\hat{b}_{jk}(s_n, \lambda_n) - z_i^{\mathrm{T}}\hat{\gamma}(s_n, \lambda_n)\Big\}^2,$$

and $g(s_n, \lambda_n)$ is the number of non-zero estimates of the regression functions, $g(s_n, \lambda_n) = \sum_{j=1}^{d} I(\hat{\beta}_j; s_n, \lambda_n)$. The degree of freedom $\text{df}(s_n, \lambda_n)$ equals $I(\hat{\gamma}; s_n, \lambda_n) + \sum_{j=1}^{d}\sum_{k=1}^{s_n}\hat{w}_{jk}I(\hat{\beta}_j; s_n, \lambda_n)$, with $I(\hat{\gamma}; s_n, \lambda_n)$ indicating the number of non-zero elements in $\hat{\gamma}$. This procedure requires estimation using the whole data only once and is computationally fast.

For the refit step, denote the index set of the selected functional predictors by $D \subset \{1, \ldots, d\}$, and the index set of the selected scalar covariates by $S \subset \{1, \ldots, p_n\}$. 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

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

and $\hat{b}^*_{jk}(s_{nj})$ and $\hat{\gamma}^*_l(s_{nj})$ are 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, and similarly for the rest of parameters. Recall that the boundedness of the covariance functions $K_j(s,t)$ and 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 eigenvalues $\{w_{jk}\}$ and regression coefficients $\{b_{jk0}\}$, similar to those adopted by Hall & Horowitz (2007) and Lei (2014). We assume that, for $j = 1, \ldots, d$:

(A1) $w_{jk} - w_{j(k+1)} \geq Ck^{-a-1}$ for $k \geq 1$.

This implies that $w_{jk} \geq Ck^{-a}$. 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

(A2) $|b_{jk0}| \leq Ck^{-b}$ for $k > 1$.

The decay conditions are needed only to control the tail behaviors 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 theoretical analysis. It is important to control s_n appropriately. On one hand s_n cannot be too large due to increasingly unstable functional principal component estimates,

(A3) $(s_n^{2a+2} + s_n^{a+4})/n = o(1)$.

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

(A4) $s_n^{2b-1}/n \to \infty$ as $n \to \infty$.

Combining (A3) and (A4) entails $b > \max(a + 3/2, a/2 + 5/2)$ as 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

(A5) $s_n^{a+2} q_n^2/n = o(1)$,

Such $q_n = o(n^{1/2} s_n^{-a/2-1})$ does exist and is allowed to diverge with the sample size given (A3). The dimension of the candidate set, p_n, is allowed to be ultra-high,

(A6) $p_n = O\{\exp(n^\alpha)\}$ for some $\alpha \in (0, 1/2)$.

Lastly we require the following to hold for the tuning parameter λ_n and the sparsity of γ to achieve consistent estimation,

(A7) $\lambda_n = o(1)$, $\max\{n^{2\alpha-1}, n^{-1}(q_n + s_n)\} = o(\lambda_n^2)$, $\min_{l=1,\ldots,q_n} |\gamma_{l0}|/\lambda_n \to \infty$.

We defer to the Supplement Material the standard conditions (B1)–(B5) on the underlying processes x_{ij}, how the data are sampled and smoothed, as well as the scalar covariates, followed by the auxiliary lemmas and proofs.

To facilitate theoretical analysis, we re-parameterize 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 variabilities. This re-parameterization is only used 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 $\tilde{b}^{(1)} = (\tilde{b}_1^{(1)\mathrm{T}}, \ldots, \tilde{b}_d^{(1)\mathrm{T}})^{\mathrm{T}}$, $\tilde{b}_j^{(1)} = A_j b_j^{(1)}$, and A_j is the $s_n \times s_n$ diagonal matrix with $A_j(k,k) = w_{jk}^{1/2}$. Then, the minimization of (4) is equivalent to minimizing

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

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

THEOREM 1. *If $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed subGaussian random variables, then under conditions (A1)–(A7) and (B1)–(B5), 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 $\|\check{\eta} - \tilde{\eta}_0\| = O_p[\{(q_n + s_n)/n\}^{1/2}]$ and $pr(\hat{\gamma}_2 = 0, \check{b}^{(1)} = 0, j = g+1, \ldots, d) \to 1$.*

The estimation consistency result is expressed in terms of $\tilde{b}^{(1)}$, not the original parameter $b^{(1)} = (b_1^{(1)'}, \ldots, b_d^{(1)'})'$. For estimation, given $\hat{b}^{(1)} = A_j^{-1} \check{b}^{(1)}$, it is easy to deduce that $\|\hat{\beta}_j - \beta_{j0}\|_{L^2}^2 = O_p\{s_n^a (q_n + s_n)/n\}$, 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 for 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 $z_i{}^{(1)} = (z_{i1}, ..., z_{iq_n})^{\mathrm{T}}$.

THEOREM 2. *If $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed subGaussian random variables and $q_n = o(n^{1/3})$, then under conditions (A1)–(A7) and (B1)–(B5), 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, where $\sigma^2 = var(\epsilon)$, $H^* = \lim_{n\to\infty} A_n \Sigma_1 A_n{}^{\mathrm{T}}$, $B^* = \lim_{n\to\infty} A_n B_n A_n^{\mathrm{T}}$ with*

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

where $\Xi_j = (\Xi_{j1}, ..., \Xi_{jq_n})^{\mathrm{T}}$, $E\{X_j(t) Z_l\} = \Xi_{jl}(t)$, 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

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

with $d = 4$ functional predictors, 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 $\phi_{2\ell} = 2^{-1/2}\sin\{(2\ell-1)\pi t\}$, $\ell = 1, \ldots, 25$, and $t \in T = [0, 1]$. 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 mean zero, 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 $\{\tilde{\xi}_{ijk}, i = 1, \ldots, n\}$ follow independent and identically distributed $N(0, 16k^{-2})$ for different i and j. The four functional predictors are then defined through the linear transformations

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

Here the first three functional predictors are correlated with each other. To be more realistic, we allow a moderate correlation between V_{i1} and z_i for $i = 1, \ldots, n$, by setting that $\tilde{\xi} = (\tilde{\xi}_{i11}, \tilde{\xi}_{i12}, \tilde{\xi}_{i13}, \tilde{\xi}_{i14})^{\mathrm{T}}$ and $z_i = (z_{i1}, \ldots, z_{ip_n})^{\mathrm{T}}$ have 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 they are realizations of $\{x_{ij}(\cdot), j = 1, 2, 3, 4\}$ at 100 equally spaced times $\{t_{ijl}, l = 1, \ldots, 100\} \in T$ 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, and the functional mean squared error $\mathrm{MSE}_f = \sum_{j=1}^{d} E(\|\hat{\beta}_j - \beta_j\|^2_{L^2})$. For the scalar covariates, we report the Monte Carlo averages for the numbers of false nonzero and false zero scalar covariates, and 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 $\mathrm{PE} = N^{-1}\sum_{i=1}^{N}(y_i^* - \hat{y}_i^*)^2 - \sigma^2$, where $\{x_{ij}^*, z_i^*, y_i^*, j = 1, \ldots, 4\}$ are the testing data generated from the same model, and the predictions are $\hat{y}_i^* = \sum_j \sum_k \hat{\xi}_{ijk}^* \hat{b}_{jk} + z_i^{*\mathrm{T}}\hat{\gamma}$ by plugging in 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 $j = 1, 2$, $b_{j1} = 1$, $b_{j2} = 0{\cdot}8$, $b_{j3} = 0{\cdot}6$, $b_{j4} = 0{\cdot}5$ and $b_{jk} = 8(k-2)^{-4}$ $(k = 5, \ldots, 50)$, $\beta_3 = \beta_4 = 0$, and $\gamma = (1_5^{\mathrm{T}}, 0_{15}^{\mathrm{T}})^{\mathrm{T}}$. Thus $p_n = 20$, $q_n = 5$. To illustrate the impact of the choice of s_n, we inspect the results for s_n ranging from 1 to 16 with λ_n chosen by BIC in Table 1. 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 square error improves until s_n reaches an optimal level, then deteriorates as s_n continues to increase. For the scalar covariates, the mean square error and prediction error appear more stable after 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 optimal mean square and prediction errors, selecting an average $\hat{s}_n = 4{\cdot}30$ with standard error $0{\cdot}050$. 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 with ultra-high dimension of scalar covariates $\gamma = (1_5^{\mathrm{T}}, 0_{995}^{\mathrm{T}})^{\mathrm{T}}$ with $p_n = 1000$, and other settings 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 square error and prediction error in Step 1 become larger than those in Design I, mainly due to the ultra-high number of insignificant scalar covariates. The

functional mean square error is also higher, as the correlation between functional predictors and scalar covariates becomes greater for a 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 summarized in Step 2 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 ultra-high, in a similar spirit to Fan et al. (2012). A 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.

Design	s_n	FZ_f	FN_f	MSE_f	FZ_s	FN_s	MSE_s	PE
I $p_n = 20$	1	0·95	0	4·1 (0·03)	0·39	7·1	4·7 (0·15)	27·9 (0·6)
	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)
	ABIC	$\hat{s}_n$=4·30 (0·050)						
		0	0	0·13 (0·007)	0	0·34	0·067 (0·004)	0·20 (0·006)
	TUNE s_{nj}	$\hat{s}_{n1}$=4·78 (0·075), $\hat{s}_{n2}$=4·87 (0·071)						
		0	0	0·09 (0·003)	0	0·28	0·065 (0·004)	0·18 (0·006)
II $p_n = 1000$		$\hat{s}_n$ =4·07 (0·034)						
	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)
	TUNE s_{nj}	$\hat{s}_{n1}$=4·76 (0·066), $\hat{s}_{n2}$ =4·62 (0·055)						
		0	0	0·076 (0·003)	0	0·09	0·046 (0·003)	0·16 (0·004)

Table 1. *Simulation results with sample size $n = 200$ based on 200 Monte Carlo replicates for Designs I and II. Shown are the Monte Carlo averages with standard errors in parentheses for the number of false zero functional predictors (FZ_f), the number of the false nonzero functional predictors (FN_f), the functional mean squared error (MSE_f), the number of false zero scalar covariates (FZ_s), the number of false nonzero scalar covariates (FN_s), the scalar mean squared error (MSE_s), and the prediction error (PE). We first use* ABIC *to choose the tuning parameter λ_n and a common truncation s_n, 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 contains the improved results by fitting the penalized procedure to the selected model in Step 1 with an additional sample of $n = 200$.*

5. APPLICATION

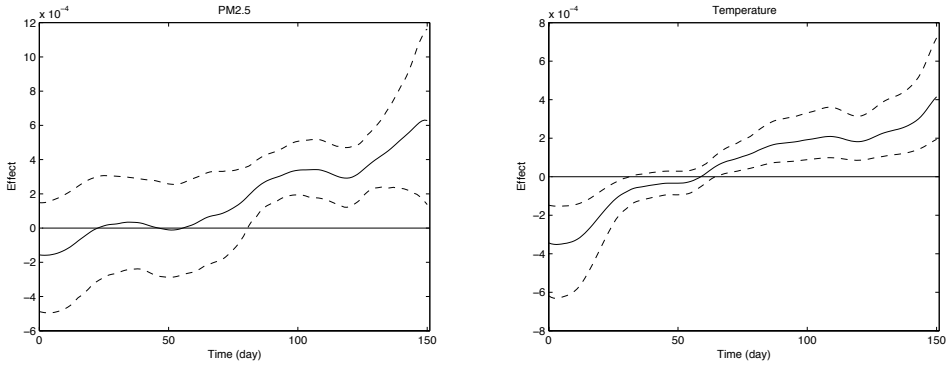
We applied our method to a dataset from the National Mortality, Morbidity, and Air Pollution Study that contains air pollution measurements and mortality counts for U.S. cities with the U.S. census information for year 2000. A main goal of the study is to investigate the impact of air pollution on the non-accidental mortality rate across different cities in the United States, when we take into account climate patterns and information from the U.S. census. Previous studies conducted a two-stage analysis: first modelling the short-term effect of certain air pollutants on the mortality count for each city, then combining the estimates across cities (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 of particulate matter with an aerodynamic diameter of less than $2{\cdot}5\mu m$, abbreviated as PM$2{\cdot}5$ and measured in $\mu g m^{-3}$, as

its negative effect on health, revealed by recent toxicological and epidemiological studies, has brought it to the public's attention. Other studies (Samoli et al., 2013; Pascal et al., 2014) have shown that PM2·5 has a larger effect on mortality in warm weather, so we focused on the daily concentration measurements of PM2·5 from April 1, 2000 to August 31, 2000, which along with the daily observations of temperature and humidity were treated as the functional predictors. After removing the cities which have more than ten consecutive missing measurements of PM2·5, we included a total of 69 cities in our analysis. The response of interest is the log-transformed total non-accidental mortality rate in the following month, September 2000, of the population of age 65 and older, which accounts for the majority of non-accidental deaths. The scalar covariates available from the U.S. census for each city are land area per individual, water area per individual, proportion of the 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 poverty line, and proportion of household owners.

The ABIC was used to first choose significant predictors with a common truncation, 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 standard error $0{\cdot}41$, indicating that household owners tend to incur a lower mortality rate; the standard error was based on 1000 bootstrap samples by fitting the selected model using ordinary least squares. Our method also selected two significant functional predictors, PM2·5 and temperature. The least squares refitting chose the truncation numbers $\hat{s}_{n1} = 2$ and $\hat{s}_{n2} = 2$. The estimated regression parameter functions with their 95% bootstrap confidence bands are shown in Figure 1. We observe that higher PM2·5 concentrations in the summer, especially in July and August, can lead to an increased mortality during the period immediately following. This coincides with the findings in Samoli et al. (2013) and Pascal et al. (2014), but needs to be interpreted with caution, as the effect might 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 Curriero et al. (2002). To better understand the effects of functional predictors, we fitted a linear regression using only the selected scalar covariate, giving $R^2 = 0{\cdot}15$. Including temperature leads to $R^2 = 0{\cdot}25$, and including both temperature and PM2·5 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 $0{\cdot}01$, and adding additional two principal component scores of PM2·5 gives a p-value of $0{\cdot}0008$. For comparison, we also fitted the marginal models containing only PM2·5 or temperature using classical functional linear regression. The marginal F tests for temperature and PM2·5 yield p-values of $0{\cdot}0001$ and $0{\cdot}004$, respectively. The regression parameter functions show similar patterns and are omitted. We conclude that, after adjusting for temperature and household ownership, summer PM2·5 concentrations have a significant impact on the near-future mortality rate of elder citizens in the U.S..

6. POTENTIAL EXTENSIONS

We conclude the paper by discussing two extensions suggested by the reviewers. The first is to consider a partially functional linear regression model when the number of functional predictors is also diverging, $d_n \to \infty$. Since each functional predictor corresponds to a group of principal component scores, the discrepancies from estimating diverging groups of principal components will be increased to a higher order of magnitude, posing additional theoretical challenges. The computation algorithm also needs to be modified or developed, especially if d_n is much larger than n.



(a) (b)

Fig. 1. The estimated regression parameter functions and 95% confidence bands based on 1000 bootstrap samples for PM2$\cdot$5 and temperature with $\hat{s}_{n1} = 2$ and $\hat{s}_{n2} = 2$, respectively. The solid line denotes the estimated regression parameter function and the dashed lines denote the 95% bootstrap confidence bands. The left and right panels are for the PM2$\cdot$5 and temperature, respectively.

Another extension concerns generalized responses y_i. For instance, with a link function $g(\cdot)$ and a variance function $V(\cdot)$, the generalized partially functional linear regression is

$$\mu_i = E(y_i \mid x_i, z_i) = g^{-1}\Big\{ \sum_{j=1}^{d} \int_T x_{ij}(t)\beta_j(t)dt + z_i^{\mathrm{T}}\gamma \Big\}, \;\; \mathrm{var}(y_i \mid x_i, z_i) = \phi V(\mu_i),$$

where ϕ is an unknown scale parameter. An immediate application is to a binary response for the purpose of classification. Model selection and estimation for generalized partially functional linear regression demand new algorithms and theoretical guarantees, which are beyond the scope of this paper.

ACKNOWLEDGEMENTS

This research is partially supported by National Science Foundation, National Institute of Health, and Natural Science and Engineering Research Council of Canada. The authors thank the editor, the associate editor and three referees for their helpful comments and suggestions.

SUPPLEMENTARY MATERIAL

Supplementary material available online includes additional simulation results, regularity conditions, auxiliary results and proofs.

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 $z_i = (z_{i1}, ..., z_{ip_n})^{\mathrm{T}}$. 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

$$\text{argmin}_\eta\{(2n)^{-1}||Y_M - N\eta||^2 + \sum_{r=1}^{R} J_{\lambda_r}\big(\|\eta_r\|\big)\},$$

where $R = d + p_n$. The tuning parameter $\lambda_r = \lambda_{rn}$ 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 square solution by $\hat{\eta}^{(0)}$, and we solve $\hat{\eta}^{(1)} = \text{argmin}_\eta\{(2n)^{-1}||Y_M - N\eta||^2 + \sum_{r=1}^{R} J'_{\lambda_r}\big(\|\eta_r^{(0)}\|\big)\|\eta_r\|\}$. Since some of the $J'_{\lambda_r}\big(\|\eta_r^{(0)}\|\big)$ are zero, we use similar algorithm proposed by Zou & Li (2008). Denote $V = \{r : J'_{\lambda_r}\big(\|\eta_r^{(0)}\|\big) = 0\}$, $W = \{r : J'_{\lambda_r}\big(\|\eta_r^{(0)}\|\big) > 0\}$, $N = (N_V, N_W)$ and $\eta^{(1)} = (\eta_V^{(1)\mathrm{T}}, \eta_W^{(1)\mathrm{T}})^{\mathrm{T}}$. Our algorithm is as follows:

1. Reparameterize the response vector by $Y_M^* = N\eta^{(0)}$, and reparameterize the observed data matrix by $N_r^* = N_r K_r^{1/2} / J'_{\lambda_r}\big(\|\eta_r^{(0)}\|\big)$ for $r \in W$; and $N_r^* = N_r$ for $r \in V$.
2. 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 $N_W^{**} = N_W^* - P_V N_W^*$.
3. Find $\hat{\eta}_W^* = \text{argmin}_\beta\{(2n)^{-1}||Y_M^{**} - N_W^{**}\eta||^2 + \sum_{r\in W} K_r^{1/2}\|\eta_r\|\}$.
4. Compute $\hat{\eta}_V^* = (N_V^{*\mathrm{T}}N_V^*)^{-1}N_V^{*\mathrm{T}}(Y_M^* - N_W^*\hat{\eta}_W^*)$.
5. Let $\eta^{(0)} = \eta^{(1)}$, where $\eta_V^{(1)} = \hat{\eta}_V^*$, and $\eta_r^{(1)} = \hat{\eta}_r^* K_r^{1/2} / J'_{\lambda_r}\big(\|\eta_r^{(0)}\|\big)$ for $r \in W$.

We repeat steps 1–5 until convergence; the final $\eta^{(0)}$ is the regularized estimator. Step 3 essentially solves a group lasso, so we adopt the shooting algorithm of Fu (1998) and Yuan & Lin (2006).

For $p_n \gg n$, it is likely that $N_V^{*\mathrm{T}}N_V^*$ in step 4 is singular, so the local linear approximation algorithm is inapplicable. We modify the concave convex procedure used in Kim et al. (2008). Let $\tilde{J}_{\lambda_r}\big(\|\eta_r\|\big) = J_{\lambda_r}\big(\|\eta_r\|\big) - \lambda_r\|\eta_r\|$ for each r. The convex part of the objective function is $C_{\text{vex}}(\eta) = (2n)^{-1}||Y_M - N\eta||^2 + \sum_{r=1}^{J}\lambda_r\|\eta_r\|$, and the concave part is $C_{\text{cav}}(\eta) = \sum_{r=1}^{J}\tilde{J}_{\lambda_r}\big(\|\eta_r\|\big)$. Begin with the initial estimator $\eta^{(0)} = 0$ and iteratively update the solution until convergence,

$$\eta^{(1)} = \text{argmin}_\eta\{C_{\text{vex}}(\eta) + \nabla C_{\text{cav}}(\eta^{(0)})^{\mathrm{T}}\eta\} = \text{argmin}_\eta\{M(\eta \mid \eta^{(0)}) + \sum_{r=1}^{R}\lambda_r\|\eta_r\|\},$$

where $M(\eta \mid \eta^{(0)}) = (2n)^{-1}||Y_M - N\eta||^2 + \nabla C_{\text{cav}}(\eta^{(0)})^{\mathrm{T}}\eta$ is quadratic in η. The proximal gradient method of Parikh & Boyd (2013) is adopted to solve the above minimization problem.

REFERENCES

CAI, T. T. & HALL, P. (2006). Prediction in functional linear regression. *Annals of Statistics* **34**, 2159–2179.

CARDOT, H., FERRATY, F., MAS, A. & SARDA, P. (2003). Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics* **30**, 241–255.

CARDOT, H. & SARDA, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis* **92**, 24–41.

CUEVAS, A., FEBRERO, M. & FRAIMAN, R. (2002). Linear functional regression: the case of fixed design and functional response. *Canadian Journal of Statistics* **30**, 285–300.

CURRIERO, F. C., HEINER, K. S., SAMET, J. M., ZEGER, S. L., STRUG, L. & PATZ, J. A. (2002). Temperature and mortality in 11 cities of the eastern united states. *American Journal of Epidemiology* **155**, 80–87.

ESCABIAS, M., AGUILERA, A. M. & VALDERRAMA, M. J. (2004). Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics* **16**, 365–384.

FAN, J., GUO, S. & HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society, Series B* **74**, 37–65.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

FAN, J., YAO, Q. & CAI, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society, Series B* **65**, 57–80.

FAN, J. & ZHANG, J.-T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society, Series B* **62**, 303–322.

FU, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.

HALL, P. & HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Annals of Statistics* **35**, 70–91.

HALL, P., MÜLLER, H. G. & WANG, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics* **34**, 1493–1517.

KIM, Y., CHOI, H. & OH, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association* **103**, 1665–1673.

LEI, J. (2014). Adaptive global testing for functional linear models. *Journal of the American Statistical Association* **109**, 624–634.

LI, Y., WANG, N. & CARROLL, R. J. (2010). Generalized functional linear models with semiparametric single-index interactions. *Journal of the American Statistical Association* **105**, 621–633.

LU, Y., DU, J. & SUN, Z. (2014). Functional partially linear quantile regression model. *Metrika* **77**, 317–332.

MORRIS, J. S., VANNUCCI, M., BROWN, P. J. & CARROLL, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association* **98**, 573–597.

MÜLLER, H.-G. & STADTMÜLLER, U. (2005). Generalized functional linear models. *Annals of Statistics* **33**, 774–805.

MÜLLER, H.-G. & YAO, F. (2008). Functional additive models. *Journal of the American Statistical Association* **103**, 1534–1544.

PARIKH, N. & BOYD, S. (2013). Proximal algorithms. *Foundations and Trends in Optimization* , 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. *Atmospheric Environment* **95**, 175–184.

PENG, R. D., DOMINICI, F. & LOUIS, T. A. (2006). Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society, Series A* **169**, 179–203.

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. *American Journal of Epidemiology* **161**, 585–594.

RAMSAY, J. O. & DALZELL, C. J. (1991). Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society, Series B* **53**, 539–572.

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. *Journal of the Royal Statistical Society, Series B* **53**, 233–243.

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. *Environmental Health Perspectives* **121**, 932.

SHIN, H. (2009). Partial functional linear regression. *Journal of Statistical Planning and Inference* **139**, 3405–3418.

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

WANG, H., LI, R. & TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **93**, 553–568.

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. *Journal of the American Statistical Association* **100**, 577–590.

YAO, F., MÜLLER, H.-G. & WANG, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *Annals of Statistics* **33**, 2873–2903.

YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

ZHANG, J.-T. & CHEN, J. (2007). Statistical inferences for functional data. *Annals of Statistics* **35**, 1052–1079.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

ZOU, H. & LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36**, 1509–1533.

[*Received January* $20XX$*. Revised June* $20XX$]