Statistica Sinica **23** (2013), 51-74
doi:http://dx.doi.org/10.5705/ss.2011.160

SHRINKAGE ESTIMATION AND SELECTION FOR MULTIPLE FUNCTIONAL REGRESSION

Heng Lian

Nanyang Technological University

Abstract: Functional linear regression is a useful extension of simple linear regression and has been investigated by many researchers. However, the functional variable selection problem when multiple functional observations exist, which is the counterpart in the functional context of multiple linear regression, is seldom studied. Here we propose a method using a group smoothly clipped absolute deviation penalty (gSCAD) which can perform regression estimation and variable selection simultaneously. We show the method can identify the true model consistently, and discuss construction of pointwise confidence intervals for the estimated functional coefficients. Our methodology and theory is verified by simulation studies as well as some applications to data.

Key words and phrases: Estimation consistency, functional linear regression, group SCAD, principal component analysis, selection consistency.

1. Introduction

In several applications, functional data appear as the basic unit of observations. Classical regression models may be inadequate for such cases because of the high correlations of the discretized data. Compared with discrete multivariate analysis, functional analysis takes into account the smoothness of the high dimensional covariates, and often suggests new approaches to the problems that have not been discovered before. Some recent developments in functional regression include Yao, Müller, and Wang (2005); Cai and Hall (2006); Crambes, Kneip, and Sarda (2009); Yuan and Cai (2010).

The literature contains an impressive range of functional analysis tools for various problems. The traditional approach, carefully documented in the monograph Ramsay and Silverman (2005), starts by representing functional data by an expansion with respect to a certain basis, and subsequent inferences are carried out on the coefficients. A line of work by the French school, taking a nonparametric point of view, extends traditional nonparametric techniques, most notably the kernel estimate, to the functional case (Ferraty and Vieu (2006)). Other methods, such as putting functional regression in the reproducing kernel Hilbert space framework, have been developed (Preda (2007); Lian (2007)).

In this paper, we are concerned with an extension of the simple functional linear regression model to the case where multiple functional observations are made on each unit. Formally, the model we consider is

$$Y_i = a + \sum_{j=1}^{p} \int_0^1 \beta_j(t) X_{ij}(t)\, dt + \epsilon_i, 1 \le i \le n, \tag{1.1}$$

where X_{ij} are random functions, a is the intercept, ϵ_i are random scalar errors and the functional coefficients $\beta_j, 1 \le j \le p$, are the objects of interest in the model.

Because functional coefficients are more complicated objects than the scalar coefficients in classical multiple linear regression, it is generally desirable to identify the significant variables in predicting the responses, even if p is small. For example, Zhu, Vannucci, and Cox (2010) investigated fluorescence spectroscopy for cervical precancer diagnosis, using a Bayesian model to select from multiple fluorescence spectra for classification of subjects.

In a non-Bayesian context, traditional methods for variable selection in classical linear models include constructing hypothesis tests or using information criteria. More recently, regularization methods have received much attention. For standard linear regression, the Lasso (Tibshirani (1996)) is probably the most popular method, using an L_1 penalty to force some of the coefficients to zero. Several subsequent works (Meinshausen and Buhlmann (2006); Zhao and Yu (2006)) have shown that Lasso is in general not consistent for model selection unless some nontrivial conditions on the covariates are satisfied. To address such shortcomings, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty and Zou (2006) proposed the adaptive lasso in the fixed p case using a weighted L_1 penalty with weights determined by an initial estimator. There are many other extensions of the regularization framework for variable selection (Yuan and Lin (2006); Wang and Leng (2007); Huang, Horowitz, and Ma (2008)).

In this article we use functional principal component analysis (PCA)-based estimation method (Cardot, Ferraty, and Sarda (1999); Hall and Horowitz (2007)) combined with group SCAD (this terminology seems to have first appeared in Wang, Chen, and Li (2007) for varying coefficients variable selection) which represents a new application of the SCAD penalty. The regularization method for variable selection in nonparametric settings has been developed in the context of smoothing spline ANOVA for nonparametric regression smoother (Lin and Zhang (2006)) and support vector machines (Zhang (2006)). For varying-coefficient models, we are aware of the work (Wang, Li, and Huang (2008)) where the authors used the basis expansion approach for estimation combined with the group

SCAD penalty on coefficients, and the work (Wang and Xia (2009)) where the group Lasso penalty was applied directly to smooth functions evaluated at sampled points.

The rest of the article is organized as follows. We describe the functional PCA and shrinkage estimation procedure in Section 2.1, and present estimation consistency and selection consistency results in Section 2.2. We discuss estimation and inference algorithms and tuning parameter selection in Sections 2.3 and 2.4, respectively. In Section 3, we present some simulation experiments and illustrate the proposed method using a spectrometrics example and a weather example. The technical details are in the Appendix.

2. Methodology and Theoretical Properties

2.1. Estimation of multiple functional regression

Suppose we have independent and identically distributed (i.i.d.) observations $(X_{i1}, \ldots, X_{ip}, Y_i)$, $1 \le i \le n$, where X_{ij} is a square integrable random function on the interval $[0, 1]$ with mean μ_j. The response variables Y_i are generated as (1.1), with i.i.d. errors ϵ_i having finite second moments. We take the errors to be independent of the predictors. We use $(X_1, \ldots, X_p, Y)$ to denote generic random variables with distribution the same as $(X_{i1}, \ldots, X_{ip}, Y_i)$. Let $S_j(s,t) = Cov\{X_j(s), X_j(t)\}$, so by Mercer's Theorem we have the spectral expansion

$$S_j(s,t) = \sum_{k=1}^{\infty} \lambda_{jk} \phi_{jk}(s) \phi_{jk}(t),$$

where $\lambda_{j1} > \lambda_{j2} > \cdots > 0$ are the eigenvalues of the linear operator associated with $S_j(s,t)$ with corresponding eigenfunctions ϕ_{jk}. We assume the eigenvalues have multiplicity one so that the eigenvectors are all identified. With $\hat{S}_j(s,t) = (1/n)\sum_{i=1}^{n}(X_{ij}(s) - \bar{X}_j(s))(X_{ij}(t) - \bar{X}_j(t))$, where $\bar{X}_j = \sum_i X_{ij}/n$, we have the empirical counterpart of this expansion as

$$\hat{S}_j(s,t) = \sum_{k=1}^{\infty} \hat{\lambda}_{jk} \hat{\phi}_{jk}(s) \hat{\phi}_{jk}(t),$$

where $\hat{\lambda}_{j1} \ge \hat{\lambda}_{j2} \ge \cdots \ge 0$. To get rid of uncertainty of signs, we assume $\int \hat{\phi}_j \phi_j \ge 0$. For the empirical operator $\hat{S}_j$, at most n eigenvalues are strictly positive.

In general, different functional predictors are not independent of each other. The Karhunen-Loève expansion of the random function X_{ij} in terms of the orthonormal basis $\phi_{jk}(t)$ is

$$X_{ij} - \mu_j = \sum_{k=1}^{\infty} \xi_{ijk} \phi_{jk}, \tag{2.1}$$

where the ξ_{ijk} are principal component scores satisfying $E\xi_{ijk} = 0, E\xi_{ijk}^2 = \lambda_{jk}$ and $E\xi_{ijk}\xi_{ijk'} = 0, k \neq k'$. Thus from (2.1) we have the covariance operator expansion

$$Cov\{X_{j_1}(s), X_{j_2}(t)\} = \sum_{k_1,k_2=1}^{\infty} \lambda_{k_1,k_2}^{j_1,j_2} \phi_{j_1k_1}(s)\phi_{j_2k_2}(t),$$

where $\lambda_{k_1,k_2}^{j_1,j_2} = E\xi_{j_1k_1}\xi_{j_2k_2}$ determines the dependency structure between different predictors. Note that with our notation, when $j_1 = j_2 = j$, $\lambda_{k_1,k_2}^{j,j} = 0$ if $k_1 \neq k_2$ and $\lambda_{k_1,k_2}^{j,j} = \lambda_{jk}$ if $k_1 = k_2 = k$. An illustration of how this dependency could arise is given in the next subsection.

The model (1.1) can be equivalently written as

$$Y_i - \mu = \sum_{j=1}^{p} \int \beta_j(X_{ij} - \mu_j) + \epsilon_i, \tag{2.2}$$

where $\mu = E[Y|X_1, \ldots, X_p] = a + \sum_j \int \beta_j\mu_j$. After β_j is estimated by $\hat{\beta}_j$, say, the intercept a can be estimated by $\hat{a} = \bar{Y} - \sum_j \int \hat{\beta}_j\bar{X}_j$, where $\bar{Y} = \sum_i Y_i/n$.

Now we consider the problem of estimating β_j. Using the orthonormal basis $\{\phi_{jk}\}$, (2.2) can be equivalently written as

$$Y_i - \mu = \sum_{j=1}^{p}\sum_{k=1}^{\infty} \xi_{ijk}b_{jk} + \epsilon_i,$$

making use of the expansion $\beta_j = \sum_k b_{jk}\phi_{jk}$. This suggests the estimator

$$\{\hat{b}_{jk}\} = \arg\min \sum_{i=1}^{n}(Y_i - \bar{Y} - \sum_{j=1}^{p}\sum_{k=1}^{K_j} \hat{\xi}_{ijk}b_{jk})^2,$$

and then $\hat{\beta}_j = \sum_{k=1}^{K_j} \hat{b}_{jk}\hat{\phi}_{jk}$ where, in the above displayed equation, $\hat{\xi}_{ijk} = \int (X_{ij} - \bar{X}_j)\hat{\phi}_{jk}$ is the principal component score estimated from data. Here the truncation point K_j is a smoothing parameter. To further select functional predictors simultaneously, we minimize the criterion function

$$J(b) = \sum_{i=1}^{n}(Y_i - \bar{Y} - \sum_{j=1}^{p}\sum_{k=1}^{K_j} \hat{\xi}_{ijk}b_{jk})^2 + n\sum_{j=1}^{p} p_{\lambda_j}(||b_j||), \tag{2.3}$$

where $||b_j||$ is the l_2 norm of $b_j = (b_{j1}, \ldots, b_{jK_j})^T$. Among many ways to specify the penalty function p_λ, we choose the SCAD penalty function of Fan and Li (2001),

$$p'_\lambda(\theta) = \lambda\left\{I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\right\},\ p_\lambda(0) = 0,$$

for $a = 3.7$ and $\theta > 0$, where $I(\cdot)$ is the indicator function. The choice of $a = 3.7$ is suggested by Fan and Li (2001) and adopted in almost all publications involving the SCAD penalty. The SCAD penalty possesses some desirable properties, such as that it results in a sparse model due to the singularity at zero, that it results in an estimator that is continuous in the observations, and that it is almost unbiased for large parameters since the derivative of the penalty is zero when θ is large. One simple property that we use in the proof is that $|p_\lambda(a) - p_\lambda(b)| \le \lambda|a - b|$ for $a, b > 0$. Other penalty functions such as the adaptive Lasso can also be used here and lead to similar consistency results as found below.

We choose to penalize $\|b_j\|$, which is probably the simplest one to use. However, as suggested by a reviewer, there are alternatives such as penalizing $(\sum_{k=1}^{K_j} \lambda_{jk} b_{jk}^2)^{1/2}$, the standard deviation of $\int \beta_j(X_j - \mu_j)$, and this seems quite natural in the functional context. Theoretically, we expect similar consistency can be obtained. Although we did not investigate this alternative, we expect the finite sample results to be similar to those found since the SCAD penalty results in almost unbiased estimation.

2.2. Consistency properties

Large sample properties of shrinkage estimation with the SCAD penalty have been established in the literature (Fan and Li (2001); Fan and Peng (2004); Wang, Li, and Huang (2008)). We show that, in our context, the estimation procedure can consistently estimate the functional coefficients as well as consistently identify the true model. However, extending these theoretical results to multiple functional regression is not trivial. Note that in criterion (2.3) two types of approximations are involved, one is the truncation of β_j to approximate the functional coefficients, the other is the unknown covariate ξ_{ijk} estimated by $\hat{\xi}_{ijk}$. While the former approximation is typical in nonparametric problems such as Wang, Li, and Huang (2008), the latter is unique to the functional regression problem. It also resembles the measurement error model in form where the covariates are not observed directly (Liang and Li (2009); Carroll, Delaigle, and Hall (2009)).

Without loss of generality, we denote the true regression coefficients by $\beta = ((\beta^{(1)})^T, (\beta^{(2)})^T)^T$, with $\beta^{(1)} = (\beta_1, \ldots, \beta_s)^T, s \le p$, containing all nonvanishing components of β and $\beta_{s+1} = \cdots = \beta_p \equiv 0$. Let Λ be the $(\sum_j K_j) \times (\sum_j K_j)$ matrix

$$\begin{pmatrix} \Lambda^{1,1} & \cdots & \Lambda^{1,p} \\ \vdots & \vdots & \vdots \\ \Lambda^{p,1} & \cdots & \Lambda^{p,p} \end{pmatrix}, \tag{2.4}$$

where Λ^{j_1,j_2} is the $K_{j_1} \times K_{j_2}$ matrix with entries $\lambda_{k_1,k_2}^{j_1,j_2}, 1 \le k_1 \le K_{j_1}, 1 \le k_2 \le K_{j_2}$. In our results, the following regularity conditions are needed.

(c1) $\int E(X_{ij}^4) < \infty$ and $E\epsilon_i^4 < \infty$.

(c2) All eigenvalues of S_j have multiplicity one; $\lambda_{jk} - \lambda_{j\,k+1} \geq C^{-1}k^{-\alpha_j - 1}$, $b_{jk} \leq Ck^{-\beta_j}$, $\alpha_j > 1, \beta_j > \alpha_j + 1/2$.

(c3) $\bar{K}^{4\bar{\alpha}+4}/n \to 0$, $\lambda_j = o(\bar{K}^{-\bar{\alpha}})$, $\bar{K}^{3\bar{\alpha}+3}/n = o(\lambda_j^2)$, and $\bar{K}^{\bar{\alpha}}\underline{K}^{-\underline{\alpha}-2\underline{\beta}+1} = o(\lambda_j^2)$, where $\bar{K} = \max_j\{K_j\}$, $\underline{K} = \min_j\{K_j\}$, and $\bar{\alpha}$, $\underline{\alpha}$, $\bar{\beta}$, $\underline{\beta}$ are similarly defined.

(c4) The minimum eigenvalue of Λ, $\rho_{\min}(\Lambda)$, is of order $\Omega(\bar{K}^{-\bar{\alpha}})$ where $a_n = \Omega(b_n)$ means $b_n = O(a_n)$.

Remark 1. We implicitly assume that it is possible to choose $\{K_j\}$ and $\{\lambda_j\}$ so that (c3) is satisfied. If all $\alpha_j \equiv \alpha$, $\beta_j \equiv \beta$, with $\beta > \alpha + 1/2$, then it is easily checked that we can do so.

Remark 2. When all $\alpha_j \equiv \alpha$ and $\beta_j \equiv \beta$, $\beta > \alpha + 1/2$ requires that β_j be sufficiently smooth relative to S_j. As β increases and α decreases, the condition (c3) becomes easier to satisfy.

Remark 3. Although we use different λ_j for each predictor in theory, in practice we fix λ_j to be the same. We use different K_j, since we find this leads to slightly better results; allowing all λ_j to be estimated from data using our method does not improve our simulation results.

Theorem 1. *Under* (c1)−(c4)*, we have*

(a) (*Estimation consistency*) $||\hat{\beta}_j - \beta_j|| = o_p(1), 1 \leq j \leq p$.

(b) (*Selection consistency*) $\hat{\beta}_{s+1} = \ldots = \hat{\beta}_p \equiv 0$ *with probability converging to* 1.

Remark 4. The proof of Theorem 1 in the Appendix actually shows that $||\hat{\beta}_j - \beta_j|| = O_p(\bar{K}^{3\bar{\alpha}+3}/n + \bar{K}^{\bar{\alpha}}\underline{K}^{-\underline{\alpha}-2\underline{\beta}+1})$. We do not believe this rate to be optimal, that some more complicated arguments could get better rates.

An illustration. Let $p = 2$. Suppose the eigenvalues of S_1 and S_2 satisfy $\lambda_{jk} = Ck^{-\alpha}, j = 1, 2$. If X_1 and X_2 are independent, then the matrix Λ at (2.4) is diagonal and its minimum eigenvalue is of order $\Omega(K^{-\alpha})$. In general, Λ can be written as a block matrix

$$\Lambda = \begin{pmatrix} E & F \\ F^T & G \end{pmatrix},$$

where E and G are $K \times K$ diagonal matrices containing the eigenvalues of S_1 and S_2 respectively. It is easy to see that the minimum eigenvalue of Λ is no bigger than $CK^{-\alpha}$, since Λ is similar to

$$\tilde{\Lambda} = \begin{pmatrix} E & 0 \\ 0 & G - F^T E^{-1} F \end{pmatrix},$$

and the eigenvalues of $G - F^T E^{-1} F$ are dominated by those of G. In assumption (c4), we assume that the minimum eigenvalue of Λ is still of order $K^{-\alpha}$ as in the independent case. This assumption thus can be thought of as a constraint on the dependence of different predictors. However, we show in the following setup this assumption is quite natural. Suppose the random functions X_1 and X_2 are specified by

$$X_1 = \sum_{j=1}^{l} a_{1j} W_j, X_2 = \sum_{j=1}^{l} a_{2j} W_j, \tag{2.5}$$

where $W_j, 1 \leq j \leq l$ are independent mean zero random functions with Karhunen -Loève expansion given by $W_j = \sum_k \omega_{jk} \phi_k$ (note that we assume the eigenfunctions are common to all W_j) with $E\omega_{jk}^2 = \kappa_{jk} > 0$. We can give a sufficient condition under which $\rho_{\min}(\Lambda) = \Omega(K^{-\alpha})$.

Proposition 1. *Suppose $ck^{-\alpha} \leq \kappa_{jk} \leq Ck^{-\alpha}, j = 1, \ldots, l$, for some constants $C \geq c > 0$. If $\{a_{1j}\}, \{a_{2j}\}$ are two non-proportional sequences, then $\rho_{\min}(\Lambda) = \Omega(K^{-\alpha})$.*

2.3. Computation and inferences

One can express the criterion function $J(b)$ in vector and matrix form. With

$$\hat{Z}_j = \begin{pmatrix} \hat{\xi}_{1j1} & \ldots & \hat{\xi}_{1jK_j} \\ \vdots & \vdots & \vdots \\ \hat{\xi}_{ij1} & \ldots & \hat{\xi}_{ijK_j} \\ \vdots & \vdots & \vdots \\ \hat{\xi}_{nj1} & \ldots & \hat{\xi}_{njK_j} \end{pmatrix},$$

$\hat{Z} = (\hat{Z}_1, \ldots, \hat{Z}_p)$, $b = (b_{11}, \ldots, b_{1K_1}, b_{21}, \ldots, b_{pK_p})^T$, and $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, (2.3) can be written as

$$J(b) = \sum_{i=1}^{n} (\mathbf{Y} - \bar{Y}\mathbf{1} - Zb)^T (\mathbf{Y} - \bar{Y}\mathbf{1} - Zb) + n \sum_{j=1}^{p} p_{\lambda_j}(||b_j||), \tag{2.6}$$

where $\mathbf{1}$ is the n-dimensional vector with all components one.

We use the local quadratic approximation idea (Fan and Li (2001)) to optimize the criterion. Specifically, if $\hat{b}^{(m)}$ is the estimate obtained in the m-th iteration, then (2.6) can be locally approximated by

$$\sum_{i=1}^{n} (\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b)^T (\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b) + \frac{n}{2} b^T R(\hat{b}^{(m)}) b, \tag{2.7}$$

where $R(\hat{b}^{(m)}) = diag\{(p'_{\lambda_1}(||\hat{b}_1^{(m)}||)/||\hat{b}_1^{(m)}||)I_{K_1}, \ldots, (p'_{\lambda_p}(||\hat{b}_p^{(m)}||)/||\hat{b}_p^{(m)}||)I_{K_p}\}$ and I_{K_j} is the $K_j \times K_j$ identity matrix. The minimizer of (2.7) is then

$$\hat{b}^{(m+1)} = (\hat{Z}^T\hat{Z} + nR(\hat{b}^{(m)}))^{-1}\hat{Z}^T(\mathbf{Y} - \bar{Y}\mathbf{1}).$$

We iterate these steps until convergence and obtain the final estimate $\hat{b}$. During the iterations, if some $||b_j||$ is smaller than a threshold (10^{-5} in our implementation), we set $b_j = 0$ and ignore the corresponding predictor in future iterations.

Now we consider the construction of pointwise confidence intervals for β_j. Following Fan and Li (2001), the sandwich formula can be used as an estimator for the variance of the nonzero components of $\hat{b}$, denoted $\hat{b}^{(1)}$ henceforth. The estimator of asymptotic variance is

$$\widehat{Cov}(\hat{b}^{(1)}) = ((\hat{Z}^{(1)})^T\hat{Z}^{(1)} + nR^{(1)})^{-1}(\hat{Z}^{(1)})^T\widehat{Cov}(Y)\hat{Z}^{(1)}((\hat{Z}^{(1)})^T\hat{Z}^{(1)} + nR^{(1)})^{-1},$$

where $\hat{Z}^{(1)}$ denotes the selected columns of $\hat{Z}$ corresponding to nonvanishing $||b_j||$, $R^{(1)}$ denotes the selected rows and columns of $R(\hat{b})$ in a similar way, and $\widehat{Cov}(Y)$ is the $n \times n$ diagonal matrix with estimated squared residuals on the diagonal. The diagonal blocks of $\widehat{Cov}(\hat{b}^{(1)})$ gives the asymptotic variance for nonvanishing $\hat{b}_j$.

Since $\hat{\beta}_j(t) = \hat{b}_j^T\hat{\phi}_j(t)$, $\hat{\phi}_j(t) = (\hat{\phi}_{j1}(t), \ldots, \hat{\phi}_{jK_j}(t))^T$. We have a natural estimator for the asymptotic variance of $\beta_j(t)$,

$$\widehat{Cov}(\beta_j(t)) = \hat{\phi}_j(t)^T\widehat{Cov}(\hat{b}_j)\hat{\phi}_j(t).$$

Note that here we ignored the uncertainty of $\hat{\phi}_j$ which is also estimated from observations. However, we think this is a reasonable first approximation. Our simulation experiments illustrate the performance of the asymptotic variance formula. Estimates of the asymptotic variance can be used to construct pointwise confidence intervals for $\beta_j(t)$ for nonzero components of the functional coefficients. Strictly speaking, the constructed intervals are for the truncated $\beta_j(t)$ at cutoff K_j in the expansion. Thus the constructed intervals have lower than targeted coverage rate for the variability in $\hat{\phi}_{jk}$ is ignored, and the interval is only for truncated functional coefficients. The bias caused will be seen from our numerical results.

2.4. Tuning parameter selection

For implementation of our method, we need to choose the truncation points $\{K_j\}$ and the regularization parameters $\{\lambda_j\}$ for group SCAD penalty.

We use generalized cross-validation (GCV) to select both K_j and λ_j. GCV can be thought of as a short-cut for leave-one-out cross-validation, and also comes

with advantageous properties (Wahba (1990)). The criterion is defined by

$$GCV(\{K_j\},\{\lambda_j\}) = \frac{1}{n}\frac{||\mathbf{Y}-\bar{Y}\mathbf{1}-\hat{\mathbf{Y}}||^2}{(1-tr(H(\{K_j\},\{\lambda_j\}))/n)^2},$$

where $\hat{\mathbf{Y}} = H(\{K_j\},\{\lambda_j\})(\mathbf{Y}–\bar{Y}\mathbf{1})$ is the fitted response values and $H(\{K_j\},\{\lambda_j\})$ $= \hat{Z}(\hat{Z}^T\hat{Z}+nR(\hat{b}))^{-1}\hat{Z}^T$ is the hat matrix.

Simultaneously choosing all parameters K_j and λ_j is computationally expensive and we use a three-step strategy instead. First we set $K_j = K$ with K determined by GCV while $\lambda_j = 0$. Then we consider each K_j in turn, with others fixed to the current values. GCV is again used to update K_j, and we use the resulting K_j as the final parameter after a complete scan through all p predictors. Finally, a single smoothing parameter $\lambda_j \equiv \lambda$ is chosen using GCV. In our simulations, we find that using different K_j gives better results than constraining them to be the same, but that using different λ_j does not lead to improvement.

3. Numerical Experiments

3.1. Simulation examples

We perform a Monte Carlo experiment to investigate the finite sample performance of the estimation method, using GCV to select the two tuning parameters. The simulated data was generated from (1.1) with $p = 4$ functional predictors, $a = 0$, and the errors ϵ distributed as $N(0,\sigma^2)$. For $1 \le j \le 4$ independently, we take $W_j = \sum_{k=1}^{50}\xi_{jk}\phi_k$, where $\xi_{jk} \sim N(0,k^{-2})$, $\phi_1 \equiv 1$, and $\phi_{k+1} = \sqrt{2}\cos(k\pi t)$ for $k \ge 1$. Then the functional predictors are defined through the linear transformations

$$\begin{aligned}
X_1 &= W_1 + \rho(W_2+W_3),\\
X_2 &= W_2 + \rho(W_1+W_3),\\
X_3 &= W_3 + \rho(W_1+W_2),\\
X_4 &= W_4.
\end{aligned}$$

Note that the scalar ρ controls the strength of dependence between different predictors, with $\rho = 0$ resulting in independent predictors. For β_1 and β_2, in terms of expansion based on $\{\phi_k\}$, we took $b_1 = (-2,1,-2,1)^T$, $b_2 = (1,-1,0.5,-0.5)^T$, and set $\beta_3 = \beta_4 = 0$. We had $n = 100$ for all our simulations, and set $\rho = 0, 0.2, 0.5$, or 0.8, and $\sigma^2 = 0.1$ or 0.3. All integrations required in the generation of the data and the estimation procedure were Riemannian sum approximations with an equally spaced grid containing 500 points on $[0,1]$.

The simulation results are summarized in Table 1 based on 500 runs in each scenario. We report the mean squared errors $||\hat{\beta}-\beta||^2$ using our regularized

multiple functional regression model (MSE), average number of correctly identified nonvanishing coefficients (TP), average number of incorrectly identified nonvanishing coefficients (FP), empirical coverage probability of pointwise 95% confidence interval for β_1 (95% Cov.Prob.1) and empirical coverage probability of pointwise 95% confidence interval for β_2 (95% Cov.Prob.2). For comparison, three other methods were applied and the corresponding errors reported in the same table: functional regression without regularization (NOPENMSE); the oracle for which the true zero coefficients were known and no shrinkage applied (OMSE); functional regression with the group adaptive Lasso applied (ALASSOMSE) (using the group Lasso as the initial estimate). The estimation errors reported for the unregularized method are the sum of errors on β_1 and β_2 only. For each scenario, the empirical coverage probabilities reported are the averages over the grid $(0.1, 0.2, \ldots, 0.9)$ for β_1 and β_2 whenever they are estimated as nonzero coefficients by the group SCAD estimation method.

With the unregularized procedure, we note that occasionally the GCV selected a large K resulting in unstable estimation and large estimation error. In the table, these unusual cases were deleted before computing the errors. However, when using the regularized procedure, we observed that even if the K selected by GCV was too large, the subsequent regularization procedure could still choose enough penalization and produce reasonable estimates, and thus for regularized estimators no cases were omitted.

As one can see from Table 1, the noise level has a significant effect on the estimation errors as well as the average number of truly relevant predictors detected. However, the number of false positives remains at a low level even for larger noise variance. Compared to noise level, the correlation between different predictors seems to have milder effects, except for the relatively high correlation level $\rho = 0.8$. Shrinkage estimation based on both the adaptive Lasso and SCAD penalty outperforms unpenalized estimation. The results also show that confidence intervals based on the sandwich formula for the asymptotic variance work well, with only a small downward bias in our simulations. As an illustration, the true functions β_1 and β_2, as well as their estimates when $\rho = 0.2$ and $\sigma = 0.1$ or 0.3, are plotted in Figure 1.

In Table 2 we give results for the SCAD penalized estimator with truncation points $K_j \equiv K$, K varying from 2 to 7, for the case $\rho = 0.2$ and $\sigma = 0.1$. λ was chosen by the GCV method. The minimum error is achieved when $K = 4$, as expected. When $K < 4$, the model underfits and the errors are much larger; as K increases beyond 4, the results are still reasonable although they get worse as K increases due to overfitting. In Table 3, we show the variable selection results as both K and λ vary (note again we take K to be independent of j for illustration purposes). The numbers in each cell are the TP/FP pairs. We see that larger values of K demand larger values of λ.

Table 1. Simulation results for penalized multiple functional regression with $p = 4$.

Scenario	$\rho = 0.0$ $\sigma = 0.1$	$\rho = 0.2$ $\sigma = 0.1$	$\rho = 0.5$ $\sigma = 0.1$	$\rho = 0.8$ $\sigma = 0.1$	$\rho = 0.0$ $\sigma = 0.3$	$\rho = 0.2$ $\sigma = 0.3$	$\rho = 0.5$ $\sigma = 0.3$	$\rho = 0.8$ $\sigma = 0.3$
NOPENMSE	0.83	0.92	1.21	1.78	1.87	2.75	3.59	5.21
ALASSOMSE	0.63	0.70	0.91	1.34	1.70	2.09	2.74	3.48
SCADMSE	0.71	0.71	0.93	1.30	1.68	2.02	2.66	3.27
OMSE	0.63	0.63	0.85	0.88	1.47	1.81	1.80	2.00
TP	2	2	2	2	1.77	1.81	1.80	1.93
FP	0.08	0.09	0.08	0.64	0.13	0.11	0.15	0.82
95% Cov.Prob.1	0.92	0.92	0.92	0.92	0.93	0.93	0.92	0.92
95% Cov.Prob.2	0.93	0.94	0.94	0.93	0.94	0.94	0.92	0.90

Table 2. MSE for different fixed K.

	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$
SCADMSE	4.22	2.92	0.73	0.80	0.89	1.17

Table 3. Variable Selection results (TP/FP)for different K and λ, with $\rho = 0.2$ and $\sigma = 0.1$.

	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$
$\lambda = 0.01$	1.76/0.64	2/0.67	2/0.82	2/0.89	2/1.03	2/2
$\lambda = 0.02$	1.52/0.27	1.83/0.34	2/0.43	2/0.64	2/1.00	2/2
$\lambda = 0.05$	1.07/0.10	1.72/0.12	2/0.17	2/0.34	2/0.72	2/1.52
$\lambda = 0.1$	0.52/0.04	1.35/0.05	1.73/0.07	1.86/0.25	2/0.51	2/1.03
$\lambda = 0.2$	0/0	0.53/0	1.02/0	1.72/0	1.80/0.33	1.95/0.57

Finally, we have simulation results for $p = 10$ with $s = 2$. The functional predictors were generated as before and β_1, β_2 the same as for $p = 4$. The results are reported in Table 4. Performance is worse than when $p = 4$, but still reasonable. In particular, there is a larger number of false positives especially when correlation is high.

3.2. Spectrometrics data

We illustrate our approach on a spectrometrics dataset that contains 215 spectra of light absorbance for meat samples as functions of wavelength. Because of the denseness of wavelengths at which the measurements were made, subjects are treated as continuous curves. Figure 2 shows the first 50 curves in the dataset. This dataset has been previously used in functional nonparametric regression studies where the covariate is the spectra curve and the response is the percentage of fat content in the piece of meat (Ferraty and Vieu (2002, 2006); Ferraty, Mas, and Vieu (2007)). In nonparametric kernel regression, the choice of semi-metric to define distance between curves is crucial to the performance

Table 4. Simulation results for penalized multiple functional regression with $p = 10$.

Scenario	$\rho = 0.0$ $\sigma = 0.1$	$\rho = 0.2$ $\sigma = 0.1$	$\rho = 0.5$ $\sigma = 0.1$	$\rho = 0.8$ $\sigma = 0.1$	$\rho = 0.0$ $\sigma = 0.3$	$\rho = 0.2$ $\sigma = 0.3$	$\rho = 0.5$ $\sigma = 0.3$	$\rho = 0.8$ $\sigma = 0.3$
NOPENMSE	3.74	3.95	4.88	5.92	5.76	5.82	5.81	6.96
ALASSOMSE	1.23	1.70	1.81	2.95	2.05	2.67	3.72	5.98
SCADMSE	1.06	1.46	1.85	3.10	2.03	2.76	3.52	5.36
OMSE	0.63	0.63	0.85	0.88	1.47	1.81	1.80	2.00
TP	2	2	2	1.73	1.82	1.88	2	2
FP	0.33	0.65	1.67	3.65	0.29	0.82	2.00	4.25
95% Cov.Prob.1	0.89	0.89	0.88	0.79	0.89	0.90	0.86	0.70
95% Cov.Prob.2	0.90	0.91	0.90	0.74	0.90	0.92	0.89	0.68

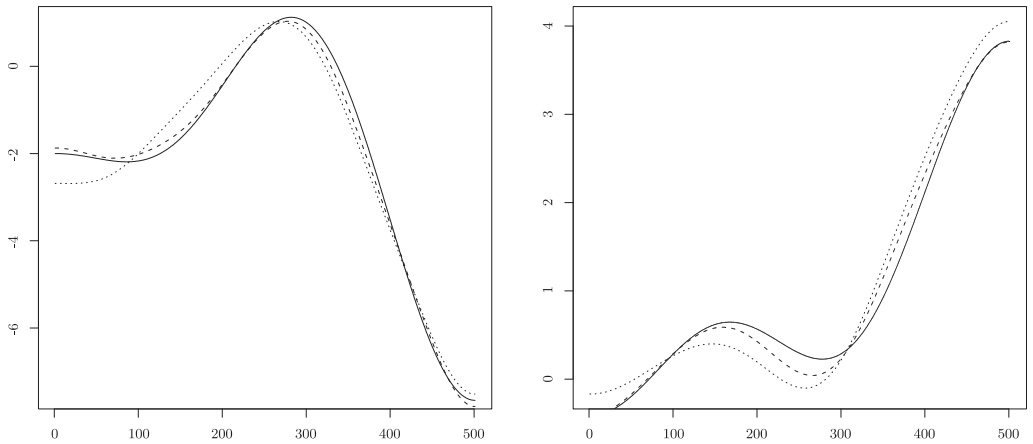

Figure 1. (a) The true coefficient β_1 (solid line) with its estimates when $\sigma = 0.1$ (dashed line) and $\sigma = 0.3$ (dotted line). (b) β_2 (solid line) with its estimates when $\sigma = 0.1$ (dashed line) and $\sigma = 0.3$ (dotted line). Here $\rho = 0.2$.

of the estimator. Previous study suggested that, for nonparametric regression function estimation, taking the L_2 distance between the second derivatives of the spectra gives favorable results based on its performance on hold-out validation data. A desirable feature of an estimation procedure would be to determine the appropriate order of derivative automatically.

Here we apply the multiple functional linear regression model to the spectrometrics data. We treat the original function, as well as up to its 3rd derivative as the predictors in our model. The idea of using different orders of derivatives of curves as covariates in the functional linear model is similar to using transformations of the original covariates in classical multiple linear regression. Our use of derivatives is directly motivated by previous studies which established

that the second derivative curve contains the most important information (Ferraty and Vieu (2006)). Compared to nonparametric functional kernel regression, the functional linear model is more easily interpretable and thus an interesting alternative.

We note that since $\int \beta X' = \beta(1)X(1) - \beta(0)X(0) - \int \beta' X$ (and similarly for higher order derivatives), mathematically it seems the model is equivalent to a standard functional linear regression. However, the current approach could still be of some interest since (i) $\beta(1)X(1) - \beta(0)X(0)$ is generally not zero and thus our model is different from standard functional linear regression; (ii) even if this intercept is incorporated, which means we try to estimate the regression function $\beta(1)X(1)-\beta(0)X(0)-\int_0^1(\beta_1(t)-\beta_2'(t))X(t)dt$, it is not clear how to take into account that the coefficients of $X(0)$ and $X(1)$ are related to the functional coefficients; (iii) we might prefer to estimate β_2 to see the effect of the derivative curve, which cannot be recovered from $\beta_1(t)-\beta_2'(t)$ alone; (iv) since the derivative of β_2 is involved, it might be harder to estimate β_2' than β_2 (β_2' is "less smooth").

For these data, we trained on the first 160 spectra and used the rest as validation. We examined the prediction accuracy of the estimated model using mean squared error on the validation data:

$$MSE = \frac{1}{55}\sum_{i=161}^{215}(Y_i - \hat{Y}_i)^2.$$

With the smoothing parameters selected by GCV, the relevant predictors were found to be the 1st and 2nd derivatives of the spectra curves, achieving an MSE of 6.89. Figure 3 shows the ability of the estimated model to predict the responses. For comparison, we also computed the nonparametric kernel regression using the **funopare.kernel.cv** function provided in the **npfda** package (it uses cross-validation to select the bandwidth), which gave a smaller MSE of 5.37. However, when using functional linear modeling, unlike kernel regression, we can visually examine the features of the functional coefficients for interpretation. For example, from Figure 4, higher fat content is seen to be related to higher values around point 160 and lower values around point 215 in the first derivative, as well as lower values around point 190 in the second derivative.

3.3. Weather data

We applied the penalized multiple functional linear regression to the analysis of Japanese weather data, available in Chronological Scientific Tables 2005. The data were collected at 79 weather stations in Japan. We used the annual total precipitation averaged over 1971 to 2000 as the response. Six functional predictors were used in our model, with monthly observations averaged from 1971 to

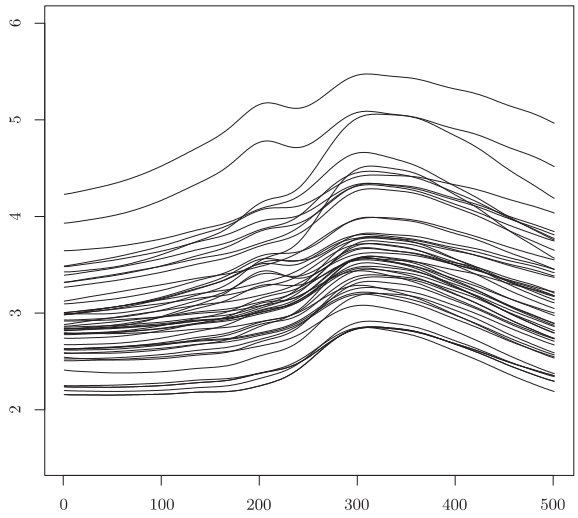

Figure 2. The spectrometric curves.

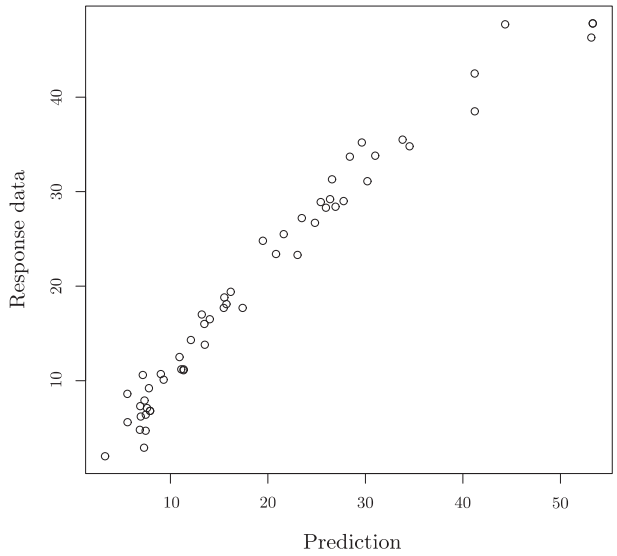

Figure 3. Prediction accuracy with penalized multiple functional regression on 55 validation samples.

2000: monthly average temperatures (TEMP), atmospheric pressure (PRESS), time of daylight (LIGHT), humidity (HUMID), monthly maximum temperature (MAX.TEMP), monthly minimum temperature (MIN.TEMP). Some of the observations on functional predictors are presented in Figure 5 (after appropriate smoothing).

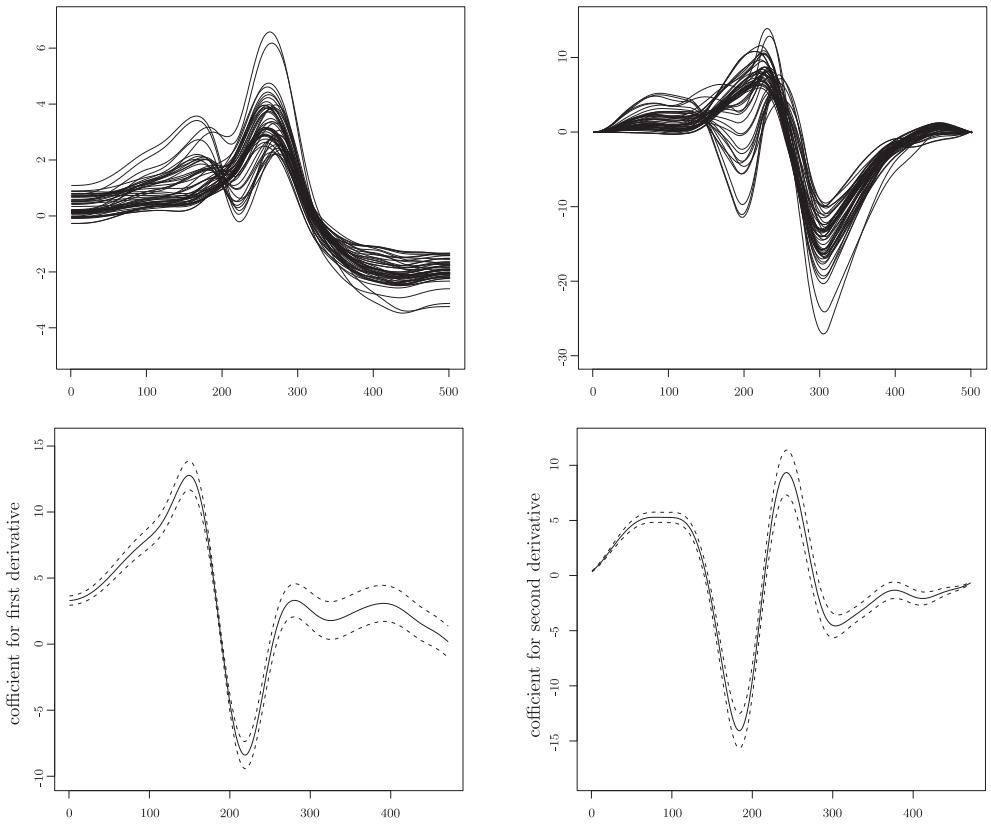

Figure 4. (a) and (b): 1st and 2nd derivative of the spectrometric data. Only 50 samples are shown in the figure. (c) and (d): Estimated functional linear coefficient corresponding to 1st and 2nd derivative curves respectively, with 95% pointwise confidence interval shown as dotted lines.

The estimated model selected MAX.TEMP, HUMID and LIGHT as predictors with nonzero coefficients and the estimated functional coefficients are shown in Figure 6. The coefficient associated with MAX.TEMP shows that higher precipitation is associated with warmer weather in the winter and colder weather in the summer. The flat coefficient for HUMID indicates that annual precipitation only depends on humidity through its annual average. The coefficient for LIGHT seems harder to interpret; the estimated coefficient suggests that it is positively correlated with precipitation in Oct-Nov and negatively correlated in Mar-Apr. As an illustration of prediction accuracy, we compared the prediction error of our multiple predictor model with a functional linear model using only one predictor. Based on five-fold cross-validation, our model gives a cross-validation MSE of 16.58. When only one functional predictor was used, the smallest error was achieved with MAX.TEMP, with an error of 23.04.

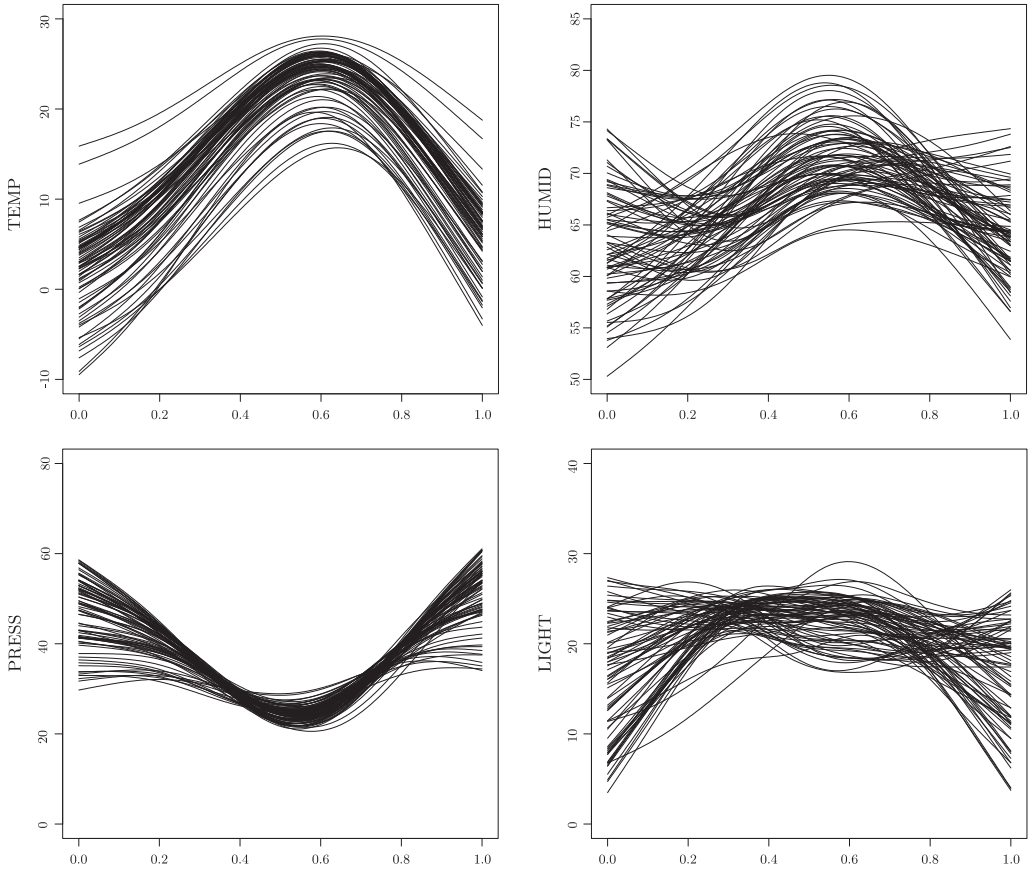

Figure 5. Some functional predictors in the Japanese weather data.

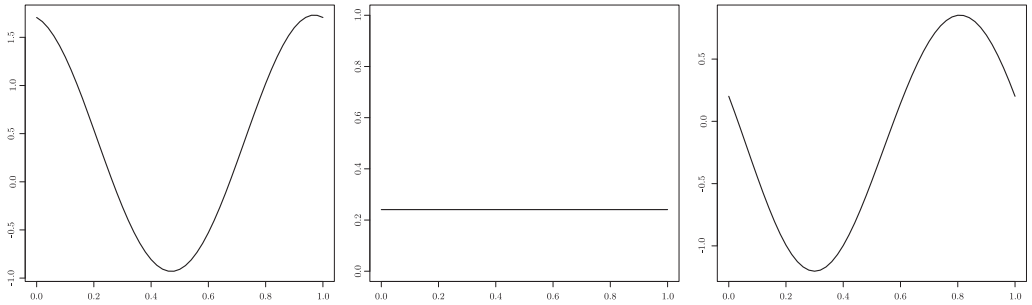

Figure 6. Estimated functional coefficients. (a) Estimated coefficient for MAX.TEMP. (b) Esimated coefficients for HUMID. (c) Estimated coefficients for LIGHT.

4. Concluding Remarks

We propose a regularization method for shrinkage estimation of multiple functional linear regression models, and show that it is consistent in estimation and variable selection. A computational algorithm based on local quadratic approximation is proposed. It is also possible to use local linear approximation (Zou and Li (2008)) and our choice is based on ease of implementation, since closed form solution exists for each iteration. Our simulation results and applications to data sets demonstrate the effectiveness of the method.

A possible topic for future study is to consider partially functional linear regression where scalar covariates are considered simultaneously. Variable selection can be applied to both the functional and non-functional part. Another direction would consider multiple functional linear regression when the number of predictors diverges with sample size; for applications that involve a large number of predictors, the diverging p case could well be more appropriate. We do not currently have such data. Extending the shrinkage estimation results to generalized functional linear model (James (2002); Müller and Stadtmüller (2005); Cardot and Sarda (2005)) is another interesting topic for study.

We became aware of an independent work by Fan and James (2011) on a similar topic after the current paper was submitted for review. Their investigations are more general than ours since they also considered additive models with unknown link functions, as well as more general basis and penalty functions. In this they assumed that the covariance information matrix ($\hat{Z}^T\hat{Z}/n$ using our notation) has eigenvalues bounded away from zero, making their theoretical analysis very similar to that of classical linear regression. Still, this assumption might not be appropriate in the functional context; we consider the magnitude of the minimum eigenvalue in our Lemma A.2 in the Appendix.

Acknowledgements

We thank the AE and reviewers for helpful comments that have improved various aspects of the manuscript. The research of Heng Lian is supported by Singapore Ministry of Education Tier 1 Grant.

Appendix

Two lemmas study properties of the estimated principal component scores. Throughout the Appendix, we follow the notation and assumptions in the main text.

Lemma A.1. We have $|\hat{\xi}_{ijk} - \xi_{ijk}| = O_p(K_j^{\alpha_j+1}/\sqrt{n})$ and $|\sum_{i=1}^n \hat{\xi}_{ij_1k_1}\hat{\xi}_{ij_2k_2}/n - \lambda_{k_1,k_2}^{j_1,j_2}| = O_p((K_{j_1}^{\alpha_{j_1}+1} + K_{j_2}^{\alpha_{j_2}+1})/\sqrt{n})$.

Proof. Given that $\hat{\phi}_{jk}$ is the eigenvector of $\hat{S}_j$, ϕ_{jk} is the eigenvector of S_j, and that $\|S_j - \hat{S}_j\| = O_p(1/\sqrt{n})$, by (5.2) of Hall and Horowitz (2007), we have $||\hat{\phi}_{jk} - \phi_{jk}|| = O_p(K_j^{2\alpha_j+2}/n)$.

Since $\xi_{ijk} = \int (X_{ij} - \mu_j)\phi_{jk}$ and $\hat{\xi}_{ijk} = \int (X_{ij} - \bar{X}_j)\hat{\phi}_{jk}$, we have $|\hat{\xi}_{ijk} - \xi_{ijk}|^2 = O_p(||\bar{X}_j - \mu_j||^2 + ||\hat{\phi}_{jk} - \phi_{jk}||^2) = O_p(K_j^{2\alpha_j+2}/n)$, using (c2) and (5.2) in Hall and Horowitz (2007). For the second part, we have

$$\begin{aligned}
&\frac{\sum_{i=1}^n \hat{\xi}_{ij_1k_1}\hat{\xi}_{ij_2k_2}}{n} - \lambda_{k_1,k_2}^{j_1,j_2} \\
&= \left(\frac{\sum_{i=1}^n \hat{\xi}_{ij_1k_1}\hat{\xi}_{ij_2k_2}}{n} - \frac{\sum_{i=1}^n \xi_{ij_1k_1}\xi_{ij_2k_2}}{n} \right) + \left(\frac{\sum_{i=1}^n \xi_{ij_1k_1}\xi_{ij_2k_2}}{n} - \lambda_{k_1,k_2}^{j_1,j_2} \right) \\
&=: \text{(I)} + \text{(II)}.
\end{aligned}$$

Obviously the second term is of order $O_p(n^{-1/2})$. The first term is further decomposed as

$$\text{(I)} = \frac{1}{n}\sum_{i=1}^n \left[(\hat{\xi}_{ij_1k_1} - \xi_{ij_1k_1})\hat{\xi}_{ij_2k_2} + (\hat{\xi}_{ij_2k_2} - \xi_{ij_2k_2})\xi_{ij_1k_1} \right].$$

Using $|\hat{\xi}_{ijk}| = O_p(1)$, since $|\hat{\xi}_{ij_1k_1} - \xi_{ij_1k_1}| = o_p(1)$, we have a bound $O_p((K_{j_1}^{\alpha_{j_1}+1} + K_{j_2}^{\alpha_{j_2}+1})/\sqrt{n})$ for (I), and the proof is complete.

Lemma A.2. For A, a subset of $\{1, 2, \ldots, p\}$, let $\hat{Z}_A$ be the columns of $\hat{Z}$ corresponding to those predictors in A, and similarly let Λ_A be the submatrix of Λ corresponding to the predictors in A. Then $\rho_{\min}(\hat{Z}_A^T\hat{Z}_A/n) = \Omega_p(\bar{K}^{-\bar{\alpha}})$.

Proof. We use $||\cdot||$ to denote also the operator norm of a matrix, and $||\cdot||_1$ for the maximum row sum of a matrix. Now $|\rho_{\min}(\hat{Z}_A^T\hat{Z}_A/n) - \rho_{\min}(\Lambda_A)| \le ||\hat{Z}_A^T\hat{Z}_A/n - \Lambda_A|| \le ||\hat{Z}_A^T\hat{Z}_A/n - \Lambda_A||_1 = O_p(\bar{K}^{\bar{\alpha}+2}/\sqrt{n})$, by Lemma A.1. This together with (c4) implies the statement of the lemma.

Proof of Theorem 1. Let the minimum eigenvalue of $\hat{Z}^T\hat{Z}/n$ be ρ^*, and thus $\rho^* = \Omega_p(\bar{K}^{-\bar{\alpha}})$ by Lemma A.2. The true functional coefficients are denoted by $\beta_j = \sum_k b_{jk}\phi_{jk}$. Then

$$\begin{aligned}
0 &\ge J(\hat{b}) - J(b) \\
&= ||\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}\hat{b}||^2 - ||\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b||^2 + n\sum_j p_{\lambda_j}(||\hat{b}_j||) - n\sum_j p_{\lambda_j}(||b_j||) \\
&= ||\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b + \hat{Z}b - \hat{Z}\hat{b}||^2 - ||\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b||^2 + n\sum_j p_{\lambda_j}(||\hat{b}_j||)
\end{aligned}$$

$$
\begin{aligned}
&-n\sum_j p_{\lambda_j}(||b_j||) \\
&= 2(\mathbf{Y}-\bar{Y}\mathbf{1}-\hat{Z}b)^T\hat{Z}(b-\hat{b}) + ||\hat{Z}(b-\hat{b})||^2 + n\sum_j p_{\lambda_j}(||\hat{b}_j||) - n\sum_j p_{\lambda_j}(||b_j||).
\end{aligned}
$$

Let $\eta = \hat{Z}(\hat{Z}^T\hat{Z})^{-1}\hat{Z}^T(\mathbf{Y}-\bar{Y}\mathbf{1}-\hat{Z}b)$ be the projection of $\mathbf{Y}-\bar{Y}\mathbf{1}-\hat{Z}b$ onto the columns of $\hat{Z}$. Lemma A.3 below shows that $||\eta||^2 = O_p(r_n^2)$, where $r_n^2 = O_p(\bar{K}^{2\bar{\alpha}+3} + n\underline{K}^{-\underline{\alpha}-2\underline{\beta}+1})$. We can then write

$$
\begin{aligned}
0 &\geq -O_p(r_n)||\hat{Z}(b-\hat{b})|| + ||\hat{Z}(b-\hat{b})||^2 + n\sum_j p_{\lambda_j}(||\hat{b}_j||) - n\sum_j p_{\lambda_j}(||b_j||) \\
&\geq -O_p(r_n^2) - \frac{1}{2}||\hat{Z}(b-\hat{b})||^2 + ||\hat{Z}(b-\hat{b})||^2 + n\sum_j p_{\lambda_j}(||\hat{b}_j||) - n\sum_j p_{\lambda_j}(||b_j||) \\
&\geq -O_p(r_n^2) + n\rho^*||b-\hat{b}||^2 - n\sum_j \lambda_j||\hat{b}_j - b_j|| \\
&\geq -O_p(r_n^2) + n\rho^*||b-\hat{b}||^2 - \frac{n\sum_j \lambda_j^2}{2\rho^*} - \frac{n\rho^*}{2}||\hat{b}-b||^2, \qquad \text{(A.1)}
\end{aligned}
$$

where we used Cauchy-Schwarz inequality on the second line, the property $|p_\lambda(a) - p_\lambda(b)| \leq \lambda|a-b|$ on the third line, and Cauchy-Schwarz inequality again on the last line. Thus $||\hat{b}-b||^2 = O_p(r_n^2/n\rho^* + \sum_j \lambda_j^2/(\rho^*)^2) = o_p(1)$ by (c3).

The convergence rate for $||\hat{b}-b||^2$ can be improved to $O_p(r_n^2/n\rho^*)$, which is useful in the proof of part (b). Since $||\hat{b}-b|| = o_p(1)$ and $\lambda_j \to 0$, we have $P(p_{\lambda_j}(||\hat{b}_j||) = p_{\lambda_j}(||b_j||), 1 \leq j \leq s) \to 1$, and thus $\sum_j p_{\lambda_j}(||\hat{b}_j||) - \sum_j p_{\lambda_j}(||b_j||) \geq 0$ with probability converging to 1. This combined with (A.1) gives $||\hat{b}-b||^2 = O_p(r_n^2/n\rho^*)$.

From $||\hat{b}_j - b_j|| = o_p(1)$,

$$
\begin{aligned}
||\hat{\beta}_j - \beta_j||^2 &\leq 2||\hat{b}_j - b_j||^2 + 2\int\Big[\sum_{k=1}^{K_j} b_{jk}(\hat{\phi}_{jk}-\phi_{jk})\Big]^2 + \sum_{k=K_j+1}^{\infty} b_{jk}^2 \\
&= 2||\hat{b}_j - b_j||^2 + 2\bar{K}\sum_{k=1}^{K_j} b_{jk}^2||\hat{\phi}_{jk}-\phi_{jk}||^2 + \sum_{k=K_j+1}^{\infty} b_{jk}^2 \\
&= 2||\hat{b}_j - b_j||^2 + O(\bar{K}\cdot\frac{\bar{K}^{2\bar{\alpha}+2}}{n}) + \sum_{k=K_j+1}^{\infty} b_{jk}^2.
\end{aligned}
$$

The last converges to zero, and this proves (a).

Now we prove part (b). Let $\hat{b}^* = (\hat{b}_1^T, \ldots, \hat{b}_s^T, 0, \ldots, 0)^T$, so $\hat{b}^*$ is obtained from $\hat{b}$ by constraining the truly irrelevant components to zero. By similar arguments

as for the proof of part (a), we have

$$\begin{aligned}0 &\geq J(\hat{b}) - J(\hat{b}^*)\\ &= 2(\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}\hat{b}^*)^T\hat{Z}(\hat{b} - \hat{b}^*) + ||\hat{Z}(\hat{b} - \hat{b}^*)||^2 + n\sum_j p_{\lambda_j}(||\hat{b}_j||)\\ &\quad -n\sum_j p_{\lambda_j}(||\hat{b}_j^*||)\\ &\geq -O_p(||\eta^*||)||\hat{Z}(\hat{b} - \hat{b}^*)|| + n\sum_{j=s+1}^{p} p_{\lambda_j}(||\hat{b}_j||)\\ &\geq -O_p(||\eta^*||)\sqrt{n}\sum_{j=s+1}^{p}||\hat{b}_j|| + n\sum_{j=s+1}^{p}\lambda_j||\hat{b}_j||, \qquad (A.2)\end{aligned}$$

where $\eta^* = \hat{Z}(\hat{Z}^T\hat{Z})^{-1}\hat{Z}^T(\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}\hat{b}^*)$. In the last line we use the fact that $||\hat{b}_j|| = O_p(r_n/\sqrt{n\rho^*}) = o_p(\lambda_j)$ when $j > s$ (from the proof of part (a)), and thus $p_{\lambda_j}(||\hat{b}_j||) = \lambda_j||\hat{b}_j||$.

We bound $||\eta^*||$ as

$$\begin{aligned}||\eta^*||^2 &\leq 2||\eta||^2 + 2||\hat{Z}(\hat{b}^* - b)||^2\\ &= O_p(r_n^2) + O_p\Big(\frac{nr_n^2}{n\rho^*}\Big) = O_p\Big(\frac{r_n^2}{\rho^*}\Big).\end{aligned}$$

Since we have that $O_p(||\eta^*||) = o_p(\sqrt{n}\lambda_j)$, there is a contradiction in (A.2) if $\sum_{j=s+1}^{p}||\hat{b}_j|| > 0$.

Lemma A.3. Let $\eta = \hat{Z}(\hat{Z}^T\hat{Z})^{-1}\hat{Z}^T(\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b)$ as in the proof of Theorem 1, then $||\eta||^2 = O_p(r_n^2)$, where $r_n^2 = \bar{K}^{2\bar{\alpha}+3} + n\underline{K}^{-\underline{\alpha}-2\underline{\beta}+1}$.

Proof. Denote by Z the matrix similar in structure to $\hat{Z}$ but with the true principal component scores ξ_{ijk} instead of $\hat{\xi}_{ijk}$. We have

$$\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b = \epsilon + (\mu - \bar{Y})\mathbf{1} + (Z - \hat{Z})b + \nu, \qquad (A.3)$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ and ν is a n-dimensional vector with

$$\nu_i = \sum_{j=1}^{p}\sum_{k=K_j+1}^{\infty}\xi_{ijk}b_{jk}.$$

Let $P_{\hat{Z}} = \hat{Z}(\hat{Z}^T\hat{Z})^{-1}\hat{Z}^T$. Now $\eta = P_{\hat{Z}}(\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b)$ is the projection of the four terms at (A.3) onto columns of $\hat{Z}$, and we bound each term in turn below.

Since $||P_{\hat{Z}}\epsilon||^2 = \epsilon^T P_{\hat{Z}}\epsilon$, using the fact $E[\epsilon^T P_{\hat{Z}}\epsilon|X] = \sigma^2 tr(P_{\hat{Z}}) = \sigma^2\sum_j K_j$ $= O(\bar{K})$, $Var(\epsilon^T P_{\hat{Z}}\epsilon|X) = 2\sigma^4 tr(P_{\hat{Z}}^2) + (E\epsilon_i^4 - 3\sigma^2)\sum_{j=1}^{n}(P_{\hat{Z}})_{jj}^2 \leq 2\sigma^4 tr(P_{\hat{Z}}^2) +$

$|E\epsilon_i^4 - 3\sigma^2| \sum_{j=1}^n (P_{\hat{Z}})_{jj} = O_p(\bar{K})$ (see, for example, (3.3) and (3.4) in Huang and Fan (1999)), where $(P_{\hat{Z}})_{jj}$ are the diagonal elements of $P_{\hat{Z}}$ which are no larger than 1, since $P_{\hat{Z}}$ is a projection matrix. Using $E\epsilon^T P_{\hat{Z}}\epsilon = E[E(\epsilon^T P_{\hat{Z}}\epsilon|X)]$ and $Var(\epsilon^T P_{\hat{Z}}\epsilon) = E[Var(\epsilon^T P_{\hat{Z}}\epsilon|X)] + Var(E[\epsilon^T P_{\hat{Z}}\epsilon|X])$, we have

$$||P_{\hat{Z}}\epsilon||^2 = O_p(\bar{K}). \tag{A.4}$$

Now $||P_{\hat{Z}}(Z - \hat{Z})b||^2 \leq ||(Z - \hat{Z})b||^2 = O(||(Z - \hat{Z})^T(Z - \hat{Z})||)$ using Lemma A.1, we get $||(Z - \hat{Z})^T(Z - \hat{Z})|| \leq ||(Z - \hat{Z})^T(Z - \hat{Z})||_1 = O_p(\bar{K}^{2\bar{\alpha}+3})$ and thus

$$||P_{\hat{Z}}(Z - \hat{Z})b||^2 = O_p(\bar{K}^{2\bar{\alpha}+3}). \tag{A.5}$$

Finally,

$$\begin{aligned} Var(\sum_{k=K_j+1}^{\infty} \xi_{ijk} b_{jk}) &= \sum_{k=K_j+1}^{\infty} \lambda_{jk} b_{jk}^2 \\ &= O(\sum_{k=K_j+1}^{\infty} k^{-\alpha_j} k^{-2\beta_j}) \\ &= O(\underline{K}^{-\underline{\alpha}-2\underline{\beta}+1}). \end{aligned}$$

Since the number of predictors p is fixed, we have $Var(\nu_i) = O(\underline{K}^{-\underline{\alpha}-2\underline{\beta}+1})$ and thus

$$||\nu||^2 = O_p(n\underline{K}^{-\underline{\alpha}-2\underline{\beta}+1}). \tag{A.6}$$

Combining (A.4), (A.5), (A.6) as well as $|\mu - \bar{Y}| = O_p(n^{-1/2})$, we get $||\eta||^2 = O_p(r_n^2)$.

Proof of Proposition 1. From (2.5), the Karhunen-Loève expansion of the predictors are

$$X_i(t) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t), \text{ with } \xi_{ik} = \sum_{j=1}^{l} a_{ij}\omega_{jk}, i = 1, 2.$$

Using the notation in the text, we have that the general entries of Λ are

$$\lambda_{k_1,k_2}^{i_1,i_2} = E\xi_{i_1k_1}\xi_{i_2k_2} = \begin{cases} \sum_{j=1}^{l} a_{i_1j}a_{i_2j}\kappa_{jk} & k_1 = k_2 = k \\ 0 & k_1 \neq k_2 \, . \end{cases}$$

Thus, in this case, in the block matrix form

$$\Lambda = \begin{pmatrix} E & F \\ F^T & G \end{pmatrix},$$

F is also diagonal. Since Λ is similar to

$$\tilde{\Lambda} = \begin{pmatrix} E & 0 \\ 0 & G - F^T E^{-1} F \end{pmatrix},$$

the eigenvalues of Λ are just the diagonal elements of E and $G - F^T E^{-1} F$. The eigenvalues of E are $\Omega(K^{-\alpha})$ by assumption, and the diagonal elements of $G - F^T E^{-1} F$ are

$$\begin{aligned}
&\sum_{j=1}^{l} a_{2j}^2 \kappa_{jk} - \frac{(\sum_{j=1}^{l} a_{1j} a_{2j} \kappa_{jk})^2}{\sum_{j=1}^{l} a_{1j}^2 \kappa_{jk}} \\
&\quad = \frac{\sum_{1 \le j_1 \ne_2 \le l} (a_{1j_1} a_{2j_2} \sqrt{\kappa_{j_1 k} \kappa_{j_2 k}} - a_{2j_1} a_{1j_2} \sqrt{\kappa_{j_1 k} \kappa_{j_2 k}})^2}{2 \sum_{j=1}^{l} a_{1j}^2 \kappa_{jk}} \\
&\quad \ge \frac{c^2 \sum_{1 \le j_1 \ne j_2 \le l} (a_{1j_1} a_{2j_2} - a_{2j_1} a_{1j_2})^2 k^{-2\alpha}}{2C \sum_{j=1}^{l} a_{1j}^2 k^{-\alpha}} \\
&\quad = \Omega(k^{-\alpha}).
\end{aligned}$$

References

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

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

Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *J. Multivariate Anal.* **92**, 24-41.

Carroll, R. J., Delaigle, A. and Hall, P. (2009). Nonparametric prediction in measurement error models. *J. Amer. Statist. Assoc.* **104**, 993-1003.

Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37**, 35-72.

Fan, J. Q. and Li, R. Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. Q. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

Fan, Y. and James, G. (2011). Functional additive regression. Technical report `http://www-bcf.usc.edu/~gareth/research/FAR.pdf`.

Ferraty, F., Mas, A. and Vieu, P. (2007). Nonparametric regression on functional data: Inference and practical aspects. *Austral. N. Z. J. Statist.* **49**, 267-286.

Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Comput. Statist.* **17**, 545-564.

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

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35**, 70-91.

Huang, J., Horowitz, J. L. and Ma, S. G. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587-613.

Huang, L. S. and Fan, J. Q. (1999). Nonparametric estimation of quadratic regression functionals. *Bernoulli* **5**, 927-949.

James, G. M. (2002). Generalized linear models with functional predictors. *J. Roy. Statist. Soc. Ser. B* **64**, 411-432.

Lian, H. (2007). Nonlinear functional models for functional responses in reproducing kernel hilbert spaces. *Canad. J. Statist.* **35**, 597-606.

Liang, H. and Li, R. Z. (2009). Variable selection for partially linear models with measurement errors. *J. Amer. Statist. Assoc.* **104**, 234-248.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34**, 2272-2297.

Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436-1462.

Müller, H. G. and Stadtmüller, U. (2005). Generalized functional linear models. *Ann. Statist.* **33**, 774-805.

Preda, C. (2007). Regression models for functional data by reproducing kernel hilbert spaces methods. *J. Statist. Plann. Inference* **137**, 829-840.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Wang, H. S. and Leng, C. L. (2007). Unified lasso estimation by least squares approximation. *J. Amer. Statist. Assoc.* **102**, 1039-1048.

Wang, H. S. and Xia, Y. C. (2009). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.* **104**, 747-757.

Wang, L. F., Chen, G. and Li, H. Z. (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics* **23**, 1486-1494.

Wang, L. F., Li, H. Z. and Huang, J. H. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103**, 1556-1569.

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

Yuan, M. and Cai, T. T. (2010). A reproducing kernel hilbert space approach to functional linear regression. *Ann. Statist.* **38**, 3412-3444.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49-67.

Zhang, H. H. (2006). Variable selection for support vector machines via smoothing spline anova. *Statist. Sinica* **16**, 659-674.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541-2563.

Zhu, H., Vannucci, M. and Cox, D. (2010). A bayesian hierarchical model for classification with selection of functional predictors. *Biometrics* **66**, 463-473.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Li, R. Z. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36**, 1509-1533.

Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, 637371.

E-mail: henglian@ntu.edu.sg

(Received July 2011; accepted April 2012)