Statistica Sinica **20** (2010), 1183-1202

SPARSE VARYING COEFFICIENT MODELS FOR LONGITUDINAL DATA

Hoh Suk Noh and Byeong U. Park

Seoul National University

Abstract: Nonparametric varying coefficient models are useful for the analysis of repeated measurements. While many procedures have been developed for estimating varying-coefficients, there have been few results on variable selection for such models. Recently, Wang, Chen and Li (2007) proposed a group SCAD procedure for model selection in varying-coefficient models, and Wang, Li and Huang (2008) established the existence of a local minimizer of the group SCAD criterion that has the oracle property. However, whether the final estimator from the gSCAD procedure via local quadratic approximation always finds the desired local minimizer is not clear. In this paper, by linearizing the gSCAD penalty we propose a one-step estimator that has the oracle property in variable selection and estimation. The proposed estimator has a much simpler implementation and gives better performance in variable selection and estimation than the ordinary gSCAD estimator.

Key words and phrases: One-step gSCAD, oracle property, spline basis, variable selection, varying-coefficient.

1. Introduction

Analysis of repeated measurement data is a recurrent challenge to statisticians engaged in biological and biomedical applications. A traditional setup for such data is to assume that the observed sequence of measurements on an individual is sampled from a realization of a continuous-time stochastic process $\{(Y(t), \mathbf{X}(t)), t \in \mathcal{T}\}$, where $Y(t)$ and $\mathbf{X}(t) = (X_1(t), \ldots, X_L(t))^\top$ denote, respectively, the response and the $\mathbb{R}^L$-valued covariate, and $\mathcal{T}$ denotes the time interval on which the measurements are taken. In practice, observations for n randomly chosen subjects are obtained as $(Y_i(t_{ij}), \mathbf{X}_i(t_{ij}))$ for the ith subject at discrete time points t_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$.

Varying-coefficient models have been used extensively for the analysis of longitudinal repeated measurement data. The linear varying-coefficient model can be written as

$$Y_i(t_{ij}) = \mathbf{X}_i(t_{ij})^\top \boldsymbol{\beta}(t_{ij}) + \epsilon_i(t_{ij}), \tag{1.1}$$

where $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_L(t))^\top$ is an L-dimensional vector of smooth functions of t, and $\epsilon_i(t),\ i = 1, \ldots, n$, are i.i.d. mean zero random processes, independent of $\mathbf{X}_i(t)$. Model (1.1) assumes a linear model for each fixed t, but allows the coefficients to vary with time. Many methods have been developed for estimating the varying coefficients in (1.1), see Hoover et al. (1998), Fan and Zhang (2000), Rice and Wu (2001), Huang, Wu and Zhou (2002, 2004), and Xue (2009), among others. However, when the number of covariates in (1.1) is very large, selection of important variables is still a challenging issue; we tackle this problem in the present paper.

Recent years have seen a few developments for regularization procedures to select important variables in Model (1.1). Fan and Li (2004) studied the SCAD penalty approach for variable selection in longitudinal data analysis, but in the case where the coefficient vector $\boldsymbol{\beta}$ is not time-dependent. Wang, Chen and Li (2007) considered a group SCAD (gSCAD) procedure, but assumed that the covariates $\mathbf{X}_i$ are not time-dependent. Wang, Li and Huang (2008) further developed the gSCAD procedure for general nonparametric varying coefficient models with time-dependent varying coefficients and time-dependent covariates. They provided theoretical justification for the gSCAD procedure by establishing the existence of a local minimizer of the gSCAD criterion that has the oracle property in variable selection and estimation. Due to the non-convexity of the gSCAD penalty, there may be multiple local minima. An important issue that deserves further attention then is whether the final estimator from the gSCAD procedure via local quadratic approximation (LQA) always finds the desired local minimizer.

Here we propose a new sparse estimation procedure for Model (1.1). The method is based on linearization of the gSCAD penalty. The idea of linearization was first used by Zou and Li (2008) to fit sparse parametric linear models. We show that the procedure leads to a sparse estimator of the varying coefficient vector that possesses the oracle property in variable selection and estimation. Our theory uses general basis functions for approximating the unknown varying coefficient functions. Thus, the results are valid for non-orthonormal basis functions, such as B-splines, as well as orthonormal basis functions. Another important advantage of our procedure is that one does not need to choose an artificial cut-off value as typically required for iterative algorithms based on LQA of the SCAD or gSCAD penalty to make the resulting estimators sparse. It was observed by Hall, Lee and Park (2009) that the performance of SCAD procedures depends crucially on the choice of the cut-off value. Since our procedure for computing the estimator is a traditional convex programming problem, its implementation is much simpler than the ordinary gSCAD estimator of Wang, Li and Huang (2008). Furthermore, our simulation study suggests that the proposed one-step gSCAD procedure has better finite sample performance than the ordinary gSCAD.

The rest of the paper is organized as follows. We first describe the one-step gSCAD estimator and an algorithm in Section 2. We then present the oracle property of our estimator in Section 3. We report simulation results in Section 4. Technical details are deferred to Section 5.

2. One-step gSCAD Procedure

In order to select significant covariates in Model (1.1), Wang, Li and Huang (2008) proposed a grouped version of the SCAD penalty based on a basis expansion of $\boldsymbol{\beta}$ and penalized estimation. We assume that each β_l, $l = 1, \ldots, L$, can be approximated by a set of basis functions, that is,

$$\beta_l(t) \approx \sum_{k=1}^{K_l} \gamma_{lk} B_{lk}(t), \ l = 1, \ldots, L, \tag{2.1}$$

where $\{B_{lk}\}_{k=1}^{\infty}$ span a function space $\mathcal{F}_l$ which is assumed to contain β_l, and K_l is the number of the basis functions used to approximate β_l. In Wang, Li and Huang (2008) it is assumed that $\{B_{lk}\}_{k=1}^{\infty}$ is a spline basis for each l. We work on general basis functions. Since, under the approximations (2.1), each function β_l in (1.1) is characterized by a set of parameters $\boldsymbol{\gamma}_l = (\gamma_{l1}, \ldots, \gamma_{lK_l})^\top$, one should not select nonzero individual components γ_{lk}, but choose the whole nonzero vector $\boldsymbol{\gamma}_l$.

For the approximation $g_l = \sum_{k=1}^{K_l} \gamma_{lk} B_{lk}$ at (2.1), its squared L_2-norm can be written as $\|g_l\|^2 = \boldsymbol{\gamma}_l^\top H_l \boldsymbol{\gamma}_l$, where H_l is a $K_l \times K_l$ matrix with entries $h_{kk'} = \int_{\mathcal{T}} B_{lk}(t) B_{lk'}(t)\, dt$. For a vector $\boldsymbol{\gamma}_l \in \mathbb{R}^{K_l}$, we use the standardized ℓ_2-norm

$$\|\boldsymbol{\gamma}_l\|_w \equiv (\boldsymbol{\gamma}_l^\top H_l \boldsymbol{\gamma}_l)^{1/2}$$

to define a group SCAD penalty function. Let $w_i \geq 0$ be the weights that are applied to each subject, and $p_\lambda(\cdot)$ be the SCAD penalty function introduced by Fan and Li (2001). The function p_λ is defined on $\mathbb{R}^+$ by its derivative

$$p'_\lambda(x) = \lambda I(x \leq \lambda) + \frac{(a\lambda - x)_+}{a-1} I(x > \lambda)$$

for some $a > 2$. Let X_{il} denote the lth component of the covariate vector process $\mathbf{X}_i$. Then, the group SCAD regularized estimator of $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_L^\top)^\top$ that we consider in this paper minimizes

$$\begin{aligned} l(\boldsymbol{\gamma}) = & \sum_{i=1}^{n} w_i \sum_{j=1}^{n_i} \left(Y_i(t_{ij}) - \sum_{l=1}^{L} \sum_{k=1}^{K_l} X_{il}(t_{ij}) B_{lk}(t_{ij}) \gamma_{lk} \right)^2 \\ & + \sum_{l=1}^{L} p_\lambda(\|\boldsymbol{\gamma}_l\|_w) \end{aligned} \tag{2.2}$$

with respect to $\boldsymbol{\gamma}$.

The SCAD penalty has the tuning parameters a and λ. Fan and Li (2001) suggested the use of $a = 3.7$. Some examples of w_i include $w_i \equiv (\sum_{i=1}^{n} n_i)^{-1}$ and $w_i = (n \times n_i)^{-1}$. Wang, Li and Huang (2008) used the latter in their objective function. Huang, Wu and Zhou (2002) showed that the latter choice yields consistency of the estimators under weaker conditions than the former in the problem of estimating the varying coefficients without penalty. Our theory in the next section treats a general choice of $\{w_i\}$.

A difficulty with the gSCAD approach described above is that the loss l often has multiple local minima due to the non-convexity of the gSCAD penalty function. The gSCAD algorithm suggested by Wang, Li and Huang (2008) is based on LQA and leads to a local minimizer whose statistical properties are not well known. As remarked by Wang, Li and Huang (2008), their gSCAD algorithm may fail to converge and the search may fall into an infinite loop surrounding several local minimizers. To overcome the difficulty, we adopt the approach of Zou and Li (2008) who worked on a linear approximation of the SCAD penalty to obtain a one-step SCAD estimator in parametric linear models. The oracle property here means that the zero coefficient functions are correctly identified with probability approaching one, and the rates of convergence of the estimators of the nonzero coefficient functions coincide with those of the oracle estimators that use knowledge of the zero coefficient functions.

2.1. Local linear approximation of gSCAD penalty

The local linear approximation (LLA) of the gSCAD penalty is

$$p_\lambda(\|\gamma_l\|_w) \approx p_\lambda(\|\gamma_l^{(0)}\|_w) + p'_\lambda(\|\gamma_l^{(0)}\|_w)(\|\gamma_l\|_w - \|\gamma_l^{(0)}\|_w)$$

for $\boldsymbol{\gamma} \approx \boldsymbol{\gamma}^{(0)}$. Set an initial value $\boldsymbol{\gamma}^{(0)}$. For $k = 0, 1, 2, \ldots$ one may repeatedly minimize $l_a(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(k)})$ to obtain the updated value $\boldsymbol{\gamma}^{(k+1)}$, where

$$\begin{aligned} l_a(\boldsymbol{\gamma}|\boldsymbol{\gamma}') = &\sum_{i=1}^{n} w_i \sum_{j=1}^{n_i} \left(Y_i(t_{ij}) - \sum_{l=1}^{L}\sum_{k=1}^{K_l} X_{il}(t_{ij}) B_{lk}(t_{ij}) \gamma_{lk} \right)^2 \\ &+ \sum_{l=1}^{L} p'_\lambda(\|\gamma'_l\|_w)\|\gamma_l\|_w. \end{aligned}$$

The following proposition is a version of Theorem 1 in Zou and Li (2008). The proof is the same as theirs, hence is omitted.

Proposition 2.1. *For the SCAD penalty function p_λ, $l(\boldsymbol{\gamma}) \le l_a(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(k)}) + c$ and $l(\boldsymbol{\gamma}^{(k)}) = l_a(\boldsymbol{\gamma}^{(k)}|\boldsymbol{\gamma}^{(k)}) + c$, where $c = \sum_{l=1}^{L}[p_\lambda(\|\gamma_l^{(0)}\|_w) - p'_\lambda(\|\gamma_l^{(0)}\|_w)\|\gamma_l^{(0)}\|_w]$. Furthermore, the LLA algorithm has satisfies $l(\boldsymbol{\gamma}^{(k+1)}) \le l(\boldsymbol{\gamma}^{(k)})$ for all $k = 0, 1, 2, \ldots$*

The LLA algorithm is thus an instance of the Majorize-Minimize (MM) algorithms and its convergence may be analyzed using general convergence results for MM algorithms, see Lange (1995), Hunter and Li (2005), and Zou and Li (2008). The fact that the LLA is the best convex majorization of p_λ is demonstrated by Theorem 2 in Zou and Li (2008).

2.2. One-step gSCAD estimator

Let $\boldsymbol{\gamma}^0$ be an initial estimator of $\boldsymbol{\gamma}$, and $\hat{\boldsymbol{\gamma}}$ be the one-step gSCAD estimator obtained by minimizing the loss

$$\begin{aligned} l_a(\boldsymbol{\gamma}) \ &\equiv l_a(\boldsymbol{\gamma}|\hat{\boldsymbol{\gamma}}^{(0)}) = \sum_{i=1}^{n} w_i \sum_{j=1}^{n_i} \left(Y_i(t_{ij}) - \sum_{l=1}^{L}\sum_{k=1}^{K_l} X_{il}(t_{ij}) B_{lk}(t_{ij}) \gamma_{lk} \right)^2 \\ &\qquad + \sum_{l=1}^{L} p'_\lambda(\|\boldsymbol{\gamma}_l^0\|_w)\|\boldsymbol{\gamma}_l\|_w. \end{aligned} \tag{2.3}$$

In Section 3, we show that, if one uses the least square estimator in Huang, Wu and Zhou (2004) as an initial value $\hat{\boldsymbol{\gamma}}^0$, the one-step gSCAD estimator is already efficient, enjoying the oracle property in variable selection and estimation.

Since minimization of l_a at (2.3) is a convex programming problem as is the lasso, the estimator is uniquely defined and can be computed by a convex optimization technique. We present a complete algorithm; alternatively, the minimization can be performed by a standard convex programming solver such as `CVX` (Grant and Boyd (2008)).

For $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, let $\mathbf{Y}_i = (Y_i(t_{i1}), \ldots, Y_i(t_{in_i}))^\top$, $\mathbf{Y} = (\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_n^\top)^\top$, $\mathbf{W}_i = w_i I_{n_i}$ with I_d denoting the d-dimensional identity matrix, and $\mathbf{W} = \operatorname{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_n)$. Also, define $\mathbf{U}_{ij}^\top = \mathbf{X}_i(t_{ij})^\top \mathbf{B}(t_{ij})$, $\mathbf{U}_i = (\mathbf{U}_{i1}, \ldots, \mathbf{U}_{in_i})^\top$, and $\mathbf{U} = (\mathbf{U}_1^\top, \ldots, \mathbf{U}_n^\top)^\top$, where

$$\mathbf{B}(t) = \begin{pmatrix} B_{11}(t) & \cdots & B_{1K_1}(t) & 0\cdots 0 & 0 & \cdots & 0 \\ & \vdots & & \vdots & & \vdots & \\ 0 & \cdots & 0 & 0\cdots 0 & B_{L1}(t) & \cdots & B_{LK_L}(t) \end{pmatrix}.$$

Then, (2.3) can be written as

$$l_a = (\mathbf{Y} - \mathbf{U}\boldsymbol{\gamma})^\top \mathbf{W} (\mathbf{Y} - \mathbf{U}\boldsymbol{\gamma}) + \sum_{l=1}^{L} p'_\lambda(\|\boldsymbol{\gamma}_l^0\|_w)\|\boldsymbol{\gamma}_l\|_w.$$

As in Zou and Li (2008), we split the quadratic term in l_a into two parts, one for those $\boldsymbol{\gamma}_l$ with l in A and the other for those with l in B:

$$A = \{l : p'_\lambda(\|\hat{\boldsymbol{\gamma}}_l^0\|_w) = 0\}, \quad B = \{l : p'_\lambda(\|\hat{\boldsymbol{\gamma}}_l^0\|_w) > 0\}.$$

Computation of $\hat{\gamma}_l$ with $l \in A$ is the standard least squares problem, and those $\hat{\gamma}_l$ with $l \in B$ can be obtained by the standard group lasso algorithm.

To be specific, we rearrange the columns of the matrix $\mathbf{U}$ so that those corresponding to $l \in A$ are positioned in the first $\sum_{l \in A} K_l$ columns. With a slight abuse of notation, we continue to refer to the rearranged matrix as $\mathbf{U}$. Thus we write $\mathbf{U} = (\mathbf{U}_A, \mathbf{U}_B)$, where $\mathbf{U}_A$ consists of those $\sum_{l \in A} K_l$ columns of $\mathbf{U}$ that correspond to $l \in A$ and $\mathbf{U}_B$ is the remaining block of $\mathbf{U}$. Likewise, we rearrange and partition the coefficient vector $\boldsymbol{\gamma}$ as $\boldsymbol{\gamma}^\top = (\boldsymbol{\gamma}_A^\top, \boldsymbol{\gamma}_B^\top)$, where $\boldsymbol{\gamma}_A^\top = (\boldsymbol{\gamma}_l^\top : l \in A)$ and $\boldsymbol{\gamma}_B^\top = (\boldsymbol{\gamma}_l^\top : l \in B)$. Let $\mathbf{U}^l$ denote the lth block of the rearranged matrix $\mathbf{U}$, that is, $\mathbf{U}^l$ is of size $(\sum_{i=1}^n n_i) \times K_l$. Let H_A be the projection matrix onto the column space of $\mathbf{U}_A$ in $\mathbb{R}^{\sum_{i=1}^n n_i}$ with the inner product $\langle \boldsymbol{y}_1, \boldsymbol{y}_2 \rangle = \boldsymbol{y}_1^\top \mathbf{W} \boldsymbol{y}_2$, that is $H_A = \mathbf{U}_A(\mathbf{U}_A^\top \mathbf{W} \mathbf{U}_A)^{-1} \mathbf{U}_A^\top \mathbf{W}$. For $l \in B$, define

$$\tilde{\mathbf{U}}^l = \mathbf{U}^l \left[\frac{\lambda}{p'_\lambda(\|\hat{\gamma}_l^0\|_w)} \right], \quad \gamma_l^* = \gamma_l \left[\frac{p'_\lambda(\|\hat{\gamma}_l^0\|_w)}{\lambda} \right].$$

Let $\tilde{\mathbf{U}}_B = (\tilde{\mathbf{U}}^l : l \in B)$ and $\boldsymbol{\gamma}_B^{*\top} = (\boldsymbol{\gamma}_l^{*\top} : l \in B)$. Then, one can decompose $\mathbf{U}\boldsymbol{\gamma} = \mathbf{U}_A \boldsymbol{\gamma}_A + \tilde{\mathbf{U}}_B \boldsymbol{\gamma}_B^*$ into a sum of orthogonal vectors as

$$\mathbf{U}\boldsymbol{\gamma} = \mathbf{U}_A \boldsymbol{\gamma}_A^* + (I - H_A)\tilde{\mathbf{U}}_B \boldsymbol{\gamma}_B^*,$$

where $\boldsymbol{\gamma}_A^* = \boldsymbol{\gamma}_A + (\mathbf{U}_A^\top \mathbf{W} \mathbf{U}_A)^{-1} \mathbf{U}_A^\top \mathbf{W} \tilde{\mathbf{U}}_B \boldsymbol{\gamma}_B^*$. Decomposing $\mathbf{Y} = H_A \mathbf{Y} + \mathbf{Y}^*$ likewise, where $\mathbf{Y}^* = (I - H_A)\mathbf{Y}$, it follows that

$$\begin{aligned}(\mathbf{Y} - \mathbf{U}\boldsymbol{\gamma})^\top \mathbf{W} (\mathbf{Y} - \mathbf{U}\boldsymbol{\gamma}) = {} & (H_A \mathbf{Y} - \mathbf{U}_A \boldsymbol{\gamma}_A^*)^\top \mathbf{W} (H_A \mathbf{Y} - \mathbf{U}_A \boldsymbol{\gamma}_A^*) \\ & + (\mathbf{Y}^* - \tilde{\mathbf{U}}_B^* \boldsymbol{\gamma}_B^*)^\top \mathbf{W} (\mathbf{Y}^* - \tilde{\mathbf{U}}_B^* \boldsymbol{\gamma}_B^*),\end{aligned}$$

where $\tilde{\mathbf{U}}_B^* = (I - H_A)\tilde{\mathbf{U}}_B$.

The foregoing arguments justify the following algorithm.

(1) Solve the typical group lasso problem

$$\hat{\boldsymbol{\gamma}}_B^* = \underset{\boldsymbol{\gamma}_B^*}{\text{argmin}} \, \frac{1}{n} (\mathbf{Y}^* - \tilde{\mathbf{U}}_B^* \boldsymbol{\gamma}_B^*)^\top \mathbf{W} (\mathbf{Y}^* - \tilde{\mathbf{U}}_B^* \boldsymbol{\gamma}_B^*) + \lambda \sum_{l \in B} \|\boldsymbol{\gamma}_l^*\|_w,$$

and then compute

$$\hat{\boldsymbol{\gamma}}_B = \hat{\boldsymbol{\gamma}}_B^* \left[\frac{\lambda}{p'_\lambda(\|\hat{\gamma}_l^0\|_w)} \right].$$

(2) Compute $\hat{\boldsymbol{\gamma}}_A = (\mathbf{U}_A^\top \mathbf{W} \mathbf{U}_A)^{-1} \mathbf{U}_A^\top \mathbf{W} (\mathbf{Y} - \tilde{\mathbf{U}}_B \hat{\boldsymbol{\gamma}}_B^*)$.

We note that the iterative algorithm of Wang, Li and Huang (2008) based on LQA of $p_\lambda(\|\gamma_l\|_w)$ requires one to select an additional cut-off parameter to produce a group-sparse solution. It was observed by Hall, Lee and Park (2009) that the performance of SCAD procedures depends crucially on the choice of the cut-off value. As demonstrated in Zou and Li (2008), if the cut-off value in LQA is properly set in the right scale, the performance of LQA may be similar to that of the one-step procedure. In the regression spline setting of the present paper, however, it is difficult to find an appropriate scale for δ. Since our method uses a group lasso algorithm, it does not need to select additional parameters to get a sparse solution. For a group lasso algorithm, one may use the one proposed by Yuan and Lin (2006), as we have done in our numerical study.

For an initial estimator $\hat{\gamma}_0$, one can take the least square estimator in Huang, Wu and Zhou (2004). The initial estimator $\hat{\gamma}_0$ is defined as the minimizer of

$$\sum_{i=1}^{n} w_i \sum_{j=1}^{n_i} \left(Y_i(t_{ij}) - \sum_{l=1}^{L} \sum_{k=1}^{K_l} X_{il}(t_{ij}) B_{lk}(t_{ij}) \gamma_{lk} \right)^2 .$$

We show that with this initial estimator the proposed one-step estimator has the oracle property. For the initial estimation one needs to choose several tuning parameters. For example, if one uses a spline basis, one needs to select the degrees of the spline functions, the numbers and locations of the knots. For simplicity, one may use the usual cubic splines with equally spaced knots and select K_l only. To choose the tuning parameters K_l one can use the "leave-one-subject-out" cross-validation (Rice and Silverman (1991)), or the K-fold cross-validation by splitting the subjects into K roughly equal-size parts. For the one-step estimator, one needs to choose K_l again for the minimization of l_a at (2.3). In addition to this, one should also select the regularization parameter λ. To avoid unnecessary complication we suggest using the same K_l used for the initial estimator. This is justified by the fact that the optimal order of K_l for the one-step estimator is the same as that of the initial estimator. For the selection of λ, one may use a cross-validatory criterion, or develop a bootstrap procedure similar to the one that was proposed by Hall, Lee and Park (2009) in a related problem.

We introduce a cross-validatory criterion that selects the regularization parameter λ as well as the number of knots $K \equiv K_l$ to be used in the initial and the one-step estimator. We have used this criterion in our numerical study presented in Sections 4 and 5. The cross-validatory criterion, which was also considered in Wang, Li and Huang (2008), is

$$CV(K, \lambda) = \sum_{i=1}^{n} w_i \|\mathbf{Y}_i - \mathbf{U}_i \hat{\boldsymbol{\gamma}}^{(-i)}\|_2^2, \tag{2.4}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm, and $\hat{\boldsymbol{\gamma}}^{(-i)}$ is the one-step gSCAD estimator obtained with the ith subject being deleted. Now, suppose that the initial estimator $\hat{\boldsymbol{\gamma}}^0$ is given as the minimizer of $(\mathbf{Y} - \mathbf{U}\boldsymbol{\gamma})^\top \mathbf{W} (\mathbf{Y} - \mathbf{U}\boldsymbol{\gamma})$. Then, the one-step gSCAD estimator $\hat{\gamma}$ is approximately $\hat{\gamma}^{\text{ridge}} = (\mathbf{U}^\top\mathbf{W}\mathbf{U} + \Sigma_\lambda(\hat{\gamma}^0))^{-1}\mathbf{U}^\top\mathbf{W}\mathbf{Y}$, where

$$\Sigma_\lambda(\gamma) = \text{diag}\Big\{\frac{I_{K_1}p'_\lambda(\|\gamma_1\|_2)}{\|\gamma_1\|_2}, \ldots, \frac{I_{K_L}p'_\lambda(\|\gamma_L\|_2)}{\|\gamma_L\|_2}\Big\}$$

and I_k denotes the identity matrix of dimension k. The latter is the solution of the ridge regression that minimizes

$$(\mathbf{Y} - \mathbf{U}\gamma)^\top\mathbf{W}(\mathbf{Y} - \mathbf{U}\gamma) + \gamma^\top\Sigma_\lambda(\hat{\gamma}^0)\gamma. \tag{2.5}$$

From Lemma 1 of Wang, Li and Huang (2008), it follows that, if $\hat{\gamma}^{\text{ridge},-i}$ are the minimizers of (2.5) computed with the ith subject being deleted, then the cross-validatory criterion at (2.4) with $\hat{\gamma}^{(-i)}$ being replaced by $\hat{\gamma}^{\text{ridge},-i}$ is

$$\sum_{i=1}^n w_i\|(I_{n_i} - \mathbf{M}_{ii}(K,\lambda))^{-1}(\mathbf{Y}_i - \mathbf{U}_i\hat{\gamma}^{\text{ridge}})\|_2^2,$$

where $\mathbf{M}(K,\lambda) = \mathbf{U}(\mathbf{U}^\top\mathbf{W}\mathbf{U} + \Sigma_\lambda(\hat{\gamma}^0))^{-1}\mathbf{U}^\top\mathbf{W}$. This motivates us to use

$$ACV(K,\lambda) = \sum_{i=1}^n w_i\|(I_{n_i} - \mathbf{M}_{ii}(K,\lambda))^{-1}(\mathbf{Y}_i - \mathbf{U}_i\hat{\gamma})\|_2^2$$

as an approximation of $CV(K,\lambda)$ at (2.4) for the selection of K and λ.

3. Oracle Property

The oracle property in a strong sense implies that the asymptotic distributions of the estimators of the nonzero coefficients coincide with those of the oracle estimators. Fan and Li (2001), Zou (2006), and Zou and Yuan (2008), among others, established this property in parametric linear models. In nonparametric settings, however, the oracle property often means that the rates of convergence, rather than the asymptotic distributions, of the nonparametric estimators are the same as those of the oracle estimators. This weaker notion of oracle property was also used in, for example, Storlie et al. (2010) for a nonparametric additive model, and Bach (2008) for multiple kernel learning.

We take the latter definition of the oracle property. For $\hat{\beta}_l \equiv \sum_{k=1}^{K_l}\hat{\gamma}_{lk}B_{lk}$, where $\hat{\gamma}_{lk}$ are the one-step gSCAD estimators of γ_{lk} defined in Section 2, we show that $\hat{\beta}_l = 0$ with probability tending to one for all the irrelevant covariates X_l, and that $\hat{\beta}_l$ converge to the true β_l at the univariate optimal rate for all the relevant covariates X_l. Our assumptions are similar to those in Wang, Li and Huang (2008), but we establish the oracle property for general basis functions B_{lk}.

We assume that only s predictors among the X_l are relevant in Model (1.1). Without loss of generality, we let $\beta_l(\cdot)$, $1 \le l \le s$, be the nonzero coefficient functions, and β_l, $(s+1) \leq l \leq L$, be identically zero. We consider the case where K_l tends to infinity as n goes to infinity, thus K_l depends on n although we suppress the dependence. We treat the case where the n_i are deterministic, as in Fan and Zhang (2000) and Huang, Wu and Zhou (2002, 2004). The case of random n_i requires a different analysis. We make the following technical assumptions.

(A1) The observation times t_{ij} are chosen independently according to a distribution F_T on $\mathcal{T}$, and independent of the response and covariate processes $\{Y_i(\cdot), \mathbf{X}_i(\cdot)\}$, $i = 1, \ldots, n$. The distribution F_T has a density f_T, with respect to Lebesgue measure, bounded away from 0 and ∞ in $t \in \mathcal{T}$.

(A2) The eigenvalues of the matrix $E\left\{\mathbf{X}(t)\mathbf{X}(t)^\top\right\}$ are bounded away from 0 and ∞ in $t \in \mathcal{T}$.

(A3) There exists a positive constant $M_1 < \infty$ such that $|X_l(t)| \leq M_1$ for all $t \in \mathcal{T}$ and for all $l = 1, \ldots, L$

(A4) There exists a positive constant M_2 such that $E\epsilon(t)^2 \leq M_2$ for all $t \in \mathcal{T}$.

(A5) $\limsup_{n\to\infty} (\max_{1\leq l\leq L} K_l / \min_{1\leq l\leq L} K_l) < \infty$.

(A6) There exist constants $\alpha \geq 0$ and $0 < M_3, M_4 < \infty$, not depending on K_l, such that

$$M_3 K_l^{-\alpha} \sum_{k=1}^{K_l} \gamma_{lk}^2 \leq \int_{\mathcal{T}} \Big[\sum_{k=1}^{K_l} \gamma_{lk} B_{lk}(t)\Big]^2 dt \leq M_4 K_l^{-\alpha} \sum_{k=1}^{K_l} \gamma_{lk}^2$$

for any sequence $\{\gamma_{lk} \in \mathbb{R} : k = 1, \ldots, K_l\}$.

Assumption (A6) is the only requirement for the system of basis functions B_{lk}. We show that under this assumption our estimator enjoys the oracle property. Assumption (A6) asks that the ℓ_2-norm of the estimated coefficient vector $\hat{\gamma}_l$ can be translated directly to the L_2-norm of the estimated function $\hat{\beta}_l$. This assumption accommodates not only orthonormal bases with $M_3 = M_4 = 1$ and $\alpha = 0$, but also non-orthonormal bases such as the Riesz basis ($\alpha = 0$) and the B-spline ($\alpha = 1$).

We now state the main theorem. For each $1 \leq l \leq L$, let $\mathbb{G}_l$ be a linear space spanned by $\{B_{lk}(\cdot) : 1 \leq k \leq K_l\}$. Let $K = \max_{1\leq l\leq L} K_l$, $\mathrm{dist}(\beta_l, \mathbb{G}_l) = \inf_{g\in\mathbb{G}_l} \sup_{t\in\mathcal{T}} |\beta_l(t) - g(t)|$, and $\rho = \max_{1\leq l\leq L} \mathrm{dist}(\beta_l, \mathbb{G}_l)$. Also, take

$$A_l = \sup_{g\in\mathbb{G}_l,\ \|g\|\neq 0} \frac{\sup_{t\in\mathcal{T}} |g(t)|}{\|g\|}, \quad A = \max_{0\leq l\leq L} A_l,$$

where $\|\cdot\|$ denotes the L_2-norm. In this notation, we suppress dependence on n in K, ρ, A_l and A.

Theorem 3.1. *Assume that $\lambda, r, \rho \to 0$ and $\lambda/\max\{r,\rho\} \to \infty$ as $n \to \infty$, where $r = (K\sum_{i=1}^n n_i^2 w_i^2)^{1/2}$, and that*

$$\lim_{n\to\infty}\left[A^2 K \max\left\{\max_{1\le i\le n}(n_i w_i), \sum_{i=1}^n n_i^2 w_i^2\right\}\right] = 0. \tag{3.1}$$

Furthermore, assume that the initial estimator $\hat{\boldsymbol{\gamma}}^0$ is given as the minimizer of $(\mathbf{Y} - \mathbf{U}\boldsymbol{\gamma}^0)^\top \mathbf{W}(\mathbf{Y} - \mathbf{U}\boldsymbol{\gamma}^0)$. Then, under (A1)−(A6),

(a) $\hat{\beta}_l = 0$ *for all $(s+1) \le l \le L$ with probability tending to* 1;

(b) $\|\hat{\beta}_l - \beta_l\| = O_p(\rho + r)$ *for all* $1 \le l \le s$.

When β_l has a bounded second derivatives and $\mathbb{G}_l$ is a space of cubic splines with K interior knots on $\mathcal{T}$, one has $A_l = O(K^{1/2})$ and $\rho = O(K^{-2})$, see Theorem 6.27 in Schumaker (1981), and also Huang (1998). Thus if $\sum_{i=1}^n n_i^2 w_i^2 = O(n^{-1})$, one gets $\|\hat{\beta}_l - \beta_l\| = O_p(n^{-2/5})$ by taking $K \sim n^{1/5}$, which is the same as the optimal rate for i.i.d. data (Stone (1982)). If one employs the weights $w_i = (nn_i)^{-1}$, then the condition on $\{n_i\}$ is satisfied automatically. If one uses $w_i = (\sum_{i=1}^m n_i)^{-1}$, then the condition is met provided the sequence $(\max_{1\le i\le n} n_i/\min_{1\le i\le n} n_i)$ is bounded.

One can derive a version of Theorem 3.1 for the one-step gSCAD estimator with the ℓ_2-norm $\|\cdot\|_2$, instead of the standardized norm $\|\cdot\|_w$, in the penalty function. To get the conclusions (a) and (b) of the theorem, one needs the following additional conditions on K and λ:

$$rK^{\alpha/2}, \rho K^{\alpha/2}, \lambda K^{\alpha} \;\to\; 0, \quad \text{and} \quad \frac{\lambda}{\max\{K^{\alpha/2}r, K^{\alpha/2}\rho\}} \;\to\; \infty$$

as $n \to \infty$. Thus, to get the oracle property for cubic splines, the regularization parameter λ should satisfy $\lambda n^{1/5} \to 0$ and $\lambda n^{3/10} \to \infty$. This means that the use of the ℓ_2-norm in the penalty function requires stronger conditions on λ than the norm-type $\|\cdot\|_w$ that requires $\lambda n^{2/5} \to \infty$. In general, the optimal order of λ for the norm-type $\|\cdot\|_2$ is greater than the one for the norm-type $\|\cdot\|_w$ by a factor of $K^{\alpha/2}$. In practical terms, this means that one should take a wider range of λ to search for the best λ with the norm-type $\|\cdot\|_2$.

Wang, Li and Huang (2008) showed that the asymptotic distribution of the ordinary gSCAD estimator is the same as that of the oracle estimator. Their work is based on spline basis functions. Using the same arguments, one can show that our one-step gSCAD estimator with spline basis functions B_{lk} enjoys the same property under the additional condition (C6) of Wang, Li and Huang (2008) on the error process.

4. Simulation Study

We conducted simulation studies to assess the effectiveness of our one-step gSCAD procedure for selecting and estimating the relevant varying coefficients. To compare the one-step gSCAD estimator with the ordinary gSCAD and the oracle estimators, we generated random samples of $n = 100$ subjects from the model (1.1). The observation time points t_{ij} were generated by the same scheme as in Huang, Wu and Zhou (2002), where each subject had a set of "scheduled"time points $\{1, \ldots, 30\}$, and each scheduled time had 60% probability of being skipped. A Uniform$(-0.5, 0.5)$ random deviate was added to a non-skipped scheduled time to obtain an actual observation time point. We took $L = 23$ with the coefficient functions

$$\beta_0(t) = 15 + 20 \sin\left(\frac{\pi t}{60}\right), \quad \beta_1(t) = 2 - 3\cos\left(\frac{\pi(t-25)}{15}\right),$$

$$\beta_2(t) = 6 - 0.2t, \qquad\qquad \beta_3(t) = -4 + \frac{(20-t)^3}{2000},$$

and $\beta_l \equiv 0$ for $4 \leq l \leq 23$. Thus, the variables X_l for $4 \leq l \leq 23$ in $\mathbf{X} = (X_0, X_1, \ldots, X_{23})^\top$ were irrelevant.

The variables X_l were generated as follows. For each observation time point t, $X_1(t)$ was sampled from the uniform distribution on $[t/10, 2+t/10]$; the second variable $X_2(t)$, conditioned on the value of $X_1(t)$, was generated from the normal distribution with mean zero and variance $(1 + X_1(t))/(2 + X_1(t))$; the third variable $X_3(t)$, independent of $X_1(t)$ and $X_2(t)$, was a Bernoulli random variable with probability of success 0.6. The remaining 20 covariate processes $X_l(\cdot)$ for $4 \leq l \leq 23$ were taken to be independent of each other, and each of them was a Gaussian process with mean zero and covariance structure $\mathrm{cov}(X_l(t), X_l(s)) = \sigma_1^2 \exp(-|t-s|)$. The error process $\epsilon(\cdot)$ was set to $\epsilon(t) = Z(t) + u(t)$, where the process $Z(\cdot)$ had the same distribution as $X^{(l)}(\cdot)$, $4 \leq l \leq 23$, and $u(t)$ for each t was an independent measurement error from $N(0, \sigma_2^2)$. For the levels of the correlations among X_l, $4 \leq l \leq 23$, we chose $\sigma_1 = 1$ and 2, and we took $\sigma_2 = 2$ and 3.

We repeated the simulation 100 times. We used cubic splines and an equal number of knots for estimating different varying coefficients. The cross-validatory criterion ACV, described at the end of Section 2, was used to choose the number of knots K and the regularization parameter λ for the one-step and the ordinary gSCAD estimators. For the tuning parameter a in the definition of p_λ, we took $a = 3.7$ as suggested in Fan and Li (2001). Since there is no confirmed guideline to choosing an appropriate δ in our setting, we adopted the approach of Hall, Lee and Park (2009): the cut-off value δ to set the ordinary gSCAD estimator

Table 1. Comparison of the one-step and the ordinary gSCAD estimators in terms of identifying relevant variables and the expected value of MADE, based on 100 pseudo-samples of size $n = 100$.

Correlation and		Avg. No of 0 Coefficients		
Noise Level	Method	Correct	Incorrect	E(MADE)
	one-step	19.90	0	0.0829
$\sigma_1 = 1,\ \sigma_2 = 2$	ordinary ($\delta = 0.01$)	20	0.02	0.0813
	ordinary ($\delta = 0.001$)	19.98	0	0.0784
	oracle	20	0	0.0795
	one-step	18.89	0	0.1870
$\sigma_1 = 1,\ \sigma_2 = 3$	ordinary ($\delta = 0.01$)	19.35	0.21	0.1975
	ordinary ($\delta = 0.001$)	18.51	0.06	0.1841
	oracle	20	0	0.1611
	one-step	19.68	0	0.0899
$\sigma_1 = 2,\ \sigma_2 = 2$	ordinary ($\delta = 0.01$)	19.97	0.07	0.0938
	ordinary ($\delta = 0.001$)	15.59	0	0.0947
	oracle	20	0	0.0857
	one-step	17.36	0	0.1838
$\sigma_1 = 2,\ \sigma_2 = 3$	ordinary ($\delta = 0.01$)	19.99	2.69	0.4202
	ordinary ($\delta = 0.001$)	11.01	0.10	0.2150
	oracle	20	0	0.1643

$\hat{\gamma}_l$ to zero for $\|\hat{\gamma}_l\| \le \delta \sum_{l=1}^{L} \|\hat{\gamma}_l\|_2$ was taken as 10^{-2} and 10^{-3}. For the weights w_i, we used $w_i = (n \times n_i)^{-1}$.

We computed the average numbers of the coefficients that were estimated to be 0. Also, we calculated the mean absolute deviation of errors

$$\text{MADE} = \sum_{l=1}^{24} n_{\text{gr}}^{-1} \sum_{r=1}^{n_{\text{gr}}} \frac{|\bar{\beta}_l(t_r) - \beta_l(t_r)|}{\text{range}(\beta_l)},$$

where the $\bar{\beta}_l$ denote the one-step gSCAD, the ordinary gSCAD, or the oracle estimates, and t_r for $1 \le r \le n_{\text{gr}}$ with $n_{\text{gr}} = 301$ are the equally spaced grid points on the support of t_{ij}. These results are reported in Table 1. In the table, those in the column labeled “Correct” indicate the average number, out of 100 replications, of $\bar{\beta}_l = 0$ among the estimates for the 20 truly zero coefficients β_l, $4 \le l \le 23$, and those in the column labeled “Incorrect” give that number among the estimates for the truly nonzero coefficients β_l, $1 \le l \le 3$.

From the table one sees that the one-step gSCAD estimator selected all relevant variables in every run. The ordinary gSCAD procedure sometimes estimated the nonzero coefficients as zero, especially when $\delta = 0.01$. In terms of identifying irrelevant variables, the results suggest that performance of the ordinary gSCAD estimator is sensitive to the choice of δ, the larger it is the higher the noise and

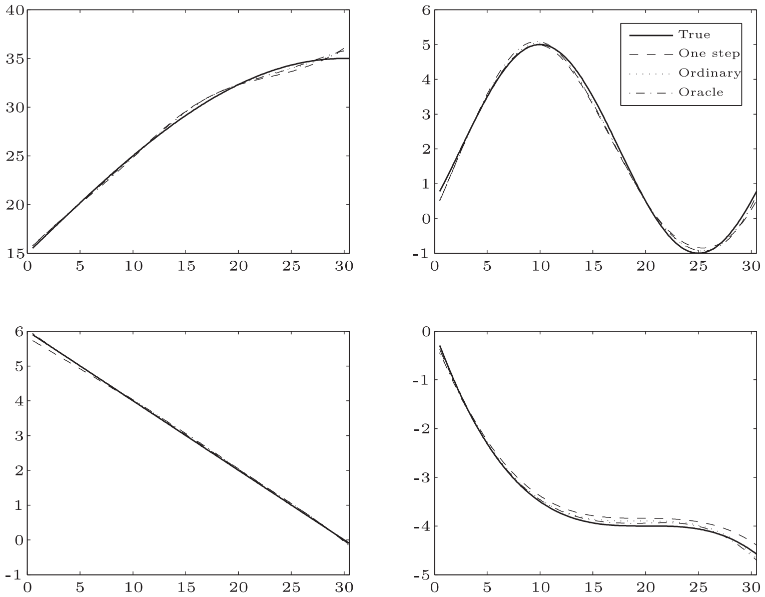

Figure 1. Mean estimated curves for the coefficient functions β_0 (top-left), β_1 (top-right), β_2 (bottom-left), β_3 (bottom-right), when $\sigma_1 = 1$, $\sigma_2 = 2$, and $n = 100$. The cut-off value δ for the ordinary gSCAD estimates was 0.01.

correlation levels. In particular, one can find that a good choice of δ for selecting relevant variables is a bad choice for identifying irrelevant variables, and vice versa. Our one-step gSCAD estimator does not need to choose the thresholding parameter δ, and its performance is competitive to the best case of the ordinary gSCAD. Furthermore, in terms of MADE, the one-step gSCAD estimator outperforms the ordinary gSCAD at higher noise and correlation levels.

Figures 1 and 2 depict the mean and variance of the estimated coefficient functions when $\sigma_1 = 1$ and $\sigma_2 = 2$. The results suggest that the bias and variance performance of the one-step gSCAD estimator is very similar to that of the oracle estimator. In particular, the variance property of the one-step gSCAD estimator is better than that of the ordinary gSCAD estimator, and is closer to that of the oracle estimator. This was also found to be the case for other correlation and noise levels.

5. A Data Example

We illustrate the one-step gSCAD method in an analysis of a dataset from the Multicenter AIDS Cohort study. The dataset consists of 283 homosexual males who were HIV positive between the years 1984 and 1991. During the study, all participants were scheduled to have their measurements taken every six months,

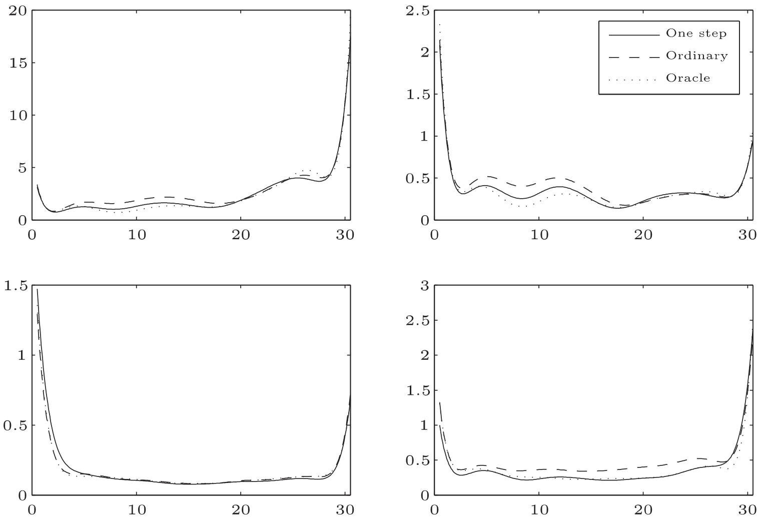

Figure 2. Variance curves of the estimators of the coefficient functions β_0 (top-left), β_1 (top-right), β_2 (bottom-left), β_3 (bottom-right), when $\sigma_1 = 1$, $\sigma_2 = 2$, and $n = 100$. The cut-off value δ for the ordinary gSCAD estimates was 0.01.

but it often happened that patients missed or rescheduled their appointments. Each individual had a different number of repeated measurements at different times. It is known that HIV destroys CD4 cells, so by measuring CD4 cell counts and percentages in the blood, doctors are able to monitor progression of the disease. The aim of our statistical analysis was to identify the covariates that influence the mean CD4 percentage depletion after infection, and to describe their effects on the mean CD4 percentage depletion over time.

Let t_{ij} be the time (in years) of the jth measurement for the ith individual after HIV infection. The response variable Y_{ij} is the CD4 percentage for the ith individual measured at t_{ij}. Three covariates were collected: X_{i1} denotes the ith individual's smoking status, 1 for smoker and 0 for nonsmoker; $X_{i2}(t_{ij})$ is the ith individual's centered age at time t_{ij}; X_{i3} is the ith individual's centered pre-infection CD4 cell percentage. For the analysis of this dataset, we considered the varying coefficient model

$$Y_{ij} = \beta_0(t_{ij}) + X_{i1}\beta_1(t_{ij}) + X_{i2}(t_{ij})\beta_2(t_{ij}) + X_{i3}\beta_3(t_{ij}) + \epsilon_{ij}.$$

For numerical tractability we used cubic splines with equally spaced knots, and used the same number of knots for estimating different varying coefficients. The cut-off value δ to set the ordinary gSCAD estimator $\check{\gamma}_l$ to zero for $\|\check{\gamma}_l\| \le \delta \sum_{l=1}^{L} \|\check{\gamma}_l\|_2$ was 10^{-3}. From the cross-validation criterion mentioned in Section 2, the number of knots was chosen to be 5 for both the one-step and the

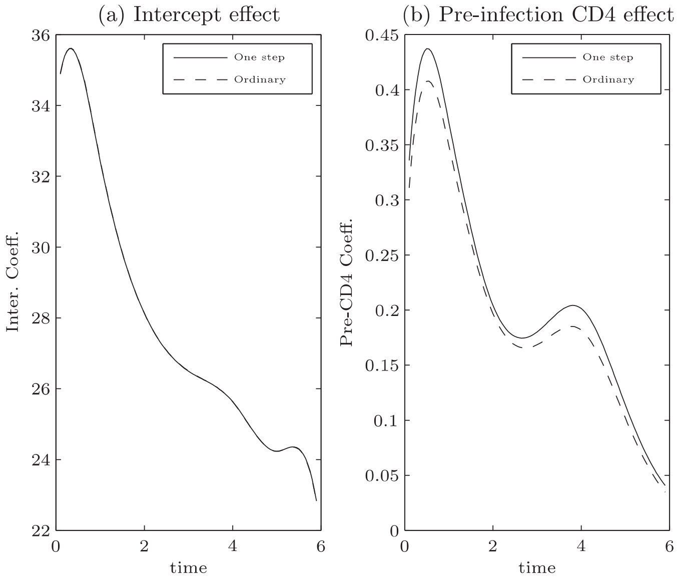

Figure 3. Fitted varying coefficient functions for the AIDS data.

ordinary gSCAD estimators, and the regularization parameters were selected to be 6.9 and 6 for the one-step and the ordinary gSCAD estimators, respectively. The weights $w_i = (n \times n_i)^{-1}$ were applied for both estimators.

Both gSCAD procedures identified the pre-infection CD4 cell percentage as an influential factor on the mean CD4 percentage depletion, but neither of them selected age and smoking status. This is consistent with the result of Huang, Wu and Zhou (2002). The fact that smoking status was not selected as an influential factor to AIDS progression coincides with the conclusion of Furber et al. (2007). While age is known to be related to AIDS progression (for example, see Zingmond et al. (2001)), age was not selected by the procedure. Figure 3 depicts the fitted coefficient functions of $\beta_0(t)$ and $\beta_3(t)$ by the two methods. We observed that if we set the cut-off value δ to 10^{-2} for the ordinary gSCAD estimator, the pre-infection CD4 cell percentage was not selected by the ordinary gSCAD procedure. This suggests that the one-step gSCAD procedure is more stable than, and improves upon, the ordinary gSCAD procedure.

6. Extensions

We have proposed and studied the one-step gSCAD procedure for model selection and estimation in varying coefficient models. The one-step gSCAD method may be extended to other nonparametric models, for example to generalized linear models. We describe the method for the *varying-coefficient logistic* *regression model*, where $Y_i(t_{ij}) \in \{0,1\}$ are the observed responses and $\mathbf{X}_i(t_{ij})$ are the observed vectors of the covariate processes. In this case, the w_i-weighted log-likelihood of the varying-coefficient vector $\boldsymbol{\beta}$ is

$$\ell(\boldsymbol{\beta},\mathbf{Z}) = \sum_{i=1}^{n} w_i \sum_{j=1}^{n_i} \Big[Y_i(t_{ij})\mathbf{X}_i(t_{ij})^\top \boldsymbol{\beta}(t_{ij}) - \log\Big\{ 1 + \exp\Big(\mathbf{X}_i(t_{ij})^\top \boldsymbol{\beta}(t_{ij})\Big)\Big\}\Big],$$

where $\mathbf{Z} \equiv \{(Y_i(t_{ij}), \mathbf{X}(t_{ij})) : 1 \le j \le n_i,\ 1 \le i \le n\}$. Approximating β_l by $\beta_l \approx \sum_{k=1}^{K_l} \gamma_{lk} B_{lk}$, we obtain the following penalized log-likelihood of $\boldsymbol{\gamma}$:

$$l_a(\boldsymbol{\gamma}) = \ell\bigg(\Big\{\sum_{k=1}^{K_l} \gamma_{lk} B_{lk}\Big\}_{l=1}^{L}, \mathbf{Z}\bigg) - \sum_{l=1}^{L} p'_\lambda(\|\boldsymbol{\gamma}_l^0\|_w)\|\boldsymbol{\gamma}_l\|_w, \qquad (6.1)$$

where $p_\lambda(\cdot)$ is the SCAD penalty function with λ as a regularization parameter and $\hat{\boldsymbol{\gamma}}^0$ is an initial estimator of $\boldsymbol{\gamma}$. The one-step gSCAD regularized estimator of $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_L^\top)^\top$ is then defined as the maximizer of l_a.

Although l_a is a convex function of $\boldsymbol{\gamma}$, its maximization involves a nonlinear optimization problem due to the first term in its definition at (6.1). One can use numerical methods for the nonlinear optimization. For example, one can linearize the problem by a Newton-Raphson method, approximating ℓ by a quadratic function locally around the initial choice of $\hat{\boldsymbol{\gamma}}^0$. This would involve an iterative algorithm. In the iteration, one can use the updated estimates of $\boldsymbol{\gamma}$ in the approximation of ℓ only, or use them as well in the linearization of the SCAD penalty. Theoretical properties of the estimator are yet to be developed.

7. Proof of Theorem 3.1

Let $\tilde{Y}_{ij} = \mathbf{X}_i(t_{ij})^\top \boldsymbol{\beta}(t_{ij})$, $\tilde{\mathbf{Y}}_i = (\tilde{Y}_{i1}, \ldots, \tilde{Y}_{in_i})^\top$, $\tilde{\mathbf{Y}} = (\tilde{\mathbf{Y}}_1^\top, \ldots, \tilde{\mathbf{Y}}_n^\top)^\top$, and $\tilde{\boldsymbol{\gamma}} = \big(\mathbf{U}^\top\mathbf{W}\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{W}\tilde{\mathbf{Y}}$, where $\mathbf{U}$ and $\mathbf{W}$ are defined in Section 2. Also, take $\boldsymbol{\epsilon}_i = (\epsilon_i(t_{i1}), \ldots, \epsilon_i(t_{in_i}))^\top$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^\top, \ldots, \boldsymbol{\epsilon}_n^\top)^\top$ and $\tilde{\boldsymbol{\beta}}(t) = \mathbf{B}(t)\tilde{\boldsymbol{\gamma}}$. We write $a_n \sim b_n$ if both a_n and b_n are positive and a_n/b_n and b_n/a_n are bounded.

Lemma 7.1. *Suppose that* (3.1) *holds. Then, there exist constants* $0 < M_5 < M_6 < \infty$ *such that, as* $n \to \infty$,

$$P\{\text{all the eigenvalues of } K^\alpha \mathbf{U}^\top\mathbf{W}\mathbf{U} \text{ fall in } [M_5, M_6]\} \longrightarrow 1.$$

Proof. For $g_l = \sum_{k=1}^{K_l} \gamma_{lk} B_{lk}$, $1 \le l \le L$, we have from (A5) and (A6) that

$$\sum_{l=1}^{L} \|g_l\|^2 = \sum_{l=1}^{L} \|\boldsymbol{\gamma}_l\|_w^2 \sim \sum_{l=1}^{L} K_l^{-\alpha}\|\boldsymbol{\gamma}_l\|_2^2 \sim K^{-\alpha}\|\boldsymbol{\gamma}\|_2^2.$$

Following the proofs of Lemmas A1 and A2 in Huang, Wu and Zhou (2002) using the approximation $\sum_{l=1}^{L} \|g_l\|^2 \sim K^{-\alpha}\|\boldsymbol{\gamma}\|_2^2$ leads to the result.

Lemma 7.2. *Suppose that* (3.1) *holds. Then,* $\|\hat{\gamma}-\tilde{\gamma}\|_w = O_p(r+\sqrt{\lambda\rho})$

Proof. Let $\hat{\gamma}-\tilde{\gamma} = \delta K^{\alpha/2}\mathbf{u}$ where δ is a scalar and $\mathbf{u}$ is a vector such that $\|\mathbf{u}\|_2 = 1$. Observe that

$$\begin{aligned} l_a(\hat{\gamma}) - l_a(\tilde{\gamma}) &= -2K^{\alpha/2}\delta\boldsymbol{\epsilon}^\top \mathbf{WUu} + K^{\alpha}\delta^2\mathbf{u}^\top\mathbf{U}^\top\mathbf{WUu} \\ &\qquad + \sum_{l=1}^{L} p'_\lambda(\|\hat{\gamma}_l^0\|_w)(\|\hat{\gamma}_l\|_w - \|\tilde{\gamma}_l\|_w). \end{aligned}$$

From (A1), (A3), (A4), and (A6), we obtain

$$\begin{aligned} E(|\mathbf{U}_i^\top \mathbf{W}_i\boldsymbol{\epsilon}_i|^2) &= E\Big[\sum_{l=1}^{L}\sum_{k=1}^{K_l} w_i^2\Big\{\sum_{j=1}^{n_i} X_{il}(t_{ij})B_{lk}(t_{ij})\epsilon_i(t_{ij})\Big\}^2\Big] \\ &\le M_1^2M_2^2 n_i w_i^2 \sum_{l=1}^{L}\sum_{j=1}^{n_i} E\Big[\sum_{k=1}^{K_l} B_{lk}(t_{ij})^2\Big] \\ &\le \sup_{t\in\mathcal{T}} f_T(t) M_1^2M_2^2M_4 n_i^2 w_i^2 \sum_{l=1}^{L} K_l^{1-\alpha}. \end{aligned}$$

This implies $E(\boldsymbol{\epsilon}^\top\mathbf{WUU}^\top\mathbf{W}\boldsymbol{\epsilon}) = O(K^{-\alpha}r^2)$, so that

$$\boldsymbol{\epsilon}^\top\mathbf{WUu}\delta = O_p\left(K^{-\alpha/2}r\right)\delta. \tag{7.1}$$

By Lemma 5.1, we have $\mathbf{u}^\top\mathbf{U}^\top\mathbf{WUu} \ge M_5K^{-\alpha}$. From (A6) and the triangular inequality for the ℓ_2-norm, and since p' is bounded by λ, we obtain

$$\sum_{l=1}^{L} p'_\lambda(\|\hat{\gamma}_l^0\|_w)(\|\hat{\gamma}_l\|_w - \|\tilde{\gamma}_l\|_w) \ge -(\text{const.})\lambda\|\hat{\gamma}-\tilde{\gamma}\|_w.$$

From $l_a(\hat{\gamma}) - l_a(\tilde{\gamma}) \le 0$, we get

$$\begin{aligned} M_5\delta &\le 2\,K^{\alpha/2}\boldsymbol{\epsilon}^\top\mathbf{WUu} - \sum_{l=1}^{L} p'_\lambda(\|\hat{\gamma}_l^0\|_w)(\|\hat{\gamma}_l\|_w - \|\tilde{\gamma}_l\|_w)/\delta \\ &\le O_p(r) + (\text{const.})\lambda. \end{aligned}$$

This yields $\|\hat{\gamma}-\tilde{\gamma}\|_w \le M_4^{1/2}\delta = O_p(r+\lambda)$.

One can show that Lemma A3 of Huang, Wu and Zhou (2002) continues to hold for a non-orthonormal basis under the assumption (A6), that is, for $\hat{\beta}_l^0 \equiv \sum_{k=1}^{K_l}\hat{\gamma}_{lk}^0 B_{lk}$,

$$\|\hat{\beta}_l^0 - \tilde{\beta}_l\| = O_p(r), \quad \|\tilde{\beta}_l - \beta_l\| = O_p(\rho). \tag{7.2}$$

This gives

$$\|\hat{\gamma}_l^0\|_w = \|\beta_l\| + o_p(1),\ 1 \le l \le s, \tag{7.3}$$

$$\|\hat{\gamma}_l^0\|_w = O_p(r+\rho),\ (s+1) \le l \le L, \tag{7.4}$$

$$\|\tilde{\gamma}_l\|_w = O_p(\rho),\ (s+1) \le l \le L. \tag{7.5}$$

From (7.3) it follows that $\|\hat{\gamma}_l^0\|_w > a\lambda$ for all $1 \le l \le s$ with probability tending to one, where a appears in the definition of p_λ. This means that, with probability tending to one, $p'_\lambda(\|\hat{\gamma}_l^0\|_w) = 0$ for all $1 \le l \le s$. Since $\|\hat{\gamma} - \tilde{\gamma}\|_w = O_p(r+\lambda) = o_p(1)$, so that $\|\hat{\gamma}_l\|_w - \|\tilde{\gamma}_l\|_w = o_p(1)$, it follows from the definition of p_λ that

$$\sum_{l=1}^{s} p'_\lambda(\|\hat{\gamma}_l^0\|_w)(\|\hat{\gamma}_l\|_w - \|\tilde{\gamma}_l\|_w) = o_p(\kappa_n) \tag{7.6}$$

for any sequence $\{\kappa_n\}$ such that $\kappa_n \to 0$ as $n \to \infty$. Also, from (7.4) and the condition that $\lambda/\max\{r,\rho\} \to \infty$ as $n \to \infty$, we have $\|\hat{\gamma}_l^0\|_w < \lambda$ with probability tending to one, so that

$$\begin{aligned}\sum_{l=s+1}^{L} p'_\lambda(\|\hat{\gamma}_l^0\|_w)(\|\hat{\gamma}_l\|_w - \|\tilde{\gamma}_l\|_w) &\ge -\sum_{l=s+1}^{L} p'_\lambda(\|\hat{\gamma}_l^0\|_w)\|\tilde{\gamma}_l\|_w \\ &= -\lambda \sum_{l=s+1}^{L} \|\tilde{\gamma}_l\|_w\end{aligned} \tag{7.7}$$

with probability tending to one. Putting (7.5)−(7.7) together, we obtain

$$\begin{aligned}M_5\delta^2 &\le 2\,K^{\alpha/2}\delta\boldsymbol{\epsilon}^\top \mathbf{WUu} - \sum_{l=1}^{L} p'_\lambda(\|\hat{\gamma}_l^0\|_w)(\|\hat{\gamma}_l\|_w - \|\tilde{\gamma}_l\|_w) \\ &\le O_p(r\delta + \lambda\rho),\end{aligned} \tag{7.8}$$

which concludes the proof of the lemma.

Proof of Theorem 3.1 (a). Suppose that there exists an $l_0, (s+1) \le l_0 \le L$, such that the probability of $\hat{\beta}_{l_0} \equiv 0$ does not converge to one. Then there exists $\varepsilon > 0$ such that, for infinitely many n, $P(\hat{\gamma}_{l_0} \ne \mathbf{0}) = P(\hat{\beta}_{l_0} \ne 0) \ge \varepsilon$. Let $\boldsymbol{\gamma}^*$ be the vector obtained from $\hat{\boldsymbol{\gamma}}$ with $\hat{\gamma}_{l_0}$ replaced by $\mathbf{0}$. Then, from Lemma 5.2, (A6), and (7.2), we have $\|\boldsymbol{\gamma}^* - \hat{\boldsymbol{\gamma}}\|_w = \|\hat{\gamma}_{l_0}\|_w = O_p(r + \sqrt{\lambda\rho} + \rho)$. This together with (7.1), (A6), the two lemmas, and the fact that $p'_\lambda(\|\hat{\gamma}_{l_0}^0\|_w) = \lambda$ with probability tending to one, gives

$$\begin{aligned}l_a(\hat{\boldsymbol{\gamma}}) - l_a(\boldsymbol{\gamma}^*) &= -2(\mathbf{Y} - \mathbf{U}\tilde{\boldsymbol{\gamma}})^\top \mathbf{WU}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) - 2(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)^\top \mathbf{U}^\top \mathbf{WU}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \\ &\quad + (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)^\top \mathbf{U}^\top \mathbf{WU}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) + p'_\lambda(\|\hat{\gamma}_{l_0}^0\|_w)\|\hat{\gamma}_{l_0}\|_w \\ &\ge O_p(r)\|\hat{\gamma}_{l_0}\|_w + O_p(\|\hat{\gamma}_{l_0}\|_w^2) + \lambda\|\hat{\gamma}_{l_0}\|_w.\end{aligned}$$

Since $\lambda\|\hat{\gamma}_{l_0}\|_w > 0$ dominates the other two terms in the above inequality, one contradicts $l(\hat{\gamma}) - l(\gamma^*) \le 0$.

Proof of Theorem 3.1 (b). Let $\tilde{\gamma}^*$ be the minimizer of $(\tilde{\mathbf{Y}} - \mathbf{U}\gamma)^\top \mathbf{W}(\tilde{\mathbf{Y}} - \mathbf{U}\gamma)$ over γ, with the constraints $\gamma_l = \mathbf{0}$ for $(s+1) \le l \le L$. From the first part of the theorem, we have $P(\hat{\gamma}_l = \tilde{\gamma}_l^* \text{ for all } (s+1) \le l \le L) \to 1$. By applying (7.8) with $\tilde{\gamma} = \tilde{\gamma}^*$, we obtain that, with probability tending to one,

$$M_5\|\hat{\gamma} - \tilde{\gamma}^*\|_w^2 \le 2\,\boldsymbol{\epsilon}^\top \mathbf{W}\mathbf{U}(\hat{\gamma} - \tilde{\gamma}^*) = O_p(r\|\hat{\gamma} - \tilde{\gamma}^*\|_w).$$

This shows $\|\hat{\gamma} - \tilde{\gamma}^*\|_w = O_p(r)$ and thus $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\| = O_p(r)$. From the triangle inequality, the second part of the theorem follows.

Acknowledgement

We thank the Editor, an associate editor, and two referees for their constructive comments on an earlier version of this paper. This work was supported by the Korea Science and Engineering Foundation (KOSEF) funded by the Korea government (MEST 2009-0052815).

References

Bach, F. (2008). Consistency of the group Lasso and multiple kernel learning. *J. Machine Learning Research* **9**, 1179-1225.

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

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99**, 710-723.

Fan, J. and Zhang, J. T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J. Roy. Statist. Soc. Ser. B* **62**, 303-322.

Furber, A. S., Maheswaran, R., Newell, J. N. and Carroll, C. (2007). Is smoking tobacco an independent risk factor for HIV infection and progression to AIDS? A systemic review. *Sexually Transmitted Infections* **83**, 41-46.

Grant, M. and Boyd, S. (2008). CVX: Matlab software for disciplined convex programming (web page and software). `http://stanford.edu/~boyd/cvx`.

Hall, P., Lee, E. R. and Park, B. U. (2009). Bootstrap-based penalty choice for the lasso, achieving oracle performance. *Statist. Sinica* **19**, 449-471.

Hoover, D. R, Rice, J. A, Wu, C. O. and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809-822.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional anova models, *Ann. Statist.* **26**, 242-272.

Huang, J. Z., Wu, C. O. and Zhou, L. (2002). Varying-coefficient models and basis function approximation for the analysis of repeated measurements. *Biometrika* **89**, 112-128.

Huang, J. Z., Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14**, 763-788.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 1617-1642.

Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **57**, 425-437.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53**, 233-243.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253-259.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1348-1360.

Storlie, C., Bondell, H., Reich, B. and Zhang, H. H. (2010). Surface estimation, variable selection, and the nonparametric oracle property. *Statist. Sinica.* In Press.

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

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

Xue, L. (2009). Consistent variable selection in additive models. *Statist. Sinica* **19**, 1281-1296.

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

Zingmond, D. S., Wenger, N. S., Crystal, S., Joyce, G. F., Liu, H., Sambamoorthi, U., Lillard, L. A., Leibowitz, A. A., Shapiro, M. F. and Bozzette, S. A. (2001). Circumstances at HIV diagnosis and progression of disease in older HIV-infected Americans. *American Journal of Public Health* **91**, 1117-1120.

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

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36**, 1509-1566.

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36**, 1108-1126.

Department of Statistics, Seoul National University, Seoul 151-747, Korea.

E-mail: word5810@snu.ac.kr

Department of Statistics, Seoul National University, Seoul 151-747, Korea.

E-mail: bupark@stats.snu.ac.kr

(Received September 2008; accepted March 2009)