Statistica Sinica **22** (2012), 1563-1588
doi:http://dx.doi.org/10.5705/ss.2010.308

VARIABLE SELECTION FOR HIGH-DIMENSIONAL GENERALIZED VARYING-COEFFICIENT MODELS

Heng Lian

Nanyang Technological University

Abstract: In this paper, we consider the problem of variable selection for high-dimensional generalized varying-coefficient models and propose a polynomial-spline based procedure that simultaneously eliminates irrelevant predictors and estimates the nonzero coefficients. In a "large p, small n" setting, we demonstrate the convergence rates of the estimator under suitable regularity assumptions. In particular, we show the adaptive group lasso estimator can correctly select important variables with probability approaching one and the convergence rates for the nonzero coefficients are the same as the oracle estimator (the estimator when the important variables are known before carrying out statistical analysis). To automatically choose the regularization parameters, we use the extended Bayesian information criterion (eBIC) that effectively controls the number of false positives. Monte Carlo simulations are conducted to examine the finite sample performance of the proposed procedures.

Key words and phrases: Diverging parameters, group lasso, polynomial splines, quasi-likelihood.

1. Introduction

Regression analysis where investigators are interested in the relationships between a set of predictors and the responses is of utmost importance in statistics, with linear regression the oldest, the simplest, and the most popular approach. Generalized linear models (GLM) provide an extension of linear models in dealing with different types of responses, including for example binary data and count data (McCullagh and Nelder (1989)). Let Y be a response variable and suppose the (conditional) mean of the response, μ, depends on the p-dimensional predictors $X = (X_1, \ldots, X_p)$ through

$$g(E[Y|X]) = g(\mu) = X^T\beta, \tag{1.1}$$

where g is a known link function and $\beta = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown regression coefficients. The variance of Y is typically a function of the mean, that is,

$$Var(Y|X) = V(\mu) = V(g^{-1}(X^T\beta)).$$

However, such parametric models are not flexible enough to capture the true underlying relationships between covariates and responses. Of particular interests to us is the generalized varying-coefficient models (GVCM) (Hastie and Tibshirani (1993); Cai, Fan, and Li (2000)) in which the coefficients β in GLMs are replaced with smooth nonparametric functions that depends on an index variable T, resulting in

$$g(E[Y|X,T]) = g(\mu) = X^T\alpha(T), \tag{1.2}$$

where $\alpha(T) = (\alpha_1(T), \ldots, \alpha_p(T))^T$. The index variable T is usually some variable related to time or age in many applications, and whose interactions with other predictors is believed to be of importance.

High-dimensionality is an important characteristic of many modern data sets. However, even with many predictors available to be included in an initial modeling, many of them may not be significant and their inclusion only decreases the accuracy of prediction.

Recent challenging topics in statistics include the development of automatic variable selection procedures intended to automatically find the relevant parameters among all candidate parameters and simultaneously estimate them. As argued in Li and Liang (2008), traditional variable selection methods, such as stepwise regression and best subset selection, are computationally infeasible when the number of predictors is large, and this is part of the reason why the penalization based method has gained popularity in recent years. Substantial progress has been made on the problem of variable selection for linear models and generalized linear models (Tibshirani (1996); Fan and Li (2001); Fan and Peng (2004); Zou (2006); Zou and Li (2008); Yuan and Lin (2007); Huang, Horowitz, and Ma (2008); Choi, Li, and Zhu (2010); Li, Peng, and Zhu (2011)). In particular, the adaptive group lasso was proposed in Wang and Leng (2008) and is the penalty we use here. More recently, variable selection methods using penalty functions in nonparametric or semiparametric settings have been developed. For example, Xie and Huang (2009) developed variable selection based on penalization for partially linear models, and Ravikumar et al. (2008); Meier, Van de Geer, and Bühlmann (2009); Huang, Horowitz, and Wei (2010) independently investigated the additive models. For generalized varying-coefficient partial linear models, Li and Liang (2008) used penalization to select the significant predictors in the parametric components while the nonparametric components were selected by hypothesis testing. Note that this work is for the fixed p case. Lam and Fan (2008) studied varying-coefficient partially linear models with a diverging number of parametric components but they did not investigate the variable selection problem. For the ordinary varying-coefficients models with quadratic loss function, Wang, Li, and Huang (2008) and Wang and Xia (2009) each proposed a group penalization method in the fixed p case, and Wei, Huang, and Li (2011)

recently extended this work to the case of diverging p. These previous works motivated us to develop a penalization based approach for variable selection in GVCMs. Thus our work is a natural extension of Wei, Huang, and Li (2011) to more general types of responses using quasi-likelihood as opposed to varying-coefficient models with least squares loss. Besides, the theoretical proofs seem more difficult with quasi-likelihood. Finally, we provide the consistency proof of eBIC while no corresponding results are stated in their work.

This paper is organized as follows. In Section 2, we propose a penalization procedure for coefficient estimation and variable selection. Unlike Li and Liang (2008) and Lam and Fan (2008), which are based on local polynomial regression, we use polynomial splines to approximate the nonparametric coefficients. This is computationally easier since it directly reduces the nonparametric model to a parametric GLM as far as computations are concerned. The regularization parameter is automatically chosen using the extended Bayesian information criterion (eBIC) (Chen and Chen (2008)). In Section 3, Monte Carlo simulation studies are carried out for the Poisson regression and logistic regression models to demonstrate the performance of the proposed method. In addition, a data set is used as an illustration of varying-coefficient logistic regression models. The Appendix contains all technical proofs.

2. Spline Estimator and Sampling Properties

The data we observe for the ith subject or unit are $(X_i, T_i, Y_i), i = 1, \ldots, n$, where the $X_i = (X_{i1}, \ldots, X_{ip})^T$ are the predictors and the T_i are index variables. The true model is assumed to be that of (1.2) with p potentially much bigger than n. (As shown at the end of Section 2.2, if the number of nonzero coefficients is bounded, we can take $p = o(\exp\{n^{d/(2d+1)}\})$. However, we assume a sparse model with only s significant predictors, denoted by $X^{(1)} = (X_1, \ldots, X_s)$, and the other $p-s$ predictors do not appear in the true model. We denote the nonzero coefficients by $\alpha_0 = (\alpha_{01}, \ldots, \alpha_{0s})^T$. Note that for simplicity the index variable T is assumed to be univariate with a distribution supported on the interval $[0, 1]$. The extension to multi-dimensional T is possible but rarely used in practice due to the so-called "curse of dimensionality".

2.1. Estimation and variable selection based on adaptive group lasso penalty

The (negative) quasi-likelihood function is defined by

$$Q(\mu, y) = \int_{\mu}^{y} \frac{y - s}{V(s)}\, ds,$$

and the negative quasi-likelihood of the observed n i.i.d. data is

$$\sum_{i=1}^{n} Q(g^{-1}(X_i^T \alpha(T_i)), Y_i).$$

The attractiveness and popularity of quasi-likelihood largely lies in that estimation is still consistent even if the variance function is misspecified. In this paper, we take "likelihood" to mean quasi-likelihood.

We use polynomial B-splines to approximate the varying coefficients $\alpha_j(t)$, $1 \le j \le p$. To approximate a function on $[0,1]$, we partition the interval $[0,1]$ into K' subintervals $[(k-1)/K', k/K']$, for $k = 1, 2, \ldots, K'$ with $K' = K'(n)$ being a sequence of natural numbers diverging to infinity as sample size n goes to infinity. A polynomial spline of order q is a function whose restriction to each subinterval is a polynomial of degree $q-1$ and is globally $q-2$ times differentiable. The collection of such polynomial splines has a normalized B-spline basis $\{B_1(t), \ldots, B_K(t)\}$ with $K = K' + q$. As in De Boor (2001), the basis satisfies $B_k \ge 0$, $k = 1, \ldots, K$, $\sum_{k=1}^{K} B_k(t) \equiv 1$, and B_j is supported inside an interval of length q/K and at most q of the basis functions are nonzero at any given t. Using spline expansions, we can approximate the coefficients by $\alpha_j(t) \approx \sum_k a_{jk} B_k(t)$. It is also possible to construct irregular subintervals based on observed values of the index variable, or to specify different K for different coefficient, but we make the above choices for computational and theoretical simplicity.

We start with a model where all coefficients are potentially nonzero, with negative likelihood given by

$$\sum_{i=1}^{n} Q(g^{-1}(\sum_{j=1}^{p}\sum_{k=1}^{K} X_{ij} a_{jk} B_k(T_i)), Y_i) = \sum_{i=1}^{n} Q(g^{-1}(Z_i^T a), Y_i),$$

where $Z_i = (X_{i1}B_1(T_i), \ldots, X_{i1}B_K(T_i), \ldots, X_{ip}B_K(T_i))^T$, and $a = (a_1^T, \ldots, a_p^T)^T$ $= (a_{11}, \ldots, a_{1K}, a_{21}, \ldots, a_{pK})^T$. The adaptive group lasso penalty is used to encourage shrinkage to zero coefficients. For any $1 \le j \le p$, since $\sum_k a_{jk} B_k(t) \equiv 0$ if and only if $a_{jk} = 0$, for all $1 \le k \le K$, or equivalently $\|a_j\| = 0$ ($\|.\|$ is the l_2 norm), the group lasso penalty $\sum_{j=1}^{n} \|a_j\|$ can be used to identify zero coefficients, as done in Yuan and Lin (2006). Thus we propose the estimation procedure based on penalized negative likelihood,

$$\hat{a} = \arg\min_{a} \sum_{i=1}^{n} Q(g^{-1}(Z_i^T a), Y_i) + n\lambda \sum_{j=1}^{p} w_j \|a_j\|, \tag{2.1}$$

where λ is a regularization parameter controlling the amount of shrinkage, $w = (w_1, \ldots, w_p)$ is a given vector of weights. Intuitively, w_j should be large if α_j is

actually zero to encourage more shrinkage. In this subsection we assume these weights are given and known, while in practice we estimate them based on an initial estimator, as discussed in the next subsection.

If we have the knowledge of which coefficients are zeros, we can optimize the similar penalized functional with all insignificant predictors removed:

$$\hat{a}^{(1)} = \arg\min_{a^{(1)}} \sum_{i=1}^{n} Q(g^{-1}(Z_i^{(1)^T} a^{(1)}), Y_i) + n\lambda \sum_{j=1}^{s} w_j \|a_j^{(1)}\|, \tag{2.2}$$

where $Z_i^{(1)}$ is the sK-dimensional subvector of Z_i corresponding to nonzero coefficients. We take $Z_i^{(2)}$, which is associated with zero coefficients, so that $Z_i^T = (Z_i^{(1)T}, Z_i^{(2)T})$. Other variables with these superscripts are interpreted similarly.

Since the weights $w_j, s+1 \le j \le p$ are associated with the zero coefficients and do not appear in the functional (2.2), it makes sense to take $\|w'\|^2 = \sum_{j=1}^{s} w_j^2$. A theorem gives the convergence rate of the estimator in (2.2), for which we need an assumption involving the dimensionality and the smoothing parameter. The parameter d that appears below is the smoothness parameter for $\alpha_{0j}, 1 \le j \le s$, as stated in condition (c7) in the Appendix.

Assumption (A).

$$(Ks)^{3/2} \sqrt{\frac{Ks}{n} + \frac{s^2}{K^{2d}} + \lambda^2 K \|w'\|^2} \to 0.$$

As discussed in Section 2.2, for appropriately chosen weights w_j, the term $\lambda^2 K\|w'\|^2$ is no bigger than Ks/n and can be ignored in (A). If s is bounded, the usual choice $K \sim n^{1/(2d+1)}$ balances bias and variance in the convergence rates stated below, and assumption (A) reduces to $n^{4/(2d+1)}/n \to 0$. Thus (A) is satisfied if $d > 3/2$, in particular if α_{0j} is twice differentiable.

Theorem 1. *Under the regularity conditions* (c1)−(c8) *in the Appendix, as well as* (A)*, the estimator* $\hat{a}^{(1)}$ *in* (2.2) *satisfies*

$$\sum_{j=1}^{s} \|\hat{\alpha}_j(t) - \alpha_{0j}(t)\|^2 = O_P\Big(\frac{Ks}{n} + \frac{s^2}{K^{2d}} + \lambda^2 Ks\|w'\|^2\Big), \tag{2.3}$$

where $\hat{\alpha}_j(t) = \sum_k \hat{a}_{jk}^{(1)} B_k(t)$.

The next theorem shows that the estimator from (2.1), which does not assume knowledge of the zero coefficients, is exactly equal to the estimator from (2.2), with probability converging to 1. Thus the convergence rates for the estimator (2.1) are as stated in Theorem 1. Extra conditions on the weights w_j are

needed as stated in the assumption below, it can be interpreted as the requirement that w_j is sufficiently large for zero coefficients.

Assumption (B).

$$\sqrt{n\log(p\vee n)}+\sqrt{\frac{n}{K}(Ks+\frac{ns^2}{K^{2d}}+nK\lambda^2\|w'\|^2)}=o(n\lambda w_j),\ s+1\le j\le p.$$

Theorem 2. *Under the regularity conditions* (c1)−(c9) *in the Appendix, as well as* (A) *and* (B), *suppose* $\hat{a}^{(1)}$ *is obtained from* (2.2) *and take* $\hat{a}=(\hat{a}^{(1)},\hat{a}^{(2)})$ *with* $\hat{a}^{(2)}_{jk}=0$ *for* $s+1\le j\le p, 1\le k\le K$. *Then* $\hat{a}$ *is the solution of* (2.1) *with probability converging to* 1.

2.2. Initial estimator based on the group Lasso

In the adaptive group lasso penalties, the weight w_j is generally desired to be large for zero coefficients and small for nonzero ones. Following Zou (2006), we can first obtain an initial estimator with the group lasso penalty (all weights set to be 1),

$$\tilde{a}=\arg\min_a\sum_i Q(g^{-1}(Z_i^Ta),Y_i)+\lambda_0\sum_{j=1}^p\|a_j\|, \tag{2.4}$$

and then set $w_j=1/\|\tilde{a}_j\|$.

Theorem 3. *Under* (c1)−(c7), (c9), (c10) *in the Appendix, if*

$$\frac{\lambda_0}{\max\{\sqrt{n\log(p\vee n)},\sqrt{n}s\}}\to\infty,\qquad\frac{s^2K^2\lambda_0}{n}\to 0,$$

then $\|\tilde{a}-a_0\|=O_P(\sqrt{s}K\lambda_0/n)$, *where* a_0 *contains the coefficients in the optimal approximation of* α_0 *in the spline basis expansion, which satisfies* $\|\sum_k a_{0jk}B_k(t)-\alpha_{0j}(t)\|=O(K^{-d}), 1\le j\le p$. *If* (c11) *holds, all coefficients except* Ms *of them are estimated as zeros for some constant* M *as* (c11) *in the Appendix.*

Now we discuss how condition (B) can be satisfied using the initial estimator (2.4). For simplicity of discussion we assume $\|a_{0j}\|/\sqrt{K}$ is bounded away from zero for $1\le j\le s$; this assumption is satisfied if, for example, the true coefficients $\alpha_0=(\alpha_{01},\ldots,\alpha_{0s})^T$ do not change with sample size. Choosing $\lambda_0\sim\sqrt{nb_n}\max\{\sqrt{\log(p\vee n)},s\}$ with some $b_n\to\infty$ arbitrarily slowly, the convergence rate of the initial estimator is $O_P(\sqrt{s}K\max\{\sqrt{\log(p\vee n)},s\}\sqrt{b_n}/\sqrt{n})$. If this convergence rate is $o(\sqrt{K})$ (if s is bounded, this condition just simplifies to $K\log(p\vee n)/n\to 0$), then $\|\tilde{a}_j-a_{0j}\|$ is of smaller order than $\|a_{0j}\|$

when $1 \le j \le s$ and thus $\|\tilde{a}_j\|/\sqrt{K}$ is also bounded away from zero, leading to $w_j = 1/\|\tilde{a}_j\| = O(1/\sqrt{K})$. This implies $\|w'\| = O_P(\sqrt{s/K})$ and $w_j, s+1 \le j \le p$ is at least of order $\sqrt{n/s}/(K \max\{\sqrt{\log(p \vee n)}, s\}\sqrt{b_n})$. Thus if

$$\lambda = O\Big(\sqrt{\frac{K}{n}}\Big), \tag{2.5}$$

the final term in (2.3) is small enough and can be ignored. Furthermore, taking $K \sim n^{1/(2d+1)}$, the usual choice that balances bias and variance in nonparametric regression, assumption (B) is equivalent to

$$\lambda \left(\frac{\sqrt{s^3}K}{n} \max\{\sqrt{\log(p \vee n)}, s\}\sqrt{b_n} \right)^{-1} \to \infty,$$
$$\lambda \left(\frac{\sqrt{s}K\sqrt{\log(p \vee n)}}{n} \max\{\sqrt{\log(p \vee n)}, s\}\sqrt{b_n} \right)^{-1} \to \infty.$$

Under different rates of divergence of s, with constraint on the size of p and s, λ can be found that satisfies the above, as well as (2.5). For example, if $s = O(1)$, then we require that p satisfies $\log p = o(n^{d/(2d+1)})$ in order for such λ to exist.

2.3. Tuning parameters selection and implementation

In practice, we need to choose some parameters including the spline order q, the number of basis terms K, as well as the regularization parameters λ_0 and λ. As is common, we fix $q = 4$ (cubic splines) in all our numerical results. When computing the initial group lasso estimator and the adaptive group lasso estimator, we fix $K = 8$. This strategy is similar to that commonly used in functional smoothing/functional data analysis literature where the number of knots is chosen to be sufficiently large so that approximation error is small and the overfitting can be effectively controlled by the penalization terms (see for example Chapter 5 of Ramsay and Silverman (2005)). Nevertheless, we conducted some simulation studies on the choice of K that suggested that the results are not too sensitive to its value.

The choice of λ in (2.1) is critical for the performance of the estimators. In our high-dimensional context, we adopt the extended Bayesian information criterion (eBIC) of Chen and Chen (2008) that was developed for parametric models. More specifically, we select λ that minimizes

$$\frac{2}{n}\sum_{i=1}^{n} Q(g^{-1}(Z_i\hat{a}_\lambda), Y_i) + d_1\frac{\log(n/K)}{n/K} + \frac{2}{n/K}\log\binom{p}{d_1}, \tag{2.6}$$

where $\hat{a}_\lambda$ is the minimizer of (2.1) for the given λ, and d_1 is the number of coefficients estimated as nonzero, also for the given λ. The final term comes from, with a total of p predictors, the number of different models with d_1 nonzero coefficients is exactly $\binom{p}{d_1}$ (see Chen and Chen (2008) for motivation of this term). For the initial estimator we use a similar criterion,

$$\frac{2}{n}\sum_{i=1}^{n} Q(g^{-1}(Z_i\tilde{a}_{\lambda_0}), Y_i) + d_1\frac{\log(n/K)}{n/K} + \frac{2}{n/K}\log\binom{p}{d_1}. \tag{2.7}$$

In (2.6) and (2.7), if the last term is omitted, we have the ordinary BIC.

We now show that eBIC can correctly separate the nonzero coefficients from zero ones with probability approaching one. For this we use the simplifying assumption that s is bounded. Furthermore, for technical reasons, we also assume we only search over models with at most D (fixed) nonzero coefficients. This means we should have some a priori knowledge about the complexity of the true model. In Chen and Chen (2008), where eBIC was first proposed, the same assumption was made. Note that we are not able to provide corresponding theoretical analysis on eBIC for the initial estimator (which is not consistent in variable selection anyway).

Theorem 4. *Suppose the number of nonzero coefficients s in the true model does not diverge with sample size and that we have an a priori upper bound D for s. Under the conditions of Theorems* 1 *and* 2, *$K \sim n^{1/(2d+1)}$, and that $\inf_{1\le j\le s}||\alpha_{0j}(t)||$ is bounded away from zero, the eBIC as* (2.6) *correctly identifies the nonzero coefficients and the constant coefficients with probability approaching* 1.

The minimization problem (2.1) (as well as (2.4)) is solved by local quadratic approximation, as adopted by Fan and Li (2001). Given the current estimate $a^{(0)}$, the local quadratic approximation procedure solves

$$\min \sum_i Q(g^{-1}(Z_i^T a), Y_i) + n\lambda\sum_{j=1}^{p} w_j\frac{\|a_j\|^2}{\|a_j^{(0)}\|},$$

which needs to be minimized by a Newton-Raphson iterative algorithm, resulting in an inner loop in our algorithm. During the iterations, we need to keep track of the zero coefficients and remove the corresponding predictor as soon as $\|a_j\|$ is smaller than a certain threshold (10^{-5} in our implementation). The pseudo-code of our algorithm for solving (2.1) is the following (the algorithm for solving (2.4) is similar):

Table 1. Model selection results of different penalized estimators for the Poisson model based on 100 replications, with $n = 150$. For adaptive group lasso estimator (AGL), the label (BIC-BIC), for example, means the ordinary BIC is used for both the initial estimator as well as the adaptive group lasso estimator.

		Avg # of varying coef.	
		correct	incorrect
$p = 50$	GL(BIC)	3	25.79
	GL(eBIC)	3	24.22
	AGL(BIC-BIC)	3	20.83
	AGL(eBIC-eBIC)	3	1.25
$p = 200$	GL(BIC)	3	66.28
	GL(eBIC)	3	59.15
	AGL(BIC-BIC)	2.98	43.29
	AGL(eBIC-eBIC)	2.97	3.61

Algorithm for computing (2.1)

initialize a^0.
for $k = 1, 2, \ldots$
  Starting from $a^{k,1} := a^{k-1}$, iterate until convergence to obtain a^k:
  $a^{k,l+1} = a^{k,l} - (q_2(Z_i^T a^{k,l}, Y_i) Z_i Z_i^T + 2n\lambda\Omega)^{-1}(q_1(Z_i^T a^{k,l}, Y_i) Z_i + 2n\lambda\Omega a^{k,l})$
    where $\Omega = diag(w_1 I_K/\|a_1^{k-1}\|, \ldots, w_p I_K/\|a_p^{k-1}\|)$
      is a $pK \times pK$ diagonal matrix (I_K denotes $K \times K$ identity matrix)
endfor

3. Numerical Examples

In this section we report on some simulations to evaluate the finite sample performance of the spline estimator for GVCM and demonstrate the effectiveness of eBIC for smoothing parameter selection. We also present an application to cancer classification.

Example 1. In this example, consider the varying-coefficient Poisson regression model where the true conditional mean function is

$$\mu = \exp\{X^T \alpha(T)\}.$$

The data sets were generated with sample size $n = 150$ and dimensionality $p = 50$ and $p = 200$, respectively. Due to the Newton update within the inner loop involving inversions of $p \times p$ matrices, the computation time was too long for larger p and thus we did not attempt larger dimensionality in our simulations. The index variable T was sampled uniformly on $[0, 1]$, and the predictors X_i were taken to be $X_{i1} = 1$ and X_{ij}'s marginally standard normal with within subject correlations $Cov(X_{ij_1}, X_{ij_2}) = (0.1)^{|j_1 - j_2|}, j_1, j_2 \neq 1$. We set $\alpha_1(t) = 4\sin(2\pi t)$,

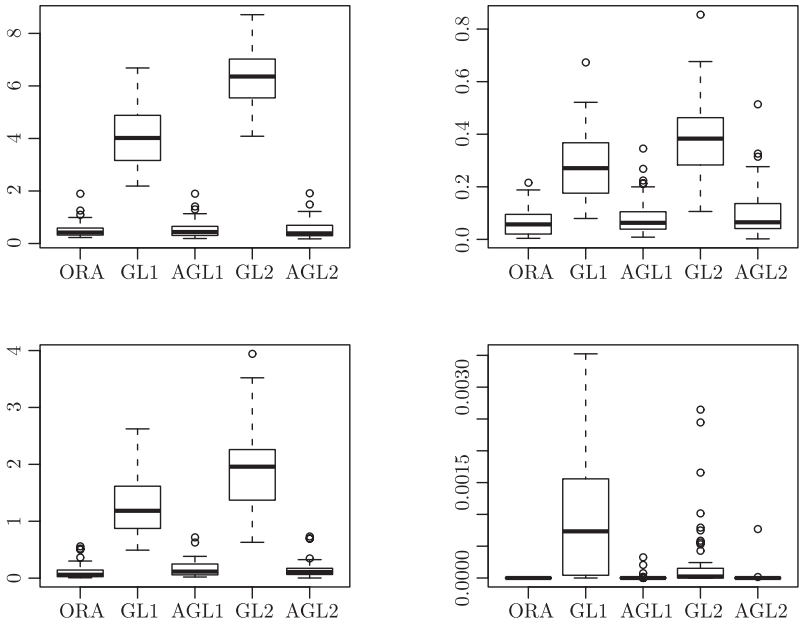

Figure 1. Boxplots showing MSEs for $\alpha_1, \ldots, \alpha_4$, for the Poisson model.

$\alpha_2(t) = 10t(1-t)$, $\alpha_3 = 3\exp\{t - 0.5\}$, and $\alpha_j = 0, j = 4, \ldots, p$. For both $p = 50$ and $p = 200$, 100 data sets were generated and fitted, with smoothing parameters λ_0, λ chosen by either ordinary BIC or eBIC for comparison. In Table 1, we show the number of identified varying coefficients. When BIC was used (for both group lasso estimator and adaptive group lasso estimator), we generally saw a large number of false positives, especially when $p = 200$. When eBIC was used, although the group lasso estimator still contained many false positives, this number was effectively controlled in the adaptive group lasso estimator. Thus we only consider the estimation accuracy when using eBIC next. In Figure 1, we show the boxplots of the mean squared errors for the coefficients $\alpha_1, \alpha_2, \alpha_3$ and $\alpha_4(= 0)$, where the calculation of MSE $\|\alpha_j(t) - \hat{\alpha}_j(t)\|$ was based on numerical approximation on a uniform grid containing 500 points on $[0, 1]$. The MSEs for five estimators are shown, including the oracle estimator (ORA, when it is known which coefficients are zeros, and 10-fold CV is used to choose K), the group lasso estimators when $p = 50$ and $p = 200$ (denoted by GL1 and GL2, respectively, in the figure), and the adaptive group lasso estimators when $p = 50$ and $p = 200$ (denoted by AGL1 and AGL2, respectively, in the figure). It is seen from the boxplots that the adaptive group lasso estimators performed much better than the group lasso estimator, and in many cases the performance was close to the oracle estimator.

Table 2. Model selection results of different penalized estimators for the logistic model, based on 100 replications, with $n = 150$.

		Avg # of varying coef.	
		correct	incorrect
$p = 50$	GL(BIC)	3	18.75
	GL(eBIC)	3	16.33
	AGL(BIC-BIC)	3	10.29
	AGL(eBIC-eBIC)	3	1.56
$p = 200$	GL(BIC)	3	38.78
	GL(eBIC)	3	32.04
	AGL(BIC-BIC)	3	25.72
	AGL(eBIC-eBIC)	2.96	2.49

Table 3. Mean squared errors for $\alpha_1, \ldots, \alpha_4$ for the data when some noise predictors are artificially added to the model. The estimates obtained from model (2.8) are taken as the truth when calculating the MSEs. GL: group lasso estimator; AGL: adaptive group lasso estimator.

		α_1	α_2	α_3	α_4
$p = 50$	GL	2.43	3.01	2.48	2.42
	AGL	0.24	0.97	1.35	1.27
$p = 200$	GL	4.03	3.80	3.79	3.37
	AGL	1.89	1.27	1.92	2.05

Example 2. Consider the varying-coefficient logistic regression model where the conditional mean function is

$$\mu = \frac{\exp\{X^T\alpha(T)\}}{(1 + \exp\{X^T\alpha(T)\})}.$$

We set $\alpha_1(t) = -4(t^3 + 2t^2 - 2t)$, $\alpha_2(t) = 4\cos(2\pi t)$, $\alpha_3 = 3\exp\{t - 0.5\}$, $\alpha_j(t) = 0, j = 4, \ldots, p$, and other aspects of the simulation set-up were the same as in Example 1. Special care was needed with binary data, since it is well-known that the algorithm does not converge to finite values when the two classes are completely separable Albert and Anderson (1984)). In our numerical studies on logistic regression models, we used Firth's bias correction to deal with this potential problem (Heinze and Schemper (2002)). The variable selection results shown in Table 2 demonstrate a similar effect as before, with eBIC effectively controlling the number of false positives, especially when $p > n$. The estimation MSE, shown in Figures 2, also demonstrated the accuracy of the adaptive group lasso estimators.

Example 3. We used the varying-coefficient logistic regression model example to study the effect of K, the same setup as in Example 2 with $p = 200$ and

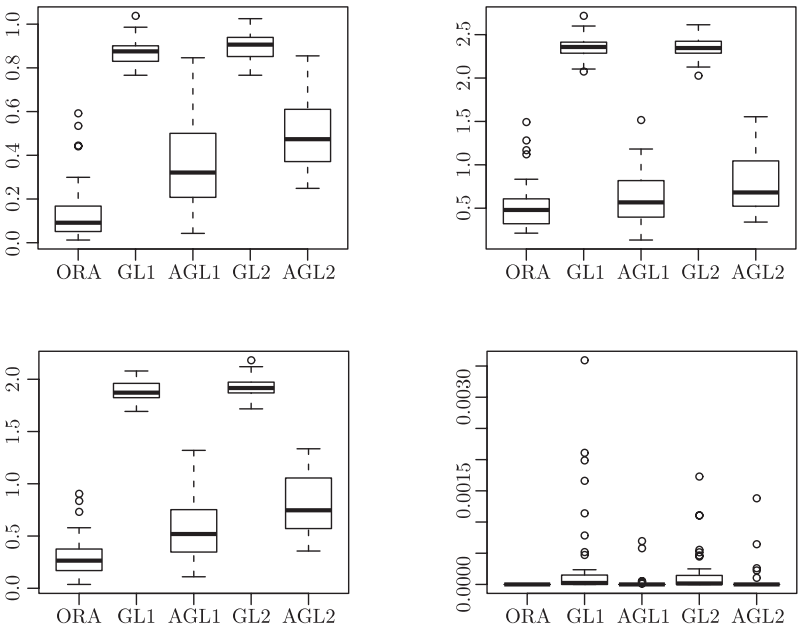

Figure 2. Boxplots showing MSEs for $\alpha_1, \ldots, \alpha_4$, for the logistic model.

$K = 5, 7, 9, 11$. We also considered using the same criteria, (2.6) and (2.7), to select λ and K simultaneously (with K ranging over $5, 7, 9, 11$). The estimation results for $\alpha_1, \ldots, \alpha_4$ are shown in Figure 3. For different K, the results look very similar, although there seems to be some weak overfitting effects when K is large. This suggests that the choice of K is not important in our model, where the overfitting can be reduced by the penalty even though K is relatively large. Also, automatically choosing K based on eBIC did not provide significant advantages over a fixed K, while it increased the computational burden. Thus we suggest fixing a relatively large K that is able to approximate the nonparametric functions reasonably well in most situations (for example, in Huang, Horowitz, and Wei (2010) the authors fixed $K = 6$).

Example 4. We applied the proposed method to a data set from the Guidelines for Urinary Incontinence Discussion and Evaluation (GUIDE) study, which assesses the impact of urinary incontinence (UI) guideline adoption by primary care providers on patient outcomes (Preisser and Qaqish (1999)). The goal is to study the factors that are predictive of the response of the 137 patients to the survey question: "Do you consider this accidental loss of urine a problem that interferes with your day to day activities or bothers you in other ways?" The binary responses are recorded as BOTHERED ($Y = 1$) if the answer is yes and $Y = 0$ if no. The predictive factors include the number of leaking accidents per day on average (DAYACC), the severity of the leaking accidents on a scale from

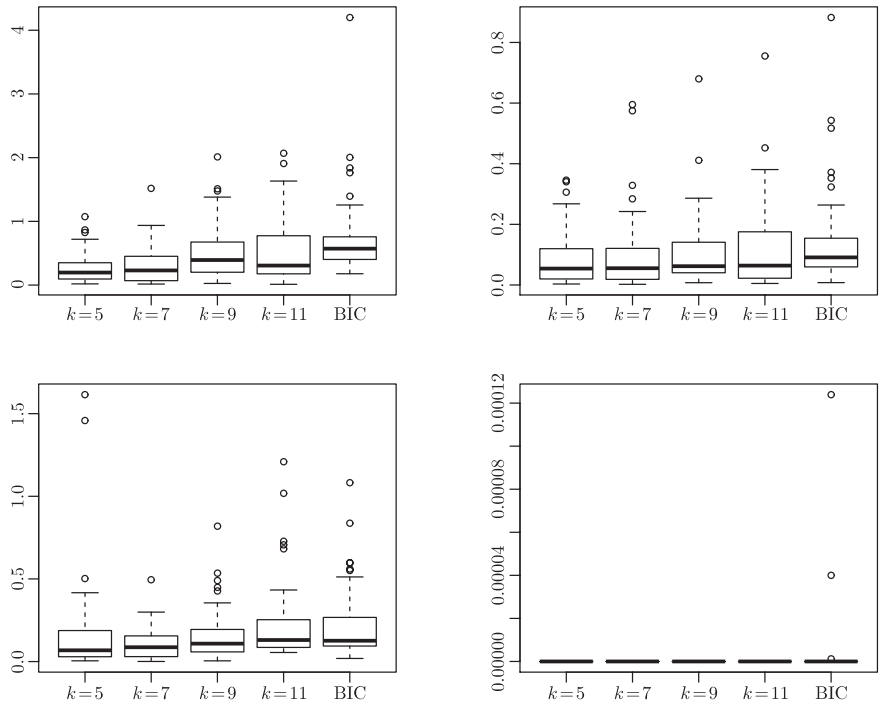

Figure 3. Boxplots showing MSEs for $\alpha_1, \ldots, \alpha_4$, for four choices of K, as well as for K chosen by BIC-type criteria.

1 to 4 with 4 the most severe (SEVERE), the number of times during the day they go to the toilet to urinate (TOILET). These three predictors are denoted by $X_2, \ldots, X_4$, respectively, while $X_1 = 1$. Using the AGE of the patient as the index variable T, and with $p = 4$, our estimation procedure produces the generalized varying-coefficient logistic regression model

$$\text{logit}(\mu) = \alpha_1(T) + X_2\alpha_2(T) + X_3\alpha_3(T) + X_4\alpha_4(T). \tag{2.8}$$

where the functions $\alpha_1, \ldots, \alpha_4$ are plotted in Figure 4. That is, all predictors are significant. It is seen that the response BOTHERED is positively correlated with DAYACC, SEVERE, and TOILET, as expected, and the correlation increases rapidly for older people (in the figure the AGE represented on the x-axis is normalized age on $[0, 1]$).

Treating the coefficients estimated as truth, we examined the effects of artificially added predictors on the estimation. The additional noise covariates were generated as in the previous examples (correlated among themselves but independent of the original covariates) and we studied the case $p = 50$ and $p = 200$ using eBIC for smoothing parameter selection. When $p = 50$, only one additional predictor was incorporated, while for $p = 200$, six additional predictors

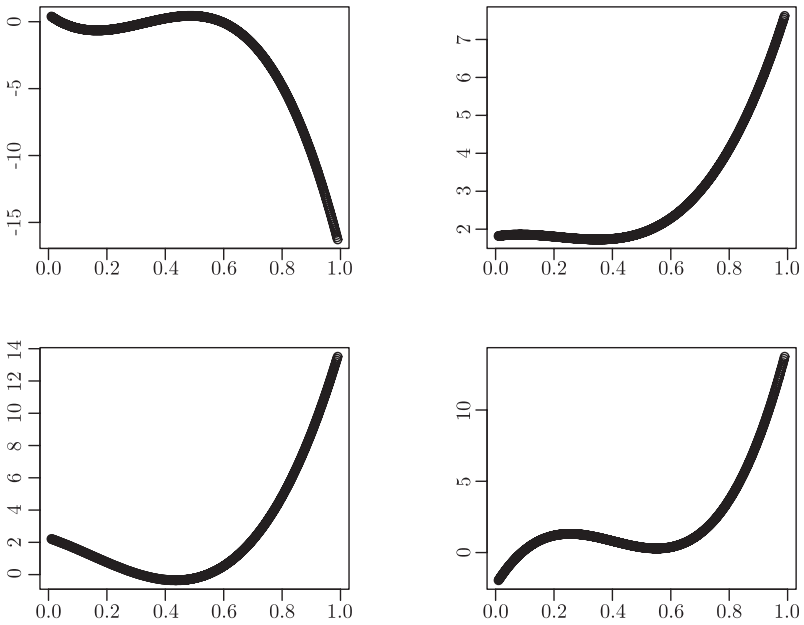

Figure 4. The estimated $\alpha_1, \ldots, \alpha_4$ for the GUIDE data set that are used as the true coefficients.

were incorporated. The MSEs for the group lasso estimator and the adaptive group lasso estimator are shown in Table 3.

Example 5. We also applied the varying-coefficient logistic regression model to cancer classification. We considered a subset of the ALL data (Chiaretti et al. (2004)) representing 79 samples from patients with B-cell acute lymphoblastic leukemia that were investigated using HG-U95Av2 Affymetrix GeneChip arrays. Of particular interest is the classification of 37 samples with the BCR/ABL fusion gene resulting from a translocation of the chromosomes 9 and 22 and 42 normal samples. Many of the genes represented by the 12,625 probesets on the array are not expressed. Thus we removed the probesets with expression measurements less than 100 fluorescence units in at least 75% of the samples, and the interquartile range (IQR) across the samples on the log base 2 scale smaller than 0.5, leaving 2401 probesets for analysis. Then we performed a t-test to rank the probesets and used the top 300 most differentially expressed ones. For classical logistic regression, we fit the data using the *glmnet* package in R and selected the tuning parameter by eBIC (we also used an adaptive lasso penalty). Besides the gene expressions, we can used the age information of the individuals. We were especially interested in a more general model where age can interact with gene expression levels, this is where our varying-coefficient logistic regression model comes in with age acting as the index variable. In the 79 samples, 3

samples have missing age information and were removed. We used leave-one-out-cross-validation to examine the classification accuracy of the two models, with 75 training samples and 1 test sample in each split. The cross-validation errors for parametric logistic regression and the more general varying-coefficient model were 6 (47 probesets selected on average) and 4 (69 probesets selected), respectively. There were on average about 35 probesets selected by both models, thus both methods identified many common probesets and our model identified more.

Acknowledgements

The author thanks the Editor, an associate editor and three anonymous referees for their insightful comments and suggestions that led to significant improvement of the manuscript. This research is supported by Singapore MOE Tier 1 Grant.

Appendix. Technical Proofs

We introduce some notation. Let $q_l(x,y) = (\partial^l/\partial x^l)Q(g^{-1}(x),y), l = 1,2,3$. We have $q_1(x,y) = -(y - g^{-1}(x))\rho_1(x)$ and $q_2(x,y) = \rho_2(x) - (y - g^{-1}(x))\rho_1'(x)$, with $\rho_l(x) = \left[dg^{-1}(x)/dx\right]^l / V(g^{-1}(x))$. For logistic and Poisson regression with canonical link function, we actually have $\rho_1(x) \equiv 1$. Denote the true varying coefficients by $\alpha_0 = (\alpha_{01}, \ldots, \alpha_{0s})^T$. Let $a_{0j} = (a_{0j1}, \ldots, a_{0jK})^T, 1 \le j \le s$, be the coefficients in spline approximation of α_{0j} that satisfies the approximation property $\| \sum_k a_{0jk} B_k(t) - \alpha_{0j}(t)\| = O(K^{-d})$ and set $a_0 = (a_{01}^T, \ldots, a_{0s}^T)^T$. With an abuse of notation, α_0 and a_0 also denote all coefficients including the zero ones. In the proofs we use a simple property of subdifferentials. For a vector b, the subdifferential of its l_2 norm is

$$\partial||b|| = \begin{cases} \frac{b}{||b||} & \text{if } b \neq 0, \\ \text{some } a \text{ with } ||a|| \le 1 & \text{if } b = 0. \end{cases}$$

Note that when $b = 0$, the subdifferential is not unique but we still use $\partial||b||$ to denote some subdifferential since its specific value plays no roles in our proofs. Finally, for any matrix A, P_A denotes the projection matrix onto the column space of A. The following regularity conditions are used in the proofs.

(c1) The covariates $X_j, 1 \le j \le p$ are bounded random variables.

(c2) The function V is twice continuously differentiable, and g is three times continuously differentiable.

(c3) The eigenvalues of $\sum_i X_i^{(1)} X_i^{(1)T}/n$ (note $X_i^{(1)} = (X_{i1}, \ldots, X_{is})^T$) are bounded away from zero and infinity.

(c4) The index variable T has a continuous density with support on $[0,1]$.

(c5) $q_2(x,y) > 0$ and $q_1(X^T\alpha_0, Y)$, $q_2(X^T\alpha_0, Y)$, and $q_3(X^T\alpha_0, Y)$ have finite second moments.

(c6) $Eq_3^2(Y, Z^T a)$ is bounded for a inside a large enough neighborhood of a_0.

(c7) For $1 \le j \le s$, $\alpha_{0j}(t)$ satisfies a Lipschitz condition of order $d > 1/2$: $|\alpha_{0j}^{(\lfloor d \rfloor)}(t) - \alpha_{0j}^{(\lfloor d \rfloor)}(s)| \le C|s-t|^{d-\lfloor d \rfloor}$, where $\lfloor d \rfloor$ is the biggest integer strictly smaller than d and $\alpha_{0j}^{(\lfloor d \rfloor)}(t)$ is the $\lfloor d \rfloor$-th derivative of $\alpha_{0j}(t)$. The order of the B-spline used satisfies $q \ge d+2$.

(c8) The eigenvalues of $\sum_i q_2(Z_i a_0, Y_i) Z_i^{(1)} Z_i^{(1)T}/(n/K)$ are bounded away from zero and infinity.

(c9) For all $1 \le k \le K$, $E[|B_k^m(T) q_1^m(X^T\alpha_0(T), Y)|] \le (m!/2) J^{m-2}/K, m = 2, 3, \ldots$, for some constant $J > 0$.

The following additional conditions are used in Theorem 3 for the initial estimator. For simplicity we assume k_1 and c_*, defined below, are bounded away from zero, while k_2 and c^* are bounded. However we keep these constants in the proofs for generality.

(c10) (Restricted eigenvalue condition) Let $S := \{1, \ldots, s\}$ be the indices of nonzero coefficients. For some $\gamma > 0$,

$$\inf_{\|v\|=1, \sum_{j\in S^c}\|v_j\| \le (1+\gamma)\sum_{j\in S}\|v_j\|} \frac{\sum_i q_2(Y_i, Z_i^T a_0) v^T Z_i Z_i^T v}{n/K} =: k_1 > 0,$$

$$\sup_{\|v\|=1, \sum_{j\in S^c}\|v_j\| \le (1+\gamma)\sum_{j\in S}\|v_j\|} \frac{\sum_i q_2(Y_i, Z_i^T a_0) v^T Z_i Z_i^T v}{n/K} =: k_2 < \infty.$$

(c11) (Zhang and Huang (2008)) The sparse Riesz condition (SRC) holds with rank s^* and spectrum bounds $0 < c_* < c^* < \infty$, that is,

$$c_*\|v\|^2 \le \frac{\sum_i q_2(Y_i, Z_i^T a_0) v^T Z_{iA} Z_{iA}^T v}{n/K} \le c^*\|v\|^2, \ \forall A \text{ with } |A| \le s^* \text{ and } v \in R^{|A|},$$

where Z_{iA} is the subvector of Z_i containing only components associated with predictors in A. Let $M = (1+2\gamma)^2 c^* k_2/k_1^2$, $s^* \ge Ms + 1$.

Most of the conditions imposed are quite standard in the literature, in particular in Lam and Fan (2008). The condition (c6) is similar to that assumed in Lam and Fan (2008) for the third derivative of the loss function and, although stronger than usually assumed in likelihood theory, it facilitates the technical derivations. For (c3), it is well-known that it implies eigenvalues of $\sum_i Z_i^{(1)} Z_i^{(1)T}/(n/K)$ are bounded away from zero and infinity (Lemma A.1 in Huang, Wu, and Zhou

(2004)). It is also possible to assume eigenvalues of $EX^{(1)}X^{(1)T}$ to be bounded away from zero and infinity, and then the eigenvalues for the sample version $\sum_i X_i^{(1)}X_i^{(1)T}/n$ will have the same property under some constraint on the size of s. However, we choose to put assumptions on the sample version so that it is directly applicable in the proofs. Since $q_1(x,y) = -(y - g^{-1}(x))\rho_1(x)$, (c9) is an assumption on the moments of the noise and, since $EB_k^m(T) = O(1/K)$, the appearance of $1/K$ in the bound is natural. The condition (c10) is a variant of ones in Bickel, Ritov, and Tsybakov (2009), and (c11) follows Zhang and Huang (2008).

Proof of Theorem 1. For the proof of Theorem 1 we write $Z_i^{(1)}$ as Z_i and $\hat{a}^{(1)}$ as $\hat{a}$. Since $\hat{a}$ minimizes

$$\sum_i Q(g^{-1}(Z_i^T a), Y_i) + n\lambda \sum_{j=1}^{s} w_j \|a_j\|$$

with respect to a, $\hat{a}$ satisfies the first-order condition

$$\sum_i q_1(Z_i^T \hat{a}, Y_i) Z_i + nv(\hat{a}) = 0, \tag{A.1}$$

where $v(\hat{a}) = (v_1(\hat{a})^T, \ldots, v_s(\hat{a})^T)^T$ is the Ks dimensional vector with $v_j(\hat{a}) = \lambda w_j \partial \|\hat{a}_j\|, 1 \le j \le s$. Obviously $\|v(\hat{a})\|^2 = O_P(\lambda^2 \|w'\|^2)$. Using a Taylor expansion at $Z_i^T a_0$ for the first term in (A.1), we get

$$\Big\| \sum_i q_1(Z_i^T a_0, Y_i) Z_i + q_2(Z_i^T a_0, Y_i) Z_i Z_i^T (\hat{a} - a_0) + \frac{1}{2} q_3(., Y_i) Z_i (Z_i^T (\hat{a} - a_0))^2 \Big\|$$
$$+ O_P(n\sqrt{\lambda^2 \|w'\|^2}) = 0,$$

or,

$$\Big| \sum_i q_1(Z_i^T a_0, Y_i) Z_i^T (\hat{a} - a_0) + q_2(Z_i^T a_0, Y_i)(\hat{a} - a_0)^T Z_i Z_i^T (\hat{a} - a_0) \tag{A.2}$$
$$+ \frac{1}{2} q_3(., Y_i)(Z_i^T (\hat{a} - a_0))^3 \Big| + O_P(n\sqrt{\lambda^2 \|w'\|^2} \|\hat{a} - a_0\|) = 0, \tag{A.3}$$

where $q_3(., Y_i)$ is evaluated at some point between $Z_i^T a_0$ and $Z_i^T \hat{a}$.

Using the notation

$$\mathbf{q_1}(Za_0, Y) = \begin{pmatrix} q_1(Z_1^T a_0, Y_1) \\ \vdots \\ q_1(Z_n^T a_0, Y_n) \end{pmatrix}$$

and $Z = (Z_1, \ldots, Z_n)^T$, the first term in (A.3) can be written as $|\mathbf{q}_1(Za_0, Y)^T Z(\hat{a} - a_0)|$, and is of order $O(\sqrt{Ks + ns^2K^{-2d}} \cdot \sqrt{n/K}\|\hat{a} - a_0\|)$ by Lemma A.1 below.

Using (c8), we have

$$c_1 \frac{n}{K}\|\hat{a} - a_0\|^2 \leq q_2(Z_i^T a_0, Y_i)(\hat{a} - a_0)^T Z_i Z_i^T (\hat{a} - a_0) \leq c_2 \frac{n}{K}\|\hat{a} - a_0\|^2 \quad \text{(A.4)}$$

for two positive constants c_1, c_2, with probability converging to 1.

Finally, the third term in (A.3) is bounded by

$$O_P(n\|Z_i\|^3\|\hat{a} - a_0\|^2) = O_P(ns^{3/2}\|\hat{a} - a_0\|^3). \quad \text{(A.5)}$$

The plan for the rest of the proof is as follows. First we consider the optimization of (2.2) inside the ball $B = \{\|a - a_0\| \leq b_n\}$ where the sequence b_n satisfies $Ks^{3/2}b_n \to 0$ and $\sqrt{K^2 s/n + s^2/K^{2d-1} + \lambda^2\|w'\|^2K^2}/b_n \to 0$ (such a sequence b_n exists due to (A)). We show that with probability converging to 1, this constrained optimization problem has a minimizer in the interior of the ball B which satisfies (2.3). By convexity this local minimizer is the global minimizer of (2.2) and the theorem is proved.

In fact, with the constraint $\|a - a_0\| \leq b_n$, it is easy to see that the third term in (A.3) is of smaller order than the second term (see (A.4) and (A.5)), and thus (A.3) implies that $\|\hat{a} - a_0\| = O_P(\sqrt{K^2 s/n + s^2/K^{2d-1} + \lambda^2\|w'\|^2K^2})$, which is of smaller order than b_n and thus $\hat{a}$ is indeed in the interior of the ball B with probability approaching 1.

Lemma A.1. $|\mathbf{q}_1(Za_0, Y)^T Z(\hat{a} - a_0)|^2 = O_P\left((n/K)(Ks + ns^2/K^{2d})\|\hat{a} - a_0\|^2\right)$.

Proof of Lemma A.1. We use $|\mathbf{q}_1(Za_0, Y)^T Z(\hat{a} - a_0)|^2 \leq \|P_Z \mathbf{q}_1(Za_0, Y)\|^2 \cdot \|Z(\hat{a} - a_0)\|^2$. Obviously $\|Z(\hat{a} - a_0)\|^2 = O_P((n/K)\|\hat{a} - a_0)\|^2)$ by (c3), and

$$\|P_Z \mathbf{q}_1(Za_0, Y)\|^2 \leq 2\|P_Z \mathbf{q}_1(X\alpha_0(T), Y)\|^2 + 2\|P_Z(\mathbf{q}_1(Za_0) - \mathbf{q}_1(X\alpha_0(T), Y))\|^2.$$

The first term here is of order $O_P(tr(P_Z)) = O_P(Ks)$, since $\mathbf{q}_1(X\alpha_0(T), Y)$ has mean zero conditional on the predictors. The second term is bounded by, using a Taylor expansion, $\|\mathbf{q}_2(X\alpha_0(T), Y)(Za_0 - X\alpha_0(T))\|^2 +$ smaller order terms $= O_P(ns^2/K^{2d})$. Note that $\mathbf{q}_2$ is an n-dimensional vector defined similar to $\mathbf{q}_1$ and that the product of two n-dimensional vectors is taken to mean the componentwise product.

Proof of Theorem 2. As before we take

$$Z_i = (X_{i1}B_1(T_i), \ldots, X_{i1}B_K(T_i), \ldots, X_{ip}B_K(T_i))^T,$$

$$Z_j = \begin{pmatrix} X_{1j}B_1(T_1) & \dots & X_{1j}B_K(T_1) \\ \vdots & \vdots & \vdots \\ X_{nj}B_1(T_n) & \dots & X_{nj}B_K(T_n) \end{pmatrix}_{n\times K}.$$

We also let $Z = (Z_1, \dots, Z_i, \dots, Z_n)^T = (Z_1, \dots, Z_j, \dots, Z_p)$, and partition it as $Z = (Z^{(1)}, Z^{(2)})$.

Since $\hat{a}^{(1)}$ solves (2.2), we have that

$$Z_j^T \mathbf{q}_1(Z^{(1)}\hat{a}^{(1)}, Y) + n\lambda w_j \partial||\hat{a}_j^{(1)}|| = 0, j = 1, \dots, s. \tag{A.6}$$

This means that "there exists some subdifferential that makes the left hand side zero" in case the subdifferential is not unique.

In order to show that the pK-dimensional vector $\hat{a} = (\hat{a}^{(1)}, \hat{a}^{(2)})$ with $\hat{a}_{jk}^{(2)} = 0, j = s+1, \dots, p, k = 1, \dots, K$, solves (2.1), we need only verify the corresponding KKT conditions,

$$Z_j^T \mathbf{q}_1(Z^{(1)}\hat{a}^{(1)} + Z^{(2)}\hat{a}^{(2)}, Y) + n\lambda w_j \partial||\hat{a}_j|| = 0, j = 1, \dots, p. \tag{A.7}$$

First, for $1 \le j \le s$, (A.7) trivially follows from (A.6), since $Z^{(2)}\hat{a}^{(2)} = 0$. Next, for $s+1 \le j \le p$, by the property of subdifferential stated at the beginning of the Appendix, (A.7) is implied by

$$(*) \quad ||Z_j^T \mathbf{q}_1(Z^{(1)}\hat{a}^{(1)} + Z^{(2)}\hat{a}^{(2)}, Y)|| \le n\lambda w_j.$$

which is shown in Lemma A.2.

Lemma A.2. $\|Z_j^T \mathbf{q}_1(Z^{(1)}\hat{a}^{(1)} + Z^{(2)}\hat{a}^{(2)}, Y)\| \le n\lambda w_j$, as used in the proof of Theorem 2.

Proof of Lemma A.2. Using a Taylor expansion, we have

$$\begin{aligned} &\max_{s+1\le j\le p} \|Z_j^T \mathbf{q}_1(Z^{(1)}\hat{a}^{(1)} + Z^{(2)}\hat{a}^{(2)}, Y)\| \\ &\quad \le \max_{s+1\le j\le p} \|Z_j^T \mathbf{q}_1(\mathbf{m}, Y)\| + \max_{s+1\le j\le p} \|Z_j^T[\mathbf{q}_2(\mathbf{m}, Y)(Z\hat{a} - \mathbf{m})]\| \\ &\qquad + \text{smaller order terms}, \end{aligned} \tag{A.8}$$

where $\mathbf{m} = (m_1, \dots, m_n)^T$ with $m_i = X_i^T \alpha_0(T_i)$.

For fixed $s+1 \le j \le p$, $1 \le k \le K$, we have

$$P(|\sum_i X_{ij}B_k(T_i)q_1(m_i, Y_i)| > c) \le 2\exp\Big\{-\frac{c^2}{(2Jc + 2n/K)}\Big\}, \forall c > 0,$$

by Bernstein's Inequality (using condition (c9) and Lemma 5.7 in van der Geer (2000)). Using a simple union bound, we have

$$P\Big(\max_{s+1\le j\le p}\|Z_j^T\mathbf{q}_1(\mathbf{m},Y)\|>c\sqrt{K}\Big)\le pK\max_{j,k}P\Big(|\sum_i X_{ij}B_k(T_i)q_1(m_i,Y_i)|>c\Big)$$
$$\le 2pK\exp\Big\{-\frac{c^2}{(2Jc+2n/K)}\Big\}.$$

Taking $c=C\sqrt{(n/K)\log(p\vee n)}$ for some $C>0$ large enough, we get

$$P(\max_{s+1\le j\le p}\|Z_j^T\mathbf{q}_1(\mathbf{m},Y)\|>c\sqrt{K})\to 0,$$

and thus

$$\max_{s+1\le j\le p}\|Z_j^T\mathbf{q}_1(\mathbf{m},Y)\|=O_P(\sqrt{n\log(p\vee n)}). \tag{A.9}$$

Furthermore, using Theorem 1, we have

$$\max_{s+1\le j\le p}\|Z_j^T\mathbf{q}_2(\mathbf{m},Y)(Z^T\hat{a}-\mathbf{m})\|=O_P\left(\sqrt{\frac{n}{K}(Ks+\frac{ns^2}{K^{2d}}+nK\lambda^2\|w'\|^2)}\right). \tag{A.10}$$

Combining (A.8), (A.9), (A.10) and (B) shows

$$\|Z_j^T\mathbf{q}_1(Z^{(1)}\hat{a}^{(1)}+Z^{(2)}\hat{a}^{(2)},Y)\|=o(n\lambda w_j).$$

Proof of Theorem 3. The proof uses similar techniques as in Bickel, Ritov, and Tsybakov (2009); Zhang and Huang (2008) only dealt with quadratic loss.

Step 1. Let $\delta=\tilde{a}-a_0$, where $a_0=(a_{01},\dots,a_{0p})$ is the coefficient in the spline basis approximation of $\alpha_0=(\alpha_1,\dots,\alpha_p)$ that satisfies $\|\sum_k a_{0jk}B_k(t)-\alpha_{0j}(t)\|=O(K^{-d})$ (for $j>s$ this approximation error is actually 0). We show that

$$\sum_{j\in S^c}\|\delta_j\|\le(1+\gamma)\sum_{j\in S}\|\delta_j\|, \tag{A.11}$$

with probability converging to 1, for any $\gamma>0$, where $S=\{1,\dots,s\}$.

By the definition of $\tilde{a}$, we have

$$\sum_i Q(Z_i^T\tilde{a},Y_i)-\sum_i Q(Z_i^Ta_0,Y_i)\le\lambda_0\sum_{j=1}^p\|a_{0j}\|-\lambda_0\sum_{j=1}^p\|\tilde{a}_j\|. \tag{A.12}$$

On the other hand, by the convexity of Q,

$$\sum_i Q(Z_i^T\tilde{a},Y_i)-\sum_i Q(Z_i^Ta_0,Y_i)\ge\sum_i q_1(Z_i^Ta_0,Y_i)Z_i^T\delta.$$

Combining, we get

$$-(\sum_{j=1}^{p}\|\delta_j\|)\cdot\max_{1\le j\le p}\|\sum_i q_1(Z_i^Ta_0,Y_i)Z_{ij}\|\le\lambda_0\sum_{j=1}^{p}\|a_{0j}\|-\lambda_0\sum_{j=1}^{p}\|\tilde{a}_j\|, \quad (A.13)$$

where $Z_{ij}=(X_{ij}B_1(T_i),\ldots,X_{ij}B_K(T_i))^T$ is a subvector of Z_i.

Using the arguments of Lemma A.2, we have

$$\max_{1\le j\le p}\|\sum_i q_1(Z_i^Ta_0,Y_i)Z_{ij}\|=O_p(\sqrt{n\log(p\vee n)})+O_P(\sqrt{\frac{n}{K}\frac{ns^2}{K^{2d}}})=o_P(\lambda_0). \quad (A.14)$$

Using (A.13), (A.14), together with

$$\sum_{j=1}^{p}\|\delta_j\|=\sum_{j\in S}\|\delta_j\|+\sum_{j\in S^c}\|\delta_j\|, \quad (A.15)$$

$$\begin{aligned}\lambda_0(\sum_{j=1}^{p}\|a_{0j}\|-\sum_{j=1}^{p}\|\tilde{a}_j\|)&=\lambda_0(\sum_{j\in S}\|a_{0j}\|-\sum_{j\in S}\|\tilde{a}_j\|-\sum_{j\in S^c}\|\tilde{a}_j\|)\\&\le\lambda_0(\sum_{j\in S}\|\delta_j\|-\sum_{j\in S^c}\|\delta_j\|), \quad (A.16)\end{aligned}$$

we obtain

$$\sum_{j\in S^c}\|\delta_j\|\le(1+o_P(1))\sum_{j\in S}\|\delta_j\|\le(1+\gamma)\sum_{j\in S}\|\delta_j\|.$$

Finally, we note that this step is an extension of the similar ones obtained for the Lasso estimate with quadratic loss function. It was shown in Bickel, Ritov, and Tsybakov (2009) that (A.11) plays a critical role in showing the convergence of the estimate.

Step 2. We have $\|\tilde{a}-a_0\|\le(1+\gamma)(\sqrt{s}K\lambda_0/nk_1)$, with probability converging to 1.

We show that with probability converging to 1, there exists a local minimizer a^* of (2.4) in the interior of the ball $\{a:\|a-a_0\|\le b_n\}$, where $\sqrt{s}K\lambda_0/(b_nnk_1)\to 0$ and $(b_nK\sqrt{s^3})/k_1\to 0$ (such a sequence b_n exists since we assume $s^2K^2\lambda_0/(nk_1^2)\to 0$), and this local minimizer satisfies $\|a^*-a_0\|\le(1+\gamma)(\sqrt{s}K\lambda_0/nk_1)$. Then by the convexity of the problem, this local minimizer is the global minimizer $\tilde{a}$.

Let $\delta=a^*-a_0$. Combining (A.12) and (A.16), we get

$$\sum_i Q(Z_i^Ta^*,Y_i)-\sum_i Q(Z_i^Ta_0,Y_i)\le\lambda_0(\sum_{j\in S}\|\delta_j\|-\sum_{j\in S^c}\|\delta_j\|).$$

Using Taylor's expansion for the left hand side here results in

$$\sum_i q_1(Z_i^T a_0, Y_i) Z_i^T \delta + \frac{1}{2} q_2(Z_i^T a_0, Y_i) \delta^T Z_i Z_i^T \delta + \frac{1}{6} q_3(z_i^*, Y_i)(Z_i^T \delta)^3 \\ \leq \lambda_0 (\sum_{j \in S} \|\delta_j\| - \sum_{j \in S^c} \|\delta_j\|), \tag{A.17}$$

where z_i^* lies between $Z_i^T \tilde{a}$ and $Z_i^T a_0$.

As shown in Step 1, $\| \sum_i q_1(Z_i^T a_0, Y_i) Z_i^T \delta \| = o_P(\lambda_0) \sum_{j=1}^p \|\delta_j\|$. Condition (c10) implies $\sum_i q_2(Z_i^T a_0, Y_i) \delta^T Z_i Z_i^T \delta \geq n k_1 \|\delta\|^2 / K$ and

$$\sum_i q_3(z_i^*, Y_i)(Z_i^T \delta)^3 = O_P(nE[q_3(z_1^*, Y_1)(\max_{1 \leq j \leq p} \|Z_{1j}\|)^3 (\sum_{j=1}^p \|\delta_j\|)^3]) \\ = O_P(n(\sum_{j=1}^p \|\delta_j\|)^3).$$

Since $\sum_{j=1}^p \|\delta_j\| \leq (2+\gamma) \sum_{j \in S} \|\delta_j\| \leq (2+\gamma)\sqrt{s} (\sum_{j \in S} \|\delta_j\|^2)^{1/2} \leq (2+\gamma)\sqrt{s}\|\delta\|$ by Step 1, and using the assumption on b_n, it is easily seen that $n(\sum_{j=1}^p \|\delta_j\|)^3 = o_P(n k_1 \|\delta\|^2 / K)$, and thus (A.17) becomes

$$\|\delta\|^2 \leq (1 + o_P(1)) \frac{\lambda_0 K}{n k_1} \sum_{j \in S} \|\delta_j\|.$$

Using again $\sum_{j \in S} \|\delta_j\| \leq \sqrt{s}\|\delta\|$, we get $\|\delta\| \leq (1+\gamma)\sqrt{s}\lambda_0 K/(n k_1)$. Finally, since $\sqrt{s}\lambda_0 K/(n k_1) = o(b_n)$, a^* is indeed an interior point of the ball with probability converging to 1.

Step 3. Under (c11), we have $\hat{s} \leq Ms$ with probability approaching 1, where $\hat{s}$ is the number of estimated nonzero coefficients. In particular, if c^* and k_2 are bounded, and k_1 is bounded away from zero, then $\hat{s}$ is of the same order as s.

Define more generally $c^*(m) = \sup_{|A|=m, \|v\|=1} \sum_i q_2(Z_i^T a_0, Y_i) v^T Z_{iA} Z_{iA}^T v / n$, and define $c_*(m)$ similarly. Let $A_1 = \{j : \tilde{a}_j \neq 0\}$ be the set of indices of estimated nonzero coefficients, and thus $\hat{s} = |A_1|$. By the first order condition of the minimization problem (2.4), we know that for $j \in A_1$,

$$\sum_i q_1(Z_i \tilde{a}, Y_i) Z_{ij} = -\lambda_0 \partial \|\tilde{a}_j\|. \tag{A.18}$$

Let A_2 be the set of indices j that satisfies (A.18) and thus $A_1 \subseteq A_2$. Suppose $\tilde{A}$ satisfies $A_1 \subseteq \tilde{A} \subseteq A_2$ (the precise choice of $\tilde{A}$ is only important toward the end of the proof). Let $\tilde{s} = |\tilde{A}| \geq \hat{s}$.

Using (A.18) for $j \in \tilde{A}$ (taking sum of squares) and setting $\delta = \tilde{a} - a_0$, we have

$$\begin{aligned}\sqrt{\tilde{s}}\lambda_0 &= \Big\| \sum_i q_1(Z_i\tilde{a}, Y_i) Z_{i\tilde{A}} \Big\| \\ &= \Big\| \sum_i q_1(Z_i^T a_0, Y_i) Z_{i\tilde{A}} + q_2(Z_i a_0, Y_i) Z_{i\tilde{A}} Z_i^T \delta + \frac{1}{2} q_3(z_i^*, Y_i) Z_{i\tilde{A}} (Z_i^T \delta)^2 \Big\|.\end{aligned}$$

The first term satisfies $\| \sum_i q_1(Z_i^T a_0, Y_i) Z_{i\tilde{A}} \| \le \sqrt{\tilde{s}} \max_j \| \sum_i q_1(Z_i^T a_0, Y_i) Z_{ij} \| = o_P(\sqrt{\tilde{s}}\lambda_0)$ as in Step 1. The second term can be bounded by $\| \sum_i q_2(Z_i a_0, Y_i) Z_{i\tilde{A}} Z_i^T \delta \| \le n\sqrt{c^*(\tilde{s})k_2}\|\delta\|/K$ by the Cauchy-Schwartz Inequality. Similar to the proof in Step 2, the third term is of smaller order than the second term if

$$\frac{s^2 K^2 \lambda_0}{n k_1 \sqrt{c_*(\tilde{s})k_2)}} \to 0. \tag{A.19}$$

Then we have, with probability converging to 1,

$$\sqrt{\tilde{s}}\lambda_0 \le (1 + o_P(1)) \frac{n}{K} \sqrt{c^*(\tilde{s})k_2}\|\delta\| \le (1 + 2\gamma) n \sqrt{s c^*(\tilde{s}) k_2} \frac{\lambda_0}{n k_1},$$

which is equivalent to

$$\tilde{s} \le \frac{(1 + 2\gamma)^2 c^*(\tilde{s}) k_2}{k_1^2} s. \tag{A.20}$$

By the continuity of $\tilde{a}$ in λ_0, we can choose $\tilde{A}$ such that its size jumps at most one each time that λ_0 decreases, beginning from $\lambda_0 = \infty$ to the lower bound. We show $\tilde{s} \le Ms$ by contradiction. In fact, suppose for some λ_0 we have $|\tilde{A}| > Ms$, then we are able to find λ_0 such that in fact $Ms < |\tilde{A}| \le Ms + 1$ since we change the size of $\tilde{A}$ one at a time. For this λ_0, since $|\tilde{A}| \le s^*$, we have $c^*(\tilde{s}) \le c^*$ and $c_*(\tilde{s}) \ge c_*$, so that (A.19) is satisfied. Thus from (A.20), $\tilde{s} \le Ms$, leading to a contradiction.

Proof of Theorem 4. For any given regularization parameter λ, we denote by $\hat{a}_\lambda$ the minimizer of (2.1), and by $\hat{a}$ the minimizer when the optimal sequence of regularization parameter is chosen such that $\hat{a}$ results in a consistent model selection. We separately consider the overfitting and underfitting cases below.

Underfitting. Assume some nonzero coefficients are estimated as zero coefficients in $\hat{a}_\lambda$ and look to establish some contradiction. Similar to the proof of

Theorem 1, we have

$$\begin{aligned}
&\frac{1}{n}\sum_i Q(g^{-1}(Z_i^T\hat{a}_\lambda),Y_i)-\frac{1}{n}\sum_i Q(g^{-1}(Z_i^T\hat{a}),Y_i)\\
&\quad=\frac{1}{n}\sum_i q_1(Z_i^T\hat{a},Y_i)Z_i^T(\hat{a}_\lambda-\hat{a})+\frac{1}{2n}\sum_i q_2(Z_i^T\hat{a},Y_i)(\hat{a}_\lambda-\hat{a})Z_iZ_i^T(\hat{a}_\lambda-\hat{a})\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad+\text{smaller order terms}\\
&\quad\geq-\frac{C_1}{n}||P_Z\mathbf{q}_1(Z\hat{a},Y)||^2+\frac{C_2}{n}||Z(\hat{a}-\hat{a}_\lambda)||^2,
\end{aligned}$$

for some constants $C_1, C_2 > 0$, where we used the Cauchy-Schwartz Inequality

$$\Big|\sum_i q_1(Z_i\hat{a},Y_i)Z_i^T(\hat{a}_\lambda-\hat{a})\Big|\leq\frac{1}{C}||P_Z\mathbf{q}_1(Z\hat{a},Y)||^2+\frac{C}{4}\|Z(\hat{a}-\hat{a}_\lambda)\|^2$$

(with a small enough constant $C > 0$), as well as (c8).

Since there is some j for which $\hat{a}_j$ represents a truly varying coefficient with convergence rate given by Theorem 1, while $\hat{a}_{\lambda j} = 0$, it is easy to show that $||Z(\hat{a}-\hat{a}_\lambda)||^2/n$ is bounded away from zero. Besides, $||P_Z\mathbf{q}_1(Z\hat{a},Y)||/n = o(1)$ (using the same arguments as in Lemma A.1, as well as the proof of convergence rate in Theorem 1) and the penalty terms in eBIC are all of order $o(1)$, thus the eBIC when λ is used is bigger than the eBIC when the optimal regularization sequence is used, leading to a contradiction.

Overfitting. Here we assume some zero coefficients are estimated as nonzero in $\hat{a}_\lambda$. Let $\hat{a}^*$ be the minimizer of $\sum_i Q(g^{-1}(Z_i^Ta),Y_i)$ under the additional constraint that the model identified by $\hat{a}_\lambda$ is used when minimizing the negative likelihood (without penalty). We have that

$$\begin{aligned}
&\frac{1}{n}\sum_i Q(g^{-1}(Z_i^T\hat{a}_\lambda),Y_i)-\frac{1}{n}\sum_i Q(g^{-1}(Z_i^T\hat{a}),Y_i)\\
&\quad\geq\frac{1}{n}\sum_i Q(g^{-1}(Z_i^T\hat{a}^*),Y_i)-\frac{1}{n}\sum_i Q(g^{-1}(Z_i^T\hat{a}),Y_i)\\
&\quad\geq\frac{1}{n}\sum_i q_1(Z_i^T\hat{a},Y_i)Z_i^T(\hat{a}^*-\hat{a}),
\end{aligned}\tag{A.21}$$

by the convexity of Q. Using the definition of $\hat{a}^*$ and the fact that we only search over models with size at most D, the convergence rate of $\hat{a}^*$ can be obtained using similar arguments as Theorem 1 (although the third term in (2.3) involving λ does not appear for the unpenalized estimator). Arguments similar to those used in the proof of Lemma A.2 can be used to show that (A.21) is bounded below by a negative term whose absolute value is of order

$$\frac{1}{n}\sqrt{n\log(p\vee n)(\frac{K^2}{n}+\frac{1}{K^{2d-1}})},$$

which is of order smaller than the BIC penalty term $\log(n/K)/(n/K) + \log p/(n/K)$. Thus BIC cannot have selected such a λ.

References

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1-10.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.* **37**, 1705-1732.

Cai, Z., Fan, J. Q. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95**, 941-956.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759-771.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J. and Foa, R. (2004). Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood* **103**, 2771-2778.

Choi, N.-H., Li, W. and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *J. Amer. Statist. Assoc.* **105**, 354-364.

De Boor, C. (2001). *A Practical Guide to Splines*. Springer-Verlag, New York.

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.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55**, 757-796.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statist. Medicine* **21**, 2409-2419.

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, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38**, 2282-2313.

Huang, J. H. 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.

Lam, C. and Fan, J. Q. (2008). Profile-kernel likelihood inference with diverging number of parameters. *Ann. Statist.* **36**, 2232-2260.

Li, G., Peng, H. and Zhu, L. X., (2011). Nonconcave penalized M-estimation with diverging number of parameters. *Statist. Sinica*, **21**, 391-419.

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.* **36**, 261-286.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd edition. Chapman and Hall, London.

Meier, L., Van de Geer, S. and Bühlmann, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37**, 3779-3821.

Preisser, J. S. and Qaqish, B. F. (1999). Robust regression for clustered data with application to binary responses. *Biometrics* **55**, 574-579.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd edition. Springer, New York.

Ravikumar, P., Liu, H., Lafferty, J. and Wasserman, L. (2008). Spam: Sparse additive models. In *Advances in Neural Information Processing Systems* **20** (Edied by J. Platt, D. Koller, Y. Singer and S. Roweis), 1201-1208, MIT Press, Cambridge, MA.

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

van der Geer, S. A. (2000). *Applications of Empirical Process Theory*. Cambridge University Press, Cambridge.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Comput. Statist. Data Anal.* **52**, 5277-5286.

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., 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.

Wei, F., Huang, J. and Li, H. Z. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statist. Sinica*, to appear.

Xie, H. L. and Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models. *Ann. Statist.* **37**, 673-696.

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

Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. *J. Roy. Statist. Soc. Ser. B* **69**, 143-161.

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

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 December 2010; accepted May 2011)