Statistica Sinica **21** (2011), 1515-1540
doi:http://dx.doi.org/10.5705/ss.2009.316

VARIABLE SELECTION AND ESTIMATION IN HIGH-DIMENSIONAL VARYING-COEFFICIENT MODELS

Fengrong Wei, Jian Huang and Hongzhe Li

University of West Georgia, University of Iowa and University of Pennsylvania

Abstract: Nonparametric varying coefficient models are useful for studying the time-dependent effects of variables. Many procedures have been developed for estimation and variable selection in such models. However, existing work has focused on the case when the number of variables is fixed or smaller than the sample size. In this paper, we consider the problem of variable selection and estimation in varying coefficient models in sparse, high-dimensional settings when the number of variables can be larger than the sample size. We apply the group Lasso and basis function expansion to simultaneously select the important variables and estimate the nonzero varying coefficient functions. Under appropriate conditions, we show that the group Lasso selects a model of the right order of dimensionality, selects all variables with the norms of the corresponding coefficient functions greater than certain threshold level, and is estimation consistent. However, the group Lasso is in general not selection consistent and tends to select variables that are not important in the model. In order to improve the selection results, we apply the adaptive group Lasso. We show that, under suitable conditions, the adaptive group Lasso has the oracle selection property in the sense that it correctly selects important variables with probability converging to one. In contrast, the group Lasso does not possess such oracle property. Both approaches are evaluated using simulation and demonstrated on a data example.

Key words and phrases: Basis expansion, group Lasso, high-dimensional data, nonparametric coefficient function, selection consistency, sparsity

1. Introduction

Consider a linear varying coefficient model with p_n variables

$$y_i(t_{ij}) = \sum_{k=1}^{p} x_{ik}(t_{ij})\beta_k(t_{ij}) + \epsilon_i(t_{ij}),\ i = 1, \ldots, n, j = 1, \ldots, n_i, \tag{1.1}$$

where $y_i(t)$ is the response variable for the ith subject at time point $t \in T$, T is the time interval on which the measurements are taken, $\epsilon_i(t)$ is the error term, $x_{ik}(t)$ is the covariate variable with time-varying effects, $\beta_k(t)$ is the corresponding smooth coefficient function. Such a model is useful in investigating the

time-dependent effects of covariates on responses measured repeatedly. One well known example is longitudinal data analysis (Hoover et al. (1998)) where the response for the ith experimental subject in the study is observed on n_i occasions, and the observations at times $t_{ij} : j = 1, \ldots, n_i$ are correlated. Another important example is the functional response model (Rice (2004)), where the response $y_i(t)$ is a smooth real function, although only $y_i(t_{ij})$, $j = 1, \ldots, n_i$ are observed in practice. In both examples, the response $y_i(t)$ is a random process and the covariate $x_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))'$ is a p-dimensional vector of random processes. In this paper, we investigate the selection of the important covariates and the estimation of their relative coefficient functions in high-dimensional settings, in the particular case $p \gg n$, under the assumption that the number of important covariates is "small" relative to the sample size. We propose penalized methods for variable selection and estimation in (1.1) based on basis expansion of the coefficient functions, and show that under appropriate conditions, the proposed methods can select the important variables with high probability and estimate the coefficient functions efficiently.

Many methods have been developed for variable selection and estimation in varying coefficient models (1.1). See, for example, Fan and Zhang (2000) and Wu and Chiang (2000), for the local polynomial smoothing method; Wang and Xia (2008) for the local polynomial method with Lasso penalty; Huang, Wu, and Zhou (2004) and Qu and Li (2006) for basis expansion and the spline method; Chiang, Rice, and Wu (2001) for the smoothing spline method; Wang, Li, and Huang (2008) for basis function approximation with SCAD penalty (Fan and Li (2001); Fan and Lv (2010)). In addition to these methods, much progress has been made in understanding such properties of the resulting estimators as selection consistency, convergence, and asymptotic distribution. However, in all these studies, the number of variables p is fixed or less than the sample size n. To the best of our knowledge, there has been no work on the problem of variable selection and estimation in varying coefficient models in sparse, $p \gg n$ situations.

There has been much work on the selection and estimation of groups of variables. For example, Yuan and Lin (2006) proposed the group Lasso, group Lars, and group nonnegative garrote methods. Kim, Kim, and Kim (2006) considered the group Lasso in the context of generalized linear models. Zhao, Rocha, and Yu (2008) proposed a composite absolute penalty for group selection that can be considered a generalization of the group Lasso. Huang et al. (2007) considered the group bridge approach which can be used for simultaneous group and within group variable selection. However, there has been no investigation of these methods in the context of high-dimensional varying coefficient models.

In this paper, we apply the group Lasso and basis expansion to simultaneously select the important variables and estimate the coefficient functions in

(1.1). With basis expansion, each coefficient function is approximated by a linear combination of a set of basis functions. Thus the selection of important variables and estimation of the corresponding coefficient functions amounts to the selection and estimation of groups of coefficients in the linear expansions. It is natural to apply the group Lasso, since it takes into account the group structure in the approximation model. We show that, under appropriate conditions, the group Lasso selects a model of the right order of dimensionality, selects all variables with coefficient function ℓ_2 norms greater than a certain threshold level, and is estimation consistent. In order to achieve selection consistency, we apply the adaptive group Lasso. We show that the adaptive group Lasso can correctly select important variables with probability converging to one based on an initial consistent estimator. In particular, we use the group Lasso to obtain the initial estimator for the adaptive group Lasso. This approach follows the idea of the adaptive Lasso (Zou (2006)). An important aspect of our results is that p can be much larger than n.

The rest of the paper is organized as follows. In Section 2, we describe the procedure for selection and estimation using the group Lasso and the adaptive group Lasso with basis expansion. In Section 3, we state the results on estimation consistency of the group Lasso and the selection consistency of the adaptive group Lasso in high-dimensional settings. Proofs are given in Section 6. In Section 4, simulations and data examples are used to illustrate the proposed methods. Summary and discussion are given in Section 5.

2. Basis Expansion and Penalized Estimation

Suppose that the coefficient function β_k can be approximated by a linear combination of basis functions,

$$g_k(t) = \sum_{l=1}^{d_k} \gamma_{kl} B_{kl}(t),\ t \in T,\ k = 1, \ldots, p, \tag{2.1}$$

where $B_{kl}(t)$, $t \in T$, $l = 1, \ldots, d_k$, are basis functions and d_k is the number of basis functions, which is allowed to increase with the sample size n.

Let G_k denote all functions that have the form $\sum_{l=1}^{d_k} \gamma_{kl} B_{kl}(t)$ for a given basis system $\{B_{kl}, l = 1, \ldots, d_k\}$. For $g_k \in G_k$, define the approximation error by

$$\rho_k(t) = \beta_k(t) - g_k(t) = \beta_k(t) - \sum_{l=1}^{d_k} \gamma_{kl} B_{kl}(t),\ t \in T,\ k = 1, \ldots, p.$$

Let $\mathrm{dist}(\beta_k, G_k) = \inf_{g_k \in G_k} \sup_{t \in T} |\rho_k(t)|$ be the L_∞ distance between β_k and G_k, and take $\rho = \max_{1 \le k \le p} \mathrm{dist}(\beta_k, G_k)$.

By the definition of ρ_k and (2.1), model (1.1) can be written as

$$y_i(t_{ij}) = \sum_{k=1}^{p}\sum_{l=1}^{d_k} x_{ik}(t_{ij})\gamma_{kl}B_{kl}(t_{ij}) + \sum_{k=1}^{p} x_{ik}(t_{ij})\rho_k(t_{ij}) + \epsilon_i(t_{ij}), \tag{2.2}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. In low-dimensional settings, we can minimize the least squares criterion

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{y_i(t_{ij}) - \sum_{k=1}^{p}\sum_{l=1}^{d_k}\gamma_{kl}x_{ik}(t_{ij})B_{kl}(t_{ij})\}^2 \tag{2.3}$$

with respect to γ_{kl}'s. The least squares estimator of β_k is $\hat{\beta}_k(t) = \sum_{l=1}^{d_k}\hat{\gamma}_{kl}B_{kl}(t), t \in T$, where $\hat{\gamma}_{kl}$'s are the minimizer of (2.3).

When the number of variables p or $\sum_{k=1}^{p} d_k$ is larger than the sample size n, however the least squares method is not applicable since there is no unique solution to (2.3). In such case, regularized methods are needed. We apply the group Lasso (Yuan and Lin (2006)),

$$\arg\min_{\gamma}\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_i(t_{ij}) - \sum_{k=1}^{p}\sum_{l=1}^{d_k}x_{ik}(t_{ij})\gamma_{kl}B_{kl}(t_{ij}))^2 + \sum_{k=1}^{p}\lambda\|\gamma_k\|_k, \tag{2.4}$$

where λ is the penalty parameter, $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kd_k})'$ is a d_k-dimensional coefficient vector corresponding to the kth variable, and $\|\gamma_k\|_k^2 = \gamma_k' R_k \gamma_k$. Here $R_k = (r_{ij})_{d_k \times d_k}$ is the kernel matrix whose (i,j)th element is

$$r_{ij} = \int_T B_{ki}(t)B_{kj}(t)dt, \quad \text{for } i = 1, \ldots, d_k, \quad j = 1, \ldots, d_k, \quad k = 1, \ldots, p, \tag{2.5}$$

it is a symmetric positive definite matrix by Lemma A.1 in Huang, Wu, and Zhou (2004).

To express the criterion function (2.4), let

$$Y = (y_1(t_{11}), \ldots, y_1(t_{1n_1}), \ldots, y_n(t_{n1}), \ldots, y_n(t_{nn_n}))', \; X = (X_1, \ldots, X_p)$$

with $X_k = (x_{1k}(t_{11}), \ldots, x_{1k}(t_{1n_1}), \ldots, x_{nk}(t_{n1}), \ldots, x_{nk}(t_{nn_n}))'$ and define

$$B(t) = \begin{pmatrix} B_{11}(t) & B_{12}(t) & \ldots & B_{1d_1}(t) & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 & B_{p1}(t) & B_{p2}(t) & \ldots & B_{pd_p}(t) \end{pmatrix},$$

and $\gamma = (\gamma_1', \ldots, \gamma_p')'$. Set $U = (U_{11}, \ldots, U_{1n_1}, \ldots, U_{n1}, \ldots, U_{nn_n})'$ with $U_{ij}' = x_i(t_{ij})'B(t_{ij})$ for $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. Then the group Lasso penalized

criterion (2.4) can be rewritten as

$$\hat{\gamma} = \arg\min_{\gamma} \frac{1}{2}(Y - U\gamma)'(Y - U\gamma) + \sum_{k=1}^{p} \lambda \|\gamma_k\|_k. \tag{2.6}$$

The group Lasso estimator is $\hat{\beta}_k(t) = \sum_{l=1}^{d_k} \hat{\gamma}_{kl} B_{kl}(t)$, $k = 1, \ldots, p$.

Let $\omega = (\omega_1, \ldots, \omega_p)'$ be a given vector of weights, where $0 \le \omega_k \le \infty, 1 \le k \le p$. Then a weighted group Lasso criterion is

$$\hat{\gamma^*} = \arg\min_{\gamma} \frac{1}{2}(Y - U\gamma)'(Y - U\gamma) + \sum_{k=1}^{p} \tilde{\lambda}\omega_k \|\gamma_k\|_k. \tag{2.7}$$

The weighted group Lasso estimator $\hat{\beta}_k^*(t) = \sum_{l=1}^{d_k} \hat{\gamma}_{kl}^* B_{kl}(t)$, $k = 1, \ldots, p$, where $\hat{\gamma}^*$ is the minimizer of (2.7). When the weights are dependent on the data through an initial estimator, such as $\hat{\beta}$, then we call the resulting $\hat{\beta}_k^*$ an adaptive group Lasso estimator.

3. Theoretical Results

In this section, we describe the asymptotic properties of the group Lasso and the adaptive group Lasso estimators defined in (2.6) and (2.7) of Section 2 when p can be larger than n, but the number of important covariates is relatively small.

In (1.1), without loss of generality, suppose that the first q_n variables are important. Let $A_0 = \{q_n+1, \ldots, p_n\}$. Here we write q_n, p_n to indicate that q and p are allowed to diverge with n. Thus all the variables in A_0 are not important. Let $|A|$ denote the cardinality of any set $A \subset \{1, \ldots, p_n\}$ and $d_A = \sum_{k \in A} d_k$. For any set $A \subset \{1, \ldots, p_n\}$, define

$$U_A = (u_{kj} : j = 1, \ldots, d_k;\ k \in A) \text{ and } \Sigma_{AA} = U_A' \frac{U_A}{n}.$$

Here U_A is a $n \times d_A$ dimensional submatrix of the 'designed' matrix U. Take $\|\beta_k\|_2 = \left[\int_T \beta_k^2(t)\right]^{1/2} dt$ whenever the integral exists.

We rely on the following conditions.

(C1) There exist constants $q^* > 0$, $c_* > 0$ and $c^* > 0$ where $0 < c_* \le c^* < \infty$ such that

$$c_* \le \frac{\|U_A \nu\|_2^2}{n\|\nu\|_2^2} \le c^*,\ \forall A \text{ with } |A| = q^* \text{ and } \nu \in \mathbb{R}^{d_A}.$$

(C2) There is a small constant $\eta_1 \ge 0$ such that $\sum_{k \in A_0} \|\beta_k\|_2 \le \eta_1$.

(C3) The random errors $\epsilon_i(t)$, $i = 1, \ldots, n$ are independent and identically distributed as $\epsilon(t)$, where $E[\epsilon(t)] = 0$ and $E[\epsilon(t)^2] \leq \sigma^2 < \infty$ for $t \in T$; moreover, the tail probabilities satisfy $P(|\epsilon(t)| > x) \leq K \exp(-Cx^2)$ for $t \in T$, $x > 0$, and some constants C and K.

(C4) There exists a positive constant M such that $|x_{ik}(t)| \leq M$ for all $t \in T$ and $i = 1, \ldots, n$, $k = 1, \ldots, p_n$.

Condition (C1) is the sparse Riesz condition for varying coefficient models, which controls the range of eigenvalues of the matrix U. This condition was formulated for the linear regression model in Zhang and Huang (2008). If the covariate matrix X satisfies (C1), then the matrix U also satisfies (C1) and $c^*, c_* \sim O(d_a^{-1})$. See Lemma A.1 in Huang, Wu, and Zhou (2004). Condition (C2) assumes that the varying coefficients of the unimportant variables are small in the ℓ_2 sense, but do not need to be exactly zero. If $\eta_1 = 0$, (C2) becomes $\beta_k(t) \equiv 0$ for all $k \in A_0$. This can be called the narrow-sense sparsity condition (NSC) (Wei and Huang (2008)). Under the NSC, the problem of variable selection is equivalent to distinguishing nonzero coefficient functions from zero coefficient functions. Under (C2), it is no longer sensible to select the set of all nonzero coefficient functions, the goal is to select the set of important variables with large coefficient functions. From the standpoint of statistical modeling and interpretation, (C2) is mathematically weaker and more realistic. Condition (C3) assumes that the error term is a mean zero stochastic process with uniformly bounded variance function and has a sub-Gaussian tail behavior. Condition (C4) assumes that all the covariates are uniformly bounded, which is satisfied in many practical situations.

3.1. Estimation consistency of group Lasso

For the matrix R_k at (2.5), by the Cholesky decomposition there exists a matrix Q_k such that $R_k = d_k Q_k' Q_k$.

Let Q_{kb} be the smallest eigenvalue of matrix Q_k, $Q_b = \min_k Q_{kb}$, $d_a = \max_k d_k$, $d_b = \min_k d_k$, $d = d_a/d_b$, $N = \sum_{i=1}^n n_i$ and $m_n = \sum_{k=1}^{p_n} d_k$. Thus N is the number of total observations, m_n is the number of all approximation coefficients in the basis expansions. Note that for $k = 1, \ldots, p_n$, d_k can increase as n increases to give a more accurate approximation. For example, as in nonparametric regression, we can choose $d_k = O(n^\tau)$ for some constant $0 < \tau < 1/2$. With $\bar{c} = c^*/c_*$, (C1), let

$$M_1 = 2 + 4d\bar{c}, \; M_2 = \frac{2}{3} + \frac{4}{9} d\bar{c}(7 + 4\bar{c}), \tag{3.1}$$

and consider the constraint

$$\lambda \geq \max\{\lambda_0, \lambda_{n,p_n}\}, \tag{3.2}$$

where $\lambda_0 = \inf\{\lambda : M_1 q_n + 1 \le q^*\}$, $\inf \emptyset = \infty$, and $\lambda_{n,p_n} = 2\sigma\rho\sqrt{8(1+c_0)d_a d^2 q^* \bar{c} N c^* \log m_n}$, $c_0 \ge 0$. Note that when q^* is fixed, $\lambda_{np_n} = O_p(\rho(N \log m_n)^{1/2})$.

Let $\hat{A} = \{k : \|\hat{\beta}_k\|_2 \ne 0, 1 \le k \le p_n\}$ represent the set of indices of the variables selected by the group Lasso. The cardinality of $\hat{A}$ is

$$\hat{q} = |\hat{A}| = \#\{k : \|\hat{\beta}_k\|_2 \ne 0, 1 \le k \le p_n\}. \tag{3.3}$$

This describes the dimension of the selected model; if $\hat{q} = O(q_n)$, then the size of the selected model is of the same order as the underlying model. To measure the important variables missing in the selected model, take

$$\xi_2^2 = \sum_{k \notin A_0} \|\beta_k\|_2^2 I\{\|\hat{\beta}_k\|_2 = 0\}. \tag{3.4}$$

Theorem 1. *Assume* (C1)−(C4) *and that* $\eta_1 \le \rho$. *Let* $\hat{q}$ *and* ξ_2 *be defined as in* (3.3) *and* (3.4), *respectively, for the model selected by the group Lasso from* (2.6). *Let* M_1 *and* M_2 *be defined as in* (3.1). *If the constraint* (3.2) *is satisfied, then, with probability converging to 1 and* $B_1^2(\lambda) = \lambda^2 d_b^2 q_n / c^* N$,

(**i**). $\hat{q} \le M_1 q_n$,

(**ii**). $\xi_2^2 \le 2\left(\dfrac{M^2 \rho q_n^{3/2}}{n c_*} + \rho\sqrt{q_n}\right)^2 + 2\dfrac{M_2 B_1^2(\lambda)}{c_* N}$.

Part (i) of Theorem 1 shows that the group Lasso selects a model whose dimension is comparable to the underlying model, regardless of the large number of unimportant variables. Part (ii) implies that all the variables with coefficient functions $\|\beta_k\|_2^2 \ge 2\left(M^2 q_n^{3/2}\rho/(nc_*) + q_n^{1/2}\rho\right)^2 + 2M_2 B_1^2/(c_* N)$ are selected in the model with high probability.

Let $\tilde{\beta}_k(t) = E(\hat{\beta}_k(t))$ be the mean of $\hat{\beta}_k(t)$ conditional on X. It is useful to consider the decomposition $\hat{\beta}_k(t) - \beta_k(t) = \hat{\beta}_k(t) - \tilde{\beta}_k(t) + \tilde{\beta}_k(t) - \beta_k(t)$, where $\hat{\beta}_k(t) - \tilde{\beta}_k(t)$ and $\tilde{\beta}_k(t) - \beta_k(t)$ contribute to the variance and bias terms, respectively. Let $\|\beta\|_2 = (\sum_{k=1}^{p_n} \|\beta_k\|_2^2)^{1/2}$, where $\beta = \beta(t) = (\beta_1(t), \ldots, \beta_{p_n}(t))'$.

Theorem 2 (Convergency of group Lasso)**.** *Let* $\{\bar{c}, c_0, \sigma, d\}$ *be fixed and* $1 \le q_n \le n \le p_n \to \infty$. *Suppose that the conditions in Theorem* 1 *hold. Then, with probability converging to one,*

$$\|\tilde{\beta} - \beta\|_2 \le \left(\frac{M^2 q_n}{n c_*} + 1\right)\sqrt{q_n}\rho + \rho,$$

$$\|\hat{\beta} - \tilde{\beta}\|_2 \le \frac{2\sigma(M_1 \log m_n q_n)^{1/2}}{Q_b\sqrt{d_a N c_*}} + \frac{\lambda(d_b M_1 q_n)^{1/2}}{Q_b N c_*}.$$

Consequently, with probability converging to one,

$$\|\hat{\beta}-\beta\|_2 \equiv \left[\sum_{k=1}^{p_n}(\|\hat{\beta}_k-\beta_k\|_2^2)\right]^{1/2} \le \frac{2\sigma\sqrt{M_1 q_n \log m_n}}{Q_b\sqrt{d_a N c_*}} + \frac{\lambda\sqrt{d_b M_1 q_n}}{Q_b N c_*} + \left(\frac{M^2 q_n}{n c_*}+1\right)\sqrt{q_n}\rho + \rho.$$

This theorem gives the rate of convergence of $\beta(t)$ as determined by four terms: the stochastic error and bias due to penalization (the first and second terms, $\|\hat{\beta}-\tilde{\beta}\|_2$), the basis approximation error (the third and fourth terms, $\|\tilde{\beta}-\beta\|_2$). Under the conditions of Theorem 1 and Theorem 2, the group Lasso is estimation consistent in model selection.

Immediately from Theorem 2, we have the following corollary.

Corollary 1. *Let $\lambda = O_p(\rho(N\log m_n)^{1/2})$. Suppose the conditions in Theorem 2 hold. Then Theorem 1 holds and, with probability converging to one, $\|\tilde{\beta}-\beta\|_2 = O_p((q_n^{3/2}d_a/n + q_n^{1/2}+1)\rho)$ and $\|\hat{\beta}-\tilde{\beta}\|_2 = O_p((d_a+\rho d_a^{5/2})(q_n\log m_n/N)^{1/2})$. Consequently, with probability converging to one,*

$$\|\hat{\beta}-\beta\|_2 = O_p\left((d_a+\rho d_a^{5/2})\sqrt{\frac{q_n\log m_n}{N}} + \left(\frac{q_n^{3/2}d_a}{n}+q_n^{1/2}+1\right)\rho\right).$$

This corollary follows by substituting the given λ value into the expression in the results of Theorem 2 and using $Q_b = O(d_a^{-1})$ by Lemma A.1 in Huang, Wu, and Zhou (2004). Note that $\tilde{\beta}_k(t)$ can be interpreted as the best approximation in the estimation space $G_k(t)$ to $\beta_k(t)$; under appropriate conditions, the bias term $\|\tilde{\beta}(t)-\beta(t)\|_2$ is asymptotically negligible to the variance $\|\hat{\beta}-\tilde{\beta}\|_2$. For example, a special extreme case: for $k=1,\ldots,p_n$, if $\beta_k(t)$ is a constant function independent of t, then (1.1) simplifies to a high-dimensional linear regression problem, $\|\tilde{\beta}-\beta\|_2 \equiv 0$. Thus by choosing appropriate λ, $\|\hat{\beta}-\beta\|_2 = O_p(\sqrt{q_n\log p_n/n})$ which is consistent with the result obtained in Zhang and Huang (2008). If we use B-spline basis functions to approximate $\beta(t)$, by Theorem 6.27 in Schumaker (1981), for $k=1,\ldots,p_n$, if $\beta_k(t)$ has bounded second derivatives and $\limsup_n d_a/d_b < \infty$, then $\rho = O(d_a^{-2})$, thus $\|\tilde{\beta}-\beta\|_2 = O_p((q_n^{3/2}d_a/n + q_n^{1/2}+1)d_a^{-2})$.

Corollary 2. *Suppose B-spline basis approximation, for $k=1,\ldots,p_n$, with coefficient functions $\beta_k(t)$ having bounded second derivatives, $\limsup_n d_a/d_b < \infty$, and the conditions in Corollary 1. Then*

$$\|\hat{\beta}-\beta\|_2 = O_p\left(\frac{d_a\sqrt{q_n\log m_n}}{\sqrt{N}} + \left(\frac{q_n^{3/2}d_a}{n}+q_n^{1/2}+1\right)d_a^{-2}\right).$$

For the conditions given in Corollary 2, the number of covariates p_n can be as large as $\exp(o(N/(d_a^2 q_n)))$, which can be much larger than n.

4. Selection Consistency of Adaptive Group Lasso

As just shown, the group Lasso has nice selection and estimation properties. It selects a model that has the same order of dimension as that of the underlying model. However, there is still room for improvement. To achieve variable selection accuracy and reduce the estimation bias of the group Lasso, we consider the adaptive group Lasso given an initial consistent estimator $\bar{\beta}(t)$. Take weights

$$\omega_k = \begin{cases} (\sqrt{d_k}\|\bar{\beta}_k\|_2)^{-1}, & \text{if} \|\bar{\beta}_k\|_2 > 0, \\ \infty, & \text{if} \|\bar{\beta}_k\|_2 = 0, \end{cases} \tag{4.1}$$

so ω_k is proportional to the inverse of the norm of $\bar{\beta}_k(t)$. Here we define $0 \cdot \infty = 0$. Thus the variables not included in the initial estimator are not included in the adaptive group Lasso. Given a zero-consistent initial estimator (Huang, Ma, and Zhang (2008)), the adaptive group Lasso penalty level λ_k goes to zero when $\|\gamma_k\|_2$ is large, which satisfies the conditions given in Lv and Fan (2009) for a penalty function having the oracle selection property.

Consider the following additional conditions.

(C5) The initial estimator $\bar{\beta}(t)$ is zero-consistent with rate r_n if

$$\max_{k \in A_0} \|\bar{\beta}_k\|_2 = o_p(1), \ r_n \max_{k \in A_0} \|\bar{\beta}_k\|_2 = O_p(1), \ r_n \to \infty$$

and there exists a constant $\xi_b > 0$ such that $P(\min_{k \in A_0^c} \|\bar{\beta}_k\|_2 > \xi_b \theta_b) \to 1$ as $n \to \infty$, where $\theta_b = \min_{k \in A_0^c} \|\beta_k\|_2$.

(C6) If $s_n = p_n - q_n$ is the number of unimportant variables,

$$\frac{\sqrt{d_a(\log q_n)}}{\sqrt{N} d_b \theta_b} + \frac{\tilde{\lambda} d_a^{3/2} q_n}{N d_b^2 \theta_b^2} + \frac{\sqrt{N d \log s_n}}{\tilde{\lambda} r_n} + \frac{d_a d^{3/2} q_n^2}{r_n \theta_b} \to 0.$$

(C7) All the eigenvalues of $\Sigma_{A_0^c A_0^c}$ are bounded away from zero and infinity.

Theorem 3. *Suppose that* (C3), (C5)–(C7) *are satisfied. Under NSC,*

$$P\big(\|\hat{\beta}_k^*\|_2 \neq 0, k \notin A_0 \text{ and } \|\hat{\beta}_k^*\|_2 = 0, k \in A_0\big) \to 1.$$

Theorem 3 shows that the adaptive group Lasso is selection consistent if an initial consistent estimator is available. Condition (C5) is critical, and is very difficult to establish. It assumes that we can consistently differentiate between important and unimportant variables. For fixed p_n and d_k, the ordinary least

squares estimator can be used as the initial estimator. However, when $p_n > n$, the least squares estimator is no longer feasible. Theorem 1 and Theorem 2 show that, under certain conditions, the group Lasso estimator is zero-consistent with rate

$$r_n = \left((d_a + \rho d_a^{5/2})(q_n \log \frac{m_n}{N})^{1/2} + (\frac{q_n^{3/2} d_a}{n} + q_n^{1/2} + 1)\rho \right)^{-1}.$$

Thus if we use the group Lasso estimator as the initial estimator for the adaptive group Lasso, we have the selection consistent property in Theorem 3. In addition, we reduce the dimensionality of the problem using this initial estimator. Condition (C6) restricts the numbers of important variables and basis functions, the penalty parameter, and the smallest important coefficient function (in the ℓ_2 sense). When d and θ_b are fixed constants and the n_i, $i = 1, \ldots, n$ are bounded, (C6) can be simplified to

$$\log \frac{q_n}{n} + \frac{\tilde{\lambda} q_n}{n} + \frac{\sqrt{n \log s_n}}{\tilde{\lambda} r_n} + \frac{q_n^2}{r_n} \to 0,$$

which can be obtained by choosing appropriate $\tilde{\lambda}$ and initial estimator. Condition (C7) assumes that the eigenvalues of $\Sigma_{A_0^c A_0^c}$ are finite and bounded away from zero; this is reasonable since the number of important variables is small in a sparse model.

Using the group Lasso result as the initial estimator for the adaptive group Lasso, we then have the following theorem.

Theorem 4. *Suppose the conditions of Theorem* 1 *hold, and* $\theta_b > t_b$ *for some constant* $t_b > 0$. *Let* $\tilde{\gamma} = O(N^\alpha)$ *for some* $0 < \alpha < 1/2$. *Then with probability converging to one,*

$$\|\hat{\beta}^* - \beta\|_2 = O_p \left((\frac{q_n}{d_a N})^{1/2} + (\frac{1}{d_a N})^{1/2} + (\frac{q_n^{3/2} d_a}{n} + q_n^{1/2} + 1)\rho \right),$$

For $k = 1, \ldots, p_n$, if all $\beta_k(t)$ are constant functions, q_n and the number of observations n_i for the ith subject are fixed, then the result of Theorem 4 is consistent with the well-known result for low-dimensional linear regression problem, $\|\hat{\beta}^* - \beta\|_2 = O_p(n^{-1/2})$. Moreover, similar to Corollary 2, if B-spline basis functions are used to approximate the regression coefficient functions, then we have the following.

Corollary 3. *Consider B-spline basis approximation and choose* $d_a = O(n^{1/5})$. *For* $k = 1, \ldots, p_n$, *the coefficient function* $\beta_k(t)$ *has a bounded second derivative,* $\limsup_n d_a/d_b < \infty$, *and the conditions in Theorem* 4 *hold. If* q_n *and* n_i *are fixed, then with probability converging to one,* $\|\hat{\beta}^* - \beta\|_2 = O_p\left(n^{-2/5}\right)$.

5. Numerical Studies

In this section, we derive a group coordinate descent algorithm to compute the group Lasso and adaptive group Lasso estimates in varying coefficient models. For the adaptive group Lasso, we use the group Lasso as the initial estimator. We compare the results from the group Lasso and the adaptive group Lasso with the results from the group SCAD (Antoniadis and Fan (2001); Wang, Li, and Huang (2008)).

5.1. The group coordinate descent algorithm

The group coordinate descent algorithm is a natural extension of standard coordinate descent, see for example, Fu (1998) and Friedman et al. (2007). Meier, Van de Geer and Bülmann (2008) also used a group coordinate descent for selecting groups of variables in high-dimensional logistic regression.

Let $\tilde{\gamma}_k = R_k^{1/2}\gamma_k$ and $\tilde{U}_k = U_k R_k^{-1/2}$ for $k = 1, \ldots, p_n$, so (2.6) can be rewritten as

$$\tilde{\gamma}^* = \arg\min_{\tilde{\gamma}} \frac{1}{2}(Y - \tilde{U}\tilde{\gamma})'(Y - \tilde{U}\tilde{\gamma}) + \sum_{k=1}^{p_n} \lambda \|\tilde{\gamma}_k\|_2. \tag{5.1}$$

The group Lasso estimates $\hat{\gamma}_k$ of (2.6) can then be obtained as $\hat{\gamma}_k = R_k^{-1/2}\tilde{\gamma}_k^*$ for $k = 1, \ldots, p_n$.

Denote by $L(\tilde{\gamma})$ the objective function in (5.1). Suppose we have estimates $\tilde{\gamma}_l$ for $l \neq j$ and wish to partially optimize with respect to $\tilde{\gamma}_j$. The gradient at $\tilde{\gamma}_j$ only exists if $\|\tilde{\gamma}_j\|_2 \neq 0$, and then

$$\frac{\partial L}{\partial \tilde{\gamma}_j} = -\frac{1}{N}\tilde{U}_j'(Y - \tilde{U}\tilde{\gamma}) + \lambda\frac{\tilde{\gamma}_j}{\|\tilde{\gamma}_j\|_2} = -\frac{1}{N}\tilde{U}_j'(Y - Y^{(-j)}) + \frac{1}{N}\tilde{U}_j'\tilde{U}_j + \lambda\frac{\tilde{\gamma}_j}{\|\tilde{\gamma}_j\|_2},$$

where $Y^{(-j)} = \sum_{l \neq j} \tilde{U}_l\tilde{\gamma}_l$ is the fitted value excluding the contribution from $\tilde{U}_j$. With $\tilde{U}_k'\tilde{U}_k/N = I_{d_k \times d_k}$ for $k = 1, \ldots, p_n$, simple calculus shows that the group coordinate-wise update has the form

$$\tilde{\gamma}_j = (1 - \frac{\lambda}{\|z_j\|_2})_+ z_j,$$

where $z_j = N^{-1}\tilde{U}_j'(Y - Y^{(-j)})$ and $(x)_+ = xI_{\{x \geq 0\}}$. Then for fixed λ, the above estimator $\tilde{\gamma}^*$ can be computed with the following iterative algorithm.

1. Center and standardize Y and $\tilde{U}$, such that $\sum_{i=1}^N Y_i = 0$, $\tilde{U}_j'\tilde{U}_j/n = I_{d_j \times d_j}$ for $j = 1, \ldots, p_n$.
2. Initialize $\tilde{\gamma}^{(0)} = 0$ and let $m = 0$, $r = Y$.

3. Calculate $z_j = N^{-1}\tilde{U}_j' r + \tilde{\gamma}_j^{(m)}$.
4. Update $\tilde{\gamma}_j^{(m+1)} = (1 - \lambda/\|z_j\|_2)_+ z_j$, for $j = 1, \ldots, p_n$.
5. Update $r = r - \tilde{U}_j(\tilde{\gamma}_j^{(m+1)} - \tilde{\gamma}_j^{(m)})$ and $m = m + 1$.
6. Repeat Steps 3-5 until convergence or a fixed number of maximum iterations has been reached. The $\tilde{\gamma}$ at convergence is the group Lasso estimate $\tilde{\gamma}^*$ of (5.1).
7. Change $\tilde{\gamma}^*$ to the original scale corresponding to original Y and $\tilde{U}$ before centering and standardization, and $\hat{\gamma}_k = R_k^{-1/2}\tilde{\gamma}_k^*$.

It can be seen that the idea of the group coordinate descent algorithm is simple but efficient, every update cycle requires only $O(Np_n)$ operations and the computational burden increases linearly with p_n. If the number of iterations is smaller than p_n, the solution is reached with even less computational burden than the Np_n^2 operations required to solve a linear regression problem by QR decomposition.

For the adaptive group Lasso, we can use the same coordinate descent algorithm by simple substitution, as in (2.6) and (5.1). We use the same set of cubic B-spline basis functions for each β_k. That is, $d_1 = \cdots = d_{p_n} \equiv d_0$, and $B_{kl} = B_{k'l}$ for $k \neq k'$, $1 \leq k, k' \leq p_n$. In our application, we apply the BIC criterion (Schwarz (1978)) to select (λ, d_0) for the group Lasso and $(\tilde{\lambda}, d_0)$ for the adaptive group Lasso. The BIC criterion is

$$BIC(\lambda, d_0) = \log(RSS_{\lambda,d_0}) + \log N \cdot \frac{df_{\lambda,d_0}}{N},$$

where RSS is the residual sum of squares, df is the number of selected variables for a given (λ, d_0). We choose d_0 from an increasing sequence of ten values, starting from 5 to 14; for any given value of d_0, we choose λ from a sequence of 100 values, starting from λ_{max} to $0.001\lambda_{max}$ with $\lambda_{max} = \max_{1\leq k\leq p_n} \|\tilde{U}_k' Y\|_2/\sqrt{d_0}$, where $\tilde{U}_k$ is the $N \times d_0$ submatrix of the "designed" matrix $\tilde{U}$ corresponding to the covariate X_k. This λ_{max} is the smallest penalty value that forces all the estimated coefficients to be zero.

5.2. Monte Carlo simulation

We used simulation to assess the performance of the proposed procedures. Because our main interest is in the case when p_n is large, we focused on the case $p_n > n$. We consider the model

$$y_i(t_{ij}) = \sum_{k=1}^{p_n} x_{ik}(t_{ij})\beta_k(t_{ij}) + \epsilon_i(t_{ij}).$$

The time points t_{ij} for each individual subject are scheduled to be $\{1, \ldots, 30\}$, each scheduled time point has some probability to be skipped, then the number of actual observed time points n_i for different subject is different. This generating model is similar to the one in Wang, Li, and Huang (2008).

The first six variables x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5} and x_{i6}, $i = 1, \ldots, 100$, are the true relevant variables, and were simulated as follows: $x_{i1}(t)$ was uniform $[t/10, 2 + t/10]$ at any given time point t; $x_{ij}(t)$, $j = 2, \ldots, 5$, conditioned on $x_{i1}(t)$, were i.i.d. from the normal distribution with mean zero and variance $(1 + x_{i1}(t))/(2 + x_{i1}(t))$; x_{i6}, independent of x_{ij}, $j = 1, \ldots, 5$, was normal with mean $3\exp(t/30)$ and variance 1. For $k = 7, \ldots, 500$, each $x_{ik}(t)$, independent of others, was multivariate normal distribution with covariance structure $cov(x_{ik}(t), x_{ik}(s)) = 4\exp(-|t - s|)$. The random error $\epsilon_i(t)$ was $Z(t) + E(t)$, where $Z(t)$ had the same distribution as $x_{ik}, k = 7, \ldots, 500$, and $E(t)$ were independent measurement errors from $N(0, 2^2)$ at each time point t. The coefficient functions were

$$\beta_1(t) = 15 + 20\sin(\frac{\pi t}{15}),\ \beta_2(t) = 15 + 20\cos(\frac{\pi t}{15}),\ \beta_3(t) = 2 - 3\sin(\frac{\pi(t-25)}{15}),$$
$$\beta_4(t) = 2 - 3\cos(\frac{\pi(t-25)}{15}),\ \beta_5(t) = 6 - 0.2t^2,\ \beta_6(t) = -4 + \frac{(20-t)^3}{2000},$$
$$\beta_7(t) = \cdots = \beta_{500}(t) \equiv 0.$$

The observation time points t_{ij} for each individual were generated from scheduled time points $\{1, \ldots, 30\}$, each scheduled time point had a probability of 60% being skipped, and the actual observation time t_{ij} was obtained by adding a random perturbation from uniform $[-0.5, 0.5]$ to the non-skipped scheduled time.

We consider the cases $n = 50, 100, 200$ with $p_n = 500$, to see the performance of our proposed methods as sample size increases. The penalty parameters were selected using BIC. The results for the group Lasso, the adaptive group Lasso, and the group SCAD methods are given in Tables 1 and 2 based on 200 replications. The columns in Table 1 include the average number of variables (NV) selected, model error (ER), percentage of occasions on which correct variables were included in the selected model (%IN), and percentage of occasions on which the exactly correct variables were selected (%CS), with standard error in parentheses. Table 2 summarizes the mean square errors for the six important coefficient functions $N^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\hat{\beta}_j(t_{ij}) - \beta_j(t_{ij}))^2$, with standard error in parentheses.

Several observations can be obtained from Tables 1 and 2. The model that was selected by the adaptive group Lasso was similar to the one selected by the group SCAD, and better than the one selected by the group Lasso in terms of model error, the percentage of occasions on which the true variables were selected and the mean square errors for the important coefficient functions. The

Table 1. Simulation study. NG, number of selected variables; ER, model error; IN%, percentage of occasions on which the correct variables were included in the selected model; CS%, percentage of occasions on which exactly correct variables were selected, averaged over 200 replications. Enclosed in parentheses are the corresponding standard errors.

Results for high dimension cases, $p = 500$												
	adaptive group Lasso				group Lasso				group SCAD			
	NG	ER	IN%	CS%	NG	ER	IN%	CS%	NG	ER	IN%	CS%
$n = 200$	6.13	15.18	99	93	6.20	21.07	99	82	6.12	14.29	100	95
	(0.38)	(3.33)	(0.01)	(0.26)	(0.64)	(5.88)	(0.01)	(0.38)	(0.37)	(2.37)	(0.00)	(0.22)
$n = 100$	6.21	15.20	87	84	6.34	26.81	87	69	6.25	15.37	90	83
	(0.72)	(3.58)	(0.34)	(0.36)	(1.25)	(4.44)	(0.34)	(0.51)	(0.78)	(2.52)	(0.10)	(0.37)
$n = 50$	7.04	15.98	72	68	10.29	27.09	72	53	6.99	15.74	78	70
	(1.43)	(3.86)	(0.48)	(0.52)	(5.64)	(4.96)	(0.48)	(0.57)	(1.38)	(3.12)	(0.42)	(0.49)

Table 2. Simulation study. Mean square errors for the important coefficient functions based on 200 replications. Enclosed in parentheses are the corresponding standard errors.

	β_1	β_2	β_3	β_4	β_5	β_6
	$n = 200$					
adaptive group Lasso	6.49	4.46	2.28	1.36	1.49	7.72
	(2.71)	(1.21)	(1.10)	(0.95)	(0.98)	(3.58)
group Lasso	16.41	9.72	8.31	4.98	5.14	10.81
	(9.09)	(4.10)	(3.08)	(2.39)	(3.92)	(11.91)
group SCAD	6.46	4.44	2.27	1.34	1.45	7.59
	(2.70)	(1.20)	(1.09)	(0.93)	(0.94)	(3.31)
	$n = 100$					
adaptive group Lasso	8.19	7.65	3.97	2.02	2.14	9.10
	(3.82)	(2.17)	(1.44)	(1.30)	(1.35)	(5.27)
group Lasso	29.99	18.04	11.12	8.51	9.69	22.03
	(15.95)	(5.15)	(3.96)	(3.85)	(5.09)	(18.55)
group SCAD	8.20	7.65	3.95	2.01	2.12	9.70
	(3.84)	(2.17)	(1.39)	(1.28)	(1.30)	(6.05)
	$n = 50$					
adaptive group Lasso	9.08	8.20	4.49	3.41	3.86	9.72
	(3.88)	(3.84)	(1.76)	(1.73)	(1.75)	(5.56)
group Lasso	34.39	26.05	14.15	12.98	12.64	33.49
	(20.25)	(17.44)	(4.92)	(4.80)	(6.66)	(22.18)
group SCAD	9.05	8.22	4.51	3.38	3.82	9.73
	(3.86)	(3.91)	(1.77)	(1.64)	(1.72)	(5.64)

group Lasso included the correct variables with high probability. For smaller sample sizes, the performance of both methods was worse. This is expected since variable selection in models with a small number of observations is more difficult. To examine the estimated time-varying coefficient functions from the adaptive

Table 3. Yeast cell cycle study. Identified cooperative pairs of TFs involved in the cell cycle process.

	adaptive group Lasso	group Lasso
	MBP1-SWI6,	MBP1-SWI6,
	MCM1-NDD1,	MCM1-NDD1,
	FKH2-MCM1,	FKH2-MCM1,
	FKH2-NDD1,	FKH2-NDD1,
	SWI4-SWI6,	SWI4-SWI6,
	FHC1-GAT3,	FHL1-GAT3
	NRG1-YAP6,	NRG1-YAP6,
	GAT3-MSN4,	GAT3-MSN4,
	REB1-SKN7,	REB1-SKN7
	ACE2-REB1,	ACE2-REB1,
	GCN4-SUM1,	GAL4-RGM1,
	FKH1-FKH2,	GCN4-SUM1
	CIN5-NRG1,	FKH1-FKH2,
	SMP1-SWI5,	CIN5-NRG1,
	FKH1-NDD1,	SMP1-SWI5
	ACE2-SWI5,	FKH1-NDD1,
	CIN5-YAP6,	ACE2-SWI5,
	STB1-SWI4,	CIN5-YAP6
	ARG81-GCN4,	STB1-SWI4,
	NDD1-STB1,	ARG81-GCN4,
	NRG1-PHD1.	NDD1-STB1,
		DAL81-STP1,
		NRG1-PHD1.
No. of pairs	21	23

group Lasso, we plot them along with the true function components in Figure 1. The estimated coefficient functions are from the adaptive group Lasso method in one run when $n = 200$. From the graph, the estimators of the time-varying coefficient functions $\beta_k(t)$, $k = 3, 4, 5$, fit the true coefficient functions well, which is consistent with the mean square errors for the coefficient functions reported in Table 2.

These simulation results have the adaptive group Lasso with good selection and estimation performance, even when p is larger than n. They also suggest that the adaptive group Lasso can better the selection and estimation results of the group Lasso.

5.3. Identification of yeast cell cycle transcription factors

We apply our procedures to investigate the transcription factors (TFs) involved in the yeast cell cycle, which is helpful for understanding the regulation

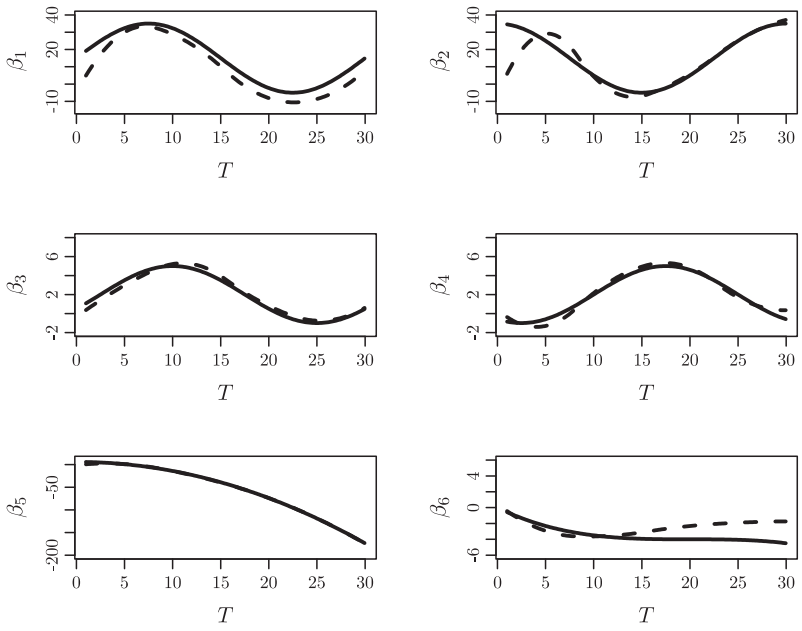

Figure 1. Adaptive group Lasso method. The estimated coefficient functions (dashed line) and true coefficient (solid line) functions in one run when $n = 200$

of yeast cell cycle genes. The cell cycle is an ordered set of events, culminating in cell growth and division into two daughter cells. Stages of the cell cycle are commonly divided into G1-S-G2-M. The G1 stage stands for "GAP 1". The S stage stands for "Synthesis"; is the stage when DNA replication occurs. The G2 stage stands for "GAP 2". The M stage stands for "mitosis", when nuclear (chromosomes separate) and cytoplasmic (cytokinesis) division occur. Coordinate expression of genes whose products contribute to stage-specific functions is a key feature of the cell cycle (Simon et al. (2001), Morgan (1997), Nasmyth (1996)). Transcription factors (TFs) have been identified that play critical roles in gene expression regulation. To understand how the cell cycle is regulated and how cell cycle regulates other biological processes, such as DNA replication and amino acids biosynthesis, it is useful to identify the cell cycle regulated transcription factors.

We apply the group Lasso and the adaptive group Lasso methods to identify the key transcription factors that play critical roles in the cell cycle regulations from a set of gene expression measurements which are captured at equally spaced sampling time points. The data set we use comes from Spellman et al. (1998). They measured the genome-wide mRNA levels for 6178 yeast ORFs simultaneously over approximately two cell cycle periods in a yeast culture synchronized

by α factor relative to a reference mRNA from an asynchronous yeast culture. The yeast cells were measured at 7-min intervals for 119 mins, with a total of 18 time points after synchronization. Using a model-based approach, Luan and Li (2003) identified 297 cell-cycle-regulated genes based on the α factor synchronization experiments. In our study, we consider 240 genes without missing values out of these 297 cell-cycle-regulated genes. Let $y_i(t_j)$ denote the log-expression level for gene i at time point t_j during the cell cycle process, for $i = 1, \ldots, 240$ and $j = 1, \ldots, 18$. We then use the chromatin immunoprecipitation (ChIP-chip) data of Lee et al. (2002) to derive the binding probabilities x_{ik} for these 240 cell-cycle-regulated genes for a total of 96 transcriptional factors with at least one nonzero binding probability in the 240 genes. We assume the following varying coefficient model to link the binding probabilities to the gene expression levels

$$y_i(t_j) = \beta_0(t_j) + \sum_{k=1}^{96} \beta_k(t_j) x_{ik} + \epsilon_{it_j},$$

where $\beta_k(t_j)$ represents the effect of the kth TF on gene expression at time t_j during the cell cycle process. Our goal is to identify the TFs that might be related to the cell cycle regulated gene expression.

We used BIC to select the tuning parameters in the group Lasso and adaptive group Lasso. The selected tuning parameters were $d_0 = 7$, $\lambda = 0.89$ and $\tilde{\lambda} = 1.07$. The group Lasso identified a total of 67 TFs related to yeast cell-cycle processes, including 19 of the 21 known and experimentally verified cell-cycle related TFs. The other two TFs, LEU3 and MET31, were not selected by the group Lasso method. Using the result from the group Lasso as the initial estimator for the adaptive group Lasso, adaptive group Lasso identified a total of 54 TFs, including the same 19 of the 21 known and experimentally verified cell-cycle related TFs. In addition, all of the identified TFs showed certain estimated periodic transcriptional effects on the cell cycle regulated gene expression, for example, MBP1, SWI4, SWI6, MCM1, FKH1, FKH2, NDD1, SWI5, and ACE2 (Simon et al. (2001)). The transcriptional effects of these 9 TFs are shown in Figure 2 and Figure 3 estimated by the group Lasso and the adaptive group Lasso methods, respectively. MBP1, SWI4, and SWI6 control late G1 genes. MCM1, together with FKH1 or FKH2, recruits the NDD1 protein in late G2, and thus controls the transcription of G2/M genes. MCM1 is also involved in the transcription of some M/G1 genes. SWI5 and ACE2 regulate genes at the end of M and early G1 (Simon et al. (2001)).

Moreover, the identified key TFs from both the group Lasso and the adaptive group Lasso include many pairs of cooperative or synergistic pairs of TFs involved in the yeast cell cycle process reported in the literature (Banerjee and Zhang

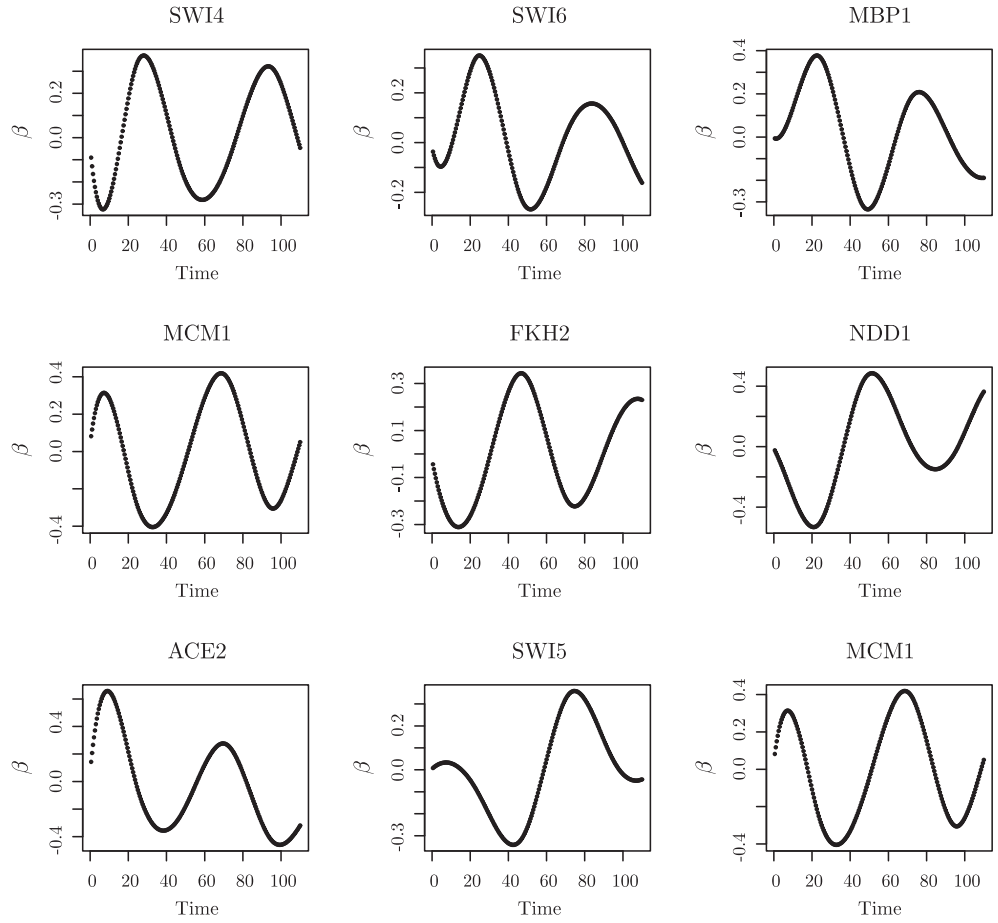

Figure 2. Application to yeast cell cycle gene expression data. The results are from the group Lasso.

(2003), Tsai, Lu, and Li (2005)). Among the 67 TFs identified by the group Lasso, 27 of them belong to the cooperative pairs of the TFs identified by Banerjee and Zhang (2003), including 23 out of 31 significant cooperative TF pairs. For the 54 TFs identified by the adaptive group Lasso, 25 of them belong to the cooperative pairs of the TFs, including 21 out of 31 significant cooperative TF pairs. The results are summarized in Table 2.

For this data set, the binding data are only available for 96 TFs. The sample size is larger than the number of variables. In order to see the performance of our proposed method with $p > n$, we artificially added 200 more variables to this data set. We randomly chose 200 values from the whole binding data set without replacement to add those 200 additional variables to each gene. We repeated this process 100 times. We first looked at the results concerning the 21 known TFs,

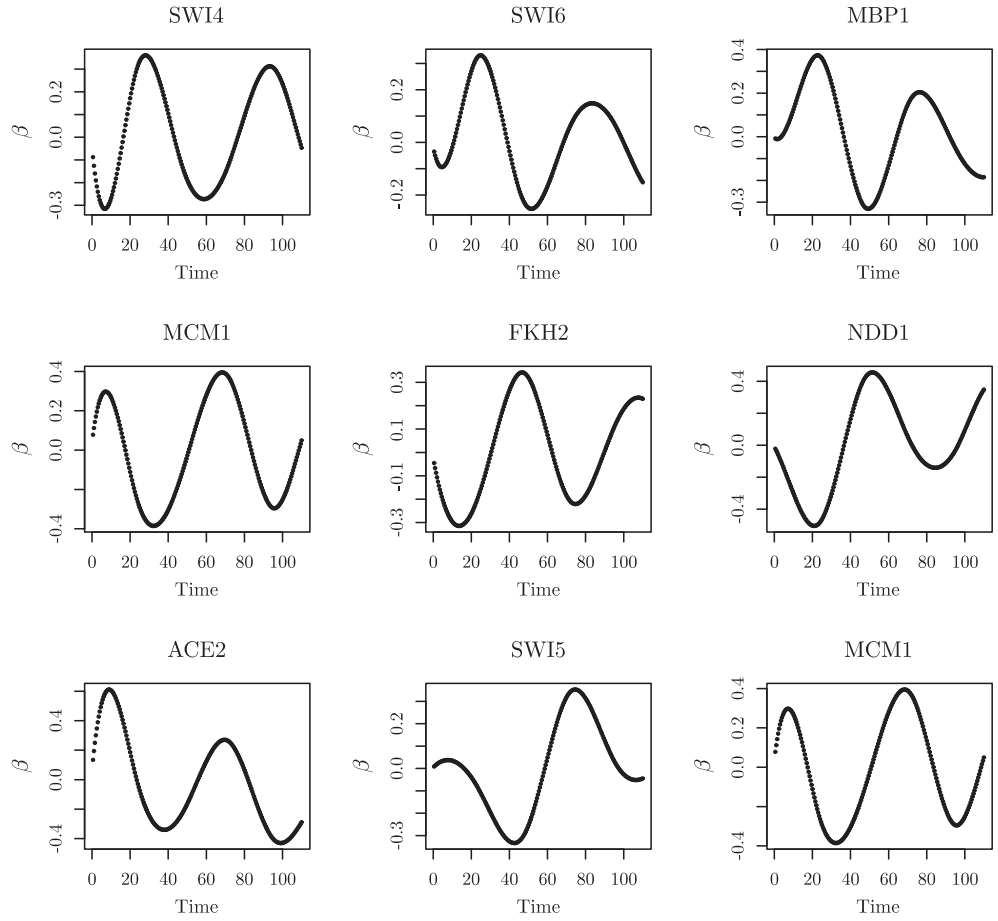

Figure 3. Application to yeast cell cycle gene expression data. The results are from the adaptive group Lasso.

since this is the only 'truth' we know about this data set. The average number of the 21 known important TFs identified was: the group Lasso 17.0 (standard deviation 0.12), the adaptive group Lasso 14.2 (standard deviation 0.42). The group Lasso tended to have a much higher false positive rate. The selected TFs sets after we artificially added 200 variables had a large intersection with the ones selected using only the data set itself. This suggests our method works well with many noisy variables in the model.

Finally, to compare the group Lasso and the adaptive group Lasso with simple linear regression with lasso penalty, we performed simple linear regression with binding probability as the predictors and the gene expression at each time point as the response with lasso penalty. We found only 1 TF significant related to cell cycle regulation. The result is not surprising, since the effects of the

TFs on gene expression levels are time-dependent. Overall, our procedures can effectively identify the important TFs that affect the gene expression over time.

6. Concluding Remarks

In this article, we studied the estimation and selection properties of the group Lasso and adaptive group Lasso in time varying coefficient models with high dimensional data. For the group Lasso method, we considered its properties in terms of the sparsity of the selected model, bias, and the convergence rate of the estimator, as given in Theorems 1 and 2. An interesting aspect in our study is that we can allow many small non-zero coefficient functions as long as the sum of their ℓ_2 norm is below a certain level. Our simulation results indicate that the group Lasso tends to select some non-important variables. An effective remedy then is to use the adaptive group Lasso. Compared with the group Lasso, the advantage of the adaptive group Lasso is that it has the oracle selection property. Moreover, the convergence rate of the adaptive group Lasso estimator is better. In addition, the computational cost is the same as the group Lasso. The adaptive group Lasso that uses group Lasso as the initial estimator is an effective way to analyze varying coefficient problems in sparse, high-dimensional settings.

In this paper, we have focused on the group Lasso and the adaptive group Lasso in the context of linear varying coefficient models. These methods can be applied in a similar way to other nonlinear and nonparametric regression models, but more work is needed.

Acknowledgements

The authors wish to thank the Editor, an associate editor and three referees for their helpful comments. The research of Jian Huang is supported in part by NIH grant R01CA120988 and NSF grant DMS 0805670. The research of Hongzhe Li is supported in part by NIH grant R01CA127334.

Appendix: Proofs

This section provides the proofs of the results in Sections 3 and 4. For simplicity, we often drop the subscript n from certain quantities, for example, we simply write p for p_n, q for q_n. Let $\tilde{y} = E(y) = X(t)\beta(t)$, $\tilde{\gamma} = \min_{\gamma=(\gamma_{A_0^c},0)'}(\tilde{y}-U\gamma)'(\tilde{y}-U\gamma)$, then $\tilde{\beta}(t) = B(t)\tilde{\gamma}$. We write $\hat{\beta}-\beta = (\hat{\beta}-\tilde{\beta})+(\tilde{\beta}-\beta)$, and find the rates of convergence of $\|\tilde{\beta}-\beta\|_2$ and $\|\hat{\beta}-\tilde{\beta}\|_2$.

For any two sequences $\{a_n, b_n, n = 1, 2, \ldots\}$, we write $a_n \asymp b_n$ if there are constants $0 < c_1 < c_2 < \infty$ such that $c_1 \le a_n/b_n \le c_2$ for all n sufficiently large, and write $a_n \asymp_p b_n$ if this inequality holds with probability converging to one.

Lemma A.1. $\|\tilde{\beta} - \beta\|_2 = O_p(\rho q^{3/2}/(nc_*) + \rho q^{1/2} + \rho)$.

Proof of Lemma A.1. By the properties of basis functions and (C2), there exists $g^*(t) = (g_1^*(t), \ldots, g_q^*(t), 0, \ldots, 0)$ for $g_k^* \in G_k(t)$, $k = 1, \ldots, q$ such that $\|g^* - \beta\|_\infty = \rho$. Thus $\exists\ \gamma^* = ({\gamma_1^*}', \ldots, {\gamma_p^*}')' = ({\gamma_{A_0^c}^*}', {\gamma_{A_0}^*}')'$ with $\gamma_k^* = (\gamma_{k1}^*, \ldots, \gamma_{kd_k}^*)'$ for $k \notin A_0$ and $\gamma_k^* = (0, \ldots, 0)'$ for $k \in A_0$, such that $g^*(t) = B(t)\gamma^*$.

By the definition of $\tilde{y}$, $\tilde{\gamma}$ and Lemma A.3 in Huang, Wu, and Zhou (2004), we have

$$\tilde{\gamma} = \begin{pmatrix} (U_{A_0^c}' U_{A_0^c})^{-1} U_{A_0^c}' \tilde{y} \\ 0 \end{pmatrix} = \begin{pmatrix} \tilde{\gamma}_{A_0^c} \\ \tilde{\gamma}_{A_0} \end{pmatrix}.$$

Thus $\tilde{\gamma}_{A_0^c} - \gamma_{A_0^c}^* = (U_{A_0^c}' U_{A_0^c})^{-1} U_{A_0^c}' \tilde{y} - (U_{A_0^c}' U_{A_0^c})^{-1} U_{A_0^c}' U_{A_0^c} \gamma_{A_0^c}^* = (U_{A_0^c}' U_{A_0^c})^{-1} U_{A_0^c}' (\tilde{y} - U_{A_0^c} \gamma_{A_0^c}^*)$. By Lemma 1 in Zhang and Huang (2008),

$$\|\tilde{\gamma}_{A_0^c} - \gamma_{A_0^c}^*\|_2^2 \le \frac{\|U_{A_0^c}'(\tilde{y} - U_{A_0^c}\gamma_{(1)}^*)\|_2^2}{n^2 c_*^2(|A_0^c|)}.$$

By (C4),

$$\begin{aligned} |\tilde{y}_i - U_{A_{0i}^c} \gamma_{A_0^c}^*| &= |x_i' \beta(t_i) - x_i' B_{A_0^c}(t_i) \gamma_{A_0^c}^*| = |x_i' \beta(t_i) - x_i' B(t_i) \gamma^*| \\ &= |x_i'(\beta(t_i) - g^*(t_i))| \le Mq\rho. \end{aligned}$$

It follows that

$$\begin{aligned} \|\tilde{\gamma}_{A_0^c} - \gamma_{A_0^c}^*\|_2^2 &\le \frac{1}{n^2 c_*^2(|A_0^c|)} \Big(\sum_{k=1}^{q} \sum_{l=1}^{d_k} \Big(\sum_{i=1}^{n} x_{ik} B_{kl}(t_i)(\tilde{y}_i - U_{A_{0i}^c} \gamma_{A_0^c}^*)\Big)^2\Big) \\ &\le \frac{M^2 q^2 \rho^2}{n^2 c_*^2(|A_0^c|)} \Big(\sum_{k=1}^{q} \sum_{l=1}^{d_k} \Big(\sum_{i=1}^{n} x_{ik} B_{kl}(t_i)\Big)^2\Big) \le \frac{(M^2 q\rho)^2 (\sum_{k=1}^{q} d_k)}{n^2 c_*^2(|A_0^c|)} \\ &\le \frac{(M^2 q\rho)^2 q d_a}{n^2 c_*^2(|A_0^c|)}. \end{aligned}$$

By Lemma A.1 in Huang, Wu, and Zhou (2004),

$$\|\tilde{\beta} - g^*\|_2 = \|B\tilde{\gamma} - B\gamma^*\|_2 \asymp_p \frac{\|\tilde{\gamma} - \gamma^*\|_2}{\sqrt{d_a}} \le \frac{M^2 \rho q^{3/2}}{n c_*(|A_0^c|)}.$$

Since $\|\tilde{\beta} - \beta\|_2 = \|\tilde{\beta} - g^* + g^* - \beta\|_2 \le \|\tilde{\beta} - g^*\|_2 + \|g^* - \beta\|_2$ and $\eta_1 \le \rho$,

$$\|\tilde{\beta} - \beta\|_2 \le \frac{M^2 q^{3/2} \rho}{n c_*(|A_0^c|)} + \sqrt{q}\rho + \rho.$$

This completes the proof of Lemma A.1.

Proof of Theorem 1. The proof of Theorem 1 is similar to the proof of the rate consistency of the group Lasso in Wei and Huang (2008). The only difference is in Step 3 of their proof of Theorem 1, where we need to consider the approximation error of the regression coefficient functions by basis expansion. Thus we omit the other details of the proof here.

From (2.6) and the definition of $\tilde{\gamma}$, we know

$$\hat{\gamma} = \arg\min_{\tilde{\gamma}} (y - U\tilde{\gamma})'(y - U\tilde{\gamma}) + \sum_{k=1}^{p} \lambda \|\tilde{\gamma}_k\|_k. \tag{A.1}$$

If $\tilde{\gamma}_k^* = Q_k \tilde{\gamma}_k$, $U_k^* = U_k Q_k^{-1}$, then (2.6) can be rewritten as

$$\hat{\gamma^*} = \arg\min_{\tilde{\gamma}^*} (y - U^*\tilde{\gamma}^*)'(y - U^*\tilde{\gamma}^*) + \sum_{k=1}^{p} \lambda \sqrt{d_k} \|\tilde{\gamma}_k^*\|_2, \tag{A.2}$$

and the estimator of (A.1) can be approximated by $\hat{\gamma_k} = Q_k^{-1}\hat{\gamma_k^*}$ where $\hat{\gamma}^*$ is the estimator of (A.2).

By the definition of $\tilde{\gamma}$, we have NSC on the regression coefficient $\tilde{\gamma}$, namely, $\sum_{k \in A_0} \|\tilde{\gamma}_k\|_2 = 0$. From Lemma A.3 in Huang, Wu, and Zhou (2004), the matrix U^* satisfies (C1). Compared with the sufficient conditions for the group Lasso problem given in Wei and Huang (2008), the only change is in the error terms in our (A.2). From (2.2), we have

$$y_i(t_{ij}) = \sum_{k=1}^{p} \sum_{l=1}^{d_k} x_{ik}(t_{ij}) \gamma_{kl} B_{kl}(t_{ij}) + \sum_{k=1}^{p} x_{ik}(t_{ij}) \rho_k(t_{ij}) + \epsilon_i(t_{ij}).$$

Define

$$\delta_n(ij) = \sum_{k=1}^{p} x_{ik}(t_{ij}) \rho_k(t_{ij}) + \epsilon_i(t_{ij}) = \rho_n(ij) + \epsilon_n(ij),$$

for $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. Let $\delta_n = (\delta_n(11), \ldots, \delta_n(nn_n))'$, $\epsilon_n = (\epsilon_n(11), \ldots, \epsilon_n(nn_n))'$ and $\rho_n = (\rho_n(11), \ldots, \rho_n(nn_n))'$. By (C4), we have $\|\rho_n\|_2 \le C_1 \sqrt{N\rho^2}$ for some constant $C_1 > 0$. Define

$$\begin{aligned} x_m &= \max_{|A|=m} \max_{\|b_{A_k}\|_2=1, k=1,\ldots,m} |\delta_n' \frac{V_A}{\|V_A\|_2}| \qquad \text{and} \\ x_m^* &= \max_{|A|=m} \max_{\|b_{A_k}\|_2=1, k=1,\ldots,m} |\epsilon_n' \frac{V_A}{\|V_A\|_2}|, \end{aligned} \tag{A.3}$$

where $V_A = U_A^*(U_A^{*\prime} U_A^*)^{-1} \bar{S}_A - (I - P_A) U^* \gamma^*$, $A_k \in A$, $\bar{S}_{A_k} = \lambda \sqrt{d_{A_k}} b_{A_k}$, and b_{A_k} is a d_{A_k}-dimensional unit vector. For a sufficiently large constant $C_2 > 0$,

also define, as Borel sets in $R^{n\times m_n}\times R^n$,

$$\Omega_{m_0}=\{(U^*,\epsilon): x_m\le \sigma\rho C_2\sqrt{(md_b\vee d_b)\log m_n}, \forall m\ge m_0\},$$
$$\Omega^*_{m_0}=\{(U^*,\epsilon): x^*_m\le \sigma\rho C_2\sqrt{(md_b\vee d_b)\log m_n}, \forall m\ge m_0\},$$

where $m_0\ge 0$. As in the proof of Theorem 1 of Wei and Huang (2008), $(U^*,\epsilon))\in \Omega_q\Rightarrow |\hat{A}|\le M_1 q$. By the triangle and Cauchy-Schwarz inequalities,

$$\frac{|\delta_n^{'}V_A|}{\|V_A\|_2}=\frac{|\epsilon_n^{'}V_A+\rho_n^{'}V_A|}{\|V_A\|_2}\le\frac{|\epsilon_n^{'}V_A|}{\|V_A\|_2}+\|\rho_n\|_2.$$

In the proof of Theorem 1 of Wei and Huang (2008), it is shown that $P\big((U^*,\epsilon_n)\in\Omega_0^*\big)\to 1$. Since $|\rho_n^{'}V_A|/\|V_A\|_2\le\|\rho_n\|_2\le C_1\sqrt{N\rho^2}$, we have for all $m\ge 0$ and p sufficiently large that $\|\rho_n\|_2\le C_1\sqrt{N\rho^2}\le\sigma\rho C_2\sqrt{(m\vee 1)d_b\log m_n}$. Then $P\left((U^*,\epsilon_n)\in\Omega_{m_0}\right)\to 1$. By the definition of $\lambda_{n,p}$, and x^*_m we have

$$\begin{aligned}(U^*,\epsilon)\in\Omega_{m_0}\Rightarrow |u^{'}\epsilon|^2\le (x^*_m)^2&\le(\sigma^2\rho^2C_2^2(md_b\vee d_b)\log m_n)\\&\le\frac{(|A_1|\vee d_b)\lambda_b^2 d_b}{4d_aNc^*},\ \text{for}|A_1|\ge m_0\ge 0,\end{aligned}\tag{A.4}$$

where u is defined as in the proof of Theorem 1 of Wei and Huang (2008). Since $\epsilon_1(t),\ldots,\epsilon_n(t)$ are iid with $E\epsilon_i(t_{ij})=0$, by (C3) and the proof of Theorem 1 of Wei and Huang (2008), we have $\hat{q}\le M_1 q$ and

$$\sum_{k\notin A_0}\|\tilde{\gamma}_k\|_2^2 I\{\|\hat{\gamma}_k\|_2=0\}\le\frac{M_2B_1^2}{c_*N}.\tag{A.5}$$

From Lemma A.1 in Huang, Wu, and Zhou (2004) and (A.5), we have,

$$\sum_{k\notin A_0}d_k\|\tilde{\beta}_k\|_2^2 I\{\|\hat{\beta}_k\|_T=0\}\le\frac{M_2B_1^2}{c_*N}.\tag{A.6}$$

By the definition of ξ_2 and the triangle inequality,

$$\begin{aligned}\xi_2^2&=\sum_{k\notin A_0}\|\beta_k\|_2^2 I\{\|\hat{\beta}_k\|_2=0\}\\&\le\sum_{k\notin A_0}(\|\tilde{\beta}_k-\beta_k\|_2+\|\tilde{\beta}_k\|_2)^2 I\{\|\hat{\beta}_k\|_2=0\}\\&\le 2\sum_{k\notin A_0}\|\tilde{\beta}_k-\beta_k\|_2^2 I\{\|\hat{\beta}_k\|_2=0\}+2\sum_{k\notin A_0}\|\tilde{\beta}_k\|_2^2 I\{\|\hat{\beta}_k\|_2=0\}.\end{aligned}$$

From Lemma A.1, (A.6) and $d_k\ge 1$, we have

$$\xi_2^2\le 2\left(\frac{M^2\rho q^{3/2}}{nc_*}+\rho\sqrt{q}\right)^2+2\frac{M_2B_1^2}{c_*N}.$$

This completes the proof of Theorem 1.

Proof of Theorem 2. From the proof of Theorem 1 and Theorem 2 of Wei and Huang (2008),

$$\|\hat{\gamma} - \tilde{\gamma}\|_2 = (\sum_{k=1}^{p} \|Q_k^{-1}(\hat{\gamma}_k^* - \tilde{\gamma}_k^*)\|_2^2)^{1/2} \leq \frac{2\sigma\sqrt{M_1 \log m_n q}}{Q_b\sqrt{Nc_*}} + \frac{\lambda d_b \sqrt{dM_1 q}}{Q_b N c_*}.$$

By Lemma A.1 in Huang, Wu, and Zhou (2004), we know that, $\|\hat{\beta} - \tilde{\beta}\|_2 = \|\hat{\gamma} - \tilde{\gamma}\|_k \asymp \|\hat{\gamma} - \tilde{\gamma}\|_2/\sqrt{d_a}$. Then

$$\|\hat{\beta} - \tilde{\beta}\|_2 \leq \frac{2\sigma\sqrt{M_1 \log m_n q}}{Q_b\sqrt{d_a N c_*}} + \frac{\lambda d_b \sqrt{dM_1 q}}{Q_b\sqrt{d_a} N c_*}.$$

By Lemma A.1, we know that $\|\tilde{\beta} - \beta\|_2 \leq M^2 \rho q^{3/2}/(nc_*(|A_0^c|)) + \sqrt{q}\rho + \rho$. Thus

$$\|\hat{\beta} - \beta\|_2 \leq \|\hat{\beta} - \tilde{\beta}\|_2 + \|\tilde{\beta} - \beta\|_2 \leq$$
$$\frac{2\sigma\sqrt{M_1 \log m_n q}}{Q_b\sqrt{d_a N c_*}} + \frac{\lambda\sqrt{d_b M_1 q}}{Q_b N c_*} + (\frac{M^2 q}{nc_*} + 1)\sqrt{q}\rho + \rho.$$

This completes the proof of Theorem 2.

Proof of Theorems 3 and 4. Theorems 3 and 4 can be obtained directly from Theorems 3 and 4 of Wei and Huang (2008) and Lemma A.1 in Huang, Wu, and Zhou (2004); we omit the proofs here.

References

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximation (with discussion). *J. Amer. Statist. Assoc.* **96**, 939-967.

Banerjee, N. and Zhang, M. Q. (2003). Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.* **31**, 7024-7031.

Chiang, C. T., Rice, J. A. and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J. Amer. Statist. Assoc.* **96**, 605-619.

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 Lv, J. (2010). A selective overview of variable selection in high dimensional feature space (invited review article). *Statist. Sinica* **20**, 101-148.

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.

Friedman, J., Hastie, Hoefling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **35**, 302-332.

Fu, W. J. (1998). Penalized regressions: the bridge versus the LASSO. *J. Comp. Graph. Statist.* **7**, 397-416.

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

Huang, J., Ma, S., Xie, H. L. and Zhang, C. H. (2007). A group bridge approach for variable selection. *Biometrika* **96**, 339-355.

Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive Lasso for high-dimensional regression models. *Statist. Sinica* **18**, 1603-1618.

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

Kim, Y., Kim, J. and Kim, Y. (2006). The blockwise sparse regression. *Statist. Sinica* **16**, 375-90.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M. and Simon, I. (2002). Transcriptional regulatory networks in S.cerevisae. *Science* **298**, 799-804.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed- effects model with B-splines. *Bioinformatics* **19**, 474-482.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498-3528.

Meier, L., Van de Geer, S. and Bülmann, P. (2008). The group lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B* **70**, 53-71.

Morgan, D. O. (1997). Cyclin-dependent kinases: engines, clocks, and microprocessors. *Annu. Rev. Cell Dev. Biol.* **13**, 261-291.

Nasmyth, K. (1996). At the heart of the budding yeast cell cycle. *Trends Genet.* **12**, 405-412.

Qu, A. and Li, R. (2006). Quadratic inference functions for varying coefficient models with longitudinal data. *Biometrics* **62**, 379-391.

Rice, J. A. (2004). Functional and longitudinal data analysis: perspectives on smoothing. *Statist. Sinica* **14**, 631-647.

Schumaker, L. (1981). *Spline Functions: basic theory.* Wiley, New York.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regualted genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of Cell* **9**, 3273-3297.

Schwarz, G.(1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakola, T. S. and Young, R. A. (2001). Serial regulation of transcriptinal regulators in the yeast cell cycle. *Cell* **106**, 697-708.

Tsai, H. K., Lu, S. H. H. and Li, W. H. (2005). Statistical methods for identifying yeast cell cycle transcription factors. *Proc. Nat. Acad. Sci.* **38**, 13532- 13537.

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

Wang, H. and Xia, Y. (2008). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association* **104**, 747-757

Wei, F. R. and Huang, J. (2008). Group selection in high-dimensional linear regression. Technical report No. 387, Department of Statistics and Actuarial Science, The University of Iowa.

Wu, C. O. and Chiang, C. T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statist. Sinica* **10**, 433-456

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

Zhang, C. H. and Huang, J. (2008). Model-selection consistency of the LASSO in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

Zhao, P., Rocha, G. and Yu, B. (2008). Grouped and hierarchical model selection through composite absolute penalties. To appear in the *Ann. Statist.*

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

Department of Mathematics, University of West Georgia, 1601 Maple Street, Carrollton, GA 30118, USA.

E-mail: fwei@westga.edu

Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA.

E-mail: jian-huang@uiowa.edu

Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

E-mail: hongzhe@upenn.edu

(Received December 2009; accepted June 2010)