Variable selection in quantile varying coefficient models with longitudinal data

Yanlin Tang\textsuperscript{a}, Huixia Judy Wang\textsuperscript{b}, Zhongyi Zhu\textsuperscript{c,}\textsuperscript{*}

\textsuperscript{a} Department of Mathematics, Tongji University, Shanghai, 200092, China
\textsuperscript{b} Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA
\textsuperscript{c} Department of Statistics, Fudan University, Shanghai, 200433, China

\begin{abstract}
In this paper, we develop a new variable selection procedure for quantile varying coefficient models with longitudinal data. The proposed method is based on basis function approximation and a class of group versions of the adaptive LASSO penalty, which penalizes the $L_\gamma$ norm of the within-group coefficients with $\gamma \geq 1$. We show that with properly chosen adaptive group weights in the penalization, the resulting penalized estimators are consistent in variable selection, and the estimated functional coefficients retain the optimal convergence rate of nonparametric estimators under the true model. We assess the finite sample performance of the proposed procedure by an extensive simulation study, and the analysis of an AIDS data set and a yeast cell-cycle gene expression data set.

\end{abstract}

\section{1. Introduction}

Varying coefficient models are popular nonparametric fitting techniques, in which the response is linear in the covariates, but the coefficients are smooth nonparametric functions of some factors such as the measurement time. The varying coefficient models were first proposed by Hastie and Tibshirani (1993) for conditional mean regression with cross sectional data, and later extended to longitudinal studies by Chiang et al. (2001), Huang et al. (2002), and Qu and Li (2006), among others.

The quantile varying coefficient model is a valuable alternative to the conditional mean varying coefficient models. While mean regression is confined to estimating the mean function of the response, quantile regression offers a systematic strategy for examining how covariates influence the location, scale, and shape of the entire response distribution. In addition, quantile regression does not assume any parametric form on the error distribution, and thus is able to accommodate non-normal errors, as often seen in longitudinal studies. Several authors studied estimation and hypothesis testing in nonparametric or semi-parametric quantile regressions. For example, He and Shi (1994) and Kim (2007) studied quantile varying coefficient models with cross-sectional data, and Wei and He (2006) and Wang et al. (2009) studied quantile partially linear models and quantile partially linear varying coefficient models for longitudinal data, respectively. However, variable selection for quantile varying coefficient models has not been well studied. This paper aims to develop a new variable selection procedure for quantile varying coefficient models with longitudinal data.

Suppose that the experiment involves $n$ subjects, and the $i$th subject has $J_i \geq 1$ repeated measurements over time. At a given quantile level $0 < \tau < 1$, we assume the following quantile varying coefficient model:

$$y_{ij} = \mathbf{x}_{ij}' \beta(t_{ij}, \tau) + \epsilon_{ij}(\tau), \quad i = 1, \ldots, n, \quad j = 1, \ldots, J_i,$$

(1)
where \( y_{ij} \) is the \( j \)th observation of the \( i \)th subject, \( \beta(t, \tau) = (\beta_0(t, \tau), \beta_1(t, \tau), \ldots, \beta_p(t, \tau))' \) is the \((p + 1)\)-vector of unknown smooth functions of \( t \), which may vary across quantile levels \( \tau \), \( x_{ij} = (x_{ij}^{(0)}, x_{ij}^{(1)}, \ldots, x_{ij}^{(p)})' \) is the design vector with \( x_{ij}^{(0)} = 1 \), and \( \epsilon_{ij}(\tau) \) is the random error whose \( \tau \)th quantile conditional on \((x, t)\) equals zero. Without loss of generality, we assume that \( t \in [0, 1] \). Furthermore, we assume that the observations, and therefore \( \epsilon_{ij}(\tau) \), are dependent within the same subjects, but independent across subjects. The form of the error distribution and the intra-subject dependence structure are left unspecified.

Existing variable selection methods for nonparametric regression are mainly confined to estimating the conditional mean functions of the response variable. For example, Wang et al. (2008) and Wang and Xia (2009) proposed variable selection approaches for varying coefficient models, via group SCAD (Fan and Li, 2001) and adaptive group LASSO (Tibshirani, 1996; Yuan and Lin, 2006; Zou, 2006) penalties, respectively, and Huang et al. (2010) proposed an adaptive group LASSO method for nonparametric additive models. In all these papers, the penalization was applied to the \( L_2 \) norm of coefficients within each group. Wu and Liu (2009), Belloni and Chernozhukov (2011), and Wang et al. (2012) studied variable selection for quantile regression but all of them are confined to parametric regression models.

In this paper, we develop a new variable selection procedure for quantile varying coefficient models with longitudinal data, where the varying coefficients are approximated by basis expansions. Different from Wang et al. (2008) and Huang et al. (2010), we consider a wider class of adaptive group LASSO penalty, where the penalization is applied to the \( L_\gamma \) norm of the basis coefficients corresponding to each varying coefficient. For computational convenience, we focused on two specific norms with \( \gamma = 1 \) and 2 in this paper.

Since quantile regression involves a non-differentiable loss function that can be considered as an asymmetric \( L_1 \) function, the computation is more challenging when penalizing the \( L_1 \) norm of the within-group coefficients. This motivated us to consider the penalization based on group \( L_1 \) norm, for which standard linear programming can be employed to solve the optimization problem; therefore, it is more computationally convenient. Even though group LASSO based on \( L_2 \) norm is more commonly used in the literature for group selection (Yuan and Lin, 2006; Wang and Leng, 2008; Wang et al., 2008; Huang et al., 2010), we demonstrate that with properly defined group adaptive weights, the penalization method with \( L_1 \) norm and that with \( L_2 \) norm are asymptotically equivalent: both are able to identify the true model consistently, and the estimated quantile functional coefficients retain the optimal convergence rate of nonparametric estimators obtained under the true model. Our theoretical investigation also indicates that the asymptotic properties can be extended to more general \( L_\gamma \) norm with any \( \gamma \geq 1 \); see Section 6 for more detailed discussions.

The rest of the paper is organized as follows. In Section 2, we introduce the penalized estimation procedure. In Section 3, we establish the theoretical properties of the estimated varying coefficients. We assess the empirical performance of the proposed method through a simulation study in Section 4 and the analysis of two real data sets in Section 5. We conclude the paper with some remarks in Section 6. The proofs of the main theoretical results are deferred to the Appendix.

2. The proposed method

2.1. Estimation

For simplicity, we omit \( t \) from \( \beta(t, \tau) \) and \( \epsilon_{ij}(\tau) \) in model (1) wherever clear from the context, but it is helpful to keep in mind that those quantities are \( t \)-specific.

Let \( \pi(t) = (B_1(t), \ldots, B_{k_n+m+1}(t))' \) be a set of normalized \( B \)-spline basis functions of order \( m + 1 \), with \( k_n \) quasi-uniform internal knots. Let \( q_n = k_n + m + 1 \). We propose to approximate each \( \beta_k(t) \) by a linear combination of \( \pi(t) \): \( \beta_k(t) \approx \sum_{l=1}^{q_n} \theta_{k,l} B_l(t) = \pi(t)' \theta_k \), where \( \theta_k = (\theta_{k,1}, \ldots, \theta_{k,q_n})' \) is the spline coefficient vector; see Schumaker (1981, Chapter 4) for details on the construction of \( B \)-spline basis functions. In what follows, we often suppress the subscript of \( q_n \) for convenience. For simplicity, define \( \Pi(t, x) = (\pi(t)'_1, \ldots, \pi(t)'_p)' \), \( \pi_{ij} = \pi(t_{ij}, x_{ij}) \), and \( \Pi_{ij} = \Pi(t_{ij}, x_{ij}) \), \( i = 1, \ldots, n \), \( j = 1, \ldots, J_i \). Model (1) can be approximated by

\[
y_{ij} \approx \Pi_{ij}' \theta + \epsilon_{ij},
\]

where \( \theta = (\theta_0', \theta_1', \ldots, \theta_p')' \). Our interest lies in selecting the covariates that are relevant to the \( \tau \)th quantile of \( y \). We will exclude a covariate, e.g., \( x^{(0)} \), from model (2) if and only if every component of \( \theta_k \) equals 0. Therefore, we treat \( \theta_k \) as a group. Instead of selecting components with nonzero \( \theta_{k,l} \), we need to select groups with nonzero \( \theta_k \). We propose a penalization method by applying adaptive LASSO penalty to the \( L_\gamma \) norm of the coefficient group. Throughout the paper, we refer to this method as GAL, meaning that group adaptive LASSO penalty is applied to the \( L_\gamma \) norm of each coefficient group. We define the proposed estimator \( \hat{\theta} \) as the minimizer of the following penalized objective function

\[
l(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{J_i} \rho_{\gamma}(y_{ij} - \Pi_{ij}' \theta) + \lambda \gamma \sum_{k=0}^{p} \omega_k \| \theta_k \|_\gamma,
\]
where $\rho_r(h) = h \{ r - l(h < 0) \}$ is the quantile check loss function (Koenker and Bassett, 1978), $\lambda_n$ is a nonnegative regularization parameter that determines the sparsity of the solution, $\| \theta_k \|_y = (\sum_{l=1}^{q} |\theta_{k,l}|^{\gamma})^1/\gamma$ is the $L_y$-norm of $\theta_k$ with $\gamma \geq 1$, and $\omega_{k,y}$, $k = 0, \ldots, p$, are the penalty weights of the coefficient groups. The specific choices of $\omega_{k,y}$ are described in the computational algorithms in Section 2.2. For each $k = 0, 1, \ldots, p$, the coefficient function $\beta_k(t)$ can be estimated by $\hat{\beta}_k(t) = \pi(t)\hat{\theta}_k$.

In the proposed procedure, every $\beta_k(t)$ is characterized by a spline coefficient vector $\hat{\theta}_k$, which can be treated as a group. In this paper, we penalize the $L_1$ norm of the within-group coefficients and consider $\gamma = 1$ and 2 for empirical investigations. The $L_2$ penalty corresponds to the group LASSO as discussed in Yuan and Lin (2006) and Wang and Leng (2008) for conditional mean regression. When $\gamma = 1$, each basis coefficient receives a LASSO penalty. However, different from the adaptive LASSO penalization in Zou (2006), all components within the same group $k$ receive the same group-specific penalty $\lambda_n\omega_{k,1}$. For quantile regression, the method based on the $L_1$ norm penalty is computationally more convenient, because linear programming can be directly applied to solve the minimization problem.

2.2. Computational algorithm

We describe the computational procedures for the penalization methods with adaptive $L_1$ and $L_2$ penalties in Algorithms 1 and 2, respectively.

Algorithm 1.

When $\lambda_n$ and the $B$-spline basis functions are given, the minimization of (3) with $L_1$ group penalty can be written as a linear programming problem with the aid of slack variables. In this paper, we solve the linear programming problem with interior-point method (Portnoy and Koenker, 1997), which was shown to be faster than the more commonly used simplex algorithm, especially when the dimension of parameters is large. The detailed algorithm is described as follows.

Step 1: Obtain the initial estimator $\hat{\theta}$ by using the function “rq” in the R package quanreg, which is designed for linear quantile regression. In this paper, we choose $\hat{\theta}_k$ as the minimizer of (3) with $\lambda_n = 0$.

Step 2: Set $\omega_{0,1} = 0$ and $\omega_{k,1} = \| \hat{\theta}_k \|_1^{-1}$, $k = 1, \ldots, p$ as the common penalty weights for the $k$th group basis coefficients $(\theta_{k,1}, \ldots, \theta_{k,q})'$. The $\delta$ is some appropriately chosen positive value, for instance $\delta = 0.5, 1, 2$ as used in Zou (2006). Throughout our empirical studies, we set $\delta = 2$.

Step 3: Use the interior-point method to solve the linear programming problem

$$
\min_{\xi^+, \xi^-, \theta^+, \theta^-} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ \tau \xi_{ij}^+ + (1 - \tau) \xi_{ij}^- \right\} + \lambda_n \sum_{k=0}^{p} \omega_{k,1} \sum_{l=1}^{q} (\theta_{k,l}^+ - \theta_{k,l}^-)
$$

s.t. $\xi_{ij}^+ - \xi_{ij}^- = y_{ij} - \Pi_{ij}(\theta^+ - \theta^-)$,

where $\theta^+ = (\theta_{0,1}^+, \ldots, \theta_{0,q}^+, \ldots, \theta_{p,1}^+, \ldots, \theta_{p,q}^+)'$, $\theta^- = (\theta_{0,1}^-, \ldots, \theta_{0,q}^-, \ldots, \theta_{p,1}^-, \ldots, \theta_{p,q}^-)'$, $N = \sum_{i=1}^{n} j_i$ is the total number of observations, and $z^+ = zI(z > 0)$ and $z^- = -zI(z < 0)$ for any variable $z$.

Algorithm 2.

The minimization of (3) with $L_2$ group penalty contains non-differentiable asymmetric $L_1$ loss and $L_2$ penalty, which leads to a computational challenge. In this paper, we extend the local linear approximation (LLA) (Zou and Li, 2008) to a group version so that it works under the quantile regression settings, and estimate $\hat{\theta}$ by an iterative algorithm, where in each iteration, the estimate is obtained by standard linear programming. The iteration makes the computation more complicated than Algorithm 1 with $L_1$ penalty.

Specifically, we first approximate $\| \theta_k \|_2$ by

$$
\| \theta_k \|_2 \approx \| \theta_k^{(0)} \|_2 + \| \theta_k^{(1)} \|_2^{-1} |\theta_k^{(1)}|' \left( |\theta_k| - |\theta_k^{(0)}| \right) = \| \theta_k^{(0)} \|_2^{-1} \sum_{i=1}^{q} |\theta_{k,l}^{(0)}| \cdot |\theta_{k,l}|,
$$

where $\theta_k^{(0)} = \tilde{\theta}_k$ is some initial estimator as in Algorithm 1, and $|\theta_k| = (|\theta_{k,1}|, \ldots, |\theta_{k,q}|)'$. Noh and Park (2010) extended LLA to the group SCAD penalty, but their method still contained the $L_2$ norm in the penalty, which does not work in our setting. Then we can obtain $\hat{\theta}$ by iteratively conducting the following penalized quantile regression until convergence

$$
\theta_k^{(j)} = \arg \min_{\theta_k} \sum_{i=1}^{n} \sum_{j=1}^{j_i} \rho_r(y_{ij} - \Pi_{ij}\theta_k) + \lambda_n \sum_{k=0}^{p} \omega_{k,2} \| \theta_k^{(j-1)} \|_2^{-1} \sum_{l=1}^{q} |\theta_{k,l}^{(j-1)}| \cdot |\theta_{k,l}|,
$$

where $\omega_{0,2} = 0$, $\omega_{k,2} = \| \tilde{\theta}_k \|_2^{-4}$ with $\delta = 2$. The minimization of (5) can be solved by using an algorithm similar to the Step 3 of Algorithm 1. During the iteration, once the whole $k$th group is shrunk to zero, we will set $\hat{\theta}_k = 0$. In our simulation studies, the algorithm often converges in less than five iterations.
2.3. Selection of tuning parameters

For practical implementation, the tuning parameters need to be decided. Throughout the empirical studies in this paper, we consider cubic spline basis \((m = 3)\) with \(k_n\) interior knots chosen equally spaced on \((0, 1)\). We employ a data-driven procedure to choose the tuning parameters \(k_n\) and \(\lambda_n\), where \(k_n\) controls the smoothness of \(\hat{\beta}(t)\) and \(\lambda_n\) determines the sparsity of the solution. Searching over two-dimensional grids, we choose the optimal pair of tuning parameters \((k_n, \lambda_n)\) as the minimizer to the following Schwarz-type Information Criterion (Schwarz, 1978; Koenker et al., 1994),

\[
SIC(k_n, \lambda_n) = \log \left\{ \sum_{i=1}^{n} \sum_{j=1}^{h_i} \rho_t(y_{ij} - \Pi_j y_{i}) \right\} + \frac{\log N}{2N} edf,
\]

where \(edf\) denotes the effective degree of freedom, \(N\) denotes the total number of observations. In this paper, \(edf\) is defined as the number of interpolated \(y_j\)'s, i.e., the number of zero residuals, which provides a plausible measure for the effective degree of freedom of the fitted quantile regression model (Koenker et al., 1994; Li and Zhu, 2008). In this paper, a residual is treated as zero if its absolute value is smaller than \(10^{-6}\).

Remark 1. Schwarz’s Bayesian information criterion, SIC, was originally given as a criterion for evaluating models estimated by the maximum likelihood methods. Machado (1993) proved that minimizing SIC yields consistent model selection in parametric \(M\)-estimation, which cover quantile regression as a special case. For semi-parametric/nonparametric quantile regression, Koenker et al. (1994) and Wang et al. (2009) showed that SIC performs well in the selection of tuning parameters. Another commonly used criterion for selecting tuning parameters is cross-validation, which was shown tending to have an over-fitting effect in the resulting model (Wang et al., 2007b). In contrast, SIC tends to choose smaller models than the cross-validation method.

3. Large-sample properties

In this section, we show that for quantile varying coefficient models, the proposed GALP method leads to consistent variable selection, and the resulting estimator achieves the optimal convergence rate under the same smoothness assumptions in quantile varying coefficient models.

For the sake of convenience, we define \(X_i = (x_{i1}, \ldots, x_{ip})'\) as the \(J_i \times (p + 1)\) design matrix on the \(i\)th subject, and \(X = (X_1', \ldots, X_n')'.\) Throughout the paper, we use \(a_n \sim b_n\) to mean that there are constants \(0 < A < B < \infty\) such that \(A \leq a_n/b_n \leq B\) when \(n\) is large enough.

Assume that there are \(s\) relevant covariates in Model \((1)\). Without loss of generality, let \(\beta_k(t), k = 0, 1, \ldots, s,\) be the nonzero coefficient functions, and \(\beta_k(t) \equiv 0, k = s + 1, \ldots, p.\) The following assumptions are needed to obtain the asymptotic properties of \(\hat{\beta}(t)\).

A1 \(\beta_k(t) \in H_r, k = 0, 1, \ldots, \) \(s\) for some \(r > 3/2,\) where \(H_r\) is the collection of all functions on \([0, 1]\) whose \(d\)th order derivative is \(Hölder\) of order \(r,\) with \(r \equiv d + v.\)

A2 The distribution of \(t_j\) is absolutely continuous with density function \(g_j(\cdot).\) Furthermore, there exist constants \(0 < b < B\) such that the density \(g_j(\cdot)\) satisfies \(b \leq g_j(t) \leq B\) for \(t \in [0, 1].\)

A3 There exists a positive integer \(J\) such that \(\max_{1 \leq j \leq n} J_i \leq J < \infty.\)

A4 For all \(i\) and \(j,\) the random design vector \(x_{ij}\) is bounded in probability. The eigenvalues of the matrix \(N^{-1}X'X\) are bounded away from zero and infinity.

A5 The regression errors \(e_i(\tau), i = 1, \ldots, n,\) are independent random vectors; while for each \(i, e_i(\tau), j = 1, \ldots, J_i,\) are dependent. The \(r\)th quantile of \(e_i(\tau)\) conditional on \((x_i, t_i)\) equals zero. The density function of \(e_i(\tau),\) denoted by \(f(\cdot),\) is uniformly bounded, and it is continuous and bounded away from zero in a neighborhood of zero.

A6 Let \(\psi(\cdot) = \tau - I_{\{x < 0\}}.\) For each \(i, A_i = E\left[\psi(e_i(\tau))\psi(e_i(\tau))'\right] > 0\) and \(A_i\) is bounded.

A7 The sparsity parameter satisfies: \(n^{-1/2}k_n^{1/2} \lambda_n \rightarrow 0\) and \(k_n^{-3/2}\lambda_n \rightarrow \infty.\)

Theorem 1. Let \(\hat{\beta}_k(t) = \tau(t) \hat{\theta}_k, k = 1, \ldots, p,\) be the estimate of the \(k\)th quantile coefficient \(\beta_k(t),\) where \(\hat{\theta}\) is the minimizer of \((3)\) with \(\gamma = 1\) or \(2.\) Suppose assumptions A1–A7 hold, \(k_n \sim n^{1/(2r+1)}\) and \(\delta \geq 1,\) then we have

(a) \(\hat{\beta}_k(t) = 0, k = s + 1, \ldots, p,\) with probability approaching 1;

(b) \(N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{h_i} \left[ \hat{\beta}_k(t_{ij}) - \beta_k(t_{ij}) \right]^2 = O_p\left(n^{-2r/(2r+1)}\right), k = 0, 1, \ldots, s.\)

Theorem 1(a) suggests that the proposed penalized procedure is consistent in variable selection. Theorem 1(b) provides the rate of convergence in estimating the varying coefficient functions, and this rate is optimal for nonparametric regression under the same smoothness assumptions (Stone, 1982).
4. Simulation study

In this section, we use two simulated examples to investigate the empirical performance of the proposed GAL procedure and to compare it with the group SCAD (gSCAD, Wang et al., 2008) method proposed for conditional mean regression. Example 1 involves 23 predictors, while Example 2 corresponds to a more challenging design with 100 predictors. For both examples, we generate 500 data sets, each consisting of 200 subjects, and each subject is supposed to be measured at each time point (excluding time 1) with a 60% probability of being skipped, resulting in different $J_i$ for each subject. Example 2 has a balanced design, where measurement is obtained at each time point. The actual measurement times are generated by adding an $U(-0.5, 0.5)$ random deviate to the scheduled times. We focus on $\tau = 0.25$ and $\tau = 0.5$ in this study. Following the procedure described in Section 2.3, we search $(k_0, \lambda_n)$ over a two-dimensional grid by minimizing SIC.

4.1. Example 1

The data sets are generated from the following model:

$$y_{ij} = \beta_0(t_{ij}) + \sum_{k=1}^{23} \beta_k(t_{ij})x^{(k)}_{ij} + \epsilon_{ij}(\tau), \quad i = 1, \ldots, 200, \quad j = 1, \ldots, J_i.$$  

The three relevant variables, $x^{(k)}_{ij}$, $k = 1, 2, 3$, are simulated in a manner analogous to that of Wang et al. (2008): $x^{(1)}_{ij}$ is generated from $U(t_{ij}/10, 2 + t_{ij}/10)$, $x^{(2)}_{ij}$, conditioning on $x^{(1)}_{ij}$, is $N\left(0, (1 + x^{(1)}_{ij})/(2 + x^{(1)}_{ij})\right)$, and $x^{(3)}_{ij}$, independent of $x^{(1)}_{ij}$ and $x^{(2)}_{ij}$, is a Bernoulli random variable with success rate 0.6. The 20 redundant variables, $x^{(k)}_{ij}$, $k = 4, \ldots, 23$, are mutually independent, and for each $k$, $x^{(k)}_{ij}$ is generated from a Gaussian process with zero mean and a decayed exponential covariance

$$\text{cov}(x^{(k)}_{ij}, x^{(k)}_{iu}) = \begin{cases} 4 \exp\left(-|t_{ij} - t_{iu}|\right) & \text{if } i = u, \\ 0 & \text{if } i \neq u. \end{cases}$$

The random error $\epsilon_{ij}(\tau)$ is given by $\epsilon_{ij}(\tau) = \epsilon_{ij} - F^{-1}(\tau)$, with $F(\cdot)$ being the common cumulative distribution function of $\epsilon_{ij}$. Here $F^{-1}(\tau)$ is subtracted from $\epsilon_{ij}$ to make the $\tau$th quantile of $\epsilon_{ij}$ zero for identifiability purpose.

We consider two cases for generating $\epsilon_{ij}$. In Case 1, $\epsilon_{ij}$ is given by $Z_{ij} + E_{ij}$, where $Z_{ij}$ has the same distribution as $x^{(4)}_{ij}$, and $E_{ij} \sim N(0, 4)$. In Case 2, $\epsilon_{ij}$ is given by $6^{1/2}u_{ij}^{-1/2}Z_{ij}$, where $u_{ij} \sim \chi^2(3)$, and $Z_{ij}$ and $u_{ij}$ are mutually independent. The coefficient functions $\beta_k(t)$ are given by

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

$$\beta_2(t) = 6 - 0.2t; \quad \beta_3(t) = -4 + (20 - t)^3/2000,$$

and $\beta_k(t) \equiv 0$ for $k = 4, \ldots, 23$.

The top part of Table 1 gives the variable selection results from GAL, at 0.25th and 0.5th quantile regressions in Case 1. We compare our results to those from the mean regression method proposed by Wang et al. (2008), which is based on the $B$-spline approximation and the group SCAD penalty (gSCAD). The GAL1 and GAL2 perform very similar, and both select the true relevant variables 1–3 in all of the 500 runs at both $\tau$'s. Compared to gSCAD, GAL performs better.
selections. On average, GAL selects the 20 irrelevant variables incorrectly 0.45% of the time at $\tau = 0.5$ and 0.70% of the time at $\tau = 0.25$. GAL selects them 0.45% at $\tau = 0.5$ and 0.60% at $\tau = 0.25$, while gSCAD selects them 1.45% of the time. In addition, GAL selects the correct model 92.0% of the time at $\tau = 0.5$ and 89.6% at $\tau = 0.25$, while gSCAD selects the correct model 91.2% of the time at $\tau = 0.5$ and 89.0% at $\tau = 0.25$, while gSCAD selects the correct model only 79% of the time.

The bottom part of Table 1 summarizes the integrated squared bias (IBias) and the integrated variance (IVar) of $\hat{\beta}_k(t)$, $k = 0, \ldots, 3$, defined as

$$IBias \left\{ \hat{\beta}_k(t) \right\} = \frac{1}{300} \sum_{a=1}^{300} \left( \hat{\beta}_k(t_a) - \beta_k(t_a) \right)^2,$$

$$IVar \left\{ \hat{\beta}_k(t) \right\} = \frac{1}{300} \sum_{a=1}^{300} \frac{1}{500} \sum_{i=1}^{500} \left( \hat{\beta}_{k,i}(t_a) - \bar{\hat{\beta}}_k(t_a) \right)^2,$$

where $t_a = a/10$, $a = 6, \ldots, 305$, $\hat{\beta}_{k,i}(t_a)$ and $\bar{\hat{\beta}}_k(t_a)$ are the estimates of $\beta_k(t_a)$ and $\theta_k$ in the $i$th replicate, respectively, and $\hat{\beta}_k(t_a) = 1/500 \sum_{i=1}^{500} \hat{\beta}_{k,i}(t_a)$. From Table 1, we can see that the GAL$_{\nu}$ estimates have slightly larger IBias, but smaller IVar, than those of the gSCAD, except for $\beta_2(t)$. Wang et al. (2008) chose the tuning parameters through approximate cross validation (ACV). Results in Table 1 confirm Remark 1, that is, SIC tends to choose smaller models than the ACV method.

Table 2 summarizes the simulation results for Case 2, where the random error has a marginal $t(3)$ distribution. The GAL$_{\nu}$ provides more accurate variable selections than the mean method gSCAD. Furthermore, the proposed median estimation has clearly smaller IVar than the gSCAD estimation; the relative efficiencies between the mean and the median estimates of $\hat{\beta}_k(t)$, $k = 0, \ldots, 3$, are around 1.6, which implies that the median estimates are more robust than the mean estimates when the error distribution has heavy tails.

### 4.2. Example 2

To reflect the dimensionality of the yeast cell-cycle gene expression data in Section 5.2, we use Example 2 to assess the performance of the proposed method in models with high dimensional covariates. The data sets are generated from the following model:

$$y_{ij} = \beta_0(t_j) + \sum_{k=1}^{100} \beta_k(t_j)x_{ij}^{(k)} + \epsilon_{ij}(\tau), \quad i = 1, \ldots, 200, \quad j = 1, \ldots, 30,$$

where $x_{ij}^{(k)}$, $k = 1, \ldots, 20$ are relevant covariates, and $x_{ij}^{(k)}$, $k = 21, \ldots, 100$ correspond to redundant covariates. The covariates $x_{ij}^{(k)}$ are mutually independent, and they are generated the same way as the relevant variables in Example 1 for $k = 1, 2, 3$, and the same way as $x_{ij}^{(4)}$ in Example 1 for $k = 4, \ldots, 100$. The random error $\epsilon_{ij}(\tau)$ is generated from the multivariate normal distribution as in Example 1. In this example, the dimension of the spline coefficients is $100(k_n + m + 1)$, which is much greater than the number of subjects ($n = 200$), but smaller than the number of observations ($N = 6000$).
Table 2
Simulation results for Case 2 of Example 1.

<table>
<thead>
<tr>
<th></th>
<th>( \tau = 0.25 )</th>
<th>( \tau = 0.5 )</th>
<th>( gSCAD )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \text{GAL}_1 )</td>
<td>( \text{GAL}_2 )</td>
<td>( \text{GAL}_1 )</td>
</tr>
<tr>
<td>Variable selection results</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oracle Perc.</td>
<td>79.6</td>
<td>77.6</td>
<td>89.4</td>
</tr>
<tr>
<td>Aver.r</td>
<td>3.00</td>
<td>3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>Aver.z</td>
<td>0.24</td>
<td>0.25</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Estimation results

<table>
<thead>
<tr>
<th></th>
<th>( \tau = 0.25 )</th>
<th>( \tau = 0.5 )</th>
<th>( gSCAD )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \text{GAL}_1 )</td>
<td>( \text{GAL}_2 )</td>
<td>( \text{GAL}_1 )</td>
</tr>
<tr>
<td>( 100 \times \text{IBias} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_0(t) )</td>
<td>10.4</td>
<td>21.5</td>
<td>5.6</td>
</tr>
<tr>
<td>( \beta_1(t) )</td>
<td>1.8</td>
<td>5.2</td>
<td>1.0</td>
</tr>
<tr>
<td>( \beta_2(t) )</td>
<td>0.1</td>
<td>1.6</td>
<td>0.1</td>
</tr>
<tr>
<td>( \beta_3(t) )</td>
<td>0.7</td>
<td>4.1</td>
<td>0.3</td>
</tr>
<tr>
<td>( 100 \times \text{IVar} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_0(t) )</td>
<td>126.3</td>
<td>124.7</td>
<td>91.2</td>
</tr>
<tr>
<td>( \beta_1(t) )</td>
<td>15.4</td>
<td>14.5</td>
<td>10.8</td>
</tr>
<tr>
<td>( \beta_2(t) )</td>
<td>11.2</td>
<td>10.0</td>
<td>7.4</td>
</tr>
<tr>
<td>( \beta_3(t) )</td>
<td>25.9</td>
<td>34.7</td>
<td>17.4</td>
</tr>
</tbody>
</table>

Oracle Perc.: the percentage of replications that the exact true model is selected; Aver.r: average number of selected variables that are truly relevant; Aver.z: average number of redundant variables selected; IBias: integrated squared biases; IVar: integrated variance.

Table 3
Simulation results for Example 2.

<table>
<thead>
<tr>
<th></th>
<th>( \tau = 0.25 )</th>
<th>( \tau = 0.5 )</th>
<th>( gSCAD )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \text{GAL}_1 )</td>
<td>( \text{GAL}_2 )</td>
<td>( \text{GAL}_1 )</td>
</tr>
<tr>
<td>Variable selection results</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oracle Perc.</td>
<td>76.4</td>
<td>74.0</td>
<td>79.8</td>
</tr>
<tr>
<td>Aver.r</td>
<td>20.00</td>
<td>20.00</td>
<td>20.00</td>
</tr>
<tr>
<td>Aver.z</td>
<td>0.26</td>
<td>0.30</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Estimation accuracy

<table>
<thead>
<tr>
<th></th>
<th>( \tau = 0.25 )</th>
<th>( \tau = 0.5 )</th>
<th>( gSCAD )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \text{GAL}_1 )</td>
<td>( \text{GAL}_2 )</td>
<td>( \text{GAL}_1 )</td>
</tr>
<tr>
<td>MREE %</td>
<td>103.78</td>
<td>104.07</td>
<td>102.77</td>
</tr>
</tbody>
</table>

Oracle Perc.: the percentage of replications that the exact true model is selected; Aver.r: average number of selected variables that are truly relevant; Aver.z: average number of redundant variables selected; MREE: median of the relative estimation error with respect to the oracle estimates that are obtained under the true model.

The intercept and the first three coefficient functions are the same as in Example 1, and the rest coefficient functions are defined as following:

\[ \beta_4(t) = -1 + (t - 10)^2/100, \quad \beta_5(t) = \exp(t/10 - 1), \quad \beta_6(t) = \log(t), \]

\[ \beta_k(t) = (-1)^k, \quad k = 7, \ldots, 14, \quad \beta_k(t) = 1.5 \times (-1)^k, \quad k = 15, \ldots, 20, \]

and \( \beta_k(t) \equiv 0 \) for \( k = 21, \ldots, 100. \)

Table 3 summarizes the simulation results for Example 2. The top part summarizes the variable selection results from \( \text{GAL}_\gamma \) at \( \tau = 0.25 \) and 0.5. We find that the proposed \( \text{GAL}_\gamma \) selects the exact true models with high proportions. However, we find that \( gSCAD \) clearly under-performs for this high dimensional example. In general, \( gSCAD \) not only tends to select a lot of irrelevant variables, but also misses some important variables. Therefore, we exclude \( gSCAD \) from the comparison.

To evaluate the estimation accuracy of the proposed \( \text{GAL}_\gamma \) estimates, we consider the relative estimation error (REE):

\[
\text{REE} = \frac{100}{p} \sum_{k=0}^{p-1} \left| \frac{\hat{\beta}_k(t_a) - \hat{\beta}_k(t_a)}{\hat{\beta}_k(t_a)} \right|
\]

where \( t_a = a/10, \ a = 6, \ldots, 305 \), \( \hat{\beta}_k(t_a) \) is the \( \text{GAL}_\gamma \) estimate and \( \hat{\beta}_k(t_a) \) is the oracle estimate from the true model. The bottom part of Table 3 summarizes the median of the REE values (denoted as MREE). We can see that our estimates are very close to the oracle estimates.

**Remark 4.** Asymptotically, both group SCAD in Wang et al. (2008) and our proposed \( \text{GAL}_\gamma \) method enjoy variable selection consistency. The competitive performance of \( gSCAD \) in Example 1 confirms their theory. However, in Example 2 with high dimensional data, \( gSCAD \) performs poorly. One possible reason is that by applying the local quadratic approximation (LQA)
to SCAD penalty, the gSCAD estimator is the limit of the iterated ridge regression. It is known that ridge regression does not result exact zero estimates, so a cutoff value is needed to set the estimated coefficients to zero, which may affect the variable selection accuracy especially in the high dimensional case. On the other hand, our proposed GAL, for quantile regression will result exact zero estimates according to the property of linear programming, which leads to natural variable selection and better finite sample performance.

5. Empirical data analysis

In this section, we apply our proposed method to a longitudinal AIDS data and a yeast cell-cycle gene expression data. From the simulation study, we find that the group adaptive $L_1$ and $L_2$ penalties lead to similar results, so here we only consider the adaptive $L_1$ penalty to save some space, and also for computational convenience.

5.1. Longitudinal AIDS study

In this section, we analyze a subset of the data from the Multi-Center AIDS Cohort study. The data set contains the human immunodeficiency virus (HIV) status of 283 homosexual men who were infected with HIV during the follow-up period between 1984 and 1991. Each patient had a different number of repeated measurements and different measurement times. Details of the study design can be found in Kaslow et al. (1987). Several researchers have studied the same data set by fitting varying coefficient models (Huang et al., 2002; Fan and Zhang, 2000) or a semi-parametric model with quadratic and interaction effects (Fan and Li, 2004). The previous analysis aimed to describe the trend of the mean CD4 percentage depletion over time and to evaluate the effects of cigarette smoking, pre-HIV infection CD4 percentage, and age at infection fitting varying coefficient models (Huangetal., 2002; FanandZhang, 2000)orasemi-parametricmodelwithquadratic

Our analysis focuses on evaluating the time dependent effects of smoking status, age, PreCD4, and the quadratic and interaction of the covariates on different quantiles of the CD4 percentage population. We consider the following quantile varying coefficient model,

$$
y_{ij} = \beta_0(t, \tau) + \beta_1(t, \tau) x_{i1} + \beta_2(t, \tau) x_{i2} + \beta_3(t, \tau) x_{i3} + \beta_4(t, \tau) x_{i2}^2 + \beta_5(t, \tau) x_{i3}^2 + \beta_6(t, \tau) x_{i2} x_{i3} + \epsilon_{ij}(\tau),
$$

where $y_{ij}$ is the $i$th individual’s CD4 percentage at time $t_{ij}$ (in years), $x_{i1}$ is the smoking status, taking a value of 1 or 0 for a smoker or a nonsmoker, $x_{i2}$ is the standardized variable for age at HIV infection, and $x_{i3}$ is the standardized PreCD4 percentage of the $i$th individual. We focus on three quantile levels, $\tau = 0.25, 0.5$ and 0.75. The baseline function $\beta_0(t, \tau)$ represents the $\tau$th quantile of CD4 percentage $t$ years after the infection, for a nonsmoker with average PreCD4 percentage and average age at HIV infection. Using the SIC criterion in Section 2.2, the tuning parameter $k_n$ is selected as 2 for all the three quantiles, and $\lambda_n$ is selected as 4.8, 4.4 and 3.9, respectively.

The baseline coefficient $\beta_0(t, \tau)$ and the PreCD4 coefficient $\beta_3(t, \tau)$ are identified as nonzero at all the three quantiles. At median, Smoking and Age tend to have a negative interaction, that is, the elder smokers tend to have lower median CD4 counts, and this agrees with the mean regression findings in Fan and Li (2004). In addition, our method suggests a positive interaction between Smoking and PreCD4 in the upper quartile for larger $t$: smokers with higher PreCD4 tend to have higher CD4 counts after a period of follow-up. An examination of the box-plots of PreCD4 for the smokers and for the nonsmokers shows that smokers in general have higher PreCD4. Note that the smoking status indicates whether a person ever or never smoked after HIV infection. One possible explanation for the positive interaction between Smoking and PreCD4 is that patients with lower PreCD4 may choose to quit smoking due to medical concerns. Fig. 1 shows the estimated coefficient functions of $\beta_0(t, \tau)$, $\beta_3(t, \tau)$, $\beta_6(t, \tau)$ and $\beta_7(t, \tau)$, respectively. We also include the 90% point-wise confidence bands for the median estimates of the coefficient functions, and also for $\beta_7(t, \tau)$ at $\tau = 0.75$, where the standard errors are estimated by applying the bootstrap method with 200 bootstrap samples to the selected model (Huang et al., 2002).

This data set involves only $p = 8$ covariates. For comparison, we carry out an all-subset selection procedure to select among $2^8 = 256$ candidate models. Specifically, we fit each candidate model, and calculate the Schwarz-type Information Criterion (SIC), and then select the model with the smallest SIC. According to Wang and Qu (2009), this SIC-based all-subset selection method provides a consistent variable selection approach. With the all-subset selection approach, only the baseline coefficient $\beta_0(t, \tau)$ and the PreCD4 coefficient $\beta_3(t, \tau)$ are identified as nonzero at the three quartiles. To better understand the performance of the proposed GAL1 method and the all-subset selection procedure, in the following, we compare the two methods by assessing their selection stability and the prediction accuracy.

To compare the selection stability, we employ the following bootstrap method. For each bootstrap, we resample the subjects ($y_{ij}, x_{ij}, j = 1, \ldots, J_i$) with replacement, and apply the proposed variable selection method to the bootstrap sample. Table 4 summarizes the frequencies that the covariates are selected in the observed data and in the bootstrap data among 200 bootstraps. Using the GAL1 method, the interaction effect $\beta_6(t, \tau)$ is selected in 164 bootstrap samples at $\tau = 0.5$ and the effect $\beta_7(t, \tau)$ is selected in 148 bootstrap samples at $\tau = 0.75$, which indicates that the interaction effects are persistent. Both the proposed GAL1 method and the all-subset selection method are stable in variable selection. The models selected with the bootstrap samples agree with the models selected with the observed data 74.5% and 60% times at $\tau = 0.5$ and 0.75 for GAL1, and they agree 54% and 63.5% times for the all-subset selection method at $\tau = 0.5$ and 0.75, respectively.
Fig. 1. Estimated coefficient functional curves at $\tau = 0.25$ (dashed), $\tau = 0.5$ (solid) and $\tau = 0.75$ (open circle) in the AIDS study. Panels (a)–(d) show the estimated effects of Baseline, PreCD4, the interaction of Smoking and Age, and the interaction of Smoking and PreCD4, respectively. Panels (c) and (d) only show the estimated curves at median and at the third quartile, respectively, since the interaction effects are not selected at the other quantile levels. The shaded area indicates the 90% point-wise confidence band for the median estimate in (a)–(c) and also for $\beta_7(t, \tau)$ at $\tau = 0.75$ in (d).

Table 4

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>Observed</th>
<th>Bootstrap</th>
<th>Observed</th>
<th>Bootstrap</th>
<th>Observed</th>
<th>Bootstrap</th>
<th>Observed</th>
<th>Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0(t, \tau)$</td>
<td>1</td>
<td>200</td>
<td>1</td>
<td>200</td>
<td>1</td>
<td>200</td>
<td>1</td>
<td>200</td>
</tr>
<tr>
<td>$\beta_1(t, \tau)$</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_2(t, \tau)$</td>
<td>0</td>
<td>35</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>28</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>$\beta_3(t, \tau)$</td>
<td>1</td>
<td>194</td>
<td>1</td>
<td>200</td>
<td>1</td>
<td>199</td>
<td>1</td>
<td>200</td>
</tr>
<tr>
<td>$\beta_4(t, \tau)$</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>$\beta_5(t, \tau)$</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>$\beta_6(t, \tau)$</td>
<td>0</td>
<td>15</td>
<td>1</td>
<td>164</td>
<td>0</td>
<td>22</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>$\beta_7(t, \tau)$</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>148</td>
</tr>
<tr>
<td>$\beta_8(t, \tau)$</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

AgreeP: the percentage of bootstrap samples for which the selected models agree with the model selected with the observed sample.

To assess the prediction accuracy, we randomly split the data into training (183 subjects) and testing (100 subjects), and calculate the averaged prediction error obtained under the selected model over 50 random splits for predicting the $\tau$th conditional quantiles of $y_{ij}$ at $\tau = 0.5$ and $0.75$. The average prediction errors and the corresponding standard errors are (4.24, 0.03) and (4.25, 0.03) for the models selected by GAL$^1$ and the all-subset approach at $\tau = 0.5$, and they are (3.33, 0.02) and (3.33, 0.02) for GAL$^1$ and the all-subset method at $\tau = 0.75$, respectively.
The comparison suggests that the proposed method GAL$_1$ performs very similarly to the all-subset approach in terms of both selection stability and prediction accuracy. However, the proposed method is much faster in computation. For example, to obtain the final model at the median, the all-subset approach required 18.08 s, while our approach, which searched $(k_n, \lambda_n)$ on a $3 \times 10$ grid, used only 2.31 s. The all-subset approach requires searching over all possible combinations of covariates, which is not feasible for data sets with large or even moderate number of covariates. For instance, in the simulation Example 1 with $p = 23$, the all-subset selection requires fitting $2^{23} = 8,388,608$ models, which is computationally very challenging.

5.2. Yeast cell-cycle gene expression data

In this section, we apply the proposed method to a subset of the microarray time-course gene expression data set (Spellman et al., 1998). Luan and Li (2003) identified 297 cell cycle regulated genes using a model-based method. Wang et al. (2007a,b) used the ChIP data of Lee et al. (2002), and derived the binding probabilities for these 297 cell-cycle-regulated genes for a total of 96 transcription factors (TFs). By applying the gSCAD method, Wang et al. (2008) identified 71 TFs that might be related to the mean expression patterns of the 297 genes.

To identify the TFs that have nonzero time-dependent effects on the quantiles of the gene expression, we consider the following model

$$y_{it} = \beta_0(t, \tau) + \sum_{k=1}^{96} \beta_k(t, \tau)x_{ik} + \epsilon_{it}(\tau),$$

where $y_{it}$ denotes the log-expression level for gene $i$ at time $t$ (in days) during the cell-cycle process for $i = 1, \ldots, 297$ and $t = 0, 7, 14, \ldots, 119$, and $\beta_k(t, \tau)$ is the transcription effect of the $k$th TF on the $\tau$th quantile of $y_{it}$. We focus on three quantiles levels, $\tau = 0.25$, 0.5 and 0.75. Using the SIC criterion in Section 2.3, the tuning parameter $k_n$ is selected as 2, and $\lambda_n$ is selected as 12.5, 13.5 and 15 at the three quantiles, respectively.

At $\tau = 0.25, 0.5$ and 0.75, GAL$_1$ identifies 47, 40 and 44 TFs related to yeast cell-cycle processes, respectively, including 15, 15 and 16 of the 21 known and experimentally verified cell-cycle-related TFs. As an illustration, Fig. 2 shows the estimated time-dependent transcriptional effects of two known TFs that are selected by our method. The shaded areas correspond to the 90% point-wise bootstrap confidence bands for the median estimates obtained under the selected model with 200 bootstrap runs. The effects of MBP1 have similar profiles at three quantiles, and three quantile coefficient curves exhibit differences mainly for $t \in [40, 80]$. The SWI5 shows very different effects at three quantiles across the main support of $t$ except for $t \geq 100$, and this suggests possible population heterogeneity.

Table 5 summarizes the results of the proposed method at $\tau = 0.25, 0.5, 0.75$ and the results of gSCAD at the mean for identifying the 21 known TFs. The top 13 TFs are identified at all three quartiles and at the mean, the next four TFs are identified at mean and at least one of the quartiles, GCR1 and LEU3 are identified at mean but missed by the quantile regressions, GCN4 is identified at $\tau = 0.75$ but missed by the mean, and CBF1 is missed by all methods. In the following, we will focus on the cases on which the mean and the quantile methods have discrepancy.

To assess the significance of each TF effect, we also employ the rank score test in Wang et al. (2009) for testing the null hypothesis that the TF effect is zero. For LEU3, the $p$-values are 0.674, 0.877 and 0.845 at the three quartiles, suggesting that LEU3 has no significant impact on any quartiles of the gene expression. This is partly caused by the two extreme binding probabilities to the genes YMR183C and YDL238C, which are 20 times larger than the rest. After these two genes are...
Table 5
The results for identifying the 21 known TFs.

<table>
<thead>
<tr>
<th>TF</th>
<th>GAL_γ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>τ = 0.25</td>
</tr>
<tr>
<td>ABF1</td>
<td>1</td>
</tr>
<tr>
<td>ACE2</td>
<td>1</td>
</tr>
<tr>
<td>FKH1</td>
<td>1</td>
</tr>
<tr>
<td>FKH2</td>
<td>1</td>
</tr>
<tr>
<td>GCR2</td>
<td>1</td>
</tr>
<tr>
<td>MBP1</td>
<td>1</td>
</tr>
<tr>
<td>MCM1</td>
<td>1</td>
</tr>
<tr>
<td>NDD1</td>
<td>1</td>
</tr>
<tr>
<td>SKN7</td>
<td>1</td>
</tr>
<tr>
<td>STB1</td>
<td>1</td>
</tr>
<tr>
<td>SWI4</td>
<td>1</td>
</tr>
<tr>
<td>SWI5</td>
<td>1</td>
</tr>
<tr>
<td>SWI6</td>
<td>1</td>
</tr>
<tr>
<td>BAS1</td>
<td>0</td>
</tr>
<tr>
<td>MET31</td>
<td>1</td>
</tr>
<tr>
<td>REB1</td>
<td>1</td>
</tr>
<tr>
<td>STE12</td>
<td>0</td>
</tr>
<tr>
<td>GCR1</td>
<td>0</td>
</tr>
<tr>
<td>LEU3</td>
<td>0</td>
</tr>
<tr>
<td>GCN4</td>
<td>0</td>
</tr>
<tr>
<td>CBF1</td>
<td>0</td>
</tr>
</tbody>
</table>

*“1” means that the TF is selected in the model, and “0” means the opposite.

Table 6
Hypothesis testing results on 6 randomly chosen TFs identified by the gSCAD procedure but missed by the proposed variable selection method at median.

<table>
<thead>
<tr>
<th>TF</th>
<th>ARG81</th>
<th>CRZ1</th>
<th>FHL1</th>
<th>GTS1</th>
<th>HAP4</th>
<th>RFX1</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>0.919</td>
<td>0.047</td>
<td>&lt;0.001</td>
<td>0.57</td>
<td>0.057</td>
<td>0.316</td>
</tr>
</tbody>
</table>

As observed in our simulation study (see Remark 1), due to different criteria used for choosing the tuning parameters, our procedure tends to choose a relatively smaller model than the mean regression procedure $gSCAD$. For this yeast data set, the $gSCAD$ procedure identifies 71 TFs, among which 19 are the known and experimentally verified TFs. In contrast, our procedure identifies 40 TFs at median, including 15 known TFs. As the $gSCAD$ procedure tends to select larger models, the identified TFs are more likely to be false discoveries. For a closer comparison, we randomly select 6 TFs from the unknown (unverified) TFs that are identified by the $gSCAD$ procedure but missed by GAL_γ, and test the functional effects of the TFs at median using the rank score test. The testing results are summarized in Table 6. For FHL1, the rank score test gives a very small $p$-value. This could be a false negative that GAL_γ fails to identify because of shrinking. For the other five TFs, the rank score test suggests no strong association with the median response distribution, and this agrees with the results from GAL_γ. To determine whether these five TFs identified by $gSCAD$ are truly false positives, more biological validation is needed.

6. Discussion

We developed a class of group adaptive penalization methods to select covariates that are relevant to the quantiles of the response variable in longitudinal studies. In our proposed method, the varying coefficients were approximated by spline expansions. The relevant variables were selected automatically by penalizing the $L_γ$ norm of the basis coefficients corresponding to each varying coefficient in an adaptive and group manner. Both simulation and real life data analysis suggested that the proposed method performs well in terms of both estimation and variable selection, and it retains good selection stability. In this paper, we focused on two specific norms with $γ = 1$ and $2$ in the group adaptive penalization. We showed that both norms lead to estimations that are consistent in variable selection, and the estimated functional coefficients retain the optimal convergence rate with the oracle nonparametric estimation obtained under the true model. The proof of Theorem 1 in the Appendix suggests that, with properly defined adaptive group weights, these asymptotic properties can be extended to more general norms with any $γ ≥ 1$. 
Acknowledgments

The research of Tang is partially supported by NNSFC grant 11171074. The research of Wang is supported by NSF grant DMS-1007420. The research of Zhu is supported by NFFC grants 10931002 and 1091112038. We would like to thank Dr. Lifeng Wang and Dr. Hongze Li for sharing the yeast cell-cycle gene expression data and their computer code for the gSCAD method. We would like to thank the Editor, an associate editor and two anonymous reviewers for their constructive comments that led to a major improvement of this article.

Appendix

We first present a few lemmas that are necessary to prove Theorem 1. Lemma 1 follows directly from Corollary 6.21 of Schumaker (1981, Chapter 6). Lemma 2 follows lines of arguments similar to those used in the proofs of Theorem 4 and Proposition 4 in Chen (1991). The arguments used in the proofs of Lemma 3 is similar to those used in the proof of Theorem 2.1 in He and Shi (1994). Lemma 4 is a basic inequality. Therefore, we omit the proofs of these lemmas.

**Lemma 1.** Suppose that assumptions A1–A4 hold. For some constants $a_1$ and $a_2$, there exists a vector $\theta^0 = (\theta_0^0, \theta_1^0, \ldots, \theta_p^0)'$ such that:

\begin{align}
(\text{i)} \; & \|\theta_k^0\|_{L_1} \neq 0 \quad \text{if} \; k \in \{1, \ldots, s\}, \quad \theta_k^0 = 0_q \quad \text{if} \; k \in \{s+1, \ldots, p\}; \\
(\text{ii)} \; & \sup_{(t,x) \in (0,1) \times \mathbb{R}^{p+1}} |\Pi(t,x)' \theta^0 - x' \beta(t)| \leq a_1 k_n^{-r}; \\
(\text{iii)} \; & \sup_t |\beta_k(t) - \pi(t)' \theta^0_k| \leq a_2 k_n^{-r}, \quad k = 0, 1, \ldots, s.
\end{align}

Let $R_{nij} = \Pi_j^0' \theta^0 - \mathbf{x}_i \beta(t_i)$, $r_{njk} = \pi_j^0 \theta^0_k - \beta_k(t_i)$. By (ii) and (iii) of Lemma 1, it is easy to see that $\max_{i,j} |R_{nij}| \leq a_1 k_n^{-r}$ and $\max_{i,j,k} |r_{njk}| \leq a_2 k_n^{-r}$.

**Lemma 2.** Suppose that assumptions A1–A4 hold and $k_n \sim n^{1/(2r+1)}$. Then the eigenvalues of $k_n \mathbf{V}_n / N$ are uniformly bounded away from 0 and $\infty$ in probability, where $\mathbf{V}_n = (\pi_{11}, \ldots, \pi_{nL_n})(\pi_{11}, \ldots, \pi_{nL_n})'$.

**Lemma 3** (Consistency of $\hat{\theta}$). Let $\tilde{\theta}$ be the minimizer of (3) with $\lambda_n = 0$. Suppose that assumptions A1–A6 hold, $k_n \sim n^{1/(2r+1)}$ and $r > 3/2$. Then

$$
\|\tilde{\theta} - \theta^0\|_2 = O_p(n^{-1/2} k_n).
$$

**Lemma 4.** Suppose $\zeta$ is a vector of dimension $\kappa$ and $\gamma \geq 1$, then

$$
\|\zeta\|_{\gamma} \leq \|\zeta\|_1 \leq \kappa^{1/2} \|\zeta\|_2.
$$

**Proof of Theorem 1.**

First, we prove part (b) of Theorem 1.

For any given $\theta \in \mathbb{R}^{(p+1)q}$, let $\theta - \theta^0 \doteq u = (u_0', u_1', \ldots, u_p')'$, where $u_k = (u_{k1}, \ldots, u_{kq})'$, $k = 0, 1, \ldots, p$. Define

$$
G_N(u) \doteq \sum_{i=1}^n \sum_{j=1}^{L_n} \left\{ \rho_t(\epsilon_{ij} - \Pi_j^0 u - R_{nij}) - \rho_t(\epsilon_{ij} - R_{nij}) \right\}
$$

$$
= \sum_{i=1}^n \sum_{j=1}^{L_n} \left[ E \left\{ \rho_t(\epsilon_{ij} - \Pi_j^0 u - R_{nij}) - \rho_t(\epsilon_{ij} - R_{nij}) \right\} - \Pi_j^0 u \left\{ \tau - I(\epsilon_{ij} < 0) \right\} \right]
$$

$$
+ \sum_{i=1}^n \sum_{j=1}^{L_n} \left[ \rho_t(\epsilon_{ij} - \Pi_j^0 u - R_{nij}) - \rho_t(\epsilon_{ij} - R_{nij}) + \Pi_j^0 u (\tau - I(\epsilon_{ij} < 0)) \right]
$$

$$
- E \left\{ \rho_t(\epsilon_{ij} - \Pi_j^0 u - R_{nij}) - \rho_t(\epsilon_{ij} - R_{nij}) \right\}
$$

$$
= G_{N,1}(u) + G_{N,2}(u),
$$

where $G_{N,1}(u)$ is the first summation and $G_{N,2}(u)$ is the second one. Following arguments similar to those used in Lemma 3.2 of He and Shi (1994), we have

$$
\sup_{\|u\|_2 \leq k_n^{1/2} n^{-1/2}} k_n^{-1} |G_{N,2}(u)| = a_p(1)
$$

(8)
where \( \{L_n\} \) is any sequence that satisfies \( 1 \leq L_n \leq h_n^{10} \) for some \( 0 < h_0 < (r - 1)/(2r + 1) \). Following arguments similar to those used in Lemma 3.3 of He and Shi (1994), we have, for any \( \epsilon > 0 \), there exist \( L_\epsilon \) (sufficiently large) such that, as \( n \to \infty \),

\[
P \left( \inf_{\|u\|_2 = L_n h_n^{-1/2}} k_n^{-1} G_N(u) > 1 \right) > 1 - \epsilon. \tag{9}\]

Combining (7)–(9), we have, for some large enough positive constant \( L \),

\[
P \left( \inf_{\|u\|_2 = Lk_n^{-1/2}} k_n^{-1} G_N(u) > 1 \right) \to 1. \tag{10}\]

By Lemma 4 and the triangle inequality of the \( L_\gamma \) norm, we have \( \|\theta_k\|_\gamma - \|\theta_0^k\|_\gamma \leq \|\theta_k - \theta_0^k\|_\gamma \leq \|\theta_k - \theta_0^k\|_1 = \|u_k\|_1 \leq q^{1/2} \|u_k\|_2 \leq q^{1/2} \|u\|_2 \) for any \( \gamma \geq 1 \). By (6) and Lemma 3, there exists a positive constant \( C_0 \) such that \( \|\hat{\theta}_k\|_\gamma \geq \|\theta_0^k\|_2 \geq C_0 \), and therefore \( \omega_{k,\gamma} \leq C_0^{-\delta} \), \( k = 1, \ldots, s \). Due to the fact that \( s \) is finite and not increasing with \( n \), when \( \|\theta - \theta_0^k\|_2 = \|u\|_2 = Lk_n^{-1/2} \), we have

\[
\lambda_n \sum_{k=1}^p \omega_{k,\gamma} (\|\theta_k\|_\gamma - \|\theta_0^k\|_\gamma) \geq -\lambda_n \sum_{k=1}^p \omega_{k,\gamma} (\|\theta_k\|_\gamma - \|\theta_0^k\|_\gamma) \geq -s\lambda_n C_0^{-\delta} q^{1/2} Lk_n^{-1/2} = - (sL C_0^{-\delta} q^{1/2} L/n) k_n. \tag{11}\]

By assumption A7, \( sL C_0^{-\delta} q^{1/2} \lambda_n \to 0 \). For any \( \gamma \geq 1 \), combining (10) and (11), we have

\[
P \left( \inf_{\|u\|_2 = Lk_n^{-1/2}} G_N(u) + \lambda_n \sum_{k=1}^p \omega_{k,\gamma} (\|\theta_k\|_\gamma - \|\theta_0^k\|_\gamma) > 0 \right) \to 1.
\]

By the definition of \( G_N(\cdot) \),

\[
l(\theta) - l(\theta^0) = l(\theta^0 + u) - l(\theta^0) = G_N(u) + \lambda_n \sum_{k=1}^p \omega_{k,\gamma} (\|\theta_k\|_\gamma - \|\theta_0^k\|_\gamma).
\]

Therefore, there exists some \( L := L_\epsilon \) for any \( \epsilon > 0 \) and any \( \gamma \geq 1 \) such that as \( n \to \infty \),

\[
P \left( \inf_{\|u\|_2 = Lk_n^{-1/2}} l(\theta^0 + u) > l(\theta^0) \right) \to 1 - \epsilon.
\]

By the convexity of \( l(\cdot) \) and Corollary 25 of Eggleston (1958, Chapter 3), we obtain as \( n \to \infty \),

\[
P \left( \inf_{\|u\|_2 = Lk_n^{-1/2}} l(\theta^0 + u) > l(\theta^0) \right) \to 1 - \epsilon. \tag{12}\]

Combining (12) and the fact that \( l(\hat{\theta}) - l(\theta^0) \leq 0 \), we have, for any \( \gamma \geq 1 \),

\[
P \left( \|\hat{\theta} - \theta^0\|_2 \leq Lk_n^{-1/2} \right) > 1 - \epsilon. \tag{13}\]

It is easy to see that, for \( k = 0, 1, \ldots, p \),

\[
\frac{1}{N} \sum_{i=1}^n \sum_{j=1}^l \left( \beta_k(t_{ij}) - \beta_k(t_{ij}) \right)^2 \leq \frac{2}{N} \sum_{i=1}^n \sum_{j=1}^l \left( \pi_j(\hat{\theta}_k - \theta_0^k) \right)^2 + \frac{2}{N} \sum_{i=1}^n \sum_{j=1}^l r_{ijk}^2
\]

\[
\leq \frac{2}{N} (\hat{\theta}_k - \theta_0^k)^t V_N (\hat{\theta}_k - \theta_0^k) + 2a_k^2 k_n^{-2r}.
\]

By (13), \( \|\hat{\theta} - \theta^0\|_2 = O_{\lambda}(n^{-1} k_n^2) \), which implies \( \|\hat{\theta}_k - \theta_0^k\|_2 = O_{\lambda}(n^{-1} k_n^2) \). As \( k_n \sim n^{1/(2r + 1)} \), by Lemma 2, we have proved part b of Theorem 1.

In the following, we prove part (a) of Theorem 1, that is, with probability approaching 1, \( \hat{\theta}_k = 0 \), \( k = s + 1, \ldots, p \). A sufficient condition is that

\[
\frac{\partial l(\theta)}{\partial \theta_{k,l}} < 0, \quad \theta_{k,l} < 0,
\]

\[
\frac{\partial l(\theta)}{\partial \theta_{k,l}} > 0, \quad \theta_{k,l} > 0
\]
holds for \( k = s + 1, \ldots, p, \) \( l = 1, \ldots, m_n. \) This suffices to prove that

\[
\left\| \sum_{i=1}^{n} \sum_{j=1}^{l} \psi_{\epsilon}(y_{ij} - \Pi_{ij}^{y} \hat{\theta}) \chi_{y}^{(k)} \pi_{ij} \right\|_2 \leq \lambda_n \omega_{k, y}
\]  

(14)

holds for \( y = 1, 2 \) and \( k = s + 1, \ldots, p. \) By Lemma 3, \( \left\| \hat{\theta}_k \right\|_2 = O_p(n^{-1/2} k_n) \), therefore there exists some positive constant \( C_1 \) such that \( \left\| \hat{\theta}_k \right\|_1 \leq C_1 q^{1/2} n^{-1/2} k_n. \) Let \( C_2 = C_1^{-1} \), then

\[
\lambda_n \omega_{k, y} = \lambda_n \left( \left\| \hat{\theta}_k \right\|_y \right)^{-\delta} \geq \lambda_n \left( \left\| \hat{\theta}_k \right\|_y \right)^{-1} \geq \lambda_n \left( \left\| \hat{\theta}_k \right\|_1 \right)^{-1} \geq C_2 \lambda_n n^{1/2} k_n^{-3/2}.
\]  

(15)

Define

\[
S_n(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{l} \psi_{\epsilon}(y_{ij} - \Pi_{ij}^{y} \theta) \Pi_{ij}, \quad S^0_n = \sum_{i=1}^{n} \sum_{j=1}^{l} \psi_{\epsilon}(\epsilon_{ij}) \Pi_{ij}.
\]  

(16)

By Lemma 2, \( \left\| S_n^0 \right\|_2 = O_p(n^{1/2} k_n^2). \) Assume that \( \left\| \theta - \theta^0 \right\|_2 = O_p(n^{-1/2} k_n). \) Using Lemma 2 and the similar arguments as used in the proofs of Lemmas 8.4 and 8.5 in Wei and He (2006), we can show that for any \( L > 0, \)

\[
\sup_{\left\| \theta - \theta^0 \right\|_2 < L n^{-1/2}} \left\| S_n(\theta) - S_n(\theta^0) \right\|_2 = o_p(n^{1/2}).
\]  

(17)

Note that \( S_n(\theta^0) = \sum_{i=1}^{n} \sum_{j=1}^{l} \psi_{\epsilon}(\epsilon_{ij} - R_{ij}) \Pi_{ij}, \) \( \max_{i,j} |R_{ij}| \leq a_1 k_n^\tau, \) and

\[
\left\| S_n(\theta^0) - S^0_n \right\|_2 = O_p \left( \left\| E \left\{ S_n(\theta^0) - S^0_n \right\} \right\|_2 + \sqrt{\left\{ E \left\{ S_n(\theta^0) - S^0_n \right\} \right\}^2} \right).
\]

Let \( H_n = \sum_{i=1}^{n} \sum_{j=1}^{l} \Pi_{ij} \Pi_{ij}^{y} \), then by Lemma 2, the maximum eigenvalue of \( H_n, \lambda_{\max}(H_n) = O_p(n/k_n). \) By the Cauchy–Schwarz inequality, there exists a large enough positive constant \( C_3 \) such that

\[
\left\{ E \left\{ S_n(\theta^0) - S^0_n \right\} \right\}_2 = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{l} \Pi_{ij}^{y} E \left\{ \psi_{\epsilon}(\epsilon_{ij} - R_{ij}) - \psi_{\epsilon}(\epsilon_{ij}) \right\} \right\}_2 = O_p \left( \left\| \sum_{i=1}^{n} \sum_{j=1}^{l} \Pi_{ij}^{y} (O(R_{ij})) \right\|_2 \right).
\]

\[
\leq O_p \left( C_3 \lambda_{\max}(H_n)(nk^{-2})^{1/2} \right) = O_p(n^{1/2}).
\]

By Lemma 2 and the independence between \( \epsilon_i, \) there exists a large enough positive constant \( C_4 \) such that

\[
E \left\{ S_n(\theta^0) - S^0_n \right\} = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{l} \Pi_{ij}^{y} E \left\{ \psi_{\epsilon}(\epsilon_{ij}) - \psi_{\epsilon}(\epsilon_{ij} - R_{ij}) \right\} \right\}^2 \leq C_4 \left( \lambda_{\max}(H_n) a_1 k_n^\tau + \lambda_{\max}(H_n) a_1^2 k_n^{-2} \right) = o_p(n).
\]

Therefore, \( \left\| S_n(\theta^0) - S^0_n \right\|_2 = O_p(n^{-1/2}). \) Combining (17) and the convergence rate of \( \hat{\theta} \) obtained in (13), we have

\[
\left\| \hat{\theta}_k \right\|_2 \leq \left\| S_n(\hat{\theta}) - S_n(\theta^0) \right\|_2 + \left\| S_n(\theta^0) - S^0_n \right\|_2 = O_p(n^{1/2} k_n^{-1/2} + n^{-1/2}) = O_p(n^{1/2}).
\]

Recall the definition of \( S_n(\theta) \) in (16), we have

\[
\left\| \sum_{i=1}^{n} \sum_{j=1}^{l} \psi_{\epsilon}(y_{ij} - \Pi_{ij}^{y} \theta) \chi_{y}^{(k)} \pi_{ij} \right\|_2 = o_p(n^{1/2}), \) which is dominated by \( \lambda_n \omega_{k, y} \), \( y = 1, 2, \) \( k = s + 1, \ldots, p. \) This will not affect the asymptotic properties of the resulting penalized estimator.

\[ \square \]

References


