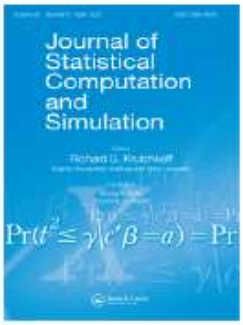

Taylor & Francis

Journal of Statistical Computation and Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Network-adaptive robust penalized estimation of time-varying coefficient models with longitudinal data

Kuangnan Fang, Xinyan Fan, Shuangge Ma & Qingzhao Zhang

To cite this article: Kuangnan Fang, Xinyan Fan, Shuangge Ma & Qingzhao Zhang (2022): Network-adaptive robust penalized estimation of time-varying coefficient models with longitudinal data, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2022.2055758

To link to this article: https://doi.org/10.1080/00949655.2022.2055758

Published online: 25 Mar 2022.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
https://doi.org/10.1080/00949655.2022.2055758

Network-adaptive robust penalized estimation of time-varying coefficient models with longitudinal data

Kuangnan Fang[a], Xinyan Fan[b], Shuangge Ma[c] and Qingzhao Zhang[a,d]

[a]Department of Statistics and Data Science, School of Economics, Xiamen University, Xiamen, People's Republic of China; [b]School of Statistics, Renmin University of China, Beijing, People's Republic of China; [c]Department of Biostatistics, Yale University, New Haven, CT, USA; [d]The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, People's Republic of China

ABSTRACT

Longitudinal data are commonly encountered in many fields. Many statistical models have been developed, among which the time-varying coefficient model has been shown to be effective for many practical problems. Long-tailed/contaminated distributions are not uncommon and cannot be accommodated using non-robust likelihood-based estimation. Another common limitation shared by many of the existing methods is the insufficient account for the interconnections among covariates. In this study, we adopt a least absolute deviation loss function to achieve robustness. For the selection of relevant covariates, a penalization approach is adopted. Significantly advancing from the existing literature, we describe the interconnections among covariates using a network structure and develop novel penalties to accommodate the network connectivity and connection measures. Consistency properties are rigorously established. Numerical studies, including both simulations and data analysis, demonstrate the competitive practical performance of the proposed method.

ARTICLE HISTORY
Received 18 August 2021
Accepted 16 March 2022

KEYWORDS
Basis function expansion; network structure; robustness; variable selection

1. Introduction

Longitudinal data arise in many scientific fields. Extensive methodological, theoretical, and numerical studies have been conducted. For detailed discussions, we refer to [1,2]. Consider a real-valued time variable t. At time t, denote $Y(t)$ as the response variable and $X(t) = (X^{(0)}(t), \ldots, X^{(p)}(t))^\top \in \mathbb{R}^{p+1}$ as the covariate vector with $X^{(0)} \equiv 1$. For simplicity of notation, assume that the support of t is $[0, 1]$. For a dataset with sample size N, denote the mth measurement of $(t, Y(t), X(t))$ for the ith subject as $(t_{im}, Y_i(t_{im}), X_i(t_{im}))$, where $1 \le m \le M_i$. With a slight abuse of notation, we also denote $Y_{im} = Y_i(t_{im})$ and $X_{im} = X_i(t_{im})$. A large number of statistical models have been developed for longitudinal data. Among them, the time-varying coefficient model has been shown to be very useful for many practical problems. The model postulates that

$$Y_{im} = X_{im}^\top \boldsymbol{\beta}(t_{im}) + \epsilon_{im}, \quad i = 1, \ldots, N;\ m = 1, \ldots, M_i, \tag{1}$$

CONTACT Qingzhao Zhang qzzhang@xmu.edu.cn, zhangqingzhao@amss.ac.cn

where $\beta(t) = (\beta_0(t), \ldots, \beta_p(t))^\top$ is the vector of time-varying coefficients and ϵ_{im} is the error term. In practice, it is usually assumed that the components of $\beta(t)$ are smooth. Estimation, inference, and computation of this model have been studied in quite a few publications. For reference, we refer to [3,4].

Literature review suggests that, despite extensive research, many of the existing methods still have limitations. Most of the existing time-varying coefficient models are based on the least-squares method and cannot accommodate long-tailed distributions/contamination. However, in practical data analysis, long-tailed distributions/contamination in the response variable are not uncommon and can be caused by special data-generating mechanisms, data mixture, human errors, as well as many other factors. Specific examples such as long-tailed gene expression distributions and the empirical distributions of stock returns are presented in [5,6], respectively. There is a demand for robust estimation for time-varying coefficient models. In more recent studies, with the cost of data collection going down, more and more variables have been collected. For example, hundreds of financial ratios, macroeconomic indices and investor sentiment indices have been selected in financial studies [7]. However, only a few of them are related to the phenomenon of interest. To improve the stability of estimation and also to identify the most relevant variables so as to create more interpretable models, regularized estimation and variable selection are often needed. There has been extensive literature on regularized estimation/variable selection with parametric models. With semi-parametric and nonparametric modelling, which is the case in this study, relevant studies include [8,9] and others. For longitudinal data, the most relevant studies may be [10] and others. However, it is noted that most of the existing studies have adopted non-robust loss functions, and research on 'robust loss function + regularized estimation/variable selection' is still much limited. Another common limitation shared by many of the existing studies is the lack of attention to the interconnections among covariates. Statistically, it is usually true that covariates are correlated. From a mechanistic perspective, some covariates can be interconnected. In some studies [11], it has been suggested that accommodating the interconnections among covariates can improve model interpretability, estimation, as well as selection of relevant variables. However, in the context of time-varying coefficient model for longitudinal data, there is still a lack of a method that can sufficiently account for such interconnections.

In this study, we consider the time-varying coefficient model for longitudinal data. Our goal is to develop a more effective data analysis method that can fill the aforementioned knowledge gaps. This study may advance from the existing literature in the following perspectives. First, to accommodate long-tailed distributions/contamination, we propose adopting a robust loss function. For the selection of relevant variables and regularized estimation, we adopt a penalization strategy. It is noted that although the strategy of 'robust loss function + penalization' may seem 'straightforward', it has not been investigated for model (1). Significantly advancing from a large number of existing literature, we propose describing the interconnections among covariates using a network structure and accommodating the most essential network properties in estimation. Although the proposed strategy has roots in some recently published studies on parametric models [11,12], the special characteristics of longitudinal data and nonparametric modelling bring significant new challenges. Overall, in terms of methodology, this study provides a practically useful new venue for analysing longitudinal data with long-tailed distributions/contamination

and multiple covariates. On the other hand, as the model/estimation strategy includes multiple others (parametric longitudinal models, varying coefficient models for cross-sectional observations, etc.) as special cases, our theoretical investigation can provide important insights beyond this study.

The rest of the article is organized as follows. The proposed analysis, its theoretical properties, and computational algorithms are described in Section 2. Numerical studies, including simulation in Section 3 and data analysis in Section 4, demonstrate its satisfactory performance. The article concludes with a discussion in Section 5. Additional technical details and numerical results are provided in Appendix.

2. Methods

2.1. Penalized estimation

We propose the penalized objective function

$$L_n(\boldsymbol{\beta}(\cdot)) = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i}\left|Y_{im} - \sum_{k=0}^{p}\beta_k(t_{im})X_{im}^{(k)}\right| + P(\boldsymbol{\beta},\lambda), \tag{2}$$

where the first term is the least absolute deviation (LAD)-based loss and can accommodate long-tailed distributions/contamination. The main advancement is the development of the penalty $P(\boldsymbol{\beta},\lambda)$. For $\lambda = (\lambda_1,\lambda_2)$ with $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$, we propose the penalty

$$\begin{aligned} P(\boldsymbol{\beta},\lambda) &= \lambda_1\sum_{k=0}^{p} w_k\|\beta_k(\cdot)\| + \frac{1}{2}\lambda_2\int_0^1 \sum_{1\leq j<k\leq p}|a_{jk}|(\beta_j(t) - s_{jk}\beta_k(t))^2\,\mathrm{d}t \\ &:= P_1(\cdot) + P_2(\cdot), \end{aligned}$$

where $\|\beta_k(\cdot)\| = (\int_0^1 \beta_k^2(t)\,\mathrm{d}t)^{1/2}$, a_{jk} is a measure of the connection between the jth and kth variables, $s_{jk} = \mathrm{sign}(a_{jk})$. We assume that a_{jk} for $j,k = 1,\ldots,p$ can be derived from prior knowledge. If not, when the observed time points are the same for all subjects, we can calculate a_{jk} by $a_{jk} = r_{jk}^3 I(|r_{jk}| > \tau)$ with

$$r_{jk} = \frac{1}{M_i}\sum_{m=1}^{M_i}\frac{\sum_{i=1}^{N}(X_{im}^{(k)} - \bar{X}_m^{(k)})(X_{im}^{(j)} - \bar{X}_m^{(j)})}{\sqrt{\sum_{i=1}^{N}(X_{im}^{(k)} - \bar{X}_m^{(k)})^2\sum_{i=1}^{N}(X_{im}^{(j)} - \bar{X}_m^{(j)})^2}},$$

for $j,k = 1,\ldots,p$, $\bar{X}_m^{(j)} = 1/N(\sum_{i=1}^{N} X_{im}^{(j)})$, and τ is the 95% quantile of the distribution of the sample correlation coefficient under the assumption that the jth and kth variables are uncorrelated. When the observed time points are subject-specific, we suggest to synchronize the data through preprocessing following [13,14].

The first penalty takes a Lasso form and penalizes the integrated ℓ_2-norms of the time-varying coefficients. Similar penalties have been considered in the literature [15]. Motivated by adaptive penalization, weights are imposed. Significantly different from the literature, the weights are not determined by initial estimates but by the network connectivity. Specifically, we define $w_0 = 1$ (for the intercept) and $w_k = 1/(\sum_{j\neq k}|a_{jk}| + \xi)$ for $k = 1,\ldots,p$. ξ is a small positive constant. This is motivated by the observation that highly

connected nodes are more likely to be important for outcome variables [12]. Using the connectivity to gauge variable selection has been considered in the context of thresholding with parametric models [12] and demonstrated to be successful.

In the second penalty, a_{jk} is the adjacency measure. For parametric models, the Laplacian penalization has been developed [11], which promotes adjacent nodes to have similar regression coefficients. The proposed penalty has a similar spirit. Different from the existing literature, it is imposed on functions. When two nodes are positively correlated, the penalty encourages similarity in their corresponding functions. When two nodes are negatively correlated, the penalty encourages their sum to be small, leading to shrinkage on both functions and/or opposite signs.

The network structure may also change over time owing to the time-dependent nature of the longitudinal data. That is, a_{jk} and other downstream measurements are functions of t. Another possibility is to use $a_{jk}(t)$ and $s_{jk}(t)$ in the penalty. However, we propose using the averages to simplify the calculation. Both P_1 and P_2 have shrinkage properties. λ_1 and λ_2 are data-dependently chosen to avoid double penalization.

Basis function expansion: Suppose that the coefficient function $\beta_k(\cdot)$ can be formed by a linear combination of basis functions, that is, $\beta_k(t) = \sum_{\ell=1}^{K_n} B_\ell(t)\gamma_{k\ell}$, where B_l's are the known basis functions, γ_{kl}'s are the unknown regression coefficients and K_n is the number of knots. For basis functions, we use natural cubic splines, which has been a popular choice in the literature. Although it is possible to have data-driven knots, to simply computation, we use equally spaced knots. Without loss of generality, for all time-varying coefficients, we use the same number of knots (denoted as K_n) and hence the same basis functions. As suggested by Fan et al. [16], we set $K_n = [N^{1/5}] + 2$, where $[b]$ denotes an integer part of b. Denote $\boldsymbol{B}(t) = (B_1(t), \ldots, B_{K_n}(t))^\top$, $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kK_n})^\top$, and $\Theta = \int_0^1 \boldsymbol{B}(t)\boldsymbol{B}(t)^\top \, \mathrm{d}t$. Then following [15], we have $\|\beta_k(\cdot)\| = (\gamma_k^\top \Theta \gamma_k)^{1/2}$ and

$$\frac{1}{2}\lambda_2 \sum_{1\le j<k\le p} |a_{jk}| \int_0^1 (\beta_j(t) - s_{jk}\beta_k(t))^2 \, \mathrm{d}t = \frac{1}{2}\lambda_2 \sum_{1\le j<k\le p} |a_{jk}|(\gamma_j - s_{jk}\gamma_k)^\top \Theta (\gamma_j - s_{jk}\gamma_k).$$

Then objective function (2) can be rewritten as

$$Q_n(\boldsymbol{\gamma}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i}\left|Y_{im} - \sum_{k=0}^{p}\sum_{l=1}^{K_n} X_{im}^{(k)} B_l(t_{im})\gamma_{kl}\right| + \lambda_1 \sum_{k=0}^{p} w_k(\gamma_k^\top \Theta \gamma_k)^{\frac{1}{2}} + \frac{\lambda_2}{2}\sum_{1\le j<k\le p} |a_{jk}|(\gamma_j - s_{jk}\gamma_k)^\top \Theta (\gamma_j - s_{jk}\gamma_k), \tag{3}$$

where $\boldsymbol{\gamma} = (\gamma_0^\top, \ldots, \gamma_p^\top)^\top$. An estimate of $\beta_k(t)$ can be obtained by $\hat{\beta}_k(t) = \sum_{l=1}^{K_n} \hat{\gamma}_{kl} B_{kl}(t)$, where $\hat{\gamma}_{kl}$'s are the minimizers of (3).

2.2. Asymptotic properties

Denote A as the adjacency matrix composed of a_{jk}'s. Let $D = \mathrm{diag}(d_1, \ldots, d_p)$, where $d_j = \sum_{k=1}^{p} |a_{jk}|$. Define $L = D - A$, which is semi-positive definite. Then

$$\boldsymbol{\gamma}_{-0}^\top (L \otimes \Theta)\boldsymbol{\gamma}_{-0} = \sum_{1\le j<k\le p} |a_{jk}|(\gamma_j - s_{jk}\gamma_k)^\top \Theta (\gamma_j - s_{jk}\gamma_k),$$

where $\boldsymbol{\gamma}_{-0} = (\gamma_1^\top, \ldots, \gamma_p^\top)^\top$. Obviously, $\boldsymbol{\gamma} = (\gamma_0^\top, \boldsymbol{\gamma}_{-0}^\top)^\top$. Note that

$$U_{im} = (X_{im}^{(0)} \boldsymbol{B}(t_{im})^\top, \ldots, X_{im}^{(p)} \boldsymbol{B}(t_{im})^\top)^\top = X_{im} \otimes \boldsymbol{B}(t_{im}).$$

Based on the discussions above, the objective function can be rewritten as

$$Q_n(\boldsymbol{\gamma}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{m=1}^{M_i} |Y_{im} - U_{im}^\top \boldsymbol{\gamma}| + \lambda_1 \sum_{k=0}^{p} w_k (\gamma_k^\top \Theta \gamma_k)^{1/2} + \frac{\lambda_2}{2} \boldsymbol{\gamma}_{-0}^\top (L \otimes \Theta) \boldsymbol{\gamma}_{-0}.$$

Denote the true value as $\boldsymbol{\beta}^*(t) = (\beta_0^*(t), \ldots, \beta_p^*(t))^\top$. Denote $\mathcal{A} = \{k : \beta_k^*(t) \neq 0, 0 \leq k \leq p\}$ as the set of important variables. Define $X_{im,\mathcal{I}}$ as the subvector of X_{im} indexed by $\mathcal{I} \subseteq \{0, 1, \ldots, p\}$. Let $\mathcal{I}^c$ and $|\mathcal{I}|$ denote the complement and cardinality of set $\mathcal{I}$, respectively. Let $\|\cdot\|_q$ be the L_q norm for vectors and $C^{(r)}([0,1])$ be the space of r-times continuously differentiable functions defined on $[0,1]$.

First consider the oracle estimator

$$\hat{\boldsymbol{\gamma}}^o = \arg\min_{\boldsymbol{\gamma}} Q_n(\boldsymbol{\gamma}), \quad \text{s.t. } \|\gamma_k\|_2 = 0, k \notin \mathcal{A}. \tag{4}$$

Then we can obtain $\widehat{\boldsymbol{\beta}}^o(t) = (\widehat{\beta}_0^o(t), \ldots, \widehat{\beta}_p^o(t))^\top$, where $\widehat{\beta}_k^o(t) = \boldsymbol{B}(t)^\top \widehat{\gamma}_k^o$. The following conditions are assumed.

(C1) There exists a positive integer M such that $\max_i M_i < M < \infty$.

(C2) The cardinality of $\mathcal{A}$ is fixed. For $k \in \mathcal{A}$, $\beta_k^*(\cdot) \in C^{(r)}([0,1])$.

(C3) Let $\Pi = N^{-1} K_n \sum_{i=1}^{N} \sum_{m=1}^{M_i} \{(X_{im,\mathcal{A}} X_{im,\mathcal{A}}^\top) \otimes (\boldsymbol{B}(t_{im}) \boldsymbol{B}(t_{im})^\top)\}$. There exist positive constants $D_1 < D_2$ such that $D_1 \leq \lambda_{\min}(\Pi) \leq \lambda_{\max}(\Pi) \leq D_2$, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalues, respectively. Moreover, $|X_{im}|_\infty < D_3$ with D_3 being a positive constant for $i = 1, \ldots, N$ and $m = 1, \ldots, M_i$.

(C4) Denote the density function and distribution function of ϵ_{im} conditional on (X_{im}, t_{im}) as $f_{im}(\cdot|X_{im}, t_{im})$ and $F_{im}(\cdot|X_{im}, t_{im})$, respectively. $F(0|X_{im}, t_{im}) = 0.5$. There exist positive constants c_1 and c_2 such that for any u satisfying $|u| < c_1$, $f_{im}(u|X_{im}, t_{im})$'s are uniformly bounded away from 0 and ∞, and

$$|F_{im}(u|X_{im}, t_{im}) - F_{im}(0|X_{im}, t_{im}) - u f_{im}(0|X_{im}, t_{im})| \leq c_2 u^2.$$

Similar conditions have been assumed in the literature (see, e.g. [17,18]). The following theorem establishes the consistency of $\widehat{\boldsymbol{\beta}}^o(t)$.

Theorem 2.1: *Assume that* (C1)–(C4) *hold. In addition,*

$$K_n \to \infty, \quad K_n \log(K_n)/N \to 0, \quad \lambda_1 \max_{k \in \mathcal{A}} w_k \to 0, \quad \textit{and} \quad \lambda_2 \to 0,$$

as $N \to \infty$. *Then* $\widehat{\boldsymbol{\beta}}^o(t)$ *satisfies*

(a) $1/n \sum_{i=1}^{N} \sum_{m=1}^{M_i} (\widehat{\beta}_k^o(t_{im}) - \beta_k^*(t_{im}))^2 = O_p\{K_n/N + K_n^{-2r} + \lambda_1^2 (\max_{k \in \mathcal{A}} w_k)^2\}$, *for* $k \in \mathcal{A}$;

(b) $\widehat{\beta}_k^o(t) = 0$, *for* $k \notin \mathcal{A}$.

The proof is given in Appendix. Hereafter, we set $L^* = \text{diag}\{1, L\}$ which is a $(p+1) \times (p+1)$ matrix, and

$$b(\lambda_1, \lambda_2) = \min_{k \in \mathcal{A}^c} \left\{ \lambda_1 w_k - 2\lambda_2 \sum_{k' \in \mathcal{A}} |L^*_{(k+1)(k'+1)}| \left\{ \int_0^1 \beta^{*2}_{k'}(t)\, \mathrm{d}t \right\}^{1/2} \right\},$$

where $L^*_{(k+1)(k'+1)}$ is the $(k+1, k'+1)$th element of L^*. The following additional conditions are needed.

(C5) $b(\lambda_1, \lambda_2) > 0, Nb(\lambda_1, \lambda_2) \to \infty$.
(C6) $\{N^{-1/2}K_n^{1/2} + K_n^{-r} + \lambda_1 \max_{k \in \mathcal{A}} w_k\}^{-1} \min_{k' \in \mathcal{A}} \int_0^1 \beta^{*2}_{k'}(t)\, \mathrm{d}t \to \infty$.

Conditions (C5) and (C6) are needed for variable selection consistency. Condition (C6) requires that the smallest signal does not decay too fast, which is similar to that in [19].

Theorem 2.2: *Assume that* (C1)–(C6) *and conditions in Theorem* 2.1 *hold. If* $\log p + \log K_n = o(Nb^2(\lambda_1, \lambda_2))$ *and* $N^{-1/2}K_n^{1/2} + K_n^{-r} + \lambda_1 \max_{k \in \mathcal{A}} w_k = o(b(\lambda_1, \lambda_2))$, *then with asymptotic probability one,* $\widehat{\gamma}^o$ *is the minimizer of the proposed penalized objective function* (3).

This theorem establishes that the proposed estimator enjoys the same asymptotic properties as the oracle estimator $\widehat{\gamma}^o$ in (4) with probability approaching one. It is worth noting that this result holds under high dimensions without restrictive conditions on the errors, which are commonly imposed in the literature. Proofs are presented in Appendix.

2.3. Computation

In a neighbourhood of an estimator $\gamma^{(s)}$, the objective function (3) can be approximated by

$$\begin{aligned} Q_n(\gamma; \gamma^{(s)}) = {} & \frac{1}{N} \sum_{i=1}^{N} \sum_{m=1}^{M_i} \frac{(Y_{im} - \sum_{k=0}^{p} \sum_{l=1}^{K_n} X_{im}^{(k)} B_l(t_{im}) \gamma_{kl})^2}{\varepsilon + |Y_{im} - \sum_{k=0}^{p} \sum_{l=1}^{K_n} X_{im}^{(k)} B_l(t_{im}) \gamma_{kl}^{(s)}|} \\ & + \lambda_1 \sum_{k=0}^{p} w_k (\gamma_k^\top \Theta \gamma_k)^{1/2} + \frac{\lambda_2}{2} \sum_{1 \le j < k \le p} |a_{jk}| (\gamma_j - s_{jk}\gamma_k)^\top \Theta (\gamma_j - s_{jk}\gamma_k), \end{aligned} \tag{5}$$

where ε is a small positive number to avoid zero in the denominator.

Define $U_{\cdot i} = (U_{i1}, \ldots, U_{iM_i})$ with $U_{im} = (X_{im}^{(0)} B(t_{im})^\top, \ldots, X_{im}^{(p)} B(t_{im})^\top)^\top$. Then we have $U = (U_{\cdot 1}, \ldots, U_{\cdot N})$. Similarly, let $Y = (Y_{11}, \ldots, Y_{1M_1}, \ldots, Y_{N1}, \ldots, Y_{1M_N})^\top$. Furthermore, define $\mathbb{Z}^{-1} = \text{diag}(|\epsilon^{(s)}| + \varepsilon)$, where $\epsilon^{(s)} = Y - U^\top \gamma^{(s)}$. Then function (5) can be rewritten as

$$Q_n(\gamma; \gamma^{(s)}) = \frac{1}{N} \left(Y - U^\top \gamma\right)^\top \mathbb{Z} \left(Y - U^\top \gamma\right) + \lambda_1 \sum_{k=0}^{p} w_k (\gamma_k^\top \Theta \gamma_k)^{1/2}$$

$$+\frac{\lambda_2}{2}\sum_{1\le j<k\le p}|a_{jk}|(\gamma_j-s_{jk}\gamma_k)^\top\Theta(\gamma_j-s_{jk}\gamma_k). \tag{6}$$

Let $\bar{\gamma}=I_{p+1}\otimes\Theta^{1/2}\gamma$ and $\bar{U}=I_{p+1}\otimes\Theta^{-1/2}U$. Then function (6) can be rewritten as

$$Q_n(\bar{\gamma};\gamma^{(s)})=\frac{1}{N}\left(Y-\bar{U}^\top\bar{\gamma}\right)^\top\mathbb{Z}\left(Y-\bar{U}^\top\bar{\gamma}\right)+\lambda_1\sum_{k=0}^{p}w_k(\bar{\gamma}_k^\top\bar{\gamma}_k)^{1/2}$$
$$+\frac{\lambda_2}{2}\sum_{1\le j<k\le p}|a_{jk}|(\bar{\gamma}_j-s_{jk}\bar{\gamma}_k)^\top(\bar{\gamma}_j-s_{jk}\bar{\gamma}_k). \tag{7}$$

Define $\check{L}=\mathrm{diag}(0,L)\otimes I_{K_n}$ with L defined in the above subsection. Then function (7) can be rewritten as

$$Q_n(\bar{\gamma};\gamma^{(s)})=\frac{1}{N}\left(\tilde{Y}-\tilde{U}^\top\bar{\gamma}\right)^\top\left(\tilde{Y}-\tilde{U}^\top\bar{\gamma}\right)+\lambda_1\sum_{k=0}^{p}w_k(\bar{\gamma}_k^\top\bar{\gamma}_k)^{1/2}, \tag{8}$$

where $\tilde{Y}=((\mathbb{Z}^{1/2}Y)^\top,\mathbf{0}^\top_{(p+1)K_n})^\top$ and $\tilde{U}=(\bar{U}\mathbb{Z}^{1/2},\sqrt{\lambda_2N/2}\check{L}^{1/2})$. This objective function can be solved using algorithms for group Lasso [20]. Overall, the proposed computational algorithm proceeds as follows.

Step 1 Initialize $\gamma^{(0)}$ and $s=0$.

Step 2 Given $\gamma^{(s)}$, for each k, update γ_k to $\gamma_k^{(s+1)}$ by optimizing (8). $s=s+1$.

Step 3 Repeat Step 3 until convergence.

In all our numerical examples, convergence is achieved with a small number of iterations. With fast convergence, the proposed algorithm is computationally much affordable. For example, for one simulation replicate with $p=500$ (more details described in the next section), the analysis takes 2.4 min on a laptop with standard configurations.

The proposed method involves two tuning parameters λ_1,λ_2, which have similar interpretations as in the existing penalization studies. We choose tuning parameters using V-fold cross-validation. This is computationally feasible as only simple updates are involved in the proposed algorithm.

3. Simulation

In the simulation, the data are generated from (1) with sample size $N=200$. The number of repeated measurements for each subject is $M_i=10$, for $i=1,\ldots,N$. The index variable t is generated from uniform[0, 1]. We consider two error distributions with (Error 1) $\mathcal{N}(0,1)$ and (Error 2) $0.7\,\mathcal{N}(0,1)+0.3$ Cauchy(0, 1). Here the normal error distribution can serve as a benchmark, and the Cauchy contamination has been commonly considered in the literature. It is noted that the contamination rate is considerably higher than in some existing studies. We set $p=50$ and 500, and generate covariates from multivariate normal distributions with marginal means 0 and variances 1. Covariates form clusters of size 10. Those in the same clusters are correlated, leading to the network structure, and different clusters are uncorrelated. The following specific examples are considered.

Example I: Covariates in the first three clusters have correlation coefficients ρ_1 at any time t. Those in the other clusters have an auto-regression correlation structure, with

covariates i and j having correlation coefficients $\rho_2^{|i-j|}$ at any time t. Set $\rho_1 = 0.7, 0.5$ and $\rho_2 = 0.3$, representing different levels of interconnections. For the time-varying coefficient functions, set $\beta_1(t) = \beta_2(t) = 0.06t, \beta_3(t) = \beta_4(t) = 0.08t, \beta_5(t) = \beta_6(t) = 0.12t, \beta_7(t) = \beta_8(t) = 0.16t, \beta_9(t) = \beta_{10}(t) = 0.20t$, and the others are zero.

Example II: Covariates in the first cluster have correlation coefficients 0.7, and those in the other clusters have correlation coefficients 0.5. Set $\beta_1(t) = 0.1\log(1+5t), \beta_2(t) = 0.1t + 0.06t^2, \beta_3(t) = 0.14\exp(t-0.1), \beta_4(t) = 0.08\cos(t)/(1+\sin(t)) + 0.12t, \beta_5(t) = 0.14\sin(\pi/2t) + 0.1t^2, \beta_6(t) = 0.08\sin(t) - 0.02\cos(\pi t) + 0.02, \beta_7(t) = -0.06t + 0.14(t+1)\log(1+t), \beta_8(t) = 0.24t, \beta_9(t) = -0.16\cos(2\pi/3(t-2.5)), \beta_{10}(t) = 0.1(t-1)^3 + 0.14t, \beta_{11}(t) = \beta_{12}(t) = 0.12t, \beta_{13}(t) = \beta_{14}(t) = 0.10t, \beta_{15}(t) = \beta_{16}(t) = 0.08t, \beta_{17}(t) = \beta_{18}(t) = 0.09t, \beta_{19}(t) = \beta_{20}(t) = 0.12t$, and the others are zero.

Example III: Covariates in the first two cluster have correlation coefficients 0.7, and those in the other clusters have correlation coefficients 0.5. Set $\beta_{21}(t) = 0.06t + 0.04\log(1+t), \beta_{22}(t) = 0.06t + 0.03\sin(\pi t/2), \beta_{23}(t) = 0.09t + 0.01t^2, \beta_{24}(t) = 0.09t + (0.05t - 0.03)^2, \beta_{25}(t) = 0.08t + 0.01\log(1+5t), \beta_{26}(t) = 0.08t + 0.02\exp(t-0.1), \beta_{27}(t) = \beta_{28}(t) = 0.10t, \beta_{29}(t) = \beta_{30}(t) = 0.12t$, and the others are zero.

Under Examples I and II, important variables have higher connectivity, satisfying the assumed model structure. Under Example III, important variables have lower connectivity. This example can test the performance of the proposed method under mis-specification.

To better gauge the performance of the proposed method, we also consider four direct competitors. The main characteristics of the five methods are summarized in Table 1. Method M4 also has a robust LAD loss. Only penalty P_1 is applied. Compared with this method the merit of accounting for network adjacency can be established. Methods M1 and M2 are parallel to M4 and M5, but with a non-robust LS (least-squared) loss. Compared with these two methods the merit of robustness can be established. In addition, we also consider method M3, which has an LAD robust loss with penalty $P_1^*(\cdot) = \lambda_1 \sum_{k=1}^{p} \|\beta_k(\cdot)\|$. Compared with this method the merit of accounting for network connectivity can be established. We acknowledge that there are other potential alternatives. The four alternatives we consider have analysis frameworks most comparable to the proposed.

To compare different methods, we consider AUC (area under the ROC curve). Using the AUC for comparison can effectively 'remove' the impact of tuning parameter selection, which can differ across methods. In addition, with tunings selected using cross-validation, we consider the integrated model error $\mathrm{ME} = \int (\hat{\beta}(t) - \beta^0(t))^\top (\hat{\beta}(t) - \beta^0(t))\,\mathrm{d}t$, where $\hat{\beta}(t)$ and $\beta^0(t)$ are the estimated and true functions, the numbers of true positive (TP) and false positive (FP).

Since the simulation results are qualitatively similar for $p = 50$ and $p = 500$, we present only the case with $p = 500$, while the results for $p = 50$ are relegated to Appendix 2. The

Table 1. Summary of all methods.

Method	Loss function	Penalty
M1	LS Loss	$P_1(\cdot)$
M2	LS Loss	$P_1(\cdot) + P_2(\cdot)$
M3	LAD Loss	$P_1^*(\cdot)$
M4	LAD Loss	$P_1(\cdot)$
M5	LAD Loss	$P_1(\cdot) + P_2(\cdot)$

Table 2. Results of Example I for $p = 500$.

	Error 1				Error 2			
Method	ME	TP	FP	AUC	ME	TP	FP	AUC
$\rho_1 = 0.7$								
M1	0.17	9.00	1.00	1.00	0.57	0.00	0.00	0.98
	(0.05)	(1.48)	(1.48)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
M2	0.08	10.00	1.00	1.00	0.57	2.00	1.00	0.98
	(0.01)	(0.00)	(1.48)	(0.00)	(0.10)	(2.97)	(1.48)	(0.01)
M3	0.37	9.00	2.50	0.90	0.40	8.00	2.00	0.89
	(0.14)	(1.48)	(2.22)	(0.07)	(0.14)	(1.48)	(2.22)	(0.05)
M4	0.38	9.00	0.00	1.00	0.42	9.00	1.00	1.00
	(0.14)	(1.48)	(0.00)	(0.00)	(0.16)	(1.48)	(1.48)	(0.00)
M5	0.08	10.00	1.00	1.00	0.09	10.00	1.00	1.00
	(0.01)	(0.00)	(1.48)	(0.00)	(0.02)	(0.00)	(1.48)	(0.00)
$\rho_1 = 0.5$								
M1	0.14	9.00	1.00	1.00	0.57	0.00	0.00	0.97
	(0.04)	(1.48)	(1.48)	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)
M2	0.09	10.00	2.00	1.00	0.57	0.00	0.00	0.97
	(0.02)	(0.00)	(1.48)	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)
M3	0.14	8.00	3.00	0.93	0.17	8.00	4.00	0.91
	(0.05)	(1.48)	(1.48)	(0.06)	(0.05)	(1.48)	(2.97)	(0.07)
M4	0.15	9.00	1.00	0.99	0.17	9.00	2.00	0.99
	(0.05)	(1.48)	(1.48)	(0.00)	(0.05)	(1.48)	(1.48)	(0.00)
M5	0.08	10.00	2.00	1.00	0.10	10.00	2.00	1.00
	(0.02)	(0.00)	(1.48)	(0.00)	(0.02)	(0.00)	(1.48)	(0.00)

Note: In each cell, median (median absolute deviation/0.6745).

results are summarized in Tables 2 and 3. Under Examples I and II, M5 has the best performance in both identification and estimation with the mixture error distribution and M2 has the best performance with the normal error distribution, as expected. Accounting for network adjacency improves the performance of the model. The accuracy of $\beta^0(t)$'s estimation is improved significantly from M1 to M2 and M4 to M5. For example, with $\rho_1 = 0.5$ and $p = 500$, the MEs of M4 and M5 are 0.17 and 0.10, respectively, with the mixture error distribution under Example I. Improvement in TP and FP is also observed. For example, with $p = 500$, the (TP, FP, AUC) of M4 and M5 are equal to (18, 4, 0.96) and (20, 2, 0.99), respectively, with the normal error distribution under Example II. Performances of non-robust M2 and M1 are inferior with the mixture error distribution, while the performances of M4 and M5 are still reasonable. For example, with $\rho_1 = 0.7$ and $p = 500$, the MEs of M1, M2, M4 and M5 are 0.57, 0.57, 0.42 and 0.09, respectively, with the mixture error distribution under Example I. Under Example III, which is not favourable for the proposed method, M5 is still observed to have competitive performance.

4. Real data analysis

To demonstrate the effectiveness of our proposed method in selecting the variables with time-varying effects as well as estimating the corresponding time-varying effects, in this section, we present a real data analysis based on China stock market. As we know, a large body of financial literature shows that the stock returns could be predictable in the long run [21–23]. However, the conclusions are quite mixed regarding the variables which can significantly predict stock returns, since various techniques, variables and different time periods of data were employed; therefore, different conclusions were derived by different

Table 3. Results of Examples II and III for $p = 500$.

	Error 1				Error 2			
Method	ME	TP	FP	AUC	ME	TP	FP	AUC
Example II								
M1	0.40	19.00	5.00	0.98	1.86	2.00	0.00	0.63
	(0.08)	(1.48)	(4.45)	(0.03)	(0.31)	(2.97)	(0.00)	(0.08)
M2	0.27	20.00	4.00	1.00	1.62	9.00	2.00	0.74
	(0.01)	(0.00)	(4.45)	(0.01)	(0.63)	(5.93)	(2.97)	(0.08)
M3	0.59	17.00	7.00	0.92	0.69	17.00	8.00	0.90
	(0.14)	(1.48)	(4.45)	(0.04)	(0.18)	(1.48)	(5.93)	(0.04)
M4	0.67	18.00	4.00	0.96	0.77	17.00	4.50	0.95
	(0.17)	(1.48)	(4.45)	(0.03)	(0.19)	(1.48)	(3.71)	(0.04)
M5	0.27	20.00	2.00	0.99	0.27	20.00	3.00	0.98
	(0.01)	(0.00)	(2.97)	(0.01)	(0.01)	(0.00)	(2.97)	(0.01)
Example III								
M1	0.16	9.00	7.00	0.98	0.36	0.00	0.00	0.47
	(0.04)	(1.48)	(4.45)	(0.03)	(0.00)	(0.00)	(0.00)	(0.11)
M2	0.01	10.00	0.00	1.00	0.36	0.00	0.00	0.55
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.17)
M3	0.13	9.00	4.00	0.97	0.16	8.00	3.00	0.92
	(0.04)	(1.48)	(4.45)	(0.05)	(0.05)	(1.48)	(2.22)	(0.05)
M4	0.16	9.00	7.00	0.96	0.19	8.00	6.00	0.91
	(0.05)	(1.48)	(2.97)	(0.05)	(0.06)	(1.48)	(2.97)	(0.06)
M5	0.01	10.00	0.00	0.99	0.01	10.00	0.00	0.98
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.02)

Note: In each cell, median (median absolute deviation/0.6745).

researchers. The empirical evidence from the literature in asset pricing demonstrates that most of the predictive models based on cross-section data may be unstable or even spurious. In particular, Welch and Goyal [23] shows that the predictability of these variables becomes weak in the late of the sample period, especially for the longitudinal dataset.

In this paper, we obtain the data from the Wind database, which is one of the most widely used and authoritative database in China. The sample period is from 2004 to 2013.

Table 4. Financial ratios.

Notation	Financial ratios	Notation	Financial ratios
X_1	Float A-Shares	X_{20}	Return on total assets
X_2	A-Shares total	X_{21}	Return on equity
X_3	Flow of equity	X_{22}	Net profit margin on sales
X_4	A-Shares of the total shares capital ratio	X_{23}	Main business ratio
X_5	Big shareholders shareholding ratio	X_{24}	Total asset cash recovery
X_6	Shareholding of the top 10 shareholders	X_{25}	Asset-liability ratio
X_7	Number of shareholders	X_{26}	Current assets ratio
X_8	Average annual turnover rate	X_{27}	Current liabilities ratio
X_9	Annual volume	X_{28}	Current ratio
X_{10}	Annual turnover	X_{29}	Quick ratio
X_{11}	Price/earn ratio	X_{30}	Cash ratio
X_{12}	Price/book value ratio	X_{31}	Long-term debt ratio
X_{13}	Price cash flow ratio	X_{32}	Current asset turnover rate
X_{14}	Price to sales ratio	X_{33}	Inventory turnover rate
X_{15}	Aggregate market value	X_{34}	Total asset turnover
X_{16}	Free float market capitalization	X_{35}	Accounts payable turnover rate
X_{17}	Net asset value per share	X_{36}	Total assets
X_{18}	Net operating cash flow per share	X_{37}	Dividends per share
X_{19}	Net cash flow per share		

The response variable is the yearly stock returns of listed firms in China, and the covariates include 37 firm financial ratios, which are displayed in Table 4. After removing missing measurements, we obtain 389 subjects and each has 10 observations from 2004 to 2013. All covariates are standardized before model fitting. The histogram given in Figure 1 shows that the stock returns have a long-tailed skewed distribution, which suggests that a robust model should be used here. In addition, Figure 2 gives the network correlation between the financial ratios. From this figure, we can see that there exist significant network modules among these variables, which suggests the necessity of considering interconnections among covariates when analysing the data. By analysing this data set, we aim to find the significant financial ratios and their time-varying effects.

To evaluate the performance of our method in variable selection and prediction, we randomly split the data into a training set, which includes two-thirds of the data, and a testing set, which includes the remaining one-third of the data. Here, we compare our proposed model (M5) with the other four alternatives (M1–M4). By computing the median of square

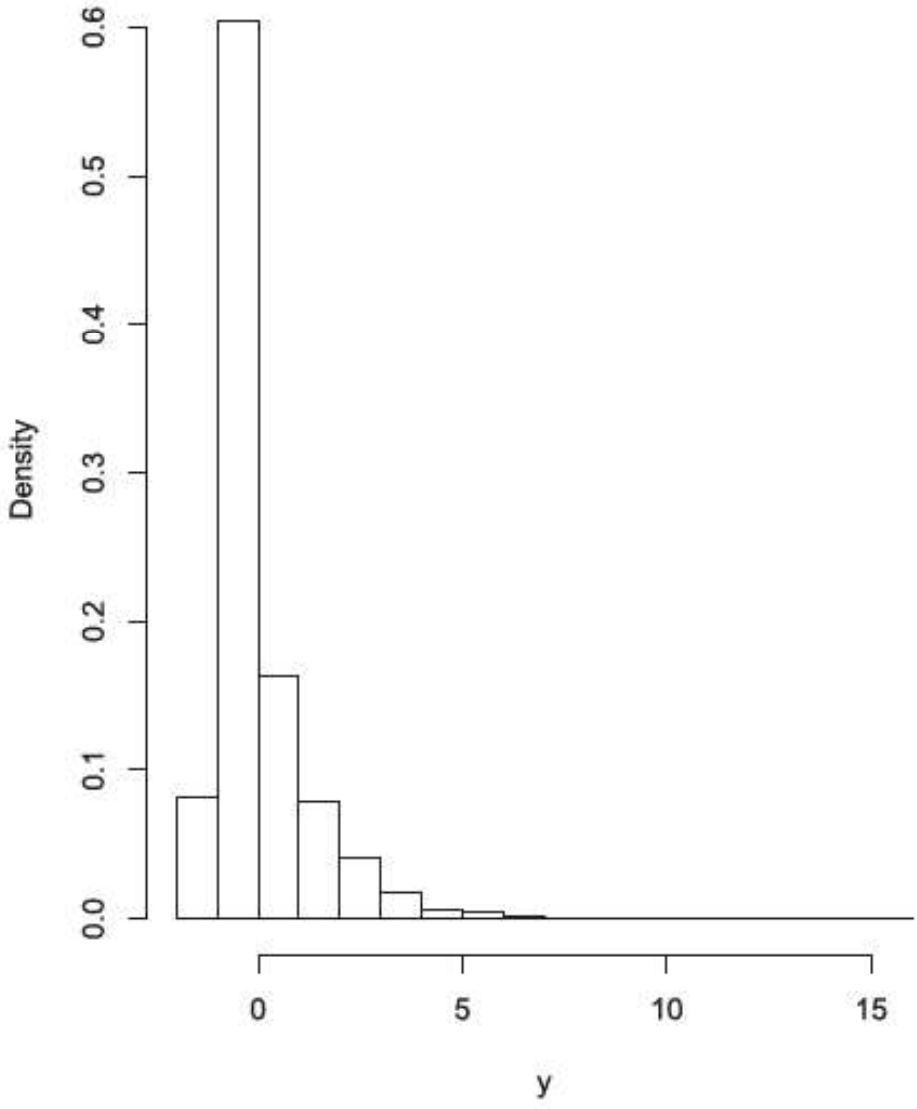

Figure 1. The histogram of stock returns.

error of 100 replications, the results are 0.869, 0.867, 0.861, 0.857 and 0.854 corresponding to M1, M2, M3, M4 and M5, respectively. It is easy to see that our proposed method has the smallest prediction error among all the competing methods as we expected. Moreover, Figure 3 provides the estimated coefficients of nine selected financial ratios using M5, that is, the float A-Shares, the A-Shares total, the A-Shares of the total shares capital ratio, the number of shareholders, the average annual turnover rate, the annual volume, the annual turnover, the aggregate market value, and the free-float market capitalization. More detailed results are available from the authors upon request. These results are consistent with much literature [24,25]. Meanwhile, Figure 3 also shows that there exist significant time-varying effects of financial ratios on the stock returns. We can also find such supporting evidence from [23] for the US stock market, which suggests that the predictability of factors varies with time. To complement the estimation and identification analysis, we also evaluate the stability of analysis by computing the observed occurrence index (OOI) [26]. For each variable selected using M5 based on the whole data, we compute its probability of being selected out of the 100 resampling and refer to this probability as the OOI. The median value of OOI is 0.94. Satisfactory stability provides certain support for this analysis.

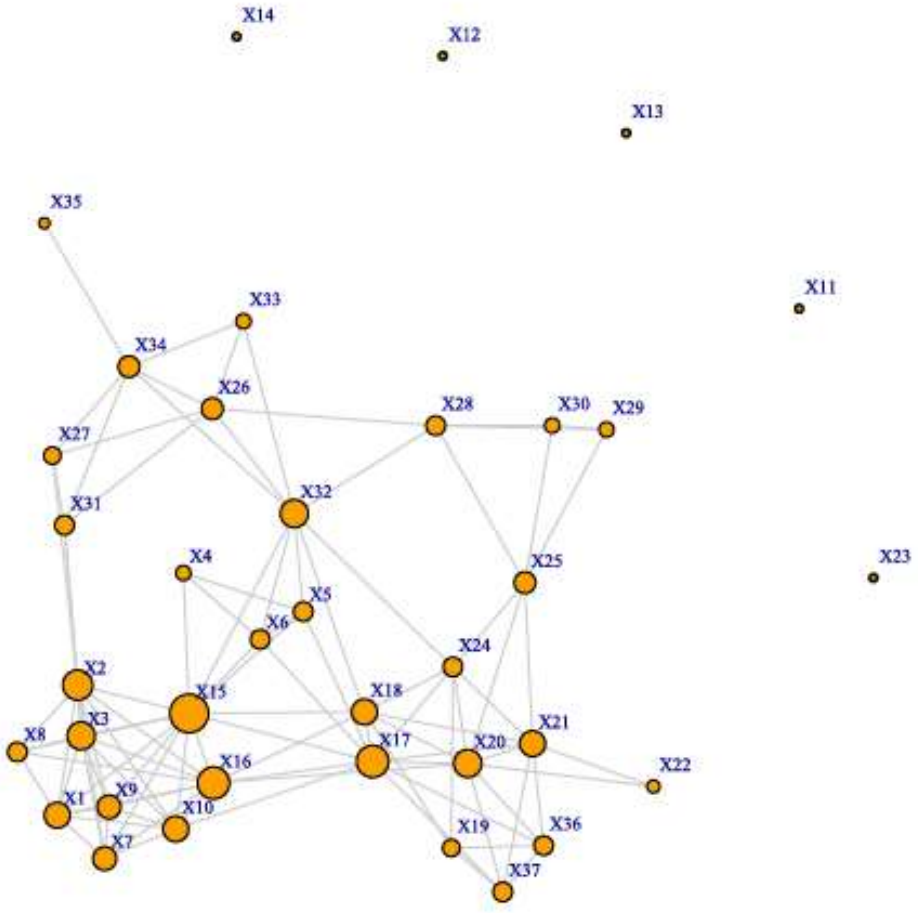

Figure 2. The network of the financial ratios: each node in the graph corresponds to a financial ratio. The absence of an edge between two nodes means the average value of the absolute correlation coefficients between them at all time points is less than 0.187 (determined using the Fisher transformation). The size of a node is proportional to its degree.

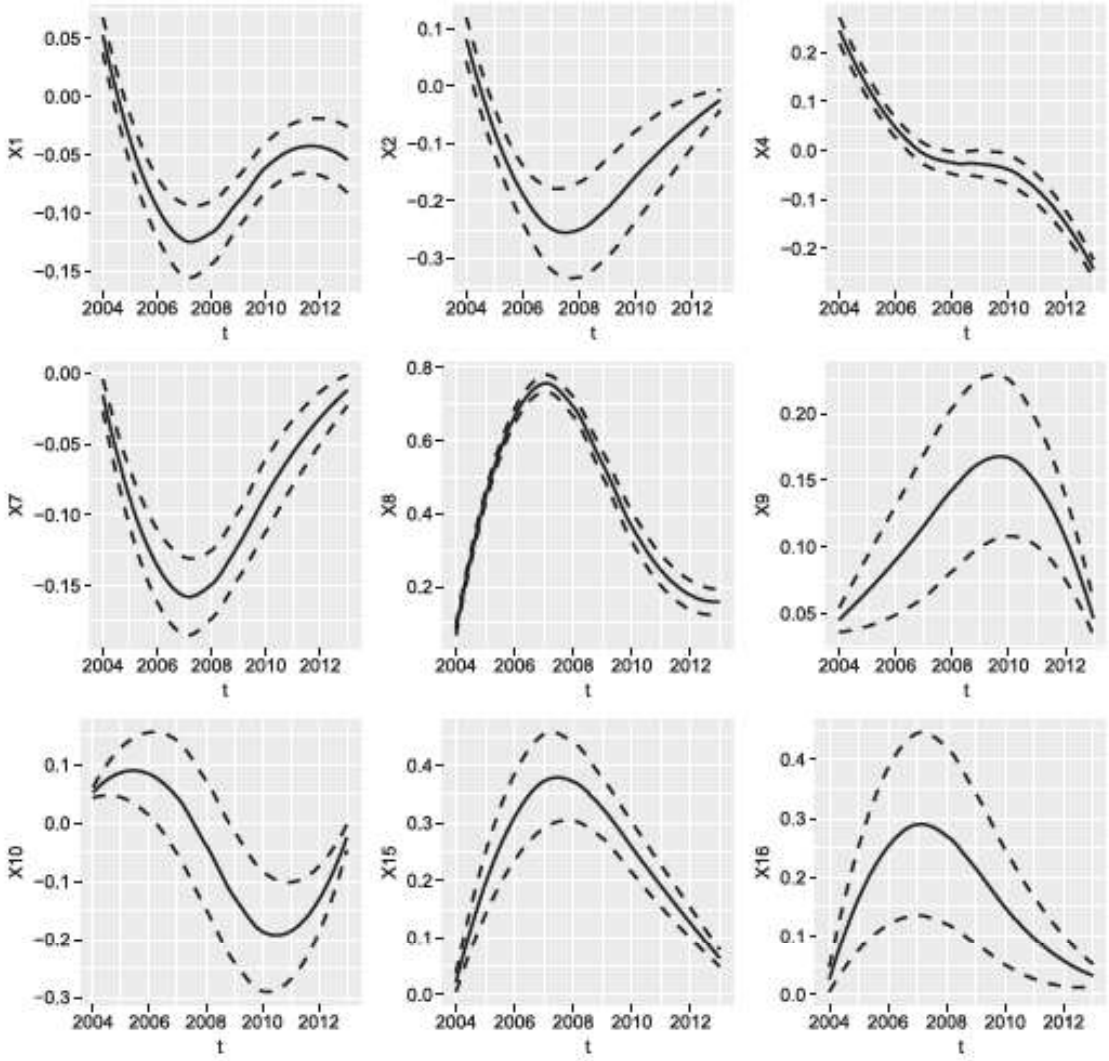

Figure 3. The estimated functions of coefficients (±1 pointwise standard deviation).

In summary, our proposed model is an effective method to select significant covariates and has better prediction performance than other alternative models. The proposed model is robust for long-tailed skewed distribution and the effects are time-varying.

5. Conclusion

Longitudinal data from long-tailed/contaminated distributions occur in many areas of modern science. In this study, we propose a network-adaptive robust penalization approach for time-varying coefficient models. The strategy of 'robust loss function + penalization' has been adopted to accommodate long-tailed distributions/contamination and select the relevant variables. An important contribution is that interconnections among covariates are described using a network structure and the most essential network properties of covariates are accommodated in estimation. Consistency properties are rigorously established. Numerical studies, including both simulations and data analysis, demonstrate the competitive practical performance of the proposed method.

There are several potential extensions to the method presented in this article. Beyond LAD-based loss, there are quite a few other robust techniques. It may be of interest to conduct time-varying coefficient models on other techniques. To accommodate the adjacency measures in the estimation, similarity in shapes instead of the values of coefficient functions of adjacent nodes can also be considered in further studies.

Acknowledgements

We thank the editor, AE, and reviewers for their insightful comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by the National Natural Science Foundation of China [11971404, 72071169], Humanity and Social Science Youth Foundation of Ministry of Education of China (20YJC910004), National Social Science Foundation of China (21 ZD146), NSF (1916251), the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China [2020030259].

References

[1] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13–22.
[2] Noh HS, Park BU. Sparse varying coefficient models for longitudinal data. Statist Sin. 2010;20:1183–1202.
[3] Huang JZ, Wu CO, Zhou L. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. Statist Sin. 2004;14:763–788.
[4] Cai Z, Li Q. Nonparametric estimation of varying coefficient dynamic panel data models. Econom Theory. 2008;24(5):1321–1342.
[5] Zang Y, Zhao Q, Zhang Q, et al. Inferring gene regulatory relationships with a high-dimensional robust approach. Genet Epidemiol. 2017;41(5):437–454.
[6] Akgiray V, Booth GG. The stable-law model of stock returns. J Bus Econom Stat. 1988;6:51–57.
[7] Hjalmarsson E. Predicting global stock returns. J Financ Quant. 2010;45(1):49–80.
[8] Wood SN, Pya N, Säfken B. Smoothing parameter and model selection for general smooth models. J Am Stat Assoc. 2016;111(516):1548–1563.
[9] Xue L, Qu A, Zhou J. Consistent model selection for marginal generalized additive model for correlated data. J Am Stat Assoc. 2010;105(492):1518–1530.
[10] Xue L, Qu A. Variable selection in high-dimensional varying-coefficient models with global optimality. J Mach Learn Res. 2012;13:1973–1998.
[11] Huang J, Ma S, Li H, et al. The sparse Laplacian shrinkage estimator for high-dimensional regression. Ann Stat. 2011;39(4):2021–2046.
[12] Ma S, Shi M, Li Y, et al. Incorporating gene co-expression network in identification of cancer prognosis markers. BMC Bioinform. 2010;11:271.
[13] Hayashi T, Yoshida N. On covariance estimation of non-synchronously observed diffusion processes. Bernoulli. 2005;11(2):359–379.
[14] Barndorff-Nielsen O, Hansen P, Lunde A. Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. J Econom. 2011;162:149–169.
[15] Wang L, Li H, Huang JZ. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. J Am Stat Assoc. 2008;103:1556–1569.

[16] Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. J Am Stat Assoc. 2011;106(494):544–557.
[17] Fan J, Fan Y, Barut E. Adaptive robust variable selection. Ann Stat. 2014;42:324–351.
[18] Lian H, Liang H. Generalized additive partial linear models with high-dimensional covariates. Econom Theory. 2013;29(6):1136–1161.
[19] Wang L, Wu Y, Li R. Quantile regression for analyzing heterogeneity in ultra-high dimension. J Am Stat Assoc. 2012;107(497):214–222.
[20] Yang Y, Zou H. A fast unified algorithm for solving group-lasso penalize learning problems. Stat Comput. 2015;25(6):1129–1141.
[21] Fama EF, French KR. Size, value, and momentum in international stock returns. J Financ Econ. 2012;105(3):457–472.
[22] Fama EF, French KR. Choosing factors. J Financ Econ. 2018;128(2):234–252.
[23] Welch I, Goyal A. A comprehensive look at the empirical performance of equity premium prediction. Rev Financ Stud. 2007;21(4):1455–1508.
[24] Amihud Y. Illiquidity and stock returns: cross-section and time-series effects. J Financ Mark. 2002;5(1):31–56.
[25] Chordia T, Swaminathan B. Trading volume and cross-autocorrelations in stock returns. J Finance. 2000;55(2):913–935.
[26] Huang J, Ma S. Variable selection in the accelerated failure time model via the bridge method. Lifetime Data Anal. 2010;16(2):176–195.
[27] Schumaker L. Spline functions: basic theory. Cambridge (UK): Cambridge University Press; 2007.
[28] Wang HJ, Zhu Z, Zhou J. Quantile regression in partially linear varying coefficient models. Ann Stat. 2009;37:3841–3866.
[29] Knight K. Limiting distributions for l_1 regression estimators under general conditions. Ann Stat. 1998;26(2):755–770.
[30] He X, Shi P. Convergence rate of b-spline estimators of nonparametric conditional quantile functions. J Nonparametr Stat. 1994;3(3–4):299–308.
[31] He X, Zhu ZY, Fung WK. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. Biometrika. 2002;89(3):579–590.
[32] Koenker R. Quantile regression. Cambridge (UK): Cambridge University Press; 2005.

Appendices

Appendix 1. Proofs of Theorems 2.1 and 2.2

Here we present the proofs for the results presented in Section 2.3. We first present some lemmas that are necessary to prove the theorems.

Lemma A.1: *Assume that Condition* (C2) *holds. Then there exists a spline approximation* $\boldsymbol{B}(t)^\top \gamma_k^*$ *to* $\beta_k^*(t)$ *for* $k \in \mathcal{A}$, *such that* $\beta_k^*(t) = \boldsymbol{B}(t)^\top \gamma_k^* - R_k(t)$ *and* $\sup_{t\in[0,1]} |R_k(t)| = O(K_n^{-r})$.

Lemma A.1 follows directly from Corollary 6.21 of [27]. Multiple published papers involving splines have also used this lemma in their theoretical investigations (see, e.g. [10,18,28]).

Proof of Theorem 2.1: Let $U_{im,\mathcal{S}} = X_{im,\mathcal{A}} \otimes \boldsymbol{B}(t_{im})$, where $\mathcal{S}$ represents the active set of U_{im}. Denote its corresponding coefficient as $\boldsymbol{\gamma}_{\mathcal{S}}$. In addition, let $\mathcal{A}_- = \mathcal{A} - \{0\}$. Similarly, we can define the corresponding active set $\mathcal{S}_-$ and regression coefficient $\boldsymbol{\gamma}_{\mathcal{S}_-}$. Objective function (3) in Section 2 is

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i} |Y_{im} - U_{im,\mathcal{S}}^\top \boldsymbol{\gamma}_{\mathcal{S}}| + \lambda_1 \sum_{k\in\mathcal{A}} w_k (\gamma_k^\top \Theta \gamma_k)^{1/2} + \frac{\lambda_2}{2} \boldsymbol{\gamma}_{\mathcal{S}_-}^\top (L_{\mathcal{A}_-\mathcal{A}_-} \otimes \Theta) \boldsymbol{\gamma}_{\mathcal{S}_-}.$$

Note that this function is strictly convex. Denote the minimizing solution as $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}$. Recall that $Y_{im} = X_{im,\mathcal{A}}^\top \boldsymbol{\beta}_{\mathcal{A}}^*(t_{im}) + \epsilon_{im}$. Set $\xi_k = \widehat{\gamma}_k - \gamma_k^*$ for $k \in \mathcal{A}$. Denote $R_{im} = (X_{im} \otimes \boldsymbol{B}(t_{im}))^\top \boldsymbol{\gamma}^* - X_{im}^\top \boldsymbol{\beta}^*(t_{im})$.

Then $|Y_{im} - U_{im,\mathcal{S}}^\top \widehat{\boldsymbol{\gamma}}_{\mathcal{S}}| = |\epsilon_{im} - U_{im,\mathcal{S}}^\top \boldsymbol{\xi}_{\mathcal{S}} - R_{im}|$. We have

$$\begin{aligned}\mathcal{O}_n(\boldsymbol{\xi}_{\mathcal{S}}) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i}\left\{|\epsilon_{im} - U_{im,\mathcal{S}}^\top \boldsymbol{\xi}_{\mathcal{S}} - R_{im}| - |\epsilon_{im} - R_{im}|\right\} \\ &\quad + \lambda_1 \sum_{k\in\mathcal{A}} w_k \left\{[(\gamma_k^* + \xi_k)^\top \Theta(\gamma_k^* + \xi_k)]^{1/2} - (\gamma_k^{*\top}\Theta\gamma_k^*)^{1/2}\right\} \\ &\quad + \frac{1}{2}\lambda_2 \left\{[\boldsymbol{\gamma}_{\mathcal{S}_-}^* + \boldsymbol{\xi}_{\mathcal{S}_-}]^\top (L_{\mathcal{A}_-\mathcal{A}_-}\otimes\Theta)[\boldsymbol{\gamma}_{\mathcal{S}_-}^* + \boldsymbol{\xi}_{\mathcal{S}_-}] - \boldsymbol{\gamma}_{\mathcal{S}_-}^{*\top}(L_{\mathcal{A}_-\mathcal{A}_-}\otimes\Theta)\boldsymbol{\gamma}_{\mathcal{S}_-}^*\right\} \\ &:= T_{n1} + T_{n2} + T_{n3}. \end{aligned} \tag{A1}$$

Obviously, $\mathcal{O}_n(\boldsymbol{\xi}_{\mathcal{S}}) \le \mathcal{O}_n(0) = 0$. Let $\|\boldsymbol{\xi}_{\mathcal{S}}\|_2 = \delta_n K_n^{1/2}$ with δ_n being a scalar.

For T_{n1}, we have

$$\begin{aligned}T_{n1} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i} E\left\{|\epsilon_{im} - U_{im,\mathcal{S}}^\top \boldsymbol{\xi}_{\mathcal{S}} - R_{im}| - |\epsilon_{im} - R_{im}|\right\} \\ &\quad + \frac{2}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i}[I(\epsilon_{im} < 0) - \frac{1}{2}]U_{im,\mathcal{S}}^\top \boldsymbol{\xi}_{\mathcal{S}} \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i}\left\{|\epsilon_{im} - U_{im,\mathcal{S}}^\top \boldsymbol{\xi}_{\mathcal{S}} - R_{im}| - |\epsilon_{im} - R_{im}| - 2\left[I(\epsilon_{im} < 0) - \frac{1}{2}\right]U_{im,\mathcal{S}}^\top \boldsymbol{\xi}_{\mathcal{S}}\right. \\ &\quad \left. - E\left[|\epsilon_{im} - U_{im,\mathcal{S}}^\top \boldsymbol{\xi}_{\mathcal{S}} - R_{im}| - |\epsilon_{im} - R_{im}|\right]\right\} \\ &:= T_{n1}^{[1]} + T_{n1}^{[2]} + T_{n1}^{[3]}. \end{aligned} \tag{A2}$$

By applying the identity [29]

$$|x - y| - |x| = 2y\left(I(x \le 0) - \frac{1}{2}\right) + 2\int_0^y [I(x \le z) - I(x \le 0)]\,\mathrm{d}z$$

and Condition (C3) for $T_{n1}^{[1]}$, we have

$$\begin{aligned}T_{n1}^{[1]} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i} E\left\{\left|\epsilon_{im} - U_{im,\mathcal{S}}^\top \boldsymbol{\xi}_{\mathcal{S}} - R_{im}\right| - |\epsilon_{im}|\right\} - E\left\{|\epsilon_{im} - R_{im}| - |\epsilon_{im}|\right\} \\ &= \frac{2}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i}\int_{R_{im}}^{U_{im,\mathcal{S}}^\top \boldsymbol{\xi}_{\mathcal{S}} + R_{im}} (F(z|X_{im}, t_{im}) - F(0|X_{im}, t_{im}))\,\mathrm{d}z \\ &= \frac{2}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i}\int_{R_{im}}^{U_{im,\mathcal{S}}^\top \boldsymbol{\xi}_{\mathcal{S}} + R_{im}} f(0|X_{im}, t_{im})z\,\mathrm{d}z \\ &\quad + \frac{2}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i}\int_{R_{im}}^{U_{im,\mathcal{S}}^\top \boldsymbol{\xi}_{\mathcal{S}} + R_{im}} (F(z|X_{im}, t_{im}) - F(0|X_{im}, t_{im}) - f(0|X_{im}, t_{im})z)\,\mathrm{d}z \\ &:= T_{n1}^{[1,1]} + T_{n1}^{[1,2]}. \end{aligned}$$

Note that $\|\boldsymbol{\xi}_{\mathcal{S}}\|_2 = \delta_n K_n^{1/2}$. By Condition (C1) and Lemma A.1, we have $\sup_{i,m}|R_{im}| = O(K_n^{-r})$. Combining Conditions (C3), (C4) and the Cauchy–Schwarz inequality, we have that there exists a

positive constant D_4,

$$T_{n1}^{[1,1]} = \boldsymbol{\xi}_{\mathcal{S}}^{\top}\frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i} f_{im}(0|X_{im},t_{im})U_{im,\mathcal{S}}U_{im,\mathcal{S}}^{\top}\boldsymbol{\xi}_{\mathcal{S}} + \frac{2}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i} f_{im}(0|X_{im},t_{im})R_{im}U_{im,\mathcal{S}}^{\top}\boldsymbol{\xi}_{\mathcal{S}}$$
$$\geq D_4\delta_n^2 - O(\delta_n K_n^{-r}).$$

Similarly, as $\delta_n \to 0$ and $K_n \to \infty$, we can derive that $T_{n1}^{[1,2]} = o(T_{n1}^{[1,1]})$ by Condition (C4) and some complicated calculations. Therefore, for a sufficiently large N,

$$T_{n1}^{[1]} \geq \frac{1}{2}D_4\delta_n^2 - |O(\delta_n K_n^{-r})|. \tag{A3}$$

Calculating the first and second moments of $T_{n1}^{[2]}$ shows that

$$T_{n1}^{[2]} = O_p(N^{-1/2}K_n^{1/2}\delta_n). \tag{A4}$$

Following arguments similar to those used in Lemma 3.2 of [30], we have

$$\sup_{\|\boldsymbol{\xi}_{\mathcal{S}}\|_2\leq 1} |T_{n1}^{[3]}| = o_p(N^{-1}K_n). \tag{A5}$$

Similar results can also be seen in [31]. Combining (A2), (A3), (A4), and (A5), we have that

$$T_{n1} > \frac{1}{2}D_4\delta_n^2 - |O_p(N^{-1/2}K_n^{1/2} + K_n^{-r})|\delta_n - |o_p(N^{-1}K_n)|. \tag{A6}$$

For T_{n2}, note that for $k \in \mathcal{A}$,

$$|[(\gamma_k^* + \xi_k)^{\top}\Theta(\gamma_k^* + \xi_k)]^{\frac{1}{2}} - (\gamma_k^{*\top}\Theta\gamma_k^*)^{\frac{1}{2}}| \leq (\xi_k^{\top}\Theta\xi_k)^{\frac{1}{2}}.$$

Then, we obtain that

$$T_{n2} \geq -|O(\lambda_1\delta_n \max_{k\in\mathcal{A}} w_k)|. \tag{A7}$$

As for T_{n3}, similar to the derivation of T_{n2}, together with $\lambda_2 \to 0$,

$$T_{n3} \geq -|O(\lambda_2\delta_n^2)| = -|o(1)|\delta_n^2. \tag{A8}$$

Therefore, combining (A1), (A6), (A7), and (A8), we can conclude that

$$0 \geq \mathcal{O}_n(\boldsymbol{\xi}_{\mathcal{S}}) \geq |O(1)|\delta_n^2 - |O(N^{-1/2}K_n^{1/2} + K_n^{-r} + \lambda_1 \max_{k\in\mathcal{A}} w_k)|\delta_n - |o(N^{-1}K_n)|$$

holds with probability approaching 1, which implies that $\delta_n = O_p(N^{-1/2}K_n^{1/2} + K_n^{-r} + \lambda_1 \max_{k\in\mathcal{A}} w_k)$. Theorem 2.1 is established. ■

Proof of Theorem 2.2: Recall that the proposed objective function is

$$Q_n(\boldsymbol{\gamma}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i}|Y_{im} - U_{im}^{\top}\boldsymbol{\gamma}| + \lambda_1\sum_{k=0}^{p} w_k(\gamma_k^{\top}\Theta\gamma_k)^{\frac{1}{2}} + \frac{\lambda_2}{2}\boldsymbol{\gamma}_{-0}^{\top}(L\otimes\Theta)\boldsymbol{\gamma}_{-0}.$$

By defining $\bar{\gamma}_k = W\gamma_k$ and $\bar{U}_{im} = \mathrm{diag}\{W^{-1},\ldots,W^{-1}\}U_{im}$ with $W = \Theta^{1/2}$, the above objective function reduces to

$$\widetilde{Q}_n(\bar{\boldsymbol{\gamma}}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M_i}|Y_{im} - \bar{U}_{im}^{\top}\bar{\boldsymbol{\gamma}}| + \lambda_1\sum_{k=0}^{p} w_k(\bar{\gamma}_k^{\top}\bar{\gamma}_k)^{\frac{1}{2}} + \frac{\lambda_2}{2}\bar{\boldsymbol{\gamma}}_{-0}^{\top}(L\otimes I)\bar{\boldsymbol{\gamma}}_{-0}.$$

Denote $\mathcal{I}_k$ as the set corresponding to $X^{(k)}$. Then $|\mathcal{I}_k| = K_n$. Following [19], we define the subgradient functions of $\sum_{i=1}^{N}\sum_{m=1}^{M_i}|Y_{im} - \bar{U}_{im}^{\top}\bar{\boldsymbol{\gamma}}|$ for $k \notin \mathcal{A}$:

$$G_k(\bar{\boldsymbol{\gamma}}) = \sum_{i=1}^{N}\sum_{m=1}^{M_i}\bar{U}_{im,\mathcal{I}_k}\left[I(Y_{im} - \bar{U}_{im}^{\top}\bar{\boldsymbol{\gamma}} \leq 0) - \frac{1}{2}\right] - \sum_{i=1}^{N}\sum_{m=1}^{M_i}\bar{U}_{im,\mathcal{I}_k}\left[v_{im} + \frac{1}{2}\right]$$

with $v_{im} = 0$ if $Y_{im} - U_{im}^{\top}\boldsymbol{\gamma} \neq 0$ and $v_{im} \in [-1, 1]$ otherwise.

Note that $G_k(\bar{\boldsymbol{\gamma}})$ is a vector of length K_n. All $G_k(\bar{\boldsymbol{\gamma}})$ with $k \notin \mathcal{A}$ make $G_{\mathcal{A}^c}(\bar{\boldsymbol{\gamma}})$, a vector of length $(p + 1 - |\mathcal{A}|)K_n$. Let $\widehat{\bar{\boldsymbol{\gamma}}}^o = (\widehat{\bar{\gamma}}_k^o)$ with $\widehat{\bar{\gamma}}_k^o = W\widehat{\gamma}_k^o$. Next consider the partial derivative of $\lambda_1 \sum_{k=1}^{p} w_k(\bar{\gamma}_k^{\top}\bar{\gamma}_k)^{1/2}$ with respect to $\bar{\gamma}_k$ at the oracle estimator $\widehat{\bar{\boldsymbol{\gamma}}}^o$. By some calculations, we have that for $j \notin \mathcal{A}$, the result is $\lambda_1 w_k[-1_{K_n}, 1_{K_n}]$. Recall that $L^* = \operatorname{diag}\{1, L\}$. For the term $\lambda_2/2\bar{\boldsymbol{\gamma}}_{-0}^{\top}(L \otimes I)\bar{\boldsymbol{\gamma}}_{-0}$, its derivative with respect to $\bar{\gamma}_k$ at the oracle estimator is $\lambda_2 \sum_{k' \in \mathcal{A}} L^*_{(k+1)(k'+1)}\Theta^{1/2}\widehat{\gamma}_{k'}^o$, where $L^*_{(k+1)(k'+1)}$ is the $(k+1, k'+1)$th element of L^*.

By the second-order sufficiency of the Karush–Kuhn–Tucher (KKT) conditions, to ensure that the minimizer is the oracle estimator, we need to check that the oracle estimator satisfies

$$\Omega^{\dagger} = \left\{ |G_{kl}(\widehat{\bar{\boldsymbol{\gamma}}}^o) + \lambda_2 N \sum_{k' \in \mathcal{A}} L^*_{(k+1)(k'+1)}\Theta_{l,}^{1/2}\hat{\gamma}_{k'}^o| < \lambda_1 w_k N,\ k \notin \mathcal{A},\ l = 1, \ldots, K_n \right\}.$$

Let $\Delta_n = N^{-1/2}K_n^{1/2} + K_n^{-r} + \lambda_1 \max_{k \in \mathcal{A}} w_k$. Note that $\|\Theta^{1/2}\gamma_k^*\|_2 \leq \{\int_0^1 \beta_{k'}^{*2}(t)\,\mathrm{d}t\}^{1/2} + |O(K_n^{-r})|$. By $\{\int_0^1 \beta_{k'}^{*2}(t)\,\mathrm{d}t\}^{1/2} \gg K_n^{-r}$ for $k' \in \mathcal{A}$ in Condition (C6) and $\|\widehat{\boldsymbol{\gamma}}^o - \boldsymbol{\gamma}^*\|_2 \leq K_n O_p(\Delta_n)$ in Theorem 2.1, we have

$$\begin{aligned}\left|\sum_{k' \in \mathcal{A}} L^*_{(k+1)(k'+1)}\Theta_{l,}^{1/2}\hat{\gamma}_{k'}^o\right| &\leq \left|\sum_{k' \in \mathcal{A}} L^*_{(k+1)(k'+1)}\Theta_{l,}^{1/2}(\hat{\gamma}_{k'}^o - \gamma_{k'}^*)\right| + \left|\sum_{k' \in \mathcal{A}} L^*_{(k+1)(k'+1)}\Theta_{l,}^{1/2}\gamma_{k'}^*\right| \\ &\leq |O_p(\Delta_n)| + \sum_{k' \in \mathcal{A}} |L^*_{(k+1)(k'+1)}|\left[\left\{\int_0^1 \beta_{k'}^{*2}(t)\,\mathrm{d}t\right\}^{1/2} + |O(K_n^{-r})|\right].\end{aligned}$$

Owing to the fact that $\Delta_n^{-1}\min_{k' \in \mathcal{A}} \int_0^1 \beta_{k'}^{*2}(t)\,\mathrm{d}t \to \infty$,

$$\left|\sum_{k' \in \mathcal{A}} L^*_{(k+1)(k'+1)}\Theta_{l,}^{1/2}\widehat{\gamma}_{k'}^o\right| \leq 2\sum_{k' \in \mathcal{A}} |L^*_{(k+1)(k'+1)}|\left\{\int_0^1 \beta_{k'}^{*2}(t)\,\mathrm{d}t\right\}^{1/2}$$

with asymptotic probability one. Recall that

$$b(\lambda_1, \lambda_2) = \min_{k \in \mathcal{A}^c}\left\{\lambda_1 w_k - 2\lambda_2 \sum_{k' \in \mathcal{A}} |L^*_{(k+1)(k'+1)}|\left\{\int_0^1 \beta_{k'}^{*2}(t)\,\mathrm{d}t\right\}^{1/2}\right\}.$$

Then it suffices to prove that the oracle estimator satisfies $\Omega = \{\|G_{\mathcal{A}^c}(\widehat{\bar{\boldsymbol{\gamma}}}^o)\|_{\infty} < Nb(\lambda_1, \lambda_2)\}$.

In fact,

$$\begin{aligned}G_{\mathcal{A}^c}(\widehat{\bar{\boldsymbol{\gamma}}}^o) &= \sum_{i=1}^{N}\sum_{m=1}^{M_i} \bar{U}_{im,\mathcal{S}^c}\left[I(Y_{im} - \bar{U}_{im}^{\top}\widehat{\bar{\boldsymbol{\gamma}}}^o \leq 0) - \frac{1}{2}\right] - \sum_{i=1}^{N}\sum_{m=1}^{M_i} \bar{U}_{im,\mathcal{S}^c}\left[v_{im} + \frac{1}{2}\right] \\ &= \sum_{i=1}^{N}\sum_{m=1}^{M_i} \bar{U}_{im,\mathcal{S}^c}\left[I(Y_{im} - U_{im}^{\top}\boldsymbol{\gamma}^* \leq 0) - \frac{1}{2}\right] \\ &\quad - \sum_{i=1}^{N}\sum_{m=1}^{M_i} \bar{U}_{im,\mathcal{S}^c}\left[v_{im} + \frac{1}{2}\right] \\ &\quad + \sum_{i=1}^{N}\sum_{m=1}^{M_i} \bar{U}_{im,\mathcal{S}^c}\left[\Pr(Y_{im} - \bar{U}_{im}^{\top}\widehat{\bar{\boldsymbol{\gamma}}}^o \leq 0) - \Pr(Y_{im} - U_{im}^{\top}\boldsymbol{\gamma}^* \leq 0)\right] \\ &\quad + \sum_{i=1}^{N}\sum_{m=1}^{M_i} \bar{U}_{im,\mathcal{S}^c}\left[I(Y_{im} - \bar{U}_{im}^{\top}\widehat{\bar{\boldsymbol{\gamma}}}^o \leq 0) - I(Y_{im} - U_{im}^{\top}\boldsymbol{\gamma}^* \leq 0)\right.\end{aligned}$$

$$- \Pr(Y_{im} - \tilde{U}_{im}^{\top}\widehat{\tilde{\boldsymbol{\gamma}}}^{o} \le 0) + \Pr(Y_{im} - U_{im}^{\top}\boldsymbol{\gamma}^{*} \le 0)\Big]$$
$$\doteq H_{\mathcal{A}^c}^{[1]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o}) + H_{\mathcal{A}^c}^{[2]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o}) + H_{\mathcal{A}^c}^{[3]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o}) + H_{\mathcal{A}^c}^{[4]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o}). \tag{A9}$$

Therefore, $\|G_{\mathcal{A}^c}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o})\|_{\infty} \le \|H_{\mathcal{A}^c}^{[1]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o})\|_{\infty} + \|H_{\mathcal{A}^c}^{[2]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o})\|_{\infty} + \|H_{\mathcal{A}^c}^{[3]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o})\|_{\infty} + \|H_{\mathcal{A}^c}^{[4]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o})\|_{\infty}$, which yields that

$$\Omega \supseteq \Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4 \quad \text{with } \Omega_q = \{\|H_{\mathcal{A}^c}^{[q]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o})\|_{\infty} < Nb(\lambda_1,\lambda_2)/4, q = 1,2,3,4\}. \tag{A10}$$

For $H_{\mathcal{A}^c}^{[1]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o})$, we can see that

$$H_{\mathcal{A}^c}^{[1]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o}) = \sum_{i=1}^{N}\sum_{m=1}^{M_i} \tilde{U}_{im,\mathcal{S}^c}\Big[I(Y_{im} - U_{im}^{\top}\boldsymbol{\gamma}^{*} \le 0) - I(\epsilon_{im} \le 0) - \Pr(Y_{im} - U_{im}^{\top}\boldsymbol{\gamma}^{*} \le 0)$$
$$+ \Pr(\epsilon_{im} \le 0)] + \sum_{i=1}^{N}\sum_{m=1}^{M_i} \tilde{U}_{im,\mathcal{S}^c}\left[I(\epsilon_{im} \le 0) - \frac{1}{2}\right]$$
$$+ \sum_{i=1}^{N}\sum_{m=1}^{M_i} \tilde{U}_{im,\mathcal{S}^c}[\Pr(Y_{im} - U_{im}^{\top}\boldsymbol{\gamma}^{*} \le 0) - \Pr(\epsilon_{im} \le 0)]. \tag{A11}$$

Note that $\varpi_{im} \doteq I(Y_{im} - U_{im}^{\top}\boldsymbol{\gamma}^{*} \le 0) - I(\epsilon_{im} \le 0) - \Pr(Y_{im} - U_{im}^{\top}\boldsymbol{\gamma}^{*} \le 0) + \Pr(\epsilon_{im} \le 0)$ are mean-zero random variables and $Var(\varpi_i^m) = O(K_n^{-r})$. Applying Bernstein's inequality, for any given $\ell \in \mathcal{S}^c$,

$$\Pr\left(\left|\sum_{i=1}^{N}\sum_{m=1}^{M_i} \tilde{U}_{im,\ell}\varpi_{im}\right| > Nb(\lambda_1,\lambda_2)/12\right) \le \exp\left(\frac{N^2b^2(\lambda_1,\lambda_2)/144}{12NK_n^{-r} + CNb(\lambda_1,\lambda_2)}\right).$$

With the condition that $K_n^{-r} = o(b(\lambda_1,\lambda_2))$, we have

$$\Pr\left(\left|\sum_{i=1}^{N}\sum_{m=1}^{M_i} \tilde{U}_{im,\ell}\varpi_{im}\right| > Nb(\lambda_1,\lambda_2)/12\right) \le \exp\left(-CNb(\lambda_1,\lambda_2)\right). \tag{A12}$$

For the term $\sum_{i=1}^{N}\sum_{m=1}^{M_i} \tilde{U}_{im,\ell}[I(\epsilon_{im} \le 0) - 1/2]$, with Hoeffding's inequality, it can be shown that

$$\Pr\left(\left|\sum_{i=1}^{N}\sum_{m=1}^{M_i} \tilde{U}_{im,\ell}\left[I(\epsilon_{im} \le 0) - \frac{1}{2}\right]\right| > Nb(\lambda_1,\lambda_2)/12\right) \le \exp\left(-CNb^2(\lambda_1,\lambda_2)\right). \tag{A13}$$

Moreover,

$$\left|\sum_{i=1}^{N}\sum_{m=1}^{M_i} \tilde{U}_{im,\ell}[\Pr(Y_{im} - U_{im}^{\top}\boldsymbol{\gamma}^{*} \le 0) - \Pr(\epsilon_{im} \le 0)]\right| = O(NK_n^{-r}) = o(Nb(\lambda_1,\lambda_2)). \tag{A14}$$

Combining (A11), (A12), (A13), and (A14), it yields

$$\Pr\{\|H_{\mathcal{A}^c}^{[1]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o})\|_{\infty} < Nb(\lambda_1,\lambda_2)/4\} \le \exp\left(-CNb^2(\lambda_1,\lambda_2) + \log p + \log K_n\right) \to 1. \tag{A15}$$

For $H_{\mathcal{A}^c}^{[2]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o})$, following Section 2.2 of [32], with probability one, there exist exactly $|\mathcal{A}|K_n$ samples that satisfy $Y_{im} - U_{im}^{\top}\widehat{\tilde{\boldsymbol{\gamma}}}^{o} = 0$. Therefore, with probability one,

$$\|H_{\mathcal{A}^c}^{[2]}(\widehat{\tilde{\boldsymbol{\gamma}}}^{o})\|_{\infty} = O(|\mathcal{A}|K_n) = o(Nb(\lambda_1,\lambda_2)). \tag{A16}$$

Table A1. Results of Example I for $p = 50$.

	Error 1				Error 2			
Method	ME	TP	FP	AUC	ME	TP	FP	AUC
$\rho_1 = 0.7$								
M1	0.19	9.00	3.00	0.97	0.57	0.00	0.00	0.73
	(0.07)	(0.00)	(2.97)	(0.04)	(0.00)	(0.00)	(0.00)	(0.07)
M2	0.09	10.00	3.00	1.00	0.57	1.00	1.00	0.77
	(0.02)	(0.00)	(2.97)	(0.00)	(0.16)	(1.48)	(1.48)	(0.07)
M3	0.33	9.00	2.00	0.93	0.35	9.00	2.00	0.91
	(0.10)	(1.48)	(1.48)	(0.07)	(0.12)	(1.48)	(1.48)	(0.05)
M4	0.35	9.00	1.00	0.97	0.36	9.00	1.00	0.97
	(0.11)	(1.48)	(1.48)	(0.03)	(0.11)	(1.48)	(1.48)	(0.03)
M5	0.08	10.00	2.00	1.00	0.08	10.00	2.00	0.99
	(0.01)	(0.00)	(1.48)	(0.01)	(0.01)	(0.00)	(1.48)	(0.01)
$\rho_1 = 0.5$								
M1	0.13	9.00	1.00	0.99	0.57	0.00	0.00	0.72
	(0.04)	(1.48)	(1.48)	(0.02)	(0.00)	(0.00)	(0.00)	(0.07)
M2	0.08	10.00	1.00	1.00	0.57	0.00	0.00	0.73
	(0.03)	(0.00)	(1.48)	(0.01)	(0.00)	(0.00)	(0.00)	(0.08)
M3	0.14	8.00	1.00	0.92	0.16	9.00	2.00	0.92
	(0.04)	(1.48)	(1.48)	(0.07)	(0.06)	(1.48)	(1.48)	(0.06)
M4	0.14	9.00	0.00	0.97	0.16	9.00	1.00	0.96
	(0.04)	(1.48)	(0.00)	(0.04)	(0.06)	(1.48)	(1.48)	(0.03)
M5	0.08	10.00	1.00	0.99	0.08	10.00	2.00	0.98
	(0.02)	(0.00)	(1.48)	(0.01)	(0.02)	(0.00)	(2.97)	(0.02)

Note: In each cell, median (median absolute deviation/0.6745).

Note that for any $\ell \in \mathcal{S}^c$

$$\sum_{i=1}^{N}\sum_{m=1}^{M_i} \bar{U}_{im,\mathcal{S}^c}\left[\Pr(Y_{im} - \bar{U}_{im}^{\top}\widehat{\bar{\boldsymbol{\gamma}}}^{o} \le 0) - \Pr(Y_{im} - U_{im}^{\top}\boldsymbol{\gamma}^{*} \le 0)\right]$$
$$\asymp \sum_{i=1}^{N}\sum_{m=1}^{M_i} \|\bar{U}_{im,\ell}\|_2 \|\widehat{\bar{\boldsymbol{\gamma}}}^{o} - \bar{\boldsymbol{\gamma}}^{*}\|_2 = O_p(N\Delta_n).$$

For $H^{[3]}_{\mathcal{A}^c}(\widehat{\bar{\boldsymbol{\gamma}}}^{o})$, by condition $\Delta_n = o(b(\lambda_1, \lambda_2))$, we have

$$\Pr\{\|H^{[3]}_{\mathcal{A}^c}(\widehat{\bar{\boldsymbol{\gamma}}}^{o})\|_\infty < Nb(\lambda_1, \lambda_2)/4\} \to 1. \tag{A17}$$

Following similar techniques as in Lemma A.3 of [19], we can obtain

$$\Pr\{\|H^{[4]}_{\mathcal{A}^c}(\widehat{\bar{\boldsymbol{\gamma}}}^{o})\|_\infty < Nb(\lambda_1, \lambda_2)/4\} \to 1. \tag{A18}$$

Combining (A10), (A15), (A16), (A17) and (A18), we can conclude that $\Pr\{\Omega\} \to 1$ as $N \to \infty$. These results complete the proof. ■

Appendix 2. Additional simulation results for $p = 50$

This section presents simulation results for $p = 50$.

Table A2. Results of Examples II and III for $p = 50$.

	Error 1				Error 2			
Method	ME	TP	FP	AUC	ME	TP	FP	AUC
Example II								
M1	0.40	19.00	1.00	0.98	2.27	2.50	0.00	0.57
	(0.08)	(1.48)	(1.48)	(0.03)	(0.71)	(2.97)	(0.00)	(0.09)
M2	0.26	20.00	0.00	1.00	1.36	10.00	1.00	0.67
	(0.01)	(0.00)	(0.00)	(0.00)	(0.64)	(8.90)	(1.48)	(0.09)
M3	0.51	18.00	2.00	0.94	0.60	18.00	2.00	0.93
	(0.10)	(1.48)	(1.48)	(0.04)	(0.15)	(1.48)	(1.48)	(0.04)
M4	0.58	18.00	1.00	0.96	0.67	18.00	1.00	0.96
	(0.11)	(1.48)	(1.48)	(0.04)	(0.17)	(1.48)	(1.48)	(0.04)
M5	0.27	20.00	0.00	1.00	0.27	20.00	0.00	1.00
	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)
Example III								
M1	0.16	9.00	7.00	0.98	0.36	0.00	0.00	0.47
	(0.04)	(1.48)	(4.45)	(0.03)	(0.00)	(0.00)	(0.00)	(0.11)
M2	0.01	10.00	0.00	1.00	0.36	0.00	0.00	0.55
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.17)
M3	0.13	9.00	4.00	0.97	0.16	8.00	3.00	0.92
	(0.04)	(1.48)	(4.45)	(0.05)	(0.05)	(1.48)	(2.22)	(0.05)
M4	0.16	9.00	7.00	0.96	0.19	8.00	6.00	0.91
	(0.05)	(1.48)	(2.97)	(0.05)	(0.06)	(1.48)	(2.97)	(0.06)
M5	0.01	10.00	0.00	0.99	0.01	10.00	0.00	0.98
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.02)

Note: In each cell, median (median absolute deviation/0.6745).