Journal of Machine Learning Research 22 (2021) 1-42 Submitted 12/19; Revised 12/20; Published 1/21

Homogeneity Structure Learning in Large-scale Panel Data with Heavy-tailed Errors

Di Xiao DI.XIAO@UGA.EDU
Department of Statistics
University of Georgia
Athens, GA 30602, USA

Yuan Ke YUAN.KE@UGA.EDU
Department of Statistics
University of Georgia
Athens, GA 30602, USA

Runze Li RIL4@PSU.EDU
Department of Statistics
The Pennsylvania State University
University Park, PA 16802, USA

Editor: Jie Peng

Abstract

Large-scale panel data is ubiquitous in many modern data science applications. Conventional panel data analysis methods fail to address the new challenges, like individual impacts of covariates, endogeneity, embedded low-dimensional structure, and heavy-tailed errors, arising from the innovation of data collection platforms on which applications operate. In response to these challenges, this paper studies large-scale panel data with an interactive effects model. This model takes into account the individual impacts of covariates on each spatial node and removes the exogenous condition by allowing latent factors to affect both covariates and errors. Besides, we waive the sub-Gaussian assumption and allow the errors to be heavy-tailed. Further, we propose a data-driven procedure to learn a parsimonious yet flexible homogeneity structure embedded in high-dimensional individual impacts of covariates. The homogeneity structure assumes that there exists a partition of regression coefficients where the coefficients are the same within each group but different between the groups. The homogeneity structure is flexible as it contains many widely assumed low-dimensional structures (sparsity, global impact, etc.) as its special cases. Non-asymptotic properties are established to justify the proposed learning procedure. Extensive numerical experiments demonstrate the advantage of the proposed learning procedure over conventional methods especially when the data are generated from heavy-tailed distributions.

Keywords: interactive effects, robust estimation, factor model, Huber's loss, change-points detection

1. Introduction

Panel data analysis has been one of the most exciting subjects in statistics and econometrics. The possibility of modeling multi-dimensional data with cross-sectional dependence and serial dynamics has led to a remarkable proliferation of applications in diversified fields,

including biology, economics, epidemiology, finance, and social science. We refer to Anderson and Hsiao (1982), Chamberlain and Rothschild (1983), Hsiao (1986), Arellano (2003), Hsiao (2007), Kneip et al. (2012), and many more representative literature therein. Let y_{it} be a univariate response variable and $\mathbf{X}_{it}$ be a p-dimensional centered covariate, a fixed-effects model for panel data analysis would be

$$y_{it} = \alpha_i + \mathbf{X}_{it}^{\mathrm{T}}\beta + e_{it}, \quad i = 1, \ \ldots, \ N, \ \ t = 1, \ \ldots, \ T, \tag{1}$$

where $\alpha_i \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ are unknown parameters to be estimated, and

$$E(e_{it}|\mathbf{X}_{it}) = 0, \quad i = 1, \ \ldots, \ N, \ \ t = 1, \ \ldots, \ T. \tag{2}$$

This fixed effects model has been extensively studied in methodological and empirical literature (e.g., Nickell, 1981; Bhargava et al., 1982; Judson and Owen, 1999; Lee and Yu, 2010; Bell and Jones, 2015).

In the era of big data, we benefit from the escalation of data availability. Meanwhile, many cutting-edge challenges in panel data analysis arise from the innovation of data collection platforms on which applications operate. For instance, sensor network has become an increasingly important data collection method for various applications like air pollution monitoring, climate study, energy consumption, earthquake detection and so on. A sensor network system can automatically collect, process, and transfer multiple time-series data from a huge number of spatially distributed nodes. To model such a panel data, the condition (2) is no longer suitable as there may exist some latent factors, which influence the covariate $\mathbf{X}_{it}$ as well as the error e_{it}. Besides, assuming β to be the same across $i = 1, \ldots, N$ ignores the individual attribute at each node and hence may lead to a model misspecificaiton. To account for the interactive effects caused by the latent factors and the individual attribute of the impact, we consider the following panel data model with interactive effects

$$\begin{cases} y_{it} &= \alpha_i + \mathbf{X}_{it}^{\mathrm{T}}\beta_i + \mathbf{f}_t^{\mathrm{T}}\lambda_i + \varepsilon_{it}, \\ \mathbf{X}_{it} &= \mathbf{B}_i\mathbf{f}_t + \mathbf{u}_{it}, \end{cases} \quad i = 1, \ \ldots, \ N, \ \ t = 1, \ \ldots, \ T, \tag{3}$$

where $\beta_i = (\beta_{i1}, \ldots, \beta_{ip}) \in \mathbb{R}^p$ are the unknown parameters of individual attributes; $\alpha_i \in \mathbb{R}$, $\lambda_i \in \mathbb{R}^q$ and $\mathbf{B}_i \in \mathbb{R}^{p\times q}$ are treated as nuisance unknown parameters; $\mathbf{f}_t \in \mathbb{R}^q$ are latent factor; $\varepsilon_{it} \in \mathbb{R}$ and $\mathbf{u}_{it} \in \mathbb{R}^p$ are random errors. In this paper, we assume $\{\mathbf{f}_t\}_{t=1}^T$, $\{\varepsilon_{it}\}_{t=1}^T$ and $\{\mathbf{u}_{it}\}_{t=1}^T$ are independent sequences, and allow p to diverge with condition $p+q+1 < T$. To keep the presentation concise, we present the theoretical results for serial independent/weakly dependent scenarios in the main document of this paper. We defer the theoretical results for strongly serial dependent scenarios to the supplemental material.

Although (3) takes into account the individual attributes $\{\beta_i\}_{i=1}^n$, it involves too many unknown parameters, which may miss the inherent low-dimensional structure among β_{ij}'s. In some modern data science applications, learning the low-dimensional structure embedded in large scale panel data has become the primary objective over the coefficient estimation and inference. For example, social media users of websites such as Twitter and Facebook generate unprecedented amounts of data on a wide range of topics (politics, sports, entertainment, etc.) on daily basis (e.g., Lerman and Ghosh, 2010; Abel et al., 2011). Furthermore, it is common for social media data to contain geographical location information, so the data is inherently a large scale panel data. A hot topic in business analytic is to

cluster the social media users across the topics and geological locations into sub-groups, such that further precise business actions can be applied to each group. To this end, we impose a parsimonious yet flexible homogeneity structure among β_{ij}'s,

$$\beta_{ij} = \begin{cases} \beta_{0,1} & \text{when } (i,j) \in A_1, \\ \beta_{0,2} & \text{when } (i,j) \in A_2, \\ \vdots & \vdots \\ \beta_{0,K+1} & \text{when } (i,j) \in A_{K+1}, \end{cases} \tag{4}$$

where K is unknown and $\{A_k : 1 \le k \le K+1\}$ is an unknown partition of $\mathcal{I} := \{(i,j) : 1 \le i \le N;\ 1 \le j \le p\}$. Notice that the global attribute assumption (i.e., $\beta_1 = \cdots = \beta_N = \beta$) and the sparsity assumption (i.e., $\beta_{ij} = 0$ when $(i,j) \in \mathcal{S}$ for some $\mathcal{S} \subset \mathcal{I}$) can be considered as two special cases of the homogeneity structure (4). Due to its flexibility, the homogeneity structure and some alternatives have been studied by Ke et al. (2015), Su et al. (2016), and Su and Ju (2018), among others. These studies mainly follow the penalized regression approach which put a penalty on sequential differences between the initial estimator of coefficients. Hence, the penalized regression approach is sensitive to the correctness of the order of the initial estimators and may not perform well in practice when the data exhibits one or more of the following features: (a) the partition is heavily imbalanced; (b) the error is heavy-tailed; (c) the signal jump between two partitions slowly converges to zero when the sample size diverges. Besides penalized approaches, there are literature that study the latent group structure in panel data model with other procedures, see Ke et al. (2016), Xu et al. (2020), Ke et al. (2020), and references therein.

Large-scale panel data also challenges existing methodologies by driving researchers out of the comfort zone. Some commonly assumed conditions, like Gaussianity (or sub-Gaussianity), may no longer be realistic in the big data regime. Indeed, heavy-tailed panel data are widely encountered in many areas including genetics, economics, and finance. Even if the Gaussian assumption holds on the population level, one may observe spurious outliers due to the large cross-sectional size of N. Over the past two decades, the flourishing of large-scale macroeconomic panel data has motivated new developments in econometric panel data analysis (e.g, Stock and Watson, 2002; Ludvigson and Ng, 2009). Consider a macroeconomic panel data set consisting of 131 time-series which are widely used to describe the macroeconomic activities in United States[1]. We calculate the sample excess kurtosis of each time-series in the panel data to assess their tail behaviors. Figure 1 shows that most time-series have positive excess kurtoses which means their tails are heavier than Gaussian distribution. Besides, there are 43 time-series whose excess kurtoses are greater than 6. This indicates that their tails are heavier than t-distribution with degrees of freedom 5 which is a heavy-tailed distribution. In this paper, we propose to relax the sub-Gaussian assumption in panel data analysis. In particular, we allow ε_{it} and $\mathbf{u}_{it}$ in (3) to follow a wide range of distributions including heavy-tailed ones with only finite moment conditions. Recently, robust covariance matrix estimation and robust factor analysis have drawn huge attentions. We refer to Pison et al. (2003), Avella-Medina et al. (2018), Fan et al. (2019a), and Ke et al. (2019), among many others.

1. A detailed description of this panel data can be found in Appendix A of Ludvigson and Ng (2009).

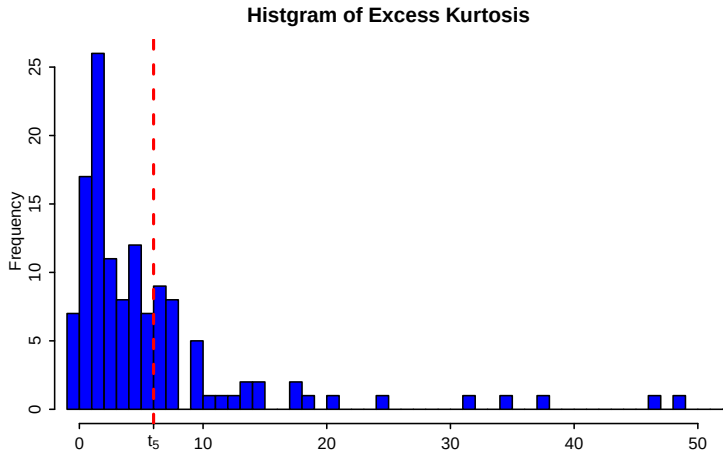

Figure 1: Histogram of excess kurtosis of 131 macroeconomic variables.

In response to the challenges discussed above, this paper studies the large-scale panel data with the interactive effects model (3) and allows the errors ε_{it} and $\mathbf{u}_{it}$ to be heavy-tailed. Learning the homogeneity structure (4) consists of three objectives: (i) estimate the number of homogeneity groups K; (ii) estimate the partition $\{A_k : 1 \leq k \leq K+1\}$; and (iii) estimate the homogeneity coefficients $\{\beta_{0,k} : 1 \leq k \leq K+1\}$. We gradually unveil the learning procedure in four steps. In the first step, we show key insights into the robust estimation of (3) through an oracle scenario that assumes the latent factors are known. In the second step, we consider a robust estimator for the covariance matrix of covariates. Then, we propose to recover the latent factors by applying eigen-decomposition to the robustly estimated covariance matrix. By plugging the estimators of latent factors back into the first step, we obtain a robust initial estimator of coefficients in (3). In the third step, we pursue the first two objectives in the homogeneity structure learning by detecting the change-points among the initial estimator of coefficients. The change-points detection process is carried out by wild binary segmentation (Fryzlewicz, 2014). In the final step, we estimate the homogeneity coefficients based upon the recovered partitions.

1.1 Our Contributions

This paper studies large-scale panel data and addresses some challenges that arise in modern applications. We model large-scale panel data with an interactive effects model where both covariates and errors are influenced by some latent factors. Besides, the response variable and covariates are allowed to be heavy-tailed. Instead of assuming a global attribute that may lead to model misspecification or individual attributes that create too many free parameters, we propose to learn a parsimonious yet flexible homogeneity structure in coefficients. The homogeneity structure assumes that there exists an unobservable partition such that coefficients are the same within each group but diverse between groups. With the limited restriction on the number and size of groups, homogeneity structure is a generalization of many widely assumed low-dimensional structures such as sparsity, grouping, and tree. We propose a data-driven procedure to robustly estimate the interactive effects model and

learn the homogeneity structure. The robustness is achieved by replacing the L_2 loss with the Huber's loss as the latter down weights outliers. The homogeneity structure is learned by detecting multiple change-points among initially estimated coefficients. Theoretically, we have shown the proposed procedure achieves the non-asymptotic robustness in the sense that the resulting estimators admit exponential-type concentration bounds with low-order finite moment conditions. Moreover, the resulting estimators are asymptotically unbiased estimates for the parameters of interest. Numerically, the proposed robust homogeneity structure learning procedure is proved to be able to improve the interpretability as well as prediction accuracy in various empirical scenarios.

1.2 Organization of the Paper

The rest of the paper is organized as follows. In Section 2, we introduce the estimation procedure of the panel data model with interactive effects and heavy-tailed errors. In Section 3, we study the homogeneity structure learning procedure. In Section 4, we summarize the proposed learning procedure by a computation algorithm and introduce a fast robust covariance matrix estimation method. In Section 5, we assess the finite sample performance of the proposed learning procedure with simulated experiments. In Section 6, we analyze an air quality panel data collected by a large out-door monitor network in the United States. The proofs of theoretical results are presented in Appendices.

1.3 Notations

We adopt the following notations throughout the paper. Let $\mathbf{A} = (A_{k\ell})_{1\le k,\ell\le p}$ be a $p \times p$ matrix. We write $\|\mathbf{A}\|_{\max} = \max_{1\le k,\ell\le p} |A_{k\ell}|$, $\|\mathbf{A}\|_\infty = \max_{1\le k\le p} \sum_{\ell=1}^p |A_{k\ell}|$ and $\|\mathbf{A}\|_{\mathrm{F}} = \big(\sum_{k=1}^p \sum_{\ell=1}^p |A_{k\ell}|^2\big)^{1/2}$. When $\mathbf{A}$ is symmetric, we have $\|\mathbf{A}\|_2 = \max_{1\le k\le p} |\lambda_k(\mathbf{A})|$, where $\lambda_1(\mathbf{A}) \ge \lambda_2(\mathbf{A}) \ge \cdots \ge \lambda_p(\mathbf{A})$ are the eigenvalues of $\mathbf{A}$. Further, we use $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ to denote the maximum and minimum eigenvalues of $\mathbf{A}$, respectively.

2. Robust Panel Data Analysis

In this section, we gradually unveil a robust estimation procedure for the panel data model with interactive effects and heavy-tailed errors.

2.1 An Oracle Estimator with Observable Factors

To begin with, we introduce the robust estimation procedure of the interactive effects model (3), through an oracle scenario such that the latent factors are assumed to be observable. When $\mathbf{f}_t$'s are observable, model (3) can be re-formulated as a linear regression problem

$$y_{it} = \alpha_i + \mathbf{X}_{it}^{\mathrm{T}}\beta_i + \mathbf{f}_t^{\mathrm{T}}\lambda_i + \epsilon_{it} := \mathbf{W}_{it}^{\mathrm{T}}\theta_i + \varepsilon_{it}, \quad i = 1,\ \ldots,\ N,\ t = 1,\ \ldots,\ T, \tag{5}$$

where $d = 1 + p + q$, $\mathbf{W}_{it} = (1, \mathbf{X}_{it}^{\mathrm{T}}, \mathbf{f}_t^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^d$, and $\theta_i = (\alpha_i, \beta_i^{\mathrm{T}}, \lambda_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^d$. The ordinary least squares (OLS) estimator of θ_i immediately follows

$$\widehat{\theta}_i^{OLS} = \Big(\sum_{t=1}^T \mathbf{W}_{it}\mathbf{W}_{it}^{\mathrm{T}}\Big)^{-1} \sum_{t=1}^T \mathbf{W}_{it} y_{it}.$$

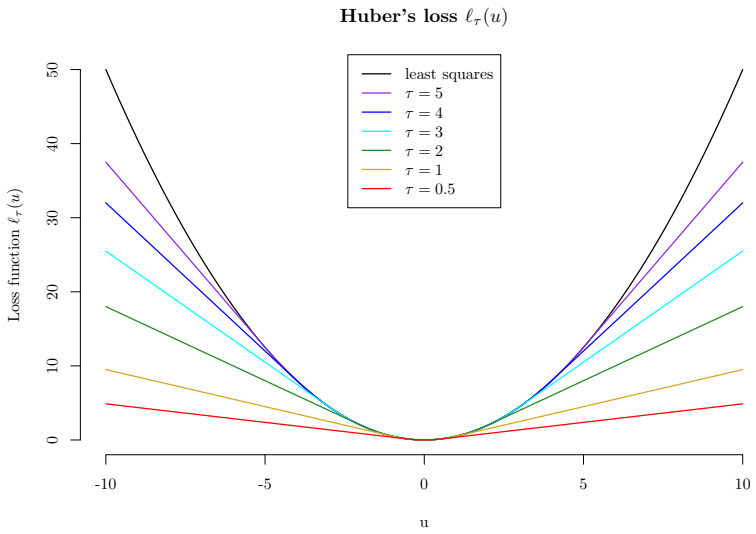

Figure 2: Huber's loss function for various choices of the tuning parameter τ. The least-squares (ℓ_τ with $\tau = \infty$)-loss is also shown for comparison.

However, the OLS estimator is not robust against outliers and/or heavy-tailed errors. This effect is amplified by high dimensionality. Even if we assume the errors follow moderate-tailed distributions, we may observe large outliers by chance which makes the OLS estimator naturally a non-robust estimator for large-scale panel data. Some recent studies (e.g., Catoni, 2012; Fan et al., 2017; Avella-Medina et al., 2018; Minsker, 2018; Sun et al., 2019) have tackled this issue by revisiting Huber's wisdom (Huber, 1984).

We introduce the Huber's loss in Definition 1 below. The parameter τ controls the shape as well as the robustness of the Huber's loss. When $\tau \to \infty$, the Huber's loss approaches the L_2 loss that leads to the least-squares estimator. On the other hand, when $\tau \to 0$, the Huber's loss approaches the L_1 loss (after proper normalization), which corresponds to the least absolute deviation (LAD) estimator. The LAD estimator is robust against outliers but can be biased when the distribution is asymmetric. Figure 2 portrays the shape of the Huber's loss with different values of τ.

Definition 1 *The Huber's loss $\ell_\tau(x)$ is defined as*

$$\ell_\tau(x) = \begin{cases} \frac{1}{2}x^2, & \text{if } |x| \le \tau, \\ \tau|x| - \frac{1}{2}\tau^2, & \text{if } |x| > \tau, \end{cases}$$

where τ is a robustification parameter that trades bias for robustness.

We define the robust estimator of θ_i through the following convex optimizaiton problem:

$$\widehat{\theta}_i(\tau, \mathbf{f}) = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{t=1}^{T} \ell_\tau(y_{it} - \mathbf{W}_{it}^{\mathrm{T}}\theta), \quad \text{for} \quad i = 1, \ \ldots, \ N, \tag{6}$$

where $(\tau, \mathbf{f})$ emphasizes the estimator depends on the choice of τ and the observation of $\mathbf{f}$.

Condition 1 *For $i = 1, \ldots, N$,*

(a) $\mathbb{E}(\epsilon_{it}|\mathbf{W}_{it}) = 0$ *and* $v_\delta = \max_i \mathbb{E}(|\epsilon_{it}|^{1+\delta})$ *is finite for some* $\delta > 1$.

(b) The empirical Gram matrix $\mathbf{S}_i = T^{-1}\sum_{t=1}^{T} \mathbf{W}_{it}\mathbf{W}_{it}^{\mathrm{T}}$ *satisfies* $\min_i \lambda_{\min}(\mathbf{S}_i) > c_l$, *for some positive constant* c_l.

The Condition 1 (a) waives the sub-Gaussian condition in conventional panel data analysis literature. Instead, we allow the errors to be heavy-tailed with finite $(1+\delta)$th moment for some $\delta > 1$. This condition is slightly stronger than a finite variance condition. The Condition 1 (b) requires the smallest eigenvalue of $\mathbf{S}_i$ to be uniformly lower bounded away from zero.

Theorem 1 *Assume the Condition 1 holds. Then, for any $s > 0$ and choosing $\tau \geq v_\delta(T/s)^{1/2}$, with probability at least $1 - N(2d+1)e^{-s}$, the estimator $\widehat{\theta}_i(\tau, \mathbf{f})$ satisfies*

$$\max_{1\leq i\leq N} \|\widehat{\theta}_i(\tau, \mathbf{f}) - \theta_i\|_2 \leq \frac{4\tau d s}{c_l T}, \tag{7}$$

as long as $T \geq 32d^2 s$,

Theorem 1 provides a uniform non-asymptotic upper bound for the estimation accuracy of the oracle estimator $\widehat{\theta}_i(\tau, \mathbf{f})$. If we choose $\tau = c_\tau T^{1/2}/s$ for some constant $c_\tau > 0$. When $T/d^2 \to \infty$ as $T \to \infty$, the upper bound (7) implies that $\widehat{\theta}_i(\tau, \mathbf{f})$ converges to θ_i at a rate approximately equals to $T^{-1/2}$. In the next subsection, we introduce the estimation procedure for the latent factor $\mathbf{f}_t$. Once the estimator $\widehat{\mathbf{f}}_t$ is available, we can plug it in (5) to obtain the estimator $\widehat{\theta}_i(\tau, \widehat{\mathbf{f}})$.

Remark 1 *The robust estimation for multiple linear regressions has been a key component in many studies. He et al. (2004) considered a robust estimator of linear regression for longitudinal data by maximizing the marginal likelihood of scaled t-type error distribution. She and Owen (2011) studied the multiple outliers detection problems from the penalized regressions point of view. More recently, Zhou et al. (2018) and Fan et al. (2019a) proposed factor-adjusted robust multiple testing procedures for large-scale multiple testing with correlated and heavy-tailed data. Similar theoretical discussions can also be found in high-dimensional sparse linear regression and large covariance matrix estimation with heavy-tailed data. The results in Theorem 2.1 are comparable to the multiple mean regression results in Fan et al. (2019a).*

2.2 Estimate Latent Factors

Denote $\mathbf{Z}_t = N^{-1}\sum_{i=1}^{N} \mathbf{X}_{it} = (Z_{t1}, \ldots, Z_{tp})^{\mathrm{T}}$, $\mathbf{B} = N^{-1}\sum_{i=1}^{N} \mathbf{B}_i$ and $\bar{\mathbf{u}}_t = N^{-1}\sum_{i=1}^{N} \mathbf{u}_{it}$. We propose to estimate the latent factors through an averaged latent factor model

$$\mathbf{Z}_t = \mathbf{B}\mathbf{f}_t + \bar{\mathbf{u}}_t, \quad t = 1, \ldots, T. \tag{8}$$

To make the model (8) identifiable, we impose the following identification conditions

$$\mathrm{cov}(\mathbf{f}_t) = \mathbf{I}_q \quad \text{and} \quad \mathbf{B}^{\mathrm{T}}\mathbf{B} \text{ is diagonal.}$$

We assume that the following condition to hold for the factor model (8).

Condition 2 *In the factor model* (8)*, we assume the latent factor* $\{\mathbf{f}_t\}_{t=1}^T$ *and the idiosyncratic noise* $\{\bar{\mathbf{u}}_t\}_{t=1}^T$ *are two i.i.d. sequences and independent with each other. Denote the covariance matrix of* $\mathbf{Z}_t$ *and* $\bar{\mathbf{u}}_t$ *as* $\mathbf{\Sigma}_Z$ *and* $\mathbf{\Sigma}_u$*, respectively. Let* $\lambda_1 \geq \cdots \geq \lambda_p$ *be the eigenvalues of* $\mathbf{\Sigma}_Z$ *in the descending order and* $\mathbf{v}_1, \ldots, \mathbf{v}_p$ *be the corresponding eigenvectors. Moreover,*

(a) (Finite kurtosis) $\max_{1\leq t\leq T,\ 1\leq \ell\leq p} \kappa_{t\ell} \leq c_1$*, where* c_1 *is a positive constant and* $\kappa_{t\ell}$ *is the kurtosis of* $Z_{t\ell}$*, for* $t = 1, \ldots, T$ *and* $\ell = 1, \ldots, p$*;*

(b) (Pervasiveness) There exist positive constants c_2*,* c_3 *and* c_4*, such that* $c_2 p \leq \lambda_\ell - \lambda_{\ell+1} \leq c_3 p$ *for* $\ell = 1, \ldots, q$*, and* $\|\mathbf{\Sigma}_u\|_2 \leq \lambda_{q+1} \leq c_4$*.*

Condition 2 (a) requires $\mathbf{Z}_t$, and hence $\bar{\mathbf{u}}_t$, to have finite fourth moments. This condition is much weaker than requiring finite fourth moments of $\mathbf{u}_{it}$'s. It allows $\mathbf{u}_{it}$'s to be strongly dependent w.r.t. i and can be checked by calculating the empirical kurtosis of $\mathbf{Z}_t$. Condition 2 (b) assumes that the first q eigenvalues of $\mathbf{\Sigma}_Z$ are much larger than the rest $p-q$ ones when the dimensionality p is large. This pervasiveness assumption is widely used in high-dimensional factor model literature (e.g., Johnstone and Lu, 2009; Fan et al., 2013; Shen et al., 2016; Wang and Fan, 2017) to identify the low-rank part from the idiosyncratic errors. Recently, literature (e.g., Fan et al., 2018b; Abbe et al., 2020) studied weaker versions of the pervasiveness assumption that allows the eigen-gap between λ_ℓ and $\lambda_{\ell+1}$, for $\ell = 1, \ldots, q$, to diverge slower than order $O(p)$. Our theoretical results can be extended to adapt the weaker version of pervasiveness assumption.

Next, we illustrate the estimation procedure of latent factors by three steps.

Step 1: Estimate $\mathbf{\Sigma}_Z$

Denote $\mathbf{\Sigma}_Z = (\sigma_{k\ell})_{1\leq k,\ell\leq p}$ and $\mathrm{sign}(x)$ the sign function of x. Define $\psi_\tau(\cdot)$ the first order derivative of Huber's loss $\ell_\tau(\cdot)$, which admits the following form

$$\psi_\tau(x) = \begin{cases} x, & \text{if } |x| \leq \tau, \\ \tau\,\mathrm{sign}(x), & \text{if } |x| > \tau. \end{cases}$$

Then, we define the element-wise estimator of $\sigma_{k\ell}$ as

$$\widehat{\sigma}_{k\ell} = \frac{2}{T(T-1)} \sum_{1\leq i<j\leq T} \psi_{\tau_{k\ell}}\left(\frac{(Z_{ik} - Z_{jk})(Z_{i\ell} - Z_{j\ell})}{2}\right), \quad 1 \leq k, \ell \leq p, \tag{9}$$

where $\tau_{k\ell}$'s are robustification parameters satisfying $\tau_{k\ell} = \tau_{\ell k}$. By definition, it is easy to see that $\widehat{\sigma}_{\ell k} = \widehat{\sigma}_{k\ell}$.

Collecting these element-wise estimators, we obtain the robust covariance estimator

$$\widehat{\mathbf{\Sigma}}_Z = \widehat{\mathbf{\Sigma}}_Z(\mathbf{\Gamma}) = (\widehat{\sigma}_{k\ell})_{1\leq k,\ell\leq p}, \tag{10}$$

where $\mathbf{\Gamma} = (\tau_{k\ell})_{1\le k,\ell\le p}$ is a symmetric matrix of robustification parameters.

To avoid trivial discussion, we assume $T \ge 2,\ p \ge 1$ and define $T_0 = \lfloor T/2 \rfloor$, the largest integer no greater than $T/2$. Let $\mathbf{V} = (v_{k\ell})_{1\le k,\ell\le p}$ be a symmetric $p \times p$ matrix with

$$v_{k\ell}^2 = \mathbb{E}((Z_{1k} - Z_{2k})(Z_{1\ell} - Z_{2\ell}))^2/4.$$

Theorem 2 *Under Condition 2 (a) and for any $0 < \delta < 1$, the covariance estimator $\widehat{\mathbf{\Sigma}}_Z = \widehat{\mathbf{\Sigma}}_Z(\mathbf{\Gamma})$ given in (10) with*

$$\mathbf{\Gamma} = \sqrt{T_0/(2\log p + \log \delta^{-1})}\,\mathbf{V}, \tag{11}$$

satisfies

$$\|\widehat{\mathbf{\Sigma}}_Z - \mathbf{\Sigma}_Z\|_{\max} \le 2\|\mathbf{V}\|_{\max}\sqrt{\frac{2\log p + \log \delta^{-1}}{T_0}}, \tag{12}$$

with probability at least $1 - 2\delta$.

Theorem 2 shows that each element of $\widehat{\mathbf{\Sigma}}_Z$ concentrates around the truth as the maximum error scales as $\sqrt{2\log(p)/T_0} \approx \sqrt{4\log(p)/T}$. Therefore, we can accurately estimate $\mathbf{\Sigma}_Z$ at a high confidence level under the condition that $\log(p)/T$ is small.

Remark 2 *Recently, estimating large scale covariance matrices from heavy-tailed data or data contaminated by outliers has become a hot topic, see Catoni (2016); Minsker (2018); Minsker and Wei (2020); Avella-Medina et al. (2018); Mendelson and Zhivotovskiy (2020); Ke et al. (2019) and references therein. Catoni (2016) proposed a robust estimator of the Gram and covariance matrices of a random vector from a spectrum-wise perspective and proved error bounds under the operator norm. Mendelson and Zhivotovskiy (2020) studied a different robust covariance estimator that admits tight deviation bounds under the finite kurtosis condition. However, both estimators involve a brute-force search and hence are computationally intractable in high-dimensional set-up. Avella-Medina et al. (2018) combined robust estimates of the first and second moments to obtain variance estimators from an element-wise perspective. The estimator proposed in Avella-Medina et al. (2018) uses cross-validation to calibrate a total number of dimension squared tuning parameters which is computationally expensive in practice. Motivated by the ideas of Minsker (2018) and Avella-Medina et al. (2018), we propose an efficient tail-robust covariance estimator that enjoys desirable finite-sample deviation bounds under weak moment conditions. The constructed estimator is computationally efficient for large-scale problems since it is based on a simple truncation technique and a novel data-driven tuning scheme. These two points distinguish our work from the aforementioned robust covariance estimators in the literature.*

Step 2: Estimate the number of latent factors

Estimating q, the number of latent factors, in (8) is an intrinsic un-supervised learning problem since factors, loading and idiosyncratic noises are all unobsevable. To avoid the ambiguity towards the definition of q, the Condition 2 (b) assumes that there exists a non-negative integer q such that the first q eigenvalues of $\mathbf{\Sigma}_Z$ are diverging with p, while the

rest $p-q$ eigenvalues are bounded. This definition is similar to the ones used in existing high-dimensional factor analysis literature, we refer to Chamberlain and Rothschild (1983), Stock and Watson (2002), Bai and Ng (2002), and more recent references.

Let $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p$ and $\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_p$ be the eigenvalues and corresponding eigenvectors of $\widehat{\mathbf{\Sigma}}_Z$ respectively. We follow the modified ratio method, e.g., equation (10) in Chang et al. (2015), to estimate the number of latent factors. Let q_{max} be a prescribed upper bound and C_T be a constant that depends on p and T. The number of factors can be estimated by

$$\widehat{q} = \underset{k \leq q_{max}}{\operatorname{argmin}} \frac{\widehat{\lambda}_{k+1} + C_T}{\widehat{\lambda}_k + C_T}. \tag{13}$$

For the special case that $\mathbf{Z}_t$ itself is weakly correlated, one can estimate q as 0. In our numerical studies, we choose $q_{max} = p/2$ and $C_T = \ln T/10T$ as recommended in Xia et al. (2015).

Lemma 1 *Under Condition 2, we have*

$$\max_{1 \leq \ell \leq q} |\widehat{\lambda}_\ell - \lambda_\ell| \leq p \|\widehat{\mathbf{\Sigma}}_Z - \mathbf{\Sigma}_Z\|_{\max}, \tag{14}$$

$$\textit{and} \quad \max_{1 \leq \ell \leq q} \|\widehat{\mathbf{v}}_\ell - \mathbf{v}_\ell\|_\infty \leq C_1 (p^{-1/2} \|\widehat{\mathbf{\Sigma}}_Z - \mathbf{\Sigma}_Z\|_{\max} + p^{-1} \|\mathbf{\Sigma}_{\mathbf{u}}\|_2), \tag{15}$$

where $C_1 > 0$ is a constant independent of (T, p).

Lemma 1 gives uniform upper bounds for estimated eigenvalues and eigenvectors. The following lemma shows that Lemma 1 together with Theorem 2 can yield a consistency argument of $\widehat{q}$ similar as Theorem 2.4 in Chang et al. (2015). Hence we omit its proof.

Lemma 2 *(Theorem 2.4 in Chang et al. (2015)) Under Condition 2, we have*

$$P(\widehat{q} \neq q) \to 0, \quad \textit{as} \quad T \to \infty.$$

Besides the modified ratio method, Bai and Ng (2002) studied the estimation of the number of factors for high-dimensional factor models. They proposed to estimate q by minimizing a family of information criteria. We refer to (9) in Bai and Ng (2002) for viable examples.

Step 3: Estimate loading and latent factors

Then, we estimate the loading $\mathbf{B}$ and latent factors $\mathbf{f}_t$ as follows. Define

$$\widehat{\mathbf{B}} = (\widehat{\lambda}_1^{1/2} \widehat{\mathbf{v}}_1, \ldots, \widehat{\lambda}_q^{1/2} \widehat{\mathbf{v}}_{\widehat{q}}) \in \mathbb{R}^{p \times \widehat{q}}$$

as an estimator of $\mathbf{B}$. Let $\{\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_p\} \in \mathbb{R}^{\widehat{q}}$ be the p rows of $\widehat{\mathbf{B}}$, and define

$$\widehat{\mathbf{f}}_t = \underset{\mathbf{f} \in \mathbb{R}^{\widehat{q}}}{\operatorname{argmin}} \sum_{j=1}^{p} \ell_\gamma (Z_{jt} - \widehat{\mathbf{b}}_j^{\mathrm{T}} \mathbf{f}), \quad t = 1, \ \ldots, \ T, \tag{16}$$

where γ is a robustification parameter. The following theorem gives uniform upper bounds for the estimated loading and latent factors.

Theorem 3 *Assume that Condition 2 holds and $C_2 - C_6$ are positive constants independent of (T, N, p). Choose $\min_{1\le k,l\le p}\tau_{k\ell} \ge C_2\sqrt{T/(\log p)}$ and $\gamma \ge C_3\sqrt{p}$. Then, we have*

$$\max_{1\le j\le p} \|\widehat{\mathbf{b}}_j - \mathbf{b}_j\| \le C_4\{(\log p)^{1/2}T^{-1/2} + p^{-1/2}\}, \tag{17}$$

$$\text{and} \quad \max_{1\le t\le T} \|\|\widehat{\mathbf{f}}_t - \mathbf{f}_t\| \le C_5(\log p/p)^{1/2}, \tag{18}$$

with probability at least $1 - C_6p^{-1}$.

Remark 3 *In the past few years, robust factor model estimation has been studied by various literature in statistics, econometrics, and finance. We refer to Fan et al. (2016, 2018a, 2019b,a, 2020a,b), among others. Most existing robust factor model estimation methods adopt a two-stage scheme: first obtain a "good enough" robust covariance estimator, and then approximate the factor model by principal component analysis. Therefore, innovation mainly resides in the first stage. For example, Fan et al. (2018a) proposed a general principal orthogonal complement thresholding to estimate elliptical factor models. Fan et al. (2016) exploited rank-based and quantile-based covariance estimators for robust factor model estimations. Fan et al. (2019b,a, 2020a) used adaptive Huber type robust covariance matrix estimators in the first stage. In this paper, we also followed this two-stage scheme. In the first stage, we proposed an efficient truncation based robust covariance estimator which is comparable to the adaptive Huber estimator used in Fan et al. (2019b,a, 2020a) but computationally less expensive.*

Then, we plug the estimated factors back to (6) and estimate θ_i by solving the following convex optimizaiton problem:

$$\widetilde{\theta}_i(\tau, \widehat{\mathbf{f}}) = \underset{\theta\in\mathbb{R}^d}{\operatorname{argmin}} \sum_{t=1}^{T} \ell_\tau(y_{it} - \widehat{\mathbf{W}}_{it}^{\mathrm{T}}\theta), \quad \text{for} \quad i = 1, \cdots, N, \tag{19}$$

where $\widehat{\mathbf{W}}_{it} = (1, \mathbf{X}_{it}^{\mathrm{T}}, \widehat{\mathbf{f}}_t^{\mathrm{T}})^{\mathrm{T}}$. Denote $M = Np$. Let $\widetilde{\beta}_{ij} := \widetilde{\beta}_{ij}(\tau, \widehat{\mathbf{f}})$ the estimator of β_{ij}, $1 \le i \le N$ and $1 \le j \le p$, which is a sub-vector of $\widetilde{\theta}_i(\tau, \widehat{\mathbf{f}})$ in (19). The corollary below gives a uniform upper bound of $\widetilde{\beta}_{ij}$'s.

Corollary 1 *Assume that Conditions 1 − 2 hold, and C_7 and C_8 are positive constants independent of (T, N, p). For $1 \le i \le N$ and $1 \le j \le p$,*

$$\max_{i,j} |\widetilde{\beta}_{ij} - \beta_{ij}| \le C_7 \{\log M/T\}^{1/2},$$

with probability at least $1 - C_8p^{-1}$.

In comparison with Theorem 1, the uniform upper bound of $\widetilde{\beta}_{i,j}(\tau, \widehat{\mathbf{f}})$ is close to the uniform upper bound for the oracle estimator with known factors, i.e., $\widetilde{\beta}_{i,j}(\tau, \mathbf{f})$. The numerical studies in Section 5 show that $\widetilde{\beta}_{i,j}(\tau, \widehat{\mathbf{f}})$ performs very similar as the oracle estimator $\widetilde{\beta}_{i,j}(\tau, \mathbf{f})$ in various finite sample scenarios.

3. Homogeneity Structure Learning

In this section, we describe a generic homogeneity structure learning procedure.

3.1 Detect the Homogeneity Structure

In this subsection, we detect the homogeneity structure embedded in the robust estimator $\widetilde{\beta}_{ij}$'s, $i = 1, \ldots, N$ and $j = 1, \ldots, p$. Denote $\{\widetilde{\beta}_{(m)}\}_{m=1}^{M}$ the sorted sequence of $\widetilde{\beta}_{ij}$'s in an ascending order. Without loss of generality, we assume $\beta_{0,1} < \ldots < \beta_{0,K+1}$ and the change points located at $\eta_{(0)} < \eta_{(1)} < \ldots < \eta_{(K)} < \eta_{(K+1)}$, where $\eta_{(0)} = 1$ and $\eta_{(K+1)} = M$. As we can see, the partition $\{A_k : 1 \le k \le K+1\}$ in (4) is uniquely defined by the number and locations of change-points among $\{\widetilde{\beta}_{(m)}\}_{m=1}^{M}$, i.e., K and $\{\eta_{(k)}\}_{k=1}^{K}$.

Binary segmentation techniques have been extensively studied for multiple change-points detection applications, see Vostrikova (1981), Bai (1997), Chen et al. (2011), Killick et al. (2012), Fryzlewicz and Subba Rao (2014), and Cho and Fryzlewicz (2015), to name but a few. As a relatively new member of the house, the wild binary segmentation (WBS) method (Fryzlewicz, 2014) detects the change-points in some randomly drawn sub-intervals instead of the whole interval. This "localizing" setup allows WBS to achieve near-optimal theoretical results with much weaker conditions on the spacing between change-points and minimal jump magnitudes. Besides, extensive numerical studies have shown that WBS is faster and more stable than the standard binary segmentation in various scenarios. Based on the success of WBS, we propose to detect the homogeneity structure with a procedure summarized in Algorithm 1 below.

Remark 4 *Algorithm 1 involves two pre-specified parameters R and ξ. R controls the number of random intervals and should be "as large as possible" subject to computational constraints as suggested in Fryzlewicz (2014). The stopping criterion ξ works as a threshold that decides if a change point should be recovered in a region or not. For a given region, if the cumulative sum statistic defined in* (20) *falls below ξ, Algorithm 1 does not detect any change point in this region and stops further splitting this region. As recommended in Fryzlewicz (2014), one should choose $\xi = C_\xi\sqrt{2\ln T}$ for some positive constant C_ξ. Notice that, the number of detected change points $\widehat{K}$ is a nonincreasing function of ξ. Therefore, one can select C_ξ through BIC or the Strengthened Schwarz information criterion (SSIC) proposed in Fryzlewicz (2014). In our numerical studies, we choose $R = 5000$ and $\xi = \sqrt{2\ln T}$ which are the default values in the WBS package* [2].

Condition 3 *Denote $\underline{\eta} = \min_{0\le k\le K}\left\{\eta_{(k+1)} - \eta_{(k)}\right\}$, the minimum separation between two neighboring change-points. Denote $\underline{\beta} = \min_{1\le k\le K}(\beta_{0,k+1} - \beta_{0,k})$, the minimum jump between two neighboring homogeneity groups. We require $\{\beta_{0,k}\}_{k=1}^{K+1}$ bounded and $\underline{\eta}^{1/2}\underline{\beta} \ge c_5\log^{1/2} M$ for some positive constant c_5.*

Condition 3 assumes that the minimum spacing between two neighboring change-points and the minimum signal jump between two neighboring homogeneity groups to diverge

2. The WBS package is available at `https://cran.r-project.org/web/packages/wbs/index.html`

Algorithm 1 Change-points detection with wild binary segmentation

Input: Ascending sorted initial estimator $\{\widetilde{\beta}_{(m)}\}_{m=1}^{M}$, number of random intervals R and stopping criterion ξ.

Step 1

Randomly draw a set of R intervals $[s_r, e_r]$, $r = 1, \ldots, R$. The start point s_r and the end point e_r are drawn uniformly from the set $\{1, \ldots, M\}$.

Step 2

2.1 For each given interval $[s_r, e_r]$, apply binary segmentation by finding the index $\tilde{\eta}_r$ that maximizes a cumulative sum statistic defined as

$$\widehat{Q}^{\eta}_{s_r,e_r} = \sqrt{\frac{e_r - \eta}{M_r(\eta - s_r + 1)}} \sum_{m=s_r}^{\eta} \widetilde{\beta}_{(m)} - \sqrt{\frac{\eta - s_r + 1}{M_r(e_r - \eta)}} \sum_{m=\eta+1}^{e_r} \widetilde{\beta}_{(m)}, \tag{20}$$

where $M_r = e_r - s_r + 1$.

2.2 Pick the index $\widehat{\eta}_1$ as the first detected change point that satisfy

$$\widehat{\eta}_1 = \underset{r\in[1,R], b\in[s_r,e_r)}{\operatorname{argmax}} |\widehat{Q}^{b}_{s_r,e_r}| \quad \text{and} \quad |\widehat{Q}^{\widehat{\eta}_1}_{s_r,e_r}| > \xi,$$

where ξ is a pre-specified stopping criterion.

Step 3

Divide the original interval $[1, M]$ into two sub-intervals $[1, \widehat{\eta}_1]$ and $[\widehat{\eta}_1 + 1, M]$. Repeat Step 1 and Step 2 on each sub-interval to detect new change points.

Step 4

Repeat Step 3 for any newly detected change points until no new change point is detected. Denote $\widehat{K}$ and $\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$ the estimated number and locations of change points respectively. Resort $\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$ in ascending order and denote the new sequence as $\{\widehat{\eta}_{(k)}\}_{k=1}^{\widehat{K}}$

Output: $\widehat{K}$ and $\{\widehat{\eta}_{(k)}\}_{k=1}^{\widehat{K}}$.

logarithmically slowly with M, which is a very mild condition. Hence, we allow the number of change points K to slowly diverge with N and p. Theorem 4 below shows that Algorithm 1 can correctly detect the number and all locations of change-points with high probability.

Theorem 4 *Assume that Conditions 1, 2 and 3 hold. Let C_9 and C_{10} be two positive constants independent of (T, N, p). Choose the threshold ξ and the number of random intervals R to satisfy*

$$c_5 \log^{1/2} M \le \xi \le 2\underline{\eta}^{1/2}\underline{\beta} \quad \textit{and} \quad R \ge 9T^2\underline{\eta}^{-2}\log(Mp/\log\underline{\eta})$$

respectively, then with probability at least $1 - C_9 p^{-1}$,

$$\widehat{K} = K \quad \textit{and} \quad \max_{1\le k\le K} |\widehat{\eta}_{(k)} - \eta_{(k)}| \le C_{10}\underline{\beta}^{-2}\log M.$$

Remark 5 *In this paper, we mainly focused on detecting the homogeneity structure and estimating homogeneity coefficients. Besides, the inference issues of the estimated coefficients are important in many economic and statistical applications. Here we briefly review*

the recent developments of inference for longitudinal data with high dimensional covariates. The seminal papers Zhang and Zhang (2014) and Van de Geer et al. (2014) proposed a general framework for constructing confidence intervals and statistical tests for single or low-dimensional components of a large parameter vector in a high-dimensional model. Later, Ning and Liu (2017) developed a novel decorrelated score function to assess the uncertainty for low dimensional components in high dimensional models. Specifically, their methods can be applied to study hypothesis tests and confidence regions for generic M-estimators. More recently, Fang et al. (2020) studied the statistical inference for longitudinal data with ultra-high dimensional covariates. They addressed the challenge of constructing a powerful test statistic in the presence of high-dimensional nuisance parameters and sophisticated within-subject correlation of longitudinal data. Follow the analysis in Fang et al. (2020), we may show that the proposed homogeneity coefficient estimator is asymptotically normal, based on which we can construct an optimal Wald test statistic. Due to the limited space, we do not pursue this direction in the paper.

3.2 Estimate Homogeneity Coefficients

In this subsection, we introduce the estimation of homogeneity coefficients $\{\beta_{0,k} : 1 \le k \le K+1\}$ with the homogeneity structure detected by Algorithm 1.

Denote $\widehat{\eta}_{(0)} = 0$, $\widehat{\eta}_{(\widehat{K}+1)} = \infty$, and

$$\widehat{A}_k = \{\beta_{(m)} \ : \ \widehat{\eta}_{(k-1)} < m \le \widehat{\eta}_{(k)}\}, \quad k = 1, \ \ldots, \ \widehat{K}+1,$$

the detected homogeneity structure. We re-parameterise β_{ij} in (3) by setting $\beta_{ij} = \beta_{0,k}$ if $\beta_{ij} \in \widehat{A}_k$, for $i = 1, \ \ldots, \ N$ and $j = 1, \ \ldots, \ p$. Through this re-parameterisation, the Np unknown parameters β_{ij}'s are reduced to $\widehat{K}+1$ unknown parameters $\beta_{0,k}$'s.

Replacing each β_{ij} in (19) by its corresponding $\beta_{0,k}$ is equivalent to minimize the following empirical Huber's loss over a reduced parameter space.

$$(\widehat{\beta}_i, \widehat{\alpha}_i, \widehat{\lambda}_i) = \underset{\beta_i \in \mathcal{B}, \alpha_i \in \mathbb{R}, \lambda_i \in \mathbb{R}^q}{\operatorname{argmin}} \sum_{t=1}^{T} \ell_\tau(y_{it} - \alpha_i - \mathbf{X}_{it}^{\mathrm{T}}\beta_i - \widehat{\mathbf{f}}_t^{\mathrm{T}}\lambda_i), \quad \text{for} \quad i = 1, \ \ldots \ , N, \tag{21}$$

where $\mathcal{B} = \{\beta_{ij} : \beta_{ij} = \beta_{0,k}$ if $\beta_{ij} \in \widehat{A}_k$; $i = 1, \ \ldots, \ N$, $j = 1, \ \ldots, \ p$ and $k = 1, \ \ldots, \ \widehat{K}+1.\}$ is a $\widehat{K}+1$ dimensional subspace of $\mathbb{R}^{N \times p}$.

Corollary 2 *Assume that Conditions 1, 2 and 3 hold. Let C_{11} and C_{12} be two positive constants independent of (T, N, p). For $1 \le k \le K$, denote $\widehat{\beta}_{0,k} = \widehat{\beta}_{ij}$, $\forall \beta_{ij} \in A_k$. Then we have*

$$\max_k |\widehat{\beta}_{0,k} - \beta_{0,k}| \le C_{11} \left\{\log K/T\right\}^{1/2}, \tag{22}$$

with probability at least $1 - C_{12}p^{-1}$.

Corollary 2 gives a uniform upper bound for the homogeneity coefficient estimator with the detected homogeneity structure. By comparing it with Corollary 1, the upper bound in (22) replaces the diverging term $\log(Np)$ with a much smaller one $\log K$, which justifies the intuition that correctly learning the homogeneity structure can avoid overfitting in panel data analysis.

Algorithm 2 Robust homogeneity structure learning

Input: Observed data $(\mathbf{X}_{it}, y_{it}) \in \mathbb{R}^{p+1}$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$. Upper bound $q_{\max}$, constant C_T, number of random intervals R and stopping criterion ξ.

1. Estimate covariance matrix

1.1 Calculate $\mathbf{Z}_t = N^{-1}\sum_{i=1}^{N}\mathbf{X}_{it} = (Z_{t1}, \ldots, Z_{tp})^{\mathrm{T}}$.

1.2 For $1 \le k \le \ell \le p$, select $\tau_{k\ell} = \tau_{\ell k}$ by solving (23).

1.3 Calculate $\widehat{\sigma}_{\ell k} = \widehat{\sigma}_{k\ell}$ by (10) and Collect $\widehat{\mathbf{\Sigma}}_Z = (\widehat{\sigma}_{k\ell})_{1\le k,\ell\le p}$.

2. Estimate latent factors

2.1 Apply eigen-decomposition to $\widehat{\mathbf{\Sigma}}_Z$. Let $\widehat{\lambda}_1 \ge \cdots \ge \widehat{\lambda}_{q_{\max}}$ be the first $q_{\max}$ eigenvalues of $\widehat{\mathbf{\Sigma}}_Z$ in a descending order, and $\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_{q_{\max}}$ be the corresponding eigenvectors.

2.2 Estimate the number of factors by (13)

2.3 Estimate the factor loading by $\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_p)^{\mathrm{T}} = (\widehat{\lambda}_1^{1/2}\widehat{\mathbf{v}}_1, \ldots, \widehat{\lambda}_q^{1/2}\widehat{\mathbf{v}}_{\widehat{q}})$.

2.4 Estimate latent factors $\{\widehat{\mathbf{f}}_t\}_{t=1}^{T}$ by (16).

2.5 Estimate the coefficients $\{\widetilde{\beta}_i\}_{i=1}^{N}$ by (19), where $\widetilde{\beta}_i = (\widetilde{\beta}_{i1}, \ldots, \widetilde{\beta}_{ip})^{\mathrm{T}}$.

3. Detect homogeneity structure

3.1 Sort $\widetilde{\beta}_{ij}$'s in an ascending order and denote the obtained sequence as $\{\widetilde{\beta}_{(m)}\}_{m=1}^{M}$ with $M = Np$.

3.1 Detect the number and location of changes points $\widehat{K}$ and $\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$ by inputting $\{\widehat{\beta}_{(m)}\}_{m=1}^{M}$, R and ξ to Algorithm 1.

4. Estimate homogeneity coefficients

Obtain the final estimator of $\{\widehat{\beta}_i\}_{i=1}^{N}$ by (21).

Output: $\widehat{K}$, $\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$ and $\{\widehat{\beta}_{0,k}\}_{k=1}^{\widehat{K}}$

4. Implementation

In this section, we summarize the proposed robust homogeneity structure learning procedure as a computational algorithm. We also present a fast robust covariance estimation method.

4.1 Computational Algorithm

To conclude Sections 2 and 3, we summarize the full robust homogeneity structure learning procedure in Algorithm 2 below. The computational complexity of Algorithm 2 mainly resides in the covariance matrix estimation step which is of the order $O(p^2T^2)$. To address this issue, we introduce a fast robust covariance matrix estimation method in Section 4.2 below.

4.2 Fast Robust Covariance Estimation

The element-wise covariance estimator (9) entails $p(p-1)/2$ robustification parameters $\tau_{k\ell}$, $1 \le k, \ell \le p$. When the dimensionality p is large, it is computationally expensive to select $\tau_{k\ell}$'s through cross-validations. Recently, Ke et al. (2019) proposed a fast data-driven approach to select the robustification parameters and estimate the covariance matrix simultaneously by solving a system of equations. Numerical studies therein suggest that the new data-driven method is considerably faster than the cross-validation while performs equally

as well. This fast data-driven covariance matrix estimation method can be implemented by an R package named FarmTest[3] (Bose et al., 2021).

For the completeness of the paper, we briefly illustrate this fast data-driven approach. Denote $\mathcal{T} = T(T-1)/2$. For the ease of presentation, we fix $1 \le k \le \ell \le p$ and define

$$\{U_1 \ \ldots, \ U_{\mathcal{T}}\} = \{\frac{(Z_{1k} - Z_{2k})(Z_{1\ell} - Z_{2\ell})}{2}, \frac{(Z_{1k} - Z_{3k})(Z_{1\ell} - Z_{3\ell})}{2}, \\ \ldots, \ \frac{(Z_{(T-1)k} - Z_{Tk})(Z_{(T-1)\ell} - Z_{T\ell})}{2}\}.$$

The dependence of $U_1, \ldots, U_{\mathcal{T}}$ on k and l has been suppressed for the simplicity of notations.

One can see that $U_1, \ldots, U_{\mathcal{T}}$ are weakly stationary with $\mathbb{E}(U_1) = \sigma_{k\ell}$ and $\mathbb{E}(U_1^2) = v_{k\ell}^2$. Suggested by (11), an "ideal" choice of $\tau_{k\ell}$ is

$$\tau_{k\ell} = v_{k\ell}\sqrt{\frac{T_0}{2\log p + \log \delta^{-1}}},$$

where δ is prespecified to control the confidence level in (12). In the presence of heavy-taildness, we expect the empirical truncated second moment

$$\mathcal{T}^{-1}\sum_{i=1}^{\mathcal{T}} \psi_{\tau_{k\ell}}^2(U_i) = \mathcal{T}^{-1}\sum_{i=1}^{\mathcal{T}} (U_i^2 \wedge \tau_{k\ell}^2)$$

to be a reasonable estimate of $\mathbb{E}(U_1^2)$. Plugging this estimator in (9) yields the following equation of τ

$$\frac{1}{\mathcal{T}}\sum_{i=1}^{\mathcal{T}} \frac{(U_i^2 \wedge \tau^2)}{\tau^2} = \frac{2\log p + \log \delta^{-1}}{T_0}, \quad \tau > 0. \tag{23}$$

We propose to use the solution of (23), namely $\widehat{\tau}_{k\ell}$, as a data-driven choice of $\tau_{k\ell}$. With $\widehat{\tau}_{k\ell}$, the calculation of (9) is straightforward and there is no optimization involved.

As δ controls the confidence level in (12), we should let $\delta = \delta(p)$ be sufficiently small so that the estimator is concentrated around the true value with a high probability. On the other hand, δ^{-1} also appears in the deviation bound that corresponds to the width of the confidence interval, it should not grow too fast as a function of p. We refer to Wang et al. (2020) for more discussions on the properties of (23). In practice, we recommend using $\delta = p^{-1}$, a typical slowly varying function of p.

5. Simulations

In this section, we use simulated examples to assess the finite sample performance of the proposed estimation procedure. Throughout this section, we set $N = 100$, $T = 200$, $p = 30$ and $q = 2$. For each scenario, we simulate 200 replications unless otherwise specified.

3. The FarmTest package is available at https://cran.r-project.org/web/packages/FarmTest/index.html.

5.1 Data Generation

Consider a panel data model with interactive effects:

$$\begin{cases} y_{it} &= \alpha_i + \mathbf{X}_{it}^{\mathrm{T}}\beta_i + \mathbf{f}_t^{\mathrm{T}}\lambda_i + \varepsilon_{it}, \\ \mathbf{X}_{it} &= \mathbf{b}_i\mathbf{f}_t + \mathbf{u}_{it}, \end{cases} \qquad i = 1,\ \ldots,\ N,\ \ t = 1,\ \ldots,\ T,$$

where $\alpha_i \in \mathbb{R}$, $\beta_i \in \mathbb{R}^p$, $\lambda_i \in \mathbb{R}^q$, $\mathbf{b}_i \in \mathbb{R}^{p\times q}$, $\mathbf{f}_t \in \mathbb{R}^q$, $\varepsilon_{it} \in \mathbb{R}$, and $\mathbf{u}_{it} \in \mathbb{R}^p$.

The intercepts α_i's are independently drawn from a uniform distribution $U(-1, 1)$. The latent factors $\mathbf{f}_t$ are independently drawn from $N(0, \mathbf{I}_q)$. The factor loading $\mathbf{b}_i = \{b_{i,kj}\}, k = 1,\ \ldots,\ p, j = 1, 2$ are generated as

$$b_{i,kj} = \begin{cases} \sin(2\pi k/p) & \text{if } j = 1, \\ \cos(2\pi k/p) & \text{if } j = 2. \end{cases}$$

Besides, the coefficients $\lambda_i = (\lambda_{i,1}, \lambda_{i,2})^{\mathrm{T}}$ are generated as

$$\lambda_{i,j} = \begin{cases} \sin(2\pi i/N) & \text{if } j = 1, \\ \cos(2\pi i/N) & \text{if } j = 2. \end{cases}$$

Each element of $\{\mathbf{u}_{it}\}$ and $\{\varepsilon_{it}\}$ are sampled independently from one of the following three distributions:

(a) Normal distribution with mean 0 and variance 3;

(b) t-distribution with mean 0 and degree of freedom 2.1;

(c) Pareto distribution with location and dispersion parameters being 1 and 2, respectively. This distribution is then re-scaled to have zero mean.

The distribution (b) is heavy-tailed. The distribution (c) is both heavy-tailed and asymmetric.

Next, we generate the regression coefficients of interest β_{ij}'s, $i = 1,\ \ldots,\ N$ and $j = 1,\ \ldots,\ p$. Each β_{ij} is independently generated from one of the following two homogeneity structures:

(i) 5-GROUPS: discrete uniform distribution with atoms $\{-2r, -r, 0, r, 2r\}$;

(ii) 9-GROUPS: discrete uniform distribution with atoms $\{-4r, -3r, -2r, -r, 0, r, 2r, 3r, 4r\}$.

The structures (i) and (ii) have 5 and 9 groups, respectively. For both structures, the signal strength r is set to be 1 (week), 2 (medium), or 4 (strong).

To sum up, we run simulations over three error distributions, two homogeneity structures, and three levels of signal strength. That is 18 scenarios in total.

5.2 Covariance and Latent Factors Estimation

In this subsection, we assess the performance of the covariance and latent factors estimation procedure as proposed in Section 2.2.

First, we compare the covariance estimation performance of our robust covariance estimator (Our), the adaptive Huber estimator (Ah), the median of means estimator (Mom) and the sample covariance estimator (Sample). Our is implemented as the Step 1 in Section 2.2. The matrix of robustification parameters $\mathbf{\Gamma} = (\tau_{k\ell})_{1\le k,\ell\le p}$ are selected by the tuning-free method introduced in Section 4.2. The implementations of Ah and Mom follow Avella-Medina et al. (2018). The tuning parameters of Ah and Mom are selected by five-fold cross-validations.

Recall that $\mathbf{Z}_t = N^{-1}\sum_{i=1}^N \mathbf{X}_{it}$ and $\mathbf{\Sigma}_Z$ is the covariance matrix of $\{\mathbf{Z}_t\}_{t=1}^T$. For each replication, we calculate the following two matrix norms,

$$\Delta_{\max}(\widehat{\mathbf{Z}}^{(l)}) = \|\widehat{\mathbf{\Sigma}}_Z^{(l)} - \mathbf{\Sigma}_Z\|_{\max} \text{ and } \Delta_{\mathrm{F}}(\widehat{Z}^{(l)}) = \|\widehat{\mathbf{\Sigma}}_Z^{(l)} - \mathbf{\Sigma}_Z\|_{\mathrm{F}}, \quad l = 1,\ \ldots,\ 200, \tag{24}$$

where $\widehat{\mathbf{\Sigma}}_Z^{(l)}$ is an estimator of $\mathbf{\Sigma}_Z$ in the lth replication. The estimation accuracy of $\mathbf{\Sigma}_Z$ is measured by the sample mean and the sample standard deviation of the norms in (24) over 200 replications. The results of the four competing methods with different error distributions are summarized in Table 1. When the data are generated with Normal errors, the performance of Our, Sample and Ah are comparable while Mom has slightly larger sample means. When the data are generated with heavy-tailed errors (e.g., t and Pareto distributions), all three robust estimators outperform Sample by big margins in terms of smaller sample means and sample standard deviations. To better compare the performance of three robust estimators under heavy-tailed scenarios, we report the boxplots of their error norms in Figure 3. According to Figure 3, Our and Ah perform comparably in all four scenarios, and Mom performs the worst among the three. Further, we make a wall-time computational cost comparisons among Our, Ah and Mom over 200 replications [4]. The average wall-time running costs of Our, Ah and Mom are 0.8, 11 and 7 seconds per replication, respectively. To sum up, Our pays a little price in light-tailed scenarios but gains a big advantage in the presence of heavy-tailed errors. Besides, Our performs the best among three competing robust covariance estimators in terms of both estimation accuracy and computational efficiency.

Next, we compare the factor estimation performance between our robust factor estimator (Our) and the estimator proposed in Pesaran (2006) (Pesaran). Our is implemented as the procedure introduced in Section 2.2. Pesaran uses the cross-sectional mean of both $\mathbf{X}_{it}$ and y_{it} to proxy the unobserved factor $\mathbf{f}_t$. Denote $\widehat{\mathbf{F}}^{(l)}$ an estimator of the latent factors $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)^{\mathrm{T}}$ in the lth replication. The estimation accuracy of $\mathbf{F}$ is measured by the canonical correlation analysis (CCA) between the estimator and the truth (the larger the better) as

$$\mathrm{CCA}(\widehat{\mathbf{F}}^{(l)}) \equiv \mathrm{CCA}(\widehat{\mathbf{F}}^{(l)}, \mathbf{F}), \quad l = 1,\ \ldots,\ 200, \tag{25}$$

where the $\mathrm{CCA}(\cdot,\cdot)$ stands for the sample canonical correlation between two matrices. The boxplots of canonical correlations over 200 replications are reported in Figure 4. Our

4. For each method, we simulate 200 replications on the same computer cluster node with Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz and 256Gb RAM.

Estimation error	Method	Normal	$t_{2.1}$	Pareto
$\Delta_{\max}(\widehat{\mathbf{Z}}^{(l)})$	Our	0.022 (0.005)	0.073 (0.086)	0.297 (1.751)
	Sample	0.022 (0.005)	0.672 (2.388)	6.662 (50.702)
	Ah	0.022 (0.005)	0.073 (0.086)	0.298 (1.751)
	Mom	0.027 (0.005)	0.082 (0.085)	0.306 (1.751)
$\Delta_{\mathrm{F}}(\widehat{\mathbf{Z}}^{(l)})$	Our	0.227 (0.053)	0.437 (0.438)	1.598 (9.576)
	Sample	0.227 (0.052)	0.917(2.397)	7.017(51.546)
	Ah	0.227 (0.053)	0.438 (0.438)	1.598 (9.576)
	Mom	0.269 (0.046)	0.497 (0.436)	1.657 (9.574)

Table 1: Sample means and sample standard deviations (numbers in parentheses) of covariance matrix estimation errors defined in (24) over 200 replications. Normal, $t_{2.1}$ and Pareto stand for the three error distributions listed in Section 5.1.

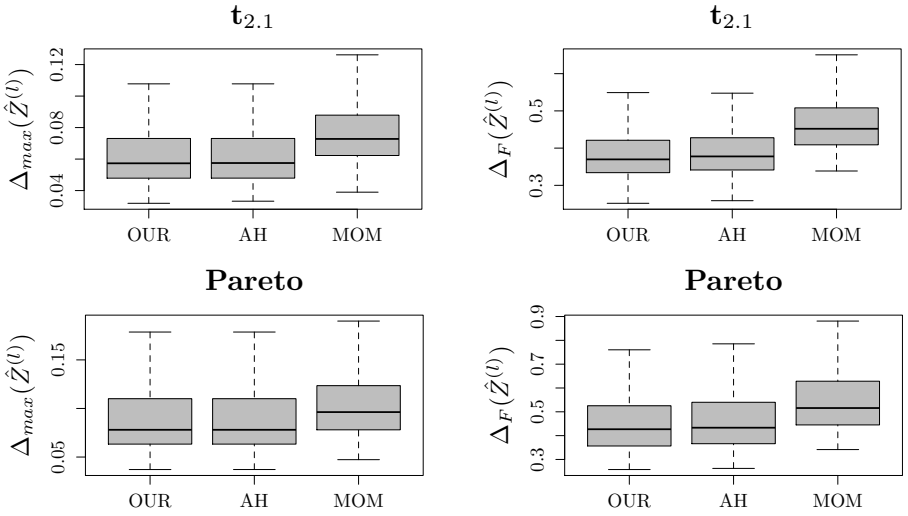

Figure 3: Covariance matrix estimation: boxplots of estimation errors for Our, Ah, and Mom. The left column reports the estimation error in max norm, while the right column reports the estimation error in Frobenius norm. The top and bottom rows represent simulation results for $t_{2.1}$ and Pareto cases, respectively.

performs as well as Pesaran in the Normal case. However, when the errors are drawn from heavy-tailed distributions, Our outperforms Pesaran as expected.

5.3 Regression Coefficients Estimation

In this subsection, we assess the homogeneity learning and robust coefficient estimation procedure introduced in Section 3.1. We propose to compare 5 estimators listed as follows.

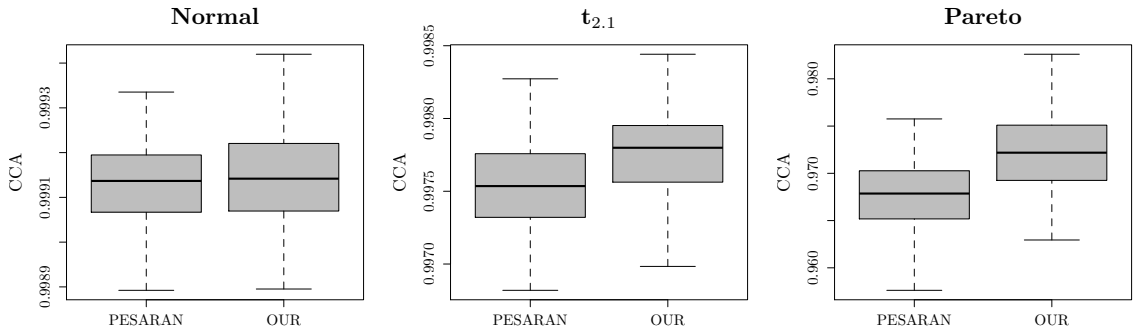

Figure 4: Factor estimation: boxplots of sample canonical correlations between the estimated factors and the truth over 200 replications (the larger the better). The left, middle and right columns present simulation results for Normal, $t_{2.1}$ and Pareto cases, respectively.

(i) Our estimator: We estimate β_i's, $i = 1, \ldots, N$, with the procedure introduced in Algorithm 2. In other words, we consider both robust estimation and homogeneity detection in the estimation procedure.

(ii) Oracle estimator: Similar to Our except that we treat the latent factors as observable and the true homogeneity structure as known. The Oracle estimator is used as a performance upper bound benchmark in the comparison.

(iii) Homogeneity estimator: Similar to Our except that we do not pursue robust estimations throughout the estimation procedure. Specifically, we replace (9), (19) and (21) by their OLS counterparts.

(iv) Robust estimator: Similar to Our except that we do not pursue the homogeneity detection procedure. To be specific, we use the estimator obtained from (19) as the final estimator.

(v) OLS estimator: Similar to Robust except that we do not pursue the robust estimations throughout the procedure. Namely, we replace (9) and (19) by their OLS counterparts.

In Table 2, we summarize the similarities and differences of the above 5 estimators according to three aspects: robust estimation; homogeneity detection; and latent factors.

Denote $\widehat{\beta}_i^{(l)}$ an estimator of β_i in the lth experiment. We measure the estimation accuracy of $\widehat{\beta}_i^{(l)}$ by calculating the root-mean-squared-error (RMSE).

$$\text{RMSE}(\widehat{\beta}_i^{(l)}) = \big\{(Np)^{-1}\sum_{i=1}^{N}\|\widehat{\beta}_i^{(l)} - \beta_i\|^2\big\}^{1/2}, \quad l = 1, \ldots, 200.$$

In Figures 5—7, we report the boxplots of RMSE of 5 estimators with errors generated from the Normal, t and Pareto distributions, respectively. For each error distribution, we

	Robust Estimation	Homogeneity Detection	Latent factors
Our	Yes	Yes	Un-observable
Oracle	Yes	Known	Observable
Homogeneity	No	Yes	Un-observable
Robust	Yes	No	Un-observable
OLS	No	No	Un-observable

Table 2: "Specificaiton" table for the 5 estimators in Section 5.3

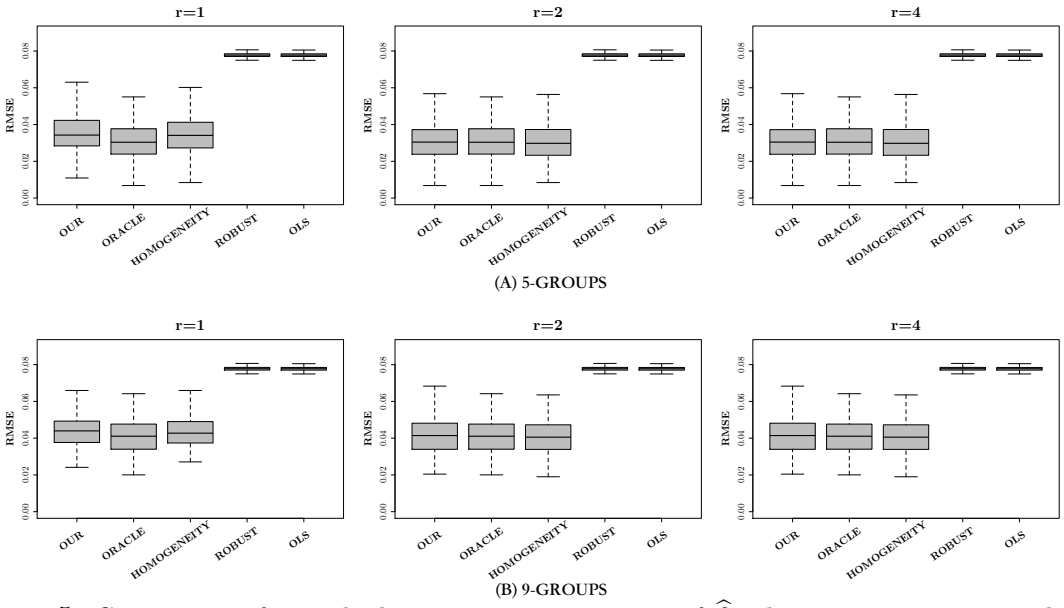

Figure 5: Comparison of 5 methods in estimation accuracy of $\widehat{\beta}_i$ when noises are generated from Normal distribution over 200 replications. The top and bottom rows represent two homogeneity structures: 5-groups and 9-groups respectively. The three columns represent the signal strengths $r = 1$, $r = 2$, and $r = 4$ respectively.

consider two homogeneity structures: 5-groups and 9-groups. For each homogeneity structure, we set the signal strength r to be 1 (week), 2 (medium), or 4 (strong). We refer to Section 5.1 for more details.

According to Figure 5, with Normally distributed errors, the estimators that use known or detected homogeneity structure (Our, Oracle and Homogeneity) outperform the estimators that ignore the homogeneity structure (Robust and OLS). When the errors follow heavy-tailed distributions, like the results in Figures 6 and 7, robust estimators (Our, Oracle and Robust) outperform the other two non-robust competitors. Under various group structures and signal strengths, the performance of Our and Oracle are fairly close to each other which indicates that Our can effectively detect the hidden homogeneity structure and robustly estimate the coefficients in the presense of heavy-tailed errors.

Next, we assess the accuracy of the detected homogeneity structure by calculating the sample mean of the adjusted Rand index (Hubert and Arabie, 1985) between Our estimator

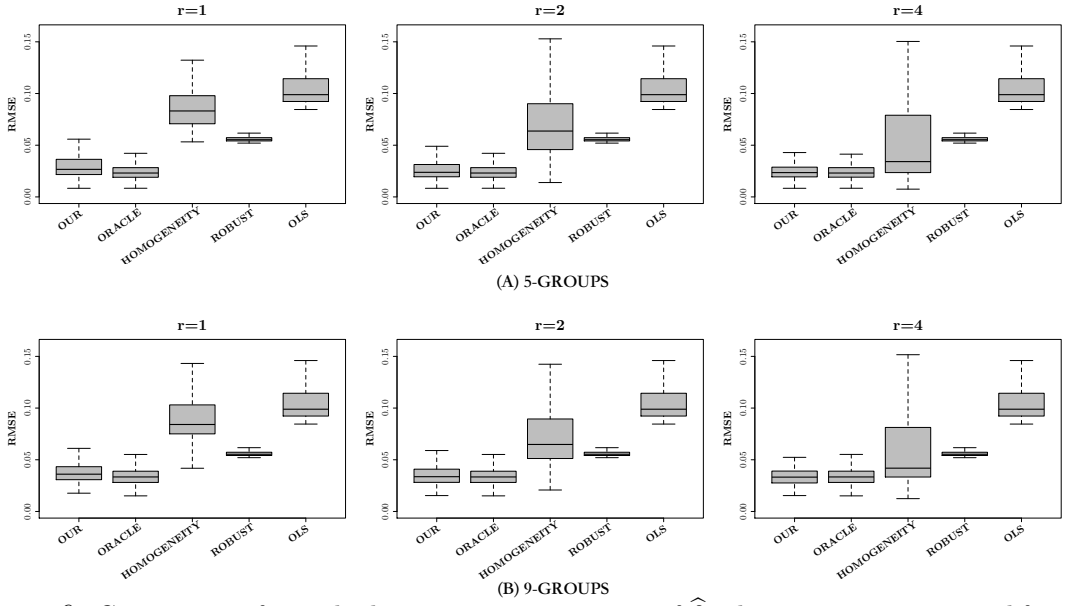

Figure 6: Comparison of 5 methods in estimation accuracy of $\widehat{\beta}_i$ when noises are generated from $t_{2.1}$ distribution over 200 replications. The top and bottom rows represent two homogeneity structures: 5-GROUPS and 9-GROUPS, respectively. The three columns represent the signal strengths $r = 1$, $r = 2$, and $r = 4$, respectively.

Structure	Signal strength	Normal	t-distribution	Pareto
5-GROUPS	$r = 1$	0.9993	0.9990	0.9967
	$r = 2$	1	0.9999	0.9990
	$r = 4$	1	0.9999	0.9997
9-GROUPS	$r = 1$	0.9995	0.9989	0.9940
	$r = 2$	1	0.9998	0.9989
	$r = 4$	1	0.9999	0.9998

Table 3: Adjusted Rand index of OUR (the higher the better).

and the truth over 200 replications. The results, presented in Table 3, are close to one in all scenarios, which indicates that the proposed homogeneity detection procedure can perfectly identify the number of groups as well as group memberships in most replications.

Further, we compare OUR with two popular homogeneity detection methods: the method proposed in Pesaran (2006) (denoted as PESARAN); and the method proposed in Su and Ju (2018) (denoted as SU). To make a fair comparison, we follow the data generating process 1 (static panel model) in Section 5.1 of Su and Ju (2018) with one latent factor. The error terms in $\mathbf{X}_{it}$ and Y_{it} follow either $N(0, 3)$ or t_3 distribution. In other words, the errors are generated from two distributions with the same variance but different tail behaviors. The box-plots of RMSE over 100 replications are presented in Figure 8. For the Normal error case, OUR performs similarly as SU which indicates that our method does not lose any

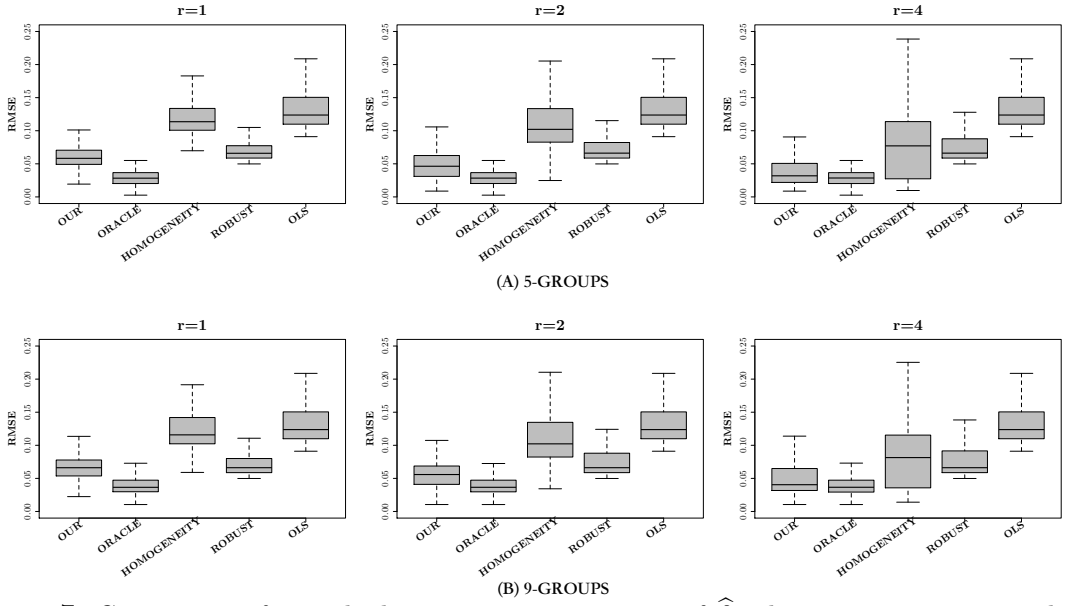

Figure 7: Comparison of 5 methods in estimation accuracy of $\widehat{\beta}_i$ when noises are generated from Pareto distribution over 200 replications. The top and bottom rows represent two homogeneity structures: 5-GROUPS and 9-GROUPS, respectively. The three columns represent the signal strengths $r = 1$, $r = 2$, and $r = 4$, respectively.

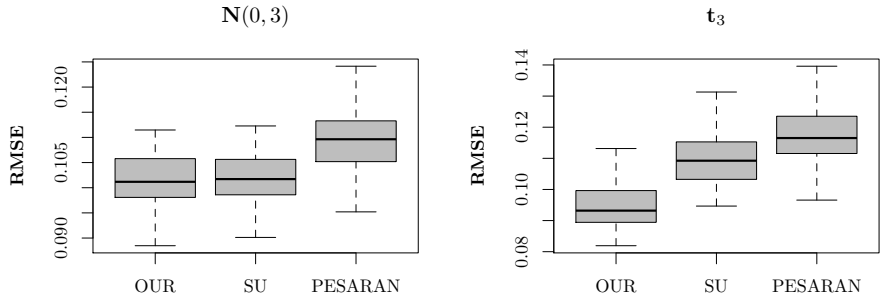

Figure 8: Comparison of Our, Su, and Pesaran in estimation accuracy of $\widehat{\beta}_i$ over 100 replications. The left and right columns represent the error terms in $\mathbf{X}_{it}$ and Y_{it} follow $N(0, 3)$ and t_3 distributions, respectively.

efficiency in the light-tailed case. In the t_3 distribution case, Our outperforms Su as Our is robust against heavy-tailed errors. In both cases, Pesaran performs the worst.

5.4 Serial Dependent Case

The simulation settings are similar as in Section 5.1 except that we generate data with serial dependent and heavy-tailed errors. Specifically, we generate $\{\mathbf{u}_{it}\}$'s and $\{\varepsilon_{it}\}$'s from

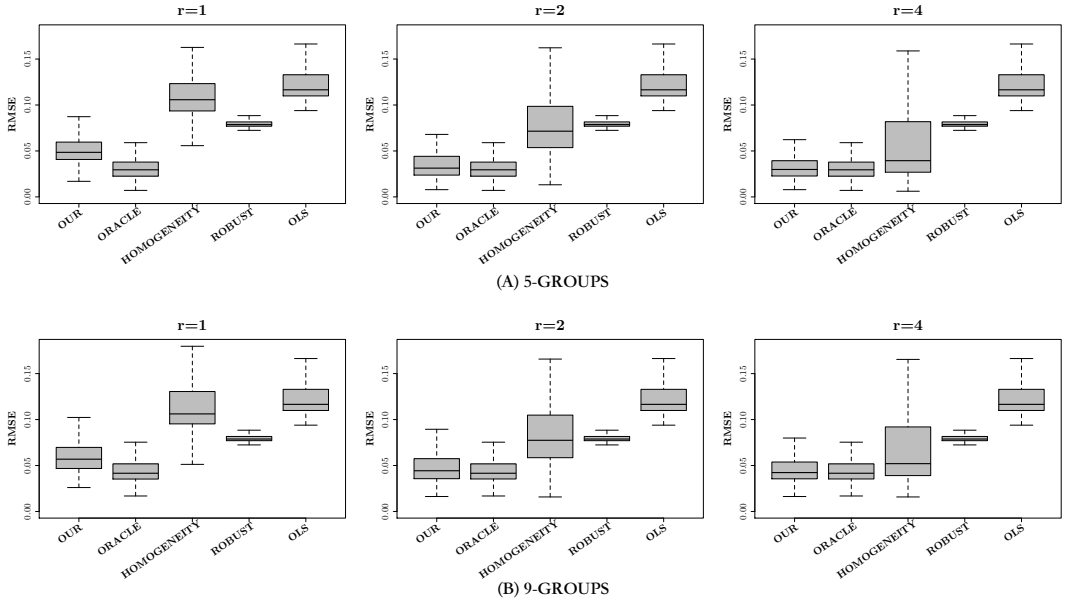

Figure 9: Comparison of 5 methods in estimation accuracy of $\widehat{\beta}_i$ when noises are generated from serial dependent t distributions. The top and bottom rows represent two homogeneity structures: 5-GROUPS and 9-GROUPS, respectively. The three columns represent the signal strengths $r = 1$, $r = 2$, and $r = 4$, respectively.

Structure	Signal strength	t-distribution	Pareto
5-GROUPS	$r = 1$	0.9950	0.9865
	$r = 2$	0.9996	0.9978
	$r = 4$	0.9999	0.9995
9-GROUPS	$r = 1$	0.9956	0.9860
	$r = 2$	0.9996	0.9977
	$r = 4$	0.9999	0.9995

Table 4: Adjusted Rand index of Our for serial dependent data.

a stationary VAR(1) model and a stationary AR(1) model as follows.

$$\mathbf{u}_{i,t} = \mathbf{\Pi}\mathbf{u}_{i,t-1} + \mathbf{v}_{i,t}, \quad \varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + \eta_{i,t}, \quad i = 1, \ldots, N, \text{ and } t = 1, \ldots, T,$$

with $\mathbf{u}_{i,0} = \mathbf{0}$, $\varepsilon_{i,0} = 0$ and $\rho = 0.5$. The (i, j)th entry of $\mathbf{\Pi}$ is set to be 0.5 when $i = j$ and $0.1^{|i-j|}$ when $i \neq j$. In addition, each element of $\{\mathbf{v}_{i,t}\}$ and $\{\eta_{i,t}\}$ is sampled independently from one of the two heavy-tailed distributions ($t_{2.1}$ and Pareto) listed in Section 5.1.

In Figures 9 and 10, we report the boxplots of RMSE of 5 estimators with errors generated from the serial dependent t and Pareto distributions, respectively. In Table 4, we report the sample mean of the adjusted Rand index between Our and the truth over 200 simulations.

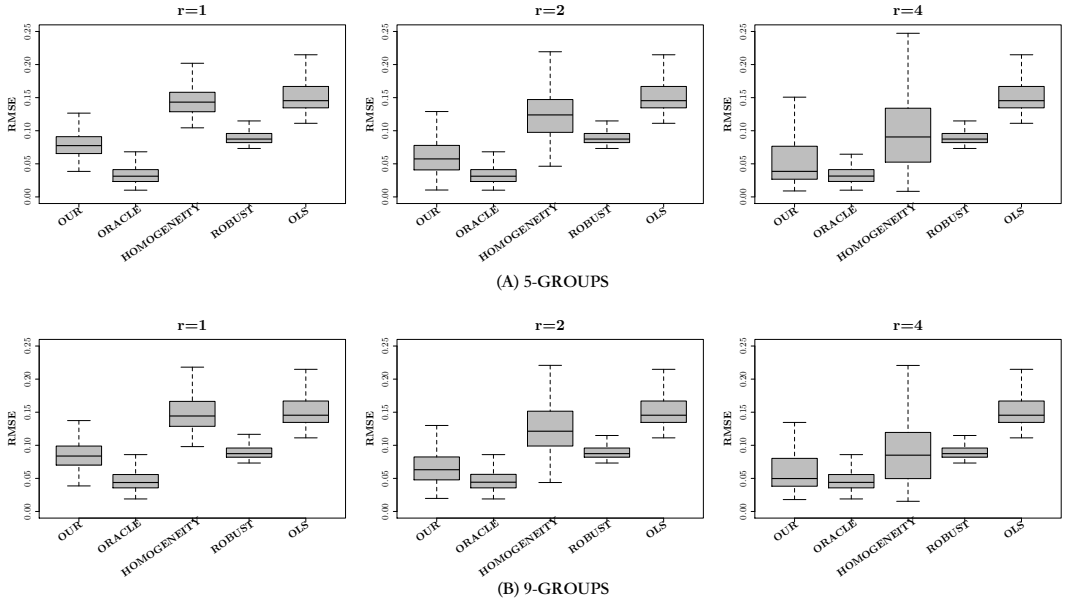

Figure 10: Comparison of 5 methods in estimation accuracy of $\widehat{\beta}_i$ when noises are generated from serial dependent Pareto distributions. The top and bottom rows represent two homogeneity structures: 5-GROUPS and 9-GROUPS, respectively. The three columns represent the signal strengths $r = 1$, $r = 2$, and $r = 4$, respectively.

5.5 Sensitivity of Robustificaiton Parameter Selection

For the proposed robust covariance matrix estimator, the robustificaiton parameters $(\tau_{kl})_{1\leq k,l\leq p}$ can be selected by the fast and data-driven method proposed in Section 4.2. The robust linear regressions in (16), (19) and (21) also involve selecting robustification parameters. Similarly, these robustification parameters can be selected in a data-driven manner. This problem has been independently studied in Wang et al. (2020). Since the computations of these robust linear regressions are relatively fast, we propose to use the 5-fold cross-validation to select their robustification parameters in our numerical studies.

In this subsection, we assess the sensitivity of the robustificaiton parameter selection. To be specific, we follow the data generating process in Section 5.1 with the error terms of $\mathbf{X}_{it}$ and y_{it} being sampled independently from the Pareto distribution. First, we calculate the proposed robust covariance matrix estimator over a sequence of robustification parameters. For ease of presentation, we set $\tau_{kl} = \tau^*$ for $1 \leq k, l \leq p$. We set τ^* to be a sequence of equally spaced grid points between 1 and 10. The estimation accuracy is measured by the Frobenius norm defined in (24). The results, presented in the left panel of Figure 11, show a flat elbow-shaped curve which indicates the proposed robust covariance estimator is not sensitive to the choice of robustificaiton parameters. Similarly, we show the relationship between the choice of the robustificaiton parameter τ and the estimation accuracy of the robust linear regression estimator proposed in (19). We set τ to be a sequence of equally spaced grid points between 0 and 3. The estimation accuracy of the robust linear regression estimator was measured by RMSE which was defined in Section 5.3. The results

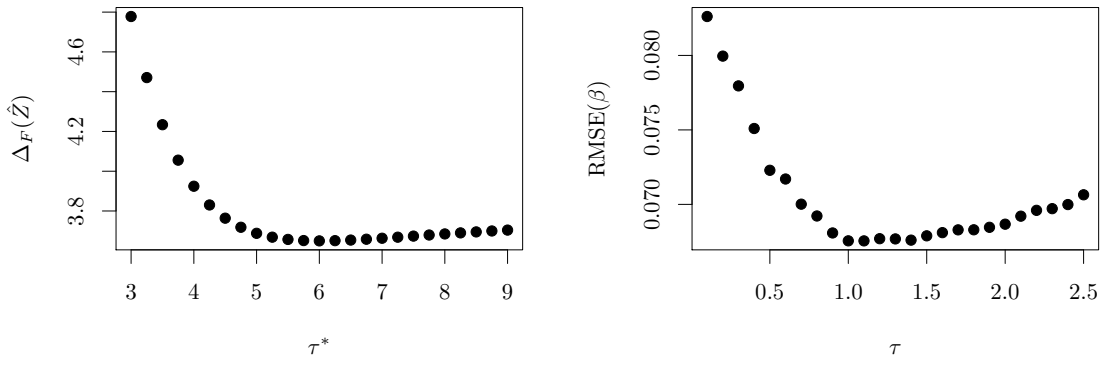

Figure 11: Sensitivity of robustificaiton parameter: (Left) Performance of robust covariance estimation with respect to τ^*. (Right) Performance of robust linear regression with respect to τ.

are presented in the right panel of Figure 11. Again, the elbow-shaped curve indicates the cross-validation approach can effectively select a τ that minimizes the empirical validation error.

6. Real Application

Particulate matter (PM) is a complex mixture of solid particles, chemicals (e.g., sulfates, nitrates) and liquid droplets in the air, which include inhalable particles that are small enough to penetrate the thoracic region of the respiratory system. The hazardous effects of inhalable PM on human health have been well-documented (e.g., Polichetti et al., 2009; Xing et al., 2016; Pun et al., 2017). Short term (days) exposure to inhalable PM can cause an increase in hospital admissions related to respiratory and cardiovascular morbidity, such as aggravation of asthma, respiratory symptoms, and cardiovascular disorders. Long term (years) exposure to inhalable PM may lead to an increase in mortality from cardiovascular and respiratory diseases, like lung cancer. Franck et al. (2011) studied the composition of PM and showed that particles of size up to $2.5\,\mu m$ (PM2.5) exerts the most significant negative impact on human health. Recently, many literature (e.g., Zheng et al., 2005; Liang et al., 2015) focused on figuring out the sources that cause PM2.5 to accumulate in the air.

In this section, we study the relationship between the concentrations of PM2.5 and the other four air pollutants: ozone, sulfur dioxide (SO2), carbon monoxide (CO), and nitrogen dioxide (NO2). The data set[5] was collected from $N = 37$ outdoor monitors across United States which consists of $T = 729$ daily observations between January 2017 to April 2019. Each time-series in the data set has been taken first-order difference and standardized to have zero mean and unit variance. We model the data set with an interactive effects model as follows

$$\begin{cases} y_{it} &= \mathbf{X}_{it}^{\mathrm{T}}\beta_i + \mathbf{f}_t^{\mathrm{T}}\lambda_i + \varepsilon_{it}, \\ \mathbf{X}_{it} &= \mathbf{b}_i\mathbf{f}_t + \mathbf{u}_{it}, \end{cases} \qquad i = 1,\ \ldots,\ N,\ \ t = 1,\ \ldots,\ T, \tag{26}$$

5. The data set is available at `https://www.epa.gov/outdoor-air-quality-data`.

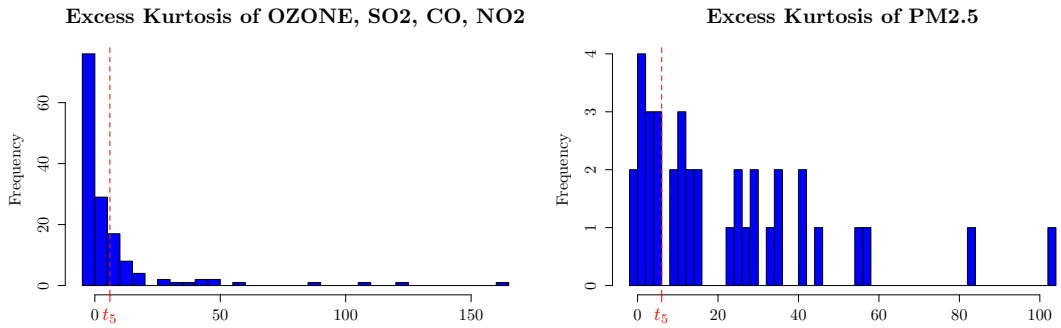

Figure 12: (Left) Histogram of excess kurtosis of 148 time series of covariates. (Right) Histogram of excess kurtosis of 37 time series of the PM2.5.

where y_{it} is the pre-processed PM2.5 data at the ith monitor and the tth day. Similarly; $\mathbf{X}_{it} \in \mathbb{R}^4$ are the pre-processed O3, SO2, CO and NO2 data at the ith monitor and the tth day. $\beta_i \in \mathbb{R}^4$ are individual attributes of covariates at each monitor. $\mathbf{f}_t$ are latent factors that affects both covariates and the response variable. $\mathbf{b}_i$ and λ_i are factor loadings. ε_{it} and $\mathbf{u}_{it}$ are random errors.

First, we calculate the sample excess kurtosis of each time-series. The left panel of Figure 12 shows that 72 out of 148 time-series in covariates have tails heavier than Gaussian distribution, and 38 time-series have tails heavier than t_5 distribution. In the right panel of Figure 12, PM2.5 time series monitored at 35 out of 37 locations have tails heavier than Gaussian distribution, and 25 of them have tails heavier than t_5 distribution. These observations indicate that both y_{it} and $\mathbf{X}_{it}$ are severely heavy-tailed.

Next, we learn the homogeneity structure in individual attributes with the proposed learning procedure. With $q_{max} = 4$ and $C_T = 0.01$, the modified ratio method (13) estimates the number of factors to be 1. Indeed, the first eigenvector of the robust covariance estimator $\widehat{\mathbf{\Sigma}}_Z$ explains 80% of its total variation. Algorithm 2 detects 6 homogeneity groups in $\{\beta_i\}_{i=1}^N$. The learning results, visualized in Figure 13, unveils a parsimonious and interpretable relationship between PM2.5 and the other four air pollutants. The attributes of each pollutant are clustered into three geological areas in the United States: west coast, central and east coast. Among the four pollutants, CO has the largest positive contribution to the concentration of PM2.5. As we know, CO is usually produced in the incomplete combustion of carbon-containing fuels, such as gasoline, natural gas, coal, and wood. Two major anthropogenic sources of CO in the United States are vehicle emissions and heating. According to Figure 13, areas in California and around New York City have high CO coefficients which are caused by the dense vehicle population. Also, we notice that the monitors with higher latitudes have higher CO coefficients which may reflect the impact of heating.

Also, we compare the prediction performance of the proposed homogeneity learning procedure (denoted as Our) with the one that ignores the homogeneity structure and heavy-tailedness (denoted as OLS). To this end, we conduct a rolling-window out-of-sample prediction procedure. Start from the first day in the data set, we use a window size of 250 days as the training set to predict the next 50 days. Each time, the window moves 20 days forward. Within each window, we first estimate $\hat{\boldsymbol{f}}_t$ in the test set. Then, we calculate the

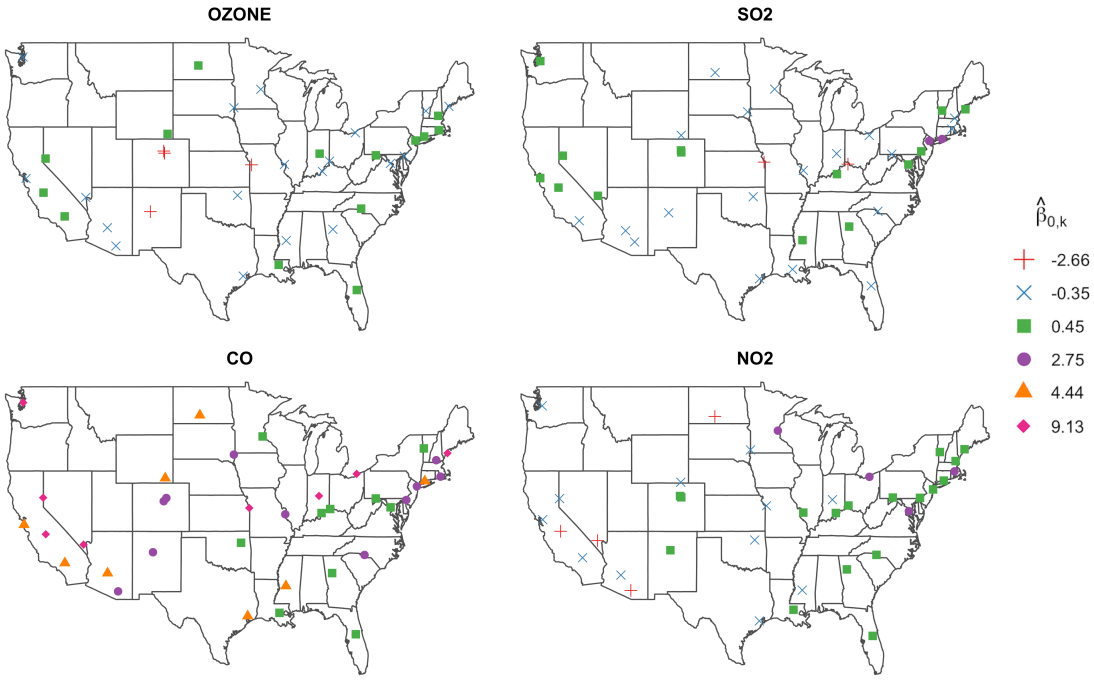

Figure 13: Homogeneity learning results for the coefficients estimate of four air pollutants.

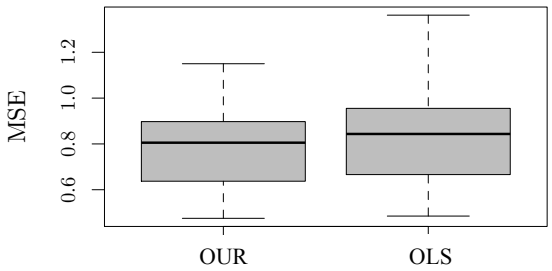

Figure 14: Rolling-window out-of-sample MSE for Our method and the OLS method.

mean squared errors (MSE) over the test set, which is defined as

$$\mathrm{MSE}_m = (Nh)^{-1} \sum_{i=1}^{N} \sum_{t=t_m}^{t_m+h} (\hat{y}_{it} - y_{it})^2, \quad m = 1, \ \ldots, \ M,$$

where $\hat{y}_{it}$ is the concentration of PM2.5 at the ith location and the tth day predicted by either Our or Ols. Besides, $N = 37$ is the number of monitors, $h = 50$ is the size of the test set, $M = 22$ is the number of rolling windows, and t_m is the start time of the test set in the mth window. Figure 14 presents the boxplots of $\{\mathrm{MSE}_m\}_{m=1}^{M}$ for Our and Ols, respectively. One can observe, in this application, learning the homogeneity structure with our robust estimation method can consistently improve prediction accuracy over Ols .

Acknowledgments

The authors would like to thank the reviewers for their constructive comments, which lead to a significant improvement of this work. Li's research was supported by NSF grants DMS 1820702, 1953196 and 2015539.

Appendix A. Proof of Section 2

In this appendix, we provide proofs for the theoretical results in Section 2.

A.1 Proof of Theorem 1

By the union bound, for any $\xi > 0$, it holds

$$\begin{aligned}&\mathbb{P}(\max_{1\le i\le N}\|\widehat{\theta}_i-\theta_i\|_2\ge c_l^{-1}\,\xi)\\&\le\sum_{1\le i\le N}\mathbb{P}(|\|\widehat{\theta}_i-\theta_i\|_2\ge c_l^{-1}\,\xi)\le N\max_{1\le k\le\ell\le p}\mathbb{P}(|\|\widehat{\theta}_i-\theta_i\|_2\ge c_l^{-1}\,\xi).\end{aligned}\tag{27}$$

In the rest of the proof, we fix $i \in \{1, \ldots, N\}$ and suppress the subscription i in (5) for the ease of notation. In addition, we write $\mathbf{S}_i = \mathbf{S}$ and $\tau_i = \tau$. Define the loss function $L_\tau(\theta) = T^{-1}\sum_{j=1}^T \ell_\tau(Y_j - \mathbf{W}_j^{\mathrm{T}}\theta)$ for $\theta \in \mathbb{R}^d$. Denote θ^* the true parameters and $\widehat{\theta} = \mathrm{argmin}_{\theta\in\mathbb{R}^d} L_\tau(\theta)$. Without loss of generality, we assume $\|\mathbf{W}\|_{\max} = 1$ for simplicity.

We can construct an intermediate estimator, denoted by $\widehat{\theta}_\eta = \theta^* + \eta(\widehat{\theta} - \theta^*)$, such that $\|\mathbf{S}^{1/2}(\widehat{\theta}_\eta - \theta^*)\|_2 \le r$ for some $r > 0$ to be specified. We take $\eta = 1$, if $\|\mathbf{S}^{1/2}(\widehat{\theta} - \theta^*)\|_2 \le r$; otherwise, we choose $\eta \in (0, 1)$ so that $\|\mathbf{S}^{1/2}(\widehat{\theta}_\eta - \theta^*)\|_2 = r$. The Lemma A.1 in Sun et al. (2019) gives

$$\langle \nabla L_\tau(\widehat{\theta}_\eta) - \nabla L_\tau(\theta^*),\ \widehat{\theta}_\eta - \theta^*\rangle \le \eta\langle \nabla L_\tau(\widehat{\theta}) - \nabla L_\tau(\theta^*),\ \widehat{\theta} - \theta^*\rangle,$$

where $\nabla L_\tau(\widehat{\mathbf{w}}) = \mathbf{0}$ according to the Karush-Kuhn-Tucker condition.

According to Lemma A.2 in Sun et al. (2019) there exists some constant $a_{\min} > 0$ such that

$$\min_{\theta\in\mathbb{R}^d:\|\theta-\theta^*\|_2\le r}\lambda_{\min}(\nabla^2 L_\tau(\theta)) \ge a_{\min}.\tag{28}$$

Then, by and the mean value theorem for vector-valued functions

$$\nabla L_\tau(\widehat{\theta}_\eta) - \nabla L_\tau(\theta^*) = \int_0^1 \nabla^2 L_\tau\big((1-t)\theta^* + t\widehat{\theta}_\eta\big)\,dt\,(\widehat{\theta}_\eta - \theta^*).$$

It follows that $a_{\min}\|\widehat{\theta}_\eta - \theta^*\|_2^2 \le -\eta\langle \nabla L_\tau(\mathbf{w}^*),\ \widehat{\theta} - \theta^*\rangle \le \|\nabla L_\tau(\theta^*)\|_2\|\widehat{\theta}_\eta - \theta^*\|_2$, or equivalently,

$$a_{\min}\|\widehat{\theta}_\eta - \theta^*\|_2 \le \|\nabla L_\tau(\theta^*)\|_2,\tag{29}$$

where $\nabla L_\tau(\theta^*) = -T^{-1}\sum_{j=1}^T \ell'_\tau(\epsilon_j)\mathbf{W}_j$ and $\mathbf{W}_j = (w_{j1},\ \ldots,\ w_{j\ell})^{\mathrm{T}}$.

Next we bound $\|\nabla L_\tau(\theta^*)\|_2$. Let $\psi_\tau(\cdot)$ be the first order derivative of the Huber's loss $\ell_\tau(\cdot)$. For every $1 \le \ell \le d$, we write $\Psi_\ell = T^{-1}\sum_{j=1}^T \psi_{j\ell} := T^{-1}\sum_{j=1}^T \tau^{-1}\psi_\tau(\epsilon_j)w_{j\ell}$, such that $\|\nabla L_\tau(\theta^*)\|_2 \le \sqrt{d}\,\|\nabla L_\tau(\theta^*)\|_\infty = \tau\sqrt{d}\,\max_{1\le\ell\le d}|\Psi_\ell|$. Observe that, for any $u \in \mathbb{R}$,

$$-\log(1 - u + u^2) \le \tau^{-1}\psi_\tau(\tau u) \le \log(1 + u + u^2).$$

After some simple algebra, we obtain that

$$\begin{aligned} e^{\psi_{j\ell}} &\le \{1+\tau^{-1}\epsilon_j+\tau^{-2}\epsilon_j^2\}^{w_{j\ell}I(w_{j\ell}\ge 0)} \\ &\quad + \{1-\tau^{-1}\epsilon_j+\tau^{-2}\epsilon_j^2\}^{-w_{j\ell}I(w_{j\ell}<0)} \\ &\le 1+\tau^{-1}\epsilon_j w_{j\ell}+\tau^{-2}\epsilon_j^2. \end{aligned}$$

Taking expectation on both sides gives

$$\mathbb{E}(e^{\psi_{j\ell}}) \le 1+\tau^{-2}\sigma_\epsilon^2.$$

Moreover, by the independence and the inequality $1+t\le e^t$, $t\in\mathbb{R}$, we get

$$\begin{aligned} \mathbb{E}(e^{p\Psi_\ell}) &= \prod_{j=1}^{T}\mathbb{E}(e^{\psi_{j\ell}}) \le \exp\left(\frac{1}{\tau^2}\sum_{j=1}^{T}\sigma_\epsilon^2\right) \\ &\le \exp\left(\frac{\sigma_\epsilon^2 T}{\tau^2}\right). \end{aligned}$$

For any $s>0$, it follows from the Markov's inequality that

$$\mathbb{P}(T\Psi_\ell \ge 2s) \le e^{-2s}\mathbb{E}(e^{T\Psi_\ell}) \le \exp\left(\frac{\sigma_\epsilon^2 T}{\tau^2}-2s\right) \le \exp(-s)$$

as long as

$$\tau \ge \sigma_\epsilon\sqrt{\frac{T}{s}}. \tag{30}$$

Under the constraint (30), it can be similarly shown that $\mathbb{P}(-T\Psi_\ell \ge 2s)\le e^{-s}$. Putting the above calculations together, we have

$$\begin{aligned} &\mathbb{P}\left\{\|\nabla L_\tau(\theta^*)\|_2 \ge \sqrt{d}\,\frac{2\tau s}{T}\right\} \\ &\le \mathbb{P}\left\{\|\nabla L_\tau(\theta^*)\|_\infty \ge \frac{2\tau s}{T}\right\} \le \sum_{\ell=1}^{d}\mathbb{P}(|T\Psi_\ell|\ge 2s) \le 2d\exp(-s). \end{aligned} \tag{31}$$

With the above preparations, now we are ready to prove the final conclusion. It follows from Lemma A.2 in Sun et al. (2019) that with probability greater than $1-e^{-s}$, (28) holds with $a_{\min}=c_l/2$, provided that $\tau\ge 4r\sqrt{d}$ and $T\ge 32d^2 s$. Hence, combining (29) and (31) with $r=4.1c_l^{-1}\sqrt{d}\,T^{-1}\tau s$ yields that, with probability at least $1-(2d+1)e^{-s}$, $\|\widehat{\theta}_\eta-\theta^*\|_2 \le 4c_l^{-1}\sqrt{d}\,T^{-1}\tau s < r$ as long as $T\ge 32d^2 s$. By the definition of $\widehat{\theta}_\eta$, we must have $\eta=1$ and thus $\widehat{\theta}=\widehat{\theta}_\eta$.

Finally, taking $\xi = 4\sqrt{d}\,T^{-1}\tau s$ in (27) finishes the proof. ■

A.2 Proof of Theorem 2

For each $1 \le k \le \ell \le p$, note that $\widehat{\sigma}_{k\ell}$ is a U-statistic with a bounded kernel of order two, say $\widehat{\sigma}_{k\ell} = \binom{T}{2}^{-1} \sum_{1 \le i < j \le T} h_{k\ell}(\mathbf{Z}_i, \mathbf{Z}_j)$. According to Huber (1984), $\widehat{\sigma}_{k\ell}$ can be represented as an average of (dependent) averages of independent random variables. Specifically, define

$$W_{k\ell}(\mathbf{Z}_1, \ldots, \mathbf{Z}_T) = \frac{h_{k\ell}(\mathbf{Z}_1, \mathbf{Z}_2) + h_{k\ell}(\mathbf{Z}_3, \mathbf{Z}_4) + \cdots + h_{k\ell}(\mathbf{Z}_{2T_0-1}, \mathbf{Z}_{2T_0})}{T_0}$$

for $\mathbf{Z}_1, \ \ldots, \ \mathbf{Z}_T \in \mathbb{R}^p$.

Denote $\sum_{\mathcal{P}}$ as the summation over all $T!$ permutations $(i_1, \ldots, i_T)$ of $[T] := \{1, \ \ldots, \ T\}$ and $\sum_{\mathcal{C}}$ denote the summation over all $\binom{T}{2}$ pairs (i_1, i_2) $(i_1 < i_2)$ from $[T]$. Then we have $\sum_{\mathcal{P}} W_{k\ell}(\mathbf{Z}_1, \ \ldots, \ \mathbf{Z}_T) = 2!(T-2)! \sum_{\mathcal{C}} h_{k\ell}(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2})$ and hence

$$\widehat{\sigma}_{k\ell} = \frac{1}{T!} \sum_{\mathcal{P}} W_{k\ell}(\mathbf{Z}_{i_1}, \ \ldots, \ \mathbf{Z}_{i_T}). \tag{32}$$

Write $\tau = \tau_{k\ell}$ and $v = v_{k\ell}$ for simplicity. For any $\eta > 0$, by the Markov's inequality, (32), convexity and independence, we derive that

$$\begin{aligned}
&\mathbb{P}(\widehat{\sigma}_{k\ell} - \sigma_{k\ell} \ge \eta) \le e^{-\tau^{-1} T_0 (\eta + \sigma_{k\ell})} \mathbb{E} e^{\tau^{-1} T_0 \widehat{\sigma}_{k\ell}} \\
&\le e^{-\tau^{-1} T_0 (\eta + \sigma_{k\ell})} \frac{1}{T!} \sum_{\mathcal{P}} \mathbb{E} e^{\tau^{-1} \sum_{j=1}^{T_0} h_{k\ell}(\mathbf{Z}_{i_{2j-1}}, \mathbf{Z}_{i_{2j}})} \\
&= e^{-\tau^{-1} T_0 (\eta + \sigma_{k\ell})} \frac{1}{T!} \sum_{\mathcal{P}} \prod_{j=1}^{T_0} \mathbb{E} e^{\tau^{-1} h_{k\ell}(\mathbf{Z}_{i_{2j-1}}, \mathbf{Z}_{i_{2j}})}.
\end{aligned}$$

Note that

$$h_{k\ell}(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}) = \psi_\tau\{(Z_{i_1,k} - Z_{i_2,k})(Z_{i_1,\ell} - Z_{i_2,\ell})/2\} = \tau \psi_1\{(Z_{i_1,k} - Z_{i_2,k})(Z_{i_1,\ell} - Z_{i_2,\ell})/(2\tau)\}. \tag{33}$$

Using the inequality that $-\log(1 - x + x^2) \le \psi_1(x) \le \log(1 + x + x^2)$ for all $x \in \mathbb{R}$, we have

$$\begin{aligned}
&\mathbb{E} e^{\tau^{-1} h_{k\ell}(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2})} \\
&\le \mathbb{E}\left\{1 + (Z_{i_1,k} - Z_{i_2,k})(Z_{i_1,\ell} - Z_{i_2,\ell})/(2\tau) + (Z_{i_1,k} - Z_{i_2,k})^2 (Z_{i_1,\ell} - Z_{i_2,\ell})^2/(2\tau)^2\right\} \\
&= 1 + \tau^{-1}\sigma_{k\ell} + \tau^{-2}\mathbb{E}\left\{(Z_{i_1,k} - Z_{i_2,k})(Z_{i_1,\ell} - Z_{i_2,\ell})/2\right\}^2 \le e^{\tau^{-1}\sigma_{k\ell} + \tau^{-2} v^2}.
\end{aligned}$$

Combining the above calculations, we arrive

$$\mathbb{P}(\widehat{\sigma}_{k\ell} - \sigma_{k\ell} \ge \eta) \le e^{-\tau^{-1} T_0 \eta + \tau^{-2} T_0 v^2} = e^{-T_0 \eta^2/(4v^2)},$$

where the equality holds by taking $\tau = 2v^2/\eta$. Similarly, it can be shown that $\mathbb{P}(\widehat{\sigma}_{k\ell} - \sigma_{k\ell} \le -\eta) \le e^{-T_0\eta^2/(4v^2)}$.

Consequently, for $\delta \in (0,1)$, taking $\eta = 2v\sqrt{(2\log p + \log \delta^{-1})/T_0}$, or equivalently, $\tau = v\sqrt{T_0/(2\log p + \log \delta^{-1})}$, we arrive at

$$\mathbb{P}\left(|\widehat{\sigma}_{k\ell} - \sigma_{k\ell}| \ge 2v\sqrt{\frac{\log \delta^{-1}}{T_0}}\right) \le \frac{2\delta}{p^2}.$$

From the union bound it follows that

$$\mathbb{P}\left(\|\widehat{\mathbf{\Sigma}}_Z - \mathbf{\Sigma}_Z\|_{\max} > 2 \max_{1\le k,\ell\le p} v_{k\ell} \sqrt{\frac{2\log p + \log \delta^{-1}}{T_0}} \right) \le (1+p^{-1})\delta,$$

which proves (2). ■

A.3 Proof of Lemma 1

By Weyl's inequality, we have $|\widehat{\lambda}_\ell - \lambda_\ell| \le \|\widehat{\mathbf{\Sigma}}_Z - \mathbf{\Sigma}_Z\|_2$ for each $1 \le \ell \le q$. Moreover, note that for any matrix $\mathbf{E} \in \mathbb{R}^{d_1 \times d_2}$,

$$\|\mathbf{E}\|_2 \le \sqrt{d_1 d_2}\|\mathbf{E}\|_{\max}.$$

Putting the above calculations together proves (14).

Next, we have the following decomposition

$$\widehat{\mathbf{\Sigma}}_Z = \widehat{\mathbf{\Sigma}}_Z - \mathbf{\Sigma}_Z + \mathbf{B}\mathbf{B}^{\mathrm{T}} + \mathbf{\Sigma}_{\mathbf{u}} = \sum_{\ell=1}^{q} \lambda_\ell \mathbf{v}_\ell \mathbf{v}_\ell^{\mathrm{T}} + \widehat{\mathbf{\Sigma}}_Z - \mathbf{\Sigma}_Z + \mathbf{\Sigma}_{\mathbf{u}}.$$

Under Condition 2, it follows from Theorem 3 and Proposition 3 in Fan et al. (2018b) that

$$\max_{1\le\ell\le q} \|\widehat{\mathbf{v}}_\ell - \mathbf{v}_\ell\|_\infty \le \frac{C_1}{p^{3/2}}(\|\widehat{\mathbf{\Sigma}}_Z - \mathbf{\Sigma}_Z\|_\infty + \|\mathbf{\Sigma}_{\mathbf{u}}\|_\infty) \le C_1(p^{-1/2}\|\widehat{\mathbf{\Sigma}}_Z - \mathbf{\Sigma}_Z\|_{\max} + p^{-1}\|\mathbf{\Sigma}_{\mathbf{u}}\|_2),$$

where we use the inequalities $\|\widehat{\mathbf{\Sigma}}_Z - \mathbf{\Sigma}_Z\|_\infty \le p\|\widehat{\mathbf{\Sigma}}_Z - \mathbf{\Sigma}_Z\|_{\max}$ and $\|\mathbf{\Sigma}_{\mathbf{u}}\|_\infty \le p^{1/2}\|\mathbf{\Sigma}_{\mathbf{u}}\|$ in the last step and $C_1 > 0$ is a constant independent of (n, p). This proves (15) . ■

A.4 Proof of Theorem 3

Recall that $\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_p)^{\mathrm{T}} = (\widehat{\lambda}_1^{1/2}\widehat{\mathbf{v}}_1, \ldots, \widehat{\lambda}_q^{1/2}\widehat{\mathbf{v}}_q)$ with $\widehat{\mathbf{v}}_\ell = (\widehat{v}_{\ell 1}, \ldots, \widehat{v}_{\ell p})^{\mathrm{T}}$ for $\ell = 1, \ldots, q$ and $\widehat{\mathbf{b}}_k = (\widehat{\lambda}_1^{1/2}\widehat{v}_{1k}, \ldots, \widehat{\lambda}_q^{1/2}\widehat{v}_{qk})^{\mathrm{T}}$ for $k = 1, \ldots, p$.

Moreover, define $\widetilde{\mathbf{b}}_k = (\lambda_1^{1/2}\widehat{v}_{1k}, \ldots, \lambda_q^{1/2}\widehat{v}_{qk})^{\mathrm{T}}$. By the triangular inequality,

$$\begin{aligned}
\|\widehat{\mathbf{b}}_k - \mathbf{b}_k\|_2 &\le \|\widehat{\mathbf{b}}_k - \widetilde{\mathbf{b}}_k\|_2 + \|\widetilde{\mathbf{b}}_k - \mathbf{b}_k\|_2 \\
&= \left\{ \sum_{\ell=1}^{q} (\widehat{\lambda}_\ell^{1/2} - \lambda_\ell^{1/2})^2 \widehat{v}_{\ell k}^2 \right\}^{1/2} + \left\{ \sum_{\ell=1}^{q} \lambda_\ell (\widehat{v}_{\ell k} - v_{\ell k})^2 \right\}^{1/2} \\
&\le q^{1/2} \left(\max_{1\le\ell\le q} |\widehat{\lambda}_\ell^{1/2} - \lambda_\ell^{1/2}| \|\widehat{\mathbf{v}}_\ell\|_\infty + \max_{1\le\ell\le q} \lambda_\ell^{1/2} \|\widehat{\mathbf{v}}_\ell - \mathbf{v}_\ell\|_\infty \right). \qquad (34)
\end{aligned}$$

According to Theorem 2, $\|\widehat{\mathbf{\Sigma}}_Z - \mathbf{\Sigma}_Z\|_{\max} \le 4\|\mathbf{V}\|_{\max}\sqrt{\log p / T_0}$ with probability at least $1 - 2p^{-1}$. Note that, under Condition 2,

$$\|\mathbf{v}_\ell\|_\infty = \|\lambda_\ell^{1/2}\mathbf{v}_\ell\|_\infty / \lambda_\ell \le \|\mathbf{B}\|_{\max}/\lambda_\ell \lesssim p^{-1/2} \quad \text{for all } \ell = 1, \ldots, q.$$

Then, under Condition 2 and using the results in Lemma 1, we have

$$\begin{aligned}\max_{1\le\ell\le q}|\widehat{\lambda}_\ell^{1/2}-\lambda_\ell^{1/2}| &= \max_{1\le\ell\le q}|\widehat{\lambda}_\ell-\lambda_\ell|/(\widehat{\lambda}_\ell^{1/2}+\lambda_\ell^{1/2})\\ &\lesssim p^{-1/2}\{p\|\widehat{\mathbf{\Sigma}}_Z-\mathbf{\Sigma}_Z\|_{\max}\}\lesssim\sqrt{\frac{p\log p}{T}},\end{aligned}\tag{35}$$

$$\|\widehat{\mathbf{v}}_\ell-\mathbf{v}_\ell\|_\infty\lesssim(p^{-1/2}\|\widehat{\mathbf{\Sigma}}_Z-\mathbf{\Sigma}_Z\|_{\max}+p^{-1}\|\mathbf{\Sigma}_{\mathbf{u}}\|_2)\lesssim\sqrt{\frac{\log p}{pT}}+\frac{1}{p},\tag{36}$$

and

$$\|\widehat{\mathbf{v}}_\ell\|_\infty\le\|\mathbf{v}_\ell\|_\infty+\|\widehat{\mathbf{v}}_\ell-\mathbf{v}_\ell\|_\infty\lesssim\frac{1}{\sqrt{p}}\quad\text{as}\quad\log p\ll T.\tag{37}$$

By plugging the results in (35)—(37) back to (34), we proves (17) by showing

$$\max_{1\le k\le p}\|\widehat{\mathbf{b}}_k-\mathbf{b}_k\|_2\lesssim\sqrt{\frac{\log p}{T}}+\frac{1}{\sqrt{p}}.$$

Denote $\widehat{\mathbf{B}}=(\widehat{\mathbf{b}}_1,\ \ldots,\ \widehat{\mathbf{b}}_p)^T$, we can rewrite (8) as $\mathbf{Z}_t=\mathbf{B}\mathbf{f}_t+\bar{\mathbf{u}}_t=\widehat{\mathbf{B}}\mathbf{g}_t+\mathbf{v}_t$ for $t=1,\ \ldots,\ T$. By the triangular inequality,

$$\|\widehat{\mathbf{f}}_t-\mathbf{f}_t\|_2\le\|\widehat{\mathbf{f}}_t-\mathbf{g}_t\|_2+\|\mathbf{g}_t-\mathbf{f}_t\|_2,$$

where the first term is the estimation error of robust estimator in (16) and the second term is the error-in-variable bias term induced by replacing $\mathbf{B}$ with $\widehat{\mathbf{B}}$.

Follow the similar arguments in the proof of Theorem 1 and choose $s=\log p$ shows that, with probability $1-c_1p^{-1}$

$$\max_{1\le t\le T}\|\widehat{\mathbf{f}}_t-\mathbf{g}_t\|_2\le c_2(\log p/p)^{1/2},\tag{38}$$

where c_1 and c_2 are positive constants independent of (T,p).

To finish the proof, we show the bias term $\|\mathbf{g}_t-\mathbf{f}_t\|$ can be ignored as long as $\widehat{\mathbf{B}}$ is a consistent estimator of $\mathbf{B}$. Denote $\widetilde{\mathbf{f}}_t=(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Z}_t$ the OLS estimator of $\mathbf{f}_t$. The consistency of $\widetilde{\mathbf{f}}_t$ leads to

$$\begin{aligned}\mathbf{f}_t=\lim_{p\to\infty}\widetilde{\mathbf{f}}_t&=\lim_{p\to\infty}\left\{(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Z}_t\right\}\\ &=\lim_{p\to\infty}\left\{(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T(\widehat{\mathbf{B}}\mathbf{g}_t+\mathbf{v}_t)\right\}\\ &=\mathbf{g}_t+\lim_{p\to\infty}\left\{(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T[(\widehat{\mathbf{B}}-\mathbf{B})\mathbf{g}_t+\mathbf{v}_t]\right\}.\end{aligned}\tag{39}$$

Further, denote $\widetilde{\mathbf{u}}_t=\mathbf{Z}_t-\mathbf{B}\widetilde{\mathbf{f}}_t$. We can show that

$$\begin{aligned}\lim_{p\to\infty}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{v}_t&=\lim_{p\to\infty}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T(\mathbf{B}\widetilde{\mathbf{f}}_t-\widehat{\mathbf{B}}\mathbf{g}_t+\widetilde{\mathbf{u}}_t)\\ &=\lim_{p\to\infty}(\widetilde{\mathbf{f}}_t-\mathbf{g}_t)+\lim_{p\to\infty}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T(\mathbf{B}-\widehat{\mathbf{B}})\mathbf{g}_t\\ &\quad+\lim_{p\to\infty}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\widetilde{\mathbf{u}}_t\\ &\equiv I_1+I_2+I_3=0.\end{aligned}$$

Since we choose $s \geq C_3\sqrt{p}$, the difference between $\widehat{\mathbf{f}}_t$ and $\widetilde{\mathbf{f}}_t$ vanishes as $p \to \infty$. This together with $\lim_{p\to\infty} \max_{1\leq t\leq T} \|\widehat{\mathbf{f}}_t - \mathbf{f}_t\|_2 = 0$ proves $I_1 = 0$. Then, we have $I_2 = 0$ since $\lim_{p\to\infty} \max_{1\leq k\leq p} \|\widehat{\mathbf{b}}_k - \mathbf{b}_k\|_2 = 0$. Further, $I_3 = 0$ since $\widetilde{\mathbf{u}}_t$ is the OLS fitting residual.

Hence, we can rewrite (39) as.

$$\mathbf{f}_t = \mathbf{g}_t + \lim_{p\to\infty} \left\{ (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T(\widehat{\mathbf{B}} - \mathbf{B})]\mathbf{g}_t \right\}.$$

In addition, we have

$$\begin{aligned}
&\|(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\left[(\widehat{\mathbf{B}} - \mathbf{B})\right]\mathbf{g}_t\|_2 \\
\leq &\|(\mathbf{B}^T\mathbf{B})^{-1}\|_2\|\mathbf{B}^T\|\|(\widehat{\mathbf{B}} - \mathbf{B})]\|_2\|\mathbf{g}_t\|_2 \\
\leq &\frac{\sqrt{\lambda_1}}{\lambda_q}\sqrt{p}\max_{1\leq k\leq p}\|\widehat{\mathbf{b}}_k - \mathbf{b}_k\|_2\|\mathbf{g}_t\|_2 \\
\lesssim &\max_{1\leq k\leq p}\|\widehat{\mathbf{b}}_k - \mathbf{b}_k\|_2. \qquad (40)
\end{aligned}$$

Therefore $\|\mathbf{f}_t - \mathbf{g}_t\|_2 = 0$ as long as $\lim_{p\to\infty} \max_{1\leq k\leq p} \|\widehat{\mathbf{b}}_k - \mathbf{b}_k\|_2 = 0$. ■

A.5 Proof of Corollary 1

Denote $\widehat{\mathbf{f}}_t$ the estimate of $\mathbf{f}_t$, $1 \leq t \leq T$, which satisfies $\|\widehat{\mathbf{f}}_t - \mathbf{f}_t\|_2 \leq c_1(\log p/p)^{1/2}$ for some positive constant c_1. By replacing $\mathbf{f}_t$ with its estimate, we can rewrite (5) as

$$y_{it} = \mathbf{W}_{it}^{\mathrm{T}}\theta_i + \varepsilon_{it} = \widehat{\mathbf{W}}_{it}^{\mathrm{T}}\theta_i^* + \varepsilon_{it}^*, \qquad (41)$$

where $\widehat{\mathbf{W}}_{it} = (1, \mathbf{X}_{it}^{\mathrm{T}}, \widehat{\mathbf{f}}_t^{\mathrm{T}})^{\mathrm{T}}$, θ_i^* and ε_{it}^* are the coefficients and errors corresponds to $\widehat{\mathbf{W}}_{it}$. By the triangular inequality,

$$\|\widehat{\theta}_i(\tau) - \theta_i\|_2 \leq \|\widehat{\theta}_i(\tau) - \theta_i^*\|_2 + \|\theta_i^* - \theta_i\|_2,$$

where the first term is the estimation error of robust estimator in (6) and the second term is the bias induced by replacing $\mathbf{f}_t$ with $\widehat{\mathbf{f}}_t$.

According to Theorem 1 and choose $t = \log Np$ yields, with probability $1 - c_2p^{-1}$,

$$\max_{1\leq i<N} \|\widehat{\theta}_i(\tau) - \theta_i^*\|_2 \leq c_3(\frac{\log Np}{T})^{1/2}, \qquad (42)$$

where c_2 and c_3 are two positive constants.

Next, with similar arguments as in the Proof of Theorem 3, the bias term $\|\theta_i^* - \theta_i\|_2$ is zero over the event that $\lim_{p\to\infty} \max_{1\leq t\leq T} \|\widehat{\mathbf{f}}_t - \mathbf{f}_t\|_2 = 0$. Under the conditions of Theorem 3, this event holds with probability at least $1 - c_4p^{-1}$ for some positive constant c_4.

To sum up, under Conditions 1 and 2

$$\max_{1\leq i<N} \|\widehat{\theta}_i(\tau) - \theta_i\|_2 \leq c_3(\frac{\log Np}{T})^{1/2}, \qquad (43)$$

with probability at least $1 - (c_2 + c_4)p^{-1}$ which completes the proof. ■

Appendix B. Proof of Section 3

In this appendix, we provide proofs for the theoretical results in Section 3.

B.1 Proof of Theorem 4

The proof of Theorem 4 follows the framework of proving Theorem 3.2 in Fryzlewicz (2014).

Denote $e_{(m)} := \widehat{\beta}_{(m)} - \beta_{(m)}$ for $1 \le m \le M$, and $\mathcal{E}_1$ the event that $\{\max_{1\le m\le M} |e_{(m)}| \le c_1(\log M/T)^{1/2}\}$ for some positive constant c_1. Under the conditions of Corollary 2, $\mathbb{P}(\mathcal{E}_1) \ge 1 - c_2 p^{-1}$ for some positive constant c_2.

Further, we partition the interval between two neighboring change points $[\eta_{(k)}, \eta_{(k+1)}]$ into three equal sized sub-intervals. Denote $\mathcal{I}_{(k)}$ the middle sub-interval, i.e.,

$$\mathcal{I}_{(k)} = \left[\eta_{(k)} + \frac{1}{3}\{\eta_{(k+1)} - \eta_{(k)}\},\ \eta_{(k)} + \frac{2}{3}\{\eta_{(k+1)} - \eta_{(k)}\}\right], \quad 0 \le k \le K.$$

Then, we define an event $\mathcal{E}_2$ in which the R randomly drawn intervals can cover all change points,

$$\mathcal{E}_2 = \{\forall\ k = 0, \ldots, K, \exists\ r = 1,\ \ldots,\ R,\ s.t.\ s_r \in \mathcal{I}_{(k)} \text{ and } e_r \in \mathcal{I}_{(k+1)}\}.$$

Notice that

$$\mathbb{P}\left\{\ s_r \in \mathcal{I}_{(k)} \text{ and } e_r \in \mathcal{I}_{(k+1)}\right\} = \frac{\{\eta_{(k+1)} - \eta_{(k)}\}}{3M} \cdot \frac{\{\eta_{(k+2)} - \eta_{(k+1)}\}}{3M} \ge \frac{\underline{\eta}^2}{9M^2},$$

where $\underline{\eta} = \min_{0\le k\le K}\left\{\eta_{(k+1)} - \eta_{(k)}\right\}$.

Then, we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_2) &\ge 1 - \sum_{k=0}^{K}\prod_{r=1}^{R}\left(1 - P\left\{\ s_r \in \mathcal{I}_{(k)} \text{ and } e_r \in \mathcal{I}_{(k+1)}\right\}\right) \\ &\ge 1 - M\underline{\eta}^{-1}(1 - \underline{\eta}^2 M^{-2}/9)^R.\end{aligned}$$

To bound the exception probability of $\mathcal{E}_2$ on the same order of $\mathcal{E}_1$, we require

$$M\underline{\eta}^{-1}(1 - \underline{\eta}^2 M^{-2}/9)^R \le p^{-1},$$

which is equivalent to choose

$$R \ge 9T^2\underline{\eta}^{-2}\log(Mp/\log\underline{\eta}). \tag{44}$$

In the calculation of (44), we used the fact that $\log(1-x) \approx -x$ when x is close to 0.

Denote the event of interest in Theorem 4 as

$$\mathcal{E}_3 = \left\{\widehat{K} = K \quad \text{and} \quad \max_{1\le k\le K}|\widehat{\eta}_k - \eta_k| \le c_3\underline{\beta}^{-2}\log M\right\},$$

where $\underline{\beta} = \min_{1\le m\le M-1}(\beta_{0,m+1} - \beta_{0,m})$ and c_3 is some positive constant.

Then, similar as the proof of Theorem 3.2 in Fryzlewicz (2014), the arguments in Theorem 4 are valid on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$. For some positive constants c_4 and c_5, we have

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \le 1 - c_4 p^{-1},$$

if we choose the threshold ξ and the number of random intervals R to satisfy $c_5\log^{1/2} M \le \xi \le \underline{\eta}^{1/2}\underline{\beta}$ and (44) respectively. ■

B.2 Proof of Corollary 2

Consider three events

$$\mathcal{E}_1 = \left\{ \max_{1\le t\le T} |\|\widehat{\mathbf{f}}_t - \mathbf{f}_t\| \le c_2(\log p/p)^{1/2} \right\},$$

$$\mathcal{E}_2 = \left\{ \widehat{K} = K \quad \text{and} \quad \max_{1\le k\le K} |\widehat{\eta}_k - \eta_k| \le c_2 \underline{\beta}^{-2} \log M \right\},$$

and

$$\mathcal{E}_3 = \left\{ \max_k |\widehat{\beta}_{0,k} - \beta_{0,k}| \le c_3 \left\{\log K/T\right\}^{1/2} \right\},$$

where c_1, c_2 and c_3 are three positive constants independent of (T, N, p).

Under Conditions 1, 2 and 3, and follow Theorems 3 and 4, we have $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \ge 1 - c_4 p^{-1}$ for some positive constant c_4.

Then, follow the proof of Corollary 1, we can show that $P(\mathcal{E}_3 | \mathcal{E}_1 \cap \mathcal{E}_2) \ge 1 - c_5 p^{-1}$ for some positive constant c_5. Therefore, we finish the proof as

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \ge \mathbb{P}(\mathcal{E}_3 | \mathcal{E}_1 \cap \mathcal{E}_2) \ge 1 - c_5 p^{-1}.$$

■

References

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics*, 48(3):1452–1474, 2020.

Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on twitter for personalized news recommendations. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 1–12. Springer, 2011.

Theodore Wilbur Anderson and Cheng Hsiao. Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18(1):47–82, 1982.

Manuel Arellano. *Panel Data Econometrics*. Oxford university press, 2003.

Marco Avella-Medina, Heather S Battey, Jianqing Fan, and Quefeng Li. Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, 105(2):271–284, 2018.

Jushan Bai. Estimating multiple breaks one at a time. *Econometric Theory*, 13(3):315–352, 1997.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Andrew Bell and Kelvyn Jones. Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, 3(1): 133–153, 2015.

Alok Bhargava, Luisa Franzini, and Wiji Narendranathan. Serial correlation and the fixed effects model. *The Review of Economic Studies*, 49(4):533–549, 1982.

Koushiki Bose, Jianqing Fan, Yuan Ke, Xiaoou Pan, and Wen-xin Zhou. Farmtest: An r package for factor-adjusted robust multiple testing. *The R Journal*, to appear, 2021.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'IHP Probabilités et Statistiques*, 48:1148–1185, 2012.

Olivier Catoni. Pac-bayesian bounds for the gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*, 2016.

Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1305–1324, 1983.

Jinyuan Chang, Bin Guo, and Qiwei Yao. High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. *Journal of Econometrics*, 189(2):297–312, 2015.

Kuo-mei Chen, Arthur Cohen, and Harold Sackrowitz. Consistent multiple testing for change points. *Journal of Multivariate Analysis*, 102(10):1339–1343, 2011.

Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507, 2015.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, 2013.

Jianqing Fan, Fang Han, Han Liu, and Byron Vickers. Robust inference of risks of large portfolios. *Journal of Econometrics*, 194(2):298–308, 2016.

Jianqing Fan, Quefeng Li, and Yuyan Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265, 2017.

Jianqing Fan, Han Liu, and Weichen Wang. Large covariance estimation through elliptical factor models. *Annals of Statistics*, 46(4):1383, 2018a.

Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An l_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42, 2018b.

Jianqing Fan, Yuan Ke, Qiang Sun, and Wen-Xin Zhou. FarmTest: Factor-adjusted robust multiple testing with approximate false discovery control. *Journal of the American Statistical Association*, 114(528):1880–1893, 2019a.

Jianqing Fan, Weichen Wang, and Yiqiao Zhong. Robust covariance estimation for approximate factor models. *Journal of Econometrics*, 208(1):5–22, 2019b.

Jianqing Fan, Yuan Ke, and Yuan Liao. Augmented factor models with applications to validating market risk factors and forecasting bond risk premia. *Journal of Econometrics*, to appear, 2020a.

Jianqing Fan, Kaizheng Wang, Yiqiao Zhong, and Ziwei Zhu. Robust high dimensional factor models with applications to statistical machine learning. *Statistical Science*, to appear, 2020b.

Ethan X Fang, Yang Ning, and Runze Li. Test of significance for high-dimensional longitudinal data. *The Annals of Statistics*, 48(5):2622–2645, 2020.

Ulrich Franck, Siad Odeh, Alfred Wiedensohler, Birgit Wehner, and Olf Herbarth. The effect of particle size on cardiovascular disorders—the smaller the worse. *Science of the Total Environment*, 409(20):4217–4221, 2011.

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

Piotr Fryzlewicz and Suhasini Subba Rao. Multiple-change-point detection for auto-regressive conditional heteroscedastic processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):903–924, 2014.

Xuming He, Hengjian Cui, and Douglas G Simpson. Longitudinal data analysis using t-type regression. *Journal of Statistical Planning and Inference*, 122(1-2):253–269, 2004.

Cheng Hsiao. *Analysis of Panel Data*. Cambridge university press, 1986.

Cheng Hsiao. Panel data analysis—advantages and challenges. *Test*, 16(1):1–22, 2007.

Peter J Huber. Finite sample breakdown of m-and p-estimators. *The Annals of Statistics*, 12(1):119–126, 1984.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2 (1):193–218, 1985.

Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486): 682–693, 2009.

Ruth A Judson and Ann L Owen. Estimating dynamic panel data models: a guide for macroeconomists. *Economics Letters*, 65(1):9–15, 1999.

Yuan Ke, Jialiang Li, and Wenyang Zhang. Structure identification in panel data analysis. *The Annals of Statistics*, 44(3):1193–1233, 2016.

Yuan Ke, Stanislav Minsker, Zhao Ren, Qiang Sun, and Wen-Xin Zhou. User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science*, 34(3):454–471, 2019.

Yuan Ke, Heng Lian, and Wenyang Zhang. High-dimensional dynamic covariance matrices with homogeneous structure. *Journal of Business & Economic Statistics*, to appear, 2020.

Zheng Tracy Ke, Jianqing Fan, and Yichao Wu. Homogeneity pursuit. *Journal of the American Statistical Association*, 110(509):175–194, 2015.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107 (500):1590–1598, 2012.

Alois Kneip, Robin C Sickles, and Wonho Song. A new panel data treatment for heterogeneity in time trends. *Econometric Theory*, 28(3):590–628, 2012.

Lung-fei Lee and Jihai Yu. Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics*, 154(2):165–185, 2010.

Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

Xuan Liang, Tao Zou, Bin Guo, Shuo Li, Haozhe Zhang, Shuyi Zhang, Hui Huang, and Song Xi Chen. Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257, 2015.

Sydney C Ludvigson and Serena Ng. A factor analysis of bond risk premia. Technical report, National Bureau of Economic Research, 2009.

Shahar Mendelson and Nikita Zhivotovskiy. Robust covariance estimation under ℓ_4—ℓ_2 norm equivalence. *Annals of Statistics*, 48(3):1648–1664, 2020.

Stanislav Minsker. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903, 2018.

Stanislav Minsker and Xiaohan Wei. Robust modifications of u-statistics and applications to covariance estimation problems. *Bernoulli*, 26(1):694–727, 2020.

Stephen Nickell. Biases in dynamic models with fixed effects. *Econometrica*, 6(1981):1417–1426, 1981.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

M Hashem Pesaran. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012, 2006.

Greet Pison, Peter J Rousseeuw, Peter Filzmoser, and Christophe Croux. Robust factor analysis. *Journal of Multivariate Analysis*, 84(1):145–172, 2003.

Giuliano Polichetti, Stefania Cocco, Alessandra Spinali, Valentina Trimarco, and Alfredo Nunziata. Effects of particulate matter (PM10, PM2.5 and PM1) on the cardiovascular system. *Toxicology*, 261(1-2):1–8, 2009.

Vivian C Pun, Fatemeh Kazemiparkouhi, Justin Manjourides, and Helen H Suh. Long-term PM2.5 exposure and respiratory, cancer, and cardiovascular mortality in older US adults. *American Journal of Epidemiology*, 186(8):961–969, 2017.

Yiyuan She and Art B Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

Dan Shen, Haipeng Shen, Hongtu Zhu, and JS Marron. The statistics and mathematics of high dimension low sample size asymptotics. *Statistica Sinica*, 26(4):1747, 2016.

James H Stock and Mark W Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002.

Liangjun Su and Gaosheng Ju. Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics*, 206(2):554–573, 2018.

Liangjun Su, Zhentao Shi, and Peter CB Phillips. Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264, 2016.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, pages 1–24, 2019.

Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Lyudmila Yur'evna Vostrikova. Detecting "disorder" in multidimensional random processes. In *Doklady Akademii Nauk*, volume 259, pages 270–274. Russian Academy of Sciences, 1981.

Lili Wang, Chao Zheng, and Wen-Xin Zhou. A new principle for tuning-free huber regression. *Statistica Sinica*, to appear, 2020.

W. Wang and J. Fan. Asymptotics of empirical eigen-structure for high dimensional spiked covariance. *The Annals of Statistics*, 45:1342–1374, 2017.

Qiang Xia, Wangli Xu, and Lixing Zhu. Consistently determining the number of factors in multivariate volatility modelling. *Statistica Sinica*, pages 1025–1044, 2015.

Yu-Fei Xing, Yue-Hua Xu, Min-Hua Shi, and Yi-Xin Lian. The impact of PM2.5 on the human respiratory system. *Journal of Thoracic Disease*, 8(1):E69, 2016.

Jinfeng Xu, Mu Yue, and Wenyang Zhang. A new multilevel modelling approach for clustered survival data. *Econometric Theory*, 36(4):707—-750, 2020.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Mei Zheng, Lynn G Salmon, James J Schauer, Limin Zeng, CS Kiang, Yuanhang Zhang, and Glen R Cass. Seasonal trends in PM2.5 source contributions in Beijing, China. *Atmospheric Environment*, 39(22):3967–3976, 2005.

Wen-Xin Zhou, Koushiki Bose, Jianqing Fan, and Han Liu. A new perspective on robust m-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Annals of Statistics*, 46(5):1904, 2018.