Journal of Statistical Planning and Inference 234 (2025) 106196

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Robust Integrative Analysis via Quantile Regression with Homogeneity and Sparsity

Hao Zeng [a], Chuang Wan [b], Wei Zhong [c], Tuo Liu [c,*]

[a] Paula and Gregory Chow Institute for Studies in Economics, Xiamen University, Xiamen, Fujian, 361005, China
[b] Department of Statistics, School of Economics, Jinan University, Guangzhou 510632, China
[c] MOE Key Lab of Econometrics, WISE and Department of Statistics and Data Science in SOE, Xiamen University, Xiamen, Fujian, 361005, China

ARTICLE INFO

Keywords:
Homogeneity
Integrative analysis
Quantile regression
Robustness
Sparsity

ABSTRACT

Integrative analysis plays a critical role in integrating heterogeneous data from multiple datasets to provide a comprehensive view of the overall data features. However, in multiple datasets, outliers and heavy-tailed data can render least squares estimation unreliable. In response, we propose a Robust Integrative Analysis via Quantile Regression (RIAQ) that accounts for homogeneity and sparsity in multiple datasets. The RIAQ approach is not only able to identify latent homogeneous coefficient structures but also recover the sparsity of high-dimensional covariates via double penalty terms. The integration of sample information across multiple datasets improves estimation efficiency, while a sparse model improves model interpretability. Furthermore, quantile regression allows the detection of subgroup structures under different quantile levels, providing a comprehensive picture of the relationship between response and high-dimensional covariates. We develop an efficient alternating direction method of multipliers (ADMM) algorithm to solve the optimization problem and study its convergence. We also derive the parameter selection consistency of the modified Bayesian information criterion. Numerical studies demonstrate that our proposed estimator has satisfactory finite-sample performance, especially in heavy-tailed cases.

1. Introduction

Multiple datasets are frequently collected from heterogeneous populations across various domains or locations. Estimation biases may be often present if the potential heterogeneity across the datasets is ignored. On the other hand, analyzing each dataset independently has the risk of overlooking common attributes across multiple datasets, leading to estimation efficiency loss. Over the past a few years, researchers and practitioners have realized the importance of simultaneously addressing homogeneity and heterogeneity when analyzing multiple datasets. For instance, in personalized marketing, the identification of market segments characterized by unique individual purchasing behaviors, coupled with the revelation of the overall market structure, enhances stakeholder satisfaction (Wedel and Kamakura, 2000; Zhu et al., 2021). Similarly, in biomedical studies such as cancer research, studying similarities and differences in the genetic basis of various cancer subtypes yields more targeted insights (Ma et al., 2011; Liu et al., 2014). Therefore, the integration of sample information from multiple datasets is crucial in exploring heterogeneity and uncovering commonalities across latent subgroups.

* Corresponding author.
E-mail address: liutuo2017@xmu.edu.cn (T. Liu).

https://doi.org/10.1016/j.jspi.2024.106196
Received 4 August 2022; Received in revised form 18 January 2024; Accepted 30 May 2024
Available online 1 June 2024

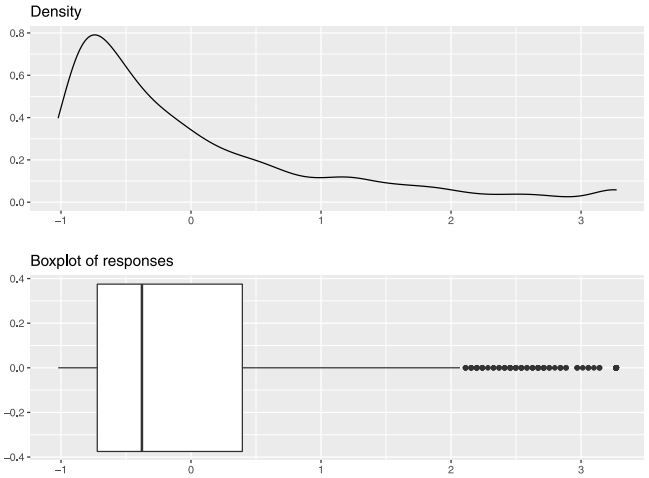

Fig. 1. Density of the total number of violent crimes per 100K population.

Meta-analysis, which traditionally analyzes each dataset separately and then combines the summary statistics across datasets, is a widely adopted method in multiple datasets analysis (Yang et al., 2010; Guerra and Goldstein, 2016). However, this approach yields suboptimal estimates, particularly concerning high-dimensional data, as it produces a large estimation error at the first step. The integrative analysis provides a more effective alternative, as it analyzes the raw data based on a joint model, allowing parameters to vary across different datasets (Zhao et al., 2015; Li et al., 2020). Recently, considerable attention has been paid to improving estimation or prediction accuracy in regression analysis by integrative analysis. For instance, Tang and Song (2016) developed a novel data integration technique to cluster homogeneous parameters without hypothesis testing. Huang et al. (2017) proposed an L_0 penalized method to promote similarity of model sparsity structures in integrative analysis for high-dimensional cancer genetic data. Dondelinger et al. (2020) proposed a joint lasso approach to identify global sparsity while permitting information sharing among subgroup-specific coefficients. Additional methods include sparse boosting techniques (Sun et al., 2020) and robust sparse quantile regression analysis (Li et al., 2022). Furthermore, Yang et al. (2019) considered high-dimensional integrative analysis based on least squares estimation.

However, as pointed out by Wang et al. (2012), high-dimensional data often suffer from heteroscedasticity and outliers, rendering least squares estimation unreliable. In our empirical example, the violent crime rate exhibits heavy tails and potential outliers, as shown in Fig. 1. To accommodate these issues, we propose a new robust integrative approach for quantile regression. This approach can simultaneously detect homogeneity, heterogeneity, and sparsity in a data-driven manner. Suppose a random sample comprises K datasets collected from different sources, where each dataset contains n_k observations. Let $y_{ik} \in \mathbb{R}$ be the scalar response and $\boldsymbol{x}_{ik} \in \mathbb{R}^p$ be the p-dimensional covariate for $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$. There are a total of $n = \sum_{k=1}^{K} n_k$ observations. For a given quantile level $\tau \in (0, 1)$, we consider the following linear model,

$$y_{ik} = \boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{\beta}_{k,\tau} + \varepsilon_{ik}^{(\tau)}, \tag{1.1}$$

where $\boldsymbol{\beta}_{k,\tau}$ is the regression coefficient for the kth subgroup, and $\varepsilon_{ik}^{(\tau)}$ is the error term satisfying $\Pr(\varepsilon_{ik}^{(\tau)} < 0 | \boldsymbol{x}_{ik}) = \tau$. We omit the dependence of $\boldsymbol{\beta}_{k,\tau}$ and $\varepsilon_{ik}^{(\tau)}$ on τ throughout this paper to avoid notational clutter. Due to potential heterogeneity, assuming that all datasets share the same coefficient is not reasonable. Conversely, assuming complete individual heterogeneity and allowing all $\boldsymbol{\beta}_k$'s to be different across the dataset ignores the latent homogeneity structure and leads to inefficient estimators. To balance subgroup-specific attributes and model parsimony, we impose a latent subgroup structure for the regression coefficients, where subgroup covariates within one cluster share the same value. Take the jth covariate as an example. We denote the $K \times 1$ vector of coefficients associated with the jth covariate as $\boldsymbol{\beta}^j = (\beta_{1j}, \ldots, \beta_{Kj})^{\mathsf{T}}$, which can be classified into several unknown subgroups $G_1^j, \ldots, G_{S_j}^j$ such that $G_s^j \cap G_{s'}^j = \emptyset$ whenever $s \neq s'$ and $\cup_{s=1}^{S_j} G_s^j = \{1, \ldots, K\}$. The coefficients in different groups are distinct, and $S_j (< K)$ is the number of distinct group effects for the jth covariate. In addition to the homogeneous and heterogeneous structures, we allow the existence of sparsity to adapt to high-dimensional settings.

In this paper, we propose a Robust Integrative Analysis via penalized Quantile regression (RIAQ) that accounts for homogeneity and sparsity in multiple datasets. Our method employs two types of penalty functions: a pairwise fusion penalty and a sparsity penalty on all $\boldsymbol{\beta}_k$'s. The pairwise fusion penalty identifies homogeneous and heterogeneous coefficients simultaneously by penalizing the difference between any pair of β_{kj}'s of the jth covariate, while the sparsity penalty selects non-zero coefficients. This approach has been extensively studied in various works, including Ma and Huang (2017), Zhu et al. (2021). Our proposed method is related to Yang et al. (2019), but we make several contributions. First, we use the quantile loss function, which is more robust in parameter estimation, by being more tolerant of heavy-tailed errors and outliers. Second, we weakened the Gaussian and sub-Gaussian error conditions required for establishing the oracle property, making our method more applicable to real-world data.

Third, we developed an ADMM algorithm for high-dimensional quantile to overcome the non-smoothness of the loss function and non-convexity introduced by the penalty function and proved its convergence. Additionally, we utilized an effective matrix inversion technique for computational efficiency. Finally, we used a modified Bayesian information criterion (BIC) combined with a data-driven procedure to determine the tuning parameters and investigate selection consistency thoroughly.

This article is organized as follows. We propose a doubly penalized quantile regression estimator in Section 2, which is solved by using the ADMM algorithm. The theoretical properties are presented in Section 3. In Section 4, we explore the finite-sample performance of our proposed method through simulation studies. Section 5 demonstrates the advantages of our method by an empirical study on the Communities and Crime Data Set. Section 6 concludes the paper. All proofs are included in the Appendix A. The code is available at https://github.com/zenghao-stat/RIAQ.

Notations. We define the Hadamard product operator "$\circ$" for a matrix $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_t)^\intercal$ and a vector $\mathbf{b} = (b_1, \ldots, b_t)^\intercal$ as $\mathbf{A} \circ \mathbf{b} = (\mathbf{a}_1 b_1, \ldots, \mathbf{a}_t b_t)^\intercal$. For two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimension, the Hadamard product $\mathbf{A} \circ \mathbf{B}$ denotes the matrix whose elements are the products of the corresponding elements of $\mathbf{A}$ and $\mathbf{B}$. Let $\|\mathbf{A}\|_F$ denote the Frobenius norm of $\mathbf{A}$, and $\mathbf{A}^+$ denotes a vector obtained by row sums of a matrix $\mathbf{A}$ and $\mathrm{vec}(\mathbf{A}) = (\mathbf{a}_1^\intercal, \ldots, \mathbf{a}_t^\intercal)^\intercal$. Let $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\intercal \mathbf{b}$ be the inner product of two vectors $\mathbf{a}$ and $\mathbf{b}$. "$\otimes$" represents the Kronecker product. Next, we define some notations related to our model. Let $\mathcal{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)^\intercal$ be a $K \times p$ coefficient matrix, $\mathbf{X} = (\mathbf{X}_{\cdot 1}^\intercal, \ldots, \mathbf{X}_{\cdot K}^\intercal)^\intercal$ be the $n \times p$ matrix of regressors where $\mathbf{X}_{\cdot k} = (X_{1k}, \ldots, X_{n_k k})^\intercal$ is an $n_k \times p$ design matrix and $\boldsymbol{y} = (\boldsymbol{y}_1^\intercal, \ldots, \boldsymbol{y}_K^\intercal)^\intercal$ be the $n \times 1$ dependent variable vector where $\boldsymbol{y}_k = (y_{1k}, \ldots, y_{n_k k})^\intercal$. Denote $\widetilde{\mathbf{X}} = \mathrm{diag}(\mathbf{X}_{\cdot 1}, \ldots, \mathbf{X}_{\cdot K})$ and $\mathcal{I} = \mathrm{diag}(\mathbf{1}_{n_1}^\intercal, \ldots, \mathbf{1}_{n_K}^\intercal)$ as two block-diagonal matrices where $\mathbf{1}_m$ is the $m \times 1$ vector whose elements are all 1. Let $\boldsymbol{\Omega} = \mathcal{E} \otimes \mathbf{I}_p$, where $\mathbf{I}_p$ denotes the $p \times p$ identity matrix, and $\mathcal{E} = (\mathbf{e}_1 - \mathbf{e}_2, \mathbf{e}_1 - \mathbf{e}_3, \ldots, \mathbf{e}_1 - \mathbf{e}_K, \mathbf{e}_2 - \mathbf{e}_3, \ldots, \mathbf{e}_2 - \mathbf{e}_K, \ldots, \mathbf{e}_{K-1} - \mathbf{e}_K)^\intercal$ be a $\frac{K(K-1)}{2} \times K$ matrix with $\mathbf{e}_i$ being the $K \times 1$ vector with the ith element being 1 and 0 elsewhere.

2. Methodologies

2.1. Quantile regression based integrative analysis

Our approach to recovering the homogeneous, heterogeneous, and sparse structures for regression across multiple datasets is based on a double penalization idea. To ensure robustness in integrative analysis, we adopt quantile regression by Koenker and Bassett (1978). In addition to promoting robustness, the use of quantile regression allows for a more comprehensive understanding of the homogeneity, heterogeneity, and sparsity patterns across different quantile levels. Further discussions on high-dimensional quantile regression can be found in Wang et al. (2012) and Fan et al. (2014). We present a doubly penalized objective function given by

$$Q_n(\boldsymbol{\beta} | \lambda, \gamma) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \rho_\tau(y_{ik} - \boldsymbol{x}_{ik}^\intercal \boldsymbol{\beta}_k) + \sum_{j=1}^{p} \sum_{k<k'} p_\lambda(|\beta_{kj} - \beta_{k'j}|) + \sum_{j=1}^{p} \sum_{k=1}^{K} p_\gamma(|\beta_{kj}|), \tag{2.2}$$

where $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\intercal, \ldots, \boldsymbol{\beta}_K^\intercal\right)^\intercal$ is a $Kp \times 1$ vector, $\rho_\tau(u) = (\tau - \mathbf{1}\{u < 0\})u$ is the quantile loss function, and $p_\lambda(\cdot)$ and $p_\gamma(\cdot)$ are two penalty functions associated with different tuning parameters λ and γ. The fusion-type penalty function $p_\lambda(\cdot)$ encourages the grouping of pairwise unit-specific effects in the data-driven manner without knowing any prior grouping information of β_{kj}'s of the jth covariate. We remark that this fusion-type penalty is different from the well-known group lasso penalty (Yuan and Lin, 2006; Meier et al., 2008) which requires the prior knowledge of the group structure of variables. The second sparsity-induced penalty function $p_\gamma(\cdot)$ is imposed to shrink the small regression coefficients to zeros and identify important covariates. The two tuning parameters λ and γ control the degree of shrinkage on the pairwise differences between any pair of β_{kj}'s of the jth covariate and the degree of shrinkage of each coefficients towards zero, respectively. λ and γ are allowed to take different values, enabling flexibility in controlling the shrinkage effects for heterogeneity and sparsity.

One of the key challenges in (2.2) is selecting an appropriate penalty function. For quantile regression, a natural choice is the L_1 penalty, $p_\lambda(\beta) = \lambda|\beta|$, such that the minimization can be easily formulated as a linear programming problem. However, it has been well-documented that the L_1 penalty tends to underestimate non-zero coefficients and does not guarantee variable selection consistency without proper assumptions. To address these limitations, we explore two alternatives: (1) the smoothly clipped absolute deviation (Fan and Li, 2001, SCAD) penalty,

$$p_\lambda(t) = \begin{cases} \lambda|t| & \text{if } |t| \le \lambda, \\ \frac{2a\lambda|t| - x^2 - \lambda^2}{2(a-1)} & \text{if } \lambda < |t| < a\lambda, \\ \frac{\lambda^2(a+1)}{2} & \text{if } |t| \ge a\lambda, \end{cases} \tag{2.3}$$

where $a > 2$ determines the concavity of the penalty function, and (2) the minimax concave penalty (Zhang, 2010, MCP),

$$p_\lambda(t) = \begin{cases} \lambda|t| - \frac{x^2}{2a}, & \text{if } |t| \le a\lambda, \\ \frac{1}{2}a\lambda^2, & \text{if } |t| > a\lambda, \end{cases} \tag{2.4}$$

for some $a > 1$. Given λ and γ, the estimator of $\boldsymbol{\beta}$ is defined as

$$\widehat{\boldsymbol{\beta}}(\lambda, \gamma) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{Kp}}{\mathrm{argmin}} \, Q_n(\boldsymbol{\beta} \mid \lambda, \gamma). \tag{2.5}$$

It should be noted that directly minimizing the penalized objective function (2.2) is computationally intractable, as the loss function is not smooth, the penalty function is not convex, and the pairwise fusion penalization term is non-separable concerning $\boldsymbol{\beta}_k$. In order to address these issues, we propose a modified ADMM algorithm to minimize (2.2).

2.2. Computation algorithm

ADMM is an optimization algorithm that is particularly useful for solving optimization problems with certain structures, such as those involving separable or block-separable objective functions (Boyd et al., 2011). It is able to decompose the optimization problems into simpler subproblems, enable parallelization and leverage the structure of the objective function for improved convergence. In our objective function (2.2), it is difficult to solve estimates $\widehat{\boldsymbol{\beta}}$ directly. In our algorithm, we first introduce a set of auxiliary parameters to reparameterize the objective function (2.2). The objective function is then equivalent to a constrained optimization problem and can be solved by the augmented Lagrangian method. Finally, we compute the respective estimates in the ADMM iterations. In details, by introducing some auxiliary parameters $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$ and $\mathbf{z}$, we reformulate the objective function as follows:

$$\begin{aligned}
&\widetilde{Q}_n(\boldsymbol{\alpha},\boldsymbol{\eta},\mathbf{z}) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\rho_\tau(z_{ik}) + \sum_{j=1}^{p}\sum_{k<k'}p_\lambda(|\alpha_{kk'}^j|) + \sum_{j=1}^{p}\sum_{k=1}^{K}p_\gamma(|\eta_{kj}|),\\
&\text{s.t.}\begin{cases}\beta_{kj}-\beta_{k'j}-\alpha_{kk'}^j=0, 1\le k<k'\le K, j=1,\dots,p,\\ \eta_{kj}-\beta_{kj}=0, 1\le k\le K, j=1,\dots,p,\\ \mathbf{y}_k-\mathbf{X}_{\cdot k}\boldsymbol{\beta}_k-\mathbf{z}_k=0, 1\le k\le K,\end{cases}
\end{aligned} \tag{2.6}$$

where $\boldsymbol{\alpha}=\{\alpha_{kk'}^j : k<k', j\in\{1,\dots,p\}\}_{\frac{K(K-1)}{2}\times p}$, $\boldsymbol{\eta}=\{\eta_{kj} : k\in\{1,\dots,K\}, j\in\{1,\dots,p\}\}_{K\times p}$ and $\mathbf{z}=(\mathbf{z}_1^\intercal,\dots,\mathbf{z}_K^\intercal)^\intercal$ is an $n\times 1$ vector with $\mathbf{z}_k=(z_{1k},\dots,z_{n_k k})^\intercal$. The corresponding augmented Lagrangian function is

$$\begin{aligned}
&\mathcal{Q}_{\rho_1,\rho_2,\rho_3}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\eta},\mathbf{z},\boldsymbol{\vartheta},\boldsymbol{\nu},\boldsymbol{\delta}) = \widetilde{Q}_n(\boldsymbol{\alpha},\boldsymbol{\eta},\mathbf{z}) + \sum_{j=1}^{p}\sum_{k<k'}\vartheta_{kk'}^j(\beta_{kj}-\beta_{k'j}-\alpha_{kk'}^j)\\
&+\frac{\rho_1}{2}\sum_{j=1}^{p}\sum_{k<k'}\left(\beta_{kj}-\beta_{k'j}-\alpha_{kk'}^j\right)^2 + \sum_{j=1}^{p}\sum_{k=1}^{K}\nu_{kj}(\eta_{kj}-\beta_{kj}) + \frac{\rho_2}{2}\sum_{j=1}^{p}\sum_{k=1}^{K}\left(\eta_{kj}-\beta_{kj}\right)^2\\
&+\sum_{k=1}^{K}\langle\boldsymbol{\delta}_k,\mathbf{y}_k-\mathbf{X}_{\cdot k}\boldsymbol{\beta}_k-\mathbf{z}_k\rangle + \frac{\rho_3}{2}\sum_{k=1}^{K}\|\mathbf{y}_k-\mathbf{X}_{\cdot k}\boldsymbol{\beta}_k-\mathbf{z}_k\|^2,
\end{aligned} \tag{2.7}$$

where the dual variables $\boldsymbol{\vartheta}=\{\vartheta_{kk'}^j : k<k', j\in\{1,\dots,p\}\}_{\frac{K(K-1)}{2}\times p}$, $\boldsymbol{\nu}=\{\nu_{kj} : k\in\{1,\dots,K\}, j\in\{1,\dots,p\}\}_{K\times p}$ and $\boldsymbol{\delta}=(\boldsymbol{\delta}_1^\intercal,\dots,\boldsymbol{\delta}_K^\intercal)^\intercal$ with $\boldsymbol{\delta}_k=(\delta_{1k},\dots,\delta_{n_k k})^\intercal$ are called Lagrange multipliers, ρ_i $(i=1,2,3)$ are the fixed augmented parameters. The ADMM algorithm aims to simplify the solution of a complex optimization problem by decomposing it into manageable subsets. In line with this, we adopt alternating minimization for each block of parameters, including $\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\eta},\mathbf{z},\boldsymbol{\vartheta},\boldsymbol{\nu}$, and $\boldsymbol{\delta}$. Subsequently, given the current values at the tth iteration, denoted as $\boldsymbol{\beta}^{(t)},\boldsymbol{\alpha}^{(t)},\boldsymbol{\eta}^{(t)},\mathbf{z}^{(t)},\boldsymbol{\vartheta}^{(t)},\boldsymbol{\nu}^{(t)},\boldsymbol{\delta}^{(t)}$, the ADMM computation at the $(t+1)$th iteration has an explicit closed form as follows.

First, $\boldsymbol{\beta}^{(t+1)}$ is the minimizer of $\mathcal{Q}_{\rho_1,\rho_2,\rho_3}(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t)},\boldsymbol{\eta}^{(t)},\mathbf{z}^{(t)},\boldsymbol{\vartheta}^{(t)},\boldsymbol{\nu}^{(t)},\boldsymbol{\delta}^{(t)})$. By several calculations,

$$\begin{aligned}
&\mathcal{Q}_{\rho_1,\rho_2,\rho_3}(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t)},\boldsymbol{\eta}^{(t)},\mathbf{z}^{(t)},\boldsymbol{\vartheta}^{(t)},\boldsymbol{\nu}^{(t)},\boldsymbol{\delta}^{(t)}) = \frac{\rho_1}{2}\|\mathcal{E}\mathcal{B}-\boldsymbol{\alpha}^{(t)}+\rho_1^{-1}\boldsymbol{\vartheta}^{(t)}\|_F^2\\
&+\frac{\rho_2}{2}\|\boldsymbol{\eta}^{(t)}-\mathcal{B}+\rho_2^{-1}\boldsymbol{\nu}^{(t)}\|_F^2 + \frac{\rho_3}{2}\sum_{k=1}^{K}\|\mathbf{y}_k-\mathbf{X}_{\cdot k}\boldsymbol{\beta}_k-\mathbf{z}_k^{(t)}+\rho_3^{-1}\boldsymbol{\delta}_k^{(t)}\|^2 + \text{constant}.
\end{aligned} \tag{2.8}$$

Since Eq. (2.8) is a quadratic function in $\boldsymbol{\beta}$, the minimizer $\boldsymbol{\beta}^{(t+1)}$ has a closed-form solution:

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{A}_1+\rho_3\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}})^{-1}\,\mathrm{vec}(\mathbf{A}_2^{\intercal(t)}), \tag{2.9}$$

where $\mathbf{A}_1=\rho_1\boldsymbol{\Omega}^\intercal\boldsymbol{\Omega}+\rho_2\mathbf{I}_{Kp}$ and $\mathbf{A}_2^{(t)}=\rho_3\mathcal{I}(\mathbf{X}\circ(\mathbf{y}-\mathbf{z}^{(t)}+\rho_3^{-1}\boldsymbol{\delta}^{(t)}))+\rho_1\mathcal{E}^\intercal(\boldsymbol{\alpha}^{(t)}-\rho_1^{-1}\boldsymbol{\vartheta}^{(t)})+\rho_2\boldsymbol{\eta}^{(t)}+\boldsymbol{\nu}^{(t)}$. In Eq. (2.9), the computation of the inverse of a matrix with dimensions $Kp\times Kp$ may impose a substantial computational burden. Fortunately, the inverse could be calculated using the Sherman–Morrison–Woodbury formula if $n<Kp$,

$$(\mathbf{A}_1+\rho_3\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}})^{-1} = \mathbf{A}_1^{-1}-\rho_3\mathbf{A}_1^{-1}\widetilde{\mathbf{X}}^\intercal(\mathbf{I}_n+\rho_3\widetilde{\mathbf{X}}\mathbf{A}_1^{-1}\widetilde{\mathbf{X}}^\intercal)^{-1}\widetilde{\mathbf{X}}\mathbf{A}_1^{-1},$$

where $\mathbf{A}_1=(\rho_1\mathcal{E}^\intercal\mathcal{E}+\rho_2\mathbf{I}_K)\otimes\mathbf{I}_p$ and $\mathcal{E}^\intercal\mathcal{E}=K\mathbf{I}_K-\mathbf{1}_K\mathbf{1}_K^\intercal$. Hence, $\mathbf{A}_1^{-1}=\mathbf{D}\otimes\mathbf{I}_p$, where

$$\mathbf{D}=[(K\rho_1+\rho_2)\mathbf{I}_K-\rho_1\mathbf{1}_K\mathbf{1}_K^\intercal]^{-1}=(\mathbf{I}_K+\rho_1\mathbf{1}_K\mathbf{1}_K^\intercal/\rho_2)/(K\rho_1+\rho_2)$$

by the Sherman–Morrison–Woodbury formula once again. Furthermore, by letting $\mathcal{A}=(\mathbf{I}_n+\rho_3(\mathbf{X}\mathbf{X}^\intercal)\circ\widetilde{\mathbf{D}})^{-1}$, where

$$\widetilde{\mathbf{D}}=\begin{pmatrix} D_{11}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^\intercal & \cdots & D_{1K}\mathbf{1}_{n_1}\mathbf{1}_{n_K}^\intercal\\ \vdots & \ddots & \vdots\\ D_{K1}\mathbf{1}_{n_K}\mathbf{1}_{n_1}^\intercal & \ddots & D_{KK}\mathbf{1}_{n_K}\mathbf{1}_{n_K}^\intercal\end{pmatrix},$$

and D_{ij} is the (i,j)th element of $\mathbf{D}$, we can update $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta}^{(t+1)} = \mathbf{D}[\mathbf{A}_2^{(t)} - \mathcal{I}\mathcal{B}^{(t)}], \tag{2.10}$$

where $\mathcal{B}^{(t)} = \rho_3 \mathbf{X} \circ \mathrm{vec}[\mathcal{A}\{\mathbf{X} \circ (\mathcal{I}^\intercal \mathbf{D}\mathbf{A}_2^{(t)})\}^+]$. Compared with (2.9), (2.10) avoids the calculation of a $Kp \times Kp$ matrix inverse, and we only need to invert an $n \times n$ matrix, which alleviates the computation burden, especially when $Kp > n$.

For $\boldsymbol{\alpha}^{(m+1)}$, it is the minimizer of

$$\frac{\rho_1}{2}\|\mathcal{E}\mathcal{B}^{(t)} + \rho_1^{-1}\boldsymbol{\vartheta}^{(t+1)} - \boldsymbol{\alpha}\|_F^2 + \sum_{j=1}^{p}\sum_{k<k'} p_\lambda(|\alpha_{kk'}^j|). \tag{2.11}$$

The updated $\boldsymbol{\alpha}^{(m+1)}$ with MCP is given by

$$\alpha_{kk'}^{j(t+1)} = \begin{cases} \theta_{kk'}^{j(t+1)} & \text{if } |\theta_{kk'}^{j(t+1)}| \geq a\lambda, \\ \frac{a\rho_1}{a\rho_1 - 1}\left(1 - \frac{\lambda}{\rho_1|\theta_{kk'}^{j(t+1)}|}\right)_+ \theta_{kk'}^{j(t+1)} & \text{if } |\theta_{kk'}^{j(t+1)}| < a\lambda, \end{cases} \tag{2.12}$$

where

$$\theta_{kk'}^{j(t+1)} \equiv \beta_{kj}^{(t+1)} - \beta_{k'j}^{(t+1)} + \rho_1^{-1}\vartheta_{kk'}^{j(t)}. \tag{2.13}$$

For the updation with SCAD penalty, it could be obtain using Eq. (7) in Ma and Huang (2017).

For $\eta_{kj}^{(t+1)}$, it is the minimizer of

$$\frac{\rho_2}{2}\|\boldsymbol{\beta}^{(t+1)} - \rho_2^{-1}\boldsymbol{\nu}^{(t)} - \boldsymbol{\eta}\|_F^2 + \sum_{j=1}^{p}\sum_{k=1}^{K} p_\gamma(|\eta_{kj}|). \tag{2.14}$$

The updation with MCP is similar to $\alpha_{kk'}^{j(t+1)}$, where $\theta_{kk'}^{j(t+1)}$ is replaced by $\xi_{kj}^{(t+1)} \equiv \beta_{kj}^{(t+1)} - \rho_2^{-1}\nu_{kj}^{(t)}$, i.e.,

$$\eta_{kj}^{(t+1)} = \begin{cases} \xi_{kj}^{(t+1)} & \text{if } |\xi_{kj}^{(t+1)}| \geq a\gamma, \\ \frac{a\rho_2}{a\rho_2 - 1}\left(1 - \frac{\gamma}{\rho_2|\xi_{kj}^{(t+1)}|}\right)_+ \xi_{kj}^{(t+1)} & \text{if } |\xi_{kj}^{(t+1)}| < a\gamma. \end{cases} \tag{2.15}$$

For $\mathbf{z}^{(t+1)}$, it is the minimizer of

$$\frac{\rho_3}{2}\sum_{k=1}^{K}\|\boldsymbol{y}_k - \mathbf{X}_{\cdot k}\boldsymbol{\beta}_k^{(t+1)} + \rho_3^{-1}\boldsymbol{\delta}_k^{(t)} - \mathbf{z}_k\|^2 + \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\rho_\tau(z_{ik}). \tag{2.16}$$

By Lemma 1 of Gu et al. (2018), $\mathbf{z}^{(t+1)}$ also has an closed form:

$$\begin{aligned} z_{ik}^{(t+1)} &= \arg\min_{z_{ik}} \frac{\rho_3}{2}(y_{ik} - \boldsymbol{x}_{ik}^\intercal\boldsymbol{\beta}_k^{(t+1)} + \rho_3^{-1}\delta_{ik}^{(t)} - z_{ik})^2 + \rho_\tau(z_{ik}) \\ &= \mathrm{Prox}_{\rho_\tau}(y_{ik} - \boldsymbol{x}_{ik}^\intercal\boldsymbol{\beta}_k^{(t+1)} + \rho_3^{-1}\delta_{ik}^{(t)}, n\rho_3), \end{aligned} \tag{2.17}$$

where the associated proximal operator is defined as

$$\mathrm{Prox}_{\rho_\tau}(\xi, \alpha) = \begin{cases} \xi - \frac{\tau}{\alpha}, & \text{if } \xi > \frac{\tau}{\alpha} \\ 0, & \text{if } \frac{\tau-1}{\alpha} \leq \xi \leq \frac{\tau}{\alpha} \\ \xi - \frac{\tau-1}{\alpha}, & \text{if } \xi < \frac{\tau-1}{\alpha}. \end{cases}$$

Finally, for $\boldsymbol{\vartheta}^{(t+1)}$, $\boldsymbol{\nu}^{(t+1)}$, and $\boldsymbol{\delta}^{(t+1)}$, they are updated by

$$\boldsymbol{\vartheta}^{(t+1)} = \boldsymbol{\vartheta}^{(t)} + \rho_1(\mathcal{E}\mathcal{B}^{(t+1)} - \boldsymbol{\alpha}^{(t+1)}), \tag{2.18}$$

$$\boldsymbol{\nu}^{(t+1)} = \boldsymbol{\nu}^{(t)} + \rho_2(\boldsymbol{\eta}^{(t+1)} - \boldsymbol{\beta}^{(t+1)}), \tag{2.19}$$

$$\boldsymbol{\delta}^{(t+1)} = \boldsymbol{\delta}^{(t)} + \rho_3(\boldsymbol{y} - \widetilde{\mathbf{X}}^\intercal\boldsymbol{\beta}^{(t+1)} - \mathbf{z}^{(t+1)}). \tag{2.20}$$

As suggested by Boyd et al. (2011), the ADMM algorithm is terminated when the primal residuals $\|\mathbf{r}^{(t+1)}\|_F = \|\mathcal{E}\mathcal{B}^{(t+1)} - \boldsymbol{\alpha}^{(t+1)}\|_F + \|\boldsymbol{\eta}^{(t+1)} - \boldsymbol{\beta}^{(t+1)}\|_F + \sum_{k=1}^{K}\|\boldsymbol{y}_k - \mathbf{X}_{\cdot k}^\intercal\boldsymbol{\beta}_k^{(t+1)} - \mathbf{z}_k^{(t+1)}\|$ is sufficiently small, i.e., $\|\mathbf{r}^{(t+1)}\|_F \leq e^{\mathrm{pri}}$, where e^{pri} is a pre-specified feasible primal tolerance. We summarize our robust integrative analysis via quantile (RIAQ) algorithm for given tuning parameters λ and γ in Algorithm 1.

2.3. Initial estimator and tuning parameters

In the context of the non-convex optimization problem (2.5), an appropriate initial value is crucial not only in producing an optimal solution but also in accelerating the iterations. We set $\boldsymbol{\beta}^{(0)}$ as the penalized quantile estimator proposed by Wang et al.

Algorithm 1: ADMM algorithm for RIAQ(λ, γ).

Step 1. Initialize $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\alpha}^{(0)}, \boldsymbol{\eta}^{(0)}, \mathbf{z}^{(0)}, \boldsymbol{\vartheta}^{(0)}, \boldsymbol{\nu}^{(0)}$, $\boldsymbol{\delta}^{(0)}$, given λ and γ.
for $t = 0, 1, 2, \cdots,$ **do**
Step 2.1. update $\boldsymbol{\beta}^{(t+1)}$ based on (2.10).
Step 2.2. update $\boldsymbol{\alpha}^{(t+1)}$, $\boldsymbol{\eta}^{(t+1)}$ based on (2.12).
Step 2.3. update $\mathbf{z}^{(t+1)}$ based on (2.17).
Step 2.4. update $\boldsymbol{\vartheta}^{(t+1)}$, $\boldsymbol{\nu}^{(t+1)}$ and $\boldsymbol{\delta}^{(t+1)}$ based on (2.18), (2.19) and (2.20).
Step 2.6. if the convergence criteria are met, stop and denote the last iteration by $\widehat{\boldsymbol{\beta}}$; otherwise, let $t = t + 1$.
end for

(2012). For the kth dataset, the initial value is computed as

$$\boldsymbol{\beta}_k^{(0)} = \arg\min_{\boldsymbol{\beta}_k} \sum_{i=1}^{n_k} \rho_\tau(y_{ik} - \mathbf{x}_{ik}^\top \boldsymbol{\beta}_k) + \sum_{j=1}^{p} p_\gamma(|\beta_{kj}|),$$

where $p_\gamma(\cdot)$ is the penalty function. We can use the function "`rq.lasso.fit`" in the R package "`rqPen`" to obtain a suitable sparse initial estimation.

To address the issue of selecting tuning parameters λ and γ, we employ a modified BIC-type criterion function. In the literature, Wang and Leng (2007), Wang et al. (2007, 2009), Chen and Chen (2008), and Fan and Tang (2013) have recommended the use of the Bayesian information criterion (BIC, Schwarz, 1978) for selecting the shrinkage parameter in penalized estimation and established the model selection consistency of the related penalized procedure. In the case of quantile regression, Lian (2012) studied its consistency in model selection when the number of variables p is finite, while Lee et al. (2014) proposed a modified version of BIC when the number of variables p is diverging. For subgroup analysis and data integration, Ma and Huang (2017), Tang and Song (2016), Tang et al. (2020), Tang and Song (2020), Yang et al. (2019), and Zhang et al. (2019) employed various modified BIC criteria. For the high-dimensional linear quantile model with potential heterogeneity and sparsity in our framework, we define a modified BIC-type criterion function for selecting the tuning parameters λ and γ as follows

$$BIC\{\widehat{\boldsymbol{\beta}}(\lambda, \gamma)\} = \log\left[\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} \rho_\tau(y_{ik} - \mathbf{x}_{ik}^\top \boldsymbol{\beta}_k)\right] + \varphi_n \widehat{S}_{\lambda,\gamma}, \tag{2.21}$$

where $\widehat{S}_{\lambda,\gamma}$ denotes the number of distinct non-zero coefficients of $\widehat{\boldsymbol{\beta}}(\lambda, \gamma)$, and $\varphi_n = n^{-1} c \log\left[\log(Kp)\right] \log(n)$. Here, a constant c serves as a balance between the degrees of freedom and the quantile fitting errors in practice while maintaining the order of φ_n. Wang et al. (2009) and Yang et al. (2019) simply used $c = 1$ in their numerical experiments while Zhang et al. (2019) considered different values of $c = 1, 2, \ldots, 5$. We run the simulations to examine the finite sample performance of the modified BIC criterion with different choices of c and the simulation results presented in the appendix show that the performance is not sensitive to the choice of c. In practice, one may set $c = 1$ for simplicity. Similar to the strategy proposed by Tang and Song (2016, 2020), we adopt the warm-start approach to obtain the solution path for a sequence of grid points (λ, γ). Specifically, we start with $\lambda = 0$ and $\gamma = 0$, compute $\widehat{\boldsymbol{\beta}}(\lambda, \gamma)$, and use it as the initial value of the iterative computation for the next tuple (λ, γ) according to a prespecified set of grid points of tuning parameters. In order to save computational resources, it is common practice to initially use coarse-grained grids to select an optimal point. Subsequently, finer grids are employed in the vicinity of this optimal point, facilitating the selection of even more refined results.

3. Theoretical results

3.1. Convergence of the algorithm

Under some regular conditions, the convergence of the ADMM algorithm is guaranteed. For the iterated sequence $\{\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\eta}^{(t)}, \mathbf{z}^{(t)}, \boldsymbol{\vartheta}^{(t)}, \boldsymbol{\nu}^{(t)}, \boldsymbol{\delta}^{(t)}\}$, there exists a global minimizer $\{\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\eta}^*, \mathbf{z}^*, \boldsymbol{\vartheta}^*, \boldsymbol{\nu}^*, \boldsymbol{\delta}^*\}$, and it satisfies the first order condition for a stationary point. We establish the convergence of the proposed ADMM algorithm in the following proposition.

Proposition 1. *For the iterated sequence $\{\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\eta}^{(t)}, \mathbf{z}^{(t)}, \boldsymbol{\vartheta}^{(t)}, \boldsymbol{\nu}^{(t)}, \boldsymbol{\delta}^{(t)}\}$, if sequences $\{\boldsymbol{\alpha}^{(t)}\}_{t=1}^{\infty}$, $\{\boldsymbol{\eta}^{(t)}\}_{t=1}^{\infty}$ and $\{\mathbf{z}^{(t)}\}_{t=1}^{\infty}$ are all bounded and $\rho_1^{-1}\| \triangle \boldsymbol{\vartheta}^{(t)}\|_F + \rho_2^{-1}\| \triangle \boldsymbol{\nu}^{(t)}\|_F + \rho_3^{-1}\| \triangle \boldsymbol{\delta}^{(t)}\| \to 0$ as $t \to +\infty$, then*

(1) $\{\boldsymbol{\beta}^{(t)}, \boldsymbol{\vartheta}^{(t)}, \boldsymbol{\nu}^{(t)}, \boldsymbol{\delta}^{(t)}\}_{t=1}^{\infty}$ is bounded;
(2) There exists a subsequence $\{\boldsymbol{\beta}^{(t_k)}, \boldsymbol{\alpha}^{(t_k)}, \boldsymbol{\eta}^{(t_k)}, \mathbf{z}^{(t_k)}, \boldsymbol{\vartheta}^{(t_k)}, \boldsymbol{\nu}^{(t_k)}, \boldsymbol{\delta}^{(t_k)}\}_{k=1}^{\infty}$ s.t. $\| \triangle \boldsymbol{\beta}^{(t_k)}\| + \| \triangle \boldsymbol{\alpha}^{(t_k)}\| + \| \triangle \boldsymbol{\eta}^{(t_k)}\| + \| \triangle \mathbf{z}^{(t_k)}\| + \| \triangle \boldsymbol{\vartheta}^{(t_k)}\|_F + \| \triangle \boldsymbol{\nu}^{(t_k)}\|_F + \| \triangle \boldsymbol{\delta}^{(t_k)}\| \to 0$ as $k \to \infty$;
(3) $\{\boldsymbol{\beta}^{(t_k)}, \boldsymbol{\alpha}^{(t_k)}, \boldsymbol{\eta}^{(t_k)}, \mathbf{z}^{(t_k)}, \boldsymbol{\vartheta}^{(t_k)}, \boldsymbol{\nu}^{(t_k)}, \boldsymbol{\delta}^{(t_k)}\}$ converges to a stationary point satisfying the KKT condition of (2.7).

We note that Shen et al. (2014), Gu et al. (2018), You et al. (2019), and Yang et al. (2019) have employed similar conditions as in Proposition 1. The detailed proofs are postponed to Appendix B.

3.2. Oracle properties

In literature, SCAD penalty and MCP's oracle properties are established in the penalized least squares or quantile regression (Fan and Li, 2001; Fan and Peng, 2004; Kim et al., 2008; Fan and Lv, 2008; Wang et al., 2012). In our paper, we find that the oracle estimator is a local solution of the doubly penalized quantile regression (2.2) and obtain the oracle properties under weaker conditions.

Let $\mathcal{G} = \{\mathcal{G}^j, j = 1, \dots, p\}$ denote a given structure, where $\mathcal{G}^j = \{G_s^j,\ s = 1, \dots, S_j\}$ and $S_j (< K)$ is the number of distinct group effects for the jth covariate. We define $\boldsymbol{\phi} = (\phi_1^1, \dots, \phi_{S_1}^1, \dots, \phi_1^p, \dots, \phi_{S_p}^p)^\intercal$, where $\phi_s^j = \beta_{kj}$ for $k \in G_s^j$. For any given coefficients vector $\boldsymbol{\beta}$, there exists a unique corresponding vector, denoted as $\boldsymbol{\phi}$, with the structure $\mathcal{G}$ such that $\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\phi}$, where $\mathbf{W}$ is a well-defined matrix. Specifically, let $S_0 = 0$, and let the kth unit's selection matrix be denoted by $\mathbf{W}_k = \{w_{k,js}\} \in \mathbb{R}^{p\times S}$, where $w_{k,js} = 1$ if $s = \sum_{m=0}^{j-1} S_m + s'$, s' s.t. $k \in G_{s'}^j$, and $w_{k,js} = 0$ otherwise. Define $\mathbf{W} = (\mathbf{W}_1^\intercal, \dots, \mathbf{W}_K^\intercal)^\intercal \in \mathbb{R}^{Kp\times S}$. Then, the model (1.1) can be written as

$$\boldsymbol{y} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \widetilde{\mathbf{X}}\mathbf{W}\boldsymbol{\phi} + \boldsymbol{\varepsilon} \equiv \mathbb{X}\boldsymbol{\phi} + \boldsymbol{\varepsilon},$$

where $\mathbb{X} = \widetilde{\mathbf{X}}\mathbf{W} \in \mathbb{R}^{n\times S}$ is the structured design matrix. Therefore, given selection matrix $\mathbf{W}$, the dimension of a candidate model is $S = \sum_{j=1}^p S_j$, and $S \in [p, Kp]$ is allowed to increase with n or be fixed.

The vector $\boldsymbol{\phi}$ could be rearranged as $\boldsymbol{\phi} = \left(\boldsymbol{\phi}_1^\intercal, \boldsymbol{\phi}_2^\intercal\right)^\intercal$, where $\boldsymbol{\phi}_1 \in \mathbb{R}^q$ and $\boldsymbol{\phi}_2 \in \mathbb{R}^{S-q}$ correspond to the nonzero and zero components of $\boldsymbol{\phi}$, respectively. Here, q represents the number of nonzero components, which is allowed to increase with the sample size n. Similarly, $\mathbb{X}$ can be divided into an $n \times q$ submatrix $\mathbb{X}^{(1)}$ and an $n \times (S-q)$ submatrix $\mathbb{X}^{(2)}$ corresponding to $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$, respectively. The true coefficients can be defined as $\boldsymbol{\beta}_0 = \mathbf{W}_0\boldsymbol{\phi}_0$, where $\boldsymbol{\phi}_0 = \left(\boldsymbol{\phi}_{01}^\intercal, \mathbf{0}_{S-q}^\intercal\right)^\intercal$ and $\mathbf{W}_0$ is the true selection matrix. The oracle quantile estimator $\widetilde{\boldsymbol{\phi}} = (\widetilde{\boldsymbol{\phi}}_1^\intercal, \mathbf{0}_{S-q}^\intercal)^\intercal$ is estimated under $\mathbf{W}_0$, namely,

$$\widetilde{\boldsymbol{\phi}}_1 = \underset{\boldsymbol{\phi}_1 \in \mathbb{R}^q}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \rho_\tau(y_i - \mathbb{X}_i^{(1)\intercal}\boldsymbol{\phi}_1), \tag{3.22}$$

where y_i and the transpose of $\mathbb{X}_i^{(1)}$ represent the ith element and the ith row of $\boldsymbol{y}$ and $\mathbb{X}^{(1)}$, respectively. Furthermore, let $|\mathcal{G}| = \sum_{j\in\{1,\dots,p\}, s\in\{1,\dots,S_j\}} \left|G_s^j\right|$, where $\left|G_s^j\right|$ denotes the cardinality of G_s^j, and $|\mathcal{G}_{\min}|$ and $|\mathcal{G}_{\max}|$ denote the true minimum and maximum group sizes among all the subgroups.

To establish the oracle property of the estimator (3.22), we impose the following regularity conditions throughout this paper.

C1. There exist two constants $M_1, M_2 > 0$ not depending on n such that $|\mathbb{X}_{ij}| \le M_1$ for all $j = q+1, \dots S, i = 1, \dots, n$, and $\frac{1}{n}\mathbb{X}_j^\intercal\mathbb{X}_j \le M_2$, for $j = 1, \dots, S$, where $\mathbb{X}_j$ is the jth column of $\mathbb{X}$. And there exist two constants $C_1, C_2 > 0$ such that $C_1 \le \lambda_{\min}(\frac{1}{n}\mathbb{X}^{(1)\intercal}\mathbb{X}^{(1)}) < \lambda_{\max}(\frac{1}{n}\mathbb{X}^{(1)\intercal}\mathbb{X}^{(1)}) \le C_2$ with probability approaching one, where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum eigenvalues, respectively.

C2. The conditional distribution of ε_i has a continuous density function $f(\cdot|\mathbb{X})$ which is uniformly bounded away from zero and infinity in a neighborhood of zero for all i.

C3. $q = \mathbf{O}(n^{c_1})$ for some $0 \le c_1 < 1/2$.

C4. There exist two constants c_2 and $M_3 > 0$, s.t. $2c_1 < c_2 \le 1$ and $n^{(1-c_2)/2}b_n \ge M_3$, where

$$b_n = \min\left\{ \min_{1\le j\le q}\{|\phi_{0,j}|\}, \min_{1\le j\le p, k\in G_s^j, k'\in G_{s'}^j \text{ with } s\ne s'}\{|\beta_{0,kj} - \beta_{0,k'j}|\} \right\}.$$

Assumptions C1–C4 are commonly utilized in high-dimensional inference literature, as documented in Wang et al. (2012). Specifically, Assumption C1 imposes a condition of uniform boundedness on the design matrix. Similarly, Assumption C2 signifies a standard condition for quantile regression literature, without requiring any Gaussian or sub-Gaussian distribution condition for the error term. Moreover, Assumption C3 allows q to augment the sample size. Last, Assumption C4 ensures that the minimum signal does not vanish too fast.

Theorem 1. *Under conditions C1–C4, let $\mathcal{B}_n(\lambda, \gamma)$ be the set of minimizers of (2.2) with penalty functions (SCAD or MCP) and tuning parameters (λ, γ). Suppose $\lambda = o(n^{-(1-c_2)/2})$, $\gamma = o(n^{-(1-c_2)/2})$, and $n\lambda\frac{|\mathcal{G}_{\min}|}{|\mathcal{G}_{\max}|} \to \infty$. Then*

$$P\left\{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_n(\lambda, \gamma)\right\} \to 1 \text{ as } n \to \infty,$$

where $\widetilde{\boldsymbol{\beta}}$ is the oracle estimator s.t. $\widetilde{\boldsymbol{\beta}}^\intercal = \mathbf{W}_0\widetilde{\boldsymbol{\phi}}$.

Theorem 1 shows that the oracle estimator is a local minimizer of the penalized quantile objective function (2.2) with a large probability. It implies that we can find a pair of (λ_0, γ_0) s.t. $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\lambda_0, \gamma_0)$. Next, we study the selection consistency of the modified BIC criterion in (2.21). For any candidate model $\mathcal{G}$, define a new BIC criterion as

$$BIC(\mathcal{G}) = \log\left\{ \frac{1}{n}\sum_{k=1}^K\sum_{i=1}^{n_k} \rho_\tau\left(y_{ik} - \boldsymbol{x}_{ik}^\intercal\widehat{\boldsymbol{\beta}}_k(\mathcal{G})\right)\right\} + \varphi_n S, \tag{3.23}$$

where $\widehat{\boldsymbol{\beta}}_k(\mathcal{G})$ is the *unpenalized* estimator given $\mathcal{G}$, and φ_n is a sequence of positive numbers that goes to zero. Let $\mathcal{G}_0$ be the true structure. Note that Eq. (3.23) is different from Eq. (2.21), as the latter is based on the penalized estimator $\widehat{\boldsymbol{\beta}}_k(\lambda,\gamma)$.

Theorem 2. *Under conditions C1–C4, and C2+ and C5 given in Appendix D, for any sequence $\varphi_n \to 0$ satisfying $\log(Kp)/n = o\left(\varphi_n\right)$, we have*

$$P\left(\inf_{\mathcal{G}\neq\mathcal{G}_0,|\mathcal{G}|<S_U} BIC(\mathcal{G}) > BIC\left(\mathcal{G}_0\right)\right) \to 1,$$

where $S_U \in \left(S_0,\infty\right)$ is an upper bound of the total number of subgroups.

For any given tuning parameters (λ,γ), the penalized BIC, $BIC(\widehat{\boldsymbol{\beta}}(\lambda,\gamma))$ is greater than or equal to $BIC(\widehat{\mathcal{G}}_{\lambda,\gamma})$, where $\widehat{\mathcal{G}}_{\lambda,\gamma}$ represents the estimated structure under the corresponding tuning parameters. According to Theorem 1, the oracle estimator $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\lambda_0,\gamma_0)$ can be generated by (λ_0,γ_0), such that $BIC\{\widehat{\boldsymbol{\beta}}(\lambda_0,\gamma_0)\} = BIC(\mathcal{G}_0)$. Furthermore, for those tuning parameters (λ,γ) that do not include the oracle model, $BIC\{\widehat{\boldsymbol{\beta}}(\lambda,\gamma)\} \geq BIC(\widehat{\mathcal{G}}_{\lambda,\gamma}) > BIC(\mathcal{G}_0) = BIC\{\widehat{\boldsymbol{\beta}}(\lambda_0,\gamma_0)\}$ with probability approaching 1. As a result, the selection consistency is achieved. We remark that the validity of this tuning selection criterion requires $\log(Kp)/n = o\left(\varphi_n\right)$ in Theorem 2, which allows $p = O(n^\kappa)$ for some constant $\kappa > 0$, which means that the proposed method can deal with the high dimensionality where p can increase at a polynomial order of n.

4. Simulation studies

In this section, we evaluate the finite sample performance of the proposed robust integrative analysis method for linear sparse quantile regression. Specifically, we investigate (i) the performance of our robust method at the median level and the mean regression-based method (Yang et al., 2019) denoted as YYH, and (ii) the performance of our method at tail quantile levels. We employ three different metrics to measure the performance of each method. First, we use the root mean square error (RMSE) to quantify the estimation accuracy of the regression coefficients, defined as $\sqrt{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^\intercal(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)}$. Second, we use the rand index (RI) to measure the difference between the detected structure and the true structure. It takes values in the range $[0,1]$, with higher values indicating greater similarity. Finally, we use the average model size (AMS), defined as the number of unique non-zero estimated coefficients. To ensure reliable results, we run each experiment 500 times with simulation. In our simulations, we use the grid points as $\lambda_i = 10^{-2.5+0.25i}$ and $\gamma_i = 10^{-3.1+0.31i}$ for $i = 0,\ldots,10$ to tuning parameters selection for simplicity.

Example 1. In this example, we investigate a simple scenario with two kinds of slope coefficients: $\boldsymbol{\beta}_1 = (-1,2,\mathbf{0}_{p-2}^\intercal)^\intercal$ and $\boldsymbol{\beta}_2 = (1,2,\mathbf{0}_{p-2}^\intercal)^\intercal$. The covariates $\mathbf{X}_k$ are generated from a multivariate Gaussian distribution with variance 1 and correlation coefficient 0.3 between any pair. The random errors are independently drawn from one of three probability distributions, including Case 1: $N(0,1)$, Case 2: $t(2)$ (the Student's t distribution with 2 degrees of freedom), and Case 3: mixture distribution $0.9N(0,1)+0.1\cdot Cauchy(0,1)$. Cases 2 and 3 aim to examine the detection performance of heavy-tailed errors. The number of individuals for all units is equal, i.e., $n_1 = n_2 = \cdots = n_K$ where $K = 10$. We consider two setups: $(n_k,p) = (40,200)$ or $(40,400)$. For the first covariate, we assume that the numbers of units in two subgroups are equivalent, i.e., $|\mathcal{G}_1^1| = |\mathcal{G}_2^1| = K/2$ and $S_1 = 2$. For the remaining covariates, all the units share the same coefficient, i.e., $\mathcal{G}_1^j = 10$ and $S_j = 1$ for $j \in \{2,\ldots,p\}$.

In this simulation study, we first compare the performance of the proposed RIAQ estimator at quantile levels $\tau = 0.5$, and that of the mean-based estimator of Yang et al. (2019) (denoted as YYH). The comparison is conducted based on the RMSE, RI1, RI2, and AMS metrics, averaged across 500 repetitions for Cases 1–3, where RI1, RI2 are the rand indices of the first and the second covariates. The MCP with tuning parameter $a = 3$ is used here. The SCAD penalty gives similar results. We use the BIC proposed in Section 2.3 to select the tuning parameters λ and γ. As is discussed in Section 2.3, the finite sample performance of the proposed method is not sensitive to the choice of c. For simplicity, we set $c = 1$ here. The results in Table 1 indicate that both estimators perform well for Case 1 with standard normal errors, with YYH having a slightly smaller RMSE than the quantile regression-based method. However, for data with heavy-tailed errors like Cases 2 and 3, our method consistently outperforms YYH in terms of all three metrics. Specifically, the AMS values of RIAQ are always close to the ideal value of 3 with a smaller RMSE, indicating that our method can effectively integrate the sparsity and homogeneity structure across multiple datasets. Moreover, the RI1 and RI2 approaches 100% greater than the least squares method, demonstrating the superiority of our robust method based on quantile regression.

To further analyze the performance of our method at different quantile levels, we also conduct simulations at quantiles, $\tau = 0.75$, and 0.9. The results in Table 1 show that the proposed method still recovers the latent model structure with an satisfactory error level, indicating that our method is robust and reliable across different quantile levels.

Example 2. We present another simulation study with more subgroups. Similar to Example 1, we consider $K = 10$ units with an equal number of observations. The true coefficient vectors for the first, second, and third covariates across the K units are $\boldsymbol{\beta}^1 = (-2,-2,-2,0,0,0,0,2,2,2)^\intercal$, $\boldsymbol{\beta}^2 = \mathbf{1}_{10}$, and $\boldsymbol{\beta}^3 = (-2,-2,-2,-2,2,2,2,2,0,0)$, respectively. For $j \in \{4,\ldots,p\}$, we set $\boldsymbol{\beta}^j = \mathbf{0}_{10}$. The estimation results of Example 2 under different error settings and sample sizes are presented in Table 2. Our method outperforms the mean-based method for heavy-tailed cases, as evidenced by the competitive RMSE and RI at the median level. Moreover, the RIs for the first three covariates are very close to 100%. Overall, our method can simultaneously recover homogeneity and sparsity, making it advantageous for data with outliers and heavy-tailed errors.

Table 1
Simulation results by RIAQ and YHH for Example 1 based on 500 replications.

Case	p	Method	RMSE	AMS	RI1	RI2
1	200	YHH	0.0071	3.0100	99.42%	100.00%
		RIAQ, $\tau = 0.5$	0.0132	3.0135	98.26%	99.87%
		RIAQ, $\tau = 0.75$	0.0109	2.9966	99.39%	100.00%
		RIAQ, $\tau = 0.9$	0.0308	2.9670	96.70%	99.55%
	400	YHH	0.0054	3.0300	99.08%	100.00%
		RIAQ, $\tau = 0.5$	0.0084	3.0322	99.62%	100.00%
		RIAQ, $\tau = 0.75$	0.0069	3.0345	99.96%	100.00%
		RIAQ, $\tau = 0.9$	0.0219	3.0971	95.51%	99.47%
2	200	YHH	0.0351	4.2273	86.64%	100.00%
		RIAQ, $\tau = 0.5$	0.0249	2.8372	93.52%	99.07%
		RIAQ, $\tau = 0.75$	0.0179	2.9400	98.12%	100.00%
		RIAQ, $\tau = 0.9$	0.0388	2.9655	92.16%	98.08%
	400	YHH	0.0340	4.8049	82.78%	99.02%
		RIAQ, $\tau = 0.5$	0.0093	3.0000	100.00%	100.00%
		RIAQ, $\tau = 0.75$	0.0106	3.0000	100.00%	100.00%
		RIAQ, $\tau = 0.9$	0.0273	3.0370	90.38%	99.98%
3	200	YHH	0.0192	3.6000	94.27%	100.00%
		RIAQ, $\tau = 0.5$	0.0160	2.9876	97.07%	99.92%
		RIAQ, $\tau = 0.75$	0.0146	2.9880	99.09%	100.00%
		RIAQ, $\tau = 0.9$	0.0372	2.9388	95.60%	99.68%
	400	YHH	0.0144	3.6528	93.28%	99.72%
		RIAQ, $\tau = 0.5$	0.0086	3.0171	99.75%	100.00%
		RIAQ, $\tau = 0.75$	0.0086	3.0989	99.32%	99.95%
		RIAQ, $\tau = 0.9$	0.0217	3.1967	96.15%	99.12%

Note: when $\tau = 0.5$, it is also the median regression with doubly penalty.

Table 2
Simulation results by RIAQ and YHH for Example 2 based on 500 replications.

Case	p	Method	RMSE	AMS	RI1	RI2	RI3
1	200	YHH	0.0079	5.0100	100.00%	100.00%	100.00%
		RIAQ, $\tau = 0.5$	0.0196	4.9745	99.91%	100.00%	99.77%
		RIAQ, $\tau = 0.75$	0.0211	4.9623	99.24%	100.00%	99.53%
		RIAQ, $\tau = 0.9$	0.0331	5.6110	98.88%	99.68%	99.16%
	400	YHH	0.0057	5.0300	100.00%	100.00%	100.00%
		RIAQ, $\tau = 0.5$	0.0115	5.0571	100.00%	100.00%	100.00%
		RIAQ, $\tau = 0.75$	0.0131	4.9709	100.00%	99.93%	100.00%
		RIAQ, $\tau = 0.9$	0.0222	5.8898	99.06%	99.99%	99.54%
2	200	YHH	0.0631	5.6575	89.04%	98.42%	84.93%
		RIAQ, $\tau = 0.5$	0.0115	5.0000	100.00%	100.00%	100.00%
		RIAQ, $\tau = 0.75$	0.0356	4.5000	95.33%	100.00%	92.89%
		RIAQ, $\tau = 0.9$	0.0417	5.3667	97.33%	99.98%	98.66%
	400	YHH	0.0413	5.7586	92.68%	98.85%	87.17%
		RIAQ, $\tau = 0.5$	0.0108	5.0000	100.00%	100.00%	100.00%
		RIAQ, $\tau = 0.75$	0.0075	5.0000	100.00%	100.00%	100.00%
		RIAQ, $\tau = 0.9$	0.0506	5.0000	93.33%	100.00%	79.44%
3	200	YHH	0.0211	5.3590	99.12%	99.06%	96.92%
		RIAQ, $\tau = 0.5$	0.0233	4.8947	99.65%	99.99%	100.00%
		RIAQ, $\tau = 0.75$	0.0183	5.0000	99.48%	100.00%	100.00%
		RIAQ, $\tau = 0.9$	0.0383	5.6875	97.58%	99.43%	99.08%
	400	YHH	0.0174	5.2436	97.89%	99.32%	96.27%
		RIAQ, $\tau = 0.5$	0.0155	5.1429	98.73%	100.00%	99.15%
		RIAQ, $\tau = 0.75$	0.0101	5.0000	100.00%	100.00%	100.00%
		RIAQ, $\tau = 0.9$	0.0318	5.7059	94.04%	99.99%	97.94%

Note: when $\tau = 0.5$, it is also the median regression with doubly penalty.

5. Empirical study

In this section, we apply our methodology to analyze the Communities and Crime Dataset obtained from the UCI machine learning repository, accessible at https://archive.ics.uci.edu/ml/datasets/communities+and+crime. The dataset comprises variables related to crime rates in 1994 communities across 49 states in the US. The dataset contains socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. After excluding variables with missing values, we retain 101 covariates and one response variable. The response variable is the total number of violent crimes per 100K population (denoted by "ViolentCrimesPerPop"). Each community pertains to a single division, with each of the 49 states

Table 3
The estimated coefficients of the selected variables by RIAQ.

	D1	D2	D3	D4	D5	D6	D7	D8	D9
racepctblack	0.276	0.276	0.276	0.000	0.000	0.000	0.276	0.556	0.556
racePctWhite	0.000	−0.339	0.000	0.000	0.000	−0.339	0.000	0.000	0.000
racePctHisp	0.000	0.000	0.300	1.319	0.000	0.000	0.000	0.300	0.300
pctWPubAsst	0.287	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
PctBSorMore	0.000	0.000	0.000	0.000	0.000	−0.187	0.000	0.000	0.000
PctFam2Par	0.000	0.000	0.000	0.000	−0.327	0.000	0.000	0.000	0.000
PctKids2Par	−0.287	−0.287	−0.287	−0.287	−0.287	−0.287	−0.287	−0.287	−0.287
NumIlleg	0.000	0.000	0.000	0.277	0.000	0.000	0.000	0.000	0.000
NumImmig	0.000	0.000	0.000	0.000	0.000	−1.351	0.000	0.000	0.000
PctNotSpeakEnglWell	0.000	0.000	0.000	0.000	0.000	0.803	0.000	−0.323	0.000
PctLargHouseFam	0.000	0.000	0.261	0.000	0.000	0.000	0.000	0.000	0.000
PctPersDenseHous	0.000	0.000	0.000	0.000	0.531	−0.647	0.000	0.000	0.000
HousVacant	0.182	0.182	0.182	0.182	0.182	0.182	0.182	0.182	0.182
MedOwnCostPctInc	0.000	0.000	0.000	−0.397	0.000	0.000	0.000	0.000	0.000
NumStreet	0.000	0.000	0.000	0.206	0.000	0.206	0.000	0.000	0.000
PctUsePubTrans	0.000	0.000	0.000	0.000	0.000	0.456	0.000	0.000	0.000

racepctblack: percentage of the population that is african american; racePctWhite: percentage of the population that is Caucasian; racePctHisp: percentage of the population that is of Hispanic heritage; pctWPubAsst: percentage of households with public assistance income in 1989; PctBSorMore: percentage of people 25 and over with a bachelor's degree or higher education; PctFam2Par: percentage of families (with kids) that are headed by two parents; PctKids2Par: percentage of kids in family housing with two parents; NumIlleg: number of kids born to never married; NumImmig: total number of people known to be foreign-born; PctNotSpeakEnglWell: percent of people who do not speak English well; PctLargHouseFam: percent of family households that are large (6 or more); PctPersDenseHous: percent of persons in dense housing (more than 1 person per room); HousVacant: number of vacant households; MedOwnCostPctInc: median owners cost as a percentage of household income for owners with a mortgage; NumStreet: number of homeless people counted in the street; PctUsePubTrans: percent of people using public transit for commuting.
D1: New England, D2: Middle Atlantic, D3: East North Central, D4: West North Central, D5: South Atlantic, D6: East South Central, D7: West South Central, D8: Mountain, D9: Pacific;
Yang et al. (2019) uses L_2 loss function with heterogeneity and sparsity recovery penalty to analyze the important factors for the violent crime rate. Compared to its result, there are 4 variables being the same. We mark these variables in bold in Table 3.

assigned a unique division number by the United States Census Bureau. To simplify our analysis, we represent these divisions as $D1, D2, \ldots, D9$ ($K = 9$), and the numbers of communities in each division are 258, 358, 217, 87, 262, 123, 239, 98, and 352, respectively, coded by the Federal Information Processing System (FIPS). Detailed information about the variables can be found on the UCI repository website.

The dataset is collected from 9 divisions. However, traditional modeling strategies tend to treat each dataset as either completely homogeneous or analyze each dataset separately. The former approach fails to account for the underlying heterogeneous structure, while the latter ignores common attributes across the 9 divisions. Additionally, Fig. 1 illustrates that the response variable exhibits a heavy-tailed distribution, indicating the presence of outliers. It is therefore advantageous to employ our proposed robust integrative approach for high-dimensional data to effectively analyze this dataset and consider the latent homogeneous, heterogeneous, and sparsity structures concurrently.

We employed the proposed RIAQ estimation method (2.2) with the MCP to estimate the model specified in (1.1). The corresponding estimation outcomes are presented in Table 3, which includes results from 16 covariates that exhibit significant effects on the response variable in at least one division. Conversely, covariates that are not displayed in Table 3 do not possess any substantial effects on the response variable across all 9 divisions given a homogeneous and sparse underlying structure. Our study yields the following key findings:

- "PctKids2Par" and "HousVacant" have homogeneous effects over all divisions. For the former one, the important role of family structure is confirmed also in the criminal justice area (Anderson, 2002; McGranahan, 2021; Singh and Kiran, 2014), and For the latter one, vacant homes are associated with a variety of negative outcomes for communities, including higher rates of some crimes (Donnelly, 1989; Roth, 2019, 2022; Taylor, 1995).
- Among the remaining variables, "racePctWhite", "racePctHisp", "pctWPubAsst", "PctBSorMore", "PctFam2Par", "NumIlleg", "NumImmig", "PctNotSpeakEnglWell", "PctPersDenseHous", "MedOwnCostPctInc", "NumStreet" and "PctUsePubTrans" have heterogeneous and sparse effects over divisions.
- Finally, "racepctblack" has a heterogeneous effect over divisions without sparsity.

To fully explore the integrated latent structure across varying quantile levels, we employed diverse quantile levels ($\tau = 0.3, 0.5, 0.7$) to fit the dataset, thereby detecting heterogeneity and sparsity phenomena. Notably, several significant variables (including "racepctblack", "PctKids2Par", "PctPersDenseHous" and "HousVacant") exhibited remarkable consistency under multiple quantile levels. Our findings, presented in Table 4, demonstrate that these crucial variables are consistently influential, particularly at the 0.5 quantile level. In comparison with the conditional mean-based approach employed by Yang et al. (2019), our method provides more comprehensive evidence for the significance of these common variables at various quantile levels.

Additionally, we compared our method with the traditional pooled quantile regression "Pooled QR" by presenting the average quantile residuals (AQR), which is defined as $\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\rho_\tau(y_{ik} - \boldsymbol{x}_{ik}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_k)$ for both methods, in Table 5. The results indicate that our method outperforms the quantile regression of the pooled data.

Table 4
Important Estimated variables under Different Quantile Levels.

Quantile	0.3	0.5	0.7
racepctblack	✓	✓	✓
racePctWhite	✓	✓	
racePctHisp	✓	✓	
pctWPubAsst	✓	✓	
PctBSorMore		✓	
PctFam2Par	✓	✓	
PctKids2Par	✓	✓	✓
NumIlleg		✓	
NumImmig		✓	
PctNotSpeakEnglWell	✓	✓	
PctLargHouseFam		✓	
PctPersDenseHous	✓	✓	✓
HousVacant	✓	✓	✓
MedOwnCostPctInc		✓	
NumStreet		✓	✓
PctUsePubTrans	✓	✓	

Table 5
The average quantile residuals (AQR) of RIAQ and the quantile regression of the pooled data (Pooled QR)

Quantile	0.3	0.5	0.7
RIAQ	0.029	0.035	0.039
Pooled QR	0.146	0.186	0.179

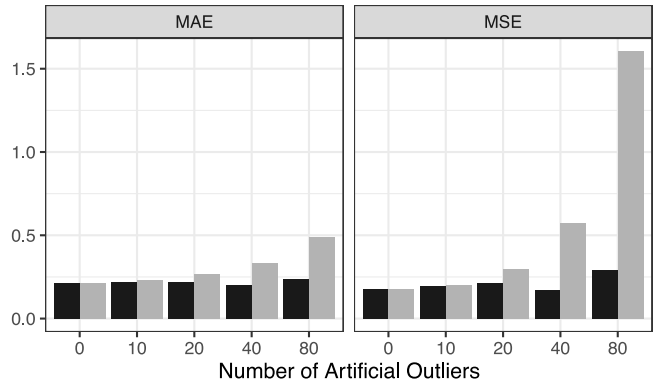

Fig. 2. Predictive error of two methods under different outlier adding levels. Black: RIAQ ($\tau = 0.5$), gray: YHH.

To assess the robustness of our proposed method, we sequentially introduced artificial outliers to the maximum values of responses in the training set. Specifically, for the 1994 responses denoted as y_i and an additional 10 outliers, we randomly distributed outliers to the maximum value in each of the 9 divisions, resulting in a total of 20 outliers. The magnitude of each outlier was set to be equal to $3SD$, where SD is the standard error of the responses. We added a total of $10, 20, 40$, and 80 artificial outliers at random to these 9 divisions. We then split the dataset into a training set and a testing set to compare the predictive accuracy of the two methods. Test mean squared error (MSE), defined as $\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ik} - \boldsymbol{x}_{ik}^{\intercal}\hat{\boldsymbol{\beta}}_k)^2$, and test mean absolute error (MAE), defined as $\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}|y_{ik} - \boldsymbol{x}_{ik}^{\intercal}\hat{\boldsymbol{\beta}}_k|$, were used to measure the predictive accuracy of the two methods. The results are presented graphically in Fig. 2, which shows that our method has considerable robustness when there are artificial outliers in the dataset.

6. Conclusion

This article introduces a new robust method for doubly penalized quantile regression that concurrently recovers homogeneous, heterogeneous, and sparse structures for multiple datasets. Our method robustly detects homogeneity and sparsity across various quantile levels, providing a more comprehensive understanding of integrative analysis for multiple datasets with potential outliers. We employ the ADMM algorithm to break down the complex computation problem into simpler components and demonstrate the parameter selection consistency of our modified Bayesian information criterion under appropriate conditions. The effectiveness of our method is verified through numerical simulations, and we illustrate its application using an empirical example that analyzes factors affecting violent crime rates and explores the impact of demographic characteristics on crime using our robust data analysis method.

Table A.6
Simulation results by RIAQ for Example 1 under different c values based on 500 replications.

Case	τ	c	$p = 200$				$p = 400$			
			RMSE	AMS	RI1	RI2	RMSE	AMS	RI1	RI2
1	0.5	0.2	0.0140	3.1904	98.96%	99.92%	0.0098	3.2048	99.63%	99.92%
		1	0.0132	3.0043	98.26%	99.87%	0.0084	3.0222	99.62%	100.00%
		1.8	0.0134	2.9765	98.15%	100.00%	0.0096	2.9615	98.94%	100.00%
	0.75	0.2	0.0117	3.3560	99.94%	99.92%	0.0084	3.3140	99.86%	99.93%
		1	0.0109	2.9966	99.39%	100.00%	0.0069	3.0345	99.96%	100.00%
		1.8	0.0122	2.9519	97.98%	100.00%	0.0089	2.9481	98.04%	100.00%
	0.9	0.2	0.0352	3.4677	92.66%	98.02%	0.0253	3.5854	90.69%	98.68%
		1	0.0308	2.9670	96.70%	99.55%	0.0219	3.0971	95.51%	99.47%
		1.8	0.0301	2.8854	96.07%	100.00%	0.0212	2.9742	95.78%	99.63%
2	0.5	0.2	0.0265	4.2597	95.26%	98.61%	0.0167	4.0504	97.82%	98.87%
		1	0.0249	2.8372	93.52%	99.07%	0.0093	3.0000	100.00%	100.00%
		1.8	0.0309	2.5714	100.00%	97.14%	0.0079	3.0000	100.00%	100.00%
	0.75	0.2	0.0257	4.5018	96.27%	99.10%	0.0162	4.1911	97.71%	99.63%
		1	0.0179	2.9400	98.12%	100.00%	0.0106	3.0000	100.00%	100.00%
		1.8	0.0218	2.7692	94.33%	100.00%	0.0107	3.0000	100.00%	100.00%
	0.9	0.2	0.0533	4.2784	82.40%	91.13%	0.0397	4.1538	76.67%	94.91%
		1	0.0388	2.9655	92.16%	98.08%	0.0273	3.0370	90.38%	99.98%
		1.8	0.0412	2.7500	92.33%	97.78%	0.0225	3.0000	97.04%	99.98%
3	0.5	0.2	0.0168	3.4179	98.62%	99.68%	0.0114	3.4736	99.40%	99.61%
		1	0.0160	2.9876	97.07%	99.92%	0.0086	3.0171	99.75%	100.00%
		1.8	0.0196	2.8718	92.76%	100.00%	0.0084	3.0000	100.00%	100.00%
	0.75	0.2	0.0160	3.6157	99.02%	99.50%	0.0109	3.6279	98.75%	99.82%
		1	0.0146	2.9880	99.09%	100.00%	0.0086	3.0989	99.32%	99.95%
		1.8	0.0173	2.8667	97.93%	100.00%	0.0089	2.9806	98.65%	100.00%
	0.9	0.2	0.0412	3.7088	90.08%	97.63%	0.0277	4.0044	88.90%	95.96%
		1	0.0372	2.9388	95.60%	99.68%	0.0217	3.1967	96.15%	99.12%
		1.8	0.0362	2.8351	96.25%	99.68%	0.0222	3.0267	95.61%	99.51%

Note: when $\tau = 0.5$, it is also the median regression with doubly penalty.

Acknowledgments

We thank the Editor, the Associate Editor and the referees for their encouragements and insightful comments which have substantially improved the paper. This work is supported by the National Key R&D Program of China (2022YFA1003800), National Natural Science Foundation of China (NNSFC) (11922117, 71903163, 71988101, 12301344), National Statistical Science Research Grants of China (Major Program 2022LD08), China Postdoctoral Science Foundation (2023M731813) and the 111 Project (B13028).

Appendix A. Additional simulation results

We simulate the proposed method and assess the selection performance of the proposed BIC in Eq. (2.21) using different values of c. The results presented in Tables A.6 and A.7 reveal that our proposed method yields stable outcomes for different values of c. Moreover, as the value of c increases, the estimated number of components K tends to be underestimated.

Appendix B. Proof of Proposition 1

Lemma 1. *The absolute value of Lagrange multipliers is bounded, (1)* $|\vartheta^{j}_{kk'}| \le \lambda$, *(2)* $|v_{kj}| \le \gamma$, *(3)* $|\delta_{ik}| \le \max\{\tau, 1-\tau\}/n$, *for* $1 \le k < k' \le K, j = 1, \ldots, p$.

Proof of Lemma 1. It follows from Eqs. (2.13) and (2.18),

$$\vartheta^{j(t+1)}_{kk'} = \rho_1 \left(\beta^{(t+1)}_{kj} - \beta^{(t+1)}_{k'j} + \vartheta^{j(t)}_{kk'}/\rho_1 \right) - \rho_1 \alpha^{j(t+1)}_{kk'} = \rho_1 (\theta^{j(t+1)}_{kk'} - \alpha^{j(t+1)}_{kk'}). \tag{B.1}$$

By Eqs. (2.12) and (2.13), we have

$$|\vartheta^{j(t+1)}_{kk'}| = \begin{cases} 0 & \text{if } |\theta^{j(t+1)}_{kk'}| \ge a\lambda, \\ \rho_1 \frac{a\lambda - |\theta^{j(t+1)}_{kk'}|}{1 - a\rho_1} & \text{if } \lambda/\rho_1 \le |\theta^{j(t+1)}_{kk'}| < a\lambda, \\ \rho_1 |\theta^{j(t+1)}_{kk'}| & \text{if } |\theta^{j(t+1)}_{kk'}| < \lambda/\rho_1, \end{cases}$$

then $|\vartheta^{j(t+1)}_{kk'}| \le \lambda$. Similarly, we can obtain $|v_{kj}| \le \gamma$.

Table A.7
Simulation results by RIAQ for Example 2 under different c values based on 500 replications.

Case	τ	c	$p = 200$					$p = 400$				
			RMSE	AMS	RI1	RI2	RI3	RMSE	AMS	RI1	RI2	RI3
1	0.5	0.2	0.0162	5.5540	99.96%	99.79%	99.91%	0.0145	6.0580	99.86%	99.97%	99.43%
		1	0.0186	5.1585	99.78%	99.75%	99.87%	0.0122	5.3464	100.00%	99.99%	99.92%
		1.2	0.0196	4.9745	99.91%	100.00%	99.77%	0.0115	5.0571	100.00%	100.00%	100.00%
		1.8	0.0360	4.6000	100.00%	100.00%	100.00%	0.0115	5.0571	100.00%	100.00%	100.00%
	0.75	0.2	0.0138	5.7140	99.98%	99.98%	99.98%	0.0103	5.8240	99.94%	99.97%	99.89%
		1	0.0191	5.0981	99.48%	100.00%	99.72%	0.0112	5.1519	99.95%	99.97%	99.96%
		1.2	0.0211	4.9623	99.24%	100.00%	99.53%	0.0131	4.9709	100.00%	99.93%	100.00%
		1.8	0.0433	4.4545	100.00%	100.00%	98.38%	0.0069	5.0000	100.00%	100.00%	100.00%
	0.9	0.2	0.0368	6.1278	98.39%	99.28%	98.77%	0.0301	7.0520	97.55%	99.53%	97.27%
		1	0.0355	5.8539	98.55%	99.58%	98.91%	0.0272	6.3299	98.04%	99.68%	98.28%
		1.2	0.0331	5.6110	98.88%	99.68%	99.16%	0.0222	5.8898	99.06%	99.99%	99.54%
		1.8	0.0273	5.0370	99.50%	100.00%	99.59%	0.0222	5.8898	99.06%	99.99%	99.54%
2	0.5	0.2	0.0445	7.0568	92.57%	99.38%	94.64%	0.0353	7.5106	92.86%	99.52%	92.23%
		1	0.0249	4.9375	99.58%	100.00%	99.44%	0.0127	5.1111	100.00%	99.96%	100.00%
		1.2	0.0115	5.0000	100.00%	100.00%	100.00%	0.0108	5.0000	100.00%	100.00%	100.00%
		1.8	0.0115	5.0000	100.00%	100.00%	100.00%	0.0108	5.0000	100.00%	100.00%	100.00%
	0.75	0.2	0.0443	7.0058	91.91%	98.79%	94.39%	0.0268	7.2765	96.17%	99.66%	95.56%
		1	0.0294	4.9000	96.83%	100.00%	97.33%	0.0117	5.0000	100.00%	100.00%	100.00%
		1.2	0.0356	4.5000	95.33%	100.00%	92.89%	0.0075	5.0000	100.00%	100.00%	100.00%
		1.8	0.0356	4.5000	95.33%	100.00%	92.89%	0.0075	5.0000	100.00%	100.00%	100.00%
	0.9	0.2	0.0697	7.6615	90.05%	97.97%	92.36%	0.0560	8.2313	90.57%	98.83%	86.10%
		1	0.0527	5.6613	93.43%	99.31%	95.86%	0.0422	5.7619	92.05%	100.00%	89.99%
		1.2	0.0417	5.3667	97.33%	99.98%	98.66%	0.0506	5.0000	93.33%	100.00%	79.44%
		1.8	0.0417	5.3667	97.33%	99.98%	98.66%	0.0506	5.0000	93.33%	100.00%	79.44%
3	0.5	0.2	0.0215	6.0500	99.22%	99.88%	99.45%	0.0202	6.4875	98.44%	99.73%	97.64%
		1	0.0206	5.1011	99.59%	99.99%	99.90%	0.0133	5.2584	99.65%	100.00%	99.73%
		1.2	0.0233	4.8947	99.65%	99.99%	100.00%	0.0155	5.1429	98.73%	100.00%	99.15%
		1.8	0.0549	4.2000	94.67%	100.00%	100.00%	0.0155	5.1429	98.73%	100.00%	99.15%
	0.75	0.2	0.0207	6.2832	98.84%	99.74%	99.29%	0.0142	6.3209	99.37%	99.93%	99.41%
		1	0.0165	5.1899	99.66%	99.89%	99.86%	0.0094	5.1496	100.00%	100.00%	100.00%
		1.2	0.0183	5.0000	99.48%	100.00%	100.00%	0.0101	5.0000	100.00%	100.00%	100.00%
		1.8	0.0183	5.0000	99.48%	100.00%	100.00%	0.0101	5.0000	100.00%	100.00%	100.00%
	0.9	0.2	0.0462	6.8231	95.88%	97.92%	97.68%	0.0415	7.2351	90.89%	99.54%	95.55%
		1	0.0422	6.0481	96.40%	98.74%	98.21%	0.0350	6.2303	93.61%	99.98%	97.17%
		1.2	0.0383	5.6875	97.58%	99.43%	99.08%	0.0318	5.7059	94.04%	99.99%	97.94%
		1.8	0.0463	4.6250	96.22%	100.00%	91.11%	0.0205	5.0000	100.00%	100.00%	100.00%

Note: when $\tau = 0.5$, it is also the median regression with doubly penalty.

By Eq. (2.20), we have

$$\boldsymbol{\delta}^{(t+1)} = \rho_3(\boldsymbol{y} - \widetilde{\mathbf{X}}^{\intercal}\boldsymbol{\beta}^{(t+1)} - \mathbf{z}^{(t+1)} + \boldsymbol{\delta}^{(t)}/\rho_3),$$

and by Eq. (2.17), we have $|\delta_{ik}^{(t+1)}| \le \max\{\tau, 1-\tau\}/n$. □

Lemma 2. *For the iterated sequence* $\{\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\eta}^{(t)}, \mathbf{z}^{(t)}, \boldsymbol{\vartheta}^{(t)}, \boldsymbol{\nu}^{(t)}, \boldsymbol{\delta}^{(t)}\}$, *denote*

$$\mathcal{Q}^{(t)} \equiv \mathcal{Q}_{\rho_1,\rho_2,\rho_3}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\eta}^{(t)}, \mathbf{z}^{(t)}, \boldsymbol{\vartheta}^{(t)}, \boldsymbol{\nu}^{(t)}, \boldsymbol{\delta}^{(t)}),$$

and $\triangle s^{(t)} \equiv s^{(t+1)} - s^{(t)}$ *for all* $s^{(t)}$. *Then* $\triangle \mathcal{Q}^{(t)} \le -\frac{c}{2}\|\triangle \boldsymbol{\beta}^{(t)}\|^2 + \rho_1^{-1}\|\triangle \boldsymbol{\vartheta}^{(t)}\|_F^2 + \rho_2^{-1}\|\triangle \boldsymbol{\nu}^{(t)}\|_F^2 + \rho_3^{-1}\|\triangle \boldsymbol{\delta}^{(t)}\|^2$ *for some constant* c.

Proof of Lemma 2. By Eqs. (2.7), and (2.18)–(2.20),

$$\begin{aligned}&\mathcal{Q}^{(t+1)} - \mathcal{Q}_{\rho_1,\rho_2,\rho_3}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\eta}^{(t+1)}, \mathbf{z}^{(t+1)}, \boldsymbol{\vartheta}^{(t)}, \boldsymbol{\nu}^{(t)}, \boldsymbol{\delta}^{(t)})\\ =&\rho_1^{-1}\|\triangle \boldsymbol{\vartheta}^{(t)}\|_F^2 + \rho_2^{-1}\|\triangle \boldsymbol{\nu}^{(t)}\|_F^2 + \rho_3^{-1}\|\triangle \boldsymbol{\delta}^{(t)}\|^2.\end{aligned}$$

Our optimization algorithm implies that

$$\begin{aligned}&\mathcal{Q}_{\rho_1,\rho_2,\rho_3}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\eta}^{(t+1)}, \mathbf{z}^{(t+1)}, \boldsymbol{\vartheta}^{(t)}, \boldsymbol{\nu}^{(t)}, \boldsymbol{\delta}^{(t)})\\ -&\mathcal{Q}_{\rho_1,\rho_2,\rho_3}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\eta}^{(t)}, \mathbf{z}^{(t)}, \boldsymbol{\vartheta}^{(t)}, \boldsymbol{\nu}^{(t)}, \boldsymbol{\delta}^{(t)}) \le 0.\end{aligned}$$

From Eq. (2.7), $\mathcal{Q}_{\rho_1,\rho_2,\rho_3}(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\eta}^{(t)}, \mathbf{z}^{(t)}, \boldsymbol{\vartheta}^{(t)}, \boldsymbol{\nu}^{(t)}, \boldsymbol{\delta}^{(t)})$ is strongly convex quadratic function of $\boldsymbol{\beta}$. As a result,

$$\begin{aligned}&\mathcal{Q}_{\rho_1,\rho_2,\rho_3}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\eta}^{(t)}, \mathbf{z}^{(t)}, \boldsymbol{\vartheta}^{(t)}, \boldsymbol{\nu}^{(t)}, \boldsymbol{\delta}^{(t)})\\ -&\mathcal{Q}_{\rho_1,\rho_2,\rho_3}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\eta}^{(t)}, \mathbf{z}^{(t)}, \boldsymbol{\vartheta}^{(t)}, \boldsymbol{\nu}^{(t)}, \boldsymbol{\delta}^{(t)}) \le -\frac{c}{2}\|\triangle \boldsymbol{\beta}^{(t)}\|^2,\end{aligned}$$

for some constant $c > 0$. Combining the above three conclusions, we have

$$\begin{aligned}\triangle Q^{(t)} &= Q^{(t+1)} - Q^{(t)} \\ &\le -\frac{c}{2}\| \triangle \boldsymbol{\beta}^{(t)}\|^2 + \rho_1^{-1}\| \triangle \boldsymbol{\vartheta}^{(t)}\|_F^2 + \rho_2^{-1}\| \triangle \boldsymbol{\nu}^{(t)}\|_F^2 + \rho_3^{-1}\| \triangle \boldsymbol{\delta}^{(t)}\|^2. \end{aligned}$$ □

Proof of Proposition 1. There are three conclusions in Proposition 1,

(1) Since sequences $\{\boldsymbol{\alpha}^{(t)}\}_{t=1}^{\infty}$, $\{\boldsymbol{\eta}^{(t)}\}_{t=1}^{\infty}$ and $\{\mathbf{z}^{(t)}\}_{t=1}^{\infty}$ are all bounded (from Eq. (2.7)), by Lemma 1, (2.9), $\{\boldsymbol{\beta}^{(t)}\}_{t=1}^{\infty}$ is also bounded.

(2) Since $\{Q^{(t)}\}_{t=1}^{\infty}$ is bounded, there exists a subsequence $\{t_k\}$ s.t. $Q^{(t_k)}, \boldsymbol{\beta}^{(t_k)}, \boldsymbol{\alpha}^{(t_k)}, \boldsymbol{\eta}^{(t_k)}, \mathbf{z}^{(t_k)}, \boldsymbol{\vartheta}^{(t_k)}, \boldsymbol{\nu}^{(t_k)}, \boldsymbol{\delta}^{(t_k)}$ converge, and we denote the limits by $Q^*, \boldsymbol{\beta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\eta}^*, \mathbf{z}^*, \boldsymbol{\vartheta}^*, \boldsymbol{\nu}^*, \boldsymbol{\delta}^*$. By Lemma 2, $\frac{c}{2}\| \triangle \boldsymbol{\beta}^{(t_k)}\|^2 \le - \triangle Q^{(t_k)} + \rho_1^{-1}\| \triangle \boldsymbol{\vartheta}^{(t_k)}\|_F^2 + \rho_2^{-1}\| \triangle \boldsymbol{\nu}^{(t_k)}\|_F^2 + \rho_3^{-1}\| \triangle \boldsymbol{\delta}^{(t_k)}\|^2 \to 0$ as $k \to \infty$. So, $\| \triangle \boldsymbol{\beta}^{(t_k)}\| \to 0$ as $k \to \infty$. Because of the condition $\rho_1^{-1}\| \triangle \boldsymbol{\vartheta}^{(t)}\|_F + \rho_2^{-1}\| \triangle \boldsymbol{\nu}^{(t)}\|_F + \rho_3^{-1}\| \triangle \boldsymbol{\delta}^{(t)}\| \to 0$ as $t \to \infty$ in this proposition, by Eqs. (2.18)–(2.20) we have $\| \triangle \boldsymbol{\alpha}^{(t_k)}\| + \| \triangle \boldsymbol{\eta}^{(t_k)}\| + \| \triangle \mathbf{z}^{(t_k)}\| \to 0$ as $k \to \infty$.

(3) Next, we will verify that $(Q^*, \boldsymbol{\beta}^*, \boldsymbol{\alpha}^*, \boldsymbol{\eta}^*, \mathbf{z}^*, \boldsymbol{\vartheta}^*, \boldsymbol{\nu}^*, \boldsymbol{\delta}^*)$ is a stationary point satisfying KKT conditions. As a result of Eqs. (2.18)–(2.20), we have

$$\mathbf{0} = \rho_1(\mathcal{E}\boldsymbol{B}^* - \boldsymbol{\alpha}^*), \tag{B.2}$$

$$\mathbf{0} = \rho_2(\boldsymbol{\eta}^* - \mathbf{B}^*), \tag{B.3}$$

$$\mathbf{0} = \rho_3(\mathbf{y} - \widetilde{\mathbf{X}}^{\intercal}\boldsymbol{\beta}^* - \mathbf{z}^*). \tag{B.4}$$

The first order conditions of optimization processes in Eqs. (2.8), (2.11), (2.14), and (2.16) are respectively

$$\begin{aligned}&\rho_1\mathcal{E}^{\intercal}(\mathcal{E}\mathbf{B}^{(t_k+1)} - \boldsymbol{\alpha}^{(t_k)} + \rho_1^{-1}\boldsymbol{\vartheta}^{(t_k)}) - \rho_2(\boldsymbol{\eta}^{(t_k)} - \mathbf{B}^{(t_k+1)} + \rho_2^{-1}\boldsymbol{\nu}^{(t_k)}) \\ &-\rho_3\mathcal{I}(\mathbf{X}\circ(\mathbf{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}^{(t_k+1)} - \mathbf{z}^{(t_k)} + \frac{\boldsymbol{\delta}^{(t_k)}}{\rho_3})) = 0,\end{aligned} \tag{B.5}$$

$$\left.\frac{\partial p_\lambda\left(\left|\alpha_{kk'}^j\right|\right)}{\partial \alpha_{kk'}^j}\right|_{\alpha_{kk'}^j = \alpha_{kk'}^{j(t_k+1)}} - \rho_1\left(\beta_{kj}^{(t_k+1)} - \beta_{k'j}^{(t_k+1)} + \rho_1^{-1}\vartheta_{kk'}^{j(t_k)} - \alpha_{kk'}^{j(t_k+1)}\right) \ni 0, \tag{B.6}$$

$$\left.\frac{\partial p_\gamma\left(\left|\eta_{kj}\right|\right)}{\partial \eta_{kj}}\right|_{\eta_{kj} = \eta_{kj}^{(t_k+1)}} - \rho_2\left(\beta_{kj}^{(t_k+1)} - \rho_2^{-1}\nu_{kj}^{(t_k)} - \eta_{kj}^{(t_k+1)}\right) \ni 0, \tag{B.7}$$

$$-\rho_3(y_{ik} - \boldsymbol{x}_{ik}^{\intercal}\boldsymbol{\beta}_k^{(t_k+1)} + \rho_3^{-1}\delta_{ik}^{(t_k)} - z_{ik}^{(t_k+1)}) + \frac{1}{n}\left.\frac{\partial \rho_\tau(z_{ik})}{\partial z_{ik}}\right|_{z_{ik} = z_{ik}^{(t_k+1)}} \ni 0. \tag{B.8}$$

Letting $k \to \infty$, by (B.2)–(B.4), the above four first order conditions are simplified into

$$\mathcal{E}^{\intercal}\boldsymbol{\vartheta}^* - \boldsymbol{\nu}^* - \mathcal{I}(\mathbf{X}\circ\boldsymbol{\delta}^*) = 0, \tag{B.9}$$

$$\left.\frac{\partial p_\lambda\left(\left|\alpha_{kk'}^j\right|\right)}{\partial \alpha_{kk'}^j}\right|_{\alpha_{kk'}^j = \alpha_{kk'}^{j(*)}} - \vartheta_{kk'}^{j*} \ni 0, \tag{B.10}$$

$$\left.\frac{\partial p_\gamma\left(\left|\eta_{kj}\right|\right)}{\partial \eta_{kj}}\right|_{\eta_{kj} = \eta_{kj}^{(*)}} + \nu_{kj}^* \ni 0, \tag{B.11}$$

$$-\delta_{ik}^* + \frac{1}{n}\left.\frac{\partial \rho_\tau(z_{ik})}{\partial z_{ik}}\right|_{z_{ik} = z_{ik}^{(*)}} \ni 0 \tag{B.12}$$

respectively, which are just the KKT conditions of the optimization problem in (2.6) concerning $\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\eta}$ and $\mathbf{z}$ respectively. □

Appendix C. Proof of Theorem 1

The optimization problem described in Eq. (2.2) is a difference convex problem that has been extensively studied by Tao (1997). Lemma 3 provides a sufficient condition for a local optimizer in a difference convex program, while Lemma 4 describes the asymptotic properties of the oracle estimator. With some additional conditions, we can prove Theorem 1 using Lemmas 3 and 4.

Lemma 3 (*Tao, 1997; Wang et al., 2012*)**.** *If there exists a neighborhood U around the point x^* such that $\partial h(x) \cap \partial g(x^*) \ne \emptyset, \forall x \in U \cap \mathrm{dom}(g)$, where g, h are both convex. Then x^* is a local minimizer of $g(x) - h(x)$.*

Lemma 4. *Suppose Assumptions C1–C4 hold, $\lambda = \mathbf{o}(n^{-(1-c_2)/2})$ and $\gamma = \mathbf{o}(n^{-(1-c_2)/2})$. Then the oracle estimator $\|\widetilde{\boldsymbol{\phi}}_1 - \boldsymbol{\phi}_{01}\| = O_p(\sqrt{q/n})$ and $|\tilde{\phi}_j| \ge (a+1/2)\gamma$ for $j = 1, \ldots, q$, $|\tilde{\beta}_{kj} - \tilde{\beta}_{k'j}| \ge (a+1/2)\lambda$ for $k \in G_s^j, k' \in G_{s'}^j, \phi_s^j \ne 0, s \ne s' \in \{1, \ldots, S_j\}$ with probability approaching 1, where a is the parameter in the penalty function.*

Remark 1. The proof of Lemma 4 is omitted as it is essentially the same as those presented in He and Shao (2000), Wang et al. (2012), Zheng et al. (2015), Zhang et al. (2019).

Proof of Theorem 1. To use Lemma 3, notice that $Q_n(\boldsymbol{\beta} \mid \lambda, \gamma) = g(\beta \mid \lambda, \gamma) - h(\beta \mid \lambda, \gamma)$, where

$$g(\boldsymbol{\beta}) = g(\boldsymbol{\beta} \mid \lambda, \gamma) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in G_k}\rho_\tau(y_i - \sum_{j=1}^{p} x_{ij}\beta_{kj}) + \lambda\sum_{j=1}^{p}\sum_{k<k'}|\beta_{kj} - \beta_{k'j}| + \gamma\sum_{j=1}^{p}\sum_{k=1}^{K}|\beta_{kj}|,$$

$$h(\boldsymbol{\beta}) = h(\boldsymbol{\beta} \mid \lambda, \gamma) = \sum_{j=1}^{p}\sum_{k<k'} H_\lambda(\beta_{kj} - \beta_{k'j}) + \sum_{j=1}^{p}\sum_{k=1}^{K} H_\gamma(\beta_{kj}),$$

where for the MCP

$$H_\lambda(x) = [x^2/(2a)]\mathbf{1}(0 \le |x| \le a\lambda) + [\lambda|x| - a\lambda^2/2]\mathbf{1}(|x| > a\lambda),$$

and for the SCAD penalty

$$H_\lambda(x) = \left[\left(x^2 - 2\lambda|x| + \lambda^2\right)/(2(a-1))\right]\mathbf{1}(\lambda \le |x| \le a\lambda) + \left[\lambda|x| - (a+1)\lambda^2/2\right]\mathbf{1}(|x| > a\lambda).$$

Here we only consider the SCAD penalty, as their proofs are similar.

First, the subderivative of $g(\boldsymbol{\beta})$ is:

$$\partial g(\boldsymbol{\beta}) = \left\{(\xi_{11}, \xi_{12}, \dots, \xi_{1p}, \dots, \xi_{Kp}) \in \mathbb{R}^{Kp} \;\middle|\; \xi_{kj} = s_{kj} + \lambda\sum_{k':k'<k} l_{k'kj} + \lambda\sum_{k':k'>k} l_{kk'j} + \gamma l_{kj}\right\}$$

where for two intervals, I_1, I_2, $I_1 + I_2 = \{y = x_1 + x_2 \mid x_1 \in I_1, x_2 \in I_2\}$ and

$$s_{kj} = s_{kj}(\boldsymbol{\nu}) = \frac{1}{n}\sum_{i\in G_k} x_{ij}\left[(1-\tau)\mathbf{1}(y_i - \sum_{j=1}^{p} x_{ij}\beta_{kj} < 0) - \tau\mathbf{1}(y_i - \sum_{j=1}^{p} x_{ij}\beta_{kj} > 0) - \nu_{ik}\right], \tag{C.1}$$

with

$$\nu_{ik} = \begin{cases}[\tau - 1, \tau], & \text{if } y_i - \sum_{j=1}^{p} x_{ij}\beta_{kj} = 0\\ 0, & \text{otherwise,}\end{cases}$$

$$l_{k'kj} = \begin{cases}\operatorname{sgn}(\beta_{kj} - \beta_{k'j}), & \text{if } \beta_{kj} \ne \beta_{k'j}\\ [-1, 1], & \text{if } \beta_{kj} = \beta_{k'j},\end{cases}$$

$$l_{kj} = \begin{cases}\operatorname{sgn}(\beta_{kj}), & \text{if } \beta_{kj} \ne 0\\ [-1, 1], & \text{if } \beta_{kj} = 0,\end{cases}$$

and $\operatorname{sgn}(x) = \mathbf{1}(x > 0) - \mathbf{1}(x < 0)$ is the sign function. And further,

$$\begin{aligned}\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_{kj}} = \sum_{k'=1}^{K}\Bigg[&\frac{(\beta_{kj} - \beta_{k'j}) - \lambda\operatorname{sgn}(\beta_{kj} - \beta_{k'j})}{a-1}\mathbf{1}(\lambda < |\beta_{kj} - \beta_{k'j}| < a\lambda) + \\ &\lambda\operatorname{sgn}(\beta_{kj} - \beta_{k'j})\mathbf{1}(|\beta_{kj} - \beta_{k'j}| \ge a\lambda)\Bigg] + \\ &\left[\frac{\beta_{kj} - \gamma\operatorname{sgn}(\beta_{kj})}{a-1}\mathbf{1}(\gamma < |\beta_{kj}| < a\gamma) + \gamma\operatorname{sgn}(\beta_{kj})\mathbf{1}(|\beta_{kj}| \ge a\gamma)\right].\end{aligned} \tag{C.2}$$

Second, we consider the oracle estimator, $\widetilde{\boldsymbol{\beta}}$ (or corresponding $\widetilde{\boldsymbol{\phi}}$ satisfying $\operatorname{vec}(\widetilde{\boldsymbol{\beta}}^\intercal) = \mathbf{W}_0\widetilde{\boldsymbol{\phi}}$), we rewrite the formula of oracle estimator in (3.22) as

$$\begin{aligned}&\widetilde{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in G_k}\rho_\tau(y_i - \sum_{j=1}^{p} x_{ij}\beta_{kj}), j = 1, \dots, p,\\ &\text{s.t. } \beta_{kj} = \beta_{k'j} \text{ for } k < k' \in G_s^j, s \in \{1, \dots, S_j\}, j = 1, \dots, p,\\ &\quad \beta_{kj} = 0 \text{ if } k \in G_s^j,\ \phi_s^j = 0.\end{aligned}$$

The corresponding Lagrange function is

$$\begin{aligned}L(\boldsymbol{\beta}, \boldsymbol{\varrho}, \boldsymbol{\varphi}) = &\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in G_k}\rho_\tau(y_i - \sum_{j=1}^{p} x_{ij}\beta_{kj}) + \sum_{j=1}^{p}\sum_{s=1}^{S_j}\sum_{k:k\in G_s^j}\sum_{k':k<k'\in G_s^j}(\beta_{kj} - \beta_{k'j})\varrho_{jskk'} + \\ &\sum_{j,s,k:\phi_s^j=0, k\in G_s^j}\beta_{kj}\varphi_{jsk},\end{aligned}$$

where $\varrho_{jskk'}$ and φ_{jsk} are Lagrange multipliers. Recall the definition of s_{kj} in Eq. (C.1), its subderivative is

$$\begin{aligned}\partial L(\boldsymbol{\beta},\boldsymbol{\varrho},\boldsymbol{\varphi}) = \Big\{ & (\pi_{kj} : k = 1,\dots,K, j = 1,\dots,p; \\ & \beta_{kj} - \beta_{k'j},\ \text{for } k < k' \in G_s^j, j = 1,\dots,p, s = 1,\dots,S_j, k < k' \in G_s^j; \\ & \beta_{kj}, j, s \text{ s.t. } \phi_s^j = 0, k \in G_s^j)\Big| \\ & \pi_{kj} = s_{kj} + \sum_{s=1}^{S_j} \sum_{k' : k<k' \in G_s^j} \varrho_{jskk'} - \sum_{s=1}^{S_j} \sum_{k' : k'<k \in G_s^j} \varrho_{jsk'k} + \sum_{s:\phi_s^j=0} \mathbf{1}\{k \in G_s^j\}\varphi_{jsk}, \Big\}.\end{aligned}$$

The truth that $0 \in \partial L(\boldsymbol{\beta},\boldsymbol{\varrho},\boldsymbol{\varphi})|_{\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\varrho}},\widetilde{\boldsymbol{\varphi}}}$ implies that $\exists \boldsymbol{\nu}^* \in [\tau-1,\tau]^n$ s.t.

$$\begin{aligned} & s_{kj}(\boldsymbol{\nu}^*) + \sum_{s=1}^{S_j} \sum_{k' : k<k' \in G_s^j} \tilde{\varrho}_{jskk'} - \sum_{s=1}^{S_j} \sum_{k' : k'<k \in G_s^j} \tilde{\varrho}_{jsk'k} + \sum_{s:\phi_s^j=0} \mathbf{1}\{k \in G_s^j\}\tilde{\varphi}_{jsk} = 0 \\ & \tilde{\beta}_{kj} = \tilde{\beta}_{k'j} \text{ for } k < k' \in G_s^j \\ & \tilde{\beta}_{kj} = 0,\ \text{for } j, s, k \text{ s.t. } \phi_s^j = 0, k \in G_s^j, \end{aligned} \tag{C.3}$$

Third, we aim to prove that for any $\boldsymbol{\beta} \in \mathbf{B}(\widetilde{\boldsymbol{\beta}}, \frac{\min\{\lambda,\gamma\}}{4})$ (a ball centered at $\widetilde{\boldsymbol{\beta}}$), $\partial h(\boldsymbol{\beta}) \cap \partial g(\widetilde{\boldsymbol{\beta}}) \neq \emptyset$, so that Lemma 3 is applicable. $\forall \boldsymbol{\beta} \in \mathbf{B}(\widetilde{\boldsymbol{\beta}}, \lambda/4)$, if k and k' belong to the same group G_s^j, then $|\beta_{kj} - \beta_{k'j}| \le |\beta_{kj} - \beta_{k'j} - (\tilde{\beta}_{kj} - \tilde{\beta}_{k'j})| + |\tilde{\beta}_{kj} - \tilde{\beta}_{k'j}| < \frac{\lambda}{2} + 0 = \frac{\lambda}{2}$; and by Lemma 4, if k and k' are not in the same group, $|\beta_{kj} - \beta_{k'j}| \ge |\tilde{\beta}_{kj} - \tilde{\beta}_{k'j}| - |\tilde{\beta}_{kj} - \tilde{\beta}_{k'j} - (\beta_{kj} - \beta_{k'j})| > (a + \frac{1}{2})\lambda - \frac{\lambda}{2} = a\lambda$, where a is defined in Lemma 4. Similarly, by Lemma 4, if $\phi_s^j = 0$ and $k \in G_s^j$, then $|\beta_{jk}| \le |\beta_{jk} - \tilde{\beta}_{jk}| + |\tilde{\beta}_{jk}| < \frac{\gamma}{2} + 0 = \frac{\gamma}{2}$, if $\phi_s^j \neq 0$ and $k \in G_s^j$, then $|\beta_{jk}| \ge |\tilde{\beta}| - |\tilde{\beta} - \beta| > (a + \frac{1}{2})\gamma - \frac{\gamma}{2} = a\gamma$. It follows from these conclusions and Eq. (C.2) that when $k \in G_s^j$,

$$\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_{kj}} = \begin{cases} \lambda \sum_{k' \notin G_s^j} \operatorname{sgn}(\beta_{kj} - \beta_{k'j}) + \gamma \operatorname{sgn}(\beta_{kj}) & \text{if } \phi_s^j \neq 0, \\ \lambda \sum_{k' \notin G_s^j} \operatorname{sgn}(\beta_{kj} - \beta_{k'j}) & \text{if } \phi_s^j = 0. \end{cases} \tag{C.4}$$

Eq. (C.3) implies that

$$\begin{aligned} & s_{kj}(\boldsymbol{\nu}^*) = \sum_{s=1}^{S_j} \sum_{k' : k'<k \in G_s^j} \tilde{\varrho}_{jsk'k} - \sum_{s=1}^{S_j} \sum_{k' : k<k' \in G_s^j} \tilde{\varrho}_{jskk'} - \sum_{s:\phi_s^j=0} \mathbf{1}\{k \in G_s^j\}\tilde{\varphi}_{jsk}, \\ & \tilde{\beta}_{kj} = \tilde{\beta}_{k'j} \text{ for } k < k' \in G_s^j, \\ & \tilde{\beta}_{kj} = 0,\ \text{when } \phi_s^j = 0 \text{ and } k \in G_s^j. \end{aligned}$$

Next we will analyze whether $\partial h(\boldsymbol{\beta}) \cap \partial g(\widetilde{\boldsymbol{\beta}}) \neq \emptyset$ or not by two cases.

(1) When $\tilde{\beta}_{kj} \neq 0$ and $k \in G_s^j$,

$$\begin{aligned} \frac{\partial g(\widetilde{\boldsymbol{\beta}})}{\partial \beta_{kj}}\bigg|_{\boldsymbol{\nu}^*} = & s_{kj}(\boldsymbol{\nu}^*) + \lambda \sum_{k' \in G_s^j : k'<k} l_{k'kj} + \lambda \sum_{k' : k' \notin G_s^j} \operatorname{sgn}(\tilde{\beta}_{kj} - \tilde{\beta}_{k'j}) \\ & + \lambda \sum_{k' \in G_s^j : k<k'} l_{kk'j} + \gamma \operatorname{sgn}(\tilde{\beta}_{kj}). \end{aligned} \tag{C.5}$$

And

$$\lambda \sum_{k' \in G_s^j : k'<k} l_{k'kj} + \lambda \sum_{k' \in G_s^j : k<k'} l_{kk'j} = \left[-\lambda(|G_s^j| - 1), \lambda(|G_s^j| - 1)\right],$$

and $s_{kj} \le \frac{\max_{i,j}|x_{ij}|}{n}|G_{\max}| = \frac{M_1}{n}|G_{\max}|$. By the condition $\lambda \ge \frac{|G_{\max}|M_1}{(|G_{\min}|-1)n}$, when n is large enough. We have $\lambda(|G_s^j| - 1) \ge \frac{|G_s^j|-1}{|G_{\min}|-1}s_{kj} \ge s_{kj}$, which implies that there exists $\boldsymbol{l} = \boldsymbol{l}^*$ s.t.

$$\lambda \sum_{k' \in G_s^j : k'<k} l^*_{k'kj} + \lambda \sum_{k' \in G_s^j : k<k'} l^*_{kk'j} + s_{kj}(\boldsymbol{\nu}^*) = 0. \tag{C.6}$$

From Lemma 4,

$$P\left\{\operatorname{sgn}\left(\widetilde{\beta}_{kj} - \widetilde{\beta}_{k'j}\right) = \operatorname{sgn}\left(\beta_{kj} - \beta_{k'j}\right), k \in G_s^j, k' \notin G_s^j\right\} \to 1, \tag{C.7}$$

and

$$P\left\{\operatorname{sgn}\left(\widetilde{\beta}_{kj}\right) = \operatorname{sgn}\left(\beta_{kj}\right), k \in G_s^j, \phi_s^j = 0\right\} \to 1. \tag{C.8}$$

Therefore, combining (C.6), (C.7) and (C.8), we have

$$\partial g(\widetilde{\boldsymbol{\beta}})/\partial \beta_{kj}|_{\boldsymbol{\nu}^*} = \partial h(\boldsymbol{\beta})/\partial \beta_{kj}$$

for $\boldsymbol{l} = \boldsymbol{l}^*$ w.p.a 1.

(2) When $\tilde{\beta}_{kj} = 0$, $k \in G_s^j$

$$\left.\frac{\partial g(\widetilde{\boldsymbol{\beta}})}{\partial \beta_{kj}}\right|_{\boldsymbol{\nu}^*} = s_{kj}(\boldsymbol{\nu}^*) + \lambda \sum_{k' : k' < k \in G_s^j} l_{k'kj} + \lambda \sum_{k' : k' \notin G_s^j} \operatorname{sgn}(\tilde{\beta}_{kj} - \tilde{\beta}_{k'j}) + \lambda \sum_{k' : k < k' \in G_s^j} l_{kk'j} + \gamma l_{kj}, \tag{C.9}$$

where $l_{kj} = [-1, 1]$. Therefore, by letting $l_{kj} = l_{kj}^* = 0$, similar to the proof of case (1) in the above, we have $\partial g(\widetilde{\boldsymbol{\beta}})/\partial \beta_{kj}|_{\boldsymbol{\nu}^*} = \partial h(\boldsymbol{\beta})/\partial \beta_{kj}$ for $\boldsymbol{l} = \boldsymbol{l}^*$ w.p.a 1.

In sum, $\partial h(\boldsymbol{\beta}) \cap \partial g(\widetilde{\boldsymbol{\beta}}) \neq \emptyset$, we complete our proof. □

Remark 2. In the proof of Lemma 2.3 in Wang et al. (2012), a similar result is obtained by assuming $|s_{kj}| \leq \gamma$, while we do not require this preliminary condition. This is because we introduce a new value $l_{k'kj}$ and set it such that $\lambda \sum_{k' < k \in G_s^j} l_{k'kj}^* + \lambda \sum_{k < k' \in G_s^j} l_{kk'j}^* + s_{kj}(\boldsymbol{\nu}^*) = 0$. This leads to a more relaxed order condition on S.

Appendix D. Proof of Theorem 2

Following Zhang et al. (2019), we classify all candidate models into four types.

True Model A model with true coefficients $\boldsymbol{\phi}_0$, true structure $\mathcal{G}_0 = \{\mathcal{G}_0^j, j = 1, \ldots, p\}$ and number of total subgroups S_0,

$$\boldsymbol{y} = \widetilde{\mathbf{X}}\mathbf{W}_0\boldsymbol{\phi}_0 + \boldsymbol{\varepsilon}. \tag{D.1}$$

Overfitted Model (OF) A model with structure $\mathcal{G}$ and coefficients $\boldsymbol{\phi} \in \mathbb{R}^S, S > S_0$, and structure $\mathcal{G}$ is split from true structure $\mathcal{G}_0$, i.e., $\forall j = 1, \ldots p, \mathcal{G}^j$ is the split structure of $\mathcal{G}_0^j$. We write the overfitted models as

$$\boldsymbol{y} = \widetilde{\mathbf{X}}\mathbf{W}\boldsymbol{\phi} + \boldsymbol{\varepsilon} = \widetilde{\mathbf{X}}(S\boldsymbol{\theta} + \mathbf{W}_0\boldsymbol{\phi}) + \boldsymbol{\varepsilon}, \tag{D.2}$$

where $S \in \mathbb{R}^{n \times (S - S_0)}$ is a sub-selection matrix, and $\boldsymbol{\theta} \in \mathbb{R}^{S - S_0}$. For examples, $p = 1$ and $\mathcal{G}_0 = \{G_0^1\}$ where $G_0^1 = \{G_{0,1}^1, G_{0,2}^1\}$ with $S_0 = 2$, and $\mathcal{G} = \{\mathcal{G}^1\}$ where $\mathcal{G}^1 = \{G_{11}^1, G_{12}^1, G_2^1\}$ and $S = 3$, $S = (a_1, \ldots, a_n)' \in \mathbb{R}^{n \times 1}$, $a_i = 1$ if $i \in G_{11}^1, a_i = 0$, otherwise. Similarly, for any overfitted model $\mathcal{G}$, we could construct the matrix S, and let augmented design matrix $\mathbf{U}_{\mathcal{G}} = (\widetilde{\mathbf{X}}S, \widetilde{\mathbf{X}}\mathbf{W}_0)$.

Underfitted Model (UF) A model with structure $\mathcal{G}$ and coefficients $\boldsymbol{\phi} \in \mathbb{R}^S, S < S_0$, and structure $\mathcal{G}$ is merged from true structure $\mathcal{G}_0$, i.e. $\forall j = 1, \ldots p, \mathcal{G}^j$ is the merged structure of $\mathcal{G}_0^j$. For an underfitted model, we could write it as a sub-model of true model $\mathcal{G}_0$,

$$\boldsymbol{y} = \widetilde{\mathbf{X}}\mathbf{W}\boldsymbol{\phi} + \boldsymbol{\varepsilon} = \widetilde{\mathbf{X}}\mathbf{W}_0(\boldsymbol{\phi}^\intercal, \mathbf{0}_{S_0 - S}^\intercal)^\intercal + \boldsymbol{\varepsilon}, \tag{D.3}$$

Wrongly Assigned Model (WA) A model does not belong to the above three classes.

We define $\mathbb{G}_{OF}$, $\mathbb{G}_{UF}$, and $\mathbb{G}_{WA}$ as the sets of all overfitted, underfitted, and wrongly assigned models, respectively. In addition, we require the following two conditions: C2+ is a stronger version of the original condition C2, while C5 ensures identifiability as presented in Belloni and Chernozhukov (2011) and Zheng et al. (2015).

C2+. There exists a constant A_0 s.t. $\forall u$, $\sup_{\mathbb{X}} |f(u|\mathbb{X}) - f(0|\mathbb{X})| \leq A_0|u|$.

C5. Let $S_U (> S_0)$ be an upper bound of the total number of subgroups. There exists a large constant N such that for all $n > N$,

$$\Lambda_{\min} := \inf_{\mathcal{G} \in \mathbb{G}_{OF}, \|\boldsymbol{\psi}\|_0 \leq S_U, \boldsymbol{\psi} \neq 0} \frac{\boldsymbol{\psi}^\intercal E\left[\mathbf{U}_{\mathcal{G},i}\mathbf{U}_{\mathcal{G},i}^\intercal\right]\boldsymbol{\psi}}{\|\boldsymbol{\psi}\|^2} > 0,$$

$$\Lambda_{\max} := \sup_{\mathcal{G} \in \mathbb{G}_{OF}, \|\boldsymbol{\psi}\|_0 \leq S_U, \boldsymbol{\psi} \neq 0} \frac{\boldsymbol{\psi}^\intercal E\left[\mathbf{U}_{\mathcal{G}}\mathbf{U}_{\mathcal{G}}^\intercal\right]\boldsymbol{\psi}}{\|\boldsymbol{\psi}\|^2} < \infty,$$

$$q' := \inf_{\mathcal{G} \in \mathbb{G}_{OF}, \|\boldsymbol{\psi}\|_0 \leq S_U, \boldsymbol{\psi} \neq 0} \frac{\left\{E\left[\left(\mathbf{U}_{\mathcal{G},i}^\intercal\boldsymbol{\psi}\right)^2\right]\right\}^{3/2}}{E\left[\left|\mathbf{U}_{\mathcal{G},i}^\intercal\boldsymbol{\psi}\right|^3\right]} > 0$$

where $\|\cdot\|_0$ counts the total number of nonzero elements of a vector, $\mathbf{U}_{\mathcal{G},i}$ is the ith row of $\mathbf{U}_{\mathcal{G}}$, and $\boldsymbol{\psi}$ is a vector whose dimension varies with the matrix $\mathbf{U}_{\mathcal{G},i}$.

Proof of Theorem 2. Under structure $\mathcal{G}$, denote $\hat{\sigma}_{\mathcal{G}} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\rho_\tau(y_{ik} - \boldsymbol{x}_{ik}^{\intercal}\widehat{\boldsymbol{\beta}}_k(\mathcal{G}))$ and $\sigma = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\rho_\tau(y_{ik} - \boldsymbol{x}_{ik}^{\intercal}\boldsymbol{\beta}_{0k})$. For an overfitted model $\mathcal{G} \in \mathbb{G}_{OF}$, w.p.a.1,

$$
\begin{aligned}
&\inf_{\mathcal{G}\in\mathbb{G}_{OF},|\mathcal{G}|<S_U} BIC(\mathcal{G}) - BIC\left(\mathcal{G}_0\right)\\
=&\inf_{\mathcal{G}\in\mathbb{G}_{OF},|\mathcal{G}|<S_U}\left[\log\left(\hat{\sigma}_{\mathcal{G}}\right) - \log\left(\hat{\sigma}_{\mathcal{G}_0}\right)\right] + \left(S - S_0\right)\varphi_n\\
(*)\geq&\inf_{\mathcal{G}\in\mathbb{G}_{OF},|\mathcal{G}|<S_U}\min\left(\log 2, \frac{1}{2}\frac{\hat{\sigma}_{\mathcal{G}} - \hat{\sigma}_{\mathcal{G}_0}}{\hat{\sigma}_{\mathcal{G}_0}}\right) + \left(S - S_0\right)\varphi_n\\
(**)\geq&-C\left(\underline{f}\Lambda_{\min}n\right)^{-1}\left(S - q_0\right)\log(S) + \left(S - S_0\right)\varphi_n\\
(***)>&(S - q_0)(o\left(\varphi_n\right) + \varphi_n)\\
>&0,
\end{aligned}
$$

where C is a positive constant and $\underline{f}$ is a uniform lower bound for $f(0|\mathbb{X})$, inequality $(*)$ is because of $\log(1+u) \geq \min(\log 2, 0.5u)$, which can be proved directly by calculus, $(**)$ is a result of Lemma 7.8 and 7.10 in Zheng et al. (2015), and $(***)$ is by the condition that $\log(Kp)/n = o\left(\varphi_n\right)$.

For an underfitted model $\mathcal{G} \in \mathbb{G}_{UF}$, since $|\mathbb{G}_{UF}| < \infty$, we only need to confirm $BIC(\mathcal{G}) - BIC\left(\mathcal{G}_0\right) > 0$ as follows. From Lemma 1 of Lian (2012) and the law of large number, $\left[\log\left(\hat{\sigma}_{\mathcal{G}}\right) - \log\left(\hat{\sigma}_{\mathcal{G}_0}\right)\right]$ is positive bounded away from zero for large enough n, and $(S - S_0)\varphi_n = o_p(1)$. Hence, $BIC(\mathcal{G}) - BIC\left(\mathcal{G}_0\right) = \left[\log\left(\hat{\sigma}_{\mathcal{G}}\right) - \log\left(\hat{\sigma}_{\mathcal{G}_0}\right)\right] + \left(S - S_0\right)\varphi_n > 0$ w.p.a.1.

For the wrongly assigned model $\mathcal{G}$, we construct an intermediate model $\mathcal{G}'$ s.t. $\mathcal{G}' \in \mathbb{G}_{OF}$ and $\mathcal{G}'$ is also overfitted for $\mathcal{G}$. Then w.p.a.1,

$$
\begin{aligned}
&\inf_{\mathcal{G}\in\mathbb{G}_{WA},|\mathcal{G}|<S_U} BIC(\mathcal{G}) - BIC\left(\mathcal{G}_0\right)\\
=&\inf_{\mathcal{G}\in\mathbb{G}_{WA},|\mathcal{G}|<S_U}\left[\log\left(\hat{\sigma}_{\mathcal{G}}\right) - \log\left(\hat{\sigma}_{\mathcal{G}_0}\right)\right] + \left(S - S_0\right)\varphi_n\\
\geq&\inf_{\mathcal{G}\in\mathbb{G}_{WA},|\mathcal{G}|<S_U}\left[\log\left(\hat{\sigma}_{\mathcal{G}}\right) - \log\left(\hat{\sigma}_{\mathcal{G}'}\right)\right] + \inf_{\mathcal{G}\in\mathbb{G}_{WA},|\mathcal{G}|<S_U}\left[\log\left(\hat{\sigma}_{\mathcal{G}'}\right) - \log\left(\hat{\sigma}_{\mathcal{G}_0}\right)\right] + \left(S - S_0\right)\varphi_n\\
>&0.
\end{aligned}
$$

□

References

Anderson, A.L., 2002. Individual and contextual influences on delinquency: The role of the single-parent family. J. Criminal Justice 30 (6), 575–587.

Belloni, A., Chernozhukov, V., 2011. L1-penalized quantile regression in high-dimensional sparse models. Ann. Statist. 39 (1).

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Mach. Learn. 3 (1), 1–122.

Chen, J., Chen, Z., 2008. Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95 (3), 759–771.

Dondelinger, F., Mukherjee, S., The Alzheimer's Disease Neuroimaging Initiative, 2020. The joint lasso: High-dimensional regression for group structured data. Biostatistics 21 (2), 219–235.

Donnelly, P.G., 1989. Individual and neighborhood influences on fear of crime. Sociol. Focus 69–85.

Fan, J., Fan, Y., Barut, E., 2014. Adaptive robust variable selection. Ann. Statist. 42 (1), 324–351.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.

Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (5), 849–911.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Ann. Statist. 32 (3).

Fan, Y., Tang, C.Y., 2013. Tuning parameter selection in high dimensional penalized likelihood. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (3), 531–552.

Gu, Y., Fan, J., Kong, L., Ma, S., Zou, H., 2018. ADMM for high-dimensional sparse penalized quantile regression. Technometrics 60 (3), 319–331.

Guerra, R., Goldstein, D.R., 2016. Meta-Analysis and Combining Information in Genetics and Genomics. Chapman and Hall/CRC.

He, X., Shao, Q.-M., 2000. On parameters of increasing dimensions. J. Multivariate Anal. 73 (1), 120–135.

Huang, Y., Liu, J., Yi, H., Shia, B.C., Ma, S., 2017. Promoting similarity of model sparsity structures in integrative analysis of cancer genetic data. Stat. Med. 36 (3), 509–559.

Kim, Y., Choi, H., Oh, H.-S., 2008. Smoothly clipped absolute deviation on high dimensions. J. Amer. Statist. Assoc. 103 (484), 1665–1673.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46 (1), 33.

Lee, E.R., Noh, H., Park, B.U., 2014. Model selection via Bayesian information criterion for quantile regression models. J. Amer. Statist. Assoc. 109 (505), 216–229.

Li, Z., Luo, Z., Sun, Y., 2022. Robust nonparametric integrative analysis to decipher heterogeneity and commonality across subgroups using sparse boosting. Stat. Med. 41 (9), 1658–1687.

Li, Y., Wang, F., Li, R., Sun, Y., 2020. Semiparametric integrative interaction analysis for non-small-cell lung cancer. Stat. Methods Med. Res. 29 (10), 2865–2880.

Lian, H., 2012. A note on the consistency of Schwarz's criterion in linear quantile regression with the SCAD penalty. Statist. Probab. Lett. 82 (7), 1224–1228.

Liu, J., Huang, J., Zhang, Y., Lan, Q., Rothman, N., Zheng, T., Ma, S., 2014. Integrative analysis of prognosis data on multiple cancer subtypes. Biometrics 70 (3), 480–488.

Ma, S., Huang, J., 2017. A concave pairwise fusion approach to subgroup analysis. J. Amer. Statist. Assoc. 112 (517), 410–423.

Ma, S., Huang, J., Song, X., 2011. Integrative analysis and variable selection with multiple high-dimensional data sets. Biostatistics 12 (4), 763–775.

McGranahan, D.A., 2021. Crime and the countryside. Rural Am./Rural Development Perspect. 2 (2), 2–8.

Meier, L., Van De Geer, S., Bühlmann, P., 2008. The group lasso for logistic regression. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (1), 53–71.

Roth, J.J., 2019. Empty homes and acquisitive crime: Does vacancy type matter? Am. J. Criminal Justice 44 (5), 770–787.

Roth, J.J., 2022. Crime and specific vacancy types in smaller cities and towns. Criminal Justice Stud. 35 (1), 93–109.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.
Shen, Y., Wen, Z., Zhang, Y., 2014. Augmented Lagrangian alternating direction method for matrix separation based on low-rank factorization. Optim. Methods Softw. 29 (2), 239–263.
Singh, A., Kiran, U.V., 2014. Effect of single parent family on child delinquency. Int. J. Sci. Res. 3 (9), 866–868.
Sun, Y., Sun, Z., Jiang, Y., Li, Y., Ma, S., 2020. An integrative sparse boosting analysis of cancer genomic commonality and difference. Stat. Methods Med. Res. 29 (5), 1325–1337.
Tang, L., Song, P.X.K., 2016. Fused lasso approach in regression coefficients clustering – learning parameter heterogeneity in data integration. J. Mach. Learn. Res. 17 (113), 1–23.
Tang, L., Song, P.X.-K., 2020. Poststratification fusion learning in longitudinal data analysis. Biometrics 77 (3), 914–928.
Tang, X., Xue, F., Qu, A., 2020. Individualized multidirectional variable selection. J. Amer. Statist. Assoc. 116 (535), 1280–1296.
Tao, P.D., 1997. Convex analysis approach to DC programming: Theory, algorithms and applications. Acta Math. Vietnam. 22 (1), 289–355.
Taylor, R.B., 1995. The impact of crime on communities. Ann. Am. Acad. Political Soc. Sci. 539 (1), 28–45.
Wang, H., Leng, C., 2007. Unified LASSO estimation by least squares approximation. J. Amer. Statist. Assoc. 102 (479), 1039–1048.
Wang, H., Li, B., Leng, C., 2009. Shrinkage tuning parameter selection with a diverging number of parameters. J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (3), 671–683.
Wang, H., Li, R., Tsai, C.-L., 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94 (3), 553–568.
Wang, L., Wu, Y., Li, R., 2012. Quantile regression for analyzing heterogeneity in ultra-high dimension. J. Amer. Statist. Assoc. 107 (497), 214–222.
Wedel, M., Kamakura, W.A., 2000. Market Segmentation: Conceptual and Methodological Foundations. Springer Science & Business Media.
Yang, P., Hwa Yang, Y., B. Zhou, B., Y. Zomaya, A., 2010. A review of ensemble methods in bioinformatics. Current Bioinf. 5 (4), 296–308.
Yang, X., Yan, X., Huang, J., 2019. High-dimensional integrative analysis with homogeneity and sparsity recovery. J. Multivariate Anal. 174, 104529.
You, J., Jiao, Y., Lu, X., Zeng, T., 2019. A nonconvex model with minimax concave penalty for image restoration. J. Sci. Comput. 78 (2), 1063–1086.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 68 (1), 49–67.
Zhang, C.H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38 (2), 894–942.
Zhang, Y., Wang, H.J., Zhu, Z., 2019. Robust subgroup identification. Statist. Sinica 29, 1873–1889.
Zhao, Q., Shi, X., Huang, J., Liu, J., Li, Y., Ma, S., 2015. Integrative analysis of '-Omics' data using penalty functions. Wiley Interdiscip. Rev. Comput. Stat. 7 (1), 99–108.
Zheng, Q., Peng, L., He, X., 2015. Globally adaptive quantile regression with ultra-high dimensional data. Ann. Statist. 43 (5), 2225–2258.
Zhu, X., Tang, X., Qu, A., 2021. Longitudinal clustering for heterogeneous binary data. Statist. Sinica 31, 603–624.