Received: 12 November 2021 | Revised: 4 June 2022 | Accepted: 5 September 2022

DOI: 10.1002/sim.9578

RESEARCH ARTICLE

Statistics in Medicine WILEY

Robust analysis of cancer heterogeneity for high-dimensional data

Chao Cheng | **Xingdong Feng** | **Xiaoguang Li** | **Mengyun Wu**

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

Correspondence
Mengyun Wu, School of Statistics and Management, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, China.
Email: wu.mengyun@mail.shufe.edu.cn

Funding information
Fundamental Research Funds for the Central Universities, Grant/Award Number: CXJJ-2019-403; National Natural Science Foundation of China, Grant/Award Numbers: 11971292, 12071273; Shanghai Research Center for Data Science and Decision Technology; Shanghai Rising-Star Program, Grant/Award Number: 22QA1403500

Cancer heterogeneity plays an important role in the understanding of tumor etiology, progression, and response to treatment. To accommodate heterogeneity, cancer subgroup analysis has been extensively conducted. However, most of the existing studies share the limitation that they cannot accommodate heavy-tailed or contaminated outcomes and also high dimensional covariates, both of which are not uncommon in biomedical research. In this study, we propose a robust subgroup identification approach based on M-estimators together with concave and pairwise fusion penalties, which advances from existing studies by effectively accommodating high-dimensional data containing some outliers. The penalties are applied on both latent heterogeneity factors and covariates, where the estimation is expected to achieve subgroup identification and variable selection simultaneously, with the number of subgroups being *apriori* unknown. We innovatively develop an algorithm based on parallel computing strategy, with a significant advantage of capable of processing large-scale data. The convergence property of the proposed algorithm, oracle property of the penalized M-estimators, and selection consistency of the proposed BIC criterion are carefully established. Simulation and analysis of TCGA breast cancer data demonstrate that the proposed approach is promising to efficiently identify underlying subgroups in high-dimensional data.

KEYWORDS
parallel computing, penalized fusion, robust estimation, subgroup analysis, variable selection

1 | INTRODUCTION

Cancer heterogeneity has received intensive attention in recent biomedical research, which is crucial for understanding tumor etiology, progression, and response to treatment.[1] For example, in precision medicines and individualized treatment designs, it is widely recognized that treatment effects may vary among different patient subgroups.[2] Cancer is extremely complex, of which the heterogeneity may be attributed to many factors, including clinical, environmental, and demographic factors with usually low-dimension, as well as genetic factors with high-dimension.[3] Thus, heterogeneity analysis for identifying cancer subgroups remains a challenging task in cancer modeling.

Recently, a number of statistical methods have been developed for cancer subgroup analysis. Among them, it is popular to consider data as coming from mixture models, including Atienza et al and Shen et al with low-dimensional covariates,[4,5] and Khalili et al and Ren et al with high-dimensional covariates.[6,7] Mixture models can easily incorporate covariate effects due to its theoretical framework, but the number of subgroups and the corresponding underlying distributions are

usually pre-specified, which is hard to be verified in practical cancer modeling. To address this issue, penalization methods have been developed which can identify hidden subgroups and estimate their centers simultaneously, without any prior information of true subgroup structures. Examples include Chi et al and Wu et al from an unsupervised clustering perspective,[8,9] and Ma et al and Chen et al based on a supervised regression model.[10,11] Here the supervised heterogeneity analysis studies effectively make use of the outcome-predictor relationship and often have more important practical implications. They usually model heterogeneity by unobserved subject-specific latent factors after adjusting for the effects of a set of observed covariates such as gender, age, and so on.[10,11] Despite considerable successes, these studies are limited by only exploring low-dimensional covariates and cannot accommodate high-dimensional data, such as genetic factors. The genetic factors also play an important role in cancer heterogeneity, and the conclusions regarding the cancer subgroups are likely confounded by inappropriately overlooking the genetic effects.[3] High dimensionality of genetic data poses many challenges for cancer subgroup analysis, both computationally and theoretically.

Another challenge for supervised heterogeneity analysis comes from the heavy-tailed distributions and contamination in responses, which is not uncommon in practical cancer studies and can be caused by multiple factors, including the inherent variability of data, biased sample selection, and mistakes in data recording, among others.[12] Take the TCGA (The Cancer Genome Atlas) breast cancer data analyzed in Section 5 of this article as an example, we observe heavier right tails for the response in Figure 1A. However, most of the existing supervised heterogeneity studies, including the aforementioned, are based on the least squares loss and cannot accommodate heavy-tails and outliers, leading to

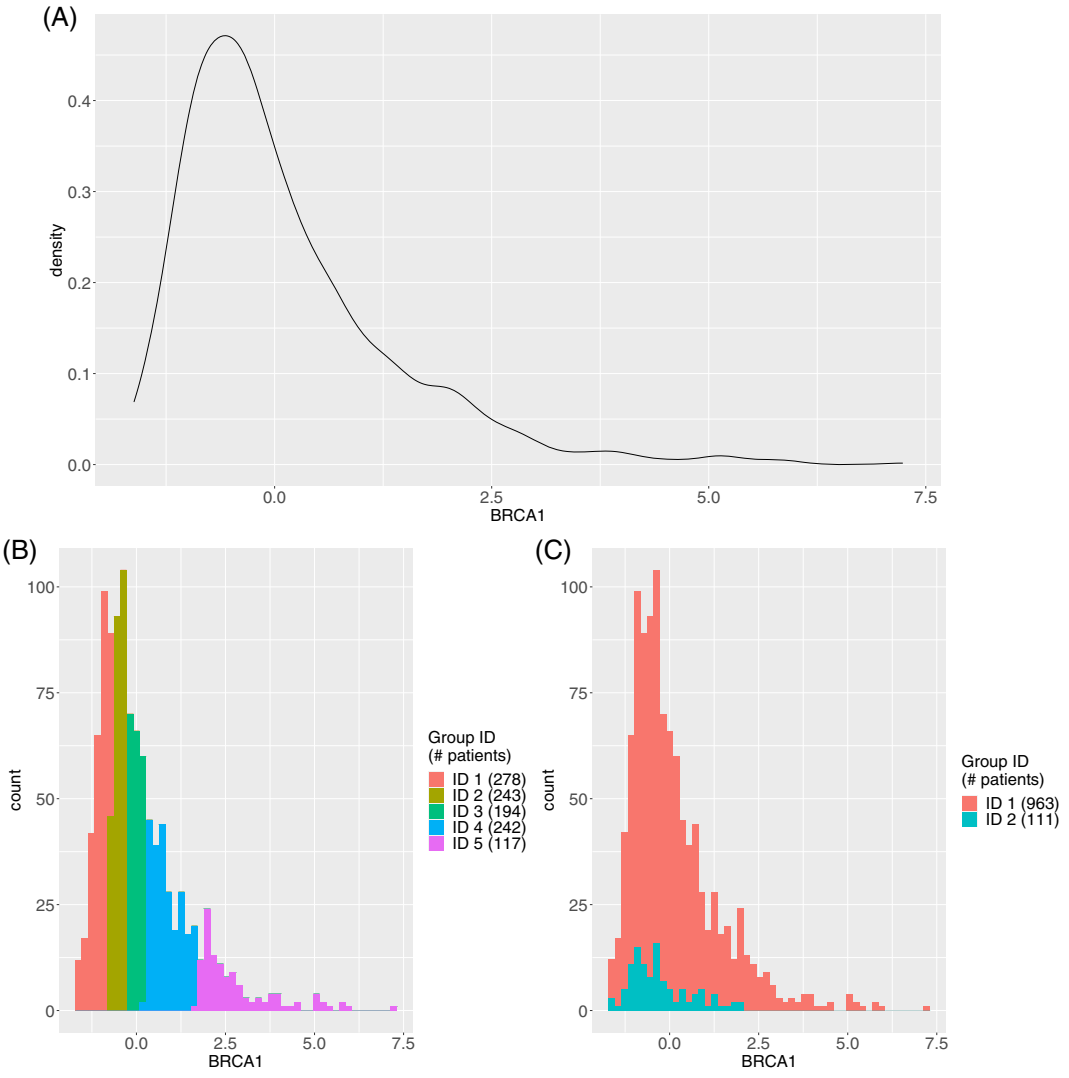

FIGURE 1 Analysis of BRCA data: density plot for mRNA expression of BRCA1 and estimated subgroup results with the proposed approach and RSI are shown in (A–C)

inaccurate subgroup identification and biased estimation. The limited robust study is proposed by Zhang et al,[13] which is an extension of Ma et al[10] and adopts the median regression to achieve robustness, but also focuses on low-dimensional covariates.

In this article, we propose a novel robust cancer subgroup identification approach with certain variable selection method to simultaneously choose useful variables from a large number of candidate covariates. The proposed approach is under the similar framework as Ma et al and Zhang et al that identifies subject subgroups via a pairwise fusion penalty.[10,13] It not only shares similar desirable properties as the existing ones, including automatically determining the number of subgroups and without making assumptions on specific underlying data distributions, but also advances from them in multiple aspects. First, the proposed approach is built on the robust M-regression which enjoys satisfactory robustness properties toward heavy-tailed errors and outliers. Although M-regression has been a popular tool in statistical analysis, its applications to cancer subgroup analysis are still limited. It has a solid statistical basis and includes the least absolute deviation estimator of Zhang et al as a special case.[13] Second, we analyze high dimensional covariates which are common in cancer studies and adopt the penalization technique for regularized estimation and variable selection, improving not only accuracy of subgroup identification but also interpretability of the model. Third, a parallel computing strategy based on an alternating direction method of multipliers (ADMM) algorithm is developed to speed up the computation. This is much desirable for large-scale cancer data, where the pairwise fusion penalty makes the complexity of ordinary algorithms in existing studies at the quadratic order of the sample size. In addition, we derive the convergence property of our algorithm and establish the oracle properties for both subgroup recovery and variable selection with a variety of penalties, including smoothly clipped absolute deviation (SCAD) and minimax concave penalty (MCP), providing theoretical results in a more general sense. Overall, this study provides an effective way of conducting robust identification of cancer subgroups for high-dimensional data.

2 | METHODS

Consider n independent subjects, which are collected from the unknown K different subgroups. For the ith subject, denote $\boldsymbol{x}_i$ as the p-dimensional covariate vector and y_i as the continuous response. We consider the following linear regression model

$$y_i = \mu_i + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{1}$$

where μ_i is the unobserved latent heterogeneity effect attributable to the i-th subject, such as the individual treatment effect, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the regression coefficient vector, ε_i's are the independently distributed errors satisfying $\mathrm{E}\psi(\varepsilon_i) = 0$ with $\psi(\cdot)$ being the derivative (or directional derivative) of $\rho(\cdot)$, and $\rho(\cdot)$ is a robust loss function which is differentiable except at finitely many points, for example, L_1 loss with $\rho(t) = |t|$ and Huber loss with $\rho(t) = \begin{cases} \frac{t^2}{2} & \text{if } |t| \le \xi \\ \xi|t| - \frac{\xi^2}{2} & \text{if } |t| > \xi \end{cases}$. Here, we assume $\mathrm{E}\psi(\varepsilon_i) = 0$, which is a commonly used condition for model identification in the M-regression literature.[14,15] The ordinary linear regression is a special case of M-regression with $\rho(t) = t^2$, and accordingly $\psi(t) = 2t$ and $\mathrm{E}\psi(\varepsilon_i) = 2\mathrm{E}(\varepsilon_i) = 0$.

For identifying subject subgroups and important covariates simultaneously, we propose the following penalized objective function,

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(y_i - \mu_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right) + \sum_{1\le i<j\le n} P_{\lambda_1,\gamma}\left(|\mu_i - \mu_j|\right) + \sum_{j=1}^{p} P_{\lambda_2,\gamma}\left(|\beta_j|\right), \tag{2}$$

where $P_{\lambda,\gamma}(\cdot)$ is the non-negative penalty function with a tuning parameter λ and a regularization parameter γ, and two parameters λ_1 and λ_2 are adopted for μ_i's and β_j's, respectively. In our study, we examine the proposed approach using the well-known SCAD and MCP penalties, with

$$P_{\lambda,\gamma}(x) = \lambda \int_0^x \min\left\{1, \frac{(\gamma - t/\lambda)_+}{(\gamma - 1)}\right\} \mathrm{d}t, \qquad x > 0, \quad \gamma > 2,$$

for SCAD and

$$P_{\lambda,\gamma}(x) = \lambda \int_0^x \left(1 - \frac{t}{\lambda\gamma}\right)_+ \mathrm{d}t, \qquad x > 0, \quad \gamma > 1,$$

for MCP, where the parameter γ controls the concavity of the penalty functions. In our study, we set $\gamma = 3.7$ for SCAD and 3.0 for MCP, suggested by Fan et al and Breheny et al, respectively.[16,17] The proposed estimates of μ_i's and β_j's are defined as the minimizer of (2). The subjects with the same values of μ_i belong to the same subgroup, and the nonzero coefficients β_j's correspond to the important covariates that are associated with the response.

In (2), as opposed to the nonrobust least squares loss developed in Ma et al and He et al,[10,18] we adopt a robust loss function based on M-regression, which has satisfactory theoretical and numerical performance even in the presence of heavy-tailed errors and outlier contamination. The proposed objective function is more general and includes L_1 loss function adopted in Zhang et al as a special example.[13] The pairwise fusion penalty is imposed on the pairwise differences $|\mu_i - \mu_j|$'s to identify subject subgroups, where $|\mu_i - \mu_j|$'s are shrunk toward zero. As such, the proposed approach can partition subjects into multiple subgroups where subjects in each subgroup has the same values of μ_i. This strategy has the advantage of automatically determining the number of subgroups, which is still a nontrivial task in heterogeneity analysis. In addition, penalization technique is also adopted to accommodate high dimensionality of covariates and identify the important covariates, which has not been well investigated in published studies, such as Ma et al and Zhang et al.[10,13] We would like to point out that different from the work developed in He et al which models the heterogeneity based on the observed covariates $\boldsymbol{x}_i$,[18] we consider the heterogeneity resulting from unobserved latent factors after adjusting for the effects of $\boldsymbol{x}_i$ and its overall effect is represented in μ_i. Regarding the penalties, we adopt SCAD and MCP, which have been popular and well-studied in the literature, and demonstrated to perform better theoretically and numerically than some popular penalties, such as Lasso.

2.1 | Computation

For optimizing the objective function (2), we propose adopting the ADMM algorithm, which has been widely used in statistical learning and optimization problems, especially for large-scale data analysis. It changes a large optimization problem into smaller ones and takes advantages of the augmented Lagrangian and coordinate descent methods. First we rewrite the objective function (2) as

$$\begin{aligned} &\min \ \frac{1}{n}\sum_{i=1}^{n}\rho(z_i) + \sum_{1\le i<j\le n} P_{\lambda_1,\gamma}\left(|s_{ij}|\right) + \sum_{j=1}^{p} P_{\lambda_2,\gamma}\left(|w_j|\right) \\ &\text{s.t.} \ \ \boldsymbol{z} = \boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}, \ \ \boldsymbol{s} = \boldsymbol{D}\boldsymbol{\mu}, \ \ \boldsymbol{w} = \boldsymbol{\beta}, \end{aligned}$$

where $\boldsymbol{X}$ is the matrix composed of $\boldsymbol{x}_i$'s, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, and

$$\boldsymbol{D} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & \cdots \\ 1 & 0 & -1 & 0 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & 1 & -1 \end{pmatrix}$$

is the $\frac{n(n-1)}{2} \times n$ pairwise difference matrix such that $s_{ij} = \mu_i - \mu_j$. The augmented Lagrangian form of this problem is then given by

$$\begin{aligned} L\left(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{w}, \boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3\right) = &\frac{1}{n}\sum_{i=1}^{n}\rho(z_i) + \sum_{1\le i<j\le n} P_{\lambda_1}\left(|s_{ij}|\right) + \sum_{j=1}^{p} P_{\lambda_2}\left(|w_j|\right) \\ &+ \frac{r_1}{2}\|\boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{z}\|_2^2 + \frac{r_2}{2}\|\boldsymbol{D}\boldsymbol{\mu} - \boldsymbol{s}\|_2^2 + \frac{r_3}{2}\|\boldsymbol{\beta} - \boldsymbol{w}\|_2^2 \\ &+ \langle \boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{z}, \boldsymbol{q}_1\rangle + \langle \boldsymbol{D}\boldsymbol{\mu} - \boldsymbol{s}, \boldsymbol{q}_2\rangle + \left\langle \boldsymbol{\beta} - \boldsymbol{w}, \boldsymbol{q}_3\right\rangle, \end{aligned} \tag{3}$$

where r_1, r_2, and r_3 are positive scalars, $\boldsymbol{q}_1$, $\boldsymbol{q}_2$ and $\boldsymbol{q}_3$ are multiplier vectors, $\|\boldsymbol{a}\|_2^2 = \boldsymbol{a}^T\boldsymbol{a}$, and $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \boldsymbol{a}^T\boldsymbol{b}$.

We use an iterative algorithm to solve (3). Specifically, at each iteration optimize the objective function (3) with respect to $\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{z}, \boldsymbol{s}$, and $\boldsymbol{w}$ sequentially while fixing the other parameters at their current estimates, and then update $\boldsymbol{q}_1, \boldsymbol{q}_2$, and $\boldsymbol{q}_3$ accordingly with r_1, r_2 and r_3. Here, r's (r_1, r_2, and r_3) balance the penalty between violations of primal and dual feasibility in ordinary ADMM algorithm.[19] We set $r = 1 + 1/(\gamma - 1)$ for SCAD penalty and $r = 1 + 1/\gamma$ for MCP to guarantee the feasibility of the ADMM algorithm. The overall algorithm is summarized in Algorithm 1 of the Supporting Information, and the detailed updating steps for various choices of loss functions and penalties are provided in Appendix A of the Supporting Information. In addition, the convergence property of the proposed algorithm is also studied, where we show that the proposed algorithm will converge to a stationary point under some mild conditions.

We further develop a parallel computing strategy for the ADMM algorithm to boost up the computational speed. Specifically, since $\boldsymbol{z}$, $\boldsymbol{s}$, and $\boldsymbol{w}$ are updated elementwisely, these steps can be naturally realized in a parallel manner. In addition, the most of the updating of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ consists of matrix multiplications, which can also benefit from parallel computing significantly for a large-scale data, as their computation complexity grows exponentially with n. In Figure S1 of the Supporting Information, we examine the speed of the single-thread v_{single} and its parallel version $v_{\text{parrallel}}$ of the proposed algorithm. Specifically, we consider the sample size n from 250 to 10 000 and the covariate size $p = n/2$ with other parameters given and fixed, and provide the speed ratio $r_v = v_{\text{parrallel}}/v_{\text{single}}$ for the overall solver, updating of $\boldsymbol{\mu}$, and updating of $\boldsymbol{s}$, respectively. These analyses are conducted on an eight core laptop. It is observed that when the sample size is small, the overhead of creating and managing multiple threads outweights the benefit. However as the sample size grows, the updating of $\boldsymbol{\mu}$ and $\boldsymbol{s}$ gets nearly two and four times speed boost over the single-thread version of the algorithm, respectively, and the speed boost for the overall solver is around 175%. The proposed parallel computing is much desirable for the analysis with a large sample size. To facilitate data analysis and applications beyond this study, we have developed an R package *rsavs* implementing the proposed parallel computing ADMM algorithm and made it publicly available at https://fenguoerbian.github.io/RSAVS.

2.2 | Selection of tuning parameters

We use a modified BIC to choose the tuning parameters λ_1 and λ_2, which is defined as

$$\text{BIC}(\hat{\boldsymbol{\mu}}(\lambda), \hat{\boldsymbol{\beta}}(\lambda)) = \log\left(\frac{1}{n}\sum_{i=1}^{n}\rho\left(y_i - \hat{\mu}_i(\lambda) - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}(\lambda)\right)\right) + \text{df}(\lambda)\phi_n,$$

where $\lambda = (\lambda_1, \lambda_2)$, $\hat{\mu}_i(\lambda)$, and $\hat{\boldsymbol{\beta}}(\lambda)$ are the estimators of μ_i's and $\boldsymbol{\beta}$ by solving (2) given λ, $\text{df}(\lambda) = \hat{K}(\lambda) + |\hat{\boldsymbol{\beta}}(\lambda)|_0$ is the degree of freedom of the model with $\hat{K}(\lambda)$ being the estimator of the number of subgroups based on $\mu_i(\lambda)$'s, and ϕ_n is a constant that can depend on the sample size n. We choose the tuning parameter combination $\lambda = (\lambda_1, \lambda_2)$ by minimizing this modified BIC with the grid search. In our numerical studies, to speed up the tuning selection procedure and improve the stability, we construct a decreasing sequence of λ values, and the estimates of the previous values of λ are used as the warm start.

3 | ASYMPTOTIC PROPERTIES

In this section, we consider the scenario where the number of covariates p can grow even faster than the sample size n, and give the asymptotic properties of the proposed estimator.

First, denote $G_1, \ldots, G_K$ as the subject index sets of the K subgroups and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^T$ as the corresponding heterogeneity effects. Then, we have $\mu_i = \alpha_k$, if $i \in G_k$ and model (1) can also be written as

$$y_i = \boldsymbol{g}_i^T\boldsymbol{\alpha} + \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i,$$

where $\boldsymbol{g}_i$'s are the K-dimensional indicator vectors with $g_{ik} = 1$ if $i \in G_k$ and $g_{ik} = 0$ if $i \notin G_k$, and $\boldsymbol{G}$ is the $n \times K$ matrix composed of $\boldsymbol{g}_i$'s.

Let $\boldsymbol{\alpha}^0$, $\boldsymbol{g}_i^0$'s, and $\boldsymbol{\beta}^0$ be the true parameter values, K_0 be the true number of the subject subgroups, and $G_1^0, \ldots, G_{K_0}^0$ be the true subject index sets of the K_0 subgroups. We assume the sparsity structure is present, which means only a

small portion of the elements of $\boldsymbol{\beta}^0$ is non-zero. To simplify notations, we write $\boldsymbol{\beta}^0 = \left((\boldsymbol{\beta}_A^0)^T, \mathbf{0}_{p-q}^T\right)^T$, and accordingly, the covariate vector is given as $\boldsymbol{x}_i = \left(\boldsymbol{x}_{i,A}^T, \boldsymbol{x}_{i,I}^T\right)^T$, where A and I refer to the true active and inactive sets of covariate indices respectively.

When the true subgroup membership indicator matrix $\boldsymbol{G}^0$ and sparse structure A of $\boldsymbol{\beta}$ are known, the oracle estimate of coefficient $\left(\tilde{\boldsymbol{\alpha}}(\boldsymbol{G}^0)^T, \tilde{\boldsymbol{\beta}}^T\right)^T = \left(\tilde{\boldsymbol{\alpha}}(\boldsymbol{G}^0)^T, \left(\tilde{\boldsymbol{\beta}}_A^T, \mathbf{0}_{p-q}^T\right)^T\right)^T$ is given by

$$\left(\tilde{\boldsymbol{\alpha}}(\boldsymbol{G}^0)^T, \tilde{\boldsymbol{\beta}}_A^T\right)^T = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}_A} \frac{1}{n}\sum_{i=1}^{n} \rho\left(y_i - \left(\mathbf{g}_i^0\right)^T\boldsymbol{\alpha} - \boldsymbol{x}_{i,A}^T\boldsymbol{\beta}_A\right), \tag{4}$$

which leads to the corresponding oracle estimates of subgroup treatment effects $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_n)^T$ with $\tilde{\mu}_i = \tilde{\alpha}_k(\boldsymbol{G}^0)$ if $i \in G_k^0$. Here, we put $\boldsymbol{G}^0$ in parenthese to indicate the dependence of $\boldsymbol{\alpha}$ on it. In Appendix B of the Supporting Information, we introduce the assumptions, which are imposed on the size of the smallest signal, the characteristics of the predictor matrix, and the orders of λ_1, λ_2, and p. Similar conditions have been considered in Ma et al, Wang et al, and some others.[10,13,20] We refer to Appendix B of the Supporting Information for more detailed discussions.

Theorem 1. *Under Assumptions A1–A5 given in the Supporting Information, if the following conditions hold,*

$$\max(\lambda_1, \lambda_2) = o\left(n^{-(1-c_2)/2}\right), \sqrt{q(K_0+q)} = o\left(\sqrt{n}\lambda_2\right), (K_0+q)\log n = o(n\lambda_2),$$
$$n\lambda_2^2 \to \infty, \log p = o\left(n\lambda_2^2\right), \text{ and } \sqrt{\log n}/(n\lambda_1 \min\{|G_k^0|, k = 1, \ldots, K_0\}) = o(1),$$

then there exists a local minimizer $\left(\hat{\boldsymbol{\mu}}^T, \hat{\boldsymbol{\beta}}^T\right)^T$ *of the objective function (2) coupled with either SCAD or MCP penalty such that*

$$P\left(\left(\hat{\boldsymbol{\mu}}^T, \hat{\boldsymbol{\beta}}^T\right)^T = \left(\tilde{\boldsymbol{\mu}}^T, \tilde{\boldsymbol{\beta}}^T\right)^T\right) \to 1$$

as $n \to \infty$, *where* $\left(\tilde{\boldsymbol{\mu}}^T, \tilde{\boldsymbol{\beta}}^T\right)^T$ *is the oracle estimate.*

Proof is provided in Appendix B of the Supporting Information. Theorem 1 shows that the oracle estimate lies in the set of local minimizers of the penalized objective function (2) with a high probability. In Appendix B of the Supporting Information, we also show that the modified BIC provides asymptotic model selection consistency, that is, for any λ that leads to the wrong model structure and λ_0 that identifies the true model, we have BIC $\left(\hat{\boldsymbol{\mu}}(\lambda), \hat{\boldsymbol{\beta}}(\lambda)\right) >$ BIC $\left(\hat{\boldsymbol{\mu}}(\lambda_0), \hat{\boldsymbol{\beta}}(\lambda_0)\right)$ with a high probability.

4 | SIMULATION STUDIES

In this section we compare the performance of the proposed method with its competitors under different scenarios. Specifically, the data are independently generated from model (1). Under different scenarios, we consider different combinations of the sample size n, the number of covariates p, the number of true signals q, the vector of heterogeneity effects $\boldsymbol{\mu}$, the vector of covariate effects $\boldsymbol{\beta}$, and the distribution of errors.

Throughout this simulation study, we examine two specific robust M-estimators based on L_1 and Huber losses in (2), respectively. Moreover, two alternatives are also analyzed. The first one is model (2) with the nonrobust L_2 loss which can hardly accommodate heavy-tailed errors and outliers. The second one is the RSI approach based on L_1 loss,[13] which also has robustness property but conduct no variable selection procedures for high-dimensional covariates. For selecting tuning parameters, we consider the modified BIC defined in (2.2) with $\phi_n = c\log n\log\log(n+p)/n$, where c is set to be 5.0 for L_1 loss following Zhang et al,[13] and also 5.0 for Huber loss, and 10.0 for L_2 loss as suggested by Ma et al[10] For Huber loss, the default value $\xi = 1.345$ is utilized to control the robustness. To reduce computational cost, for the proposed approach, we set the max number of iterations to be 50. In this article, we only report the results of these approaches with the SCAD penalty, and the performance of those with the MCP penalty is quite similar and omitted to save the space.

For each scenario, 500 replicates are simulated and the following measures are adopted for evaluation. (a) $\text{MAE}_{\boldsymbol{\mu}}$ and $\text{MAE}_{\boldsymbol{\beta}}$: the mean absolute errors for the heterogeneity and covariate effects, given by $\text{MAE}_{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} |\hat{\mu}_i - \mu_i|$ and $\text{MAE}_{\boldsymbol{\beta}} = \frac{1}{p}\sum_{i=1}^{p} \left|\hat{\beta}_i - \beta_i\right|$, respectively. The average values over 500 replicates are reported. (b) $\overline{K}$ and $\tilde{K}$: the average and median values of the estimated number of subgroups over 500 replicates. (c) $\overline{q}$: the average value of the estimated number of active covariates over 500 replicates. (d) $\overline{q}_{TP}$: the average value of the number of identified true positives over 500 replicates. (e) $\overline{RI}$: the average value of *RI* (Rand Index) over 500 replicates, which describes how close two grouping results are. Specifically, for the subjects $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, and the estimated and true subgroup memberships $\mathcal{G} = \{G_1, \ldots, G_K\}$ and $\mathcal{G}^0 = \left\{G_1^0, \ldots, G_{K_0}^0\right\}$, *RI* is defined as $RI\left(\mathcal{G}, \mathcal{G}^0\right) = \frac{2}{n(n-1)} \times \sum_{1\le i<j\le n} \gamma_{ij}$, where $\gamma_{ij} = 1$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same subgroup or different subgroups in both $\mathcal{G}$ and $\mathcal{G}^0$, and $\gamma_{ij} = 0$ otherwise. The *RI* value ranges from 0 to 1 with a larger value indicating better clustering performance.

4.1 | Subgroup analysis for low-dimensional cases

We first consider $n = 200$ or 400 and $p = q = 5$ to examine the feasibility of the proposed approach under the scenarios with low dimensional covariates. Note that these low dimensional scenarios favor RSI since the variable selection is not necessary. Specifically, consider $K = 2$ with centers $\{-1, 1\}$ and $K = 3$ with centers $\{-2, 0, 2\}$ to generate subgroups respectively. The values of μ_i's are independently generated from the multinomial distribution, with equal probability at each group center. The covariates $\boldsymbol{x}_i$'s are independently generated from the standard normal distribution and the corresponding coefficients are set to be $\boldsymbol{\beta} = \mathbf{1}_5$. Three settings for the errors ε_i's are considered: (i) the standard normal distribution $\text{N}\left(0, 0.5^2\right)$; (ii) $0.5 \times t(5)$ with $t(5)$ being the t-distribution with five degrees of freedom; (iii) Gaussian mixture $0.95 \times \text{N}\left(0, 0.5^2\right) + 0.05 \times \text{N}\left(0, 5^2\right)$. We would like to note that under settings (ii) and (iii) with heavy-tailed errors or outliers, approximately 2% to 10% of the data will be closer to other subgroup centers instead of its real center.

The summary results under settings with errors drawn from $0.5 \times t(5)$ are reported in Table 1 and the rest results are summarized in Appendix C of the Supporting Information. We can observe that under settings with normally distributed errors, all approaches perform well and comparably. When the errors are generated from the distributions with thick tails, the nonrobust approach with L_2 loss is seriously affected by outliers, while L_1 and Huber losses provide much more robust results. Although penalization for covariates is not necessary under these scenarios, the proposed approaches still behave competitively compared to RSI and can always identify the five active covariates accurately.

Here, we also conduct analysis using the proposed approach with L_1 loss and up to 1000 iterations in the ADMM approximations (denoted as L_1^{1000} in Table 1). Although more iterations often lead to better results, the improvement is relatively limited comparing to the one up to 50 iterations (denoted as L_1 in Table 1). Thus, we conduct analysis with the max number of iterations being 50 in the following numerical studies to reduce computational cost.

4.2 | Subgroup analysis for high-dimensional cases

In this subsection, we consider $n \in \{200, 400\}$, $p \in \{50, 100, 500\}$, $q = 5$ and the covariate coefficient $\boldsymbol{\beta} = \left(\mathbf{1}_5^T, \mathbf{0}_{p-5}^T\right)^T$. The generation mechanism for the covariates $\boldsymbol{x}_i$'s, subgroup intercepts μ_i's, and errors are the same as those in Section 4.1, where the errors with heavy-tails or outliers are considered. The summary results under settings with errors generated from the distribution $0.5 \times t(5)$ are summarized in Tables 1 and 2 (the results for RSI are not available when $p = 500$ since it is no longer feasible when $p > n$). The rest results are given in Appendix C of the Supporting Information.

As demonstrated in Tables 1 and 2, the method with L_2 loss performs poorly since it can hardly tolerate outliers, where it always miss the true subgroup structure and fails to recover the active covariates. The proposed method with either L_1 or Huber losses performs well, which can effectively identify both the subgroup structure and the active covariates. In the high-dimensional cases, the performance of RSI decays significantly with the increasing of the covariate dimension p due to the absence of an appropriate variable selection procedure. Under the scenarios with $n < p$ in Table 2, the proposed method with either L_1 or Huber losses can still have satisfactory performance.

TABLE 1 Simulation results under the scenarios with low dimensional covariates $p = 5$ and high dimensional covariates $p = 50$, $q = 5$, and $0.5 \times t(5)$ distributed error over 500 replicates

K	Method	MAE_μ	MAE_β	$\overline{K}$	$\tilde{K}$	$\overline{q}$	$\overline{q}_{TP}$	$\overline{RI}$
$n = 200, p = 5$								
2	L_1	0.172 (0.044)	0.045 (0.017)	2.000 (0.000)	2	5.000 (0.000)	5.000 (0.000)	0.890 (0.031)
	L_2	0.473 (0.385)	0.056 (0.026)	1.944 (1.226)	2	5.000 (0.000)	5.000 (0.000)	0.748 (0.186)
	Huber	0.169 (0.068)	0.044 (0.018)	1.996 (0.063)	2	5.000 (0.000)	5.000 (0.000)	0.890 (0.039)
	RSI	0.179 (0.126)	0.069 (0.059)	2.010 (0.184)	2	5.000 (0.000)	5.000 (0.000)	0.891 (0.040)
	L_1^{1000}	0.193 (0.078)	0.049 (0.022)	2.156 (0.478)	2	5.000 (0.000)	5.000 (0.000)	0.877 (0.051)
3	L_1	0.734 (0.155)	0.105 (0.042)	2.078 (0.283)	2	5.000 (0.000)	5.000 (0.000)	0.747 (0.049)
	L_2	0.738 (0.501)	0.087 (0.038)	3.504 (2.326)	3	5.000 (0.000)	5.000 (0.000)	0.672 (0.246)
	Huber	0.332 (0.192)	0.075 (0.038)	3.038 (0.567)	3	5.000 (0.000)	5.000 (0.000)	0.868 (0.057)
	RSI	0.400 (0.274)	0.106 (0.077)	2.746 (0.508)	3	5.000 (0.000)	5.000 (0.000)	0.853 (0.081)
	L_1^{1000}	0.323 (0.125)	0.080 (0.041)	3.420 (0.871)	3	5.000 (0.000)	5.000 (0.000)	0.866 (0.045)
$n = 400, p = 5$								
2	L_1	0.146 (0.029)	0.031 (0.011)	2.000 (0.000)	2	5.000 (0.000)	5.000 (0.000)	0.899 (0.019)
	L_2	0.365 (0.180)	0.041 (0.015)	4.916 (1.969)	6	5.000 (0.000)	5.000 (0.000)	0.752 (0.099)
	Huber	0.146 (0.029)	0.030 (0.010)	2.000 (0.000)	2	5.000 (0.000)	5.000 (0.000)	0.898 (0.019)
	RSI	0.179 (0.189)	0.061 (0.100)	2.122 (0.851)	2	5.000 (0.000)	5.000 (0.000)	0.889 (0.061)
	L_1^{1000}	0.157 (0.036)	0.033 (0.013)	2.110 (0.313)	2	5.000 (0.000)	5.000 (0.000)	0.890 (0.024)
3	L_1	0.323 (0.236)	0.048 (0.027)	2.788 (0.409)	3	5.000 (0.000)	5.000 (0.000)	0.871 (0.073)
	L_2	0.427 (0.355)	0.054 (0.024)	4.482 (2.149)	3	5.000 (0.000)	5.000 (0.000)	0.806 (0.171)
	Huber	0.219 (0.106)	0.042 (0.018)	3.002 (0.233)	3	5.000 (0.000)	5.000 (0.000)	0.901 (0.032)
	RSI	0.302 (0.333)	0.103 (0.171)	3.080 (0.975)	3	5.000 (0.000)	5.000 (0.000)	0.885 (0.069)
	L_1^{1000}	0.597 (0.110)	0.067 (0.024)	3.056 (0.744)	3	5.000 (0.000)	5.000 (0.000)	0.772 (0.026)
$n = 200$ and $p = 50$								
2	L_1	0.220 (0.070)	0.006 (0.005)	2.000 (0.000)	2	4.990 (0.161)	4.990 (0.161)	0.850 (0.044)
	L_2	0.994 (0.014)	0.100 (0.000)	1.000 (0.000)	1	0.000 (0.000)	0.000 (0.000)	0.501 (0.000)
	Huber	0.289 (0.236)	0.009 (0.008)	1.970 (0.371)	2	5.322 (0.805)	4.992 (0.126)	0.822 (0.115)
	RSI	0.591 (0.181)	0.094 (0.025)	4.118 (2.561)	4	50.000 (0.000)	5.000 (0.000)	0.658 (0.092)
3	L_1	0.812 (0.080)	0.018 (0.013)	2.000 (0.000)	2	4.788 (0.572)	4.768 (0.554)	0.722 (0.024)
	L_2	1.426 (0.038)	0.100 (0.000)	1.004 (0.063)	1	0.000 (0.000)	0.000 (0.000)	0.333 (0.004)
	Huber	0.626 (0.278)	0.019 (0.016)	2.486 (0.520)	2	5.058 (1.112)	4.762 (0.728)	0.773 (0.083)
	RSI	0.976 (0.152)	0.142 (0.032)	4.204 (2.363)	4	50.000 (0.000)	5.000 (0.000)	0.674 (0.030)
$n = 400$ and $p = 50$								
2	L_1	0.158 (0.034)	0.003 (0.001)	2.000 (0.000)	2	5.000 (0.000)	5.000 (0.000)	0.887 (0.024)
	L_2	1.000 (0.001)	0.008 (0.017)	1.000 (0.000)	1	4.840 (0.881)	4.840 (0.881)	0.499 (0.000)
	Huber	0.169 (0.068)	0.004 (0.003)	2.020 (0.189)	2	5.006 (0.077)	5.000 (0.000)	0.881 (0.037)
	RSI	0.519 (0.331)	0.079 (0.040)	4.146 (3.109)	2	50.000 (0.000)	5.000 (0.000)	0.728 (0.138)
3	L_1	0.771 (0.024)	0.009 (0.003)	2.000 (0.000)	2	5.016 (0.126)	5.000 (0.000)	0.731 (0.005)
	L_2	1.405 (0.023)	0.100 (0.000)	1.002 (0.045)	1	0.000 (0.000)	0.000 (0.000)	0.333 (0.001)
	Huber	0.349 (0.227)	0.008 (0.004)	2.820 (0.385)	3	5.544 (1.224)	5.000 (0.000)	0.857 (0.063)
	RSI	1.090 (0.300)	0.137 (0.059)	4.986 (3.208)	3	50.000 (0.000)	5.000 (0.000)	0.687 (0.035)

Note: The standard errors are reported in parentheses for MAE_μ, MAE_β, $\overline{K}$, $\overline{q}$, $\overline{q}_{TP}$, and $\overline{RI}$.

TABLE 2 Simulation results under the scenarios with high dimensional covariates $p = 100$ and $p = 500$, $q = 5$, and $0.5 \times t(5)$ distributed error over 500 replicates

K	Method	MAE_μ	MAE_β	$\overline{K}$	$\tilde{K}$	$\overline{q}$	$\tilde{q}_{TP}$	$\overline{RI}$
$n = 200$ and $p = 100$								
2	L_1	0.232 (0.066)	0.003 (0.002)	2.000 (0.000)	2	5.000 (0.063)	4.998 (0.045)	0.842 (0.044)
	L_2	0.994 (0.013)	0.050 (0.000)	1.000 (0.000)	1	0.000 (0.000)	0.000 (0.000)	0.501 (0.000)
	Huber	0.371 (0.286)	0.005 (0.005)	1.922 (0.461)	2	5.418 (1.007)	4.798 (0.240)	0.779 (0.135)
	RSI	0.893 (0.231)	0.123 (0.035)	6.398 (2.666)	6	100.000 (0.000)	5.000 (0.000)	0.535 (0.024)
3	L_1	0.817 (0.084)	0.009 (0.007)	2.000 (0.000)	2	4.780 (0.670)	4.742 (0.629)	0.720 (0.026)
	L_2	1.425 (0.036)	0.050 (0.000)	1.004 (0.063)	1	0.000 (0.000)	0.000 (0.000)	0.333 (0.005)
	Huber	0.720 (0.261)	0.011 (0.008)	2.314 (0.494)	2	4.844 (1.011)	4.688 (0.851)	0.742 (0.087)
	RSI	1.325 (0.302)	0.186 (0.050)	6.470 (2.599)	7	100.000 (0.000)	5.000 (0.000)	0.610 (0.043)
$n = 400$ and $p = 100$								
2	L_1	0.161 (0.035)	0.002 (0.001)	2.000 (0.000)	2	5.000 (0.000)	5.000 (0.000)	0.883 (0.024)
	L_2	1.000 (0.001)	0.004 (0.010)	1.000 (0.000)	1	4.790 (1.004)	4.790 (1.004)	0.499 (0.000)
	Huber	0.179 (0.076)	0.002 (0.001)	2.018 (0.133)	2	5.004 (0.063)	5.000 (0.000)	0.873 (0.037)
	RSI	0.853 (0.349)	0.093 (0.041)	6.976 (2.929)	8	100.000 (0.000)	5.000 (0.000)	0.578 (0.079)
3	L_1	0.772 (0.025)	0.004 (0.002)	2.000 (0.000)	2	5.036 (0.186)	5.000 (0.000)	0.731 (0.005)
	L_2	1.406 (0.026)	0.050 (0.000)	1.006 (0.100)	1	0.000 (0.000)	0.000 (0.000)	0.333 (0.004)
	Huber	0.439 (0.273)	0.004 (0.002)	2.672 (0.470)	3	5.350 (0.981)	5.000 (0.000)	0.832 (0.075)
	RSI	1.321 (0.512)	0.145 (0.059)	7.500 (2.651)	8.5	100.000 (0.000)	5.000 (0.000)	0.655 (0.025)
$n = 200$ and $p = 500$								
2	L_1	0.389 (0.304)	0.002 (0.003)	2.000 (0.000)	2	4.360 (1.375)	4.360 (1.375)	0.782 (0.105)
	L_2	0.995 (0.031)	0.010 (0.000)	1.010 (0.100)	1	0.000 (0.000)	0.000 (0.000)	0.501 (0.002)
	Huber	0.585 (0.462)	0.003 (0.003)	2.140 (0.450)	2	4.290 (1.924)	4.050 (1.666)	0.732 (0.131)
3	L_1	0.938 (0.140)	0.004 (0.003)	2.000 (0.000)	2	3.610 (1.456)	3.580 (1.408)	0.677 (0.049)
	L_2	1.425 (0.036)	0.010 (0.000)	1.000 (0.000)	1	0.000 (0.000)	0.000 (0.000)	0.333 (0.004)
	Huber	0.925 (0.236)	0.005 (0.003)	2.300 (0.461)	2	3.410 (1.718)	3.380 (1.674)	0.687 (0.065)
$n = 400$ and $p = 500$								
2	L_1	0.175 (0.037)	0.003 (0.001)	2.000 (0.000)	2	5.000 (0.000)	5.000 (0.000)	0.875 (0.028)
	L_2	0.999 (0.001)	0.002 (0.004)	1.000 (0.000)	1	4.060 (1.953)	4.060 (1.953)	0.499 (0.000)
	Huber	0.242 (0.162)	0.001 (0.001)	2.120 (0.409)	2	5.040 (0.243)	5.000 (0.000)	0.841 (0.071)
3	L_1	0.776 (0.038)	0.001 (0.001)	2.000 (0.000)	2	4.960 (0.315)	4.960 (0.315)	0.729 (0.010)
	L_2	1.404 (0.021)	0.010 (0.000)	1.000 (0.000)	1	0.000 (0.000)	0.000 (0.000)	0.333 (0.000)
	Huber	0.587 (0.274)	0.001 (0.001)	2.450 (0.500)	2	5.350 (0.968)	5.000 (0.000)	0.788 (0.075)

Note: The standard errors are reported in parentheses for MAE_μ, MAE_β, $\overline{K}$, $\overline{q}$, $\overline{q}_{TP}$, and $\overline{RI}$.

4.3 | Subgroup analysis for relatively large-scale data

In this subsection, we further examine the performance of the proposed approach for analyzing large-scale data, which is quite challenging as the algorithm has to deal with the heterogeneity effects for all pairwise subjects. Zhang et al pointed out that the method RSI would have to take a divide-and-conquer strategy when n is large.[13] At the dividing step, the method RSI is performed on each batch of data. At the conquering step, another procedure RSI is performed on the unique intercepts $\hat{\mu}_i$'s gathered from each batch. On the contrary, we conduct parallel computing which can potentially deal with large-scale datasets and more importantly, is well ready for data stored in batches. In the following study, the batch size is set to be 200.

We consider relatively large-scale datasets with $n = 1000$, $p = 5$ or $p = 200$, and $q = 5$. The covarites, covarite coefficients, individual intercepts, and errors are generated in the same manner as described in previous subsections, where the errors with heavy-tails or outliers are considered. The results under the scenarios with the errors generated from the distribution $0.5 \times t(5)$ are summarized in Table 3, and the rest results are given in Appendix C of the Supporting Information. In general, the proposed robust approaches computationally implemented in a parallel manner can accommodate these large-scale data effectively and have favorable subgroup and covariate identification performance, as well as satisfactory estimation. L_1 loss may give smaller estimates of the number of subgroups than that of Huber loss, while it is challenging for the method RSI coupled with divide-and-conquer strategy to identify the true subgroup structure although with a larger enough batch size of the divide-and-conquer (270 in this numerical study). It is not surprising because the batch size is still limited and the communications between batches are insufficient for the divide-and-conquer strategy, especially when the real number of subgroups is not small, which implies the divide-and-conquer strategy should be carefully used in the subgroup analysis.

TABLE 3 Simulation results under the scenarios with large sample size $n = 1000$, $p = 5$ and $p = 200$, $q = 5$, and $0.5 \times t(5)$ distributed error over 500 replicates

K	Method	MAE_μ	MAE_β	$\overline{K}$	$\tilde{K}$	$\overline{q}$	$\overline{q}_{TP}$	$\overline{RI}$
$n = 1000, p = 5$								
2	L_1	0.129 (0.019)	0.019 (0.006)	2.000 (0.000)	2	5.000 (0.000)	5.000 (0.000)	0.901 (0.013)
	L_2	0.375 (0.046)	0.029 (0.009)	6.734 (0.660)	7	5.000 (0.000)	5.000 (0.000)	0.706 (0.031)
	Huber	0.129 (0.018)	0.018 (0.006)	2.000 (0.000)	2	5.000 (0.000)	5.000 (0.000)	0.901 (0.013)
	RSI	0.723 (0.376)	0.304 (0.181)	6.594 (2.761)	7	5.000 (0.000)	5.000 (0.000)	0.677 (0.090)
3	L_1	0.177 (0.036)	0.021 (0.008)	2.990 (0.045)	3	5.000 (0.000)	5.000 (0.000)	0.911 (0.013)
	L_2	0.371 (0.057)	0.039 (0.014)	6.972 (0.368)	7	5.000 (0.000)	5.000 (0.000)	0.821 (0.026)
	Huber	0.171 (0.022)	0.022 (0.008)	3.000 (0.000)	3	5.000 (0.000)	5.000 (0.000)	0.912 (0.010)
	RSI	0.789 (0.338)	0.280 (0.172)	6.400 (2.444)	6	5.000 (0.000)	5.000 (0.000)	0.732 (0.056)
$n = 1000, p = 200$								
2	L_1	0.136 (0.019)	0.001 (0.000)	2.000 (0.000)	2	5.020 (0.141)	5.000 (0.000)	0.896 (0.013)
	L_2	1.000 (0.000)	0.001 (0.000)	1.000 (0.000)	1	5.000 (0.000)	5.000 (0.000)	0.500 (0.000)
	Huber	0.141 (0.020)	0.001 (0.000)	2.000 (0.000)	2	5.000 (0.000)	5.000 (0.000)	0.891 (0.014)
	RSI	1.002 (0.029)	0.075 (0.006)	8.590 (1.164)	8	200.000 (0.000)	5.000 (0.000)	0.502 (0.001)
3	L_1	0.781 (0.016)	0.002 (0.001)	2.000 (0.000)	2	5.890 (1.145)	5.000 (0.000)	0.725 (0.003)
	L_2	1.366 (0.006)	0.001 (0.000)	1.000 (0.000)	1	5.000 (0.000)	5.000 (0.000)	0.333 (0.000)
	Huber	0.213 (0.094)	0.001 (0.000)	2.980 (0.141)	3	5.180 (0.458)	5.000 (0.000)	0.895 (0.027)
	RSI	1.391 (0.046)	0.113 (0.008)	10.810 (1.440)	11	200.000 (0.000)	5.000 (0.000)	0.467 (0.041)

Note: The standard errors are reported in parentheses for MAE_μ, MAE_β, $\overline{K}$, $\overline{q}$, $\overline{q}_{TP}$, and $\overline{RI}$.

5 | DATA ANALYSIS

We analyze the BReast CAncer (BRCA) data from The Cancer Genome Atlas (TCGA), which is a hallmark genomics program for cancer research. For high-dimensional covariates, we consider the mRNA gene expression data. Specifically, the z-score data is analyzed and downloaded from TCGA via the R package *cgdsr*, which has been normalized against diploid samples and quantifies the relative expressions of tumor samples with respect to normal.[21] We consider the expression of gene BRCA1 as the response since BRCA1 is a well-studied tumor suppressor for BRCA,[22] and its mRNA expression level has been demonstrated to be predictive for the breast cancer risk, and patients' response/sensitivity to certain chemotherapy treatment.[23,24] The breast cancer heterogeneity attributed to the differential expressions of BRCA1 has been well studied in the literature.[25] Thus, it is of interest to model the associations of BRCA1 and other genes to further investigate the heterogeneity of breast cancer, such as the various treatment effects among different patient subgroups. In our study, after subject matching, 15 301 gene expression measurements are available for 1074 female patients. As the number of cancer-related genes is expected to be not large, a marginal screening based on the linear regression ($P < 0.01$, with FDR correction) is performed, resulting in 8571 genes for the downstream analysis.

We first take a close look at the distribution of BRCA1 expression in Figure 1A. It shows a skewed and heavy tail pattern which indicates a robust method would be desired. In this analysis, we use Huber loss coupled with $\xi = 1.345$, and for both subgroup identification and variable selection, SCAD penalty function is considered. The proposed approach identifies five subgroups, where the estimates of (group effect, group size) are $(-1.010, 278)$, $(-0.491, 243)$, $(-0.021, 194)$, $(0.801, 242)$, and $(2.384, 117)$, respectively. Under the assumption of the proposed method, it is expected that BRCA1 expressions of the whole subjects can be separated into several levels due to the underlying heterogeneity. Thus, to provide an intuitionistic exhibition of the five subgroups, we plot the histogram of all subjects' BRCA1 expressions in Figure 1B, and for each interval, we visualize the number of subjects in each subgroup with the corresponding range of BRCA1 expression levels via the height of bar, where different subgroups are represented by different colors. The number of subjects in each subgroup is also provided in the legend. As we can see, these five subgroups are relatively balanced with each subgroup having a similar level of size and well separated in terms of BRCA1 expressions, where almost all bars are represented by one color with subjects only from one subgroup.

To get a deeper insight into the identified subgroups, we examine whether these subgroups are associated with clinical outcomes and important genetic characteristics. First, in Figure 2A, we present the Kaplan-Meier (KM) curves of survival time for these five subgroups. Some differences across subgroups are observed with the *P*-value computed from the log-rank test being 0.091. Moreover, we can see that the overall survival outcome of the patients gets worse as the BRCA1 mRNA expression increases and the 5th subgroup, which has the highest BRCA1 expression, experiences much worse overall survival time than the other groups. We further study the tumor size distribution of these subgroups in Figure 2B, where the tumor size is coded into binary variables with tumor size less than 2cm (T0, T1) being coded as "small" and otherwise (T2—T4) being coded as "big". Interestingly, the proportion of big tumor is increasing in these subgroups and the p-value of the χ^2-test for the tumor size distribution is 0.079, indicating certain differences across these subgroups. In addition, we also compare the subgroups identified by the proposed method with Prediction Analysis of Microarray 50 (PAM50) subtypes, which are well-known breast cancer subtypes based on gene expression arrays, referred to as Luminal-A, Luminal-B, Her2-positive, Basel-like, and Normal breast-like.[26] Among these subtypes, many results in the literature have shown that patients with Luminal-A breast cancer have a better prognosis than ones with the other subtypes.[27] We compute the PAM50 results using the R package *genefu*[28] and show the composition of PAM50 subtypes in each subgroup identified by the proposed method in Figure 2C. It is interesting that the proportion of patients with Luminal-A breast cancer declines while the proportion of patients with Luminal-B breast cancer increases from subgroup one to five. The *P*-value of the χ^2-test for the PAM50 distribution among different subgroups is less than 10^{-15}, indicating apparently different composition among these subgroups. These biologically sensible findings provide support to the validity of the proposed subgroup analysis.

In addition, twelve genes are selected by our model and their estimated effects are summarized in Figure 3A. Literature search suggests the biological implications of these identified genes. For example, there have been many works showing that ID4 is a down regulator of BRCA1 in breast cancer.[29,30] The upregulation of ADAM9 has been suggested to contribute to the aggressiveness of triple-negative breast cancer.[31] In addition, Zhang et al has found that overexpression of NKAPL in cancers can induce cell death and cell cycle arrest, and it is a potential diagnostic and prognostic marker in triple negative breast cancer.[32] ZNHIT6/BCD1 has been referred to encoding a proto-oncogene.[33] Published studies have also demonstrated that USP32 is involved in the proliferation of tumor cells in breast cancer.[34] LAPTM4B has been shown to be amplified in breast cancer compared with normal tissues, which contributes to chemotherapy resistance and recurrence

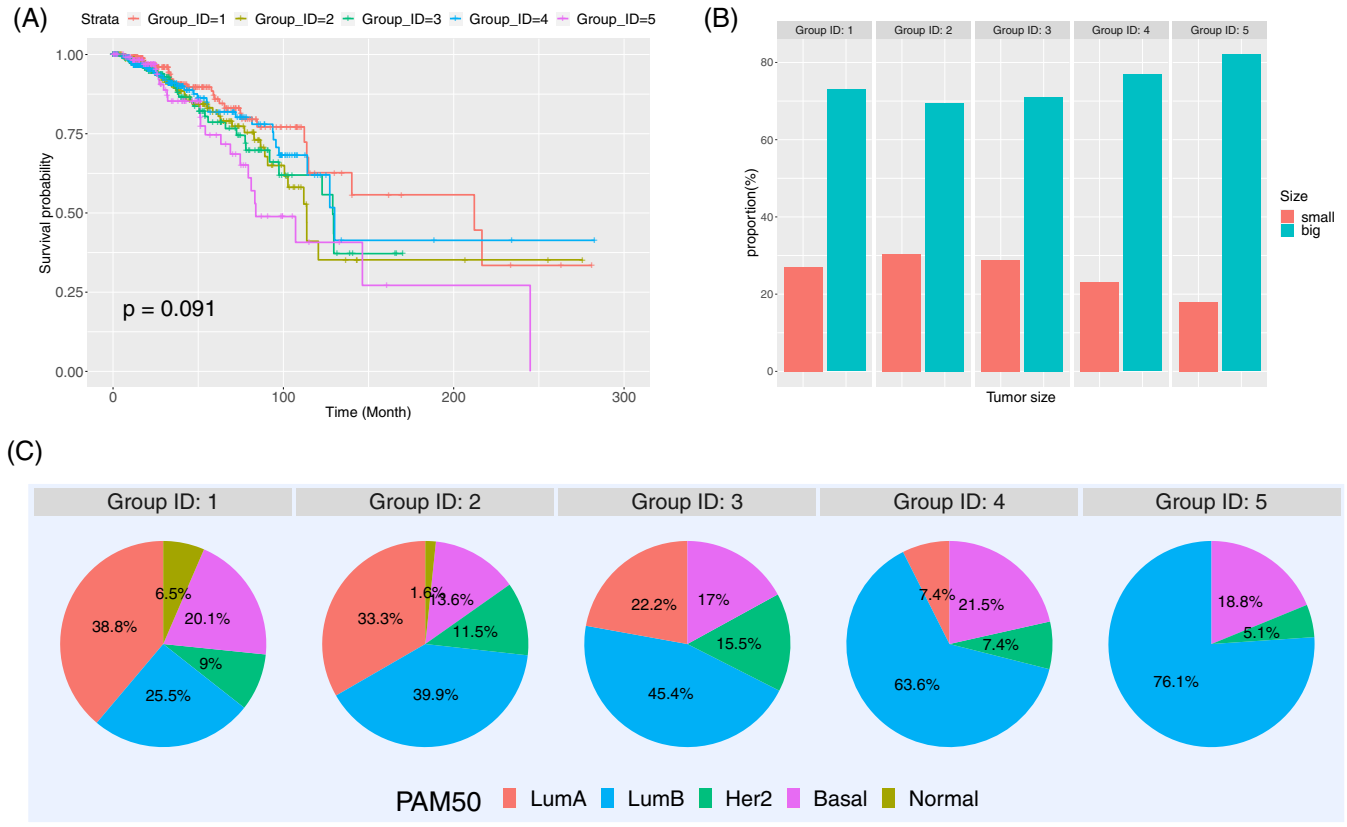

FIGURE 2 Analysis of BRCA data: important clinical outcomes and genetic characteristics for the five subgroups. (A) Kaplan-Meier curves of survival time for the five subgroups. (B) Distributions of the tumor size for the five subgroups. (C) Composition of PAM50 subtypes for the five subgroups

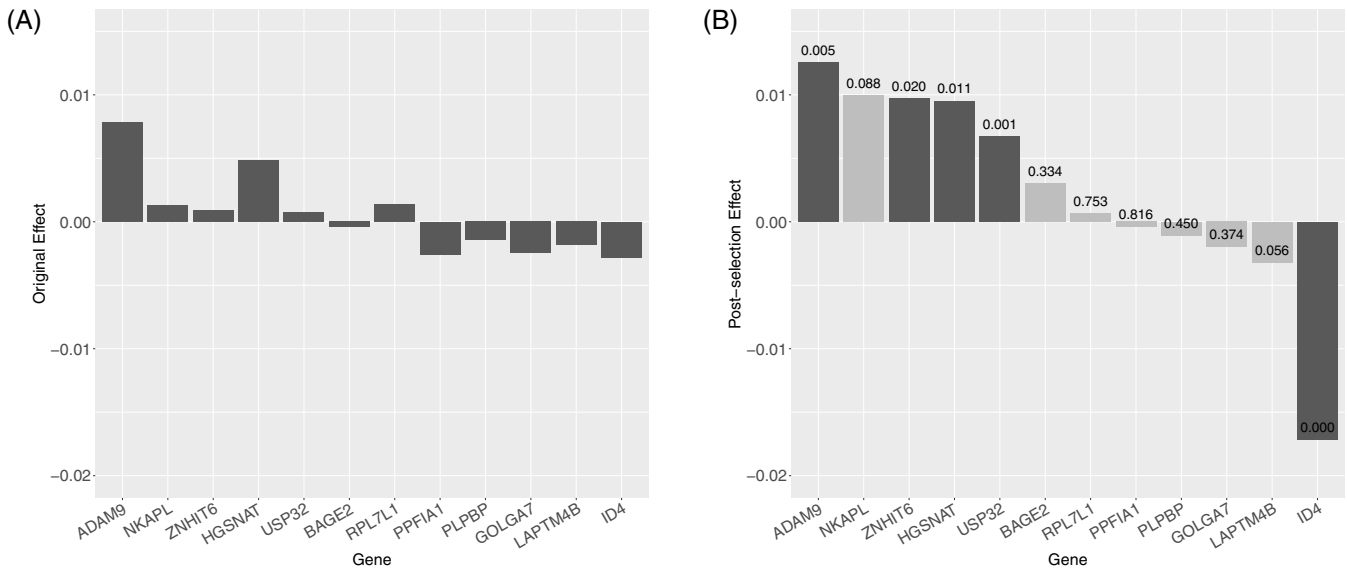

FIGURE 3 Analysis of BRCA data: estimated effects of the genes chosen by the proposed approach. (A) Original estimated effects. (B) Re-estimated effects with the post-selection strategy and *P*-values computed from a wald test

of breast cancer.[35] We further examine the statistical significance of the selected genes using a post-selection strategy. Specifically, we first refit the model (1) with pre-identified five subgroup structures and twelve important genes using Huber loss, in a similar manner as that of Belloni and Chernozhukov,[36] and then perform a Wald test[37] based on the refitted model to examine the statistical significance of the re-estimated coefficients. The *P*-values of twelve genes as well as their re-estimated effects are provided in Figure 3B. It can be seen that five of them are significantly non-zero at significant level 0.05.

We also conduct analysis using the nonrobust approach with L_2 loss and RSI. For the method with L_2 loss, since the response is heavily skewed with a long tail, the estimated subgroups are extremely unbalanced and meaningless, where 1054 patients are classified as one main group while nine and eleven patients are categorized as other two small subgroups. For RSI, we have to limit the covariates to the top 250 genes with the highest marginal correlations with BRCA1 since RSI is not capable of performing variable selection. Due to the large size of this dataset, we apply the divide-and-conquer strategy as suggested by Zhang et al and set the batch size to 270.[13] RSI identifies two subgroups, one with 963 patients and the other with 111 patients. The graphical representation is provided in Figure 1C. The identified subgroup structure is observed to be not much meaningful, which is the consequence of insufficient communications among batches for the divide-and-conquer strategy and absence of variable selection. To provide an indirect evaluation of these models, we further compute the Mean Squared Error and Mean Absolute Error (MAE) with the predicted and true BRCA1 expressions, which are given as 0.205 and 0.253 for the proposed robust method coupled with Huber loss, 0.532 and 0.525 for the nonrobust approach with L_2 loss, and 0.267 and 0.282 for the RSI method, respectively. The proposed method is observed to offer the best fitness to the data.

6 | DISCUSSION

Cancer heterogeneity analysis is a still widely open problem with the complexity of cancer. In this study, we have proposed a robust method for identifying cancer subgroups, which is built on and also advances from the existing heterogeneity analysis by effectively accommodating the heavy-tailed errors and outliers. Furthermore, our method can deal with high-dimensional data which have been less touched in the current literature, and theoretical results have been well established in this article. It also has an advantage of potentially dealing with large-scale data in subgroup analysis. An approximation algorithm is carefully designed by adopting the spirit of the ADMM method, so its estimation procedure can be implemented in a parallel or even distributed manner which is computationally efficient. Our simulation and data analysis have demonstrated its satisfactory performance. In our data analysis, we have focused on the high dimensional genetic factors. It will be of interest to include additional clinical, demographic, and other covariates.

In this study, we have modeled the heterogeneity by unobserved latent factors, following those of Ma and Huang,[10] Chen et al,[11] and other studies. This strategy has been popular and shown great success in recent biomedical heterogeneity studies. For example, Tzala and Best applied the Bayesian latent variable model to uncover the spatial and temporal patterns of latent factors underlying the cancer data.[38] Liu et al revealed the heterogeneity of Alzheimer's disease progression through observed or unobserved latent covariates.[39] The proposed method can be extended to the scenario where the factors contributing to the heterogeneity are available. Some modifications for the estimation method and theoretical properties are needed and deferred to further investigation. When the sample size is small or moderate, we can similarly consider the local linear approximation in subgroup analysis of Zhang et al to improve its practical computing efficiency.[13] However, for large-scale data, the divide-and-conquer strategy should be carefully used because sufficient communications among those data subsets are probably necessary to provide a meaningful clustering result. Furthermore, it is still challenging to well identify true subgroup structures when there are too many real subgroups, which is intrinsically difficult, and we leave this in the future research.

In data analysis, we have used the continuous expression of gene BRCA1 as an indirect marker which may reflect the heterogeneity of breast cancer to some extents, as suggested in the literature. The proposed method can be extended to accommodate categorical clinical markers, which may be more favorable for practical heterogeneity analysis, based on generalized linear model. Unlike in simulation, the true underlying heterogeneity structure is not available in practical data analysis, thus it is difficult to objectively evaluate the accuracy of subgroup analysis. To provide an indirect evaluation, we have studied the associations of the identified subgroups with survival time, tumor size, and PAM50 subtypes. This indirect evaluation strategy has been popular in published heterogeneity analysis studies. For example, Cyll et al[40] investigated prostate cancer heterogeneity with DNA ploidy analysis, Gleason grading, and PTEN expression analysis, respectively, and studied the association of the identified tumor heterogeneity with tumor volume. In addition,

Zhang et al[41] identified three subgroups in pancreatic cancer samples with gene expressions and investigated the associations of theses subgroups with multiple clinical factors, including survival time and tumor diameter. We also note that differences of the KM curves among subgroups are not statistically significant at the 0.05 level of significance. This phenomenon has been not uncommon in published heterogeneity studies, such as Zhang et al[41] for pancreatic cancers and Li et al[42] for glioblastoma multiforme, as subgroup identification procedures have been not tailored to the survival time, and have been demonstrated to be still able to provide partial support to the heterogeneity analysis. In practical data analysis, investigation on biological or clinical implications of subgroups is still challenging. We defer more definitive confirmation from other functional validations to further studies. As high-dimensional inference is a nontrivial task, especially under the proposed heterogeneity analysis framework with penalization, and still an open problem, we have conducted a post-selection strategy to examine the statistical significance of the selected genes. We acknowledge that this strategy is not rigorous, but can provide a rough insight into the significance of the selected genes. More rigorous inference will be studied in the future work.

ACKNOWLEDGEMENTS

The authors thank the editors and reviewers for their careful review and insightful comments. This research was supported by National Natural Science Foundation of China (11971292 and 12071273), Shanghai Rising-Star Program (22QA1403500), Shanghai Research Center for Data Science and Decision Technology, and Fundamental Research Funds for the Central Universities (CXJJ-2019-403).

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data set used for this study is publicly available at the TCGA repository and can be retrieved via the R package *cgdsr*.

ORCID

Mengyun Wu https://orcid.org/0000-0002-0970-4712

REFERENCES

1. Rosa SL, Rubbia-Brandt L, Scoazec JY, Weber A. Editorial: tumor heterogeneity. *Front Med*. 2019;6:156. doi:10.3389/fmed.2019.00156
2. Bhatia S, Frangioni JV, Hoffman RM, Iafrate AJ, Polyak K. The challenges posed by cancer heterogeneity. *Nat Biotechnol*. 2012;30(7):604-610. doi:10.1038/nbt.2294
3. Turajlic S, Sottoriva A, Graham T, Swanton C. Resolving genetic heterogeneity in cancer. *Nat Rev Genet*. 2019;20(7):404-416. doi:10.1038/s41576-019-0114-6
4. Atienza N, García-Heras J, Muñoz-Pichardo J, Villa R. On the consistency of MLE in finite mixture models of exponential families. *J Stat Plan Inference*. 2007;137(2):496-505. doi:10.1016/j.jspi.2005.12.014
5. Shen J, He X. Inference for subgroup analysis with a structured logistic-Normal mixture model. *J Am Stat Assoc*. 2015;110(509):303-312. doi:10.1080/01621459.2014.894763
6. Khalili A, Lin S. Regularization in finite mixture of regression models with diverging number of parameters. *Biometrics*. 2013;69(2):436-446. doi:10.1111/biom.12020
7. Ren M, Zhang S, Zhang Q, Ma S. Gaussian graphical model-based heterogeneity analysis via penalized fusion. *Biometrics*. 2022;78(2):524-535. doi:10.1111/biom.13426
8. Chi EC, Lange K. Splitting methods for convex clustering. *J Comput Graph Stat*. 2015;24(4):994-1013. doi:10.1080/10618600.2014.948181
9. Wu C, Kwon S, Shen X, Pan W. A new algorithm and theory for penalized regression-based clustering. *J Mach Learn Res*. 2016;17(188):1-25.
10. Ma S, Huang J. A concave pairwise fusion approach to subgroup analysis. *J Am Stat Assoc*. 2017;112(517):410-423. doi:10.1080/01621459.2016.1148039
11. Chen J, Tran-Dinh Q, Kosorok MR, Liu Y. Identifying heterogeneous effect using latent supervised clustering with adaptive fusion. *J Comput Graph Stat*. 2021;30(1):43-54. doi:10.1080/10618600.2020.1763808
12. Osborne JW, Overbay A. The power of outliers (and why researchers should always check for them). *Pract Assess Res Eval*. 2004;9:1-12. doi:10.7275/QF69-7K43
13. Zhang Y, Wang HJ, Zhu Z. Robust subgroup identification. *Stat Sinica*. 2019;29:1873-1889. doi:10.5705/ss.202017.0179
14. Peter J, EMR H. *Robust Statistics*. New York: WILEY; 2009.
15. Wu WB. M-estimation of linear models with dependent errors. *Anna Stat*. 2007;35(2):495-521. doi:10.1214/009053606000001406
16. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96(456):1348-1360. doi:10.1198/016214501753382273

17. Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Anna Appl Stat*. 2011;5(1):232-253. doi:10.1214/10-aoas388
18. He B, Zhong T, Huang J, Liu Y, Zhang Q, Ma S. Histopathological imaging-based cancer heterogeneity analysis via penalized fusion with model averaging. *Biometrics*. 2021;77(4):1397-1408. doi:10.1111/biom.13357
19. Boyd S. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found Trends Mach Learn*. 2010;3(1):1-122. doi:10.1561/2200000016
20. Wang L, Wu Y, Li R. Quantile regression for analyzing heterogeneity in ultra-high dimension. *J Am Stat Assoc*. 2012;107(497):214-222. doi:10.1080/01621459.2012.656014
21. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2(5):401-404. doi:10.1158/2159-8290.cd-12-0095
22. Kennedy RD, Quinn JE, Johnston PG, Harkin DP. BRCA1: mechanisms of inactivation and implications for management of patients. *Lancet*. 2002;360(9338):1007-1014. doi:10.1016/s0140-6736(02)11087-7
23. Kennedy RD, Quinn JE, Mullan PB, Johnston PG, Harkin DP. The role of BRCA1 in the cellular response to chemotherapy. *JNCI J National Cancer Ins*. 2004;96(22):1659-1668. doi:10.1093/jnci/djh312
24. Xu Y, Ouyang T, Li J, et al. Predictive value of BRCA1/2 mRNA expression for response to neoadjuvant chemotherapy in BRCA-negative breast cancers. *Cancer Sci*. 2017;109(1):166-173. doi:10.1111/cas.13426
25. Skibinski A, Kuperwasser C. The origin of breast tumor heterogeneity. *Oncogene*. 2015;34(42):5309-5316.
26. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747-752. doi:10.1038/35021093
27. Prat A, Bianchini G, Thomas M, et al. Research-based PAM50 subtype predictor identifies higher responses and improved survival outcomes in HER2-positive breast cancer in the NOAH study. *Clin Cancer Res*. 2014;20(2):511-521. doi:10.1158/1078-0432.ccr-13-0239
28. Gendoo DMA, Ratanasirigulchai N, Schröder MS, et al. Genefu: an R/bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics*. 2015;32(7):1097-1099. doi:10.1093/bioinformatics/btv693
29. Beger C, Pierce LN, Kruger M, et al. Identification of Id4 as a regulator of BRCA1 expression by using a ribozyme-library-based inverse genomics approach. *Proc Natl Acad Sci*. 2001;98(1):130-135. doi:10.1073/pnas.98.1.130
30. Welcsh PL, Lee MK, Gonzalez-Hernandez RM, et al. BRCA1 transcriptionally regulates genes involved in breast tumorigenesis. *Proc Natl Acad Sci*. 2002;99(11):7560-7565. doi:10.1073/pnas.062181799
31. Zhou R, Cho WCS, Ma V, et al. ADAM9 mediates triple-negative breast cancer progression via AKT/NF-κB pathway. *Front Med*. 2020;7:214. doi:10.3389/fmed.2020.00214
32. Zhang X, Kang X, Jin L, et al. ABCC9, NKAPL, and TMEM132C are potential diagnostic and prognostic markers in triple-negative breast cancer. *Cell Biol Int*. 2020;44(10):2002-2010. doi:10.1002/cbin.11406
33. Rouby SE, Rao P, Newcomb EW. Assignment of the human B-cell-derived (BCD1) proto-oncogene to 10p14–p15. *Genomics*. 1997;43(3):395-397. doi:10.1006/geno.1997.4824
34. Akhavantabasi S, Akman HB, Sapmaz A, Keller J, Petty EM, Erson AE. USP32 is an active, membrane-bound ubiquitin protease overexpressed in breast cancers. *Mamm Genome*. 2010;21(7-8):388-397. doi:10.1007/s00335-010-9268-4
35. Li Y, Zou L, Li Q, et al. Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. *Nat Med*. 2010;16(2):214-218. doi:10.1038/nm.2090
36. Belloni A, Chernozhukov V. Least squares after model selection in high-dimensional sparse models. *Ther Ber*. 2013;19(2):521-547.
37. Wilcox RR. Introduction to robust estimation and hypothesis testing. London: Elsevier; 2022.
38. Tzala E, Best N. Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Stat Methods Med Res*. 2008;17(1):97-118.
39. Liu M, Yang J, Liu Y, et al. A fusion learning method to subgroup analysis of Alzheimer's disease. *J Appl Stat*. 2022. doi:10.1080/02664763.2022.2036953
40. Cyll K, Ersvær E, Vlatkovic L, et al. Tumour heterogeneity poses a significant challenge to cancer biomarker research. *Br J Cancer*. 2017;117(3):367-375.
41. Zhang H, Zeng J, Tan Y, et al. Subgroup analysis reveals molecular heterogeneity and provides potential precise treatment for pancreatic cancers. *Onco Targets Ther*. 2018;11:5811-5819.
42. Li Y, Xu S, Ma S, Wu M. Network-based cancer heterogeneity analysis incorporating multi-view of prior information. *Bioinformatics*. 2022;38(10):2855-2862.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Cheng C, Feng X, Li X, Wu M. Robust analysis of cancer heterogeneity for high-dimensional data. *Statistics in Medicine*. 2022;41(27):5448-5462. doi: 10.1002/sim.9578