Check for updates

Received: 6 February 2020 | Revised: 15 August 2020 | Accepted: 17 August 2020

DOI: 10.1111/biom.13357

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Histopathological imaging-based cancer heterogeneity analysis via penalized fusion with model averaging

Baihua He[1] | **Tingyan Zhong**[2,3] | **Jian Huang**[4] | **Yanyan Liu**[1] | **Qingzhao Zhang**[5] | **Shuangge Ma**[3]

[1] School of Mathematics and Statistics, Wuhan University, Wuhan, China

[2] SJTU-Yale Joint Center for Biostatistics, Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

[3] Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut

[4] Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa

[5] Department of Statistics, School of Economics, Key Laboratory of Econometrics, Ministry of Education, The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, China

Correspondence
Qingzhao Zhang, Department of Statistics, School of Economics; Key Laboratory of Econometrics, Ministry of Education, The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, China.
Email: zhangqingzhao@amss.ac.cn; shuangge.ma@yale.edu

Funding information
National Institutes of Health, Grant/Award Numbers: CA241699, CA204120; Yale Cancer Center Pilot Award; MOE (Ministry of Education in China) Project of Humanities and Social Sciences, Grant/Award Number: 19YJC910010; Natural Science Foundation, Grant/Award Number: 1916251; National Natural Science Foundation of China, Grant/Award Number: 11971404

Abstract

Heterogeneity is a hallmark of cancer. For various cancer outcomes/phenotypes, supervised heterogeneity analysis has been conducted, leading to a deeper understanding of disease biology and customized clinical decisions. In the literature, such analysis has been oftentimes based on demographic, clinical, and omics measurements. Recent studies have shown that high-dimensional histopathological imaging features contain valuable information on cancer outcomes. However, comparatively, heterogeneity analysis based on imaging features has been very limited. In this article, we conduct supervised cancer heterogeneity analysis using histopathological imaging features. The penalized fusion technique, which has notable advantages—such as greater flexibility—over the finite mixture modeling and other techniques, is adopted. A sparse penalization is further imposed to accommodate high dimensionality and select relevant imaging features. To improve computational feasibility and generate more reliable estimation, we employ model averaging. Computational and statistical properties of the proposed approach are carefully investigated. Simulation demonstrates its favorable performance. The analysis of The Cancer Genome Atlas (TCGA) data may provide a new way of defining/examining breast cancer heterogeneity.

KEYWORDS
heterogeneity, histopathological imaging, model averaging, penalized fusion

 © 2020 The International Biometric Society

1 | INTRODUCTION

Heterogeneity is a hallmark of cancer. For many cancers, heterogeneity analysis has been extensively conducted and can be roughly classified as unsupervised and supervised. Supervised heterogeneity analysis directly concerns with outcomes/phenotypes and can be clinically more relevant. Such analysis has led to a deeper understanding of disease biology, new ways of classifying/defining diseases, and more informed clinical decision-making (Dagogo-Jack and Shaw, 2018). In "classic" studies, heterogeneity analysis has oftentimes been based on clinical and demographic features. Lately, there have also been many heterogeneity studies built on high-throughput omics data (Lawrence *et al.*, 2013).

Different from many existing studies, here we consider cancer heterogeneity analysis based on histopathological imaging data. Histopathological images are generated in biopsy. They differ from radiological images (which contain information on "macro" properties of tumors) and describe "micro" properties. In particular, they contain essential information on the histological organization and morphological characteristics of tumor cells and their surrounding microenvironment. They have been traditionally used as the gold standard for definitive diagnosis/staging. Recent studies have explored building imaging-based models for cancer prognosis and other outcomes/phenotypes (Wang *et al.*, 2019; Zhong *et al.*, 2019). There have also been a handful of recent studies exploring heterogeneity analysis based on histopathological imaging features (Belhomme *et al.*, 2015; Luo *et al.*, 2017). However, they are oftentimes built on a small number of imaging features (which are not sufficiently informative) and/or simple statistical techniques.

For supervised heterogeneity analysis, the most popular technique is perhaps the finite mixture regression (FMR; McLachlan and Peel, 2000). When the number of input variables is large and/or noises are present, regularization and other techniques have been coupled with the FMR (Khalili and Chen, 2007; Städler *et al.*, 2010). There are also more recent developments. For example, Wager and Athey (2018) develops a nonparametric causal forest for estimating heterogeneous treatment effects. For binary responses, Foster *et al.* (2011) develops a virtual twins method. A recent technique, which has attracted extensive attention and is advantageous in multiple aspects, is penalized fusion (Tibshirani *et al.*, 2005; Ma and Huang, 2017). Specifically, it has a more intuitive definition, more conveniently determines the number of subgroups, and can in principle accommodate subgroups as small as size one. On the negative side, it involves a much larger number of parameters, which leads to challenging computation and unreliable estimation.

In this article, we conduct histopathological imaging-based cancer heterogeneity analysis. The significance of cancer heterogeneity analysis does not need to be reiterated, and the demand for more effective analysis methods has been noted (Dagogo-Jack and Shaw, 2018). Compared to some other types of measurements, histopathological imaging features contain "more direct" information on tumors and are much more cost-effective and simpler to obtain. However, heterogeneity analysis built on such features remains scarce. This study can complement the existing literature by providing a new way of modeling cancer heterogeneity and a new way of utilizing histopathological imaging data. In addition, our data analysis can also provide a new way of looking at breast cancer heterogeneity. The penalized fusion technique (Ma and Huang, 2017; Zhu and Qu, 2018) is adopted for determining heterogeneity. Advancing from the "standard" penalized fusion, sparse penalization is introduced to accommodate high dimensionality and distinguish signals from noises. Our preliminary examination (described later) suggests that a direct application of double penalization—with one for heterogeneity and the other for sparsity—leads to significant computational challenges and unsatisfactory estimation. To overcome this hurdle, we resort to model averaging, divide a big analysis problem into multiple small ones, tackle each separately, and ensemble results to generate the final analysis. Although some components of the proposed approach have been examined to a certain extent, effectively "assembling" them in a novel way to tackle the present challenging analysis is new and demands extensive and challenging numerical and theoretical investigations. Our numerical study suggests favorable performance of the proposed approach. With significant practical, methodological, computational, and theoretical advancements, this study is warranted beyond the existing literature.

2 | METHODS

2.1 | Penalized fusion with model averaging

Denote y as the response variable and $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ as the p-dimensional imaging features. Consider a continuous response. Discussions on other types of response are provided later. Let $\{\mathbf{x}_i, y_i\}_{i=1}^n$ be n independent copies of $\{\mathbf{x}, y\}$, and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. Under the penalized fusion framework, we consider the models

$$y_i = \mathbf{x}_i^{\mathrm{T}}\theta_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\theta_i = (\theta_{i1}, \ldots, \theta_{ip})^{\mathrm{T}}$ is the vector of unknown regression coefficients, and ϵ_i is the random error with $\mathrm{E}(\epsilon_i) = 0$

and $\mathrm{Var}(\epsilon_i) = \sigma^2$. Here each subject has its own regression model/coefficients, which renders the penalized fusion technique greater flexibility than the FMR and some other techniques. For example, the number of subgroups does not need to be assumed *a priori*. In addition, penalized fusion can potentially accommodate small subgroups (in principle, as small as size one). The downside is that the number of parameters, $n \times p$, is much higher than in an ordinary regression.

In regression-based heterogeneity analysis, two subjects belong to the same subgroup if and only if they have the same regression model/coefficients. As such, heterogeneity analysis amounts to determining which θ_i's are equal to each other. With low-dimensional covariates, the "standard" penalized fusion has objective function:

$$\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\theta_i\right)^2 + \sum_{1\le i<m\le n} p(\|\theta_i - \theta_m\|, \lambda), \qquad (2)$$

where $p(\cdot, \lambda)$ is a penalty function with tuning parameter λ, and $\|\cdot\|$ is the L_2 norm. Note that the penalty component involves $n(n-1)/2$ terms. With proper penalization, there is a nonzero probability $\widehat{\theta}_i = \widehat{\theta}_j$, and so subgrouping can be realized.

When p is large and there are noises in covariates, additional regularization needs to be imposed to (2). Although seemingly straightforward, a direct application may lead to challenging computation and unsatisfactory estimation. For example, this may involve manipulating matrices with size $\frac{n(n-1)p}{2} \times np$ (details in Section 2.2.1). Each term in the fusion penalty involves p-dimensional vectors, and a small change of tuning may cause a big change of the objective function, leading to instability. We adopt model averaging to tackle computational challenges. Overall, the proposed approach involves the following steps:

Step 1: Partition $\{1, \ldots, p\}$ into B_n nonoverlapping sets with equal sizes. More information on the partition is provided below in the theoretical development. Denote A_b as the bth index set and $|A_b|$ as its size. For the simplicity of notation, assume $p = B_n \times |A_b|$. For a p-dimensional vector $\mathbf{a} = (a_1, \ldots, a_p)^{\mathrm{T}}$, denote $\mathbf{a}_{(b)}$ as its subvector indexed by A_b. According to A_b's, partition $\{\mathbf{x}_i, y_i\}_{i=1}^n$ into B_n subsets $\{(\mathbf{x}_{i(1)}, y_i)_{i=1}^n\}, \ldots, \{(\mathbf{x}_{i(B_n)}, y_i)_{i=1}^n\}$.

Step 2: For data subset $b(= 1, \ldots, B_n)$, conduct the double penalized fusion analysis. Denote the estimate as $\{\widehat{\theta}_{1(b)}, \ldots, \widehat{\theta}_{n(b)}\}$.

Step 3: With $\{(\mathbf{x}_{i(1)}^{\mathrm{T}}\widehat{\theta}_{i(1)}, y_i)_{i=1}^n, \ldots, (\mathbf{x}_{i(B_n)}^{\mathrm{T}}\widehat{\theta}_{i(B_n)}, y_i)_{i=1}^n\}$, minimize the prediction error, and obtain the optimal weight $\widehat{\omega} = (\widehat{\omega}_1, \ldots, \widehat{\omega}_{B_n})^{\mathrm{T}}$.

Step 4: For $i = 1, \ldots, n$, compute the model-averaged estimate $\widehat{\theta}_{i\widehat{\omega}} = \sum_{b=1}^{B_n} \widehat{\omega}_b \pi_b^{\mathrm{T}} \widehat{\theta}_{i(b)}$, where π_b is the matrix $(\mathbf{I}_{|A_b|}, \mathbf{0}_{|A_b|\times(p-|A_b|)})$ column permutation correspond to A_b. Based on $\{\widehat{\theta}_{i\widehat{\omega}}\}_{i=1}^n$, conduct subgrouping, and identify the heterogeneity structure.

In some model averaging studies, random sampling is adopted. In Step 1, we use partition (Ando and Li, 2017), which has a lower computational cost. Our numerical exploration described below suggests that the ordering of variables in the partition is not critical, as long as certain conditions are satisfied. Step 2 can be conducted on multiple CPUs in a highly parallel manner to reduce computer time. More details of Steps 2-4 are as follows.

2.1.1 | Details of Step 2

Consider the bth data subset, which contains all n samples and $|A_b|$ covariates. Consider the submodel:

$$y_i = \mathbf{x}_{i(b)}^{\mathrm{T}}\theta_{i(b)} + \epsilon_{i(b)}, \qquad (3)$$

where $\theta_{i(b)} = (\theta_{i(b)}^1, \ldots, \theta_{i(b)}^{|A_b|})^{\mathrm{T}}$ is the $|A_b|$-vector of unknown coefficients. As $p \to \infty$, both $|A_b|$ and B_n can go to infinity (details provided in the theoretical development). When applying penalized fusion to submodel (3), we note that $|A_b|$ may still be moderate to large compared to n, and there may be noises in the $|A_b|$ covariates. As such, we propose additionally applying a sparsity penalty. Specifically, consider the objective function:

$$\begin{aligned} Q_n(\{\theta_{i(b)}\}, \lambda_1, \lambda_2) = {} & \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_{i(b)}^{\mathrm{T}}\theta_{i(b)}\right)^2 \\ & + \sum_{i=1}^{n}\sum_{j=1}^{|A_b|} p_1(|\theta_{i(b)}^j|, \lambda_1) \\ & + \sum_{1\le i<m\le n} p_2(\|\theta_{i(b)} - \theta_{m(b)}\|, \lambda_2), \end{aligned} \qquad (4)$$

where $p_1(\cdot,\cdot)$ and $p_2(\cdot,\cdot)$ are two penalties, and λ_1 and λ_2 are (vectors of) tunings. In our implementation, we take p_1 and p_2 as SCAD (which also involves a regularization parameter γ) and note that some alternatives may be equally applicable. Loosely speaking, the two penalties in (4) share similar spirit as those in fused penalization, with the first for sparsity and the second for equality. Key differences are: the proposed approach involves a much larger number of parameters, all pair-wise (as opposed to the adjacent) differences are taken in the second penalty,

and different θ's correspond to different subjects. The B_n sets of estimates, as opposed to the individual subgrouping results, will be used in downstream analysis.

2.1.2 | Details of Step 3

Step 2 generates $\{\widehat{\theta}_{1(b)}, \ldots, \widehat{\theta}_{n(b)}\}_{b=1}^{B_n}$. Denote $\widehat{y}_{i(b)} = \mathbf{x}_{i(b)}^{\mathrm{T}} \widehat{\theta}_{i(b)}$ and $\widehat{\mathbf{y}}_i = (\widehat{y}_{i(1)}, \ldots, \widehat{y}_{i(B_n)})^{\mathrm{T}}$. Consider the weight vector $\omega = (\omega_1, \ldots, \omega_{B_n})^{\mathrm{T}}$ with $0 \le \omega_b \le 1$ and $\sum_b \omega_b = 1$. Here ω_b is the weight of the bth submodel. For a given ω, let $\widehat{y}_i(\omega) = \omega^{\mathrm{T}} \widehat{\mathbf{y}}_i$. For choosing ω, consider the loss function:

$$\mathcal{L}_n(\omega) = \sum_{i=1}^{n} \{y_i - \widehat{y}_i(\omega)\}^2.$$

When B_n is large, this function may not have a unique solution. In addition, directly optimizing it leads to a dense estimate. In Section 2.2.2, we adopt a greedy optimization algorithm that leads to a unique and sparse estimate. Denote $\widehat{\omega}$ as the estimated weight vector and $\widehat{\theta}_{i\widehat{\omega}} = \sum_{b=1}^{B_n} \widehat{\omega}_b \pi_b^{\mathrm{T}} \widehat{\theta}_{i(b)}$ as the corresponding estimate.

2.1.3 | Details of Step 4

With the estimates, the heterogeneity structure can be determined following the standard penalized fusion strategy and examining the equality of estimates (Ma and Huang, 2017). In practice, as the computation is terminated when the adjacent estimates are close enough (details below), thresholding may be needed to conclude equality when two estimates are close enough.

2.1.4 | Remarks

With other models and other types of response, the first term in the objective function can be replaced by a general lack-of-fit measure, and the proposed approach can then be applied. When the lack-of-fit measure is continuously differentiable, the computational algorithm described below can be applied by invoking Taylor expansion. Theoretical investigation may need further data/model-specific adjustments.

2.2 | Computation

2.2.1 | Computation of Step 2

We reparameterize by introducing $\eta_{im(b)} = \theta_{i(b)} - \theta_{m(b)}$ and $\mu_{i(b)}^j = \theta_{i(b)}^j$. Then minimizing (4) is equivalent to the constrained optimization problem:

$$\begin{aligned} S_n(\{\theta_{i(b)}\}, \{\mu_{i(b)}\}, \{\eta_{im(b)}\}) = & \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_{i(b)}^{\mathrm{T}} \theta_{i(b)}\right)^2 \\ & + \sum_{i=1}^{n} \sum_{j=1}^{|A_b|} p_1(|\mu_{i(b)}^j|, \lambda_1) \\ & + \sum_{i<m} p_2(\|\eta_{im(b)}\|, \lambda_2), \quad (5) \end{aligned}$$

$$\text{subject to} \quad \mu_{i(b)} = \theta_{i(b)}, \eta_{im(b)} = \theta_{i(b)} - \theta_{m(b)},$$

where $\mu_{i(b)} = (\mu_{i(b)}^1, \ldots, \mu_{i(b)}^{|A_b|})^{\mathrm{T}}$. The augmented Lagrangian function is

$$\begin{aligned} & T_n(\{\theta_{i(b)}\}, \mu_{(b)}, \{\eta_{im(b)}\}, \{\mathbf{v}_{i(b)}\}, \bar{\mathbf{v}}_{(b)}) \\ & = S_n(\{\theta_{i(b)}\}, \mu_{(b)}, \{\eta_{im(b)}\}) + \sum_{i<m} \mathbf{v}_{im(b)}^{\mathrm{T}} \{\theta_{i(b)} - \theta_{m(b)} - \eta_{im(b)}\} \\ & + \bar{\mathbf{v}}_{(b)}^{\mathrm{T}} \{\theta_{(b)} - \mu_{(b)}\} + \frac{\kappa}{2} \sum_{i<m} \|\theta_{i(b)} - \theta_{m(b)} - \eta_{im(b)}\|^2 \\ & + \frac{\kappa}{2} \|\theta_{(b)} - \mu_{(b)}\|^2, \quad (6) \end{aligned}$$

where $\mu_{(b)} = (\mu_{1(b)}^{\mathrm{T}}, \ldots, \mu_{n(b)}^{\mathrm{T}})^{\mathrm{T}}$, $\{\mathbf{v}_{im(b)}\}$ and $\bar{\mathbf{v}}_{(b)} = \{(\bar{v}_{i(b)}^1, \ldots, \bar{v}_{i(b)}^{|A_b|}), i = 1, \ldots, n\}^{\mathrm{T}}$ are the Lagrange multipliers, and κ is the penalty parameter. Let $\mathbf{X}_{(b)} = \mathrm{diag}(\mathbf{x}_{1(b)}^{\mathrm{T}}, \ldots, \mathbf{x}_{n(b)}^{\mathrm{T}})$, $\eta_{(b)} = (\eta_{im(b)}^{\mathrm{T}}, i < m)^{\mathrm{T}}$, $\mathbf{v}_{(b)} = (\mathbf{v}_{im(b)}^{\mathrm{T}}, i < m)^{\mathrm{T}}$, $\Delta = \{(\mathbf{e}_i - \mathbf{e}_j), i < m\}^{\mathrm{T}}$, $\mathbf{A} = \Delta \otimes \mathbf{I}_{|A_b|}$, $\mathbf{e}_i$ be the ith canonical basis of $\mathbb{R}^{|A_b|}$, $\mathbf{I}_{|A_b|}$ be the $|A_b| \times |A_b|$ identity matrix, and $\otimes$ denote the Kronecker product. Then (6) can be rewritten as

$$\begin{aligned} & L_n(\theta_{(b)}, \mu_{(b)}, \eta_{(b)}, \mathbf{v}_{(b)}, \bar{\mathbf{v}}_{(b)}) = \frac{1}{2} \left\|\mathbf{y} - \mathbf{X}_{(b)} \theta_{(b)}\right\|^2 \\ & + \sum_{i=1}^{n} \sum_{j=1}^{|A_b|} p_1(|\mu_{i(b)}^j|, \lambda_1) + \sum_{i<m} p_2(\|\eta_{im(b)}\|, \lambda_2) \\ & + \frac{\kappa}{2} \left[\left\| \mathbf{A}\theta_{(b)} - \eta_{(b)} + \frac{\mathbf{v}_{(b)}}{\kappa} \right\|^2 + \left\| \theta_{(b)} - \mu_{(b)} + \frac{\bar{\mathbf{v}}_{(b)}}{\kappa} \right\|^2 \right] + C. \quad (7) \end{aligned}$$

We adopt the alternating direction method of multipliers technique and provide additional details in the Supporting Information. Note that here $\mathbf{A}$ is a $\frac{n(n-1)|A_b|}{2} \times n|A_b|$ matrix. As such, without model averaging, the dimension of $\mathbf{A}$ would be $\frac{n(n-1)p}{2} \times np$, which can lead to challenging computation.

Given the estimate $(\{\theta_{i(b)}^{(\ell-1)}\}, \boldsymbol{\mu}_{(b)}^{(\ell-1)}, \boldsymbol{\eta}_{(b)}^{(\ell-1)}, \{\mathbf{v}_{im(b)}^{(\ell-1)}\}, \bar{\mathbf{v}}_{(b)}^{(\ell-1)})$ at the $\ell-1$th iteration, the ℓth iteration estimates are

$$\boldsymbol{\eta}_{im(b)}^{(\ell)} = \begin{cases} S(\boldsymbol{\delta}_{im(b)}^{(\ell)}, \lambda_2/\kappa) & \text{if } \|\boldsymbol{\delta}_{im(b)}^{(\ell)}\| \le \lambda_2 + \lambda_2/\kappa \\ \dfrac{S\left(\boldsymbol{\delta}_{im(b)}^{(\ell)}, \gamma\lambda_2/((\gamma-1)\kappa)\right)}{1-1/((\gamma-1)\kappa)} & \text{if } \lambda_2+\lambda_2/\kappa < \|\boldsymbol{\delta}_{im(b)}^{(\ell)}\| \le \gamma\lambda_2 \\ \boldsymbol{\delta}_{im(b)}^{(\ell)} & \text{if } \|\boldsymbol{\delta}_{im(b)}^{(\ell)}\| > \gamma\lambda_2, \end{cases} \tag{8}$$

where $\boldsymbol{\delta}_{im(b)}^{(\ell)} = \theta_{i(b)}^{(\ell-1)} - \theta_{m(b)}^{(\ell-1)} + \frac{1}{\kappa}\mathbf{v}_{im(b)}^{(\ell-1)}$ and $S(t,\lambda) = (1-\lambda/\|t\|)_+ t$,

$$\begin{aligned} \theta_{(b)}^{(\ell)} = &\left(\mathbf{X}_{(b)}^{\mathrm{T}}\mathbf{X}_{(b)} + \kappa\mathbf{A}^{\mathrm{T}}\mathbf{A} + \kappa\mathbf{I}_{n|A_b|}\right)^{-1} \\ &\times\left\{\mathbf{X}_{(b)}^{\mathrm{T}}\mathbf{y} + \kappa\mathbf{A}^{\mathrm{T}}\left(\boldsymbol{\eta}_{(b)}^{(\ell)} - \mathbf{v}_{(b)}^{(\ell-1)}/\kappa\right)\right. \\ &\left.+\kappa\left(\boldsymbol{\mu}_{(b)}^{(\ell-1)} - \bar{\mathbf{v}}_{(b)}^{(\ell-1)}/\kappa\right)\right\}, \end{aligned} \tag{9}$$

$$\mu_{i(b)}^{j(\ell)} = \begin{cases} ST(\xi_{i(b)}^{j(\ell)}, \lambda_1/\kappa) & |\xi_{i(b)}^{j(\ell)}| \le \lambda_1 + \lambda_1/\kappa \\ \dfrac{ST(\xi_{i(b)}^{j(\ell)}, \gamma\lambda_1/((\gamma-1)\kappa))}{1-1/((\gamma-1)\kappa)} & \lambda_1+\lambda_1/\kappa < |\xi_{i(b)}^{j(\ell)}| \le \gamma\lambda_1, \\ \xi_{i(b)}^{j(\ell)} & |\xi_{i(b)}^{j(\ell)}| > \gamma\lambda_1 \end{cases} \tag{10}$$

$$\begin{aligned} \mathbf{v}_{(b)}^{(\ell)} &= \mathbf{v}_{(b)}^{(\ell-1)} + \kappa(\mathbf{A}\theta_{(b)}^{(\ell)} - \boldsymbol{\eta}_{(b)}^{(\ell)}), \\ \bar{\mathbf{v}}_{(b)}^{(\ell)} &= \bar{\mathbf{v}}_{(b)}^{(\ell-1)} + \kappa(\theta_{(b)}^{(\ell)} - \boldsymbol{\mu}_{(b)}^{(\ell)}), \end{aligned} \tag{11}$$

where $\xi_{i(b)}^{j(\ell)} = \theta_{i(b)}^{j(\ell)} + \bar{v}_{i(b)}^{j(\ell-1)}/\kappa$ and $ST(t,\lambda) = \mathrm{sign}(t)(|t|-\lambda)_+$.

Overall, we propose the following algorithm: (a) Initialization: $\ell = 0$, $\boldsymbol{\mu}_{(b)}^{(0)} = \theta_{(b)}^{(0)}, \boldsymbol{\eta}_{(b)}^{(0)} = \mathbf{A}\theta_{(b)}^{(0)}, \mathbf{v}_{(b)}^{(0)} = 0, \bar{\mathbf{v}}_{(b)}^{(0)} = 0$. In numerical study, we use estimates from an FMR (with the number of subgroups determined by Bayesian information criterion (BIC)) as the initial $\theta_{(b)}^{(0)}$. Our exploration suggests that the estimates are not too sensitive to the initial values as long as they are not "too off." (b) Update $\ell = \ell+1$, $\boldsymbol{\eta}_{(b)}^{(\ell)}$ via (8), $\theta_{(b)}^{(\ell)}$ via (9), $\boldsymbol{\mu}_{(b)}^{(\ell)}$ via (10), $\mathbf{v}_{(b)}^{(\ell)}$ and $\bar{\mathbf{v}}_{(b)}^{(\ell)}$ via (11). We set $\kappa = 1$ and $\gamma = 3$. (c) Repeat Step (b) until convergence, which is concluded if $\|\tilde{\mathbf{r}}_{(b)}^{(\ell^*)}\| \le \varepsilon$ with $\varepsilon = 0.001$. Here $\tilde{\mathbf{r}}_{(b)}^{(\ell)} = ((\mathbf{r}_{(b)}^{(\ell)})^{\mathrm{T}}, (\bar{\mathbf{r}}_{(b)}^{(\ell)})^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{r}_{(b)}^{(\ell)} = \mathbf{A}\theta_{(b)}^{(\ell)} - \boldsymbol{\eta}_{(b)}^{(\ell)}$, and $\bar{\mathbf{r}}_{(b)}^{(\ell)} = \theta_{(b)}^{(\ell)} - \boldsymbol{\mu}_{(b)}^{(\ell)}$. The following convergence result is proved in the Supporting Information.

Corollary 1. *Let $\{\theta_{(b)}^{(\ell)}, \boldsymbol{\mu}_{(b)}^{(\ell)}, \boldsymbol{\eta}_{(b)}^{(\ell)}, \mathbf{v}_{(b)}^{(\ell)}, \bar{\mathbf{v}}_{(b)}^{(\ell)}\}_{\ell=1}^{\infty}$ be the sequence of estimates. If $\{\boldsymbol{\mu}_{(b)}^{(\ell)}, \boldsymbol{\eta}_{(b)}^{(\ell)}\}_{\ell=1}^{\infty}$ are bounded and $\|\mathbf{v}_{(b)}^{(\ell)} - \mathbf{v}_{(b)}^{(\ell-1)}\| + \|\bar{\mathbf{v}}_{(b)}^{(\ell)} - \bar{\mathbf{v}}_{(b)}^{(\ell-1)}\| \to 0$, then $\{\theta_{(b)}^{(\ell)}, \boldsymbol{\mu}_{(b)}^{(\ell)}, \boldsymbol{\eta}_{(b)}^{(\ell)}, \mathbf{v}_{(b)}^{(\ell)}, \bar{\mathbf{v}}_{(b)}^{(\ell)}\}_{\ell=1}^{\infty}$ is bounded. Furthermore, there exists a sequence $\{\theta_{(b)}^{(\ell_j)}, \boldsymbol{\mu}_{(b)}^{(\ell_j)}, \boldsymbol{\eta}_{(b)}^{(\ell_j)}, \mathbf{v}_{(b)}^{(\ell_j)}, \bar{\mathbf{v}}_{(b)}^{(\ell_j)}\}_{\ell_j=1}^{\infty}$ such that*

$$\begin{aligned} &\|\theta_{(b)}^{(\ell_j)} - \theta_{(b)}^{(\ell_j-1)}\| + \|\boldsymbol{\mu}_{(b)}^{(\ell_j)} - \boldsymbol{\mu}_{(b)}^{(\ell_j-1)}\| + \|\boldsymbol{\eta}_{(b)}^{(\ell_j)} - \boldsymbol{\eta}_{(b)}^{(\ell_j-1)}\| \\ &+ \|\mathbf{v}_{(b)}^{(\ell_j)} - \mathbf{v}_{(b)}^{(\ell_j-1)}\| + \|\bar{\mathbf{v}}_{(b)}^{(\ell_j)} - \bar{\mathbf{v}}_{(b)}^{(\ell_j-1)}\| \to 0 \end{aligned}$$

as $\ell_j \to \infty$. Thus $\{\theta_{(b)}^{(\ell)}, \boldsymbol{\mu}_{(b)}^{(\ell)}, \boldsymbol{\eta}_{(b)}^{(\ell)}, \mathbf{v}_{(b)}^{(\ell)}, \bar{\mathbf{v}}_{(b)}^{(\ell)}\}_{\ell=1}^{\infty}$ has a sequence that converges to the stationary point $\{\theta_{(b)}^{\#}, \boldsymbol{\mu}_{(b)}^{\#}, \boldsymbol{\eta}_{(b)}^{\#}, \mathbf{v}_{(b)}^{\#}, \bar{\mathbf{v}}_{(b)}^{\#}\}$ that satisfies the first-order conditions:

$$\begin{cases} \mathbf{X}_{(b)}^{\mathrm{T}}(-\mathbf{y} + \mathbf{X}_{(b)}\theta_{(b)}^{\#}) + \mathbf{A}^{\mathrm{T}}\mathbf{v}_{(b)}^{\#} + \bar{\mathbf{v}}_{(b)}^{\#} = \mathbf{0} \\ 0 \in -\bar{v}_{i(b)}^{j\#} + \dfrac{\partial p_1(|\mu_{i(b)}^{j}|, \lambda_1)}{\partial \mu_{i(b)}^{j}}\Big|_{\mu_{i(b)}^{j} = \mu_{i(b)}^{j\#}}, j = 1, \dots, |A_b|, \\ \qquad\qquad i = 1, \dots, n \\ \mathbf{0} \in -\mathbf{v}_{im(b)}^{\#} + \dfrac{\partial p_2(\|\boldsymbol{\eta}_{im(b)}\|, \lambda_2)}{\partial \boldsymbol{\eta}_{im(b)}}\Big|_{\boldsymbol{\eta}_{im(b)} = \boldsymbol{\eta}_{im(b)}^{\#}}, i < m \\ \mathbf{A}\theta_{(b)}^{\#} - \boldsymbol{\eta}_{(b)}^{\#} = \mathbf{0} \\ \theta_{(b)}^{\#} - \boldsymbol{\mu}_{(b)}^{\#} = \mathbf{0}. \end{cases} \tag{12}$$

2.2.2 | Computation of Step 3

We adopt a greedy algorithm to generate a unique and sparse estimate: (a) Initialize $\ell = 0$ and $\boldsymbol{\omega}^{(0)} = \mathbf{0}$; (b) update $\ell = \ell+1$, $\lambda^{(\ell)} = \frac{2}{\ell+1}$, $\boldsymbol{\gamma}^{(\ell)} \in \arg\min_{\gamma\in\Omega_n}\{\boldsymbol{\gamma}^{\mathrm{T}}\nabla\mathcal{L}_n(\boldsymbol{\omega}^{(\ell-1)})\}$, and $\boldsymbol{\omega}^{(\ell)} = \boldsymbol{\omega}^{(\ell-1)} + \lambda^{(\ell)}(\boldsymbol{\gamma}^{(\ell)} - \boldsymbol{\omega}^{(\ell-1)})$; and (c) repeat step (b) until $(\boldsymbol{\omega}^{(\ell-1)} - \boldsymbol{\gamma}^{(\ell-1)})^{\mathrm{T}}\nabla\mathcal{L}_n(\boldsymbol{\omega}^{(\ell)}) \le \varepsilon$, where $\varepsilon = 0.001$ in our numerical study. Properties such as convergence can be established following Dai *et al.* (2012).

2.2.3 | Remarks

With the convergence properties of steps 2 and 3, the overall convergence property can be established. In all of our

numerical studies, convergence is achieved within a moderate number of iterations. Following Wang *et al.* (2009) and Ma and Huang (2017), we choose the tuning parameters by minimizing a modified BIC:

$$\text{BIC}(\lambda_1,\lambda_2) = \log\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{i\widehat{\omega}}(\lambda_1,\lambda_2))^2\right\} + C_n\frac{\log n}{n}S,$$

where $C_n = \log(np)$ and S is the number of nonzero coefficients in $\widehat{\boldsymbol{\alpha}}_{\widehat{\omega}} = (\widehat{\boldsymbol{\alpha}}_{1\widehat{\omega}}^{\mathrm{T}},\ldots,\widehat{\boldsymbol{\alpha}}_{n\widehat{\omega}}^{\mathrm{T}})^{\mathrm{T}}$. We acknowledge the importance of tuning parameter selection (eg, optimality). As BIC has been extensively adopted in the literature, we choose not to discuss further. When B_n is too large, computational difficulty may arise. When B_n is too small, conditions specified in the following subsection may be violated. On the other hand, our numerical study below suggests that when B_n is in a reasonable range, its value is not critical.

2.3 | Statistical properties

Assume K subgroups, and denote $\mathcal{G} = (\mathcal{G}_1,\ldots,\mathcal{G}_K)$ as the subgroup set. Denote $\boldsymbol{\alpha}_k$ as the shared regression coefficient vector for all subjects in $\mathcal{G}_k$. Let $\widetilde{\mathbf{G}} = \{g_{ik}\}$ be the $n\times K$ matrix with $g_{ik} = 1$ for $i\in\mathcal{G}_k$ and $g_{ik} = 0$ otherwise, and $\mathbf{G}_b = \widetilde{\mathbf{G}}\otimes\mathbf{I}_{|A_b|}$. Let $C_{\mathcal{G}}^b = \{\boldsymbol{\theta}_{(b)}\in\mathbb{R}^{n|A_b|}, \boldsymbol{\theta}_{i(b)} = \boldsymbol{\theta}_{m(b)}, \text{for any } i,m\in\mathcal{G}_k, 1\le k\le K\}$. For each $\boldsymbol{\theta}\in C_{\mathcal{G}}^b$, it can be written as $\boldsymbol{\theta}_{(b)} = \mathbf{G}_b\boldsymbol{\alpha}_{(b)}$, where $\boldsymbol{\alpha}_{(b)} = (\boldsymbol{\alpha}_{1(b)}^{\mathrm{T}},\ldots,\boldsymbol{\alpha}_{K(b)}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\alpha}_{k(b)} = (\alpha_{k(b)}^1,\ldots,\alpha_{k(b)}^{|A_b|})^{\mathrm{T}}$ is a $|A_b|\times 1$ vector of the kth subgroup-specific parameter for $k = 1,\ldots,K$. Denote $|\mathcal{G}_{\min}| = \min_k|\mathcal{G}_k|$ and $|\mathcal{G}_{\max}| = \max_k|\mathcal{G}_k|$. Further denote the scaled penalty functions as $p_1(t) = \lambda_1^{-1}p_1(t,\lambda_1)$ and $p_2(t) = \lambda_2^{-1}p_2(t,\lambda_2)$. When a model includes all and only covariates with nonzero coefficients, we call it the true model. When a model omits at least one covariate with a nonzero coefficient, we call it an underfitted model, and denote the set of underfitted models as $\mathcal{M}$. When a model is not underfitted, we call it fitted. With respect to the partition, we require that at least one submodel is fitted. This requirement has been extensively imposed in the model averaging literature (Zhang *et al.*, 2020).

For the kth subgroup, consider

$$\boldsymbol{\alpha}_{k(b)}^* = \arg\min E\{(y - \mathbf{x}_{(b)}^{\mathrm{T}}\boldsymbol{\alpha}_{k(b)})^2\}.$$

When subject i belongs to the kth subgroup, $\boldsymbol{\theta}_{i(b)}^* = \boldsymbol{\alpha}_{k(b)}^*$. So the underlying submodel for the kth subgroup is

$$y_i = \mathbf{x}_{i(b)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b)}^* + \epsilon_{i(b)},\quad i\in\mathcal{G}_k.$$

When the submodel corresponding to data subset b does not belong to $\mathcal{M}$, $\boldsymbol{\alpha}_{k(b)}^*$ equals $\boldsymbol{\alpha}_{k(b)}^0$, where $\boldsymbol{\alpha}_{k(b)}^0$ contains the corresponding elements of the true coefficient $\boldsymbol{\alpha}_k^0$. Let $b_n = \min_b\min_{k\neq k'}\|\boldsymbol{\alpha}_{k(b)}^* - \boldsymbol{\alpha}_{k'(b)}^*\|$.

Denote $\boldsymbol{Y} = (y_1,\ldots,y_n)^{\mathrm{T}}$ and $\mathbf{X}_{(b)} = \text{diag}\{\mathbf{x}_{1(b)}^{\mathrm{T}},\ldots,\mathbf{x}_{n(b)}^{\mathrm{T}}\}$. If the underlying subgroups $\mathcal{G}_1,\ldots,\mathcal{G}_K$ are known, the oracle estimator $\boldsymbol{\alpha}_{(b)}$ can be defined as

$$\widehat{\boldsymbol{\alpha}}_{(b)}^{\text{or}} = \arg\min_{\boldsymbol{\alpha}_{(b)}\in\mathbb{R}^{KA_b}}\left\{\frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}_{(b)}\mathbf{G}_b\boldsymbol{\alpha}_{(b)}\|^2 + \lambda_1\sum_{k=1}^{K}\sum_{j=1}^{|A_b|}|\mathcal{G}_k|p_1(|\alpha_{k(b)}^j|)\right\}. \tag{13}$$

Then $\widehat{\boldsymbol{\theta}}_{(b)}^{\text{or}} = \mathbf{G}_b\widehat{\boldsymbol{\alpha}}_{(b)}^{\text{or}}$.

We assume several mild and sensible conditions, which are described in detail in the Supporting Information. We then can establish the following consistency results.

Theorem 1. *Under Conditions C1-C4 and C6 (Supporting Information), if $b_n > a\lambda_2$ for some constant $a > 0$, then there exists a local minimizer $\widehat{\boldsymbol{\theta}}_{(b)}$ of objective function (4) satisfying:*

$$P\left(\widehat{\boldsymbol{\theta}}_{(b)} = \widehat{\boldsymbol{\theta}}_{(b)}^{\text{or}}\right)\to 1 \text{ and } \sup_i\|\widehat{\boldsymbol{\theta}}_{i(b)} - \boldsymbol{\theta}_{i(b)}^*\| = O_p(\sqrt{|A_1|/|\mathcal{G}_{\min}|}).$$

Theorem 2. *Denote $\Omega^* = \{\omega : \sum_{b\notin\mathcal{M}}\omega_b = 1\}$. Under Conditions C1-C6 (Supporting Information),*

$$\lim_{n\to\infty}P(\widehat{\omega}\in\Omega^*)\to 1. \tag{14}$$

With these two theorems, $\widehat{\boldsymbol{\theta}}_{\widehat{\omega}}$ converges to $\boldsymbol{\theta}^0$ in probability, where $\widehat{\boldsymbol{\theta}}_{\widehat{\omega}} = (\widehat{\boldsymbol{\theta}}_{1\widehat{\omega}}^{\mathrm{T}},\ldots,\widehat{\boldsymbol{\theta}}_{n\widehat{\omega}}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\theta}^0 = \{(\boldsymbol{\theta}_1^0)^{\mathrm{T}},\ldots,(\boldsymbol{\theta}_n^0)^{\mathrm{T}}\}^{\mathrm{T}}$, and $\boldsymbol{\theta}_i^0$ is the true coefficient. The proofs and additional discussions are provided in the Supporting Information.

3 | SIMULATION

For $n = 100$ independent samples, we generate $p = 100$ dimensional $\mathbf{x}_i$'s from a multivariate normal distribution with marginal means 0 and marginal variances 1. For covariance, we consider an auto-regressive structure with parameter $\rho = 0$, 0.3, and 0.7. The random errors are generated from $N(0, 0.5)$. The response y_i's are generated from the linear regression models. The first four covariates have nonzero coefficients. To examine the "robustness" of partitioning, we randomly shuffle the unimportant covariates so that different subsets can be correlated. We consider

multiple values of B_n. The following simulation settings have been partly motivated by Liu *et al.* (2020) and Städler *et al.* (2010).

Simulation 1 There are two subgroups with coefficients $(-\beta,-\beta,-\beta,-\beta,\mathbf{0}_{p-s})$ and $(\beta,\beta,\beta,\beta,\mathbf{0}_{p-s})$. We use the vector pr to denote the proportions of subjects in different subgroups and consider both balanced and unbalanced designs. Specifically, we consider all combinations by $\beta =$ 1 and 2, pr = (0.5, 0.5) and (0.3, 0.7), and $B_n = 10$ and 5.

Simulation 2 There are three subgroups with coefficients $(-\beta,-\beta,-\beta,-\beta,\mathbf{0}_{p-s})$, $(\beta,\beta,\beta,\beta,\mathbf{0}_{p-s})$, and $(2\beta,2\beta,2\beta,2\beta,\mathbf{0}_{p-s})$, where $\beta = 2$. For their relative proportions, we consider pr $= (1/3, 1/3, 1/3)$ and (0.3,0.3,0.4). Set $B_n = 10$.

Simulation 3 There are two subgroups with coefficients $(-\beta,-\beta,-\beta,-\beta,\mathbf{0}_{p-s})$ and $(\beta,\beta,0,0,\beta,\beta,\mathbf{0}_{p-s-2})$. pr $=$ $(0.5, 0.5)$ and $(0.3, 0.7)$, $\beta = 1$ and 2, and $B_n = 10$.

To assess subgrouping performance, we examine the number of identified subgroups and accuracy of subject subgrouping results (Accuracy). To assess variable selection performance, we consider the rates of TP (true positive) and FP (false positive). In addition, estimation performance is evaluated using the MSE (mean squared error).

We consider the following alternatives: (a) The FMR approach developed in Khalili and Chen (2007) (referred to as "KC"), where Lasso is applied for accommodating high dimensionality and selecting relevant variables. It is realized using the R package *fmrs*. This is the most relevant competitor and represents sparse FMR approaches. (b) We consider three low-dimensional FMR approaches, which apply the FMR technique without sparsity. First, we consider the "True" approach, under which the truly important covariates are known, and only such covariates are used (but the subgrouping structure needs to be determined). Second, we consider a model with p/B_n covariates, which contain all the important covariates along with a few unimportant ones. As this is a fitted model, this approach is referred to as "Fitted." Third, this approach is similar to the above one, with the difference that it contains half of the important covariates. As this is an underfitted model, this approach is referred to as "Underfitted." (c) We consider a partially "oracle" approach, under which the subgrouping structure is known (as such, penalty P_2 is not needed), and a model averaging approach similar to the proposed is adopted for estimation. The FMR-based approaches, both high- and low-dimensional, need to determine the number of subgroups. We set the number of subgroups as the true value, which leads to favorable performance. Note that this is not needed with the proposed approach and not practical in practice. We have also experimented with the sparse Kmeans and other sparse clustering methods but found unacceptable results. Such methods are omitted from our reporting.

For each setting, we simulate 100 replicates. In Table 1, we examine the number of identified subgroups using mean, median, standard deviation, and percentage of correct identification. Across the whole spectrum, the proposed approach can satisfactorily identify the number of true subgroups. Quite a few scenarios have 100% correct identification. In contrast, literature suggests that, with the FMR and many other heterogeneity analysis techniques, it is extremely difficult to determine the number of subgroups. We further examine three representative settings with $\rho = 0, \beta = 1$, and $(K, B_n) = (2, 10)$ and $(2, 5)$, as well as with $\rho = 0, \beta = 2$, and $(K, B_n) = (3, 10)$. In Figure A1 (Supporting Information), we plot the average weights of the B_n candidate models. Note that to improve presentation, we always keep the first submodel as the one that includes all of the important covariates. In all plots, a spike of the first submodel is observed. The rest of the submodels have almost or exactly zero weights. Results in Table 1 clearly demonstrate the superiority of penalized fusion in this aspect. Subgrouping, estimation, and variable selection accuracy results are summarized in Table 2 for Simulation 1 and Table A1 (Supporting Information) for Simulation 2 and 3. With the complexity brought by heterogeneity, the proposed approach behaves inferior to the oracle as expected. It has significant advantages over the KC approach. Consider, for example, Simulation 1, $B_n = 5$, $\rho = 0$, and $\alpha = 1$. The proposed approach has (Accuracy, MSE, TP, FP) equal to (0.781, 0.785, 0.962, 0.010), compared to (0.500, 7.483, 0.331, 0.223) for the KC approach. Compared to the True approach (which is oracle in terms of variable selection), it has slightly inferior Accuracy and much inferior MSE. But it has advantageous performance over the Fitted and Underfitted approaches. Table 2 also suggests that the value of B_n does not have a substantial impact. Simulation 2 and 3 lead to similar findings.

With the proposed approach, partition of variables is needed. In our implementation, we partition consecutively. Different orderings of the variables can lead to different partitions. To examine the impact of ordering/partition, we consider Simulation 1 with $K = 2$, $B_n = 5, \rho = 0, \beta = 1$, and pr $= (0.5, 0.5)$. For each simulated replicate, we permute the variables 10 times. With each permutation, the proposed approach is applied. Then the mean and standard deviation of the summary statistics considered in Tables 1 and 2 are computed. We further compute the averages of such mean and standard deviation values across 100 replicates. For the mean, median, SD, and per values as considered in Table 1, the average mean (standard deviation) values are 1.942 (0.018), 2 (0), 0.243 (0.021), and 0.938 (0.011), respectively. For the Accuracy, MSE, TP,

TABLE 1 Simulation: mean, median, standard deviation (SD) of $\widehat{K}$, and percentage (per) of $\widehat{K}$ equal to the true number of subgroups

K	B_n	ρ	β	Mean	Median	SD	per	Mean	Median	SD	per
Simulation 1				pr = (0.5, 0.5)				pr = (0.3, 0.7)			
2	10	0	1	2	2	0	1	1.74	2	0.443	0.74
			2	2.02	2	0.141	0.98	2.04	2	0.198	0.96
		0.3	1	2	2	0	1	1.96	2	0.198	0.96
			2	2	2	0	1	2.04	2	0.198	0.96
		0.7	1	1.98	2	0.141	0.98	1.98	2	0.141	0.98
			2	2	2	0	1	2	2	0	1
	5	0	1	1.95	2	0.221	0.95	1.82	2	0.388	0.82
			2	2	2	0	1	2	2	0	1
		0.3	1	2	2	0	1	1.96	2	0.198	0.96
			2	2	2	0	1	2	2	0	1
		0.7	1	2	2	0	1	1.94	2	0.24	0.94
			2	2	2	0	1	2	2	0	1
Simulation 2				pr = $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$				pr = (0.3, 0.3, 0.4)			
3	10	0	2	2.84	3	0.468	0.82	2.98	3	0.589	0.88
		0.3	2	3	3	0	1	3	3	0	1
		0.7	2	3.02	3	0.141	0.98	3.62	3	3.276	0.88
Simulation 3				pr = (0.5, 0.5)				pr = (0.3, 0.7)			
2	10	0	1	1.88	2	0.328	0.88	1.68	2	0.471	0.68
			2	2	2	0	1	2	2	0	1
		0.3	1	1.96	2	0.198	0.96	1.82	2	0.388	0.82
			2	2	2	0	1	2.02	2	0.141	0.98
		0.7	1	1.82	2	0.431	0.783	1.78	2	0.465	0.74
			2	2	2	0	1	2	2	0	1

and FP values as considered in Table 2, the average mean (standard deviation) values are 0.813 (0.005), 0.800 (0.058), 0.956 (0.017), and 0.005 (0.001), respectively. The average means are very close to their counterparts in Tables 1 and 2, and the small standard deviations suggest the stability of results. This analysis suggests that the ordering of the variables is not critical.

In the Supporting Information, we additionally (a) examine a sequence of P-values, compare with the direct application of double penalization, and "re-establish" the advantage of model averaging, (b) consider higher dimensionality and show that the proposed approach still has advantageous performance, (c) examine performance when the sub-Gaussian assumption is not satisfied, and (d) evaluate the sensitivity of the analysis results to tuning parameter selection. Overall, satisfactory performance is observed.

4 | DATA ANALYSIS

The Cancer Genome Atlas (TCGA) is a collective effort organized by the NIH and has published high-quality clinical, omics, and imaging data on multiple cancer types. Compared to the clinical and omics data, the TCGA imaging data have been much less analyzed. However, several recent publications have shown that the analysis of TCGA histopathological imaging data can lead to important insights on disease classification, prognosis, and other outcomes (Noorbakhsh *et al.*, 2019). Here we consider the breast cancer (BRCA) data. The response variable of interest is the ratio between "Positive Finding Lymph Node Hematoxylin and Eosin Staining Microscopy Count" and "Lymph Node(s) Examined Number". It reflects the degree of treatment. In the literature, it has been suggested that treatment decisions depend on tumor properties, which are reflected in histopathological images, and the heterogeneity in breast cancer treatment has been observed. As such, it is biologically sensible to conduct heterogeneity analysis based on imaging features for this specific outcome. We focus on "nontrivial" ratios, which fall between 0 and 1, and conduct the transformation $\log(\frac{\text{ratio}}{1-\text{ratio}})$. The histopathological images are downloaded from the TCGA website. The pipeline for extracting imaging features has been implemented in recent studies (Zhong *et al.*, 2019) and briefly summarized

TABLE 2 Simulation 1: accuracy rate of correctly identifying subgroup memberships (accuracy), mean squared error, TP, and FP rates

					pr = (0.5, 0.5)						pr= (0.3, 0.7)					
K	B_n	ρ	β	Index	Proposed	Oracle	KC	True	Fitted	Underfitted	Proposed	Oracle	KC	True	Fitted	Underfitted
2	10	0	1	Accuracy	0.826		0.498	0.837	0.777	0.514	0.798		0.551	0.872	0.839	0.546
				MSE	0.541	0.054	7.563	1.206	2.185	6.605	2.318	0.068	6.757	0.059	1.676	7.061
				TP	1	1	0.260				0.960	1	0.408			
				FP	0.004	0.014	0.202				0.007	0.023	0.187			
			2	Accuracy	0.914		0.503	0.910	0.828	0.519	0.939		0.549	0.942	0.908	0.561
				MSE	0.525	0.074	29.675	1.618	6.879	24.996	0.684	0.088	26.868	0.064	36.437	25.509
				TP	1	1	0.288				1	1	0.490			
				FP	0.002	0.014	0.206				0.005	0.014	0.211			
		0.3	1	Accuracy	0.868		0.499	0.863	0.815	0.541	0.879		0.553	0.901	0.858	0.578
				MSE	0.293	0.056	7.804	1.451	2.149	7.069	0.655	0.057	6.803	0.064	1.764	7.013
				TP	1	1	0.205				1	1	0.390			
				FP	0.001	0.014	0.208				0.005	0.014	0.209			
			2	Accuracy	0.914		0.495	0.922	0.886	0.495	0.951		0.537	0.952	0.935	0.591
				MSE	0.577	0.026	31.438	0.035	1.054	28.642	0.x326	0.106	27.596	0.071	2.191	30.680
				TP	1	1	0.375				1	1	0.510			
				FP	0.003	0.010	0.427				0.004	0.013	0.249			
		0.7	1	Accuracy	0.864		0.513	0.901	0.836	0.627	0.904		0.594	0.921	0.852	0.676
				MSE	0.613	0.109	7.687	0.114	10.881	9.202	0.560	0.149	7.421	0.134	12.588	10.278
				TP	0.940	1	0.272				0.970	0.998	0.410			
				FP	0.008	0.017	0.232				0.010	0.020	0.200			
			2	Accuracy	0.940		0.500	0.951	0.899	0.655	0.960		0.531	0.959	0.930	0.699
				MSE	0.262	0.131	30.561	0.107	12.624	30.766	0.315	0.177	32.172	0.128	40.788	43.618
				TP	1	1	0.348				1	1	0.415			
				FP	0.027	0.016	0.349				0.009	0.017	0.312			
	5	0	1	Accuracy	0.781		0.500	0.853	0.577	0.503	0.781		0.558	0.880	0.671	0.526
				MSE	0.785	0.047	7.483	0.048	5.053	6.705	1.951	0.055	6.534	0.062	2.928	6.753
				TP	0.962	1	0.331				0.998	1	0.395			
				FP	0.010	0.020	0.223				0.013	0.020	0.208			
			2	Accuracy	0.906		0.503	0.922	0.675	0.506	0.895		0.542	0.924	0.753	0.516
				MSE	0.592	0.071	29.489	0.05	15.422	26.435	0.781	0.080	26.542	0.056	9.353	30.065
				TP	1	1	0.292				1	1	0.445			
				FP	0.002	0.009	0.204				0.008	0.017	0.187			
		0.3	1	Accuracy	0.848		0.503	0.885	0.627	0.509	0.826		0.562	0.903	0.699	0.524
				MSE	0.295	0.057	7.475	0.060	5.829	7.500	0.938	0.125	6.397	0.075	4.557	8.343
				TP	1	1	0.238				1	0.998	0.430			
				FP	0.002	0.011	0.231				0.008	0.015	0.212			
			2	Accuracy	0.923		0.501	0.935	0.695	0.509	0.922		0.527	0.943	0.823	0.528
				MSE	0.291	0.071	29.631	0.051	20.094	37.952	2.721	0.072	26.921	0.070	7.417	29.544
				TP	1	1	0.264				0.986	1	0.500			
				FP	0.001	0.011	0.261				0.020	0.013	0.299			
		0.7	1	Accuracy	0.864		0.525	0.911	0.660	0.545	0.859		0.586	0.918	0.744	0.556
				MSE	0.295	0.084	7.958	0.102	8.008	10.778	1.079	0.348	7.122	0.173	15.904	10.603
				TP	1	1	0.293				0.982	0.980	0.388			
				FP	0.003	0.013	0.233				0.012	0.026	0.216			
			2	Accuracy	0.934		0.509	0.931	0.702	0.537	0.911		0.536	0.941	0.751	0.560
				MSE	0.307	0.135	32.608	8.184	93.655	51.923	1.255	0.188	34.588	0.141	38.747	62.379
				TP	1	1	0.338				1	1	0.533			
				FP	0.026	0.015	0.312				0.138	0.016	0.361			

TABLE 3 Data analysis using the proposed approach: identified imaging features for the three subgroups

Imaging feature	Group 1	Group 2	Group 3
Texture-AngularSecondMoment-ImageAfterMath-3-01		5.620	7.587
Texture-SumAverage-ImageAfterMath-3-03		−5.473	−3.318
Texture-SumVariance-ImageAfterMath-3-02	−0.047	2.873	2.873
AreaShape-Zernike-7-3	4.988	−0.098	
Granularity-10-ImageAfterMath.1		−1.579	−1.048
Granularity-15-ImageAfterMath.1		−1.278	0.028
Threshold-WeightedVariance-Identifyhemasub2		−0.114	−0.245
AreaShape-Zernike-5-5		−0.259	
Location-Center-Y.3	−0.099	−0.099	
AreaShape-Zernike-4-2		0.168	
Texture-Entropy-ImageAfterMath-3-02			0.039
AreaShape-Zernike-4-0	−0.622	−0.622	0.047
Texture-InfoMeas1-maskosingray-3-00	−0.384	−0.384	
Granularity-1-ImageAfterMath	0.815	0.958	
Texture-SumVariance-maskosingray-3-03	0.059	0.069	
AreaShape-FormFactor	0.059	0.059	0.059
AreaShape-Zernike-3-3	0.038	0.038	−0.011
Granularity-1-ImageAfterMath.1	−0.012		−0.054
Texture-InfoMeas2-ImageAfterMath-3-00			0.062
Granularity-4-ImageAfterMath.1	−0.016	−0.016	0.017
Texture-SumVariance-ImageAfterMath-3-00	−0.017	−0.017	0.011
Texture-SumEntropy-maskosingray-3-01			−0.014
Texture-InverseDifferenceMoment-ImageAfterMath-3-03			−0.011

in Figure A2 (Supporting Information). It includes four main steps, namely image chopping, subimage selection, feature extraction, and feature averaging. We refer to Zhong *et al.* (2019) for more details on each step and quality control. The final analyzed data set contains measurements on 139 subjects and 248 imaging features, which describe tumor properties including texture, granularity, size and shape, neighbor distribution, occupation, and areafraction.

Three distinct subgroups are identified, with sizes 49, 35, and 55. The identified imaging features and their estimates are shown in Table 3. Among the identified features, 10 are related to texture, six are related to area shape, and five are related to granularity. Overall, the three subgroups have significantly different models. It is also noted that some features have identical estimates in different subgroups, which can be caused by the promotion of equality by P_2 and termination of calculation when two consecutive estimates are close enough. Compared to clinical and molecular data, the biological implications of high-dimensional imaging features are still largely unclear (Luo *et al.*, 2017). As such, we defer biological interpretations to future research.

We consider the following alternatives: KC, Alt.1 which uses the 23 selected imaging features and applies penalized fusion for subgrouping, Alt.2 that uses the 23 selected imaging features and applies FMR for subgrouping, sparse Kmeans, and sparse hierarchical clustering. To make different approaches comparable, we fix the number of subgroups as three with the alternatives. We compute a subgrouping similarity measure, with range [0,1] and a larger value indicating a higher degree of similarity, and find that the proposed approach has moderate similarity with the alternatives: 0.495 (KC), 0.621 (Alt.1), 0.604 (Alt.2), 0.556 (sparse Kmeans), and 0.465 (sparse hierarchical clustering). The KC approach identifies 29 features, which have three overlapping with the proposed. Alt.1 and Alt.2 use the same set of 23 imaging features as the proposed approach. With the sparse Kmeans and hierarchical clusterings, tiny subgroups are generated, making the assessment of variable selection unreliable. We evaluate prediction using a random splitting approach (with training: testing = 3:1 and 100 splits). The prediction MSEs are 0.804 (proposed), 1.076 (KC), 1.082 (Alt.1), and 1.181 (Alt.2). Prediction with the sparse clustering approaches

is not possible as estimation cannot be reliably conducted. Overall, the proposed approach makes different subgrouping and identification, with improved prediction performance.

5 | DISCUSSION

We have conducted cancer heterogeneity analysis using high-dimensional imaging features and the penalized fusion technique. We have applied additional penalization to accommodate high data dimension and screen out noises. Another significant advancement is the adoption of model averaging to tackle computational challenges. Beyond providing a solid ground, the theoretical investigation can also shed light on high-dimensional penalized fusion and model averaging in general. Simulation has demonstrated competitive performance. In the analysis of TCGA data, findings different from the alternatives have been made, and improved prediction is observed. Overall, this study has delivered an alternative technique for supervised heterogeneity analysis and a new venue for modeling cancer heterogeneity.

Beyond imaging features, the proposed analysis can also be conducted with other high-dimensional variables. It will also be of interest to adapt the proposed technique and apply to other data distributions/models, which can be achieved by replacing the lack-of-fit measure. The proposed computational algorithm will be applicable with minor revisions, but additional theoretical developments may be needed. In data analysis, we have identified three subgroups with significantly different regression models. In the literature, there is still a lack of commonly accepted approaches for validating heterogeneity analysis results under the mixture regression framework. It is noted that the identified subgroups differ in the relationship between imaging features and a clinical outcome. However, they may or may not differ in other clinical aspects. As the functional implications of imaging features have not been well examined, we are unable to make further biological interpretations. Nevertheless, the satisfactory simulation results and improved prediction can support the validity of our findings to a great extent. Further analysis and validation will be needed prior to any application of the findings.

ACKNOWLEDGMENTS

We thank the editors and reviewers for their careful review and insightful comments. This work was partly supported by the National Natural Science Foundation of China [11971404], MOE (Ministry of Education in China) Project of Humanities and Social Sciences [19YJC910010], Natural Science Foundation [1916251], a Yale Cancer Center Pilot Award, and National Institutes of Health [CA241699, CA204120].

DATA AVAILABILITY STATEMENT

Data analyzed in this paper are publicly available at the TCGA website (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga).

ORCID

Yanyan Liu https://orcid.org/0000-0002-1405-7345
Shuangge Ma https://orcid.org/0000-0001-9001-4999

REFERENCES

Ando, T. and Li, K.C. (2017) A weight-relaxed model averaging approach for high dimensional generalized linear models. *The Annals of Statistics*, 45, 2645–2679.

Belhomme, P., Toralba, S., Plancoulaine, B., Oger, M., Gurcan, M.N. and Bor-Angelier, C. (2015) Heterogeneity assessment of histological tissue sections in whole slide images. *Computerized Medical Imaging and Graphics*, 42, 51–55.

Dagogo-Jack, I. and Shaw, A.T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15, 81–94.

Dai, D., Rigollet, P. and Zhang, T. (2012) Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics*, 40, 1878–1905.

Foster, J. C., Taylor, J. M. G. and Ruberg, S. J. (2011) Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30, 2867–2880.

Khalili, A. and Chen, J. (2007) Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102, 1025–1038.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., et al., (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499, 214–218.

Liu, M., Zhang, Q., Fang, K. and Ma, S. (2020) Structured analysis of the high-dimensional FMR model. *Computational Statistics & Data Analysis*, 144, 106883.

Luo, X., Zang, X., Yang, L., Huang, J., Liang, F., Rodriguez-Canales, J., et al., (2017) Comprehensive computational pathological image analysis predicts lung cancer prognosis. *Journal of Thoracic Oncology*, 12, 501–509.

Ma, S. and Huang, J. (2017) A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112, 410–423.

McLachlan, G.J. and Peel, D. (2000) *Finite Mixture Models*. Hoboken, NJ: John Wiley & Sons.

Noorbakhsh, J., Farahmand, S., Soltanieh-ha, M., Namburi, S., Zarringhalam, K. and Chuang, J. (2019) Pan-cancer classifications of tumor histological images using deep learning. Preprint, https://www.biorxiv.org/content/10.1101/715656v1

Städler, N., Bühlmann, P. and van de Geer, S. A. (2010) ℓ_1-penalization for mixture regression models. *Test*, 19, 209–256.

Tibshirani, S., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *Journal of Royal Statistical Society, Series B (Statistical Methodology)*, 67, 91–108.

Wager, S. and Athey, S. (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113, 1228–1242.

Wang, H. S., Li, B. and Leng, C. L. (2009) Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of Royal Statistical Society, Series B (Statistical Methodology)*, 71, 671–683.

Wang, S., Yang, D.M., Rong, R., Zhan, X. and Xiao, G. (2019) Pathology image analysis using segmentation deep learning algorithms. *The American Journal of Pathology*, 189, 1686–1698.

Zhang, X., Zou, G., Liang, H. and Carroll, R. J. (2020) Parsimonious model averaging with a diverging number of parameters. *JASA*, 115, 972–984.

Zhong, T., Wu, M. and Ma, S. (2019). Examination of independent prognostic power of gene expressions and histopathological imaging features in cancer. *Cancers*, 11, 361.

Zhu, X. and Qu, A. (2018) Cluster analysis of longitudinal profiles with subgroups. *Electronic Journal of Statistics*, 12, 171–193.

SUPPORTING INFORMATION

Details of the computational algorithms, proofs, and additional numerical results are available at the Biometrics website on Wiley Online Library. R programs implementing the proposed method are available at www.github.com/shuanggema.

How to cite this article: 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*. 2020;1–12. https://doi.org/10.1111/biom.13357