Statistica Sinica **33** (2023), 729-758
doi:https://doi.org/10.5705/ss.202021.0002

HETEROGENEITY ANALYSIS VIA INTEGRATING MULTI-SOURCES HIGH-DIMENSIONAL DATA WITH APPLICATIONS TO CANCER STUDIES

Tingyan Zhong, Qingzhao Zhang, Jian Huang, Mengyun Wu and Shuangge Ma

Shanghai Jiao Tong University, Xiamen University, The Hong Kong Polytechnic University, Shanghai University of Finance and Economics and Yale University

Abstract: This study has been motivated by cancer research, in which heterogeneity analysis plays an important role and can be roughly classified as unsupervised or supervised. In supervised heterogeneity analysis, the finite mixture of regression (FMR) technique is used extensively, under which the covariates affect the response differently in subgroups. High-dimensional molecular and, very recently, histopathological imaging features have been analyzed separately and shown to be effective for heterogeneity analysis. For simpler analysis, they have been shown to contain overlapping, but also independent information. In this article, our goal is to conduct the first and more effective FMR-based cancer heterogeneity analysis by integrating high-dimensional molecular and histopathological imaging features. A penalization approach is developed to regularize estimation, select relevant variables, and, equally importantly, promote the identification of independent information. Consistency properties are rigorously established. An effective computational algorithm is developed. A simulation and an analysis of The Cancer Genome Atlas (TCGA) lung cancer data demonstrate the practical effectiveness of the proposed approach. Overall, this study provides a practical and useful new way of conducting supervised cancer heterogeneity analysis.

Key words and phrases: Cancer heterogeneity, data integration, FMR, molecular and imaging features.

1. Introduction

Heterogeneity is a hallmark of cancer, and thus has gained extensive research (Turajlic et al. (2019)). Heterogeneity analysis can be roughly classified as unsupervised or supervised. In unsupervised analysis, outcomes/phenotypes are not involved, and clustering and other techniques are adopted (Wiwie, Baumbach and Röttger (2015)). Unsupervised analysis can be useful, for example, for identifying new disease subtypes, but it is often difficult to associate clinical implications

Corresponding author: Mengyun Wu. E-mail: wu.mengyun@mail.shufe.edu.cn. Shuangge Ma. E-mail: shuangge.ma@yale.edu.

with findings. In contrast, supervised analysis directly addresses the heterogeneity associated with a clinical outcome/phenotype, and often has more important practical implications (Bair (2013)). In such analysis, it is postulated that covariates affect the response differently in subject subgroups (Stadler, Buhlmann and van De Geer (2010); Hui, Warton and Foster (2015)). This may manifest as different covariates being associated with the response and/or the same covariates having different magnitudes of effects. Note that, here, subject subgroups are unknown a priori and need to be estimated. This is different to the analysis that considers interactions between known subject groups and biomarkers, which is sometimes also referred to as "heterogeneity analysis" and is often conducted to study treatment effects (Coppock, Leeperand Mullinix (2018)).B

In "classic" heterogeneity analysis, clinical/demographic/environmental variables have been considered. In the past two decades, molecular data have played an increasingly important role in cancer research, and, in particular, in supervised heterogeneity analysis (Ahmad and Fröhlich (2017)). Another type of data, recently suggested as informative for modeling cancer outcomes/phenotypes, comes from histopathological images. Such images are generated in a biopsy, which is ordered for most suspected cases, and are used extensively for definitive diagnosis and staging. They contain information on a tumor's "micro" properties and surrounding microenvironment. They differ significantly from radiological images, which are generated by CT, PET, and other techniques, and provide information on a tumor's "macro" properties, such as location, size, and density. Recent studies, such as Luo et al. (2017), have analyzed high-dimensional histopathological imaging features for modeling biomarkers, survival, and other outcomes. Furthermore, a handful of studies, such as Kothari et al. (2013) and Althobiti et al. (2018), conduct imaging-based heterogeneity analysis. However, they often analyze low-dimensional imaging features and adopt relatively simple techniques.

A tumor's properties and microenvironment, as reflected in histopathological images, are affected but not fully regulated by molecular changes. As such, molecular and imaging data contain *overlapping and independent* information. This is supported by recent studies that have explicitly analyzed the relationship between the two types of data. For example, Yu et al. (2017) use a random forest to correlate molecular data with histopathological imaging data, finding that these two types of data have overlapping information, with some significant associations detected. Zhong, Wu and Ma (2019) adopt a hypothesis testing approach, showing that the two types of data have independent information, when modeling cancer prognosis. Under the homogeneity assumption, studies such as Sun et al. (2018) and Mobadersany et al. (2018) show that integrating the two types of data

leads to biologically sensible models with improved estimation/prediction performance. Complementing and advancing the existing literature, in this study, we take the natural next step and conduct cancer heterogeneity analysis by integrating high-dimensional molecular and imaging data.

In supervised heterogeneity analysis, the finite mixture of regression (FMR) technique has been adopted extensively because of its lucid interpretations and satisfactory statistical and numerical properties (McLachlan and Peel (2000)). Here, the conditional distribution of the response y given the covariates $\boldsymbol{X}$ is a mixture with multiple components, and the relationship between y and $\boldsymbol{X}$ varies across such components. For example, under the "classic" mixture of two normal distributions, $y|\boldsymbol{X} \sim \mu N(\boldsymbol{X}\boldsymbol{\alpha}_1, \sigma^2) + (1-\mu)N(\boldsymbol{X}\boldsymbol{\alpha}_2, \sigma^2)$ with different coefficient vectors $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$. Examples of FMR-based studies with low-dimensional covariates include Chen, Chen and Kalbfleisch (2001) and Atienza et al. (2007), and those with high-dimensional covariates include Khalili and Chen (2007) and Hui, Warton and Foster (2015). Note that these and other similar studies in the literature are limited to a single type of covariate.

When there are two or more types of covariates from different sources and with different properties, the simplest solution is to stack them together, after which variable selection or dimension reduction techniques can be applied. Examples include the Lasso-based approach in Boulesteix et al. (2017) and the elastic net and sparse principal component analysis in Jiang et al. (2016). However, such a strategy fails to account for overlapping information, which can manifest statistically as correlation. Approaches such as collaborative regression (Gross and Tibshirani (2015)) and canonical variate regression (Luo et al. (2016)) can accommodate overlapping information via canonical correlation analysis. As another example, the assisted robust marker identification (ARMI) approach developed in Chai et al. (2017) borrows overlapping information from one type of covariate to assist more accurate identification on the other type(s) of covariates. However, these approaches model the response using each type of covariate separately, and cannot effectively accommodate independent information contained in multiple types of covariates. In addition, they have not been applied to heterogeneity analysis. There are approaches that decompose data and use only non-overlapping information in modeling based on penalization (Zhu et al. (2016)) and Bayesian (Wang et al. (2013)) techniques. However, the decomposed data do not have clear interpretations, and these studies are also limited to the homogeneity case.

This study has been motivated by the critical importance of supervised cancer heterogeneity analysis, the increase in the number of studies that collect both molecular and histopathological imaging data, the overlapping and independent

information contained in such data, and a lack of studies that integrate them for heterogeneity analysis. Our study complements and advances the existing literature in multiple ways. In particular, we extend those works limited to a single type of covariate by effectively integrating molecular and histopathological imaging data. We also extend studies limited to low-dimensional covariates (and thus limited information) by accommodating high-dimensional and noisy covariates using a penalization technique. Furthermore, we advance the collaborative regression and ARMI by building models using both types of data (thus, using more information). In addition, without data decomposition, the resulting models can be biologically more interpretable. We also rigorously show that the proposed approach has satisfactory theoretical and numerical properties. Overall, this study provides a new and practically useful way of modeling cancer heterogeneity. Note that supervised heterogeneity analysis is not limited to cancer, and data integration is not limited to molecular and imaging data. As such, the proposed approach can enjoy broad applicability far beyond that proposed here.

2. Methods

2.1. Integrated heterogeneity analysis

Assume n independent subjects. For the ith subject, denote y_i as the response of interest, and $\boldsymbol{X}_{i\cdot} = (x_{i1}, \ldots, x_{ip})$ and $\boldsymbol{Z}_{i\cdot} = (z_{i1}, \ldots, z_{iq})$ as the p- and q-dimensional molecular and imaging measurements, respectively. The conditional density of y_i given $\boldsymbol{X}_{i\cdot}$ and $\boldsymbol{Z}_{i\cdot}$ is

$$f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \mu_k g\left(y_i; h(\boldsymbol{X}_{i\cdot}\boldsymbol{\alpha}_k + \boldsymbol{Z}_{i\cdot}\boldsymbol{\beta}_k), \sigma_k\right). \tag{2.1}$$

Here, K is the number of mixture components (subgroups), $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)'$ is the vector of mixing proportions satisfying $\mu_k > 0$ and $\sum_{k=1}^{K} \mu_k = 1$, $g(\cdot)$ is the known density function, $h(\cdot)$ is the known link function, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_K)'$ is an unknown parameter vector usually related to the variance, $\boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{kp})'$ and $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kq})'$ are the coefficient vectors for the molecular and imaging measurements, respectively, and $\boldsymbol{\theta} = (\boldsymbol{\mu}', \boldsymbol{\sigma}', \boldsymbol{\alpha}', \boldsymbol{\beta}')' = (\boldsymbol{\mu}', \boldsymbol{\sigma}', \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_K', \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_K')' \triangleq (\theta_j)_{(2K+Kp+Kq)\times 1}$.

We propose the following penalized objective function:

$$Q_{L_0}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \mu_k g\left(y_i; h(\boldsymbol{X}_{i\cdot}\boldsymbol{\alpha}_k + \boldsymbol{Z}_{i\cdot}\boldsymbol{\beta}_k), \sigma_k\right) \right\} - n \sum_{k=1}^{K} \sum_{j=1}^{p} \rho(|\alpha_{kj}|; \gamma, \lambda_1)$$

$$-n\sum_{k=1}^{K}\sum_{l=1}^{q}\rho(|\beta_{kl}|;\gamma,\lambda_1)-n\lambda_2\sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{l=1}^{q}c_{jl}1(\alpha_{kj}\neq 0)1(\beta_{kl}\neq 0), \quad (2.2)$$

where $\rho(|\nu|;\gamma,\lambda_1) = \lambda_1\int_0^{|\nu|}(1-x/\lambda_1\gamma)_+dx$ is the Minimax Concave Penalty (MCP) with regularization parameter γ, $(a)_+ = \max\{a,0\}$, $1(\cdot)$ is the indicator function, and λ_1 and λ_2 are the tuning parameters. Here, γ controls the unbiasedness and concavity of the estimator, with a larger value leading to a smoother estimation, but a larger bias and less accurate variable selection (Zhang (2010)). In addition, c_{jl} describes the amount of overlapping information between the jth component of $\boldsymbol{X}$ and the lth component of $\boldsymbol{Z}$, with a larger value indicating a higher level of overlapping. In the literature, there are multiple ways of quantifying overlapping information. Given that overlapping information can manifest as correlation, we propose $c_{jl} = |c_{jl}^{Pcorr}|1(|c_{jl}^{Pcorr}| \geq c^{Pcorr})$, where c_{jl}^{Pcorr} is the Pearson's correlation between the jth molecular and lth imaging variables, and c^{Pcorr} is the cutoff. Correlation perhaps provides the simplest and most straightforward quantification of overlapping information, and has been used extensively. The cutoff c^{Pcorr} is introduced to remove (a large number of) spurious correlations. With the maximizer of (2.2), the nonzero components of $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ correspond to the important molecular and imaging variables that are associated with the response for the kth subgroup.

The discontinuity of the L_0 penalty makes optimization challenging. To improve computational feasibility, we further propose

$$\begin{aligned}Q(\boldsymbol{\theta}) = &\sum_{i=1}^{n}\log\left\{\sum_{k=1}^{K}\mu_k g\left(y_i;h(\boldsymbol{X}_{i\cdot}\boldsymbol{\alpha}_k+\boldsymbol{Z}_{i\cdot}\boldsymbol{\beta}_k),\sigma_k\right)\right\}-n\sum_{k=1}^{K}\sum_{j=1}^{p}\rho(|\alpha_{kj}|;\gamma,\lambda_1)\\ &-n\sum_{k=1}^{K}\sum_{l=1}^{q}\rho(|\beta_{kl}|;\gamma,\lambda_1)-n\lambda_2\sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{l=1}^{q}c_{jl}\left(1-e^{-\alpha_{kj}^2/\tau}\right)\left(1-e^{-\beta_{kl}^2/\tau}\right),\end{aligned} \quad (2.3)$$

where τ is a small positive constant that controls the goodness and smoothness of the approximation.

Rationale. In contrast to existing FMR models, the proposed model includes two distinct types of high-dimensional variables. Furthermore, in contrast to, for example, the collaborative regression and ARMI, both molecular and imaging data are included in a single model to take advantage of their independent information. Penalization is adopted for regularization and sparsity. We adopt the MCP because of its satisfactory statistical properties, such as unbiasedness,

and better numerical performance than some other penalties, such as Lasso. In (2.2), the key advancement is the last term, which *promotes the identification of molecular and imaging variables with smaller correlations (weaker overlapping information).* In particular, the indicator functions $1(\alpha_{kj} \neq 0)$ and $1(\beta_{kl} \neq 0)$ pick up the selected molecular and imaging variables, and the penalty is defined as the sum of the absolute values of their pair-wise correlations. This way, the proposed approach directly encourages the selection of molecular and imaging variables with weak correlations, and effectively accommodates their overlapping information. Note that directly including two types of covariates in a single model without properly accommodating their high correlations may lead to unreliable and inaccurate estimation and identification. For two molecular (imaging) variables with similar contributions to the model, the proposed correlation-based penalty selects the one less correlated with important imaging (molecular) variables. As a result, the identified model contains less redundant information, leading to more reliable and accurate estimation and identification. In addition, these important molecular and imaging variables have more independent contributions, and may provide richer information for understanding the response. The smooth approximation of the indicator function simplifies the computation, and the exponential-based approximation can be replaced by other smooth approximations.

2.2. Statistical properties

Assume K is known. Determining its value under FMR is nontrivial, but has been discussed in the literature (Khalili and Lin (2013)). Let $\boldsymbol{\theta}^0 = ((\boldsymbol{\mu}^0)', (\boldsymbol{\sigma}^0)', (\boldsymbol{\alpha}_1^0)', \ldots, (\boldsymbol{\alpha}_K^0)', (\boldsymbol{\beta}_1^0)', \ldots, (\boldsymbol{\beta}_K^0)')'$ be the vector of true parameter values. Let $\mathcal{A}_k = \{j : \alpha_{kj}^0 \neq 0\}$, $\mathcal{B}_k = \{l : \beta_{kl}^0 \neq 0\}$, $\mathcal{C} = \{k : \theta_k^0 \neq 0\}$, and $\mathcal{C}^c = \{k : \theta_k^0 = 0\}$, where θ_k^0 is the kth element of $\boldsymbol{\theta}^0$. Note that μ_k^0 and σ_k^0 are nonzero. Denote $|\mathcal{A}|$ as the cardinality of set $\mathcal{A}$. Let $a_k = |\mathcal{A}_k|$, $b_k = |\mathcal{B}_k|$, and $s = 2K + \sum_{k=1}^K a_k + \sum_{k=1}^K b_k$. Assume that the nonsparsity size $s \ll n$. For a vector $\boldsymbol{\nu}$ and index set $\mathcal{S}$, denote $\boldsymbol{\nu}_{\mathcal{S}}$ as the components of $\boldsymbol{\nu}$ indexed by $\mathcal{S}$. For a matrix $\boldsymbol{M}$ and two index sets $\mathcal{S}_1$ and $\mathcal{S}_2$, denote $\boldsymbol{M}_{\cdot\mathcal{S}_1}$ and $\boldsymbol{M}_{\mathcal{S}_1\cdot}$ as the columns and rows, respectively, of $\boldsymbol{M}$ indexed by $\mathcal{S}_1$, and denote $\boldsymbol{M}_{\mathcal{S}_1,\mathcal{S}_2}$ as the submatrix of $\boldsymbol{M}$ indexed by $\mathcal{S}_1$ and $\mathcal{S}_2$.

Denote $\boldsymbol{\theta}_{\mathcal{C}}^* = ((\boldsymbol{\mu}^*)', (\boldsymbol{\sigma}^*)', (\boldsymbol{\alpha}_{1,\mathcal{A}_1}^*)', \ldots, (\boldsymbol{\alpha}_{K,\mathcal{A}_K}^*)', (\boldsymbol{\beta}_{1,\mathcal{B}_1}^*)', \ldots, (\boldsymbol{\beta}_{K,\mathcal{B}_K}^*)')'$ as the maximizer of

$$\tilde{Q}_n(\boldsymbol{\theta}_{\mathcal{C}}) = \sum_{i=1}^n \log \left\{ \sum_{k=1}^K \mu_k g\left(y_i; h(\boldsymbol{X}_{i,\mathcal{A}_k}\boldsymbol{\alpha}_{k,\mathcal{A}_k} + \boldsymbol{Z}_{i,\mathcal{B}_k}\boldsymbol{\beta}_{k,\mathcal{B}_k}), \sigma_k\right) \right\}$$

$$-n\lambda_2 \sum_{k=1}^{K} \sum_{j\in\mathcal{A}_k} \sum_{l\in\mathcal{B}_k} c_{jl} \left(1 - e^{-\alpha_{kj}^2/\tau}\right) \left(1 - e^{-\beta_{kl}^2/\tau}\right).$$

Let $f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta}_{\mathcal{C}}) = \sum_{k=1}^{K} \mu_k g(y_i; h(\boldsymbol{X}_{i,\mathcal{A}_k}\boldsymbol{\alpha}_{k,\mathcal{A}_k} + \boldsymbol{Z}_{i,\mathcal{B}_k}\boldsymbol{\beta}_{k,\mathcal{B}_k}), \sigma_k)$, $c_0 = \max\{ |corr(X_j, Z_l)|,\ j \in \mathcal{A}_k, l \in \mathcal{B}_k, k = 1, \ldots, K\}$, with $corr(X_j, Z_l)$ being the correlation between X_j and Z_l, and $b_0 = \min\{\{|\alpha_{kj}^0|, j \in \mathcal{A}_k\}, \{|\beta_{kl}^0|,\ l \in \mathcal{B}_k\}, k = 1, \ldots, K\}$. We first establish the estimation consistency of $\boldsymbol{\theta}_{\mathcal{C}}^*$ when the true sparsity structure is known. Assume the following conditions:

(C1) The density function $f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta})$ has a common support, is identifiable in $\boldsymbol{\theta}$ up to the permutation of the component labels, and satisfies $E[\partial \log f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta})/\partial\theta_j]|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} = 0$, $E[(\partial \log f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta})/\partial\theta_j)(\partial \log f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta})/\partial\theta_l)] = E[-\partial^2 \log f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta})/\partial\theta_j\partial\theta_l]$.

(C2) The Fisher information matrix for $\boldsymbol{\theta}_{\mathcal{C}}$,

$$I(\boldsymbol{\theta}_{\mathcal{C}}) = E\left\{\left[\frac{\partial \log f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta}_{\mathcal{C}})}{\partial \boldsymbol{\theta}_{\mathcal{C}}}\right]\left[\frac{\partial \log f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta}_{\mathcal{C}})}{\partial \boldsymbol{\theta}_{\mathcal{C}}}\right]'\right\},$$

is finite and positive definite at $\boldsymbol{\theta}_{\mathcal{C}} = \boldsymbol{\theta}_{\mathcal{C}}^0$.

(C3) There exists an open set $\mathcal{N}_0$ that contains the true parameter $\boldsymbol{\theta}^0$, such that for almost all $V_i = (y_i, \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot})$, the density $f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta})$ admits all third derivatives for all $\boldsymbol{\theta} \in \mathcal{N}_0$. There exist two functions $M_1(V_i)$ and $M_2(V_i)$, for all $\boldsymbol{\theta} \in \mathcal{N}_0$, such that $\left|\partial^2 \log f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta})/\partial\theta_j\partial\theta_l\right| \leq M_1(V_i)$, $\left|\partial^3 \log f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta})/\partial\theta_j\partial\theta_l\partial\theta_m\right| \leq M_2(V_i)$, where $E[M_1(V_i)] < \infty$ and $E[M_2(V_i)] < \infty$.

(C4) For any constant $\epsilon > 0$, there exists a finite positive constant κ_1, such that for $j \in \mathcal{A}_k, l \in \mathcal{B}_k, k = 1, \ldots, K$, $P(|c_{jl}^{Pcorr} - corr(X_j, Z_l)| \geq \epsilon) \leq 2\exp(-n\epsilon^2/2\kappa_1)$. Moreover, $b_0^2 \geq \varrho\tau$ with $\varrho > 2$ and $\sqrt{n}\lambda_2 b_0 e^{-b_0^2/2\tau}/\tau = o(1)$, if $c_0 \geq c^{Pcorr}$.

Conditions (C1)–(C3) are commonly assumed in the literature (Khalili and Lin (2013); Hui, Warton and Foster (2015)). As suggested by Khalili and Chen (2007), the identifiability of FMR models generally depends on the component density $g(\cdot)$, maximum order K, and design matrix. We refer to the aforementioned publications for detailed discussions and sufficient conditions on identifiability. Condition (C4) restricts the rate of λ_2 when the maximum value of the absolute correlations between the important molecular and imaging variables under the true model is larger than the cutoff c^{Pcorr}. Condition (C4) also provides

a constraint on the error between the estimated sample correlations and the true population correlations.

Theorem 1. *Under Conditions* (C1)–(C4)*, there exists a strict local maximizer* $\boldsymbol{\theta}^*_{\mathcal{C}}$ *of* $\tilde{Q}_n(\boldsymbol{\theta}_{\mathcal{C}})$ *such that* $||\boldsymbol{\theta}^*_{\mathcal{C}} - \boldsymbol{\theta}^0_{\mathcal{C}}|| = O_p(\sqrt{s/n})$.

The proof is provided in Appendix. Theorem 1 shows that $\boldsymbol{\theta}^*_{\mathcal{C}}$ has the usual $O_p(\sqrt{s/n})$ convergence rate. Define $\hat{\boldsymbol{\theta}}$ with $\hat{\boldsymbol{\theta}}_{\mathcal{C}} = \boldsymbol{\theta}^*_{\mathcal{C}}$ and $\hat{\boldsymbol{\theta}}_{\mathcal{C}^c} = 0$. Next, we show that $\hat{\boldsymbol{\theta}}$ is a strict local maximizer of $Q(\boldsymbol{\theta})$ in (2.3). Assume the following additional conditions:

(C5) $b_0\lambda_1^{-1} \to \infty$, $(\lambda_1/(s/\sqrt{n})) \to \infty$, and $\lambda_1/(n^{a/2-1/2}\sqrt{\log n}) \to \infty, a \in (0, 1/2)$.

(C6) $\log(p) = O(n^a), \log(q) = O(n^a)$.

Condition (C5) puts constraints on the rate of λ_1, and similar conditions have been commonly assumed in high-dimensional studies (Fan and Lv (2011)). In particular, the first subcondition establishes the rate at which the nonzero coefficients can be distinguished from zero, and the other two restrict the rate of λ_1 with respect to the sample size. Condition (C6) allows the dimensionality p and q to grow exponentially fast.

Theorem 2. *Under Conditions* (C1)–(C6)*, with probability tending to one,* $\hat{\boldsymbol{\theta}}$ *is a strict local maximizer of* $Q(\boldsymbol{\theta})$.

The proof is provided in Appendix. Theorem 2 establishes the selection and estimation consistency under high-dimensional settings. This result shows that the proposed approach has consistency comparable to that of simpler models, although its objective and form are much more complicated.

2.3. Computation

We develop an expectation-maximization (EM) algorithm. First, for subject $i(= 1, \ldots, n)$, we introduce an unobserved indicator vector $\boldsymbol{\Delta}_i = (\Delta_{i1}, \ldots, \Delta_{iK})$, where $\Delta_{ik} = 1$ if subject i belongs to subgroup k, and $\Delta_{ik} = 0$ otherwise. The complete-data objective function is

$$Q_c(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_{ik} \log\left\{\mu_k g\left(y_i; h(\boldsymbol{X}_{i\cdot}\boldsymbol{\alpha}_k + \boldsymbol{Z}_{i\cdot}\boldsymbol{\beta}_k), \sigma_k\right)\right\} - n\sum_{k=1}^{K}\sum_{j=1}^{p}\rho(|\alpha_{kj}|; \gamma, \lambda_1)$$
$$-n\sum_{k=1}^{K}\sum_{l=1}^{q}\rho(|\beta_{kl}|; \gamma, \lambda_1) - n\lambda_2\sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{l=1}^{q} c_{jl}\left(1 - e^{-\alpha_{kj}^2/\tau}\right)\left(1 - e^{-\beta_{kl}^2/\tau}\right).$$

With fixed tuning parameters, the proposed algorithm proceeds as follows:

Initialization : Set $t = 0$. Initialize $\mu_k^{(0)} = 1/K$, for $k = 1, \ldots, K$, and randomly partition subjects into K subgroups with equal sizes. For each k, initialize $\boldsymbol{\alpha}_k^{(0)}$ and $\boldsymbol{\beta}_k^{(0)}$ using the MCP and $\sigma_k^{(0)}$ as the MLE.

E-step : Update $t = t + 1$. For $k = 1, \ldots, K$, and $i = 1, \ldots, n$, compute:

$$\delta_{ik}^{(t)} = \mathbb{E}_{\boldsymbol{\theta}^{(t-1)}}[\Delta_{ik}] = \frac{\mu_k^{(t)} g\left(y_i; h\left(\boldsymbol{X}_{i\cdot}\boldsymbol{\alpha}_k^{(t-1)} + \boldsymbol{Z}_{i\cdot}\boldsymbol{\beta}_k^{(t-1)}\right), \sigma_k^{(t-1)}\right)}{f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta}^{(t-1)})}.$$

M-step : Optimize $\mathbb{E}_{\boldsymbol{\theta}^{(t-1)}}[Q_c(\boldsymbol{\theta})]$ with respect to $\boldsymbol{\theta}$. For $k = 1, \ldots, K$, carry out the following steps sequentially:

(a) Compute $\mu_k^{(t)} = (1/n)\sum_{i=1}^n \delta_{ik}^{(t)}$.

(b) Optimize $Q_E(\sigma_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) = (1/n)\sum_{i=1}^n \delta_{ik}^{(t)} \log[g(y_i; h(\boldsymbol{X}_{i\cdot}\boldsymbol{\alpha}_k + \boldsymbol{Z}_{i\cdot}\boldsymbol{\beta}_k), \sigma_k)] - \sum_{j=1}^p \rho(|\alpha_{kj}|; \gamma, \lambda_1) - \sum_{l=1}^q \rho(|\beta_{kl}|; \gamma, \lambda_1) - \lambda_2 \sum_{j=1}^p \sum_{l=1}^q c_{jl}(1 - e^{-\alpha_{kj}^2/\tau})(1 - e^{-\beta_{kl}^2/\tau})$ with respect to σ_k, $\boldsymbol{\alpha}_k$, and $\boldsymbol{\beta}_k$. This varies with $h(\cdot)$ and $g(\cdot)$. Below, we take the Gaussian distribution $g(y_i; h(\boldsymbol{X}_{i\cdot}\boldsymbol{\alpha}_k + \boldsymbol{Z}_{i\cdot}\boldsymbol{\beta}_k), \sigma_k) = (1/\sqrt{2\pi})\sigma_k \exp[-(\sigma_k y_i - \boldsymbol{X}_{i\cdot}\boldsymbol{\alpha}_k - \boldsymbol{Z}_{i\cdot}\boldsymbol{\beta}_k)^2/2]$ as an example, and develop a coordinate descent (CD) algorithm. Algorithms for other distributions can be developed accordingly.

(b.1) With $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ fixed at $\boldsymbol{\alpha}_k^{(t-1)}$ and $\boldsymbol{\beta}_k^{(t-1)}$, optimize Q_E with respect to σ_k. Let $r_{ik}^{(t-1)} = \boldsymbol{X}_{i\cdot}\boldsymbol{\alpha}_k^{(t-1)} + \boldsymbol{Z}_{i\cdot}\boldsymbol{\beta}_k^{(t-1)}$, $\tilde{a}_k^{(t)} = \sum_{i=1}^n \delta_{ik}^{(t)} y_i^2$, and $\tilde{b}_k^{(t)} = \sum_{i=1}^n \delta_{ik}^{(t)} r_{ik}^{(t-1)} y_i$. Then, $\sigma_k^{(t)} = \tilde{b}_k^{(t)} + (\sqrt{(\tilde{b}_k^{(t)})^2 + 4n\tilde{a}_k^{(t)}\mu_k^{(t)}}/2\tilde{a}_k^{(t)})$.

(b.2) With σ_k and $\boldsymbol{\beta}_k$ fixed at $\sigma_k^{(t)}$ and $\boldsymbol{\beta}_k^{(t-1)}$, optimize Q_E with respect to $\boldsymbol{\alpha}_k$. For $j = 1, \ldots, p$, carry out the following steps sequentially. Compute $\eta_{kj}^{(t)} = (1/n)\sum_{i=1}^n \delta_{ik}^{(t)} x_{ij}^2$, $res_{-kj}^{(t)} = (1/n)\sum_{i=1}^n \delta_{ik}^{(t)}(\sigma_k^{(t)} y_i - r_{ik}^{(t-1)})x_{ij} + \eta_{kj}^{(t)}\alpha_{kj}^{(t-1)}$, and $u_{kj}^{(t)} = (2/\tau)e^{-\left(\alpha_{kj}^{(t-1)}\right)^2/\tau} \sum_{l=1}^q c_{jl}(1 - e^{-\left(\beta_{kl}^{(t-1)}\right)^2/\tau})$. Update

$$\alpha_{kj}^{(t)} = \begin{cases} \frac{res_{-kj}^{(t)}}{\eta_{kj}^{(t)} + \lambda_2 u_{kj}^{(t)}}, & \left|res_{-kj}^{(t)}\right| > \lambda_1\gamma\left(\eta_{kj}^{(t)} + \lambda_2 u_{kj}^{(t)}\right) \\ \frac{res_{-kj}^{(t)} - sgn\left(res_{-kj}^{(t)}\right)\lambda_1}{\eta_{kj}^{(t)} + \lambda_2 u_{kj}^{(t)} - 1/\gamma}, & \lambda_1 < \left|res_{-kj}^{(t)}\right| \le \lambda_1\gamma\left(\eta_{kj}^{(t)} + \lambda_2 u_{kj}^{(t)}\right) \\ 0, & \text{else} \end{cases}$$

and $r_{ik}^{(t-1)} = r_{ik}^{(t-1)} + x_{ij}\alpha_{kj}^{(t)} - x_{ij}\alpha_{kj}^{(t-1)}$.

(b.3) With σ_k and $\boldsymbol{\alpha}_k$ fixed at $\sigma_k^{(t)}$ and $\boldsymbol{\alpha}_k^{(t)}$, optimize Q_E with respect to $\boldsymbol{\beta}_k$. For $l = 1, \ldots, q$, carry out the following steps sequentially. Compute $\eta_{kl}^{(t)} = (1/n)\sum_{i=1}^n \delta_{ik}^{(t)} z_{il}^2$, $res_{-kl}^{(t)} = (1/n)\sum_{i=1}^n \delta_{ik}^{(t)}(\sigma_k^{(t)} y_i - r_{ik}^{(t-1)}) z_{il} + \eta_{kl}^{(t)} \beta_{kl}^{(t-1)}$, and $u_{kl}^{(t)} = (2/\tau) e^{-\left(\beta_{kl}^{(t-1)}\right)^2/\tau} \sum_{j=1}^p c_{jl}(1 - e^{-\left(\alpha_{kj}^{(t)}\right)^2/\tau})$. Update

$$\beta_{kl}^{(t)} = \begin{cases} \frac{res_{-kl}^{(t)}}{\eta_{kl}^{(t)} + \lambda_2 u_{kl}^{(t)}}, & \left|res_{-kl}^{(t)}\right| > \lambda_1 \gamma \left(\eta_{kl}^{(t)} + \lambda_2 u_{kl}^{(t)}\right) \\ \frac{res_{-kl}^{(t)} - sgn\left(res_{-kl}^{(t)}\right)\lambda_1}{\eta_{kl}^{(t)} + \lambda_2 u_{kl}^{(t)} - 1/\gamma}, & \lambda_1 < \left|res_{-kl}^{(t)}\right| \leq \lambda_1 \gamma \left(\eta_{kl}^{(t)} + \lambda_2 u_{kl}^{(t)}\right) \\ 0, & \text{else} \end{cases},$$

and $r_{ik}^{(t-1)} = r_{ik}^{(t-1)} + z_{il}\beta_{kl}^{(t)} - z_{il}\beta_{kl}^{(t-1)}$.

We iterate the E and M steps until convergence, which is concluded in our numerical study if $||\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}||_\infty < 10^{-4}$. In the literature, the convergence properties of the EM and CD algorithms are well established, and convergence is achieved in all of our numerical examples with a moderate number of iterations. To improve the performance, as in published studies, multiple random initializations of the subjects' subgroup memberships are considered, and the final estimator is chosen as the one with the smallest BIC.

The proposed approach involves a few parameters. We set τ in the L_0 penalty approximation as 0.01, and note that its value is not critical, as long as it is sufficiently small. We set the cutoff $c^{Pcorr} = 0.15$, which leads to satisfactory numerical results. For the regularization parameter γ in the MCP, following the literature (Zhang (2010)), we examine a few values, including 1.8, 3, 6, and 10, and find that $\gamma = 6$ has satisfactory performance (see Table S1 of the Supplementary Materials). The two tuning parameters λ_1 and λ_2 are selected using the BIC and a grid search, which is common practice.

To facilitate the data analysis and broad utilization, we provide R code and an example using The Cancer Genome Atlas (TCGA) lung cancer data. The code and example are available at `https://github.com/shuanggema/fmrGI`.

3. Simulation

Consider the following settings: (a) $n = 300, p = 1000, q = 500$, and $K = 2$. (b) $\boldsymbol{X}_{i\cdot}$ is generated from a multivariate normal distribution with marginal means zero and covariance matrix $\boldsymbol{\Sigma}$. Here, $\boldsymbol{\Sigma}$ has diagonal elements equal to one and a block-diagonal structure, with two blocks corresponding to the important and

unimportant variables, of which the sizes are p_0 and $p - p_0$, respectively. Detailed values of p_0 are provided in Table 5 (Appendix). Within each block, variables have an autoregressive (AR) correlation structure, where the jth and kth variables have correlation coefficient $\rho^{|j-k|}$, with $\rho = 0.3, 0.5$, and 0.7. (c) To describe the overlapping information between molecular and imaging variables, a set of 200 imaging variables $\mathcal{C}$ is generated using a linear regression model $\boldsymbol{Z}_{i\mathcal{C}} = \boldsymbol{X}_{i\mathcal{C}}\boldsymbol{\vartheta} + N(0, 0.01^2)$. Four settings of $\boldsymbol{\vartheta}$ are considered, where $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$ have 20 blocks with equal sizes, and $\boldsymbol{\vartheta}_3$ and $\boldsymbol{\vartheta}_4$ have 10 blocks with equal sizes. In each block, $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_3$ have all elements equal to one, and $\boldsymbol{\vartheta}_2$ and $\boldsymbol{\vartheta}_4$ have an AR structure with $\rho = 0.7$. The rest of the imaging features are generated similarly to $\boldsymbol{X}_{i\cdot}$ and independent of the molecular variables. (d) Three settings (P1, P2, and P3 in Table 5 of Appendix) of important variables are considered. In particular, we consider two subgroups with the same and different sets of important variables, with different settings. (e) We consider the continuous response computed from the FMR model, with $\sigma_k = 0.5$ and $\mu_1 = \mu_2 = 0.5$ (balanced) and $\mu_1 = 0.4, \mu_2 = 0.6$ (imbalanced). There are 72 scenarios, comprehensively covering a wide spectrum with different levels of within- and between-type correlations, as well as heterogeneity.

We consider the following alternatives. [FMR-MCP] analyzes the stacked data $(\boldsymbol{X}, \boldsymbol{Z})$ under the FMR model (2.1) with the MCP for regularized estimation and selection. This is the most direct competitor, and does not account for overlapping information. [Kmeans-MCP] first applies Kmeans to the residuals computed from an MCP-penalized linear regression model, with $(\boldsymbol{X}, \boldsymbol{Z})$ to identify subgroups, and then applies the MCP to each subgroup separately. This approach accommodates heterogeneity using the clustering technique, and there is no accounting for overlapping information. [CoRe] conducts collaborative regression (Gross and Tibshirani (2015)) that accommodates overlapping information and encourages $\boldsymbol{X}$ and $\boldsymbol{Z}$ to generate similar estimated effects. [DC-SVD] conducts a decomposition of $\boldsymbol{X}$ and $\boldsymbol{Z}$ using a singular value decomposition (SVD) to extract overlapping and independent information, and then conducts modeling (Zhu et al. (2016)). Both CoRe and DC-SVD are limited to the homogeneity case. [MCP-MI], [MCP-M], and [MCP-I] analyze $(\boldsymbol{X}, \boldsymbol{Z})$, $\boldsymbol{X}$, and $\boldsymbol{Z}$, respectively, using an MCP-penalized linear regression. We acknowledge that there are other potential alternatives. However, the above are likely the most relevant.

To get more intuition, we first simulate one dataset under AR(0.5), $\mu_1 = \mu_2 = 0.5$, P3, and $\boldsymbol{\vartheta}_2$. Beyond the proposed approach, we also consider its most direct competitor, FMR-MCP. The identification results are presented in Figure S1 (Supplementary Materials). For this specific dataset, both approaches correctly

identify the important variables, with FMR-MCP having more false positives. The molecular-imaging variable pairs identified using the proposed approach have weaker correlations (fewer connections), suggesting its effectiveness in promoting non-overlapping information.

To evaluate the identification performance, we adopt the true and false positive rates computed for the molecular (M:TPR and M:FPR) and imaging variables (I:TPR and I:FPR) separately. The estimation performance is evaluated using the root sum of squared errors (RSSE), defined as $||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0||$ and $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0||$ for molecular and imaging variables, respectively, where $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ and $(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0)$ are the estimated and true values of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, respectively. Note that, with the decomposition strategy, DC-SVD cannot generate the estimated values of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. For the proposed approach, FMR-MCP, and Kmeans-MCP, we also use classification accuracy (Accuracy) to evaluate the performance of the heterogeneity analysis. Moreover, an independent set with 100 subjects is generated, and the prediction median squared error (PMSE) is computed.

For each scenario, 500 replicates are simulated, and the medians and median absolute deviations (MADs) of the evaluation measures are summarized. The results for the scenarios with AR(0.5), $\mu_1 = \mu_2 = 0.5$, and $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$ are summarized in Tables 1 and 2. The rest of the results are provided in the Supplementary Materials. Across all simulation scenarios, the proposed approach has favorable performance. For example, in Table 1, under the scenario with correlation AR(0.5), balanced heterogeneity design, P1, and $\boldsymbol{\vartheta}_1$, the proposed approach identifies the majority of true positives and only a few false positives with (M:TPR, M:FPR, I:TPR, I:FPR)=(1.00, 0.02, 1.00, 0.03), compared to (0.70, 0.02, 0.70, 0.04) for FMR-MCP, (0.15, 0.05, 0.05, 0.03) for Kmeans-MCP, (0.30,0.02, 0.10, 0.02) for CoRe, (0.40, 0.02, 0.20, 0.01) for DC-SVD, (0.10, 0.05, 0.10, 0.03) for MCP-MI, (0.30, 0.14, –, –) for MCP-M, and (–, –, 0.00, 0.00) for MCP-I. It also performs better in terms of estimation with, for example, (M:RSSE, I:RSSE)=(0.67, 0.49) under the scenario with P2 and $\boldsymbol{\vartheta}_1$ in Table 1, compared to (1.65, 1.36), (8.48, 5.15), (3.64, 3.00), (–,–), (6.34, 4.00), (11.97,–), and (–, 2.86) for the alternatives. More satisfactory prediction accuracy is observed. Take the scenario with P2 and $\boldsymbol{\vartheta}_2$ in Table 2 as an example. The PMSE values are 1.76 (proposed), 2.81 (FMR-MCP), 8.78 (Kmeans-MCP), 10.52 (CoRe), 11.55 (DC-SVD), 19.00 (MCP-MI), 40.61 (MCP-M), and 13.78 (MCP-I). The proposed approach also outperforms FMR-MCP and Kmeans-MCP in the heterogeneity analysis. For example, under the scenario with P3 and $\boldsymbol{\vartheta}_2$ in Table 2, the Accuracy values are 0.85 (proposed), 0.68 (FMR-MCP), and 0.52 (Kmeans-MCP).

Table 1. Simulation results under the scenarios with AR(0.5), $\mu_1 = \mu_2 = 0.5$, and overlapping pattern $\boldsymbol{\vartheta}_1$ with 20 blocks. In each cell, median (MAD) based on 500 replicates.

Method	Accuracy	M:TPR	M:FPR	M:RSSE	I:TPR	I:FPR	I:RSSE	PMSE
				P1				
proposed	0.95(0.0)	1.00(0.0)	0.02(0.0)	0.63(0.2)	1.00(0.0)	0.03(0.0)	0.48(0.2)	1.82(1.2)
FMR-MCP	0.89(0.1)	0.70(0.4)	0.02(0.0)	2.03(2.6)	0.70(0.4)	0.04(0.0)	1.80(2.3)	7.84(10.7)
Kmeans-MCP	0.52(0.0)	0.15(0.1)	0.05(0.0)	9.14(0.5)	0.05(0.1)	0.03(0.0)	5.48(0.7)	10.38(3.0)
CoRe	--	0.30(0.1)	0.02(0.0)	3.80(0.0)	0.10(0.1)	0.02(0.0)	3.00(0.0)	10.75(2.2)
DC-SVD	--	0.40(0.1)	0.02(0.0)	--	0.20(0.1)	0.01(0.0)	--	12.06(3.7)
MCP-MI	--	0.10(0.1)	0.05(0.0)	6.54(0.5)	0.10(0.1)	0.03(0.0)	4.10(0.5)	19.94(5.8)
MCP-M	--	0.30(0.1)	0.14(0.0)	12.68(1.0)	--	--	--	42.00(9.8)
MCP-I	--	--	--	--	0.00(0.0)	0.00(0.0)	3.00(0.0)	10.50(2.4)
				P2				
proposed	0.94(0.0)	1.00(0.0)	0.02(0.0)	0.67(0.3)	1.00(0.0)	0.03(0.0)	0.49(0.2)	1.57(1.1)
FMR-MCP	0.91(0.1)	0.80(0.3)	0.02(0.0)	1.65(1.9)	0.85(0.2)	0.04(0.0)	1.36(1.6)	4.55(5.7)
Kmeans-MCP	0.52(0.0)	0.15(0.1)	0.04(0.0)	8.48(0.6)	0.10(0.1)	0.03(0.0)	5.15(0.7)	8.99(2.4)
CoRe	--	0.28(0.1)	0.02(0.0)	3.64(0.1)	0.15(0.1)	0.02(0.0)	3.00(0.0)	10.67(2.3)
DC-SVD	--	0.35(0.1)	0.02(0.0)	--	0.15(0.1)	0.02(0.0)	--	11.40(3.2)
MCP-MI	--	0.15(0.1)	0.05(0.0)	6.34(0.3)	0.10(0.1)	0.03(0.0)	4.00(0.4)	18.15(3.9)
MCP-M	--	0.25(0.1)	0.12(0.0)	11.97(0.8)	--	--	--	40.61(9.8)
MCP-I	--	--	--	--	0.05(0.1)	0.00(0.0)	2.86(0.1)	10.67(2.3)
				P3				
proposed	0.88(0.1)	0.75(0.2)	0.02(0.0)	1.92(1.0)	0.70(0.3)	0.04(0.0)	1.59(0.9)	4.83(3.6)
FMR-MCP	0.69(0.2)	0.40(0.2)	0.04(0.0)	4.15(3.5)	0.25(0.2)	0.05(0.0)	2.96(1.6)	12.34(10.6)
Kmeans-MCP	0.52(0.0)	0.15(0.1)	0.04(0.0)	8.52(0.6)	0.10(0.1)	0.03(0.0)	5.11(0.5)	8.84(2.6)
CoRe	--	0.50(0.1)	0.01(0.0)	3.73(0.0)	0.20(0.1)	0.02(0.0)	3.00(0.0)	7.97(2.2)
DC-SVD	--	0.50(0.1)	0.02(0.0)	--	0.30(0.1)	0.01(0.0)	--	9.25(3.0)
MCP-MI	--	0.20(0.1)	0.04(0.0)	6.29(0.5)	0.10(0.1)	0.02(0.0)	4.01(0.4)	16.32(3.2)
MCP-M	--	0.30(0.1)	0.13(0.0)	11.57(0.9)	--	--	--	33.72(8.2)
MCP-I	--	--	--	--	0.10(0.1)	0.00(0.0)	3.00(0.1)	7.83(2.2)

Overall, the proposed approach exhibits better performance with a moderate within correlation AR(0.5). Compared to settings P1 and P2, which have a higher level of heterogeneity, under P3, the performance of the proposed approach and FMR-MCP decays. The two homogeneity-based alternatives CoRe and DC-SVD, which accommodate overlapping information, have improved performance. However, the proposed approach remains superior. The superiority of the proposed approach over FMR-MCP and Kmeans-MCP provides direct support for the L_0-based penalty for accommodating overlapping information. The improvement over CoRe and DC-SVD suggests the necessity of accounting for heterogeneity. The proposed approach performs much better than MCP-MI, MCP-M, and MCP-I, re-establishing the value of data integration.

We conduct additional simulations under setting AR(0.5) for within-block

Table 2. Simulation results under the scenarios with AR(0.5), $\mu_1 = \mu_2 = 0.5$, and overlapping pattern $\boldsymbol{\vartheta}_2$ with 20 blocks. In each cell, median (MAD) based on 500 replicates.

Method	Accuracy	M:TPR	M:FPR	M:RSSE	I:TPR	I:FPR	I:RSSE	PMSE
				P1				
proposed	0.95(0.0)	1.00(0.0)	0.01(0.0)	0.66(0.3)	1.00(0.0)	0.04(0.0)	0.50(0.3)	1.80(1.5)
FMR-MCP	0.93(0.1)	0.90(0.1)	0.02(0.0)	1.06(1.1)	0.90(0.1)	0.04(0.0)	1.03(1.1)	3.14(3.5)
Kmeans-MCP	0.52(0.0)	0.10(0.1)	0.05(0.0)	9.29(0.6)	0.05(0.1)	0.03(0.0)	5.43(0.6)	10.41(2.8)
CoRe	--	0.30(0.1)	0.02(0.0)	3.80(0.0)	0.10(0.1)	0.02(0.0)	3.00(0.0)	10.75(2.4)
DC-SVD	--	0.40(0.1)	0.02(0.0)	--	0.20(0.1)	0.02(0.0)	--	12.68(4.7)
MCP-MI	--	0.10(0.0)	0.04(0.0)	6.43(0.5)	0.10(0.0)	0.03(0.0)	4.13(0.5)	20.48(6.2)
MCP-M	--	0.30(0.1)	0.14(0.0)	12.68(1.0)	--	--	--	42.00(9.8)
MCP-I	--	--	--	--	0.40(0.3)	0.36(0.1)	17.83(6.1)	58.34(60.6)
				P2				
proposed	0.94(0.0)	1.00(0.0)	0.01(0.0)	0.62(0.2)	1.00(0.0)	0.04(0.0)	0.49(0.2)	1.76(1.3)
FMR-MCP	0.92(0.1)	0.85(0.2)	0.02(0.0)	1.36(1.6)	0.80(0.3)	0.04(0.0)	1.26(1.4)	2.80(3.4)
Kmeans-MCP	0.52(0.0)	0.10(0.1)	0.04(0.0)	8.53(0.5)	0.10(0.1)	0.03(0.0)	5.33(0.6)	8.78(2.8)
CoRe	--	0.30(0.1)	0.02(0.0)	3.65(0.1)	0.15(0.1)	0.02(0.0)	3.00(0.0)	10.52(2.5)
DC-SVD	--	0.35(0.1)	0.02(0.0)	--	0.20(0.1)	0.02(0.0)	--	11.55(3.3)
MCP-MI	--	0.15(0.1)	0.04(0.0)	6.43(0.5)	0.10(0.1)	0.03(0.0)	4.16(0.4)	19.00(5.5)
MCP-M	--	0.25(0.1)	0.12(0.0)	11.97(0.8)	--	--	--	40.61(9.8)
MCP-I	--	--	--	--	0.10(0.1)	0.01(0.0)	3.05(0.4)	13.78(9.1)
				P3				
proposed	0.85(0.1)	0.70(0.3)	0.02(0.0)	2.18(1.4)	0.65(0.4)	0.05(0.0)	1.65(1.0)	5.70(4.5)
FMR-MCP	0.68(0.2)	0.40(0.3)	0.04(0.0)	4.85(3.5)	0.25(0.3)	0.06(0.0)	3.05(1.6)	11.56(10.8)
Kmeans-MCP	0.52(0.0)	0.15(0.1)	0.04(0.0)	8.57(0.6)	0.10(0.1)	0.03(0.0)	4.99(0.6)	8.49(2.3)
CoRe	--	0.50(0.1)	0.02(0.0)	3.73(0.0)	0.20(0.1)	0.02(0.0)	3.00(0.0)	7.95(2.3)
DC-SVD	--	0.50(0.1)	0.02(0.0)	--	0.30(0.1)	0.01(0.0)	--	9.23(3.2)
MCP-MI	--	0.20(0.1)	0.04(0.0)	6.13(0.4)	0.10(0.1)	0.03(0.0)	4.05(0.5)	17.15(4.3)
MCP-M	--	0.30(0.1)	0.13(0.0)	11.57(0.9)	--	--	--	33.72(8.2)
MCP-I	--	--	--	--	0.40(0.3)	0.35(0.1)	16.63(5.6)	49.28(54.1)

correlation, and settings $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$ for the overlapping pattern. First, we consider two additional settings (P4 and P5 in Table 5 of Appendix) of important variables. Specifically, P4 has different important variables for the two subgroups, which may closely mimic the real data example (Tables 3 and 4). P5 is a more homogeneous case, where more than half of the important variables have the same effects for the two subgroups, and the remaining effects have different magnitudes, but the same directions. Second, a more imbalanced heterogeneity design with $\mu_1 = 0.1$ and $\mu_2 = 0.9$ is considered. Summary results are presented in Tables S13–S21 of the Supplementary Materials, where for the highly imbalanced heterogeneity scenarios, we also provide the sensitivity and specificity results of the heterogeneity analysis and consider the two subgroups separately. Patterns similar to those described above are observed. Specifically, under the

most homogeneous setting P5 (Table S14 of the Supplementary Materials), the proposed approach behaves slightly worse than DC-SVD, as expected, but still better than the other alternatives. Under the highly imbalanced setting with $\mu_1 = 0.1$ and $\mu_2 = 0.9$, the proposed approach still performs well in identifying the two subgroups, with high accuracy. In addition, it has satisfactory identification and estimation performance for the second subgroup, but worse performance for the first subgroup, which has a very limited sample size, compared to the homogeneity-based alternatives. Because the two subgroups share the same important variables under settings P1, P3, and P5, it is not surprising that the homogeneity-based alternatives behave well by considering the two subgroups together.

A closer look at the number of mixture components. In the above simulations, we assume that $K = 2$ is known, as in published studies such as Khalili and Chen (2007), Hui, Warton and Foster (2015), and Liu et al. (2020). We propose adopting the BIC when the value of K needs to be estimated. Here, we additionally take the scenarios with settings AR(0.5) and $\boldsymbol{\vartheta}_1$ as an example, and examine the performance of the BIC for selecting K. Specifically, with candidate $K = 1, 2, 3, 4, 5$, we simulate 500 replicates, compute the frequency that a particular value is selected, and report the results in Table S22 (Supplementary Materials). In general, the BIC has satisfactory performance. The setting P3 has a higher degree of homogeneity compared to P1 and P2, making it more difficult to identify the true value of $K = 2$. We report the summary identification and estimation results of the proposed approach with BIC-selected K in Table S23 (Supplementary Materials), where the true and estimated subgroups are matched by minimizing the RSSE when K is overestimated (Khalili and Lin (2013)). The observed patterns are similar to those in Table 1 and Table S3 (Supplementary Materials), suggesting that estimating K does not significantly affect the performance of the proposed approach.

Computer time. We examine the computer time under the above simulation settings and $n = 300$, $p = 1000$, and $q = 500$. With fixed tuning parameters and one initialization, the average computer time of the proposed analysis is 17.09 seconds, using a laptop with standard configurations, compared to 16.58, 1.05, 0.18, 5.69, 0.81, 0.73, and 0.71 seconds for FMR-MCP, Kmeans-MCP, CoRe, DC-SVD, MCP-MI, MCP-M, and MCP-I, respectively. With more complex analysis, the proposed approach has a higher computational cost, but is still affordable.

4. Data Analysis

TCGA is a hallmark genomics program organized by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), and has significantly advanced cancer research in multiple aspects. It has published high-quality outcome/phenotype, clinical, molecular, and histopathological imaging data. There have been both unsupervised (Li et al. (2018)) and supervised (Ahmad and Fröhlich (2017)) heterogeneity studies conducted using TCGA data.

In our analysis, we combine data on lung adenocarcinoma (LUAD) and lung squamous cell (LUSC), two major subtypes of non-small-cell lung cancer (NSCLC), to increase the sample size. We acknowledge their differences. However, as the proposed analysis is designed to accommodate heterogeneity, this does not pose a problem. The response variable is the reference value for the pre-bronchodilator forced expiratory volume in one second in percent (FEV1), which is an important biomarker for lung capacity and tightly associated with prognosis and other outcomes. For the molecular variables, we analyze mRNA gene expressions downloaded from cBioPortal. For the imaging variables, we adopt a recently developed data extraction and processing pipeline (Zhong, Wu and Ma (2019)). Briefly, we first download the diagnostic slides using the GDC Data Transfer Tool from the TCGA website, and then extract high-dimensional imaging variables using CellProfiler. These imaging features represent objective attributes of histopathological images, including the area and perimeter of the nucleus and cytoplasm, mean and standard deviation of these measures, and other general image attributes. After subject matching, a total of 370 subjects with 20,440 gene expression measurements and 221 imaging features are available. Brief information is provided in Figure S2 (Supplementary Materials). Our preliminary exploration suggests that if the dimensions of the two types of data differ significantly, performance may be inferior. In addition, the number of lung capacity-related genes is not expected to be large. As such, we conduct a marginal screening, and the top 500 genes with the smallest p-values computed from a marginal linear regression are selected for downstream analysis.

In Figure S3 (Supplementary Materials), we show the histogram and estimated density of FEV1. We observe two modes, which may reasonably suggest heterogeneity with two subgroups. FEV1 has also been examined in the literature (Liu et al. (2020)), which suggests continuous and close-to-normal distributions. As such, we model it using a mixture of two normal distributions. Note that such exploratory analysis based on a histogram has previously been conducted in the literature (Khalili and Chen (2007)). To be cautious, we have also conducted

Table 3. Data analysis using the proposed approach: identified genes and estimates.

Gene	α_1	α_2	Gene	α_1	α_2	Gene	α_1	α_2	Gene	α_1	α_2
EIF4A3	-0.58		ECHDC2		-0.19	RFC5	0.03		NFU1	-0.16	
DHX36	0.21		N4BP1		0.12	POLR3C	0.37		FAM160A2	-0.26	
KIAA0141		0.25	FAM210A		-0.04	PSMD12		-0.08	PCDHB4	0.17	
WDR43		-0.07	SST	0.15		RPL23AP53	0.35		CD81	0.16	
CNPPD1	0.49		ZNF596		0.12	PSME4		0.01	MAK16	-0.34	
METTL5	-0.22	0.09	RNF115		-0.16	NCBP1	0.20		GCSH	0.08	
DCUN1D1	-0.10		FBXO28	-0.49		LRRC31	0.13		GCFC2		-0.07
DBR1	-0.20		GPN1		0.06	RHBDF1		0.08	L2HGDH	-0.11	
B9D2		0.09	ADSL	0.07		TCTEX1D2		-0.09	DTX2	0.38	
DNAJB4	-0.28		IGIP	-0.22		RAD51		0.32	RGL2	0.27	
RPSAP58		-0.10	MEGF6		0.06	BIN3	-0.07		OR6C6	-0.15	
CNKSR1	0.11		IL22RA2	0.10		RSL24D1		-0.02	KIAA1109	-0.20	
CCT5		-0.04	TCF25	-0.15		DYNC1I2	-0.25		TMEM50B		0.10
IRX2	0.42		METTL21C		0.18	EPT1	0.29		TRAPPC10		-0.10
CTNNAL1		-0.14	CENPO	0.24		SCNN1D		0.03	CSNK2A1		0.01
TRMT61B		0.10	C1ORF112		-0.03	DEFB4A	0.16		PDLIM2		-0.02
MRPL3		0.35	CCDC92		0.03	ZIC1		-0.04	ZC3H6	0.01	
ASTN1		-0.05	UGT1A7	-0.09		MAP3K6	-0.14	0.12	KIAA1715		-0.06
CD1E	-0.16		EML3		-0.05	ZNF487		0.23	RPA3	-0.12	
TOMM5	-0.10		HMGXB4		-0.07	PRDM9		-0.17	KCNK18	-0.28	
FAM86JP		-0.17	RNF168		0.10	CDC73		0.11	CAMTA2		0.05
LDLRAD2		-0.03	RNPEPL1		-0.08	PAK1		-0.01	YPEL3	-0.19	
IARS		0.04	POLR2D		-0.04	GPATCH3		0.08	SNX5		-0.10

analyses with 3, 4, and 5 mixture components. However, the results are not as sensible, with the extra components having very small numbers of important variables and/or small mixture probabilities. With two mixture components, the proposed approach identifies 92 genes and 24 imaging variables, and the detailed information is provided in Tables 3 and 4. Almost all of the important variables contribute to the response in only one subgroup, except for genes METTL5 and MAP3K6, which have effects in both subgroups, but with different signs. More information on the identified gene expressions and imaging features is provided graphically in Figure S4 (Supplementary Materials). For the two subgroups separately, there are only 49 and 43, respectively, gene-imaging variable pairs with absolute correlations larger than 0.1, which again shows the proposed approach's effectiveness in identifying non-overlapping information.

A literature search suggests that many of the identified genes show strong evidence of being associated with lung capacity and cancer. More details are provided in the Supplementary Materials. We also examine the 24 identified imaging features more closely. These features measure tissue area shape, texture,

Table 4. Data analysis using the proposed approach: identified imaging features and estimates.

Group	Imaging Feature Name	Abbreviation	β_1	β_2
Geometry	AreaShape_Center_X	ASCX		-0.022
Geometry	AreaShape_EulerNumber	ASEN		-0.003
Geometry	AreaShape_Zernike_5_3	ASZ53		-0.212
Geometry	AreaShape_Zernike_5_5	ASZ55	-0.012	
Geometry	AreaShape_Zernike_7_1	ASZ71	-0.202	
Geometry	AreaShape_Zernike_7_3	ASZ73		0.209
Geometry	AreaShape_Zernike_8_0	ASZ80	-0.001	
Holistic	Count_Identifyeosinprimarycytoplasm	CIPC	0.113	
Texture	Granularity_10_ImageAfterMath	G10M	0.220	
Texture	Granularity_11_ImageAfterMath.1	G11M1	-0.200	
Texture	Granularity_12_ImageAfterMath.1	G12M1	-0.168	
Texture	Granularity_13_ImageAfterMath	G13M		0.005
Texture	Granularity_13_ImageAfterMath.1	G13M1	0.111	
Texture	Granularity_9_ImageAfterMath	G9M	0.005	
Texture	Granularity_9_ImageAfterMath.1	G9M1		0.041
Geometry	Location_Center_Y.1	LCY1		-0.066
Texture	Texture_Correlation_ImageAfterMath_3_00	TCM300		0.243
Texture	Texture_Correlation_ImageAfterMath_3_01	TCM301	-0.329	
Texture	Texture_Correlation_ImageAfterMath_3_02	TCM302		0.066
Texture	Texture_DifferenceEntropy_maskosingray_3_02	TDM302	0.280	
Texture	Texture_DifferenceVariance_maskosingray_3_02	TDVM302		-0.182
Texture	Texture_Entropy_ImageAfterMath_3_03	TEM303		-0.015
Holistic	Threshold_WeightedVariance_Identifyeosinprimarycytoplasm	TWPC		-0.132
Holistic	Threshold_WeightedVariance_identifyhemaprimarynuclei	TWPN		0.002

nuclear, and cytoplasm parameters. In particular, 13 are texture related. Similar findings have been made in previous studies (Luo et al. (2017)). However, our literature review suggests that the biological implications of high-dimensional imaging features are not well understood. As such, interpretation is not pursued further.

Analysis is also conducted using the alternatives. Summary comparison results are presented in Table S24 (Supplementary Materials), including the numbers of genes and imaging variables identified by the different approaches, and their overlaps and RV coefficients. The RV coefficient measures the similarity between two data matrices, with a larger value indicating higher similarity. The various approaches identify different sets of features with moderate overlapping, as suggested by the RV coefficients. To provide additional support to the analysis, we evaluate the prediction performance and selection stability of the proposed approach and the alternatives. Specifically, we conduct 100 random splits to gen-

erate training and testing data. Estimation is conducted using the training data, prediction is made on the testing data, and the median values of the PMSE and Pearson's correlation (COR, between the estimated and observed values of FEV1) are computed. The proposed approach has (PMSE,COR)=(0.32,0.49), compared to (0.34,0.48) for FMR-MCP, (0.55,0.13) for Kmeans-MCP, (0.33,0.15) for CoRe, (0.32,0.19) for DC-SVD, (1.02,0.15) for MCP-MI, (0.46,0.12) for MCP-M, and (0.34,0.15) for MCP-I. When using the observed occurrence index (OOI), which is the probability of being identified in multiple splits, to evaluate stability, the proposed approach has a mean OOI value for the identified genes and imaging features of 0.265, compared to 0.254, 0.181, 0.152, 0.132, 0.104, 0.098, and 0.200 for the alternatives. The improved prediction and stability performance supports the proposed analysis to a certain extent.

5. Discussion

Heterogeneity analysis is a "classic", yet still highly important topic in cancer research. In this article, we have advanced cancer heterogeneity analysis by integrating molecular and histopathological imaging features. We have adopted penalization for regularized estimation and selection, and, equally importantly, the promotion of non-overlapping information. The proposed analysis and approach are biologically well motivated and intuitive. Theoretical investigation, simulation, and data analysis have demonstrated satisfactory performance. Overall, this study can enrich the family of cancer analytics and suggest a new data integration direction for development. Furthermore, the proposed analysis can be applied to a wide variety of data types, models, and molecular and other measurements.

The proposed approach can be extended further to accommodate more than two types of covariates. This will require revising the last penalty term to include all pairs (of covariate types), and the extension of the other steps will be mostly straightforward. In computation, we have adopted multiple random initializations, and chosen the final estimator as the one with the smallest BIC. Other initialization techniques can also be adopted. In particular, our brief exploration has suggested that the Kmeans initializations lead to similar results. This is a "classic" problem, and we have chosen not to reiterate the literature. The adopted FMR technique can reveal important differences across subgroups in modeling a response. However, as has been noted in the literature, such differences may or may not be associated with disease subtypes or other clinical characteristics. In our data analysis, although FEV1 is an important biomarker for prognosis

and other outcomes, it is still unclear what other clinical significance the identified heterogeneity and models have. Being beyond our scope, this aspect is not pursued further. It will also be of interest to explore other and more complex measures of overlapping information. For methodological development, we have focused on molecular and imaging variables. It will be of interest to expand the scope of the analysis to include clinical/demographic and other variables.

Supplementary Material

The Supplementary Material include additional simulation and data analysis results referenced in Sections 2, 3, and 4.

Acknowledgments

We thank the editors and reviewers for their careful review and insightful comments. This work was supported by the NIH [CA204120, CA241699, CA196530]; NSF [1916251]; Pilot Award from Yale Cancer Center; Bureau of Statistics of China [2018LD02]; "Chenguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission [18CG42]; Program for Innovative Research Team of Shanghai University of Finance and Economics; Shanghai Pujiang Program [19PJ1403600]; National Natural Science Foundation of China [12071273, 11971404]; Humanity and Social Science Youth Foundation of Ministry of Education of China [19YJC910010]; and 111 Project [B13028].

Appendix

Proof of Theorem 1

Let $\delta_n = \sqrt{s/n}$ and $\boldsymbol{w} = (\boldsymbol{g}'_{K\times 1}, \boldsymbol{t}'_{K\times 1}, \boldsymbol{u}'_{1_{|\mathcal{A}_1|\times 1}}, \ldots, \boldsymbol{u}'_{K_{|\mathcal{A}_K|\times 1}}, \boldsymbol{v}'_{1_{|\mathcal{B}_1|\times 1}}, \ldots, \boldsymbol{v}'_{K_{|\mathcal{B}_K|\times 1}})'$. To prove Theorem 1, it suffices to show that under Conditions (C1)-(C4), $\tilde{Q}_n(\boldsymbol{\theta}_{\mathcal{C}}) < \tilde{Q}_n(\boldsymbol{\theta}^0_{\mathcal{C}})$ on the boundary of set $\{\boldsymbol{\theta}_{\mathcal{C}} : ||\boldsymbol{\theta}_{\mathcal{C}} - \boldsymbol{\theta}^0_{\mathcal{C}}|| \le C\delta_n\}$, where C is a sufficiently large positive constant. It is equivalent to show that $\tilde{Q}_n(\boldsymbol{\theta}^0_{\mathcal{C}} + \delta_n \boldsymbol{w}) - \tilde{Q}_n(\boldsymbol{\theta}^0_{\mathcal{C}})$ is strictly negative everywhere on the boundary $\{\boldsymbol{w} : ||\boldsymbol{w}|| = C\}$.

Let $L_n(\boldsymbol{\theta}_{\mathcal{C}}) = \sum_{i=1}^n l_i(\boldsymbol{\theta}_{\mathcal{C}})$ with $l_i(\boldsymbol{\theta}_{\mathcal{C}}) = \log f(y_i; \boldsymbol{X}_{i\cdot}, \boldsymbol{Z}_{i\cdot}, \boldsymbol{\theta}_{\mathcal{C}})$. Then

$$\begin{aligned} D_n(\boldsymbol{w}) &= \tilde{Q}_n(\boldsymbol{\theta}^0_{\mathcal{C}} + \delta_n \boldsymbol{w}) - \tilde{Q}_n(\boldsymbol{\theta}^0_{\mathcal{C}}) \\ &= L_n(\boldsymbol{\theta}^0_{\mathcal{C}} + \delta_n \boldsymbol{w}) - L_n(\boldsymbol{\theta}^0_{\mathcal{C}}) \\ &\quad - n\lambda_2 \sum_{k=1}^K \sum_{j\in\mathcal{A}_k} \sum_{l\in\mathcal{B}_k} c_{jl} \left(1 - e^{-\left(\alpha^0_{kj} + \delta_n u_{kl}\right)^2/\tau}\right) \left(1 - e^{-(\beta^0_{kl} + \delta_n v_{kl})^2/\tau}\right) \end{aligned}$$

$$+n\lambda_2 \sum_{k=1}^{K} \sum_{j\in\mathcal{A}_k} \sum_{l\in\mathcal{B}_k} c_{jl} \left(1 - e^{-\left(\alpha_{kj}^0\right)^2/\tau}\right) \left(1 - e^{-(\beta_{kl}^0)^2/\tau}\right)$$
$$= L_n(\boldsymbol{\theta}_{\mathcal{C}}^0 + \delta_n \boldsymbol{w}) - L_n(\boldsymbol{\theta}_{\mathcal{C}}^0) + IV.$$

We have

$$L_n(\boldsymbol{\theta}_{\mathcal{C}}^0 + \delta_n \boldsymbol{w}) - L_n(\boldsymbol{\theta}_{\mathcal{C}}^0) = \delta_n \boldsymbol{w}' \left(\frac{\partial L_n(\boldsymbol{\theta}_{\mathcal{C}})}{\partial \boldsymbol{\theta}_{\mathcal{C}}}\bigg|_{\boldsymbol{\theta}_{\mathcal{C}}^0} \right) + \frac{1}{2}\delta_n^2 \boldsymbol{w}' \left(\frac{\partial^2 L_n(\boldsymbol{\theta}_{\mathcal{C}})}{\partial^2 \boldsymbol{\theta}_{\mathcal{C}}}\bigg|_{\boldsymbol{\theta}_{\mathcal{C}}^0} \right) \boldsymbol{w}$$
$$+ \frac{\delta_n^3}{6} \sum_{j,l,m\in\mathcal{C}} \frac{\partial^3 L_n(\boldsymbol{\theta}_{\mathcal{C}})}{\partial\theta_j \partial\theta_l \partial\theta_m}\bigg|_{\tilde{\boldsymbol{\theta}}_{\mathcal{C}}} w_j w_l w_m$$
$$= I + II + III,$$

where $\tilde{\boldsymbol{\theta}}_{\mathcal{C}}$ lies on the line segment connecting $\boldsymbol{\theta}_{\mathcal{C}}^0 + \delta_n \boldsymbol{w}$ and $\boldsymbol{\theta}_{\mathcal{C}}^0$. With Condition (C1), we have

$$\frac{1}{\sqrt{n}} \frac{\partial L_n(\boldsymbol{\theta}_{\mathcal{C}})}{\partial \boldsymbol{\theta}_{\mathcal{C}}}\bigg|_{\boldsymbol{\theta}_{\mathcal{C}}^0} = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\theta}_{\mathcal{C}})}{\partial \boldsymbol{\theta}_{\mathcal{C}}}\bigg|_{\boldsymbol{\theta}_{\mathcal{C}}^0} \right)$$
$$= \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\theta}_{\mathcal{C}})}{\partial \boldsymbol{\theta}_{\mathcal{C}}}\bigg|_{\boldsymbol{\theta}_{\mathcal{C}}^0} - E\left[\frac{\partial l(\boldsymbol{\theta}_{\mathcal{C}})}{\partial \boldsymbol{\theta}_{\mathcal{C}}} \right]\bigg|_{\boldsymbol{\theta}_{\mathcal{C}}^0} \right) \to N(0, \tilde{\Sigma}).$$

Thus $\partial L_n(\boldsymbol{\theta}_{\mathcal{C}})/\partial \boldsymbol{\theta}_{\mathcal{C}}|_{\boldsymbol{\theta}_{\mathcal{C}}^0} = O_P(\sqrt{n}) = O_p(\sqrt{ns})$. Then,

$$|I| \le O_P(n\delta_n^2)||\boldsymbol{w}||.$$

For II,

$$II = -\frac{1}{2} n\delta_n^2 \boldsymbol{w}' I(\boldsymbol{\theta}_{\mathcal{C}}^0) \boldsymbol{w} + \frac{1}{2} n\delta_n^2 \boldsymbol{w}' \left(\frac{1}{n} \frac{\partial^2 L_n(\boldsymbol{\theta}_{\mathcal{C}})}{\partial^2 \boldsymbol{\theta}_{\mathcal{C}}}\bigg|_{\boldsymbol{\theta}_{\mathcal{C}}^0} + I(\boldsymbol{\theta}_{\mathcal{C}}^0) \right) \boldsymbol{w}.$$

Following Lemma 8 in Fan and Peng (2004), with Conditions (C1) and (C2), we have

$$\left\| \frac{1}{n} \frac{\partial^2 L_n(\boldsymbol{\theta}_{\mathcal{C}})}{\partial^2 \boldsymbol{\theta}_{\mathcal{C}}}\bigg|_{\boldsymbol{\theta}_{\mathcal{C}}^0} + I(\boldsymbol{\theta}_{\mathcal{C}}^0) \right\| = o_p\left(\frac{1}{s}\right).$$

Therefore,

$$II = -\frac{1}{2} n\delta_n^2 \boldsymbol{w}' I(\boldsymbol{\theta}_{\mathcal{C}}^0) \boldsymbol{w} + \frac{1}{2} n\delta_n^2 ||\boldsymbol{w}||^2 \times o_p(1).$$

For III, by Condition (C3) and the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
|III| &= \frac{\delta_n^3}{6}\left|\sum_{j,l,m\in\mathcal{C}} \frac{\partial^3 L_n(\boldsymbol{\theta}_{\mathcal{C}})}{\partial\theta_j\partial\theta_l\partial\theta_m}\bigg|_{\tilde{\boldsymbol{\theta}}_{\mathcal{C}}} w_j w_l w_m\right| \\
&= \frac{\delta_n^3}{6}\left|\sum_{j,l,m\in\mathcal{C}}\sum_{i=1}^{n} \frac{\partial^3 \log f(y_i;\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot},\boldsymbol{\theta}_{\mathcal{C}})}{\partial\theta_j\partial\theta_l\partial\theta_m}\bigg|_{\tilde{\boldsymbol{\theta}}_{\mathcal{C}}} w_j w_l w_m\right| \\
&\le \frac{\delta_n^3}{6}\sum_{i=1}^{n}\left[\sum_{j,l,m\in\mathcal{C}} M_2^2(V_i)\right]^{1/2} ||\boldsymbol{w}||^3 = O_p(s^{3/2}\delta_n)\times n\delta^2\times||\boldsymbol{w}||^2.
\end{aligned}$$

Since $s \ll n$, we have

$$III = o_p(n\delta_n^2)||\boldsymbol{w}||^2.$$

Moreover,

$$\begin{aligned}
|IV| = n\lambda_2\Bigg|\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k}\sum_{l\in\mathcal{B}_k}\Big\{c_{jl}\Big[&\Big(e^{-(\alpha_{kj}^0+\delta_n u_{kj})^2/\tau} - e^{-(\alpha_{kj}^0)^2/\tau}\Big) \\
&+\Big(e^{-(\beta_{kl}^0+\delta_n v_{kl})^2/\tau} - e^{-(\beta_{kl}^0)^2/\tau}\Big) \\
&-\Big(e^{-((\alpha_{kj}^0+\delta_n u_{kj})^2+(\beta_{kl}^0+\delta_n v_{kl})^2)/\tau} - e^{-((\alpha_{kj}^0)^2+(\beta_{kl}^0)^2)/\tau}\Big)\Big]\Big\}\Bigg|.
\end{aligned}$$

Let $c_0 = \max\{|corr(X_j, Z_l)|, j \in \mathcal{A}_k, l \in \mathcal{B}_k, k = 1,\ldots,K\}$ with $corr(X_j, Z_l)$ being the correlation between X_j and Z_l. If $c_0 < c^{Pcorr}$, with Condition (C4), we have

$$\begin{aligned}
&P\left(\max_{\substack{j\in\mathcal{A}_k, l\in\mathcal{B}_k,\\ k=1,\ldots,K}} |c_{jl}^{Pcorr}| < c^{Pcorr}\right) \\
&\ge 1 - P\left(\max_{\substack{j\in\mathcal{A}_k, l\in\mathcal{B}_k,\\ k=1,\ldots,K}} |c_{jl}^{Pcorr}| \ge c^{Pcorr}\right) \\
&\ge 1 - \sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k}\sum_{l\in\mathcal{B}_k} P\left(|corr(X_j,Z_l) + c_{jl}^{Pcorr} - corr(X_j,Z_l)| \ge c^{Pcorr}\right) \\
&\ge 1 - \sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k}\sum_{l\in\mathcal{B}_k} P\left(|c_{jl}^{Pcorr} - corr(X_j,Z_l)| \ge c^{Pcorr} - c_0\right)
\end{aligned}$$

$$\geq 1 - 2s^2 \exp\left(-\frac{n(c^{Pcorr} - c_0)^2}{2\kappa_1}\right) \to 1.$$

Thus, if $c_0 < c^{Pcorr}$, with probability approaching 1, $\max_{j\in\mathcal{A}_k, l\in\mathcal{B}_k, k=1,\ldots,K} |c_{jl}^{Pcorr}| < c^{Pcorr}$. That is, $IV = 0$. Next, consider the scenario with $c_0 \geq c^{Pcorr}$. With a first order Taylor's expansion,

$$e^{-\left(\alpha_{kj}^0+\delta_n u_{kj}\right)^2/\tau} - e^{-\left(\alpha_{kj}^0\right)^2/\tau} = -\frac{2}{\tau} e^{-(\tilde{\alpha}_{kj})^2/\tau} \tilde{\alpha}_{kj}\delta_n u_{kj}, \tag{A.1}$$

and

$$e^{-(\beta_{kl}^0+\delta_n v_{kl})^2/\tau} - e^{-(\beta_{kl}^0)^2/\tau} = -\frac{2}{\tau} e^{-\left(\tilde{\beta}_{kl}\right)^2/\tau} \tilde{\beta}_{kl}\delta_n v_{kl}. \tag{A.2}$$

Denote $\boldsymbol{\eta}_{kjl}^0 = (\alpha_{kj}^0, \beta_{kl}^0)'$ and $\boldsymbol{\psi}_{kjl} = (u_{kj}, v_{kl})'$. Then we have

$$\begin{aligned}
&e^{-((\alpha_{kj}^0+\delta_n u_{kj})^2+(\beta_{kl}^0+\delta_n v_{kl})^2)/\tau} - e^{-((\alpha_{kj}^0)^2+(\beta_{kl}^0)^2)/\tau} \\
&\triangleq e^{-((\boldsymbol{\eta}_{kjl}^0+\delta_n \boldsymbol{\psi}_{kjl})'(\boldsymbol{\eta}_{kjl}^0+\delta_n \boldsymbol{\psi}_{kjl}))/\tau} - e^{-(\boldsymbol{\eta}_{kjl}^0)'(\boldsymbol{\eta}_{kjl}^0)/\tau} \\
&= -\frac{2}{\tau} e^{-(\tilde{\boldsymbol{\eta}}_{kjl})'(\tilde{\boldsymbol{\eta}}_{kjl})/\tau} \delta_n \left(\tilde{\boldsymbol{\eta}}_{kjl}\right)' \boldsymbol{\psi}_{kjl},
\end{aligned} \tag{A.3}$$

where $\tilde{\boldsymbol{\theta}}_{\mathcal{C}} = ((\tilde{\boldsymbol{\mu}})', (\tilde{\boldsymbol{\sigma}})', (\tilde{\boldsymbol{\alpha}}_{1,\mathcal{A}_1})', \ldots, (\tilde{\boldsymbol{\alpha}}_{K,\mathcal{A}_K})', (\tilde{\boldsymbol{\beta}}_{1,\mathcal{B}_1})', \ldots, (\tilde{\boldsymbol{\beta}}_{K,\mathcal{B}_K})')'$ lies on the line segment connecting $\boldsymbol{\theta}_{\mathcal{C}}^0 + \delta_n \boldsymbol{w}$ and $\boldsymbol{\theta}_{\mathcal{C}}^0$. Denote $b_0 = \min\{\{|\alpha_{kj}^0|, j \in \mathcal{A}_k\}, \{|\beta_{kl}^0|, l \in \mathcal{B}_k\}, k = 1, \ldots, K\}$. First consider (A.1). With $\delta_n = \sqrt{s/n}$, we have $\tilde{\alpha}_{kj} > b_0/\sqrt{2}$. Then with Condition (C4), we have $\tilde{\alpha}_{kj}^2/\tau > b_0^2/2\tau > 1$. Since $e^{-x}x$ is monotonically decreasing when $x > 1$,

$$\begin{aligned}
\left|e^{-\left(\alpha_{kj}^0+\delta_n u_{kj}\right)^2/\tau} - e^{-\left(\alpha_{kj}^0\right)^2/\tau}\right| &= \frac{2}{\tau} e^{-(\tilde{\alpha}_{kj})^2/\tau} |\tilde{\alpha}_{kj}|\delta_n |u_{kj}| \\
&= \frac{(\tilde{\alpha}_{kj})^2}{\tau} e^{-(\tilde{\alpha}_{kj})^2/\tau} \frac{2}{|\tilde{\alpha}_{kj}|} \delta_n |u_{kj}| \\
&\leq \frac{b_0^2}{\tau} e^{-b_0^2 2\tau} \frac{\sqrt{2}}{b_0} \delta_n |u_{kj}| = \frac{\sqrt{2} b_0}{\tau} e^{-b_0^2/2\tau} \delta_n |u_{kj}|.
\end{aligned}$$

Similar conclusions can be drawn for (A.2) and (A.3). With $\sqrt{n}\lambda_2 b_0 e^{-b_0^2/2\tau}/\tau = o(1)$ in Condition (C4), we have

$$\begin{aligned}
|IV| \leq n\lambda_2 \sum_{k=1}^{K} \sum_{j\in\mathcal{A}_k} \sum_{l\in\mathcal{B}_k} \Big\{ &\left|e^{-\left(\alpha_{kj}^0+\delta_n u_{kj}\right)^2/\tau} - e^{-\left(\alpha_{kj}^0\right)^2/\tau}\right| \\
&+ \left|e^{-(\beta_{kl}^0+\delta_n v_{kl})^2/\tau} - e^{-(\beta_{kl}^0)^2/\tau}\right| \\
&+ \left|e^{-((\alpha_{kj}^0+\delta_n u_{kj})^2+(\beta_{kl}^0+\delta_n v_{kl})^2)/\tau} - e^{-((\alpha_{kj}^0)^2+(\beta_{kl}^0)^2)/\tau}\right| \Big\}
\end{aligned}$$

$$\leq 2n\lambda_2 \frac{\sqrt{2}b_0}{\tau} e^{-b_0^2/2\tau} \delta_n \sqrt{s} ||\boldsymbol{w}||$$
$$= 2\sqrt{n}\lambda_2 \frac{\sqrt{2}b_0}{\tau} e^{-b_0^2/2\tau} n\delta_n^2 ||\boldsymbol{w}|| = o_p(n\delta_n^2)||\boldsymbol{w}||.$$

It is observed that II dominates I, III, and IV, and is negative, since $I(\boldsymbol{\theta}_{\mathcal{C}})$ is positive definite at $\boldsymbol{\theta}_{\mathcal{C}} = \boldsymbol{\theta}_{\mathcal{C}}^0$. This completes the proof.

Proof of Theorem 2

Let $\hat{\boldsymbol{\theta}}$ have $\hat{\boldsymbol{\theta}}_{\mathcal{C}} = \boldsymbol{\theta}_{\mathcal{C}}^*$, a strict local maximizer of $\tilde{Q}_n(\boldsymbol{\theta}_{\mathcal{C}})$, and $\hat{\boldsymbol{\theta}}_{\mathcal{C}^c} = 0$. First, consider $\hat{\boldsymbol{\alpha}}_{k,\mathcal{A}_k^c}$. Following Theorem 1 in Fan and Lv (2011), with Condition (C5) and Theorem 1, it suffices to check condition (8) in Fan and Lv (2011). Let

$$h_1 = (n\lambda_1)^{-1} \left[\frac{\partial L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}_{k,\mathcal{A}_k^c}} \bigg|_{\hat{\boldsymbol{\theta}}} - \lambda_2 n \, \frac{\partial \rho_2(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha}_{k,\mathcal{A}_k^c}} \bigg|_{\hat{\boldsymbol{\theta}}} \right],$$

where $\rho_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{k=1}^K \sum_{j=1}^p \sum_{l=1}^q c_{jl}(1 - e^{-\alpha_{kj}^2/\tau})(1 - e^{-\beta_{kl}^2/\tau})$.

For $j \in \mathcal{A}_k^c$, $\partial \rho_2(\boldsymbol{\alpha}, \boldsymbol{\beta})/\partial a_{kj} = \sum_{l=1}^q (2/\tau) c_{jl}(1 - e^{-\beta_{kl}^2/\tau}) e^{-\alpha_{kj}^2/\tau} \alpha_{kj}$. As $\hat{\boldsymbol{\alpha}}_{k,\mathcal{A}_k^c} = 0$, $\partial \rho_2(\boldsymbol{\alpha}, \boldsymbol{\beta})/\partial a_{kj}|_{\hat{\boldsymbol{\theta}}} = 0$ for $j \in \mathcal{A}_k^c$. Therefore, $\lambda_2 n \, \partial \rho_2(\boldsymbol{\alpha}, \boldsymbol{\beta})/\partial \boldsymbol{\alpha}_{k,\mathcal{A}_k^c}|_{\hat{\boldsymbol{\theta}}} = 0$. Then, we have

$$\|h_1\|_\infty = (n\lambda_1)^{-1} \max_{j \in \mathcal{A}_k^c} \left| \frac{\partial L_n(\boldsymbol{\theta})}{\partial \alpha_{kj}} \bigg|_{\hat{\boldsymbol{\theta}}} \right|.$$

For $j \in \mathcal{A}^c$, we have

$$\begin{aligned} \frac{\partial L_n(\boldsymbol{\theta})}{\partial \alpha_{kj}} \bigg|_{\hat{\boldsymbol{\theta}}} &= \frac{\partial L_n(\boldsymbol{\theta})}{\partial \alpha_{kj}} \bigg|_{\boldsymbol{\theta}^0} + \left(\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0 \right)' \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \alpha_{kj} \partial \boldsymbol{\theta}_{\mathcal{C}}} \bigg|_{\boldsymbol{\theta}^0} \\ &\quad + \left(\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0 \right)' \frac{\partial^3 L_n(\boldsymbol{\theta})}{\partial \alpha_{kj} \partial^2 \boldsymbol{\theta}_{\mathcal{C}}} \bigg|_{\tilde{\boldsymbol{\theta}}} \left(\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0 \right), \end{aligned} \tag{A.4}$$

where $\tilde{\boldsymbol{\theta}}$ lies on the line segment connecting $\boldsymbol{\theta}^0$ and $\hat{\boldsymbol{\theta}}$.

For the first term of (A.4), consider the event

$$\Omega_1 = \left\{ \max_{j \in \mathcal{A}_k^c} \left| \frac{\partial L_n(\boldsymbol{\theta})}{\partial \alpha_{kj}} \bigg|_{\boldsymbol{\theta}^0} \right| \leq \zeta_n \sqrt{n} \right\},$$

with $\zeta_n = n^a (\log(n))^{1/2}, a \in (0, 1/2)$. With Condition (C3) and Bernstein's inequality, we have

$$P(\Omega_1) = 1 - P\left\{ \max_{j \in \mathcal{A}_k^c} \left| \frac{\partial L_n(\boldsymbol{\theta})}{\partial \alpha_{kj}} \bigg|_{\boldsymbol{\theta}^0} \right| > \zeta_n \sqrt{n} \right\}$$

$$\geq 1 - \sum_{j\in\mathcal{A}_k^c} P\left\{\left|\frac{1}{\sqrt{n}}\frac{\partial L_n(\boldsymbol{\theta})}{\partial \alpha_{kj}}\bigg|_{\boldsymbol{\theta}^0}\right| > \zeta_n\right\}$$
$$\geq 1 - 2(p-a_k)\exp\left(-\frac{\zeta_n^2}{2\kappa_2}\right)$$
$$\geq 1 - 2p\exp\left(-\frac{\zeta_n^2}{2\kappa_2}\right) \to 1,$$

as $\log(p) = O(n^a)$ in Condition (C6). Thus, with probability approaching 1,

$$\max_{j\in\mathcal{A}_k^c}\left|\frac{\partial L_n(\boldsymbol{\theta})}{\partial \alpha_{kj}}\bigg|_{\boldsymbol{\theta}^0}\right| = O(n^{a/2+1/2}\sqrt{\log n}).$$

For the second term of (A.4), by Condition (C3) and Cauchy-Schwartz inequality,

$$\max_{j\in\mathcal{A}_k^c}\left|\left(\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0\right)' \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \alpha_{kj}\partial \boldsymbol{\theta}_{\mathcal{C}}}\bigg|_{\boldsymbol{\theta}^0}\right|$$
$$\leq \max_{j\in\mathcal{A}_k^c}\sum_{i=1}^n\left|\sum_{l\in\mathcal{C}}\frac{\partial^2 \log f(y_i;\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot},\boldsymbol{\theta})}{\partial\alpha_{kj}\partial\theta_l}(\hat{\theta}_l - \theta_l^0)\right|$$
$$\leq \max_{j\in\mathcal{A}_k^c}\sum_{i=1}^n\left(\sum_{l\in\mathcal{C}}\left(\frac{\partial^2 \log f(y_i;\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot},\boldsymbol{\theta})}{\partial\alpha_{kj}\partial\theta_l}\right)^2\right)^{1/2}\|\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0\|$$
$$\leq \sum_{i=1}^n\left(s\left(M_1(V_i)\right)^2\right)^{1/2}\left\|\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0\right\| = O_p(s\sqrt{n}).$$

For the third term of (A.4), by Condition (C3) and Cauchy-Schwartz inequality,

$$\max_{j\in\mathcal{A}_k^c}\left|\left(\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0\right)' \frac{\partial^3 L_n(\boldsymbol{\theta})}{\partial \alpha_{kj}\partial^2 \boldsymbol{\theta}_{\mathcal{C}}}\bigg|_{\tilde{\boldsymbol{\theta}}}\left(\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0\right)\right|$$
$$\leq \max_{j\in\mathcal{A}_k^c}\sum_{i=1}^n\left(\sum_{l,m\in\mathcal{C}}\left(\frac{\partial^3 \log f(y_i;\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot},\boldsymbol{\theta})}{\partial\alpha_{kj}\partial\theta_l\partial\theta_m}\bigg|_{\tilde{\boldsymbol{\theta}}}\right)^2\right)^{1/2}\left\|\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0\right\|^2$$
$$\leq \sum_{i=1}^n\left(s^2\left(M_2(V_i)\right)^2\right)^{1/2}\left\|\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0\right\|^2 = o_p(s\sqrt{n}).$$

Then, Condition (C5) gives $||h_1||_\infty \leq o_p(1)$. Next, consider $\hat{\boldsymbol{\beta}}_{k,\mathcal{B}_k^c}$. Similar to above, let

$$h_2 = (n\lambda_1)^{-1}\left[\frac{\partial L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_{k,\mathcal{B}_k^c}}\bigg|_{\hat{\boldsymbol{\theta}}} - \lambda_2 n\,\frac{\partial \rho_2(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{k,\mathcal{B}_k^c}}\bigg|_{\hat{\boldsymbol{\theta}}}\right].$$

For $l \in \mathcal{B}_k^c$, $\partial\rho_2(\boldsymbol{\alpha},\boldsymbol{\beta})/\partial\beta_{kl} = \sum_{j=1}^p (2/\tau)c_{jl}(1-e^{-\alpha_{kj}^2/\tau})e^{-\beta_{kl}^2/\tau}\beta_{kl}$. As $\hat{\boldsymbol{\beta}}_{k,\mathcal{B}_k^c} = 0$,

$\partial\rho_2(\boldsymbol{\alpha},\boldsymbol{\beta})/\partial\beta_{kl}|_{\hat{\boldsymbol{\theta}}} = 0$ for $l \in \mathcal{B}_k^c$. Therefore, $\lambda_2 n\ (\partial\rho_2(\boldsymbol{\alpha},\boldsymbol{\beta})/\partial\boldsymbol{\beta}_{k,\mathcal{B}_k^c})\big|_{\hat{\boldsymbol{\theta}}} = 0$. Then, we have

$$\|h_2\|_\infty = (n\lambda_1)^{-1} \max_{l\in\mathcal{B}_k^c} \left| \frac{\partial L_n(\boldsymbol{\theta})}{\partial\beta_{kl}}\bigg|_{\hat{\boldsymbol{\theta}}} \right|.$$

For $l \in \mathcal{B}^c$, we have

$$\begin{aligned}\frac{\partial L_n(\boldsymbol{\theta})}{\partial\beta_{kl}}\bigg|_{\hat{\boldsymbol{\theta}}} &= \frac{\partial L_n(\boldsymbol{\theta})}{\partial\beta_{kl}}\bigg|_{\boldsymbol{\theta}^0} + \left(\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0\right)' \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial\beta_{kl}\partial\boldsymbol{\theta}_{\mathcal{C}}}\bigg|_{\boldsymbol{\theta}^0} \\ &\quad + \left(\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0\right)' \frac{\partial^3 L_n(\boldsymbol{\theta})}{\partial\beta_{kl}\partial^2\boldsymbol{\theta}_{\mathcal{C}}}\bigg|_{\tilde{\boldsymbol{\theta}}} \left(\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0\right), \end{aligned} \tag{A.5}$$

where $\tilde{\boldsymbol{\theta}}$ lies on the line segment connecting $\boldsymbol{\theta}^0$ and $\hat{\boldsymbol{\theta}}$. For the first term of (A.5), consider the event

$$\Omega_2 = \left\{ \max_{l\in\mathcal{B}_k^c} \left| \frac{\partial L_n(\boldsymbol{\theta})}{\partial\beta_{kl}}\bigg|_{\boldsymbol{\theta}^0} \right| \le \zeta_n\sqrt{n} \right\},$$

with $\zeta_n = n^a(\log(n))^{1/2}$. Similar to the analysis of Ω_1, we have

$$\begin{aligned} P(\Omega_2) &= 1 - P\left\{ \max_{l\in\mathcal{B}_k^c} \left| \frac{\partial L_n(\boldsymbol{\theta})}{\partial\beta_{kl}}\bigg|_{\boldsymbol{\theta}^0} \right| > \zeta_n\sqrt{n} \right\} \\ &\ge 1 - \sum_{l\in\mathcal{B}_k^c} P\left\{ \left| \frac{1}{\sqrt{n}} \frac{\partial L_n(\boldsymbol{\theta})}{\partial\beta_{kl}}\bigg|_{\boldsymbol{\theta}^0} \right| > \zeta_n \right\} \\ &\ge 1 - 2(q-b_k)\exp\left(-\frac{\zeta_n^2}{2\kappa_2}\right) \\ &\ge 1 - 2q\exp\left(-\frac{\zeta_n^2}{2\kappa_2}\right) \to 1, \end{aligned}$$

as $\log(q) = O(n^a)$ in Condition (C6). Thus, with probability approaching 1,

$$\max_{l\in\mathcal{B}_k^c} \left| \frac{\partial L_n(\boldsymbol{\theta})}{\partial\beta_{kl}}\bigg|_{\boldsymbol{\theta}^0} \right| = O(n^{a/2+1/2}\sqrt{\log n}).$$

For the second term of (A.5), by Condition (C3) and Cauchy-Schwartz inequality,

$$\begin{aligned} &\max_{l\in\mathcal{B}_k^c} \left| \left(\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^0\right)' \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial\beta_{kl}\partial\boldsymbol{\theta}_{\mathcal{C}}}\bigg|_{\boldsymbol{\theta}^0} \right| \\ &\le \max_{l\in\mathcal{B}_k^c} \sum_{i=1}^n \left| \sum_{j\in\mathcal{C}} \frac{\partial^2 \log f(y_i;\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot},\boldsymbol{\theta})}{\partial\beta_{kl}\partial\theta_j}(\hat{\theta}_j - \theta_j^0) \right| \end{aligned}$$

Table 5. Simulation settings: regression coefficients of important variables.

Setting	p_0		Regression coefficient																			
P1	10	$\boldsymbol{\alpha}_1$	1.5	1.5	1.5	1.5	1.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0	0
		$\boldsymbol{\alpha}_2$	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	0	0	0	0	0	0	0	0	0	0
		$\boldsymbol{\beta}_1$	1.5	1.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0	0
		$\boldsymbol{\beta}_2$	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	0	0	0	0	0	0	0	0	0	0
P2	13	$\boldsymbol{\alpha}_1$	1.5	1.5	1.5	1.5	1.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0	0
		$\boldsymbol{\alpha}_2$	0	0	0	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	0	0	0	0	0	0	0
		$\boldsymbol{\beta}_1$	1.5	1.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0	0
		$\boldsymbol{\beta}_2$	0	0	0	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	0	0	0	0	0	0	0
P3	10	$\boldsymbol{\alpha}_1$	1.5	1.5	1.5	1.5	1.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0	0
		$\boldsymbol{\alpha}_2$	−0.5	−0.5	−0.5	−0.5	−0.5	0.5	0.5	−0.5	−0.5	−0.5	0	0	0	0	0	0	0	0	0	0
		$\boldsymbol{\beta}_1$	1.5	1.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0	0
		$\boldsymbol{\beta}_2$	−0.5	−0.5	−0.5	−0.5	−0.5	0.5	0.5	−0.5	−0.5	−0.5	0	0	0	0	0	0	0	0	0	0
P4	20	$\boldsymbol{\alpha}_1$	1.5	1.5	1.5	1.5	1.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0	0
		$\boldsymbol{\alpha}_2$	0	0	0	0	0	0	0	0	0	0	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5
		$\boldsymbol{\beta}_1$	1.5	1.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0	0
		$\boldsymbol{\beta}_2$	0	0	0	0	0	0	0	0	0	0	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5
P5	10	$\boldsymbol{\alpha}_1$	1.5	1.5	1.5	1.5	1.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0	0
		$\boldsymbol{\alpha}_2$	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0	0
		$\boldsymbol{\beta}_1$	1.5	1.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0	0
		$\boldsymbol{\beta}_2$	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0	0

$$\leq \max_{l\in\mathcal{B}_k^c}\sum_{i=1}^{n}\left(\sum_{j\in\mathcal{C}}\left(\frac{\partial^2\log f(y_i;\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot},\boldsymbol{\theta})}{\partial\beta_{kl}\partial\theta_j}\right)^2\right)^{1/2}\|\hat{\boldsymbol{\theta}}_{\mathcal{C}}-\boldsymbol{\theta}_{\mathcal{C}}^0\|$$

$$\leq \sum_{i=1}^{n}\left(s\left(M_1(V_i)\right)^2\right)^{1/2}\left\|\hat{\boldsymbol{\theta}}_{\mathcal{C}}-\boldsymbol{\theta}_{\mathcal{C}}^0\right\| = O_p(s\sqrt{n}).$$

For the third term of (A.5), by Condition (C3) and Cauchy-Schwartz inequality,

$$\max_{l\in\mathcal{B}_k^c}\left|\left(\hat{\boldsymbol{\theta}}_{\mathcal{C}}-\boldsymbol{\theta}_{\mathcal{C}}^0\right)'\left.\frac{\partial^3 L_n(\boldsymbol{\theta})}{\partial\beta_{kl}\partial^2\boldsymbol{\theta}_{\mathcal{C}}}\right|_{\tilde{\boldsymbol{\theta}}}\left(\hat{\boldsymbol{\theta}}_{\mathcal{C}}-\boldsymbol{\theta}_{\mathcal{C}}^0\right)\right|$$

$$\leq \max_{l\in\mathcal{B}_k^c}\sum_{i=1}^{n}\left(\sum_{j,m\in\mathcal{C}}\left(\left.\frac{\partial^3\log f(y_i;\boldsymbol{X}_{i\cdot},\boldsymbol{Z}_{i\cdot},\boldsymbol{\theta})}{\partial\beta_{kl}\partial\theta_j\partial\theta_m}\right|_{\tilde{\boldsymbol{\theta}}}\right)^2\right)^{1/2}\left\|\hat{\boldsymbol{\theta}}_{\mathcal{C}}-\boldsymbol{\theta}_{\mathcal{C}}^0\right\|^2$$

$$\leq \sum_{i=1}^{n}\left(s^2\left(M_2(V_i)\right)^2\right)^{1/2}\left\|\hat{\boldsymbol{\theta}}_{\mathcal{C}}-\boldsymbol{\theta}_{\mathcal{C}}^0\right\|^2 = o_p(s\sqrt{n}).$$

Thus, $||h_2||_\infty \leq o_p(1)$. This completes the proof.

References

Atienza, N., Garciaheras, J., Munozpichardo, J. M. and Villa, R. (2007) On the consistency of MLE in finite mixture models of exponential families. *Journal of Statistical Planning and Inference* **137**, 496–505.

Ahmad, A. and Fröhlich, H. (2017). Towards clinically more relevant dissection of patient heterogeneity via survival-based Bayesian clustering. *Bioinformatics* **33**, 3558–3566.

Althobiti, M., Aleskandarany, M. A., Joseph, C., Toss, M., Mongan, N., Diez-Rodriguez, M. et al. (2018). Heterogeneity of tumour-infiltrating lymphocytes in breast cancer and its prognostic significance. *Histopathology* **73**, 887–896.

Bair, E. (2013). Semi-supervised clustering methods. *Wiley Interdisciplinary Reviews Computational Statistics* **5**, 349–361.

Boulesteix, A., De, B. R., Jiang, X. and Fuchs, M. (2017). IPF-LASSO: Integrative L1-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and Mathematical Methods in Medicine* 7691937.

Coppock, A., Leeper, T. J. and Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences* **115**, 12441–12446.

Chai, H., Shi, X., Zhang, Q., Zhao, Q., Huang, Y. and Ma, S. (2017). Analysis of cancer gene expression data with an assisted robust marker identification approach. *Genetic Epidemiology* **41**, 779–789.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of The Royal Statistical Society Series B (Statistical Methodology)* **63**, 19–29.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* **57**, 5467–5484.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32**, 928–961.

Gross, S. M. and Tibshirani, R. (2015). Collaborative regression. *Biostatistics* **16**, 326–338.

Hui, F., Warton, D. and Foster, S. (2015). Multi-species distribution modeling using penalized mixture of regressions. *The Annals of Applied Statistics* **9**, 866–882.

Jiang, Y., Shi, X., Zhao, Q., Krauthammer, M., Rothberg, B. E. and Ma, S. (2016). Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis. *Genomics* **107**, 223–230.

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

Khalili, A. and Lin, S. (2013). Regularization in finite mixture of regression models with diverging number of parameters. *Biometrics* **69**, 436–446.

Kothari, S., Phan, J. H., Young, A. N. and Wang, M. D. (2013). Histological image classification using biologically interpretable shape-based features. *BMC Medical Imaging* **13**, 9.

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

Li, Y., Bie, R., Teran Hidalgo, S. J., Qin, Y., Wu, M. and Ma, S. (2018). Assisted gene expression-based clustering with AWNCut. *Statistics in Medicine* **37**, 4386–4403.

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.

Luo, C., Liu, J., Dey, D. K. and Chen, K. (2016). Canonical variate regression. *Biostatistics* **17**, 468–483.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Vega, J. E. V. et al. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences* **115**, E2970–E2979.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Stadler, N., Buhlmann, P. and van De Geer, S. (2010). L_1-penalization for mixture regression models. *Test* **19**, 209–256.

Sun, D., Li, A., Tang, B. and Wang, M. (2018). Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. *Computer Methods and Programs in Biomedicine* **161**, 45–53.

Turajlic, S., Sottoriva, A., Graham, T. and Swanton, C. (2019). Resolving genetic heterogeneity in cancer. *Nature Reviews Genetics* **20**, 404–416.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., Do, K.-A. et al. (2013). IBAG: Integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* **29**, 149–159.

Wiwie, C., Baumbach, J. and Röttger, R. (2015). Comparing the performance of biomedical clustering methods. *Nature Methods* **12**, 1033–1038.

Yu, K., Berry, G. J., Rubin, D. L., Re, C., Altman, R. B. and Snyder, M. (2017). Association of omics features with histopathology patterns in lung adenocarcinoma. *Cell Systems* **5**, 620–627.

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, R., Zhao, Q., Zhao, H. and Ma, S. (2016). Integrating multidimensional omics data for cancer outcome. *Biostatistics* **17**, 605–618.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.

Tingyan Zhong

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

E-mail: tyzhong@sjtu.edu.cn

Qingzhao Zhang

School of Economics and Wang Yanan Institute for Studies in Economics, Xiamen University, Fujian, China.

E-mail: zhangqingzhao@amss.ac.cn

Jian Huang

Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong.

E-mail: jian-huang@uiowa.edu

Mengyun Wu

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

E-mail: wu.mengyun@mail.shufe.edu.cn

Shuangge Ma

Department of Biostatistics, Yale School of Public Health, Yale University, New Haven, CT 06520-0834, USA.

E-mail: shuangge.ma@yale.edu

(Received October 2020; accepted August 2021)