Statistica Sinica Preprint No: SS-2025-0147	
Title	Effects-Nested Multi-Level Supervised Heterogeneity Analysis
Manuscript ID	SS-2025-0147
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0147
Complete List of Authors	Ruiyue Wang, Sanguo Zhang and Shuangge Ma
Corresponding Authors	Shuangge Ma
E-mails	shuangge.ma@yale.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

EFFECTS-NESTED MULTI-LEVEL SUPERVISED HETEROGENEITY ANALYSIS

Ruiyue Wang[1,2], Sanguo Zhang[1], and Shuangge Ma[2]

[1] *University of Chinese Academy of Sciences*

[2] *Yale University*

Abstract: Under supervised heterogeneity analysis, samples within a population form groups, and different groups have different regression models. In most of the existing analyses, a single level of heterogeneity structure is considered. Partly motivated by multi-level unsupervised analysis such as hierarchical clustering, we consider multi-level supervised heterogeneity analysis. Consider for example a two-level analysis. At the higher level, "coarse" information is used, and samples form a smaller number of groups. At the lower level, "more subtle" information is used, and samples form a larger number of subgroups. To achieve more lucid interpretations, we further consider the scenario where only some variables are relevant at each level, different groups (subgroups) have the same set of relevant variables, and the important variables at the higher level are nested in those at the lower level. A penalized estimation and selection approach is developed, and its theoretical and computational properties are established. Simulation demonstrates competitive performance of the proposed approach. In the analysis of TCGA breast cancer data, the proposed approach leads to sensible

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

grouping/subgrouping, identification, and estimation results. Overall, this study expands the scope of heterogeneity analysis and delivers a practically useful tool.

Key words and phrases: Multi-level, Nested effects, Penalized estimation, Supervised heterogeneity analysis.

1. Introduction

Heterogeneity analysis is routinely conducted, and most of the existing heterogeneity analyses can be classified as unsupervised (which does not involve a response variable) and supervised (which involves a response variable). In this study, we conduct supervised heterogeneity analysis, under which samples form groups, and different groups have different regression models. For such analysis, popular techniques include finite mixture regression (FMR) (Khalili and Chen, 2007), penalized fusion (Ma and Huang, 2017), Bayesian (Qin et al., 2024), and others. When only a subset of variables is relevant, selection using regularization can be conducted. There is vast literature, and we refer to Fan and Lv (2010); Huang et al. (2012) for reviews.

In most of the existing supervised heterogeneity analyses, there is only one level of grouping, and it is assumed that there is a single true modeling and grouping structure. This study has been partly motivated multi-level unsupervised heterogeneity analyses such as hierarchical clustering. Under

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

such analyses, the same samples are grouped at multiple levels. For the simplicity of description, we consider a two-level analysis and note that the discussion can be easily extended to more than two levels. At the higher level, "coarse" information from variables is used, and samples are separated into a smaller number of groups. Note that, it is possible that only a subset of variables contributes information. At the lower level, "more subtle" information is additionally used, and samples are separated into a larger number of subgroups. As more information is needed, a bigger subset of variables (than at the higher level) needs to contribute. In a sense, the modeling and grouping structures at both levels are "true" – here we use a quotation mark to indicate that the true structures may not be interpreted as strictly as for a single-level analysis.

To further elaborate, we consider the popular setting under the penalized fusion-based and some other supervised heterogeneity analyses. For a set of n independent samples, denote $\boldsymbol{\beta}_i$ as the regression coefficient for the i-th sample. In studies such as Ma and Huang (2017); Wang and Su (2021); Tang et al. (2021), the goal is to identify a single true structure, under which samples i and j belong to the same group if and only if $\boldsymbol{\beta}_i$ exactly equals $\boldsymbol{\beta}_j$. If $\boldsymbol{\beta}_i$'s were known, then instead of being limited to a single level of grouping, we could conduct for example hierarchical clustering of $\boldsymbol{\beta}_i$'s and

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

group samples at as many as n levels. A user can specify, for example, that two levels of grouping are needed, and three and five clusters are needed at the two levels, respectively. The second user can also specify two levels but request two and six clusters (or even a different number of levels). Here, in a sense, there is not a single true number of clusters (and corresponding grouping structure). Or, we can also say that whatever number of clusters specified by a user, as long as certain conditions are satisfied (see the theoretical development below for more details), is a "true" number.

The scheme of the proposed analysis is presented in Figure 1. At the higher level, the samples form two groups based on five (out of 15) variables that have nonzero coefficients. At the lower level, they form four subgroups, and this subgrouping is defined based on the first five as well as five additional variables. Intuitively, those additional important variables at the lower level have weaker signals and contribute more subtle information. For comparison, we also present two single-level analyses.

In the literature, a relevant study is Ren et al. (2022), which also conducts a multi-level supervised heterogeneity analysis. In Ren et al. (2022), one set of coarse imaging features is used to define the higher level heterogeneity, and a separate set of refined imaging features is used to define the lower level heterogeneity. It is reinforced that the lower-level subgroups are

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

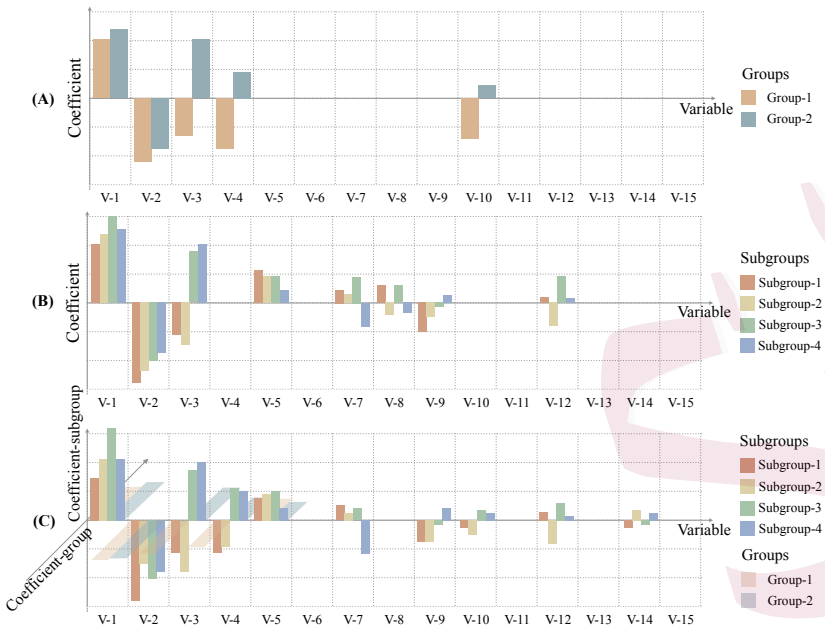

Figure 1: **Top**(**A**): A single-level analysis. Two groups are formed based on five variables with nonzero coefficients. **Middle**(**B**): A single-level analysis. Four subgroups are formed based on eight variables, lacking a effects-nested structure with respect to (A). **Bottom**(**C**): Proposed two-level analysis. Two (four) groups (subgroups) are formed at the higher (lower) level based on five (ten) variables. The sets of nonzero effects have a nested structure.

nested in the higher-level groups, forming a nested structure in terms of samples. This study and Ren et al. (2022) share some conceptual common ground. However, their data settings and analysis goals significantly differ

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

in multiple critical ways, which subsequently lead to significant differences in theoretical developments and applications. First, in Ren et al. (2022), two separate sets of variables are needed for two-level heterogeneity analysis. With this requirement in data sources, the number of analysis levels can be very limited. In contrast, in this study, multiple levels of analysis use the same variables. This is more coherent with hierarchical clustering and some other multi-level analyses. In principle, there can be easily more than two (or even many) levels. Second, in Ren et al. (2022), for the two levels, with two different sets of variables, two sets of true underlying data generating models are defined in a strict way. In contrast, in this study, as the same variables are used for multiple levels of analysis, the definition of "true models" can be somewhat "vague". Different levels demand different regression models, all of which can be viewed as "true". This is also coherent with hierarchical clustering and some other multi-level analyses. Building multiple sets of "true" models at different levels using the same variables significantly increases the complexity of methodological and theoretical developments. Third, in Ren et al. (2022), it is reinforced that the lower-level subgroups and higher-level groups have a nested structure, using a sample-based fusion strategy. In this study, we alternatively reinforce that the higher-level important variables and lower-level important vari-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

ables have a nested structure. The two nested structures are significantly different, demand different methodologies, and complement each other. It is additionally noted that the proposed effect-nested analysis may lead to samples without having a nested structure. On the other hand, as observed in our data analysis, samples can often have a close-to-nested structure.

In many practical fields, heterogeneity analysis can be conducted at multiple levels with different "resolutions". The unsupervised type of such analysis has been well developed. This study can fill an important knowledge gap by conducting supervised analysis. With the high significance of supervised heterogeneity analysis and multi-level unsupervised heterogeneity analysis, the analysis developed in this study can have important implications. Compared to Khalili and Chen (2007); Hui et al. (2015); Ren et al. (2022), it can have weaker and more realistic data requirements while more challenging methodological and theoretical developments. Additionally, this study can deliver a useful tool for many practical data scenarios.

2. Model

2.1 Base Data and Model Settings

We first consider a base setting. Consider data with n independent observations $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$. For the i-th subject, y_i denotes the response, and

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^\top$ denotes the p-dimensional vector of covariates. Consider the FMR framework, and denote K as the number of mixing components. The conditional density for y_i given $\boldsymbol{x}_i$ is:

$$f\left(y; \boldsymbol{x}, \boldsymbol{\Omega}, \boldsymbol{\pi}\right) = \sum_{k=1}^{K} \pi_k f_k\left(y; \boldsymbol{x}, \theta_k(\boldsymbol{x}), \sigma_k\right); \quad \theta_k(\boldsymbol{x}) = h(\boldsymbol{x}^\top \boldsymbol{\beta}_k), \tag{2.1}$$

where $f_k\left(y; \boldsymbol{x}, \theta_k(\boldsymbol{x}), \sigma_k\right)$ is the density for the k-th component, and $h(\cdot)^{-1}$ is a known link function. Denote $\boldsymbol{\Omega} = \{\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \cdots, \boldsymbol{\beta}_K^\top, \boldsymbol{\sigma}^\top\}^\top$, where $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kp})^\top$ represents the covariate effects for the k-th component and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_K)^\top$ contains the dispersion parameters. $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)^\top$ contains the mixing probabilities satisfying $\pi_k > 0$, and $\sum_{k=1}^{K} \pi_k = 1$. Additionally, denote $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \cdots, \boldsymbol{\beta}_K^\top\}^\top$. The observed log-likelihood is: $\mathcal{L}(\boldsymbol{\Omega}, \boldsymbol{\pi}) = \frac{1}{n} \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_k\left(y_i; \boldsymbol{x}_i, \theta_k(\boldsymbol{x}_i), \sigma_k\right)\right)$. For discussion on identifiability (up to a permutation of component labels), we refer to McLachlan et al. (2019). In what follows, we focus on linear models – meaning that f_k is the Gaussian density for $y - \boldsymbol{x}^\top \boldsymbol{\beta}_k$. Conceptually, the proposed analysis can be easily extended to other models. With a moderate to large number of variables and the sparsity assumption, penalization and other regularization techniques can be applied. For references, we refer to Khalili and Lin (2013); Hui et al. (2015).

Under this setting, we consider a simplified two-level analysis. The lower level is as defined above, and subject i belongs to the k-th subgroup

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

if $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_k + \epsilon_i$. For the upper level, we specify the number of groups as one (that is, the special case of a homogeneity model). We can write $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_{co} + \boldsymbol{x}_i^\top (\boldsymbol{\beta}_k - \boldsymbol{\beta}_{co}) + \epsilon_i$. Here, $\boldsymbol{\beta}_{co}$ is the "true" regression coefficient vector shared by all the samples, and $\boldsymbol{x}_i^\top (\boldsymbol{\beta}_k - \boldsymbol{\beta}_{co})$ can be viewed as an additional noise. A special case, which is intuitively reasonable and is considered in this study, is $\boldsymbol{\beta}_k = \boldsymbol{\beta}_{co} + \boldsymbol{\alpha}_k$ and $\beta_{coj} \cdot \alpha_{kj} = 0$, $j = 1, 2, \cdots, p$, indicating that the additional signals corresponding to $\boldsymbol{\alpha}_k$ are used to define heterogeneity for the k-th subgroup at the lower level.

2.2 Effects-nested Multi-level Analysis

Our goal is to simultaneously conduct multi-level analysis. With the higher-level important variable set nested in the lower-level one, we "decouple" the covariate effects at each level into a leading component (which likely has higher signals and is consistent with the higher-level important effects) and a level-specific component. To fix ideas and simplify notation, here, we specifically consider a two-level analysis and denote K_u and K_l as the numbers of upper-level groups and lower-level subgroups, respectively, with $K_u < K_l$ (noting that we no longer use a single K). We denote the effects for the k-th upper-level group by $\boldsymbol{\beta}_k^u = (\beta_{k1}^u, \beta_{k2}^u, \cdots, \beta_{kp}^u)^\top$, $k = 1, 2, \cdots, K_u$, and for the k-th lower-level subgroup by $\boldsymbol{\beta}_k^l = (\beta_{k1}^l, \beta_{k2}^l, \cdots, \beta_{kp}^l)^\top$, $k =$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

$1, 2, \cdots, K_l$. The lower-level effects are further decoupled as:

$$\boldsymbol{\beta}_k^l = \boldsymbol{\alpha}_k^l + \boldsymbol{\gamma}_k^l, \quad s.t. \quad \boldsymbol{\alpha}_k^l \odot \boldsymbol{\gamma}_k^l = \mathbf{0}, \quad k = 1, 2, \cdots, K_l, \tag{2.2}$$

where, for the k-th subgroup, $\boldsymbol{\alpha}_k^l$ denotes the leading effect, and $\boldsymbol{\gamma}_k^l$ denotes the level-specific effect. "$\odot$" denotes element-wise multiplication.

For the selection of important effects, estimation, and (sub)grouping, we propose objective function:

$$\mathcal{L}_{ma}(\boldsymbol{\Omega}, \boldsymbol{\pi}^u, \boldsymbol{\pi}^l) = \mathcal{L}(\boldsymbol{\Omega}^u, \boldsymbol{\pi}^u) + \mathcal{L}(\widetilde{\boldsymbol{\Omega}}^l, \boldsymbol{\pi}^l) - \mathcal{P}_{ma}\left(\boldsymbol{\beta}^u, \boldsymbol{\alpha}^l, \boldsymbol{\gamma}^l\right), \tag{2.3}$$

where $\boldsymbol{\Omega} = (\boldsymbol{\Omega}^{u\top}, \widetilde{\boldsymbol{\Omega}}^{l\top})^\top$, $\boldsymbol{\Omega}^u = \left(\boldsymbol{\Omega}_1^{u\top}, \cdots, \boldsymbol{\Omega}_{K_u}^{u\top}\right)^\top$, $\boldsymbol{\Omega}_k^u = \text{vec}\left(\boldsymbol{\beta}_k^u, \sigma_k^u\right) = \left(\beta_{k1}^u, \beta_{k2}^u, \cdots, \beta_{kp}^u, \sigma_k^u\right)^\top$, and $\widetilde{\boldsymbol{\Omega}}^l = (\boldsymbol{\Omega}^{l\top}, \boldsymbol{\gamma}^\top)^\top$, $\boldsymbol{\Omega}^l = \left(\boldsymbol{\Omega}_1^{l\top}, \cdots, \boldsymbol{\Omega}_{K_l}^{l\top}\right)^\top$, $\boldsymbol{\Omega}_k^l = \text{vec}\left(\boldsymbol{\alpha}_k^l, \sigma_k^l\right) = \left(\alpha_{k1}^l, \alpha_{k2}^l, \cdots, \alpha_{kp}^l, \sigma_k^l\right)^\top$, $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_1^{l\top}, \boldsymbol{\gamma}_2^{l\top}, \cdots, \boldsymbol{\gamma}_{K_l}^{l\top}\}^\top$. Also, $\boldsymbol{\pi}^u = \left(\pi_1^u, \pi_2^u, \ldots, \pi_{K_u}^u\right)^\top$ and $\boldsymbol{\pi}^l = \left(\pi_1^l, \pi_2^l, \ldots, \pi_{K_l}^l\right)^\top$. Here, operator "vec$(\cdot)$" straightens column vectors, and $\mathcal{L}(\boldsymbol{\Omega}^u)$ and $\mathcal{L}(\widetilde{\boldsymbol{\Omega}}^l)$ denote the observed log-likelihood functions for the upper and lower levels, respectively. The key development is the Multilevel Aligning (MA) penalty:

$$\mathcal{P}_{ma}\left(\boldsymbol{\beta}^u, \boldsymbol{\alpha}^l, \boldsymbol{\gamma}^l\right) = \sum_{j=1}^{p} P\left(\sqrt{\sum_{k=1}^{K_u}(\beta_{kj}^u)^2 + \sum_{k=1}^{K_l}(\alpha_{kj}^l)^2}, \lambda_1\right) + \sum_{j=1}^{p} P\left(\sqrt{\sum_{k=1}^{K_l}(\gamma_{kj}^l)^2}, \lambda_2\right), \tag{2.4}$$

where $P(\cdot, \lambda)$ is a penalty function, and λ_1 and λ_2 are tuning parameters. Choices for the penalty include Lasso, SCAD, MCP, and others. Note

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

that in the proposed multi-level analysis, at each level, once the number of (sub)groups at that level is set, the "true" parameters can be accordingly determined. As such, the estimation of parameters as well as the determination of grouping and subgrouping structures can be conducted in a similar way as under the FMR framework.

Rationale A sparse group type penalty is proposed. It ensures that, if a variable is identified as important at the higher level (with nonzero estimates corresponding to $\boldsymbol{\beta}$), then its lower-level estimates (corresponding to $\boldsymbol{\alpha}$) are automatically nonzero, making it identified as the lower level. The level-specific effect $\boldsymbol{\gamma}$ is only identified in the more heterogeneous lower-level subgroups. However, this is not true the other way around. At the same level, all the estimates corresponding to one specific variable across the groups/subgroups are viewed as a group. With group penalization, they are either all zero or all nonzero. In addition, our model degenerates into the simplified analysis in Section 2.1 when $K_u = 1$ and $\beta_{1j}^u = \alpha_{kj}^l$, $k = 1, \cdots, K_l$. Our approach allows for more flexible, or "more heterogeneous", analysis, and we also consider the "more homogeneous" case in Section 7.

Remarks: The proposed approach can be generalized to more than two levels in a rather straightforward way. Consider for example a three-level analysis (with an additional middle level denoted by "m"). Denote the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

numbers of mixture components at the three levels as $K_u < K_m < K_l$. For decoupling the effects, we write: $\boldsymbol{\beta}_k^m = \boldsymbol{\alpha}_k^m + \boldsymbol{\gamma}_k^m,\ s.t.\ \boldsymbol{\alpha}_k^m \odot \boldsymbol{\gamma}_k^m = 0, k = 1, 2, \cdots, K_m$. Consider the penalty:

$$\mathcal{P}_{ma}\left(\boldsymbol{\beta}^u, \boldsymbol{\alpha}^m, \boldsymbol{\gamma}^m, \boldsymbol{\alpha}^l, \boldsymbol{\gamma}^l\right) = \sum_{j=1}^{p} P\left(\sqrt{\sum_{k=1}^{K_u}(\beta_{kj}^u)^2 + \sum_{k=1}^{K_m}(\alpha_{kj}^m)^2}, \lambda_1\right)$$
$$+ \sum_{j=1}^{p} P\left(\sqrt{\sum_{k=1}^{K_m}(\alpha_{kj}^m + \gamma_{kj}^m)^2 + \sum_{k=1}^{K_l}(\alpha_{kj}^l)^2}, \lambda_2\right) + \sum_{j=1}^{p} P\left(\sqrt{\sum_{k=1}^{K_l}(\gamma_{kj}^l)^2}, \lambda_3\right), \tag{2.5}$$

where λ_1, λ_2, and λ_3 are tuning parameters. In a similar manner, the analysis can be extended to more levels. Obviously, computation is expected to be more complicated. Our preliminary exploration suggests that statistical properties similar to those described below can be established. Also noted that more levels of analysis introduce more model parameters, inevitably bringing computational challenges. In unsupervised hierarchical clustering and other analyses, sometimes, certain levels/numbers of grouping lead to insensible results. This is also expected to hold for the proposed analysis.

3. Statistical Properties

For the simplicity of notation, we still consider two levels. As discussed above, it is not assumed that there is a single true model. However, our exploration suggests that, for the purpose of presentation, it can be easier if

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

we call the lower-level model as the true, with its corresponding parameters indicated using superscript "*".

Define $\mathcal{S} = \{j : \sum_{k=1}^{K_l} \beta_{kj}^{l*2} \neq 0, 1 \leqslant j \leqslant p\} = \{j : \beta_{kj}^{l*} \neq 0, 1 \leqslant j \leqslant p, 1 \leqslant k \leqslant K_l\}$, under the assumption that all the subgroups share the same set of important variables. There exists a decomposition $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ with $\mathcal{S}_1 = \{j : \alpha_{kj}^{l*} \neq 0, 1 \leqslant j \leqslant p, 1 \leqslant k \leqslant K_l\}$ and $\mathcal{S}_2 = \{j : \gamma_{kj}^{l*} \neq 0, 1 \leqslant j \leqslant p, 1 \leqslant k \leqslant K_l\}$, and $\mathcal{S}_1 \cap \mathcal{S}_2 = \varnothing$. For the upper level, the set of important variables $\mathcal{S}_1$ corresponds to model parameters $\boldsymbol{\Omega}^{u*}$, which can be viewed as a "shrunk version" of the lower-level ones. Specifically,

$$\begin{aligned} \boldsymbol{\Omega}^{u*} &= \underset{\boldsymbol{\Omega}^u}{\operatorname{argmax}} \{\mathbb{E}[\sum_{k=1}^{K_u} \pi_k f_k\left(y; \boldsymbol{x}, \boldsymbol{\beta}_k^u, \sigma_k^u\right)]\} \\ s.t. \quad &\{j : \sum_{k=1}^{K_u} \beta_{kj}^{u*2} \neq 0, 1 \leqslant j \leqslant p\} = \mathcal{S}_1. \end{aligned} \tag{3.1}$$

Remarks: To get some intuitive insight, we consider a model with two subgroups $\boldsymbol{Y}_1 \in \mathbb{R}^n, \boldsymbol{X}_1 \in \mathbb{R}^{n\times 5}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^n, \boldsymbol{X}_2 \in \mathbb{R}^{n\times 5}$. For the two subgroups, the first four regression coefficients are the same, and the fifth regression coefficients have the same absolute magnitude but opposite signs. Additionally, we consider the special case of orthogonal designs. The separate OLS estimates are $\beta_{1j} = \frac{\boldsymbol{X}_{1j}^\top \boldsymbol{Y}_1}{\boldsymbol{X}_{1j}^\top \boldsymbol{X}_{1j}}, \beta_{2j} = \frac{\boldsymbol{X}_{2j}^\top \boldsymbol{Y}_2}{\boldsymbol{X}_{2j}^\top \boldsymbol{X}_{2j}}, j = 1, 2, \cdots, 5$, where $\boldsymbol{X}_{1j}$ and $\boldsymbol{X}_{2j}$ are the j-th columns of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, respectively. For the upper level, if we set the number of groups as one (that is, a

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

homogeneity model), with $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top)^\top$ and $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top$, $\beta_j = \frac{\boldsymbol{X}_j^\top \boldsymbol{Y}}{\boldsymbol{X}_j^\top \boldsymbol{X}_j} = \frac{\boldsymbol{X}_{1j}^\top \boldsymbol{Y}_1 + \boldsymbol{X}_{2j}^\top \boldsymbol{Y}_2}{\boldsymbol{X}_{1j}^\top \boldsymbol{X}_{1j} + \boldsymbol{X}_{2j}^\top \boldsymbol{X}_{2j}}, j = 1, 2, \cdots, 5$, where $\beta_j = \beta_{1j} = \beta_{2j} \neq 0$ for $j = 1, 2, 3, 4$, and $\beta_5 = 0$. Here, the "more common" effects of the lower level are "averaged" to form the upper-level effects, while the "more specific" effects of the lower level are canceled out. Although this is a special example, it is not hard to see that many data/model settings have similar properties.

We introduce the following notations. For simplicity, use $\boldsymbol{\alpha}_k$ and $\boldsymbol{\gamma}_k$ to denote $\boldsymbol{\alpha}_k^l$ and $\boldsymbol{\gamma}_k^l$, respectively, for $k = 1, 2, \cdots, K_l$. Similarly, use $\boldsymbol{\beta}_k$ to denote $\boldsymbol{\beta}_k^u$ for $k = 1, 2, \cdots, K_u$ (here we note the different implications from Section 2.1). Denote $\boldsymbol{\Omega}^*$ as the "true" value of $\boldsymbol{\Omega}$. $\boldsymbol{\gamma}^*$, as a part of $\boldsymbol{\Omega}^*$, is similarly defined. Let $|\cdot|$ denote the cardinality of a set, and define the sparsity parameters as $s_1 = |\mathcal{S}_1|$, $s_2 = |\mathcal{S}_2|$, and $s = s_1 + s_2$. Write $x \simeq y$ if $x = Dy$ for some positive constant D. The 2-norm of a vector $z = (z_1, \cdots, z_q)^\top$ is defined as $\|z\|_2 = \sqrt{\sum_{j=1}^q z_j^2}$. The assumed conditions and their implications are described in Supplementary Materials.

Theorem 1. *Suppose that Conditions 1-6 (Supplementary Materials) hold and* $\sqrt{\frac{K_l^3 s \log p}{n}} = O(\lambda_1)$, $\sqrt{\frac{K_l^3 s_2 \log p}{n}} = O(\lambda_2)$. *The minimum signals in* $\boldsymbol{\beta}^*$ *and* $\boldsymbol{\alpha}^*$ *are larger than* $(a+0.5)\cdot\lambda_1$, *and that in* $\boldsymbol{\gamma}^*$ *is larger than* $(a+0.5)\cdot\lambda_2$, *where* a *is the regularization parameter defined in Condition 5. Then, there exists a local maximizer of (2.3), with probability tending to 1:*

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

1. *(Estimation consistency):* $\left\|\widehat{\mathbf{\Omega}}-\mathbf{\Omega}^*\right\|_2=O\left(\sqrt{\frac{K_l^3 s\log p}{n}}\right)$;

2. *(Heterogeneous effects consistency):* $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2=O\left(\sqrt{\frac{K_l^3 s_2\log p}{n}}\right)$;

3. *(Selection consistency):* $\widehat{\mathcal{S}}_1=\mathcal{S}_1,\widehat{\mathcal{S}}_2=\mathcal{S}_2$, *where* $\widehat{\mathcal{S}}_1=\{j:\widehat{\alpha}_{kj}\neq 0,1\leqslant j\leqslant p,1\leqslant k\leqslant K_l\}$, $\widehat{\mathcal{S}}_2=\{j:\widehat{\gamma}_{kj}\neq 0,1\leqslant j\leqslant p,1\leqslant k\leqslant K_l\}$.

The proof is presented in Supplementary Materials. Note that the commonly encountered reshuffling issue is also applicable here. It is "reassuring" that, with a more complicated analysis goal and formulation, the proposed approach enjoys similar much-desired properties as Fan and Lv (2011); Zhang et al. (2016). Compared to the multi-level heterogeneity analysis in Ren et al. (2022), the proposed analysis enjoys stronger variable selection properties by better accommodating high-dimensional variables. In addition, unlike Ren et al. (2022) which is concerned with identifying the exact number of (sub)groups (using penalized fusion to aggregate samples), the proposed analysis can accommodate any reasonably specified number of (sub)groups and is more concerned with differences in parameters at different levels. There are also differences/advancements from the existing mixture regression literature (e.g., Hui et al. (2015) and Sun et al. (2022)). For example, when $s_2=o(s_1)$, the analysis can detect subgroup-level effects with smaller signals. Although with more levels in our analysis, K_l

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

is allowed to diverge like in Hao et al. (2018). Our analysis can accommodate slightly more subgroups as sample size grows and, thus, can be advantageous over some existing ones like Li et al. (2026).

4. Computation

We develop an effective algorithm based on the EM technique. Denote $\boldsymbol{\mu}_k^u$ and $\boldsymbol{\mu}_{k'}^l$ as the means of the k-th group and the k'-th subgroup, respectively. Denote $\boldsymbol{\omega}^u = (\omega_{ik}^u)_{n\times K_u}$ and $\boldsymbol{\omega}^l = \left(\omega_{ik}^l\right)_{n\times K_l}$, where ω_{ik}^u and ω_{ik}^l correspond to the upper and lower levels, respectively, as the latent indicator variables representing the component memberships of the $i-$th sample in the mixtures. Use superscript (t) to denote the updated parameters in the t-th iteration. With initial values from the standard FMR or K-means with regression estimates, the algorithm is sketched in Algorithm 1.

The EM technique has been popular in mixture modeling. There are additional complexities for our model. For example, the optimization of $\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\gamma}^{(t)}$ in the M-step is not straightforward. For this purpose, we develop an Iterative Hierarchical Shooting Algorithm. Here, we describe it for MCP and note that it can be modified for other penalties.

We denote $z_{jk}^u := \frac{\sum_{i=1}^n \omega_{ik}^u\left(-x_{ij}(y_i-\mu_k^u-\boldsymbol{x}_i^\top\boldsymbol{\beta}_k)\right)}{\sum_{i=1}^n \omega_{ik}^u}$ for $k = 1, \cdots, K_u$ and $z_{jk}^l := \frac{\sum_{i=1}^n \omega_{ik}^l\left(-x_{ij}(y_i-\mu_k^l-\boldsymbol{x}_i^\top(\boldsymbol{\alpha}_k+\boldsymbol{\gamma}_k))\right)}{\sum_{i=1}^n \omega_{ik}^l}$ for $k = 1, \cdots, K_l$, $j = 1, \cdots, p$. A necessary

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

Algorithm 1 EM-based Algorithm

Initialize: $\boldsymbol{\pi}^{u(0)} \in \mathbb{R}^{K_u}, \boldsymbol{\pi}^{l(0)} \in \mathbb{R}^{K_l}, \boldsymbol{\mu}^{u(0)} \in \mathbb{R}^{K_u}, \boldsymbol{\mu}^{l(0)} \in \mathbb{R}^{K_l}$, $\boldsymbol{\Omega}^{(0)} \in \mathbb{R}^{(K_u+2K_l)\times p+K_u+K_l}$, given tuning parameters λ_1, λ_2, regularization parameter a, and a sufficiently small constant ζ.

for $t = 1, \cdots, T$ **do**

Input: $\boldsymbol{\pi}^{u(t-1)}, \boldsymbol{\pi}^{l(t-1)}, \boldsymbol{\mu}^{u(t-1)}, \boldsymbol{\mu}^{l(t-1)}, \boldsymbol{\Omega}^{(t-1)}, \{y_i, \boldsymbol{x}_i\}_{i=1}^n$.

for $k = 1, \cdots, K_u$ **do**

for $i = 1, \cdots, n$ **do**

$$\omega_{ik}^{u(t)} = \frac{\pi_k^{u(t-1)} f_k\left(y_i; \boldsymbol{x}_i, \left(1, \boldsymbol{x}_i^\top\right)\left(\mu_k^{u(t-1)}, \boldsymbol{\beta}_k^{(t-1)\top}\right)^\top, \sigma_k^{u(t-1)}\right)}{\sum_{k=1}^{K_u} \pi_k^{u(t-1)} f_k\left(y_i; \boldsymbol{x}_i, \left(1, \boldsymbol{x}_i^\top\right)\left(\mu_k^{u(t-1)}, \boldsymbol{\beta}_k^{(t-1)\top}\right)^\top, \sigma_k^{u(t-1)}\right)}.$$

end for

$$\pi_k^{u(t)} = \frac{1}{n}\sum_{i=1}^n \omega_{ik}^{u(t)}; \quad \mu_k^{u(t)} = \frac{\sum_{i=1}^n \omega_{ik}^{u(t)}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_k^{(t-1)}\right)}{\sum_{i=1}^n \omega_{ik}^{u(t)}}$$

end for

for $k = 1, \cdots, K_u$ **do**

for $i = 1, \cdots, n$ **do**

$$\omega_{ik}^{l(t)} = \frac{\pi_k^{l(t-1)} f_k\left(y_i; \boldsymbol{x}_i, \left(1, \boldsymbol{x}_i^\top\right)\left(\mu_k^{l(t-1)}, \boldsymbol{\alpha}_k^{(t-1)\top} + \boldsymbol{\gamma}_k^{(t-1)\top}\right)^\top, \sigma_k^{l(t-1)}\right)}{\sum_{k=1}^{K_l} \pi_k^{l(t-1)} f_k\left(y_i; \boldsymbol{x}_i, \left(1, \boldsymbol{x}_i^\top\right)\left(\mu_k^{l(t-1)}, \boldsymbol{\alpha}_k^{(t-1)\top} + \boldsymbol{\gamma}_k^{(t-1)\top}\right)^\top, \sigma_k^{l(t-1)}\right)}.$$

end for

$$\pi_k^{l(t)} = \frac{1}{n}\sum_{i=1}^n \omega_{ik}^{l(t)}; \quad \mu_k^{l(t)} = \frac{\sum_{i=1}^n \omega_{ik}^{l(t)}\left(y_i - \boldsymbol{x}_i^\top\left(\boldsymbol{\alpha}_k^{(t-1)} + \boldsymbol{\gamma}_k^{(t-1)}\right)\right)}{\sum_{i=1}^n \omega_{ik}^{l(t)}}$$

end for

Apply Algorithm 2 to get $\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\gamma}^{(t)}$

for $k = 1, \cdots, K_u$ **do**

$$\sigma_k^{u(t)} = \frac{\sum_{i=1}^n \omega_{ik}^{u(t)}\left(y_i - \mu_k^{u(t)} - \boldsymbol{x}_i^\top \boldsymbol{\beta}_k^{(t)}\right)^2}{\sum_{i=1}^n \omega_{ik}^{u(t)}}$$

end for

for $k = 1, \cdots, K_l$ **do**

$$\sigma_k^{l(t)} = \frac{\sum_{i=1}^n \omega_{ik}^{l(t)}\left(y_i - \mu_k^{l(t)} - \boldsymbol{x}_i^\top\left(\boldsymbol{\alpha}_k^{(t)} + \boldsymbol{\gamma}_k^{(t)}\right)\right)^2}{\sum_{i=1}^n \omega_{ik}^{l(t)}}$$

end for

if convergence criterion is satisfied **then**

Break;

end if

Output: $\boldsymbol{\pi}^{u(t)}, \boldsymbol{\pi}^{l(t)}, \boldsymbol{\mu}^{u(t)}, \boldsymbol{\mu}^{l(t)}, \boldsymbol{\Omega}^{(t)}$.

end for

return $\widehat{\boldsymbol{\pi}}^u, \widehat{\boldsymbol{\pi}}^l, \widehat{\boldsymbol{\Omega}}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

and sufficient condition for $\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}$ after an E-step is:

$$\begin{aligned} &z_{jk}^u + D_j \beta_{kj} = 0, \quad \beta_{kj} \neq 0, \quad \text{for} \quad k = 1, \cdots, K_u; \\ &z_{jk}^l + D_j \alpha_{kj} = 0, \quad \alpha_{kj} \neq 0, \quad \text{for} \quad k = 1, \cdots, K_l; \\ &z_{jk}^l + F_j \gamma_{kj} = 0, \quad \gamma_{kj} \neq 0, \alpha_{kj} = 0 \quad \text{for} \quad k = 1, \cdots, K_l; \end{aligned} \tag{4.1}$$

$$\begin{aligned} &A_j := \left\| \left(z_{j1}^u, \cdots, z_{jK_u}^u, z_{j1}^l, \cdots, z_{jK_l}^l \right) \right\|_2 \leqslant \lambda_1, \quad \beta_{kj} = 0 \quad \text{for} \quad k = 1, \cdots, K_u \\ &\quad \text{and} \quad \alpha_{kj} = 0 \quad \text{for} \quad k = 1, \cdots, K_l; \\ &B_j := \left\| \left(z_{j1}^l, \cdots, z_{jK_l}^l \right) \right\|_2 \leqslant \lambda_2, \quad \gamma_{kj} = 0 \quad \text{for} \quad k = 1, \cdots, K_l. \end{aligned} \tag{4.2}$$

where $D_j = \left(\frac{\lambda_1}{\left\| (\boldsymbol{\beta}_{\cdot j}^\top, \boldsymbol{\alpha}_{\cdot j}^\top) \right\|_2} - \frac{1}{a} \right)_+$, $F_j = \left(\frac{\lambda_2}{\left\| \boldsymbol{\gamma}_{\cdot j} \right\|_2} - \frac{1}{a} \right)_+$, and $(\cdot)_+ = \max(\cdot, 0)$ is the ReLU function. A closed-form solution is not available here when the orthonormal condition is not satisfied. Consider that the first line of condition (4.1) is equivalent to:

$$\frac{\sum_{i=1}^n \omega_{ik}^u \left(x_{ij} \left(y_i - \mu_k^u - \boldsymbol{x}_i^\top \boldsymbol{\beta}_{k(-j)} \right) \right)}{\sum_{i=1}^n \omega_{ik}^u} = \left(\frac{\sum_{i=1}^n \omega_{ik}^u x_{ij}^2}{\sum_{i=1}^n \omega_{ik}^u} + D_j \right) \beta_{kj},$$

where $\boldsymbol{\beta}_{k(-j)} = \left(\beta_{k1}, \ldots, \beta_{k(j-1)}, 0, \beta_{k(j-1)}, \ldots, \beta_{kp} \right)^\top$, $k = 1, \cdots, K_u$. Consider the m-th iteration:

$$\beta_{kj}^{(m)} = \left(\frac{\sum_{i=1}^n \omega_{ik}^u x_{ij}^2}{\sum_{i=1}^n \omega_{ik}^u} + D_j^{(m-1)} \right)^{-1} \frac{\sum_{i=1}^n \omega_{ik}^u \left(x_{ij} \left(y_i - \mu_k^u - \boldsymbol{x}_i^\top \boldsymbol{\beta}_{k(-j)}^{(m-1)} \right) \right)}{\sum_{i=1}^n \omega_{ik}^u}, \tag{4.3}$$

where $D_j^{(m-1)} = \left(\frac{\lambda_1}{\left\| \left(\boldsymbol{\beta}_{\cdot j}^{(m-1)\top}, \boldsymbol{\alpha}_{\cdot j}^{(m-1)\top} \right) \right\|_2 + \zeta} - \frac{1}{a} \right)_+$, and ζ is a sufficiently small

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

constant (and set as 0.01 in our numerical studies). Similarly, we can conduct the iteration for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ as follows:

$$\alpha_{kj}^{(m)} = \frac{\sum_{i=1}^{n} \omega_{ik}^{l} \left(x_{ij} \left(y_i - \mu_k^l - \boldsymbol{x}_i^\top \left(\boldsymbol{\alpha}_{k(-j)}^{(m-1)} + \boldsymbol{\gamma}_k^{(m-1)} \right) \right) \right)}{\left(\frac{\sum_{i=1}^{n} \omega_{ik}^{l} x_{ij}^2}{\sum_{i=1}^{n} \omega_{ik}^{l}} + D_j^{(m-1)} \right) \sum_{i=1}^{n} \omega_{ik}^{l}}, \tag{4.4}$$

$$\gamma_{kj}^{(m)} = \frac{\sum_{i=1}^{n} \omega_{ik}^{l} \left(x_{ij} \left(y_i - \mu_k^l - \boldsymbol{x}_i^\top \left(\boldsymbol{\alpha}_{k}^{(m-1)} + \boldsymbol{\gamma}_{k(-j)}^{(m-1)} \right) \right) \right)}{\left(\frac{\sum_{i=1}^{n} \omega_{ik}^{l} x_{ij}^2}{\sum_{i=1}^{n} \omega_{ik}^{l}} + F_j^{(m-1)} \right) \sum_{i=1}^{n} \omega_{ik}^{l}}, \tag{4.5}$$

where $F_j^{(m-1)} = \left(\frac{\lambda_1}{\left\| \boldsymbol{\gamma}_{\cdot j}^{(m-1)} \right\|_2 + \zeta} - \frac{1}{a} \right)_+$. It is noted that the decoupling strategy in (2.2) can be satisfied by considering both A_j and B_j in (4.2) simultaneously. Specifically, $A_j > \lambda_1$ for a nonzero α_{kj}, $A_j \leqslant \lambda_1$ and $B_j > \lambda_2$ for a nonzero γ_{kj}. The Iterative Hierarchical Shooting Algorithm is summarized in Algorithm 2.

Here, $A_j^{(m-1)}$ and $B_j^{(m-1)}$ denote (4.2) with $\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}$ from the $(m-1)$-th iteration. The iteration is similar to the iterative group shooting algorithm (Yan and Huang, 2012) and additionally accommodates the nested structure, and it can be viewed as a special case of a block coordinate descent algorithm. Unlike the algorithm proposed by Yan and Huang (2012), this nested setup realizes a "think twice" mechanism rather than a "no drawback" approach (which means that regression coefficients, once zero, remain zero). In each iteration, the important variables at the upper level are re-evaluated within the important variable set at the lower level, and ζ can

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

Algorithm 2 Iterative Hierarchical Shooting Algorithm

Initialize: $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^{K_u \times p}, \boldsymbol{\alpha}^{(0)} \in \mathbb{R}^{K_l \times p}, \boldsymbol{\gamma}^{(0)} \in \mathbb{R}^{K_l \times p}$. Given $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$, $\boldsymbol{\omega}^u \in \mathbb{R}^{n \times K_u}, \boldsymbol{\omega}^l \in \mathbb{R}^{n \times K_l}, \boldsymbol{\mu}^u \in \mathbb{R}^{K_u}, \boldsymbol{\mu}^l \in \mathbb{R}^{K_l}$, tuning parameters λ_1, λ_2, regularization parameter a, and a sufficiently small constant ζ:
for $m = 1, \cdots, M$ **do**
  Input: $\boldsymbol{\beta}^{(m)}, \boldsymbol{\alpha}^{(m)}, \boldsymbol{\gamma}^{(m)}$.
  for $j = 1, \cdots, p$ **do**
    Calculate $A_j^{(m-1)}$ and $B_j^{(m-1)}$ following (4.2);
    if $A_j^{(m-1)} > \lambda_1$ **then**
      Update $\boldsymbol{\beta}_{\cdot j}^{(m)}$ and $\boldsymbol{\alpha}_{\cdot j}^{(m)}$ following (4.3) and (4.4), respectively;
    else
      $\boldsymbol{\beta}_{\cdot j}^{(m)} = \boldsymbol{\alpha}_{\cdot j}^{(m)} = 0$;
      if $B_j^{(m-1)} > \lambda_2$ **then**
        Update $\boldsymbol{\gamma}_{\cdot j}^{(m)}$ following (4.5);
      else
        $\boldsymbol{\gamma}_{\cdot j}^{(m)} = 0$;
      end if
    end if
  end for
  if convergence criterion is satisfied **then**
    Break;
  end if
  Output: $\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\alpha}^{(m+1)}, \boldsymbol{\gamma}^{(m+1)}$.
end for
return $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}$

be considered a "compensation signal", pushing the unselected variables to still be considered. It can help avoid information loss associated with the no-drawback approach. Once we have the parameter estimates, for a new sample, we can calculate the posterior probability that it belongs to a

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

specific (sub)group. Then, it can be assigned to the one with the highest posterior probability.

In most of the existing supervised heterogeneity analyses, the number of groups is a tuning parameter and is selected, for example, using AIC/BIC. In contrast, in the proposed analysis, the number of (sub)groups is user-defined, which can be based on analysis needs and/or data. It is also noted that there can be certain constraints. For example, to generate consistent estimation, sizes of the groups/subgroups cannot be too small. Additionally, as in hierarchical clustering, sometimes, some numbers of groups/subgroups cannot be realized. We suspect that it may be possible to revise AIC/BIC and "more objectively" select the number of (sub)groups. However, this will involve assigning different weights (associated with degrees of freedom) for different levels. This is beyond of the scope of this work and deferred to future research. With MCP, the regularization parameter can be data-dependently selected or prefixed.

5. Simulation

To complement the theoretical investigation, we conduct simulation to examine practical performance. The settings have been designed to closely resemble those in existing heterogeneity studies, to sufficiently demonstrate

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

both strengths and possible limitations of the proposed approach.

We set $K_l = 4$ and $K_u = 2$ and $p = 100$. We consider a balanced case with all the subgroups having 150 samples and an unbalanced case with $(190, 170, 130, 110)$ samples. For the subgroups, we consider three scenarios, which all have $s = 12$ variables with nonzero coefficients. Specifically,

- (S1)
$$\begin{aligned}
\boldsymbol{\beta}_1^l &= (\boldsymbol{\rho}_4, \frac{4}{3}\boldsymbol{\rho}_2, -\frac{1}{6}\boldsymbol{\xi}_2, \frac{1}{10}\xi, \frac{1}{8}\boldsymbol{\xi}_2, \frac{1}{7}\xi, \mathbf{0}_{88})^\top,\\
\boldsymbol{\beta}_2^l &= (\boldsymbol{\rho}_4, \frac{2}{3}\boldsymbol{\rho}_2, \frac{1}{7}\boldsymbol{\xi}_2, -\frac{1}{6}\xi, \frac{1}{10}\boldsymbol{\xi}_2, \frac{1}{8}\xi, \mathbf{0}_{88})^\top,\\
\boldsymbol{\beta}_3^l &= (-\boldsymbol{\rho}_4, -\frac{2}{3}\boldsymbol{\rho}_2, \frac{1}{8}\boldsymbol{\xi}_2, \frac{1}{7}\xi, -\frac{1}{6}\boldsymbol{\xi}_2, \frac{1}{10}\xi, \mathbf{0}_{88})^\top,\\
\boldsymbol{\beta}_4^l &= (-\boldsymbol{\rho}_4, -\frac{4}{3}\boldsymbol{\rho}_2, \frac{1}{10}\boldsymbol{\xi}_2, \frac{1}{8}\xi, \frac{1}{7}\boldsymbol{\xi}_2, -\frac{1}{6}\xi, \mathbf{0}_{88})^\top.
\end{aligned}$$

- (S2)
$$\begin{aligned}
\boldsymbol{\beta}_1^l &= (\boldsymbol{\rho}_5, \frac{4}{3}\boldsymbol{\rho}_3, -\frac{1}{6}\xi, \frac{1}{10}\xi, \frac{1}{8}\xi, \frac{1}{7}\xi, \mathbf{0}_{88})^\top,\\
\boldsymbol{\beta}_2^l &= (\boldsymbol{\rho}_5, \frac{2}{3}\boldsymbol{\rho}_3, \frac{1}{7}\xi_2, -\frac{1}{6}\xi, \frac{1}{10}\xi, \frac{1}{8}\xi, \mathbf{0}_{88})^\top,\\
\boldsymbol{\beta}_3^l &= (-\boldsymbol{\rho}_5, -\frac{2}{3}\boldsymbol{\rho}_3, \frac{1}{8}\xi, \frac{1}{7}\xi, -\frac{1}{6}\xi, \frac{1}{10}\xi, \mathbf{0}_{88})^\top,\\
\boldsymbol{\beta}_4^l &= (-\boldsymbol{\rho}_5, -\frac{4}{3}\boldsymbol{\rho}_3, \frac{1}{10}\xi, \frac{1}{8}\xi, \frac{1}{7}\xi, -\frac{1}{6}\xi, \mathbf{0}_{88})^\top.
\end{aligned}$$

- (S3)
$$\begin{aligned}
\boldsymbol{\beta}_1^l &= (\boldsymbol{\rho}_6, \frac{\xi}{4}, -\frac{\xi}{6}, \frac{\xi}{8}, -\frac{\xi}{10}, \frac{\xi}{12}, -\frac{1}{14}, \mathbf{0}_{88})^\top,\\
\boldsymbol{\beta}_2^l &= (\boldsymbol{\rho}_6, -\frac{\xi}{4}, \frac{\xi}{6}, -\frac{\xi}{8}, \frac{\xi}{10}, -\frac{\xi}{12}, \frac{\xi}{14}, \mathbf{0}_{88})^\top,\\
\boldsymbol{\beta}_3^l &= (-\boldsymbol{\rho}_6, \frac{\xi}{4}, -\frac{\xi}{6}, \frac{\xi}{8}, -\frac{\xi}{10}, \frac{\xi}{12}, -\frac{\xi}{14}, \mathbf{0}_{88})^\top,\\
\boldsymbol{\beta}_4^l &= (-\boldsymbol{\rho}_6, -\frac{\xi}{4}, \frac{\xi}{6}, -\frac{\xi}{8}, \frac{\xi}{10}, -\frac{\xi}{12}, \frac{\xi}{14}, \mathbf{0}_{88})^\top.
\end{aligned}$$

Here, for example, $\boldsymbol{\rho}_4$ denotes a vector of length 4 with all components equal to ρ. The relative signal strengths are controlled by ρ and ξ. More

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

specifically, ρ describes the upper-level effects, and the weaker heterogeneous signals corresponding to ξ are "truncated" at the upper levels. The heterogeneity structures and upper-level models (or $\boldsymbol{\beta}_k^u$'s) are derived based on the lower-level ones. As can be intuitively seen from the above settings, subgroups 1 and 2 form the first group, and the rest subgroups form the second group. Specifically, in S1 and S3, there are 6 important variables at the upper level, with $\boldsymbol{\beta}_1^u = (\boldsymbol{\rho}_6, \mathbf{0}_{94})^\top$. In S2, there are 8 important variables, with $\boldsymbol{\beta}_2^u = (-\boldsymbol{\rho}_8, \mathbf{0}_{92})^\top$. We consider $\rho = \xi = 1$, $\rho = \xi = 1.5$, and $\rho = 1.5$ and $\xi = 1$. We further set the intercept terms as $\boldsymbol{\mu}^l = (4, \frac{4}{3}, -\frac{4}{3}, -4)^\top$. $\boldsymbol{x}_i$'s are generated from a multivariate standard Gaussian distribution. ϵ_i's are generated from a Gaussian distribution $\mathcal{N}(0, 0.5^2)$. We have also experimented with a few other ways of generating the covariates and random errors and made similar observations.

Our literature review does not suggest any approaches that closely align with ours. We consider the following single-level approaches (which conduct analysis at different levels separately) and two-step approaches as relevant competitors. Specifically, for the former, we consider: (i) Sparse Finite Mixture Model **(S-FMR)**, which is implemented using R package "flexmix". This approach conducts FMR analysis and applies adaptive Lasso for sparsity. The numbers of groups and subgroups are set as the true values. This

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

approach is considered as the baseline. (ii) Group Finite Mixture Model **(G-FMR)**: This approach applies group Lasso to simultaneously select variables across multiple groups/subgroups. It is based on the development in Hui et al. (2015), with the numbers of groups and subgroups set as the true values. (iii) MCP Group Finite Mixture Model **(MG-FMR)**: this approach is similar to (ii), with group Lasso replaced by group MCP (which may have more favorable performance).

For two-step approaches, we consider: (iv) Ascending order **(Order-A)**. First, we perform clustering at the upper level. Then, within each group, we perform clustering to generate subgroups. The model and penalized estimation in each step follow (iii). (v) Descending order **(Order-D)**: its strategy is similar to (iv), but first generates subgroups and then merge them to generate upper-level groups. Different from the proposed approach, (iv) and (v) generate sample-nested structures. We acknowledge that there can be other alternatives and the above may be more relevant and can provide direct insights into operating characteristics of the proposed approach.

We adopt the following metrics for evaluation: (i) True and false positive rates (TPR and FPR), where $\text{TPRL} = \frac{1}{K_l}\sum_{k=1}^{K_l}\frac{\sum_{j=1}^{p} I\left(\beta_{kj}^{l*}\neq 0, \widehat{\beta}_{kj}^{l}\neq 0\right)}{\sum_{j=1}^{p} I\left(\beta_{kj}^{l*}\neq 0\right)}$, $\text{FPRL} = \frac{1}{K_l}\sum_{k=1}^{K_l}\frac{\sum_{j=1}^{p} I\left(\beta_{kj}^{l*}= 0, \widehat{\beta}_{kj}^{l}\neq 0\right)}{\sum_{j=1}^{p} I\left(\beta_{kj}^{l*}= 0\right)}$ and $\text{TPRU} = \frac{1}{K_u}\sum_{k=1}^{K_u}\frac{\sum_{j=1}^{p} I\left(\beta_{kj}^{u*}\neq 0, \widehat{\beta}_{kj}^{u}\neq 0\right)}{\sum_{j=1}^{p} I\left(\beta_{kj}^{u*}\neq 0\right)}$, $\text{FPRU} = \frac{1}{K_u}\sum_{k=1}^{K_u}\frac{\sum_{j=1}^{p} I\left(\beta_{kj}^{u*}= 0, \widehat{\beta}_{kj}^{u}\neq 0\right)}{\sum_{j=1}^{p} I\left(\beta_{kj}^{u*}= 0\right)}$ for the **L**ower and **U**pper levels, re-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

spectively. (ii) Mean squared errors (MSE) for the two levels defined as MSEL $= \sum_{k=1}^{K_l} \left\| \widehat{\boldsymbol{\beta}}_k^l - \boldsymbol{\beta}_k^{*l} \right\|_2$ and MSEU $= \sum_{k=1}^{K_u} \left\| \widehat{\boldsymbol{\beta}}_k^u - \boldsymbol{\beta}_k^{*u} \right\|_2$. (iii) Rand Index (RI) and Adjusted Rand Index (ARI) for the two levels, which measure grouping accuracy.

The proposed analysis is computationally very affordable. The analysis of one replicate takes less than one minute on a standard laptop. For each setting, we generate 100 replicates. The results for the unbalanced subgroup design with $\rho = 1.5$ and $\xi = 1$ are presented in Table 1. Those for the other settings are presented in Supplementary Materials.

In FMR analysis, estimation accuracy of regression coefficients is tightly connected to clustering performance. This is also observed in our simulation study. For subgroups, S-FMR fails to identify important variables, leading to poor subgrouping results. High FPR and low RI/ARI values are observed. G-FMR can distinguish unimportant variables more effectively compared to S-FMR, as shown by significant reduction in FPR. However, with its tendency of excessive shrinkage, G-FMR suffers from high bias, leading to low TPR, RI/ARI, and high MSE values. MG-FMR demonstrates notable superiority over the previous two methods, which is likely caused by the advantage of MCP in estimation – this subsequently leads to improved variable selection and grouping performance.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

Table 1: Simulation results for the unbalanced subgroup design. $\rho = 1.5$ and $\xi = 1$. In each cell: mean (sd).

		TPRL	FPRL	MSEL	RIL	ARIL	TPRU	FPRU	MSEU	RIU	ARIU
S1	Proposed	0.918(0.080)	0.033(0.019)	0.432(0.084)	0.845(0.012)	0.601(0.029)	1(0)	0(0)	0.225(0.096)	0.844(0.016)	0.688(0.033)
	S-FMR	0.792(0.066)	0.485(0.088)	17.546(17.823)	0.692(0.023)	0.228(0.053)	-	-	-	-	-
	(separately)	-	-	-	-	-	0.964(0.098)	0.225(0.098)	2.100(5.442)	0.823(0.049)	0.647(0.098)
	G-FMR	0.698(0.072)	0.132(0.062)	13.334(16.014)	0.675(0.036)	0.199(0.119)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0.001(0.002)	1.642(0.345)	0.810(0.022)	0.621(0.044)
	MG-FMR	0.750(0.092)	0.062(0.050)	5.810(4.698)	0.760(0.065)	0.442(0.120)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0.001(0.003)	0.223(0.096)	0.838(0.019)	0.674(0.038)
	Order-A	0.806(0.055)	0.054(0.019)	0.596(0.195)	0.788(0.039)	0.457(0.099)	1(0)	0.001(0.002)	0.250(0.116)	0.839(0.017)	0.677(0.034)
	Order-D	0.753(0.079)	0.049(0.082)	6.364(13.429)	0.776(0.076)	0.470(0.151)	0.995(0.025)	0.055(0.076)	1.883(5.073)	0.839(0.059)	0.677(0.118)
S2	Proposed	0.885(0.084)	0.049(0.018)	0.451(0.098)	0.849(0.014)	0.611(0.035)	1(0)	0(0)	0.188(0.079)	0.828(0.019)	0.656(0.038)
	S-FMR	0.741(0.076)	0.518(0.084)	9.379(12.374)	0.696(0.036)	0.221(0.092)	-	-	-	-	-
	(separately)	-	-	-	-	-	0.981(0.131)	0.206(0.073)	1.705(4.039)	0.818(0.048)	0.633(0.099)
	G-FMR	0.583(0.062)	0.053(0.064)	10.300(12.456)	0.738(0.066)	0.339(0.159)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0(0)	2.016(3.039)	0.783(0.066)	0.567(0.088)
	MG-FMR	0.664(0.113)	0.049(0.044)	3.847(4.597)	0.773(0.062)	0.464(0.106)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0(0)	0.199(0.087)	0.826(0.018)	0.651(0.036)
	Order-A	0.738(0.064)	0.054(0.015)	0.627(0.123)	0.797(0.038)	0.482(0.095)	1(0)	0(0)	0.191(0.079)	0.829(0.021)	0.656(0.043)
	Order-D	0.705(0.106)	0.042(0.027)	2.077(3.205)	0.808(0.064)	0.535(0.116)	1(0)	0.043(0.023)	0.439(0.814)	0.841(0.024)	0.682(0.048)
S3	Proposed	0.807(0.090)	0.052(0.027)	0.517(0.303)	0.854(0.045)	0.632(0.100)	1(0)	0(0)	0.140(0.068)	0.812(0.025)	0.623(0.050)
	S-FMR	0.740(0.082)	0.525(0.089)	12.519(10.182)	0.681(0.014)	0.189(0.037)	-	-	-	-	-
	(separately)	-	-	-	-	-	0.976(0.149)	0.207(0.075)	1.579(4.480)	0.797(0.050)	0.591(0.104)
	G-FMR	0.562(0.073)	0.045(0.031)	12.813(5.001)	0.678(0.035)	0.300(0.055)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0(0)	1.149(3.034)	0.759(0.056)	0.519(0.112)
	MG-FMR	0.629(0.068)	0.074(0.045)	9.344(4.306)	0.690(0.039)	0.334(0.068)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0(0)	0.154(0.059)	0.818(0.023)	0.635(0.046)
	Order-A	0.741(0.062)	0.049(0.018)	0.548(0.169)	0.817(0.048)	0.530(0.124)	1(0)	0(0)	0.148(0.054)	0.819(0.021)	0.636(0.043)
	Order-D	0.649(0.089)	0.042(0.038)	7.354(7.661)	0.756(0.081)	0.451(0.148)	1(0)	0.058(0.031)	1.802(1.976)	0.814(0.025)	0.627(0.050)

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

Among those based on group MCP penalization (MG-FMR, Order-A, and Order-D), the approach that first performs upper-level grouping and then conducts within-group analysis has better performance. For example, under S1 in Table 1, it achieves TPRL = 0.806, FPRL = 0.054, and RIL = 0.788. The proposed method consistently demonstrates the best performance. For example, under S1, S2, and S3 in Table 1, it achieves TPRL = 0.918, 0.885, and 0.807, respectively, while maintaining low FPR and high RI (≈ 0.85) and ARI values. Particularly, in detecting weak signals at the lower level, it excels with high TPR and low MSE values.

When strong signals are only present at the upper level, the methods with group penalization also perform well, and the proposed method has performance comparable to the best. An interesting observation is that, by taking the results with $\rho = \xi = 1$ as a "baseline", an increase of the overall signal level ($\rho = \xi = 1.5$) leads to an improvement in variable selection for all the methods. However, when the relative difference is larger ($\rho = 1.5, \xi = 1$), our method maintains consistent performance, whereas variable selection performance of the other methods deteriorates. Also note that in regular FMR analysis, it is of interest to examine performance when the number of groups (subgroups) is mis-determined. This is not as applicable to the proposed analysis – when the number of groups (subgroups) changes,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

the level of analysis also changes, and so does the "true" model.

6. Data Analysis

We analyze breast cancer (BRCA) data collected by The Cancer Genome Atlas (TCGA) project and refer to many published studies for information on TCGA. The response variable of interest is the ratio between "Positive Finding Lymph Node Hematoxylin and Eosin Staining Microscopy Count" and "Lymph Node Examined Number". It is an indicator of tumor burden, disease progression, and recurrence risk, and can describe the degree of treatment (Braunstein et al., 2017). It has been suggested that this ratio may be correlated with gene expressions. In genetic studies of breast cancer (and many other cancers), heterogeneity has been suggested, and heterogeneity analysis has been shown to have important implications for disease prevention, treatment, and management (Reis-Filho and Pusztai, 2011).

We adopt a dynamic smoothing approach based on empirical Bayes to correct the response ratio. Let "Positive Finding Lymph Node Hematoxylin and Eosin Staining Microscopy Count" of the i-th sample be denoted as "positive$_i$", "Lymph Node Examined Number" as "test$_i$", and ratio$_i$ $= \frac{\text{positive}_i}{\text{test}_i}$. Define $a_i = \frac{\sum_{i=1}^{n} \text{test}_i}{n\cdot(\text{test}_i+1)}$, $b_i = \frac{\sum_{i=1}^{n} \text{positive}_i}{\sum_{i=1}^{n} \text{test}_i} \cdot a_i$. Then, the adjusted ratio is $\widetilde{\text{ratio}}_i = \frac{\text{positive}_i+b_i}{\text{test}_i+a_i}$. When the numbers of tests are low,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

directly calculating the ratios is prone to extreme values of 0 or 1, which may lead to instability in subsequent analysis. This method effectively mitigates the occurrence of extreme values and has been frequently adopted. It reflects the Bayesian principle of "the more data available, the less influence the prior has", ensuring that for the samples with more tests, the ratio estimates rely more on the actual data, while for the samples with fewer tests, the ratio estimates are more influenced by global information. We then retain the samples with $\text{test}_i > 1$ and non-missing values and perform the transformation $\log\left(\frac{\widetilde{\text{ratio}_i}}{1-\widetilde{\text{ratio}_i}}\right)$ following He et al. (2021). For the gene expression data, we focus on the "breast cancer" pathway (hsa05224 in KEGG), considering a limited sample size. With the above preprocessing, we have 139 gene expression measurements for 841 samples.

With the proposed approach, the numbers of groups/subgroups need to be specified. As discussed above, in a sense, there is no wrong or right specification. As observed in unsupervised hierarchical clustering and some other analysis, some numbers and their corresponding clustering structures may be less satisfactory. Taking into account the sample size, biological (especially subtyping) information about the disease, and desired grouping results, we choose five subgroups and three groups. We have examined other options, for example four or six subgroups, and observed inferior results.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

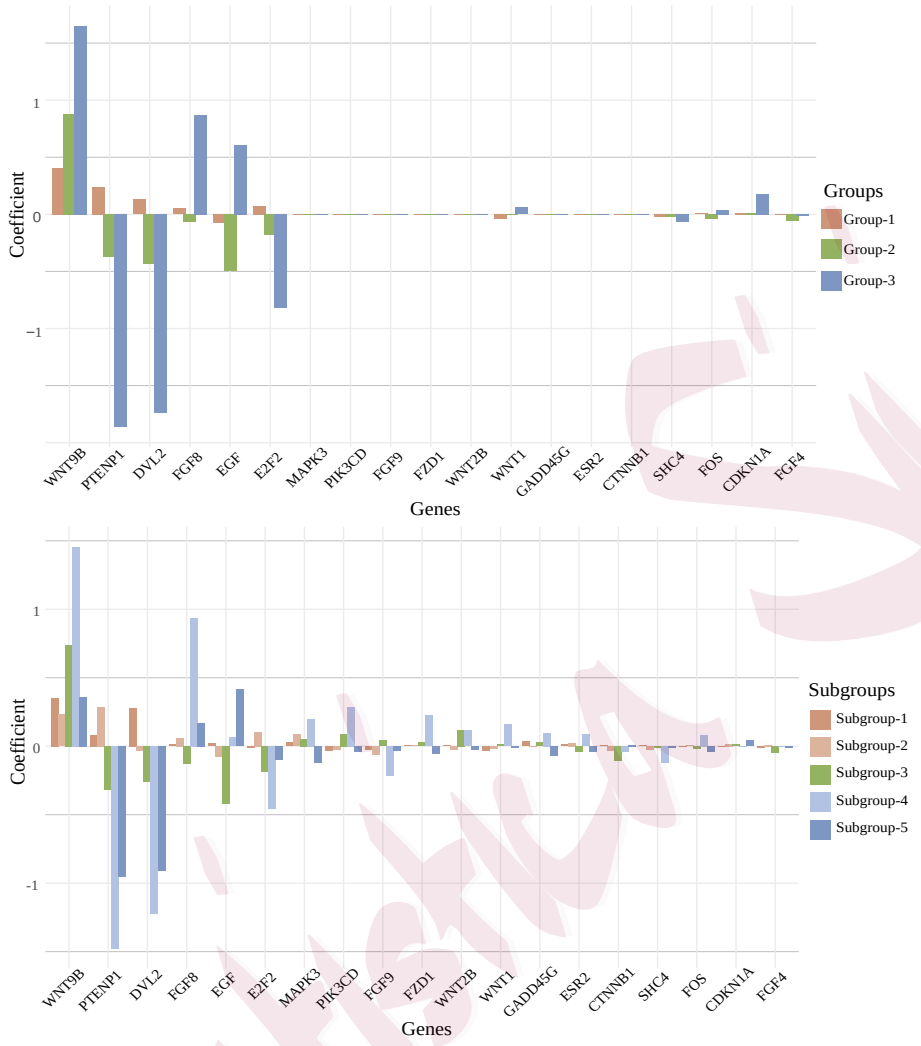

Figure 2: Analysis using the proposed approach: coefficients of the identified genes, sorted in descending order based on magnitudes of the coefficients for the subgroups. Top (bottom): upper (lower) level.

The five subgroups have sizes of 241, 289, 176, 57, and 78, and 19 genes are identified. The three groups have sizes of 552, 146, and 143, and 11 (out

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

of those 19) genes are identified. The identified genes are presented in Figure 2. Some interesting findings are made. First, even though not reinforced, the samples at the two levels roughly satisfy a hierarchical structure: the first two subgroups overlap largely with the first group (orange in Figure 2), the last two subgroups overlap largely with the second group (blue in Figure 2), and the third subgroup overlaps largely with the third group (green in Figure 2). More specifically, the samples in subgroup-1 and -2, with a total size of 530, belong almost exclusively to group-1. Those in subgroup-4 and -5, with a total size of 135, belong almost exclusively to group-3. The correspondence between group-2 and subgroup-3 is slightly ambiguous, with subgroup-3 also containing a small number of samples mainly from group-1. This may be caused by some small heterogeneous effects – the effects in group-1 are small in general. Second, for the important genes identified at the upper level, their estimates are similar for the two levels. Here, it is noted that the blue subgroup-4 and -5 in Figure 2 have significant differences in the magnitudes of some regression coefficients. However, the directions (signs) of these coefficients are highly consistent, which may lead to them identified in the same group at a higher level. Additionally, the important genes identified only at the lower level have weaker signals, and their estimates differ significantly across the subgroups.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

For the upper level, genes WNT9B, PTENP1, DVL2, FGF8, EGF, and E2F2 are clearly identified. They are involved in key signaling pathways that regulate cell proliferation, differentiation, apoptosis, and metastasis, which are critical factors influencing tumor aggressiveness and lymphatic spread. WNT9B is part of the Wnt signaling pathway, which plays a role in the determination and proliferation of cell fate. Its association with LNR implies that it may influence the ability of tumor cells to invade lymphatic vessels (Lu et al., 2021). Dishevelled-2 (DVL2) is another member of the Wnt signaling pathway, mediating signal transduction. Its association with the ratio suggests that it may modulate activity of the Wnt pathway, affecting tumor invasiveness (Sharma et al., 2018). Variations in PTENP1 affect PTEN levels, with PTEN being a well-known tumor suppressor involved in cell cycle regulation, thus altering tumor growth and metastatic potential (Ghafouri-Fard et al., 2022). FGF8 is involved in cell growth and angiogenesis, while EGF promotes cell proliferation and survival, potentially facilitating tumor expansion and lymph node involvement (Masuda et al., 2012). E2F2 regulates genes essential for cell cycle progression.

Similar insightful findings are made at the lower level. Genes WNT2B and WNT1 can influence overall activity of the Wnt signaling pathway, and their effects can accumulate and significantly affect tumor behaviors (Xu

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

et al., 2020). MAPK3, a key member of the MAPK signaling pathway, may exert effects on cell signal transduction even with minor changes. And PIK3CD indicates different degrees of activity of the PI3K/AKT pathway, potentially related to subgroup-specific molecular mechanisms influencing tumor invasiveness. These two pathways play important roles in treatment of various cancers (Ruchi Sharma et al., 2017). FGF8 and FGF9 may exhibit different affinities across subtypes, leading to variations in activated signaling pathways and biological effects.

To get some additional insight, in Figure S2 (Supplementary Materials), we compare overall survival across the subgroups and groups. Log-rank tests lead to p-values 1×10^{-4} for the upper level and 4×10^{-5} for the lower level. In Figure S3 (Supplementary Materials), we compare lymph node positive ratio across the subgroups and groups. Significant differences in mean and variation are observed. Such differences suggest that the identified heterogeneity can have meaningful biomedical implications.

Data is analyzed using the alternatives, which lead to significantly different results. S-FMR identifies significantly different variables across the subgroups (details omitted). It leads to relatively balanced subgroups with sizes of 141, 135, 193, 173, and 199, and upper-level groups with sizes of 284, 389, and 168. MG-FMR identifies subgroups with sizes of 225, 307,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

177, 43, and 89, and 16 genes are identified. The upper-level groups have sizes of 544, 152, and 142, and 11 genes are identified. The subgroups are approximately nested in the groups. Among the 11 genes identified at the upper level, 8 overlap with those identified by the proposed approach. Genes ERBB2, WNT7B, and DDB2 are newly identified, FOS, WNT1, and SHC4 are missed. Among the 16 genes identified at the lower level, only 4 (PTENP1, FGF8, EGF, and FGF4) overlap with those at the upper level – this is significantly different from our proposed analysis. More results are presented in Supplementary Materials Figure S1. It is interesting to see that subgroup-3 and its counterparts in the other methods overlap significantly. They include samples with a very low proportion of lymph node positivity (Figure S3 in Supplementary Materials). G-FMR, Order-A, and Order-D lead to results inferior to MG-FMR, and the details are omitted.

7. Extension: Modeling Homogeneity in Effects Across Levels

With the proposed approach, the sets of important variables identified at different levels have a nested structure. However, there is no special attention to the estimated effects of the same identified variables at different levels. Consider for example the case of linear regression. For the important variables identified at the higher level, if they are orthogonal to those ad-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

ditionally identified at the lower level, then their estimates are the same at the two levels. Here, we explore modeling homogeneity in estimated effects, under which estimates at multiple levels are better aligned in magnitude. The scheme of this estimation is presented in Figure S4 (Supplementary Materials). For this purpose, an additional penalty term is introduced. The proposed approach and its theoretical properties are described in Section S3 of Supplementary Materials. We have experimented with a few simulations and found that this approach has satisfactory identification, grouping, and estimation performance, in a way similar to the above. In addition, when there is a strong alignment in the magnitudes of important effects, it achieves superior performance. Details are omitted here.

8. Discussion

For supervised heterogeneity analysis, we have developed a simultaneous analysis at multiple levels, where the sets of identified important effects have a nested structure. A novel penalization approach has been developed, with its theoretical and numerical properties rigorously established. We have also considered an extension, under which there is a better alignment in the magnitudes of estimated effects. Overall, the proposed approach is conceptually and statistically sound, and as shown in the numerical studies,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

can provide a practically useful tool.

As elaborated above, it can be of interest to conduct multi-level supervised heterogeneity analysis. Although significant advancements have been made in this study and some others, it is still relatively underdeveloped compared to under the unsupervised paradigm. For example, it can be of interest to "combine" the existing techniques to achieve nested structures in both identified important effects and samples. It can also be of interest to combine the proposed nested structure with the penalized fusion technique, which has been a popular alternative to the FMR technique. In hierarchical clustering and other multi-level analysis, the numbers of groups are often specified in a heuristic way. This also applies to the proposed analysis. In general, it is of interest to more deeply examine the numbers of groups/subgroups. Last but not least, more applications may be pursued.

Supplementary Material

The online Supplementary Material contains additional theoretical (referenced in Sections 3), numerical (referenced in Section 5 and Section 6), and methodological (referenced in Section 7) developments, which is available on the Statistica Sinica website.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

Acknowledgments

We thank the editors and reviewers for their careful review and insightful comments. This work was supported by the National Natural Science Foundation of China No.12571298, Fundamental Research Funds for the Central Universities, NIH CA204120, and NSF 220968.

References

Braunstein, L. Z., A. G. Taghian, A. Niemierko, L. Salama, A. Capuco, J. R. Bellon, J. S. Wong, R. S. Punglia, S. M. MacDonald, and J. R. Harris (2017). Breast-cancer subtype, age, and lymph node status as predictors of local recurrence following breast-conserving therapy. *Breast cancer research and treatment 161*, 173–179.

Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*(1), 101–148.

Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory 57*(8), 5467–5484.

Ghafouri-Fard, S., T. Khoshbakht, B. M. Hussen, M. Taheri, and N. Akbari Dilmaghani (2022). A review on the role of ptenp1 in human disorders with an especial focus on tumor suppressor role of this lncrna. *Cancer Cell International 22*(1), 1–12.

Hao, B., W. W. Sun, Y. Liu, and G. Cheng (2018). Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research 18*(217), 1–58.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

REFERENCES

He, B., T. Zhong, J. Huang, Y. Liu, Q. Zhang, and S. Ma (2021). Histopathological imaging-based cancer heterogeneity analysis via penalized fusion with model averaging. *Biometrics 77*(4), 1397–1408.

Huang, J., P. Breheny, and S. Ma (2012). A selective review of group selection in high-dimensional models. *Statistical Science 27*(4), 481–499.

Hui, F. K., D. I. Warton, and S. D. Foster (2015). Multi-species distribution modeling using penalized mixture of regressions. *The Annals of Applied Statistics 9*(2), 866–882.

Khalili, A. and J. Chen (2007). Variable selection in finite mixture of regression models. *Journal of the american Statistical association 102*(479), 1025–1038.

Khalili, A. and S. Lin (2013). Regularization in finite mixture of regression models with diverging number of parameters. *Biometrics 69*(2), 436–446.

Li, R., Q. Zhang, and S. Ma (2026). Regulation-incorporated gene expression network-based heterogeneity analysis. *Statistica Sinica 36*, 373–394.

Lu, S., E. Yakirevich, D. Yang, Y. Xiao, L. J. Wang, and Y. Wang (2021). Wnt family member 9b (wnt9b) is a new sensitive and specific marker for breast cancer. *The American journal of surgical pathology 45*(12), 1633–1640.

Ma, S. and J. Huang (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association 112*(517), 410–423.

Masuda, H., D. Zhang, C. Bartholomeusz, H. Doihara, G. N. Hortobagyi, and N. T. Ueno

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

(2012). Role of epidermal growth factor receptor in breast cancer. *Breast cancer research and treatment 136*, 331–345.

McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2019). Finite mixture models. *Annual review of statistics and its application 6*, 355–378.

Qin, X., X. Liu, S. Ma, and M. Wu (2024). Supervised bayesian joint graphical model for simultaneous network estimation and subgroup identification. *arXiv preprint arXiv:2403.19994* .

Reis-Filho, J. S. and L. Pusztai (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet 378*(9805), 1812–1823.

Ren, M., Q. Zhang, S. Zhang, T. Zhong, J. Huang, and S. Ma (2022). Hierarchical cancer heterogeneity analysis based on histopathological imaging features. *Biometrics 78*(4), 1579–1591.

Ren, M., S. Zhang, Q. Zhang, and S. Ma (2022). Gaussian graphical model-based heterogeneity analysis via penalized fusion. *Biometrics 78*(2), 524–535.

Ruchi Sharma, V., G. Kumar Gupta, A. K Sharma, N. Batra, D. K Sharma, A. Joshi, and A. K Sharma (2017). Pi3k/akt/mtor intracellular pathway and breast cancer: factors, mechanism and regulation. *Current pharmaceutical design 23*(11), 1633–1638.

Sharma, M., I. Castro-Piedras, G. E. Simmons Jr, and K. Pruitt (2018). Dishevelled: A masterful conductor of complex wnt signals. *Cellular signalling 47*, 52–64.

Sun, Y., Z. Luo, and X. Fan (2022). Robust structured heterogeneity analysis approach for

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

high-dimensional data. *Statistics in Medicine 41*(17), 3229–3259.

Tang, X., F. Xue, and A. Qu (2021). Individualized multidirectional variable selection. *Journal of the American Statistical Association 116*(535), 1280–1296.

Wang, W. and L. Su (2021). Identifying latent group structures in nonlinear panels. *Journal of Econometrics 220*(2), 272–295.

Xu, X., M. Zhang, F. Xu, and S. Jiang (2020). Wnt signaling in breast cancer: biological mechanisms, challenges and opportunities. *Molecular cancer 19*(1), 165.

Yan, J. and J. Huang (2012). Model selection for cox models with time-varying coefficients. *Biometrics 68*(2), 419–428.

Zhang, Q., S. Zhang, J. Liu, J. Huang, and S. Ma (2016). Penalized integrative analysis under the accelerated failure time model. *Statistica Sinica 26*(2), 493–508.

Ruiyue Wang, School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China; Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, U.S.A.

E-mail: wangruiyue21@mails.ucas.ac.cn

Sanguo Zhang, School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China.

E-mail: sgzhang@ucas.ac.cn

Shuangge Ma, Department of Biostatistics, Yale School of Public Health, New Haven, Connecti-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0147

cut, U.S.A.

E-mail: shuangge.ma@yale.edu

Statistica Sinica