BIOMETRICS 000, 1–26 DOI: 000
000 0000

Subgroup identification and membership prediction

Lu Chen,[1], Xuerong Chen,[1,*] , Xinzhou, Guo[2] and Yi Li[3]

[1]Center of Statistical Research, Southwestern University of Finance and Economics

[2]Department of Mathematics, The Hong Kong University of Science and Technology

[3]Department of Biostatistics, University of Michigan

email: chenxuerong@swufe.edu.cn

SUMMARY: In clinical trials, a treatment rarely benefits every patient, underscoring the need to identify subgroups that are more likely to respond. Traditional subgroup analysis approaches, including finite mixture and threshold models, often rely on stringent distributional assumptions and prespecified subgroup structures that may be unrealistic in practice. Moreover, the resulting subgroups can be difficult to interpret and may not generalize well to new patients. To address these challenges, we propose a new least-squares regression framework that accommodates flexible subgroup structure in heterogeneous data. Our model is distribution-free and allows subgroup membership to depend on covariates, while permitting both the number and the organization of coefficient groups to vary across covariates. Building on regularization, we develop a computationally efficient procedure to detect subgroup structure in linear regression coefficients and then use a support vector machine to recover the corresponding partitions, enabling subgroup membership prediction for future individuals. Relative to pairwise fused regularization, our approach substantially reduces computational complexity. We also establish theoretical guarantees for estimation of group-specific parameters and recovery of the underlying partitions. Simulation studies and a real-data application illustrate the practical effectiveness of the proposed method.

KEY WORDS: Group membership prediction; Regularization method; Subgroup identification.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Subgroup analysis, often used to extract as much information as possible while minimizing effort and cost, has attracted substantial attention because of its broad applications in clinical trials (Assmann et al. 2000, Alosh et al. 2015), personalized medicine (Dahabreh et al., 2016), and personalized marketing (Guelman et al., 2015). Its value is well illustrated by the case of panitumumab for metastatic colorectal cancer: the initial application was rejected by the European Medicines Agency (EMA) due to insufficient overall efficacy, but subsequent re-analysis by Peeters et al. (2010) and Peeters et al. (2015) established efficacy in the wild-type KRAS patients subgroup, ultimately leading to EMA approval for this distinct population.

Subgroup identification involves two key tasks: determining the number of subgroups and assigning each subject to a subgroup. Early work largely focused on analyses with pre-specified subgroups (Amado et al., 2008; Peeters et al., 2015) , whereas more recent research has emphasized discovering latent subgroups in heterogeneous data. One line of work models the population as a mixture of subgroups, each characterized by distinct parameter values, and proceeds via finite mixture modeling (Everitt and Hand, 1981). For example, Shen and He (2015) proposed a structured logistic-normal mixture model to test for subgroup existence under homogeneous subgroup variances, and Shen et al. (2017) extended this framework to allow heterogeneous subgroup variances (for additional details, see Zhou et al. 2022). A second line of work targets subgroups induced by structural breaks in covariate effects among observed subjects, using covariate threshold models, single-index threshold models (Zhang et al., 2022), and multi-threshold change-plane models (Li et al., 2021). A third line of work relies on regularization to detect subgroups. A common strategy penalizes pairwise differences between individual-specific regression coefficient vectors to control model complexity and encourage fusion. For instance, Ma and Huang (2017) introduced a concave penalty on intercept differences, and Tang et al. (2021) developed a heterogeneous modeling framework using a cluster-centered lasso penalty to group covariate effects, see He et al. (2023) for

details. In all these methods, the penalty function is central: it shrinks parameter differences toward zero, promotes coefficient fusion, and enables automatic subgroup clustering.

These existing methods have limitations. In particular, an erroneous pre-specification of the subgroup number or structure can introduce substantial bias and lead to adverse conclusions. In the panitumumab trials, defining subgroups solely by wild-type KRAS status, as in (Amado et al., 2008; Peeters et al., 2015), would have missed the critical finding that female patients aged $\geqslant 65$ derived no benefit (Luo and Guo, 2023). This illustrates the risk of oversimplifying a pre-specified subgroup structure. Finite mixture models require specification of the outcome distribution and a parametric model for the latent subgroup indicator, and can be biased under model misspecification. Covariate threshold and single-index threshold models effectively pre-determine the number of groups, and threshold models typically constrain subgroup boundaries across covariates to share the same linear form. Such linearity assumptions can be overly restrictive and are often violated in practice; for example, as shown in our real-data analysis, subgroup boundaries in the panitumumab trial are distinctly nonlinear. Regularization-based methods, while useful for discovery, generally cannot predict subgroup membership for new individuals. In addition, the pairwise fused penalty method of Ma and Huang (2017) can be computationally burdensome for large samples, with complexity $O(n^2)$. Most existing approaches also impose a common subgroup structure across all covariate effects. These limitations motivate the development of a more flexible, data-driven, and computationally efficient strategy for detecting latent subgroups, guided by clinically meaningful covariates.

In this paper, we propose a novel heterogeneous regression model for continuous outcomes with an unknown group structure. For greater flexibility, we allow regression coefficients to vary across blocks defined by partitioning the support of selected covariates, where the partitions are determined by unknown boundary curves. As a result, the latent grouping of regression coefficients is governed by hidden partitions of the covariate support. While related to existing models, including

the covariate multi-threshold model (Zhang et al., 2014), multi-threshold change-plane model (Li et al., 2021), two-way truncated linear regression model (Teng and Zhang, 2024), and varying coefficient model (Hastie and Tibshirani, 1993), our framework differs in a key way: it does not assume a particular outcome distribution, pre-specify the number of groups, or restrict group boundaries to be linear functions of covariates. To fit the model, we develop a computationally efficient three-stage procedure. In the first stage, we enforce clustering among individuals within the same neighborhood (with respect to selected covariates) by penalizing pairwise differences in their regression coefficients, thereby detecting latent group structure. In the second stage, we assign labels to the estimated groups from the first stage, and recover subgroup partitions via a support vector machine, enabling prediction of subgroup membership for future samples. Finally, to improve coefficient estimation, we propose a post-group-recovery step that re-estimates regression coefficients given the recovered group structure.

In summary, our main contributions are: (a) We propose a novel and general framework that encompasses several established models (e.g., the threshold model) as special cases. (b) Our method addresses key limitations of existing approaches by avoiding strict distributional assumptions, a pre-specified number of groups, and linear boundary constraints, while allowing subgroup structures to vary across covariates. (c) Our procedure is computationally more efficient than the pairwise fused penalty method of Ma and Huang (2017) and, unlike existing regularization-based methods, enables prediction of subgroup membership for new individuals based on the recovered partitions. (d) A re-analysis of the panitumumab trial data identifies a novel subgroup that differentiates non-responders due to lack of efficacy from those who experience adverse effects.

The paper is organized as: In Section 2, we introduce the proposed model. Section 3 presents the methodology, including subgroup detection method, partitions estimation and Post-group-recovery estimators. We then state the theoretical properties of the proposed approach in Section 4. Section 5 and Section 6 illustrates the proposed method through the numerical simulation studies and real

data analysis. The paper is concluded with some discussions in Section 7. All technical details and additional numerical results are given in the Supplementary Material.

2. Data and Model

Consider a random sample $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{U}_i), i = 1, \ldots, n$, where Y_i is the outcome variable, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are q- and p-dimensional vectors of covariates respectively. Here, $\boldsymbol{X}_i$ represents covariates with homogeneous effects on the response, such as baseline characteristics (e.g., age, gender), while $\boldsymbol{Z}_i$ captures covariates with potentially heterogeneous effects, like treatment status or dosage in a medical study. Let $\boldsymbol{U}_i$ be an r-dimensional vector, which may share common variables with $(\boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$. For example, $\boldsymbol{U}_i$ is subset of $(\boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$. Consider following heterogeneous linear model

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}^0 + \boldsymbol{Z}_i^T \boldsymbol{\alpha}_i^0(\boldsymbol{U}_i) + \epsilon_i, \tag{1}$$

where $\boldsymbol{\beta}$ is a q-dimensional vector of regression coefficients, $\boldsymbol{\alpha}_i(\boldsymbol{U}_i)$ is a p-dimensional vector of subject-specific regression coefficients depending on $\boldsymbol{U}_i$, ϵ_i is a random error term with mean 0 and variance σ_ϵ^2, and the superscript '0' (also hereafter) denotes the true values. We further assume that there exists a latent group structure in $\boldsymbol{\alpha}_i(\boldsymbol{U}_i)$ of the following form

$$\boldsymbol{\alpha}_i(\boldsymbol{U}_i) = \left(\sum_{m=1}^{M_1^0} \theta_{1m}^0 I(\boldsymbol{U}_i \in \mathcal{U}_{1m}^0), \ldots, \sum_{m=1}^{M_p^0} \theta_{pm}^0 I(\boldsymbol{U}_i \in \mathcal{U}_{pm}^0) \right)^T, \tag{2}$$

where $\{\mathcal{U}_{gm}^0, m = 1, \ldots, M_g^0\}$ is a latent partition of $\mathcal{U}$, $g = 1, \ldots p$, in which $\mathcal{U}$ denote the support for random variable $\boldsymbol{U}$. For subjects whose $\boldsymbol{U}_i$ falls into the partition block $\mathcal{U}_{gm}^0$, the indicator function $I(\boldsymbol{U}_i \in \mathcal{U}_{gm}^0)$ equals 1, and the g-th component of $\boldsymbol{\alpha}_i^0(\boldsymbol{U})$ takes the value θ_{gm}^0. Namely, all individuals whose $\boldsymbol{U}_i$ falls within $\mathcal{U}_{gm}^0$ share the same regression coefficient θ_{gm}^0. Hence, the subgroup structure is determined by the latent partitions $\{\mathcal{U}_{gm}^0, m = 1, \ldots, M_g^0\}$.

For more clarity, Figure 1 present some simple examples of the proposed group structure when $r = 2$, $p = 2$. In this figure, we have marked each partitioned block $\mathcal{U}_{gm}^0$ along with the regression coefficient values θ_{gm}^0 on each block, for $g = 1, 2$, $m = 1, \ldots, M_g^0$, as well as the boundary curves that form these partitions. For example, under Scenario *I*, the subgroup structure of the first

component of $\boldsymbol{\alpha}_i^0(\boldsymbol{U}_i)$ is determined by the partition $\{\mathcal{U}_{1m}^0, m = 1, 2\}$. When $\boldsymbol{U}_i$ falls into $\mathcal{U}_{11}^0, \mathcal{U}_{12}^0$, its regression coefficents takes the values $\theta_{11}^0, \theta_{12}^0$ respectively.

[Figure 1 about here.]

While the proposed model (1) closely resembles the varying coefficient model (Hastie and Tibshirani, 1993), the group structure in (2) is assumed to depend on the partitions of the support for the covariates $\boldsymbol{U}_i$. This model has three key advantages: first, it allows for predicting group membership for new individuals; second, it facilitates analyzing and explaining the reasons behind subgroup formation based on the estimated potential partitions; third, the group structure reflects the "similarity" of the grouping, meaning that individuals with more similar features are more likely to belong to the same group. This "similarity" property is essential for the effectiveness of the subsequent regularization technique-based group identification approach.

Note that, if we change the order of any two items $\theta_{gm}^0 I\{\boldsymbol{U}_i \in \mathcal{U}_{gm}^0\}$ and $\theta_{gm'}^0 I\{\boldsymbol{U}_i \in \mathcal{U}_{gm'}^0\}$ for $m \neq m'$, model (2) will remain unchanged. To ensure the identifiability of the θ_{gm}^0, we assume without loss of generality that $\theta_{g1}^0 < \cdots < \theta_{gM_g^0}^0$ for any $g = 1, \ldots, p$. Let $\boldsymbol{\theta^0} = (\theta_{11}^0, \ldots, \theta_{1M_1^0}^0, \ldots, \theta_{p1}^0, \ldots, \theta_{pM_p^0}^0)^T$ is a $\sum_{g=1}^p M_g^0$-dimensional vector.

The proposed model is associated with the following models.

Example 1. (Covariate multi-threshold model) Consider the following model

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}^0 + \boldsymbol{Z}_i^T \boldsymbol{\alpha}_i^0(U_i) + \epsilon_i = \boldsymbol{X}_i^T \boldsymbol{\beta}^0 + \sum_{g=1}^{p} \sum_{m=1}^{M^0} Z_{gi} \theta_{gm}^0 I(\gamma_{m-1}^{\star 0} < U_i \leqslant \gamma_m^{\star 0}) + \epsilon_i. \tag{3}$$

This model is a special case of the proposed model, corresponding to the case when U_i is a scalar, $M_g^0 = M^0$, $\mathcal{U}_{gm}^0 = \mathcal{U}_m^0$ for $g = 1, \ldots, p$, $m = 1, \ldots M^0$. $\mathcal{U}_1^0 = \{U : U \leqslant \gamma_1^{\star 0}\}$, $\mathcal{U}_2^0 = \{U : \gamma_1^{\star 0} < U \leqslant \gamma_2^{\star 0}\}, \ldots, \mathcal{U}_{M^0}^0 = \{U : \gamma_{M^0-1}^{\star 0} < U\}$, in which $\gamma_0^{\star 0} = -\infty$, $\gamma_{M^0}^{\star 0} = +\infty$ and $\gamma_0^{\star 0} < \gamma_1^{\star 0} < \ldots < \gamma_{M^0}^{\star 0}$. When $M^0 = 2$, $\theta_g^{\star 0} = \theta_{g2}^0 - \theta_{g1}^0$, the model (3) can be rewritten as

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}^0 + \sum_{g=1}^{p} Z_{gi} \theta_{g1}^0 + \sum_{g=1}^{p} Z_{gi} \theta_g^{\star 0} I(U_i > \gamma_1^{\star 0}) + \epsilon_i, \tag{4}$$

Model (4) is the linear regression covariate threshold model considered in Lee et al. (2011).

Example 2.(Parallel threshold change plane model) Consider the following model

$$Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta}^0 + \boldsymbol{Z}_i^T\boldsymbol{\alpha}_i^0(\boldsymbol{U}_i) + \epsilon_i = \boldsymbol{X}_i^T\boldsymbol{\beta}^0 + \sum_{g=1}^{p}\sum_{m=1}^{M^0} Z_{gi}\theta_{gm}^0 I(\gamma_{m-1}^{\star 0} < \boldsymbol{U}_i^T\boldsymbol{\gamma}^{*0} \leqslant \gamma_m^{\star 0}) + \epsilon_i. \tag{5}$$

It is a special case of the proposed model, corresponding to the case when $M_g^0 = M^0, \mathcal{U}_{gm}^0 = \mathcal{U}_m^0$ for $g = 1, \ldots, p$, $m = 1, \ldots M^0$. And $\mathcal{U}_1^0 = \{\boldsymbol{U} : \boldsymbol{U}^T\boldsymbol{\gamma}^{*0} \leqslant \gamma_1^{\star 0}\}, \mathcal{U}_2^0 = \{\boldsymbol{U} : \gamma_1^{\star 0} < \boldsymbol{U}^T\boldsymbol{\gamma}^{*0} \leqslant \gamma_2^{\star 0}\}, \ldots, \mathcal{U}_{M^0}^0 = \{\boldsymbol{U} : \gamma_{M^0-1}^{\star 0} < \boldsymbol{U}^T\boldsymbol{\gamma}^{*0}\}$, in which $\gamma_0^{\star 0} = -\infty, \gamma_{M^0}^{\star 0} = +\infty$ and $\gamma_0^{\star 0} < \gamma_1^{\star 0} < \ldots < \gamma_{M_p^0}^{\star 0}$. When $M^0 = 2$, $\theta_g^{*0} = \theta_{g2}^0 - \theta_{g1}^0$, the model (5) reduce to

$$Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta}^0 + \sum_{g=1}^{p} Z_{gi}\theta_{g1}^0 + \sum_{g=1}^{p} Z_{gi}\theta_g^{*0} I(\boldsymbol{U}_i^T\boldsymbol{\gamma}^{*0} > \gamma_1^{\star 0}) + \epsilon_i, \tag{6}$$

Model (6) is single-threshold change plane model proposed by Li et al. (2021).

These existing models are implicitly or explicitly based on the assumptions: (i) the number of groups M_g^0 are same for all $g = 1, \ldots, p$; (ii) The group partitions $\mathcal{U}_{gm}^0$ for any $g = 1, \ldots, p$ are same; (iii) the group boundary of two adjacent partitions $\mathcal{U}_{gm}^0$ and $\mathcal{U}_{gm'}^0$ are linear function of U_i for $g = 1, \ldots, p, m = 1, \ldots, M_g^0$; (iv) the partitions $\mathcal{U}_{gm}^0$ boundary curves are parallel straight lines that are non-intersecting within the interior of $\mathcal{U}$. These assumptions, while simplifying model estimation, are too restrictive and may be violated. For example, scenarios III violates the assumptions (ii) and (iv), while scenarios IV in Figure 1 breach all these assumptions. In contrast, the model structure (2) proposed in this paper makes no such assumptions and can be applied to any of the scenarios in Figure 1. Specifically, the proposed model structure (2) allows the number of groups M_g^0 and the group partitions $\mathcal{U}_{gm}^0, m = 1, \ldots, M_g^0$ to vary for different $g \in 1, \ldots, p$. The group boundary between two adjacent partitions can be either linear or nonlinear, and the boundary curves may intersect within the support $\mathcal{U}$. The mixed-type threshold model in the following example, more general than those in Examples 1-2 and previously unstudied, is also a special case of the proposed model.

Example 3.(Mixed type threshold model) Consider the following model when $p = 2$,

$$Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta}^0 + \boldsymbol{Z}_i^T\boldsymbol{\alpha}_i^0(\boldsymbol{U}_i) + \epsilon_i = \boldsymbol{X}_i^T\boldsymbol{\beta}^0 + \sum_{g=1}^{p}\sum_{m=1}^{M_g^0} Z_{gi}\theta_{gm}^0 I(\boldsymbol{U}_i \in \mathcal{U}_{gm}^0) + \epsilon_i, \tag{7}$$

$$\mathcal{U}_{11}^0 = \left\{\boldsymbol{U}_i : \{f_{11}^{*0}(\boldsymbol{U}_i) > \gamma_1^{\star 0}\} \cap \{\boldsymbol{U}_i^T\boldsymbol{\gamma}_{12}^{*0} > \gamma_2^{\star 0}\}\right\}, \mathcal{U}_{12}^0 = \left\{\boldsymbol{U} : \{f_{11}^{*0}(\boldsymbol{U}_i) > \gamma_1^{\star 0}\} \cap \{\boldsymbol{U}_i^T\boldsymbol{\gamma}_{12}^{*0} \leqslant \gamma_2^{\star 0}\}\right\},$$

$$\mathcal{U}_{13}^0 = \left\{\boldsymbol{U} : f_{11}^{*0}(\boldsymbol{U}_i) \leqslant \gamma_1^{\star 0}\right\}, \mathcal{U}_{21}^0 = \left\{\boldsymbol{U} : \{f_{21}^{*0}(\boldsymbol{U}_i) \leqslant \gamma_3^{\star 0}\}\right\}, \mathcal{U}_{22}^0 = \left\{\boldsymbol{U} : \{f_{21}^{*0}(\boldsymbol{U}_i) > \gamma_3^{\star 0}\}\right\},$$

where $M_1^0 = 3$, $M_2^0 = 2$, $f_{11}^{*0}(\boldsymbol{U}_i)$, $f_{21}^{*0}(\boldsymbol{U}_i)$ are the nonlinear function of $\boldsymbol{U}_i$. Similar to Scenario *IV* of Figure 1, this region is obtained from the intersection of linear and nonlinear boundaries.

It is worth noting that the proposed model is more general and flexible than finite mixture models (Shen and He 2015, Zhou et al. 2022), threshold models (Li et al. 2021, Zhang et al. 2022) and grouping models in Ma and Huang (2017) and He et al. (2023). However, its model structure is also more complex than those of these models. In addition to the unknown regression coefficient $\boldsymbol{\beta}^0, \boldsymbol{\theta}^0$, the latent partitions $\{\mathcal{U}_{gm}^0, m = 1, \ldots, M_g^0\}$ need to be estimated. In fact, the membership prediction depends on the partition recovery step. Therefore, our goal is to estimate the unknown regression parameters $\boldsymbol{\beta}^0, \boldsymbol{\theta}^0$, as well as the group memberships, and then attempt to recover the latent partitions, making membership prediction possible.

To achieve this goal, we propose a three-step estimation procedure. In the first step, we identify the subgroup structure by using neighborhood-based regularization technique. In the second step, we employ a support vector machine to recover the latent partitions based on the estimated subgroup structure from the first step, making the prediction of subgroup memberships possible. In the third step, we enhance the performance of the estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ through a post-group-recovery estimation procedure. To clearly present the proposed method, we focus on the main ideas in the body of the paper, while leaving many notation definitions and formula details in the Supplementary Material. The implementation details of these three steps will be discussed in the following Sections.

3. Methodology

3.1 *Latent group structure detection*

As mentioned earlier, the group structure defined in (2) satisfies the "similarity" assumption for grouping. To capture and utilize this "similarity" in group detection, we need to define the neighborhood for each subject. For subject i, a commonly used neighborhood is defined as $\mathcal{N}_h^i = \{j : \|\boldsymbol{U}_i - \boldsymbol{U}_j\|_2 < h\}$ or $\mathcal{N}_K^i = \{j : \boldsymbol{U}_j$ is one of the K nearest to $\boldsymbol{U}_i$ among all the subjects$\}$, where h is bandwidth, K is positive integer, and $\|\cdot\|_2$ is Euclidean norm. Unlike the traditional pairwise fusion method in Ma and Huang (2017), which connects each subject to all other $n-1$ subjects, our approach only connects each subject to those within its neighborhood, thus potentially reducing computational complexity. For convenience, we adopt the second definition of the neighborhood. We explore the group structure by minimizing the loss function with a neighborhood-based fused penalty:

$$L_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{1}{2}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} - \boldsymbol{Z}_i^T\boldsymbol{\alpha}_i)^2 + \sum_{i=1}^{n}\sum_{j\in\mathcal{N}_K^i}\sum_{g=1}^{p} p(|\alpha_{ig} - \alpha_{jg}|, \lambda_g), \tag{8}$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{ip})'$, $p(\cdot, \lambda)$ is a penalty function with a tuning parameter $\lambda \geqslant 0$.

The rationale behind the optimization given in (8) is to fuse similar α_{ig}'s into subgroups, where each subgroup has a common value $\boldsymbol{\theta}_g$, obtained from the fused penalty function $p(\cdot, \lambda)$. With a suitable choice of λ_g, $[\boldsymbol{\alpha}]_g$ are forced to form subgroups in the fused regularization estimator, where $[\boldsymbol{\alpha}]_g$ denote the g-th component of vector $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n)^T = ([\boldsymbol{\alpha}]_1, \ldots, [\boldsymbol{\alpha}]_p)$. In this paper, we choose the minimax concave penalty function (Zhang, 2010) due to its proven advantages in terms of low bias and stable numerical performance (Ma and Huang, 2017). Additionally, the neighborhood-based minimax concave penalty effectively eliminates spurious associations while reinforcing genuine within-subgroup linkages. This process simultaneously resolves the ambiguities induced by overlapping neighborhoods, thereby enabling the clear assignment of each sample to its most relevant subgroup. The minimax concave penalty takes the form, $p_\gamma(t, \lambda) = \lambda \int_0^t \{1 - s/(\gamma\lambda)\}_+ ds$ for $\gamma > 1$, where $(x)_+ = x$ if $x > 0$ and $(x)_+ = 0$. The computational complexity

of the proposed method is $O(nK)$, which is more computationally efficient than the pairwise regularization method in (Ma and Huang, 2017), whose computational complexity is $O(n^2)$.

By minimizing the penalized loss function (8), we can obtain $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}_g$ for $g = 1, \cdots, p$. Let $\{\widehat{\theta}_{g1}, \ldots, \widehat{\theta}_{g\widehat{M}_g}\}$ be $\widehat{M}_g$ distinct values of $[\widehat{\boldsymbol{\alpha}}]_g$, due to the previously assumed identification assumption, let $\widehat{\theta}_{g1} < \ldots < \widehat{\theta}_{g\widehat{M}_g}$. Then we can obtain $\widehat{M}_g$ groups based on these $\widehat{M}_g$ distinct values.

3.2 *Group membership prediction via partitions recovery*

In subgroup analysis, predicting group membership is another important issue. Existing regularization methods, such as those in Ma and Huang (2017),Tang et al. (2021) and He et al. (2023), cannot predict the group membership of new subjects unless we include them in the original sample and repeat the group identification step. In contrast, our proposed method enables us to predict group membership based on the recovered group partitions, without redoing the group identification step.

To proceed, we first obtain the estimated group structure including the estimated number of group $\widehat{M}_g$ and the corresponding group memberships, $g = 1, \ldots, p$, after detecting the latent group structure. Define the true group memberships, $G^0_{gm} = \left\{i \in \{1, \ldots, n\} : \alpha^0_{ig} = \theta^0_{gm}\right\}, g = 1, \ldots, p, m = 1, \ldots, M^0_g$. Based on the obtained $\widehat{\theta}_{g1} < \ldots < \widehat{\theta}_{g\widehat{M}_g}$, the estimated membership set is defined as $\widehat{G}_{gm} = \left\{i \in \{1, \ldots, n\} : \widehat{\alpha}_{ig} = \widehat{\theta}_{gm}\right\}, g = 1, \ldots, p, m = 1, \ldots, \widehat{M}_g$. When $\widehat{M}^0_g = 2$, the g-th partition estimation could be a binary classification problem, otherwise could be a multi-class classification problem. We propose an support vector machine based partition recovery method due to its superior performance with moderate sample sizes, compared to other classification methods such as random forests and neural networks (Koo et al., 2008; Zhang et al., 2022).To save space, we do not distinguish between linear and nonlinear support vector machine. If the partition boundary is unknown, we prefer linear support vector machine. If the classification accuracy is low, we then consider using nonlinear support vector machine.

Two-group partition recovery and group membership prediction

We introduce the partition estimation method for the element with 2 groups, i.e $\widehat{M}_g = 2$ for any $g \in \{1, \ldots, p\}$. We assign artificial labels to two groups firstly. Let $\widetilde{Y}^*_{gi} = -1$ when i-th subjects'

belongs to $\widehat{G}_{g1}$ and $\widetilde{Y}^*_{gi} = 1$, otherwise. Then the binary classifier $\widehat{f}_g : \boldsymbol{U} \to \{-1, 1\}$ could be obtained by solving following optimization problem (Vapnik, 1998), i.e:

$$\min_{f_g \in \overline{\mathcal{H}}_{\mathcal{K}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(1 - \widetilde{Y}^*_{gi} f_g(\boldsymbol{U}_i)\right)_+ + \lambda_n \|f^*_g(\boldsymbol{U}_i)\|^2_{\mathcal{K}} \right\}, \tag{9}$$

where "+" indicates positive part, λ_n is a regularization parameter, $\overline{\mathcal{H}}_{\mathcal{K}} = \mathcal{H}_{\mathcal{K}} + R$, $f_g(\boldsymbol{U}_i) = f^*_g(\boldsymbol{U}_i) - \gamma^\star_g$ with $f^*_g(\boldsymbol{U}) \in \mathcal{H}_{\mathcal{K}}$ and $\gamma^\star_g \in R$, in which $\mathcal{H}_{\mathcal{K}}$ is the reproducing kernel Hilbert space associate with kernel function $\mathcal{K}$. After obtaining $\widehat{f}_g(\boldsymbol{U})$, define $C_{gm} = \#\left\{i \in \{1, \cdots, n\} : \widehat{\alpha}_{ig} = \widehat{\theta}_{gm}, \widehat{f}_g(\boldsymbol{U}_i) \leqslant 0\right\}, m = 1, 2$. The partition estimation $\widehat{\mathcal{U}}_{g1} = \{\boldsymbol{U} : \widehat{f}_g(\boldsymbol{U}) \leqslant 0\}$, $\widehat{\mathcal{U}}_{g2} = \{\boldsymbol{U} : \widehat{f}_g(\boldsymbol{U}) > 0\}$ for $C_{g1} > C_{g2}$. Otherwise, $\widehat{\mathcal{U}}_{g1} = \{\boldsymbol{U} : \widehat{f}_g(\boldsymbol{U}) > 0\}$, $\widehat{\mathcal{U}}_{g2} = \{\boldsymbol{U} : \widehat{f}_g(\boldsymbol{U}) \leqslant 0\}$ for $C_{g1} < C_{g2}$. When a new subject arrives, we can predict its group membership by determining which partition block its $\boldsymbol{U}_i$ falls into. To save space, details of multi-group recovery and membership prediction appear in Supplementary Materials A1.

3.3 *Post-group-recovery estimation*

We can obtain the estimators, $\widehat{\boldsymbol{\beta}}$, $\widehat{\theta}_{gm}$, and $\widehat{G}_{gm}$, for $g = 1, \ldots, p$ and $m = 1, \ldots, \widehat{M}_g$, from the previous two steps. Belloni and Chernozhukov (2013) showed that the ordinary least squares post-lasso estimators tend to have smaller bias than the original lasso estimator. Motivated by this, we adopt a post-group recovery estimation procedure to obtain more accurate estimators for $\boldsymbol{\beta}$ and θ_{gm}.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_p^T)^T$, $\boldsymbol{W}_{i,g} = (I(i \in G^0_{g1}), \ldots, I(i \in G^0_{gM^0_g}))$. It is easy to obtain that $\alpha_{ig} = \boldsymbol{W}_{i,g}\boldsymbol{\theta}_g$. Hence, we can get post-group recovery estimators based on the estimated group structure relationship $\widehat{\boldsymbol{W}}_g, g = 1, \ldots, p$,

$$(\widehat{\boldsymbol{\beta}}^{post}, \widehat{\boldsymbol{\theta}}^{post}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^q, \boldsymbol{\theta} \in \mathbb{R}^{\widehat{M}}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta} - \sum_{g=1}^{p} Z_{ig} \widehat{\boldsymbol{W}}_{i,g} \boldsymbol{\theta}_g)^2.$$

where $\widehat{\boldsymbol{W}}_g = (\widehat{\boldsymbol{W}}_{1,g}, \ldots, \widehat{\boldsymbol{W}}_{n,g})^T$ is the estimator of $\boldsymbol{W}_g = (\boldsymbol{W}_{1,g}, \ldots, \boldsymbol{W}_{n,g})^T$, obtained by replacing the G^0_{gm}, M^0_g with $\widehat{G}_{gm}, \widehat{M}_g$, and $\sum_{g=1}^{p} \widehat{M}_g = \widehat{M}$.

3.4 *Extension to high-dimensional data*

In high-dimensional settings, our method maintains its validity provided that the condition $(\sum_{g=1}^{p} M^0_g + q) = o(n)$ in Theorem 1 is satisfied. Although the dimension r of $\boldsymbol{U}$ may increase

with that of $\boldsymbol{X}$, it must not grow too rapidly-specifically, $r = o(\log n)$. To make the model estimable and interpretable, we add a sparsity penalty of β to the objective (8),

$$L_{hn}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{1}{2}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} - \boldsymbol{Z}_i^T\boldsymbol{\alpha}_i)^2 + \sum_{i=1}^{n}\sum_{j\in\mathcal{N}_K^i}\sum_{g=1}^{p} p(|\alpha_{ig} - \alpha_{jg}|, \lambda_g) + \sum_{g=1}^{q} p(|\beta_g|, \lambda_{s2}), \quad (10)$$

where $p(\cdot, \lambda_{s2})$ is a sparsity-induced penalty function and we consider the minimax concave penalty (Zhang, 2010). Similarly, the post-group-recovery estimator in the high dimensional data, $(\widehat{\boldsymbol{\beta}}_s^{post}, \widehat{\boldsymbol{\theta}}^{post}) = \underset{\boldsymbol{\beta}_s\in\mathbb{R}^{q_s}, \boldsymbol{\theta}\in\mathbb{R}^{\widehat{M}}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_{i,s}^T\boldsymbol{\beta}_s - \sum_{g=1}^{p} Z_{ig}\widehat{\boldsymbol{W}}_{i,g}\boldsymbol{\theta}_g)^2$, where $\boldsymbol{X}_{i,s}$ represent the important covariates selected in previous step and $\boldsymbol{\beta}_s$ are corresponding coefficients.

4. Theoretical properties

In this Section, we establish theoretical properties of the proposed estimator under the proposed heterogeneous model. Initial results focus on the oracle properties of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ estimators given known group structure. Then, we prove that the group structure could be recovered with probability approaching 1, and based on it, we show that the proposed estimators are asymptotically equivalent to the oracle estimators. In addition, we establish the Fisher consistency of the proposed partition recovery method based on the support vector machine 1-norm soft margin classifier within the regularization framework. Before that, we introduce some useful notations firstly.

For $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n) = ([\boldsymbol{\alpha}]_1, \ldots, [\boldsymbol{\alpha}]_p)$, take $[\boldsymbol{\alpha}]_g$ as an example to illustrate the symbols'. Let $\mathcal{M}_{\mathcal{G}}^g$ be the subspace of $\mathbb{R}^n$, $\mathcal{M}_{\mathcal{G}}^g = \{[\boldsymbol{\alpha}]_g \in \mathbb{R}^n : \alpha_{ig} = \alpha_{jg}, \text{for any } i, j \in G_{gm}^0, 1 \leqslant m \leqslant M_g^0\}$. For each $[\boldsymbol{\alpha}]_g \in \mathcal{M}_{\mathcal{G}}^g$, it can be written as $[\boldsymbol{\alpha}]_g = \boldsymbol{W}_g\boldsymbol{\theta}_g$, where $\boldsymbol{W}_g = (\boldsymbol{W}_{1,g}, \ldots, \boldsymbol{W}_{n,g})^T$. By matrix calculation, $\boldsymbol{D}_g = \boldsymbol{W}_g^T\boldsymbol{W}_g = \text{diag}(|G_{g1}^0|, \ldots, |G_{gM_g^0}^0|)$, where $|G_{gm}^0|$ denotes the group size of G_{gm}^0. Define $|G_{\min}^0| = \min_{1\leqslant g\leqslant p, 1\leqslant m\leqslant M_g^0} |G_{gm}^0|, |G_{\max}^0| = \max_{1\leqslant g\leqslant p, 1\leqslant m\leqslant M_g^0} |G_{gm}^0|$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T, \boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$, and $\boldsymbol{Z}^g = diag(Z_{g1}, \ldots, Z_{gn})$. Denote $\widetilde{\boldsymbol{Z}}_g = \boldsymbol{Z}^g\boldsymbol{W}_g$ and $\boldsymbol{E} = (\boldsymbol{X}, \boldsymbol{Z}^1\boldsymbol{W}_1, \ldots, \boldsymbol{Z}^p\boldsymbol{W}_p)$. To ensure the identifiability of $\boldsymbol{\theta}$, we assume that $\theta_{g1}^0 < \cdots < \theta_{gM_g^0}^0$ for any $g = 1, \ldots, p$ as discuss in Section 2. We study the theoretical properties of the proposed estimator under the heterogeneous model, where there are at least two subgroups, i.e., $\max\{M_g^0, g =$

$1,\ldots,p\} \geqslant 2$. If the underlying groups $G^0_{11},\ldots,G^0_{1M^0_1},\ldots,G^0_{p1},\ldots,G^0_{pM^0_p}$ were known, the oracle estimator for $(\boldsymbol{\beta},\boldsymbol{\alpha})$ would be $(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\alpha}}^{or}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^q,[\boldsymbol{\alpha}]_g\in\mathbb{R}^n}{\arg\min}\ \frac{1}{2}\|Y-\boldsymbol{X}\boldsymbol{\beta}-\sum_{g=1}^p \boldsymbol{Z}^g[\boldsymbol{\alpha}]_g\|^2$, where $\widehat{\boldsymbol{\alpha}}^{or} = ([\widehat{\boldsymbol{\alpha}}^{or}]_1,\ldots,[\widehat{\boldsymbol{\alpha}}^{or}]_p)^T$. Correspondingly, the oracle estimators for the common coefficient $\boldsymbol{\beta}$ and the group $\boldsymbol{\theta}$ coefficient are $(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\theta}}^{or}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^q,\boldsymbol{\theta}_g\in\mathbb{R}^{M^0_g}}{\arg\min}\ \frac{1}{2}\|Y-\boldsymbol{X}\boldsymbol{\beta}-\sum_{g=1}^p \widetilde{\boldsymbol{Z}}_g\boldsymbol{\theta}_g\|^2 = (\boldsymbol{E}^T\boldsymbol{E})^{-1}\boldsymbol{E}^T\boldsymbol{Y}$. where $\widehat{\boldsymbol{\theta}}^{or} = (\widehat{\boldsymbol{\theta}}^{or}_1,\ldots,\widehat{\boldsymbol{\theta}}^{or}_p)$. The following theorem provides the asymptotic property of the oracle estimators.

THEOREM 1: *Suppose* $\left|G^0_{min}\right| \gg \sqrt{(q+\sum_{g=1}^p M^0_g)n\log n}$ *and* $(\sum_{g=1}^p M^0_g+q) = o(n)$. *Then under Conditions* $(C1)-(C3)$, *we have with probability at least* $1-2(\sum_{g=1}^p M^0_g+q)n^{-1}$,
(1) $\|(\widehat{\boldsymbol{\beta}}^{or}-\boldsymbol{\beta}^0)^T,(\widehat{\boldsymbol{\theta}}^{or}-\boldsymbol{\theta}^0)^T\| \leqslant \phi_n$, *where* $\phi_n = c^{-1/2}C_1^{-1}\sqrt{q+\sum_{g=1}^p M^0_g}\left|G^0_{\min}\right|^{-1}\sqrt{n\log n}$ *and* c *is a constant. Besides,* $\|(\widehat{\boldsymbol{\alpha}}^{or}-\boldsymbol{\alpha}^0)^T\| \leqslant \sqrt{\left|G^0_{\max}\right|}\phi_n, \sup_i\|\widehat{\boldsymbol{\alpha}}^{or}_i-\boldsymbol{\alpha}^0_i\| \leqslant \phi_n$. *(2) For any vector* $\boldsymbol{a}_n \in R^{q+\sum_{g=1}^p M^0_g}$ *with* $\|\boldsymbol{a}_n\| = 1$, *as* $n\to\infty$, $\sigma_n(\boldsymbol{a}_n)^{-1}\boldsymbol{a}_n^T\left((\widehat{\boldsymbol{\beta}}^{or}-\boldsymbol{\beta}^0)^T,(\widehat{\boldsymbol{\theta}}^{or}-\boldsymbol{\theta}^0)^T\right)\to_D N(0,1)$ *where* $\sigma_n(\boldsymbol{a}_n) = \sigma[\boldsymbol{a}_n^T(\boldsymbol{E}^T\boldsymbol{E})^{-1}\boldsymbol{a}_n]^{1/2}$.

REMARK 1: By letting $\left|G^0_{min}\right| = \delta n/(\max_{1\leqslant g\leqslant p} M^0_g)$ for some constant $0<\delta\leqslant 1$, the bound ϕ_n is $\phi_n = c^{-1/2}C_1^{-1}\delta^{-1}(\max_{1\leqslant g\leqslant p} M^0_g)\sqrt{q+\sum_{g=1}^p M^0_g}\sqrt{\log n/n}$. Moreover, if $q, M^0_g (g=1,\ldots,p)$ and p are fixed quantities, then $\phi_n = C^*\sqrt{\log n/n}$ for some $0<C^*<\infty$. Let $b_n = \min_{1\leqslant g\leqslant p}\min_{i\in G^0_{gm},j\in G^0_{gm'},m\neq m'}|\alpha^0_{ig}-\alpha^0_{jg}| = \min_{1\leqslant g\leqslant p}\min_{m\neq m'}|\theta^0_{gm}-\theta^0_{gm'}|$ be the minimal difference of the common values between two groups. And $K' = \min_{i,g}|\{j:\alpha^0_{jg}=\alpha^0_{ig}, j\in\mathcal{N}^i_K\}|$ are the number of K-nearest-neighbors of each individual that belong to the same group.

THEOREM 2: *Suppose the conditions in Theorem 1 and Condition (C4)-(C5) hold. If* $b_n > a\lambda$, $K\geqslant K'\gg\max\{\log(n^2p), p\phi_n\}$ *and* $\lambda\gg\max\{\frac{\sqrt{p\log n}}{K'},\phi_n\}$, *for some constant* $a>0$, *where* ϕ_n *is given in Theorem 1, then there exists a local minimizer* $(\widehat{\boldsymbol{\beta}}(\lambda),\widehat{\boldsymbol{\alpha}}(\lambda))$ *of the objective function given (8) satisfying* $pr\left(\widehat{\boldsymbol{\beta}}(\lambda),\boldsymbol{\alpha}(\lambda) = (\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\alpha}}^{or})\right)\to 1$.

REMARK 2: Theorem 2 shows that the oracle estimator $(\widehat{\boldsymbol{\beta}}^{or},\widehat{\boldsymbol{\alpha}}^{or})$ is a local minimizer of the objective function with high probability, and thus the true groups can be recovered. Specifically,

the common value for group m at the g-th component is given by $\widehat{\theta}_{gm} = \widehat{\alpha}_{ig}^{or}$ for $i \in G_{gm}^0$. This result holds under the condition that $K \geqslant K' \gg \max\{\log(n^2 p), p\phi_n\}$. When all M_g^0, and p, q are finite and fixed, K can be of the order of $\log n$, which is consistent with Madrid Padilla et al. (2020).

Let $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_{11}, \ldots, \widehat{\theta}_{1\widehat{M}_1}, \ldots, \widehat{\theta}_{p1}, \ldots, \widehat{\theta}_{p\widehat{M}_p})^T$ be the estimated treatment effects such that $\widehat{\theta}_{gm} = \widehat{\alpha}_{gm}$ for $i \in \widehat{G}_{gm}$, $g = 1, \ldots, p, m = 1, \ldots, \widehat{M}_g$, and $\widehat{\alpha}$ is the local minimizer given in Theorem 2. Based on the results in Theorems 1 and 2, we obtain the asymptotic distribution of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ and the consistency of post-recovery estimators giving in the following corollary.

Corollary 1: *Under the conditions in Theorem 2, for any* $\boldsymbol{a}_n \in R^{q+\sum_{g=1}^p M_g}$ *with* $\|\boldsymbol{a}_n\| = 1$, *as* $n \to \infty$, *we have (1)* $pr(\widehat{\boldsymbol{W}}^g = \boldsymbol{W}^g, g = 1, \ldots, p) \to 1$. $\sigma_n(\boldsymbol{a}_n)^{-1}\boldsymbol{a}_n^T \left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)^T\right) \to_D N(0,1)$ *where* $\sigma_n(\boldsymbol{a}_n) = \sigma[\boldsymbol{a}_n^T(\boldsymbol{E}^T\boldsymbol{E})^{-1}\boldsymbol{a}_n]^{1/2}$. *(2) Besides, for any vectors* $\boldsymbol{a}_{n1} \in R^q$, $\boldsymbol{a}_{n2} \in R^{\sum_{g=1}^p M_g^0}$ *with* $\|\boldsymbol{a}_{n1}\| = 1$, $\|\boldsymbol{a}_{n2}\| = 1$, *as* $n \to \infty$, $\sigma_{n1}(\boldsymbol{a}_{n1})^{-1}\boldsymbol{a}_{n1}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \to_D N(0,1)$, $\sigma_{n2}(\boldsymbol{a}_{n2})^{-1}\boldsymbol{a}_{n2}^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \to_D N(0,1)$ *where* $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_1^T, \ldots, \widehat{\boldsymbol{\theta}}_p^T)$, $\sigma_{n1}(\boldsymbol{a}_{n1}) = \sigma[\boldsymbol{a}_{n1}^T[\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{X}^T\widetilde{\boldsymbol{Z}}(\widetilde{\boldsymbol{Z}}^T\widetilde{\boldsymbol{Z}})^{-1}\widetilde{\boldsymbol{Z}}^T\boldsymbol{X}]^{-1}\boldsymbol{a}_{n1}]^{1/2}$ *and* $\sigma_{n2}(\boldsymbol{a}_{n2}) = \sigma[\boldsymbol{a}_{n2}^T[\widetilde{\boldsymbol{Z}}^T\widetilde{\boldsymbol{Z}} - \widetilde{\boldsymbol{Z}}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\widetilde{\boldsymbol{Z}}]^{-1}\boldsymbol{a}_{n2}]^{1/2}$, *in which* $\widetilde{\boldsymbol{Z}} = (\widetilde{\boldsymbol{Z}}_1, \ldots, \widetilde{\boldsymbol{Z}}_p)$. *(3)* $pr\left((\widehat{\boldsymbol{\beta}}^{post}, \widehat{\boldsymbol{\theta}}^{post}) = (\widehat{\boldsymbol{\beta}}^{or}, \widehat{\boldsymbol{\theta}}^{or})\right) \to 1$.

Next, we consider the high-dimensional situation. Without loss of generality, let $\boldsymbol{\beta}^0 = ((\boldsymbol{\beta}^0)_s^T, (\boldsymbol{\beta}^0)_{s^c}^T)^T$, where the first q_s components of $\boldsymbol{\beta}$, denote by $\boldsymbol{\beta}_s^0$, are nonzero, the remaining $q - q_s$ coefficients, $\boldsymbol{\beta}_{s^c}^0$ are 0. The oracle estimator when true memberships is known, which is defined as $(\widehat{\boldsymbol{\beta}}_s^{or}, \widehat{\boldsymbol{\beta}}_{s^c}^{or}, \widehat{\boldsymbol{\theta}}^{or}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^q, \boldsymbol{\theta}_g\in\mathbb{R}^{M_g^0}}{\arg\min} \frac{1}{2}\|Y - \boldsymbol{X}\boldsymbol{\beta} - \sum_{g=1}^p \widetilde{\boldsymbol{Z}}_g\boldsymbol{\theta}_g\|^2 + \sum_{g=1}^q p(|\beta_g|, \lambda_{s2})$.

Theorem 3: *Under the conditions of Theorem 2 and Conditions (C6)-(C7) hold, if* $\lambda_{s2}/n \to 0, \lambda_{s2}/\sqrt{nq} \to \infty$ *and* $q^3/n \to 0$, *then with probability tending to 1,*

(1) $\widehat{\boldsymbol{\beta}}_{s^c}^{or} = \boldsymbol{\beta}_{s^c}^0 = 0$. *(2) For any vector* $\boldsymbol{a}_n \in R^{q_s+\sum_{g=1}^p M_g^0}$ *with* $\|\boldsymbol{a}_n\| = 1$, *as* $n \to \infty$, $\sigma_n(\boldsymbol{a}_n)^{-1}\boldsymbol{a}_n^T \left((\widehat{\boldsymbol{\beta}}_s^{or} - \boldsymbol{\beta}_s^0)^T, (\widehat{\boldsymbol{\theta}}^{or} - \boldsymbol{\theta}^0)^T\right) \to_D N(0,1)$. $\sigma_n(\boldsymbol{a}_n) = \sigma[\boldsymbol{a}_n^T(\boldsymbol{E}_s{}^T\boldsymbol{E}_s)^{-1}\boldsymbol{a}_n]^{1/2}$ *and* $\boldsymbol{E}_s = (\boldsymbol{X}_s, \boldsymbol{Z}^1\boldsymbol{W}_1, \cdots, \boldsymbol{Z}^p\boldsymbol{W}_p)$. *(3)* $pr\left((\widehat{\boldsymbol{\beta}}_s^{post}, \widehat{\boldsymbol{\theta}}^{post}) = (\widehat{\boldsymbol{\beta}}_s^{or}, \widehat{\boldsymbol{\theta}}^{or})\right) \to 1$.

In the following part, we establish the Fisher consistency of the proposed partition recovery

method based on the support vector machine within the regularization framework. Define the generalization error $\mathcal{E}(f_g(\boldsymbol{U})) = E((1 - \widetilde{Y}_g^{*0} f_g(\boldsymbol{U}))_+)$ (Zhang, 2004), if the underlying group memberships G_{g1}^0 and G_{g2}^0 were konwn, the ture oracle Bayes rule f_{gb}^{or} is represented as $f_{gb}^{or} = \text{sgn}\left(P_g(\widetilde{Y}_g^{*0} = 1 \mid \boldsymbol{U} = \boldsymbol{u}) - P_g(\widetilde{Y}_g^{*0} = -1 \mid \boldsymbol{U} = \boldsymbol{u})\right)$, see Lee et al. (2004). The following theorem provides the estimated classifier rule $\text{sgn}(\widehat{f_g}(\boldsymbol{U}))$ is Fisher consistent and asymptotically equivalent to the true oracle Bayes rule as the sample size tend to infinity.

THEOREM 4: *When* $\widehat{M} = 2$. *Under the conditions in Theorem 2, for every* $\lambda_n > 0$ *satisfy* $\lambda_n \to 0$ *and* $\frac{1}{n\sqrt{\lambda_n}} \log\left(\mathcal{N}\left(\sqrt{\lambda_n}\right)\right) \to 0$, *where* $\mathcal{N}(\epsilon)$ *is the covering number. Besides, define the regularization error* $\mathcal{D}(\lambda_n) = \inf_{f_g \in \overline{\mathcal{H}}_{\mathcal{K}}} \left\{\mathcal{E}(f_g(\boldsymbol{U})) - \mathcal{E}(f_{gb}^{or}(\boldsymbol{U})) + \lambda_n \|f_g^*\|_{\mathcal{K}}^2\right\}$, *where* $f_{gb}^{or}(\boldsymbol{U})$ *is the minimizer of* $\mathcal{E}(f_g(\boldsymbol{U}))$. *As* $\lim_{\lambda_n \to 0} \mathcal{D}(\lambda_n) = 0$, *we have*

(1) $pr\left\{\mathcal{R}(sgn(\widehat{f_g}(\boldsymbol{U}))) = \mathcal{R}((f_{gb}^{or}(\boldsymbol{U})))\right\} \to 1$. *(2)* $pr(\widehat{\mathcal{U}}_1 = \mathcal{U}_1^0, \widehat{\mathcal{U}}_2 = \mathcal{U}_2^0) \to 1$.

Similarly, we prove that multiple groups classifier is Fisher consistent and, as the sample size tends to infinity, asymptotically equivalent to the true oracle multi-class Bayes rule. For details, see Supplementary Materials A2.2.

5. Simulation study

We evaluate the finite sample performance of the proposed method using five data generation cases. Case 1 represents the ordinary covariate threshold model, while Case 2 corresponds to the parallel threshold change plane model. Case 3 allows the boundary curves to intersect within the support of $\boldsymbol{U}$ and different partition boundaries for different covariates. Case 4 is the most general, with both linear and nonlinear boundary functions, intersecting curves, and varying numbers of subgroups and partition boundaries for different covariates. The existing methods cannot address general cases like case 3 and 4, so we compare our method with Li et al. (2021) under cases 1 and 2. In Case 5, we also consider a high-dimensional homogeneous covariate setting. All data are generated from the model: $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}^0 + \boldsymbol{Z}_i^T \boldsymbol{\alpha}_i^0(\boldsymbol{U}_i) + \epsilon_i, i = 1, \ldots, n$. Simulation experiments are based on sample size $n = 400$ or 800 and $B = 200$ replicates. The specific setting is given as follows.

In Cases 1-4, the random error values of ϵ_i are generated from $N(0, 0.5^2)$, $\boldsymbol{X}_i = (1, X_{1i}, X_{2i})^T$, $\boldsymbol{Z}_i = (Z_{1i}, Z_{2i})^T$, the population parameters $\boldsymbol{\beta}^0 = (1, 1, 1)^T$. Two settings for the covariate (Z_{1i}, Z_{2i}) are considered: Case 1: a binary predictor $(0/1)$ with a probability of 0.5 for Z_{1i} each value, with a probability of 0.6 for Z_{2i} each value; Case 2: a continuous predictor are generated from $N((1, 2)^T, \boldsymbol{\Sigma}_1)$ where $\boldsymbol{\Sigma}_1 = \{\sigma_{jj'}\}, \sigma_{jj} = 1$ and $\sigma_{jj'} = 0$ for $j \neq j'$. All possible configurations of U_i and the probability distribution of $\boldsymbol{X}_i$ were incorporated in our setting. In Case 1, we assume X_{1i} generated from $U(0, 1)$, X_{2i} generated from $N(0, 1)$ and $U_i = X_{1i}$. Similar to Figure 1 Scenario I group structure, $\alpha^0_{1i} = \theta^0_{1m}$ for $i \in \mathcal{U}^0_{1m}$, $\alpha^0_{2i} = \theta^0_{2m}$ for $i \in \mathcal{U}^0_{2m}$, $m = 1, 2$. Where $(\theta^0_{11}, \theta^0_{12}) = (-1, 1), (\theta^0_{21}, \theta^0_{22}) = (-1.5, 0.5)$. In Case 2, we assume (X_{1i}, X_{2i}) generated from $N((0, 0)^T, \boldsymbol{\Sigma}_2), \boldsymbol{\Sigma}_2 = \{\sigma_{jj'}\}, \sigma_{jj} = 1$ and $\sigma_{jj'} = 0.3$ for $j \neq j'$, U_{1i}, U_{2i} independent generated from $U(0, 1)$. As shown in Figure 1 Scenario II, $\alpha^0_{1i} = \theta^0_{1m}$ for $i \in \mathcal{U}^0_{1m}$, $\alpha^0_{2i} = \theta^0_{2m}$ for $i \in \mathcal{U}^0_{2m}$, $m = 1, 2, 3$. Where $(\theta^0_{11}, \theta^0_{12}, \theta^0_{13}) = (-2, 0, 2), (\theta^0_{21}, \theta^0_{22}, \theta^0_{23}) = (-1.5, 0.5, 2.5)$. Under Case 2, we consider Case 3 with shown in Figure 1 Scenario III, the remaining settings are same as Case 2. In Case 4, we assume X_{1i}, X_{2i} independent generated from $U(0, 1)$ and $(U_{1i}, U_{2i}) = (X_{1i}, X_{2i})$, the group structure as shown in Figure 1 Scenario IV. Two settings for the covariate (Z_{1i}, Z_{2i}) are similar to Case 2. $\alpha^0_{1i} = \theta^0_{1m}$ for $i \in \mathcal{U}^0_{1m}$, $m = 1, 2, 3$, $\alpha^0_{2i} = \theta^0_{2m'}$ for $i \in \mathcal{U}^0_{2m'}$, $m' = 1, 2$. Where $(\theta^0_{11}, \theta^0_{12}, \theta^0_{13}) = (-2, 0, 2), (\theta^0_{21}, \theta^0_{22}) = (-1.5, 0.5)$. In Case 5, we consider a high-dimensional homogeneous covariate setting $q = \lfloor n^{1/3} \rfloor + 3$, the population parameters $\boldsymbol{\beta}^0 = (1, 1, 1, 0, \cdots, 0)$. We assume X_{1i} generated from $U(0, 1)$, X_{gi} generated from $N(0, 1)$, $g = 2, \cdots, \lfloor n^{1/3} + 3 \rfloor$, and $U_i = X_{1i}$. Assuming the dimensionality of heterogeneous covariates equal to the $\lfloor n^{1/3} \rfloor$, we verify that our method maintains numerical validity under moderate dimensionality regimes, provided $(\sum_{g=1}^{p} M^0_g + q) = o(n)$ is satisfied. Z_{gi} independent generated from $N(1, 1)$, $g = 2, \cdots, \lfloor n^{1/3} \rfloor$. Similar to Figure 1 Scenario I group structure, $\alpha^0_{1i} = \theta^0_{1m}$ for $i \in \mathcal{U}^0_{1m}$, $\alpha^0_{gi} = \theta^0_{gm}$ for $i \in \mathcal{U}^0_{gm}$, $m = 1, 2$, where $(\theta^0_{11}, \theta^0_{12}) = (-1, 1), (\theta^0_{g1}, \theta^0_{g2}) = (-2, 0), g = 2, \cdots, \lfloor n^{1/3} \rfloor$.

We adopt the alternating directions method of multipliers algorithm (Boyd et al., 2011) to

obtained the proposed estimate, details see Supplementary Materials A3.1. The tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^T$ and K is selected by minimizing a modified Bayes Information Criterion (BIC):

$$\text{BIC}(\boldsymbol{\lambda}, K) = \log\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}(\lambda_{s2}, K) - \sum_{g=1}^{p} Z_{gi}\widehat{\alpha}_{ig}(\lambda_g, K)\right)^2\right] + C_n\frac{\log n}{n}\left(\sum_{g=1}^{p}\widehat{M}_g(\lambda_g, K) + q_s\right),$$

where $C_n = \log(np + q_s)$ is a positive number, $\widehat{\boldsymbol{\beta}}(\lambda_{s2}, K)$, $\widehat{\alpha}_{ig}(\lambda_g, K)$ and $\widehat{M}_g(\lambda_g, K), g = 1, \ldots, p, i = 1, \ldots, n$ are estimated regression parameters and estimated number of groups.

The performance of our method is evaluated in three aspects: subgroup identification accuracy, parameter estimation, and partition recovery with group membership prediction. To assess subgroup identification, Table 1 reports the median, bias, and standard deviation (s.d.) of the estimated number of groups, the average Rand Index (RI) for clustering accuracy, and the percentage (per) of $\widehat{M}_g$ matching the true number of subgroups, where $per = B^{-1}\sum_{i=1}^{B} I(\widehat{M}_i = M^0)$. The Rand Index (RI) is computed by RI = (TP+TN)/(TP+FP+FN+TN), where TP,TN,FP and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively. The Rand Index (RI), which ranges between 0 and 1, reflects clustering accuracy, with higher values indicating better performance in identifying the latent group structure. From Table 1, we observe that the medians of $\widehat{M}_g(g = 1, \cdots, p)$ match the true number of subgroups across all cases. As the sample size n increases, both the RI values and the proportion of correctly selecting the number of subgroups approach 1, indicating enhanced clustering performance.

[Table 1 about here.]

To evaluate the performance of partition recovery, Table 1 also presents the average accuracy (ACC) of the partition recovery. The value of $\text{ACC}_g, g = 1, \cdots, p$ defined by $\text{ACC}_g = \frac{1}{n}\sum_{m=1}^{M_g^0}\sum_{i\in G_{gm}^0} I\{w_{im}^g = \widehat{w}_{im}^g\}$, in which $w_{im}^g = 1$ for $i \in G_{gm}^0$, $w_{im}^g = 0$ otherwise, and $\widehat{w}_{im}^g = 1$ for $U_i \in \widehat{\mathcal{U}}_{gm}$ and $\widehat{w}_{im}^g = 0$ otherwise. As n increases, the accuracy approaches 1. To assess the accuracy of the group membership predictions, we tested the proposed method by predicting the group membership of 200 test subjects using the recovered partitions. ACC_g^p $(g = 1, \cdots, p)$ show the average prediction accuracy for the 200 testing samples. From Table 1, it is clear that ACC-

related indices are very close to 1, indicating that both the proposed partition recovery method and the membership prediction based on it perform exceptionally well. Additionally, Figure 2 visualizes the recovered partitions for cases 1-5 when $n = 400$. Results for $n = 800$ are provided in the Supplementary Materials A3.4. to conserve space. In this figure, the dotted curves represent the true boundary curves, while the differently colored areas depict the estimated partition blocks. The boundaries formed by adjacent estimated blocks represent the estimated boundaries. Figure 2 shows the estimated partition blocks closely align with the true partitions in most cases. For Case 5, the recovered partitions of $[\boldsymbol{\alpha}]_3$ through $[\boldsymbol{\alpha}]_{\lfloor n^{1/3} \rfloor}$ exhibit patterns similar to those of $[\boldsymbol{\alpha}]_1, [\boldsymbol{\alpha}]_2$. For conciseness, we present only the first two dimensions as representative cases in Figure 2.

[Figure 2 about here.]

[Table 2 about here.]

Finally, to assess the performance of the proposed estimator for the regression coefficients, Table 2 presents simulation results for the post-group-recovery estimator across cases 1-5. We report the bias, empirical standard deviation (ESD), and mean squared error (MSE) of the post-group-recovery estimator. It is evident that the biases are very small across all cases. As n increases, the ESDs and MSEs decrease, demonstrating improved precision. To facilitate the comparison, Figure 3 presents the boxplots of the estimation errors (i.e., estimate minus true value) for all regression parameters obtained using the method in Li et al. (2021) and our proposed method based on 200 replications. From this figure, it is evident that the median of our estimates is closer to 0 compared to the estimates obtained using the method of Li et al. (2021), under both cases. Under Case 2, the variability of our estimates is also slightly smaller. Overall, both methods perform similarly under the classical covariate threshold model and the threshold change plane model (such as in Cases 1 and 2). However, for the remaining cases, our method continues to perform effectively, while their method fails to identify the subgroup structure and exhibits significant estimation errors. For the remaining comparison results, please refer to A3.4. in the supplementary materials.

[Figure 3 about here.]

6. Panitumumab trial data analysis

In this Section, we evaluate the proposed methodology using data from the panitumumab trial. Panitumumab is a fully human monoclonal antibody targeting the epidermal growth factor receptor (EGFR) that has been shown to improve progression-free survival (PFS) in patients with chemotherapy-refractory metastatic colorectal cancer (mCRC). The trial evaluated the efficacy and safety of panitumumab in combination with fluorouracil, leucovorin, and irinotecan (FOLFIRI) versus FOLFIRI alone among patients with mCRC after failure of initial treatment (Peeters et al., 2010). Patients were randomly assigned to one of two arms: panitumumab plus FOLFIRI (treatment 1) or FOLFIRI alone (treatment 0). The primary objective was to assess which regimen provides superior efficacy.

After excluding 95 records with missing data and 47 right-censored records, a total of 804 subjects were included in the analysis. Let the log-transformed progression-free survival day be Y_i, and let $Z_i \in \{0, 1\}$ indicate whether subject i received treatment 0 or 1. Baseline covariates include age (X_1), gender (X_2; female = 0, male = 1), prior bevacizumab exposure (X_3; no = 0, yes = 1), primary tumor site (X_4; rectal = 0, colon = 1), Kirsten rat sarcoma virus (KRAS) mutation status (X_5; mutant = 0, wild = 1), Eastern Cooperative Oncology Group (ECOG) performance status (X_6; fully active = 0, symptoms but ambulatory/in bed less than 50% of the time = 1), number of baseline metastatic sites (X_7), and liver metastasis status (X_8; no = 0, yes = 1). We first fit a homogeneous linear regression model with Y_i as the response and the intercept, the eight baseline covariates, and the treatment indicator Z_i as predictors. The estimated coefficient under the homogeneous model is 0.085 ($p < 0.001$), corresponding to a hazard ratio of approximately 1.09. While the overall effect is modest in magnitude, suggesting limited average clinical benefit at the population level, the statistically significant signal may still reflect underlying biological heterogeneity and supports further investigation of subgroup-specific treatment responses. This is consistent with the

limited overall efficacy that initially led the European Medicines Agency to decline panitumumab's application for mCRC.

[Figure 4 about here.]

Existing studies (e.g., (Amado et al., 2008; Peeters et al., 2015)) defined subgroups based solely on wild-type KRAS status, thereby establishing the efficacy of panitumumab in wild-type KRAS patients. However, this strategy would not capture the more nuanced finding reported by Luo and Guo (2023): among female patients with wild-type KRAS, those aged 65 years or older derived no clinical benefit. This motivates us to identify potential subgroups by incorporating additional covariates, such as age, beyond wild-type KRAS status. Here, we fit the proposed heterogeneous linear model $Y_i = \boldsymbol{X}_i^T\beta + Z_i\alpha_i(\boldsymbol{U}_i) + \epsilon_i$, $i = 1, \cdots, n$, where $\boldsymbol{X}_i$ includes the intercept and the eight baseline covariates described above, Z_i is the treatment indicator, and $\boldsymbol{U}_i$ denotes the threshold variables. Motivated by Luo and Guo (2023), we set $\boldsymbol{U}_i = (U_{1i}, U_{2i}, U_{3i}, U_{4i}, U_{5i})^T$, where U_1 is age, U_2 is gender (female = 0, male = 1), U_3 is KRAS mutation status (mutant = 0, wild = 1), U_4 is ECOG performance status (fully active = 0, symptoms but ambulatory/in bed less than 50% of the time = 1), and U_5 is the number of baseline metastatic sites. All predictors are centered and standardized before applying the regularization methods.

Patients were assigned to three subgroups ($\widehat{G}_1, \widehat{G}_2, \widehat{G}_3$) with sizes 134, 321, and 349, respectively. Performing post-grouping estimation based on the identified subgroup structure, we obtain the following treatment-effect estimates: $\widehat{\theta}_1 = -0.369$ (statistically significant, $p = 0.000$), $\widehat{\theta}_2 = 0.027$ (not significant, $p = 0.327$), and $\widehat{\theta}_3 = 0.352$ (statistically significant, $p = 0.000$). These results suggest that panitumumab plus FOLFIRI provides superior efficacy relative to FOLFIRI alone for patients in subgroup $\widehat{G}_3$. In contrast, no significant difference is observed between the two regimens for patients in subgroup $\widehat{G}_2$, whereas for those in subgroup $\widehat{G}_1$, panitumumab is associated with adverse effects. These findings differ from the prevailing view that panitumumab provides clinical benefit in wild-type KRAS patients.

In addition, we recover the subgroup partitions using a Gaussian-kernel support vector machine.

The prediction accuracy for subgroup membership is 92.04%, indicating strong performance of the proposed partition-recovery step. We further examine the distributions of the threshold variables across subgroups. Figures 4(a)–(d) display the counts for gender, KRAS mutation status, ECOG performance status, and number of baseline metastatic sites within $\widehat{G}_1$, $\widehat{G}_2$, and $\widehat{G}_3$. The boxplots of the threshold variable age at baseline across subgroups are presented in Figure 4(e). Based on Figures 4(a)–(e), the distributions of the discrete threshold variables (gender, KRAS mutation status, and ECOG performance status) in subgroup $\widehat{G}_1$ exhibit a distinct pattern relative to subgroups $\widehat{G}_2$ and $\widehat{G}_3$, whereas the latter two subgroups are broadly similar for these variables. By contrast, the number of baseline metastatic sites shows markedly different patterns between $\widehat{G}_2$ and $\widehat{G}_3$. Figure 4(e) further shows that both the mean and median baseline age in subgroup $\widehat{G}_2$ are higher than those in the other two subgroups, which is partially consistent with the pattern noted by Luo and Guo (2023).

For younger male patients with wild-type KRAS, if their ECOG score is 1 (indicating mild symptoms) and they have a high baseline number of metastatic sites, treatment 1 demonstrates clear superiority over treatment 0. In contrast, for female patients with mutant KRAS who are fully active (ECOG = 0) and have fewer than two baseline metastatic sites (suggesting a favorable prognosis), treatment 0 is recommended. Partial results align with previous findings. Luo and Guo (2023) showed, via formal statistical testing, that among patients with more severe disease (ECOG = 1, number of metastatic sites $>$ 2) and wild-type KRAS, panitumumab–FOLFIRI achieved greater efficacy. This concordance further supports the validity and feasibility of our approach. Importantly, existing assessments of treatment-effect heterogeneity for panitumumab–FOLFIRI (Peeters et al., 2010, Peeters et al., 2015) typically presuppose a binary subgroup structure based solely on KRAS status. In contrast, by incorporating a broader set of patient characteristics, including gender, overall health status, and other clinically relevant covariates, the proposed framework provides a data-driven basis for identifying more refined subgroups beyond conventional genetic markers. This

perspective also prompts reconsideration of whether apparent non-responsiveness reflects true treatment indifference or potential adverse effects, thereby enabling more nuanced clinical decision-making. Even with a 5.5% censoring rate, excluding censored subjects may lead to biased results. As a sensitivity analysis, we assigned Y_i to be the censoring time for censored observations and repeated the analysis. Results of this additional analysis are presented in the Supplementary Material and show broadly similar patterns. Extending the proposed method to directly accommodate right-censored outcomes remains an important direction for future research.

7. Discussion

It is important to note that our current focus is on linear regression with i.i.d. data. An important direction for future research is to extend this method to a broader range of models and/or complex data structures. Promising generalizations notably include survival (e.g., Cox) and spatial (e.g., autoregression) models, as well as applications to right-censored, longitudinal, panel, missing, and spatiotemporal data. Moreover, in real applications, it is crucial to distinguish among the covariate sets $\boldsymbol{U}$, $\boldsymbol{Z}$, and $\boldsymbol{X}$ before applying the proposed method. First, the threshold variables $\boldsymbol{U}_i$ can be chosen based on the scientific objective, domain knowledge, and preliminary analyses. For example, in our real-data analysis, age and KRAS mutation status are selected as U_1 and U_3 based on findings in Peeters et al. (2010), Peeters et al. (2015), and Amado et al. (2008). Gender and ECOG performance status are selected as U_2 and U_4 based on Luo and Guo (2023). The number of baseline metastatic sites is selected as U_5, reflecting our interest in whether patients with a more favorable prognosis (ECOG $= 0$, metastatic sites < 3) tend to exhibit different treatment effects from those with a poorer prognosis.

Next, to distinguish covariates with homogeneous versus heterogeneous effects, one pragmatic strategy is to initially treat all remaining covariates as candidates in $\boldsymbol{Z}_i$ and apply the fused regularization procedure to all of them. The proposed subgroup identification step can then estimate the number of subgroups associated with each covariate effect. Covariates whose effects are estimated

to be homogeneous (i.e., forming a single group) can be assigned to $\boldsymbol{X}$, whereas covariates whose effects split into two or more subgroups can be retained in $\boldsymbol{Z}$.

In our approach, the statistical significance (p-values) associated with subgroup treatment effects does not account for the uncertainty in subgroup detection. Addressing how to incorporate this type of uncertainty is an important direction for future research. We also acknowledge that, in some settings, the identified subgroups may overlap substantially in covariate distributions and therefore lack clean, clinically interpretable boundaries. In our data analysis, for example, although age and gender exhibit some distributional differences across subgroups, subgroup membership is not easily characterized by simple rules. Consequently, assigning a new patient to a subgroup may still require evaluating their covariates through the fitted SVM classifier rather than relying on bedside heuristics, which limits immediate clinical interpretability. More broadly, as with many flexible, data-driven subgroup discovery methods, the treatment-effect heterogeneity detected by our approach may be statistically meaningful yet difficult to translate into clinically actionable guidance, and increasingly fine partitions of the covariate space can yield subgroups with opposing average effects. Future work will therefore focus on improving interpretability and clinical usability, for example by developing parsimonious surrogate rules, post hoc explanations of the learned partitions, and stability assessments to help distinguish robust heterogeneity from modeling artifacts.

Acknowledgements

Chen's work was supported by National Key R&D Program of China (2022YFA1003702), the National Natural Science Foundation of China (NSFC) (12371296,12531011), and Guanghua Talent Project of Southwestern University of Finance and Economics. Guo's work was partially supported by grants from Research Grants Council of the Hong Kong Special Administrative Region, China (HKUST 26308323 and HKUST 16310125), the Seed fund of the Big Data for Bio-Intelligence Laboratory (Z0428) and the grant L0438 from the Hong Kong University of Science

and Technology. We thank the Co-Editor, Associate Editor and two reviewers for their constructive comments that have substantially improved the paper.

Supplementary materials

Web Appendices, Tables, and Figures referenced in Sections 3, 4, 5 are available at the Biometrics website on Oxford Academic. Supplementary Materials A1 and A2 includes some notation and formulas from the Sections 3, 4. Supplementary Materials A3 present the alternating directions method of multipliers algorithm, the choice of its initial values, extra simulation results, real data analyses and detailed proofs of theoretical results. A Zip file containing the code and data used in Sections 5 and 6 is available with this paper at the Biometrics website on Oxford Academic.

Data Availability Statement

The data supporting the findings in this paper are available in Project Data Sphere at https://doi.org/10.34949/0ws2-w454 and can be accessed upon request, subject to platform approval.

References

Alosh, M., Fritsch, K., Huque, M., Mahjoob, K., Pennello, G., Rothmann, M., and et al. (2015). Statistical considerations on subgroup analysis in clinical trials. *Statistics in Biopharmaceutical Research* **7,** 286–303.

Amado, R. G., Wolf, M., Peeters, M., Van Cutsem, E., Siena, S., Freeman, D. J., and et al. (2008). Wild-type kras is required for panitumumab efficacy in patients with metastatic colorectal cancer. *Journal of Clinical Oncology* **26,** 1626–1634.

Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet* **355,** 1064–1069.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19,** 521–547.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., and et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3,** 1–122.

Dahabreh, I. J., Hayward, R., and Kent, D. M. (2016). Using group data to treat individuals:

Understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *International Journal of Epidemiology* **45,** 2184–2193.

Everitt, B. S. and Hand, D. J. (1981). Finite mixture distributions. *Monographs on Applied Probability and Statistics* .

Guelman, L., Guillén, M., and Pérez-Marín, A. M. (2015). A decision support framework to implement optimal personalized marketing interventions. *Decision Support Systems* **72,** 24–32.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **55,** 757–779.

He, Y., Zhou, L., Xia, Y., and Lin, H. (2023). Center-augmented ℓ_2-type regularization for subgroup learning. *Biometrics* **79,** 2157–2170.

Koo, J.-Y., Lee, Y., Kim, Y., and Park, C. (2008). A bahadur representation of the linear support vector machine. *Journal of Machine Learning Research* **9,** 1343–1368.

Lee, S., Seo, M. H., and Shin, Y. (2011). Testing for threshold effects in regression models. *Journal of the American Statistical Association* **106,** 220–231.

Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association* **99,** 67–81.

Li, J., Li, Y., Jin, B., and Kosorok, M. R. (2021). Multithreshold change plane model: Estimation theory and applications in subgroup identification. *Statistics in Medicine* **40,** 3440–3459.

Luo, Y. and Guo, X. (2023). Inference on tree-structured subgroups with subgroup size and subgroup effect relationship in clinical trials. *Statistics in Medicine* **42,** 5039–5053.

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

Madrid Padilla, O. H., Sharpnack, J., Chen, Y., and Witten, D. M. (2020). Adaptive nonparametric regression with the k-nearest neighbour fused lasso. *Biometrika* **107,** 293–310.

Peeters, M., Oliner, K. S., Price, T. J., Cervantes, A., Sobrero, A. F., Ducreux, M., and et al. (2015). Analysis of KRAS/NRAS mutations in a phase III study of panitumumab with FOLFIRI compared with FOLFIRI alone as second-line treatment for metastatic colorectal

cancer. *Clinical Cancer Research* **21,** 5469–5479.

Peeters, M., Price, T. J., Cervantes, A., Sobrero, A. F., Ducreux, M., Hotko, Y., and et al. (2010). Randomized phase III study of panitumumab with fluorouracil, leucovorin, and irinotecan (FOLFIRI) compared with FOLFIRI alone as second-line treatment in patients with metastatic colorectal cancer. *Journal of Clinical Oncology* **28,** 4706–4713.

Shen, J. and He, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association* **110,** 303–312.

Shen, J., Wang, Y., and He, X. (2017). Penalized likelihood for logistic-normal mixture models with unequal variances. *Statistica Sinica* **27,** 711–731.

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

Teng, H. Y. and Zhang, Z. (2024). Two-way truncated linear regression models with extremely thresholding penalization. *Journal of the American Statistical Association* **119,** 887–903.

Vapnik, V. (1998). The support vector method of function estimation. In *Nonlinear Modeling: Advanced Black-Box Techniques*, pages 55–85. Springer.

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

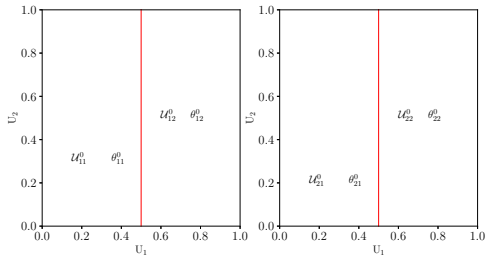
Zhang, L., Wang, H. J., and Zhu, Z. (2014). Testing for change points due to a covariate threshold in quantile regression. *Statistica Sinica* **24,** 1859–1877.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics* **32,** 56–85.

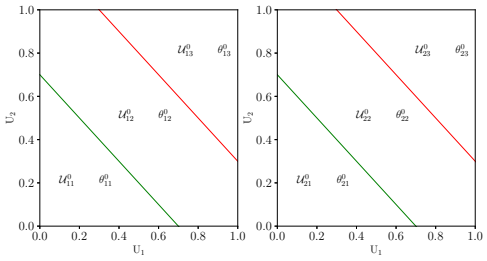
Zhang, Y., Wang, H. J., and Zhu, Z. (2022). Single-index thresholding in quantile regression. *Journal of the American Statistical Association* **117,** 2222–2237.

Zhang, Y., Zhao, Y.-Y., and Lian, H. (2022). Statistical rates of convergence for functional partially linear support vector machines for classification. *Journal of Machine Learning Research* **23,** 1–24.

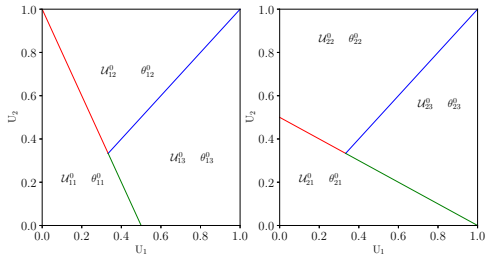
Zhou, L., Sun, S., Fu, H., and Song, P. X.-K. (2022). Subgroup-effects models for the analysis of personal treatment effects. *Annals of Applied Statistics* **16,** 80–103.



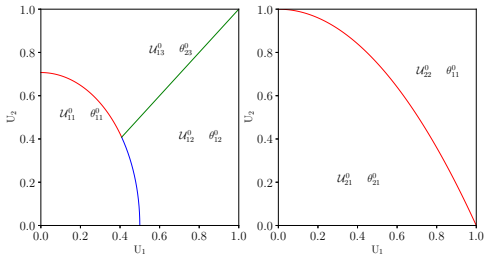
(a) Scenario *I*



(b) Scenario *II*



(c) Scenario *III*



(d) Scenario *IV*

Figure 1: Simple examples of group structure when $r = 2$, $p = 2$: $\{\mathcal{U}^0_{gm}, m = 1, \ldots, M^0\}$ is a partition and θ^0_{gm} is the corresponding subgroup coefficient. The solid lines are the partition boundaries. In (a) Scenario *I*: red lines in both correspond to $U_1 = 0.5$; In (b) Scenario *II*: red lines in both $U_1 + U_2 = 1.3$, green lines in both $U_1 + U_2 = 0.7$; In (c) Scenario *III*: left: red and green line $U_1 + 0.5U_2 = 0.25$, blue line $U_1 - U_2 = 0$, right: red and green line $U_1 + 2U_2 = 1$, blue line $U_1 - U_2 = 0$; In (d) Scenario *IV*: left: red and blue line $U_1^2 + 0.5U_2^2 = 0.5$, green line $U_1 - U_2 = 0$, right: red line $U_1^2 + U_2 = 1$.

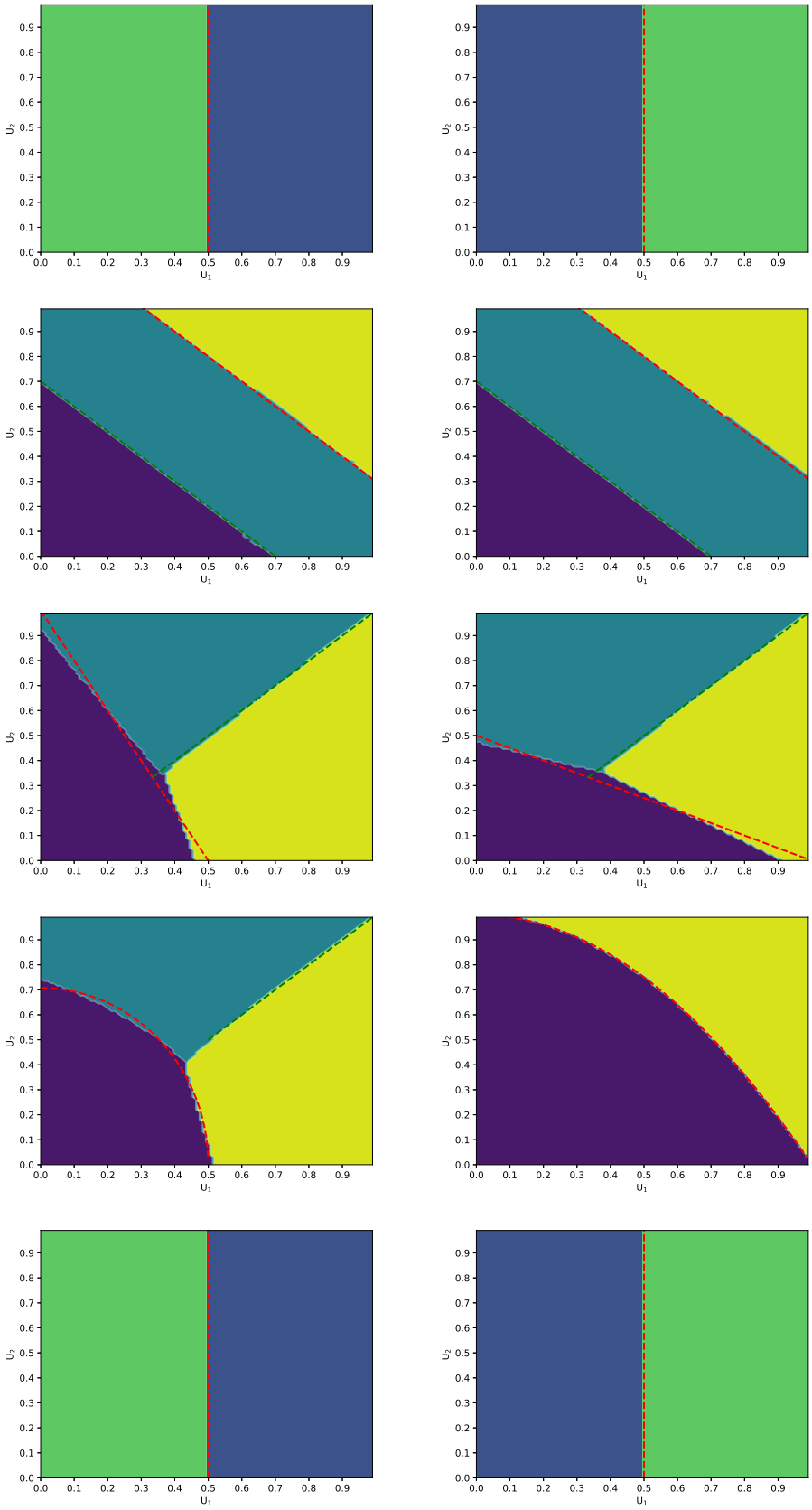
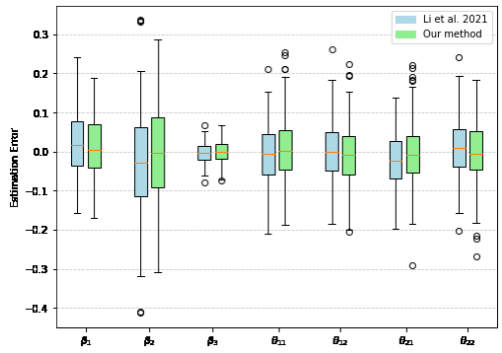
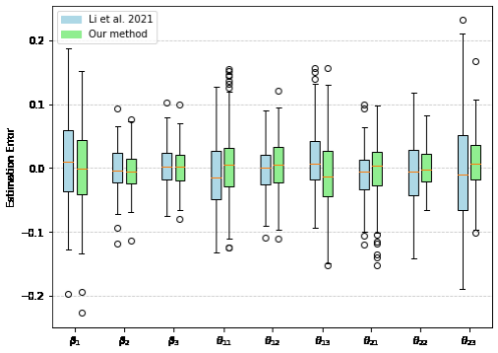


Figure 2: The partition recovery results based on $n = 400$ for Case 1-5. The first column are graphs for $[\alpha]_1$ in case 1-5 while second row are for $[\alpha]_2$.

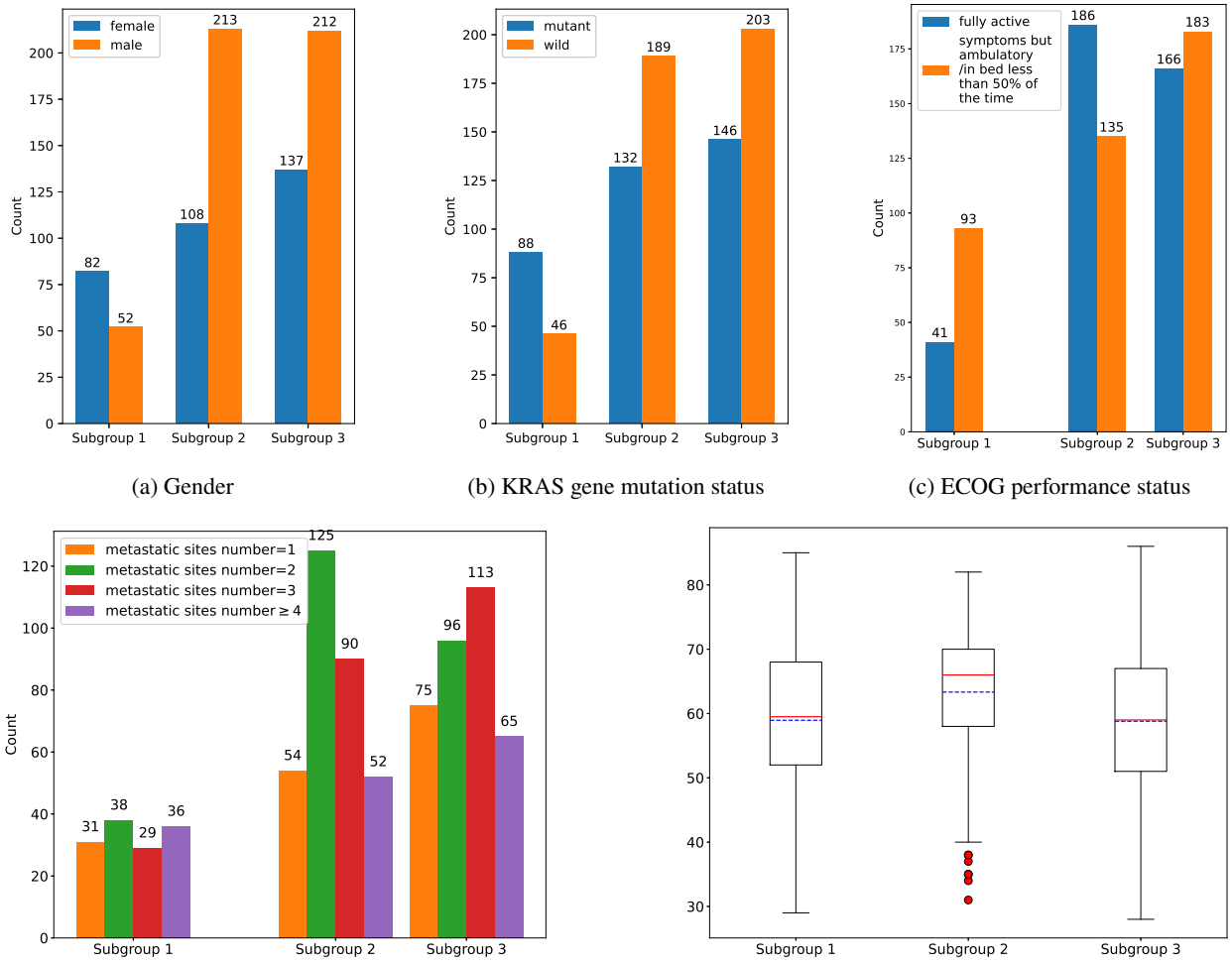


(a) Case 1



(b) Case 2

Figure 3: Boxplots of parameter estimation errors in different methods. (a) Case 1 (b) Case 2.



(a) Gender (b) KRAS gene mutation status (c) ECOG performance status

(d) Baseline metastatic site number (e) Age

Figure 4: The counts of discrete threshold variables across subgroups. (a) Gender, (b) KRAS gene mutation status, (c) ECOG performance status and (d) Baseline metastatic site number. (e) Box Plots of Baseline Age Across Subgroups, solid line: mean; dashed line: median.

Table 1: The sample median,bias and standard deviation (s.d.) of $\widehat{M}_1, \widehat{M}_2$, the Randex Index(RI) value and the percentage (per) of $\widehat{M}_1, \widehat{M}_2$ equaling to the true number of subgroups; the average accuracy (ACC) of the classifier's based on test sample $n' = 100$ in Case 1-5.

		n = 400							n = 800						
Case		median	bias	s.d.	RI	per	ACC_g	ACC_g^p	median	bias	s.d.	RI	per	ACC_g	ACC_g^p
1	$\widehat{M}_1$	2.000	0.035	0.184	0.971	0.979	0.994	0.965	2.000	0.000	0.000	0.983	1.000	0.996	0.993
	$\widehat{M}_2$	2.000	0.015	0.122	0.989	0.985	0.989	0.987	2.000	0.000	0.000	0.984	1.000	0.993	0.991
2	$\widehat{M}_1$	3.000	0.000	0.000	0.972	1.000	0.972	0.970	3.000	0.000	0.000	0.980	1.000	0.980	0.984
	$\widehat{M}_2$	3.000	0.000	0.000	0.986	1.000	0.982	0.977	3.000	0.000	0.000	0.991	1.000	0.988	0.987
3	$\widehat{M}_1$	3.000	-0.005	0.071	0.958	0.995	0.958	0.956	3.000	0.000	0.000	0.961	1.000	0.969	0.968
	$\widehat{M}_2$	3.000	-0.035	0.210	0.948	0.970	0.969	0.963	3.000	-0.010	0.099	0.956	0.990	0.976	0.973
4	$\widehat{M}_1$	3.000	0.000	0.000	0.967	1.000	0.975	0.966	3.000	0.000	0.000	0.974	1.000	0.979	0.977
	$\widehat{M}_2$	2.000	-0.025	0.157	0.970	0.975	0.990	0.975	2.000	0.000	0.000	0.959	1.000	0.992	0.985
5	$\widehat{M}_1$	2.000	-0.005	0.071	0.969	0.995	0.986	0.985	2.000	0.000	0.000	0.969	1.000	0.990	0.990
	$\widehat{M}_2$	2.000	0.025	0.157	0.975	0.975	0.987	0.987	2.000	-0.005	0.071	0.978	0.995	0.990	0.990
	$\widehat{M}_3$	2.000	0.030	0.171	0.974	0.970	0.986	0.986	2.000	-0.005	0.071	0.979	0.995	0.991	0.991
	$\widehat{M}_4$	2.000	0.000	0.000	0.975	1.000	0.988	0.986	2.000	0.000	0.000	0.976	1.000	0.992	0.991
	$\widehat{M}_5$	2.000	0.000	0.000	0.975	1.000	0.987	0.987	2.000	0.000	0.000	0.973	1.000	0.991	0.990
	$\widehat{M}_6$	2.000	0.010	0.100	0.974	0.990	0.986	0.986	3.000	-0.010	0.100	0.976	0.990	0.990	0.990
	$\widehat{M}_7$	2.000	0.000	0.000	0.974	1.000	0.987	0.975	2.000	0.000	0.000	0.980	1.000	0.990	0.990
	$\widehat{M}_8$								2.000	-0.005	0.071	0.978	0.995	0.989	0.981
	$\widehat{M}_9$								2.000	0.000	0.000	0.979	1.000	0.989	0.982

Table 2: The bias, empirical standard deviation (ESD) and mean squared error (MSE) of the post-group-recovery estimator

		n = 400			n = 800		
Case		bias	ESD	MSE	bias	ESD	MSE
	$\widehat{\beta}_0^{post}$	-0.0015	0.0952	0.0091	-0.0041	0.0598	0.0036
	$\widehat{\beta}_1^{post}$	0.0087	0.1480	0.0220	0.0137	0.0968	0.0096
	$\widehat{\beta}_2^{post}$	-0.0005	0.0259	0.0007	-4e-05	0.0178	0.0003
1	$\widehat{\theta}_{11}^{post}$	-0.0153	0.0767	0.0061	-0.0151	0.0505	0.0028
	$\widehat{\theta}_{12}^{post}$	0.0225	0.0722	0.0057	0.0055	0.0530	0.0028
	$\widehat{\theta}_{21}^{post}$	-0.0294	0.0832	0.0078	-0.0118	0.0542	0.0031
	$\widehat{\theta}_{22}^{post}$	0.0167	0.0806	0.0068	0.0137	0.0576	0.0035
	$\widehat{\beta}_0^{post}$	-0.0038	0.0627	0.0040	0.0037	0.0462	0.0022
	$\widehat{\beta}_1^{post}$	-0.0029	0.0297	0.0009	-0.0013	0.0201	0.0004
	$\widehat{\beta}_2^{post}$	0.0006	0.0301	0.0009	0.0023	0.0204	0.0004
	$\widehat{\theta}_{11}^{post}$	0.0042	0.0541	0.0029	0.0056	0.0385	0.0015
2	$\widehat{\theta}_{12}^{post}$	0.0048	0.0388	0.0015	-0.0008	0.0256	0.0007
	$\widehat{\theta}_{13}^{post}$	-0.0088	0.0572	0.0033	-0.0056	0.0387	0.0015
	$\widehat{\theta}_{21}^{post}$	-0.0018	0.0418	0.0017	-0.0059	0.0288	0.0009
	$\widehat{\theta}_{22}^{post}$	-0.0005	0.0309	0.0010	-0.0003	0.0205	0.0004
	$\widehat{\theta}_{23}^{post}$	0.0096	0.0437	0.0020	0.0036	0.0266	0.0007
	$\widehat{\beta}_0^{post}$	0.0067	0.0694	0.0049	-0.0010	0.0464	0.0022
	$\widehat{\beta}_1^{post}$	0.0019	0.0307	0.0009	-0.0021	0.0207	0.0004
	$\widehat{\beta}_2^{post}$	-0.0038	0.0319	0.0010	-0.0011	0.0201	0.0004
	$\widehat{\theta}_{11}^{post}$	0.0077	0.0515	0.0027	0.0035	0.0354	0.0013
3	$\widehat{\theta}_{12}^{post}$	-0.0055	0.0570	0.0033	0.0013	0.0391	0.0015
	$\widehat{\theta}_{13}^{post}$	0.0099	0.0494	0.0025	0.0073	0.0300	0.0010
	$\widehat{\theta}_{21}^{post}$	-0.0095	0.0483	0.0024	-0.0027	0.0307	0.0010
	$\widehat{\theta}_{22}^{post}$	-0.0032	0.0398	0.0016	-0.0002	0.0254	0.0006
	$\widehat{\theta}_{23}^{post}$	-0.0059	0.0364	0.0014	-0.0033	0.0266	0.0007
	$\widehat{\beta}_0^{post}$	-0.0683	0.1449	0.0276	-0.0521	0.1009	0.0129
	$\widehat{\beta}_1^{post}$	0.0835	0.1725	0.0367	0.0623	0.1164	0.0174
	$\widehat{\beta}_2^{post}$	0.0421	0.1548	0.0257	0.0433	0.1013	0.0121
	$\widehat{\theta}_{11}^{post}$	0.0184	0.0476	0.0026	0.0137	0.0321	0.0012
4	$\widehat{\theta}_{12}^{post}$	-0.0022	0.0561	0.0032	-0.0050	0.0369	0.0014
	$\widehat{\theta}_{13}^{post}$	-0.0146	0.0423	0.0020	-0.0069	0.0300	0.0009
	$\widehat{\theta}_{21}^{post}$	0.0044	0.0333	0.0011	0.0015	0.0231	0.0005
	$\widehat{\theta}_{22}^{post}$	-0.0082	0.0411	0.0018	-0.0082	0.0294	0.0009
	$\widehat{\beta}_0^{post}$	-0.0055	0.1538	0.0237	-0.0042	0.1102	0.0122
	$\widehat{\beta}_1^{post}$	0.0025	0.2136	0.0456	0.0084	0.1422	0.0203
	$\widehat{\beta}_2^{post}$	-0.0006	0.0357	0.0013	0.0005	0.0249	0.0006
	$\widehat{\theta}_{11}^{post}$	-0.0216	0.0915	0.0088	-0.0283	0.0785	0.0070
	$\widehat{\theta}_{12}^{post}$	0.0051	0.1038	0.0108	0.0164	0.0803	0.0067
	$\widehat{\theta}_{21}^{post}$	0.0505	0.1146	0.0157	0.0499	0.0954	0.0116
	$\widehat{\theta}_{22}^{post}$	-0.0456	0.1075	0.0136	-0.0488	0.0859	0.0098
	$\widehat{\theta}_{31}^{post}$	0.0433	0.1001	0.0119	0.0456	0.0930	0.0107
	$\widehat{\theta}_{32}^{post}$	-0.0566	0.1121	0.0164	-0.0447	0.0849	0.0092
	$\widehat{\theta}_{41}^{post}$	0.0521	0.1219	0.0176	0.0461	0.0785	0.0083
5	$\widehat{\theta}_{42}^{post}$	-0.0305	0.0961	0.0109	-0.0412	0.0807	0.0082
	$\widehat{\theta}_{51}^{post}$	0.0371	0.0901	0.0091	0.0603	0.1071	0.0151
	$\widehat{\theta}_{52}^{post}$	-0.0389	0.1118	0.0140	-0.0428	0.0840	0.0089
	$\widehat{\theta}_{61}^{post}$	0.0455	0.1177	0.0168	0.0504	0.0922	0.0110
	$\widehat{\theta}_{62}^{post}$	-0.0588	0.1167	0.0172	-0.0542	0.0840	0.0100
	$\widehat{\theta}_{71}^{post}$	0.0334	0.1077	0.0127	0.0446	0.0864	0.0095
	$\widehat{\theta}_{72}^{post}$	-0.0252	0.0904	0.0088	-0.0436	0.0767	0.0078
	$\widehat{\theta}_{81}^{post}$				0.0584	0.1023	0.0139
	$\widehat{\theta}_{82}^{post}$				-0.0474	0.0907	0.0115
	$\widehat{\theta}_{91}^{post}$				0.0254	0.0789	0.0069
	$\widehat{\theta}_{92}^{post}$				-0.0254	0.0747	0.0062