Submitted to the Annals of Statistics

INDIVIDUALIZED MULTI-DIRECTIONAL VARIABLE SELECTION*

BY XIWEI TANG*,† AND ANNIE QU*,‡

University of Illinois at Urbana-Champaign†, University of Illinois at Urbana-Champaign‡

*Abstract*In this paper we propose an individualized variable selection approach to select different relevant variables for different individuals. In contrast to conventional model selection approaches, the key component of the new approach is to construct a separation penalty with multi-directional shrinkages including zero, which facilitates individualized modeling to distinguish strong signals from noisy ones. As a byproduct, the proposed model identifies subgroups among which individuals share similar effects, and thus improves estimation efficiency and personalized prediction accuracy. Another advantage of the proposed model is that it can incorporate within-subject correlation for longitudinal data. We provide a general theoretical foundation under a double-divergence modeling framework where the number of subjects and the number of repeated measurements both go to infinity, and therefore involves high-dimensional individual parameters. In addition, we present the oracle property for the proposed estimator to ensure its optimal large sample property. Simulation studies and an application to HIV longitudinal data are illustrated to compare the new approach to existing penalization methods.

1. Introduction. In recent years there has been a growing demand for exploring individualized modeling, which has broad applications in personalized medicine, personalized education and personalized marketing. The traditional one-model-fits-the-whole-population approach is unable to detect important patterns and make personalized predictions for specific individuals. In addition, the rise of precision medicine and wide-spread electronic health record data also motivate us to develop more effective personalized treatment. The collection of rich data information makes it feasible and compelling to utilize individualized models as traditional population models cannot incorporate heterogeneous effects from different individuals.

In this paper, we consider an individualized model based on a double-divergence framework, where the number of subjects and the amount of individual information increase together. Consequently, this introduces a diverging number of parameters as the sample size of subjects increases. In addition, one unique challenge of individualized model selection is that there could be different relevant or important predictors for different subjects. For instance, different individuals may have different prognostic factors associated with the same disease. Therefore it is important to develop new statistical methodology and theory for variable selection and estimation for individualized modeling.

*Research is supported in part by National Science Foundation Grants DMS-1308227 and DMS-1415308.

Keywords and phrases: correlated data, double-divergence, Lasso, oracle property, separation penalty, subject-wise modeling, subpopulation

In the past two decades several penalized model selection methods have been developed, e.g., the Lasso [27], the smoothly clipped absolute deviation (SCAD) [6], the elastic net [36], the adaptive Lasso [37], the group Lasso [34], the minimax concave penalty (MCP) [35] and the truncated L_1-penalty (TLP) [25]. However, the above methods are based on a homogeneous model setting which selects predictors for entire populations. For the individualized model, we can employ traditional variable selection methods separately for each subject, if there are multiple observations from each subject as in longitudinal data settings. However, in practice, the number of measurements for particular individuals could be limited. In addition, it is likely that some variables are invariant for the same subject, such as demographic information variables, e.g., race and gender, which impose restrictions and additional obstacles to performing individualized variable selection based on a standard subject-wise model framework.

Another limitation of applying standard subject-wise variable selection is that it ignores information from other subjects which might share similar effects on important predictors of interest. Moreover, assuming each individual to have unique effects for all covariates is practically unrealistic and computationally infeasible. In contrast, it is more sensible to assume that a subpopulation of individuals share common effects on selected predictors. In addition, borrowing information from homogeneous subgroups allows one to increase estimation efficiency and model selection accuracy.

In order to utilize cross-subject information, one may assume that an underlying subpopulation structure depends on unobserved covariates. Existing approaches dealing with clustering on regression coefficients include mixture modeling for regression, such as the mixture-of-experts model [12]. However, most model selection approaches under this framework including [22], [20] and [9] only focus on choosing informative variables to distinguish different subgroups, rather than on selecting relevant predictors for different individuals.

Alternative approaches to model-based clustering on regression coefficients employ grouping penalization. For example, [29] propose a fused Lasso by adding an L_1-penalty to the pair of adjacent coefficients; [3] propose a clustering algorithm for regression by imposing a special octagonal shrinkage penalty on each pair of coefficients; [24] develop a grouping pursuit algorithm utilizing the truncated L_1-penalty for fusions, and [14] propose a data-driven segmentation method to explore homogeneous groups with regression. Nevertheless, these are all still under the population-regression model, and do not allow different individuals to have different features. For the purpose of subgrouping different individuals, [11] and [17] formulate clustering as a penalized regression problem by adopting an L_p-fusion penalty. [21] and [18] apply non-convex fusion penalties to solve the bias problem. However, the fusion-type of penalty focuses on subgrouping rather than on model selection for individual coefficients.

In this paper, we propose an effective individualized model selection approach utilizing multi-directional shrinkage to select unique relevant variables for different individuals. To the best of our knowledge, this is a new approach which has not been offered in the existing literature.

Specifically, the proposed penalty allows multiple possible shrinking directions including the one towards zero, which differs from conventional penalty functions with shrinking direction towards zero only. The consequence of conventional penalty functions is that non-zero signals could suffer from zero-directional shrinkage, although a variety of penalty methods have been proposed

to solve the bias problem such as non-concave penalties (e.g., SCAD, MCP and TLP) or adaptive weights (e.g., adaptive Lasso). Instead we propose a rather different approach which shrinks penalized parameters to one of the multiple directions including zero, where the best shrinking direction is determined by the data itself. One advantage of the proposed method is that, as long as the candidate directions contain the one closest to the truth, the optimal large sample properties such as the oracle property hold by applying the L_1-type of penalty function in each direction.

Another advantage of the proposed method is that it separates different groups of individuals based on their effects on the same covariates. Indeed, the proposed penalty function is analogous to an objective function from center-based clustering, which can be viewed as a "separation penalty" among different individuals. As a byproduct, we identify subgroups with individuals sharing similar covariate effects, where the centers of subgroups provide a set of estimated shrinking directions. In addition, through utilizing cross-subject information, the proposed model improves estimation efficiency and thus enhances personalized prediction power.

Another contribution of this paper is that we lay out a theoretical framework for the double-divergence individualized model with serial correlation. [33] and [1] established rigorous large sample theory for the generalized estimating equation [16] (GEE) estimator when the number of clusters and the cluster size are both large while the dimension of parameters is fixed; and [32] investigate the GEE model with high-dimensional covariates, but bounded cluster size. In contrast we establish theoretical properties in a framework when the number of clusters and the cluster size are both increasing, which involves high-dimensional parameters. We develop asymptotic theory for the oracle estimator and demonstrate the subpopulation effects on model estimation. In addition, we show the advantage of utilizing the multi-directional penalty for establishing the oracle property. Moreover, the proposed method is capable of incorporating within-subject correlation to achieve efficient estimation.

The paper is organized as follows. Section 2 introduces the model framework and presents the proposed methodology. Section 3 establishes the theoretical results. Section 4 proposes an efficient algorithm with implementation. Section 5 provides simulation studies. Section 6 illustrates an application for HIV data. The last section provides concluding remarks and discussion.

2. Model Framework and Methodology.

2.1. *The individualized model and subject-wise variable selection.* We formulate the problem under the clustered data setting, where each subject has multiple observations. Let $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{im_i})'$ be an m_i-dimensional response variable for the ith individual, $\boldsymbol{X_i} = (\boldsymbol{x_{i,1}}, \ldots, \boldsymbol{x_{i,p}})$ be an $m_i \times p$ covariates matrix corresponding to individual predictors, and $\boldsymbol{Z_i} = (\boldsymbol{z_{i,1}}, \ldots, \boldsymbol{z_{i,q}})$ be an $m_i \times q$ covariates matrix corresponding to population-shared predictors, where $i = 1, \ldots, N$. For ease of notation, we assume that the clustered data is balanced with cluster size $m_i = m$, although the development of the method does not require a balanced data structure.

We consider a regression model:

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta_i} + \boldsymbol{Z_i}\boldsymbol{\alpha} + \boldsymbol{\varepsilon_i}, \quad i = 1, \ldots, N,$$

where each individual has its own regression parameter vector $\boldsymbol{\beta_i} = (\beta_{i1}, \ldots, \beta_{ip})'_{p\times 1}$, in addition to the population-shared parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)'_{q\times 1}$, and random errors $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})'_{m\times 1}$ independent over different subjects. Within a subject, ε_{ij}'s ($j = 1, \ldots, m$) have mean 0 and variance σ^2, and they could be correlated such as in the longitudinal date setting.

In general, to identify unique features for different individuals, we select and estimate the regression parameters $\boldsymbol{\beta_i}$'s and $\boldsymbol{\alpha}$ through minimizing the penalized objective function

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) = \underset{\boldsymbol{\beta},\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{N} L(\boldsymbol{y_i} - \boldsymbol{\mu_i}) + \sum_{i=1}^{N}\sum_{k=1}^{p} h^{(1)}_{\lambda_1}(\beta_{ik}) + \sum_{l=1}^{q} h^{(2)}_{\lambda_2}(\alpha_l), \tag{1}$$

where $\boldsymbol{\mu_i}(\boldsymbol{\beta_i}, \boldsymbol{\alpha}) = \boldsymbol{X_i}\boldsymbol{\beta_i} + \boldsymbol{Z_i}\boldsymbol{\alpha}$, $L(\cdot)$ is a loss function, $h^{(1)}_{\lambda_1}(\cdot)$ and $h^{(2)}_{\lambda_1}(\cdot)$ are feature-selection penalties for individualized parameters and population-shared parameters respectively, and λ_1, λ_2 are the corresponding tuning parameters. The selection of population parameter $\boldsymbol{\alpha}$ is regular and thus, in this paper, we focus on individualized variable selection. To simplify the model, with a squared-error loss, the objective function in (1) becomes

$$\frac{1}{2}\sum_{i=1}^{N} \parallel \boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta_i} - \boldsymbol{Z_i}\boldsymbol{\alpha} \parallel_2^2 + \sum_{i=1}^{N}\sum_{k=1}^{p} h_{\lambda_{N,m}}(\beta_{ik}), \tag{2}$$

where $\parallel \cdot \parallel_2$ is the Euclidean norm. Then we could employ different penalties $h_{\lambda_{N,m}}(\cdot)$ to adopt traditional penalized selection approaches (e.g. Lasso, adaptive Lasso, MCP and SCAD).

Without the penalty term $h_{\lambda_{N,m}}(\cdot)$, minimizing (2) leads to the ordinary least squares (OLS) estimator. Let $\boldsymbol{\beta} = (\boldsymbol{\beta_1'}, \ldots, \boldsymbol{\beta_N'})'$ be the individualized coefficients vector and $\boldsymbol{Y} = (\boldsymbol{y_1'}, \ldots, \boldsymbol{y_N'})'$. We denote $\boldsymbol{X} = \operatorname{diag}(\boldsymbol{X_1}, \ldots, \boldsymbol{X_N})$, a block-diagonal matrix, and $\boldsymbol{Z} = (\boldsymbol{Z_1'}, \ldots, \boldsymbol{Z_N'})'$. The OLS estimator is

$$(\hat{\boldsymbol{\beta}}^{Sub'}, \hat{\boldsymbol{\alpha}}^{Sub'})' = [(\boldsymbol{X}, \boldsymbol{Z})^T(\boldsymbol{X}, \boldsymbol{Z})]^{-1}(\boldsymbol{X}, \boldsymbol{Z})^T\boldsymbol{Y},$$

whose dimension $(Np + q)$ diverges as subject size N increases.

Note that if there are no population-shared predictors, minimizing (2) is the same as minimizing the objective function for each individual (subject) separately. We call this approach subject-wise modeling; however, it only utilizes within-subject information. As a result, this leads to inefficient estimation and over-fitting of a model, especially when the sample size N is large and m is relatively small.

2.2. *The proposed model with multi-directional separation penalty.* We propose a novel penalized variable selection approach by providing multiple shrinking directions for individualized parameters and utilizing homogeneity information within the subpopulation, which performs parameter estimation, variable selection and subgrouping simultaneously.

For the kth ($k = 1, \ldots, p$) individualized predictor corresponding to the ith subject, we assume that there are $G_k + 1$ subgroups in the population such that

$$\beta_{ik} = \begin{cases} \boldsymbol{\gamma}_k^{(g)}, & \text{if} \quad i \in \mathcal{G}_k^{(g)}, \quad g = 1, \ldots, G_k \\ 0, & \text{if} \quad i \in \mathcal{G}_k^{(0)} \end{cases}, \tag{3}$$

where $\boldsymbol{\gamma}_k^{(g)}$ ($g \neq 0$) is an unknown non-zero parameter corresponding to the homogeneous coefficient for the gth subgroup, and $\mathcal{G}_k^{(g)}$'s are the index sets representing the subgroup memberships with respect to the kth predictor.

For ease of notation, in the following, we focus on the setting where there are two subgroups with respect to each individualized covariate: the non-zero-coefficient group ($\beta_{ik} = \gamma_k$) and the zero-coefficient group ($\beta_{ik} = 0$). We denote $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ as the sub-homogeneous effect vector. The extension to multiple subgroups is straightforward.

We first consider a model assuming within-subject independence. The extension to correlated data will be discussed later. The main idea is to encourage grouping of the subjects with similar effects on specific individualized predictors, by inducing the sub-homogeneous effect $\boldsymbol{\gamma}$ in the proposed objective function

$$Q_{N,m}^{ind}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \frac{1}{2} \sum_{i=1}^{N} \parallel \boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta_i} - \boldsymbol{Z_i}\boldsymbol{\alpha} \parallel_2^2 + \lambda_{N,m} \sum_{i=1}^{N} \sum_{k=1}^{p} s(\beta_{ik}, \gamma_k), \tag{4}$$

where $\lambda_{N,m}$ is the tuning parameter. Here the key part is the proposed multi-directional separation penalty (MDSP) function $s(\beta_{ik}, \gamma_k)$, defined as

$$s(\beta_{ik}, \gamma_k) = \min\Big(|\beta_{ik}|, |\beta_{ik} - \gamma_k| \Big), \tag{5}$$

which is a piece-wise L_1-penalization function (Figure 1).

The multi-directional penalty term in 4 essentially contains a double-summation providing two different perspectives of the proposed model. From a subject's point of view, the penalty term is $\sum_{k=1}^{p} s(\beta_{ik}, \gamma_k)$. In contrast to the traditional penalized variable selection approaches, the proposed MDSP function $s(\cdot)$ provides an alternative shrinking direction in addition to 0. Given γ_k, the $s(\cdot)$ penalty can be viewed as shrinking a weak signal of β_{ik} towards zero, while pulling the strong magnitude signals to γ_k. This reduces the bias for large coefficient estimators introduced by the L_p-penalty. Figure 1 illustrates the MDSP function $s(\beta_{ik}, \gamma_k)$ for a given γ_k, and Figure 2 provides plots of the thresholding functions of the Lasso and the proposed method. Without loss of generality, we assume $\gamma_k > 0$. Figure 2 indicates that when $\beta_{ik} > \gamma_k$ or $\beta_{ik} < 0$, $|\beta_{ik}|$ and $|\beta_{ik} - \gamma_k|$ have the same shrinking effect; and when $0 < \beta_{ik} < \gamma_k$, the two penalties produce different shrinking directions, which separates strong signals from weak signals.

From the other perspective, for one individualized predictor over different subjects, the MDSP term is $\sum_{i=1}^{N} s(\beta_{ik}, \gamma_k)$. Given $\beta_{ik}'s$, the proposed method leads to subgrouping the coefficients of individuals, where the separation-penalty term serves the role of centering, similar to K-means clustering. Compared to pairwise grouping penalization such as the fusion penalty, the center-based one has less computational cost, with $O(Np)$ penalty terms in contrast to the fusion-type of clustering containing $O(N^2p)$ penalty terms. This also implies that the computational cost of the proposed approach increases more slowly as the sample size N increases.

In addition, the unknown true effects γ_k's can be obtained simultaneously through minimizing the objective function in (4), where the estimation of γ_k utilizes information from individuals

within the subgroup. By pulling the coefficients' estimators towards the center $\hat{\gamma}_k$, it allows us to borrow cross-subject information for individuals' coefficient estimation, and therefore reduces the estimation bias and variance for non-zero coefficients.

Furthermore, the above two-subgroup model can be extended to multiple subgroups and even with additional constraints in practice. We illustrate the extension of three subgroups which allows positive and negative effects of personalized coefficients. The separation penalty imposed for three groups is

$$s(\beta_{ik}, \gamma_k^+, \gamma_k^-) = \min\left(|\beta_{ik}|, |\beta_{ik} - \gamma_k^+|, |\beta_{ik} - \gamma_k^-|\right), \quad \text{s.t.} \quad \gamma_k^+ > 0, \quad \gamma_k^- < 0, \tag{6}$$

which shrinks the coefficient of the individualized predictor either to zero, a positive effect γ_k^+, or a negative effect γ_k^-.

For correlated data structure, we can incorporate correlations of errors to obtain more efficient estimation ([16]), and introduce within-subject correlations through a weighting matrix $\boldsymbol{V}_i$ to the weighted squared-loss in the objective function

$$Q_{N,m}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{N} (\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta}))^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})) + \lambda_{N,m} \sum_{i=1}^{N} \sum_{k=1}^{p} s(\beta_{ik}, \gamma_k) \tag{7}$$

$$= L_{N,m}(\boldsymbol{\alpha}, \boldsymbol{\beta}) + S_{\lambda_{N,m}}(\boldsymbol{\beta}, \boldsymbol{\gamma}), \tag{8}$$

where $\boldsymbol{V}_i = \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}_i \boldsymbol{A}_i^{\frac{1}{2}}$, $\boldsymbol{A}_i$ is a diagonal matrix of marginal variance of $\boldsymbol{y}_i$ and $\boldsymbol{R}_i$ is a working correlation matrix.

3. Theoretical Results. In this section, we establish the theoretical properties of the proposed estimator, and the connection to the oracle estimator and the subject-wise least squares estimator. One unique aspect here is that our theory is established under a general double-divergence framework which assumes that both sample size N and cluster size m go to infinity, and therefore the number of individualized parameters also diverges.

We introduce some notation as follows. For any symmetric matrix $\boldsymbol{A}_{n\times n}$, let $\lambda_{min}(\boldsymbol{A})$ and $\lambda_{max}(\boldsymbol{A})$ be the smallest and the largest eigenvalues of $\boldsymbol{A}$, respectively. For an arbitrary matrix $\boldsymbol{A}_{m\times n}(b_{ij})$, denote $\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{max}(\boldsymbol{A}^T\boldsymbol{A})}$ as its L_2-norm, $\|\boldsymbol{A}\|_1 = \max\limits_{1\le j\le n}(\sum_{i=1}^{m}|b_{ij}|)$ as its L_1-norm and $\|\boldsymbol{A}\|_\infty = \max\limits_{1\le i\le m}(\sum_{j=1}^{n}|b_{ij}|)$ as its L_∞-norm. For a vector $\boldsymbol{a} = (a_1, \ldots, a_n)'$, $\|\boldsymbol{a}\|_2$ reduces to its Euclidean norm and $\|\boldsymbol{a}\|_\infty = \max\limits_{1\le i\le n}(|a_i|)$. Moreover, we denote $\|\boldsymbol{a}\|_0 = \sum_{i=1}^{n} I_{\{a_i\neq 0\}}$.

In addition, we define the order between two $n \times n$ square matrices as $\boldsymbol{A} > \boldsymbol{B}$ if $\forall \boldsymbol{x} \in \mathbf{R}^n$, $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}$ holds. Let $\boldsymbol{A} \asymp \boldsymbol{B}$ denote $c_1\boldsymbol{A} \le \boldsymbol{B} \le c_2\boldsymbol{A}$ for some constants $0 < c_1 \le c_2 < \infty$. Then we define a sequence of $m \times m$ matrices $\boldsymbol{A_n}$ as $\boldsymbol{A_n} = O(n)$ if $c_1 n\boldsymbol{I}_m \le \boldsymbol{A}_n \le c_2 n\boldsymbol{I}_m$ when n is large. Moreover, let $\boldsymbol{A} \circ \boldsymbol{B}$ denote the entrywise Hadamard product between two same-dimension matrices (see details in Appendix A.1), and "$\otimes$" denote the Kronecker product.

For unbalanced data, we define $\min(m_i) = m$ and assume $m_i = O(m)$ for $1 \leq i \leq N$. To simplify the notation, we let $m_i = m$ in the following discussion. In addition, without loss of generality, we consider the two-subpopulation structure with respect to each individualized predictor. The theory for a structure with more than two subpopulations can be shown similarly. Let $\mathcal{G}_k \subset \{i : 1 \leq i \leq N\}$ denote a signal-group index set for the kth individualized predictor such that $\beta_{ik} = \gamma_k \neq 0$ if $i \in \mathcal{G}_k$ and $\beta_{ik} = 0$ otherwise. For any set $\mathcal{G}$, let $|\mathcal{G}|$ be the cardinal of $\mathcal{G}$. Moreover, we denote $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ and let $\boldsymbol{\theta}^0 = ((\boldsymbol{\beta}^0)', (\boldsymbol{\alpha}^0)')'$ be its true value. Let the true value of $\boldsymbol{\beta}_i$ be $\boldsymbol{\beta}_{\boldsymbol{i}}^{\mathbf{0}} = (\boldsymbol{\beta}_{\boldsymbol{i},\boldsymbol{\mathcal{A}_i}}^{\mathbf{0}}{}', \boldsymbol{\beta}_{\boldsymbol{i},\boldsymbol{\mathcal{A}_i^c}}^{\mathbf{0}}{}')'$, where $\mathcal{A}_i$ and $\mathcal{A}_i^c$ denote the index sets such that $\boldsymbol{\beta}_{i,\mathcal{A}_i}^0 = \boldsymbol{\gamma}_{\mathcal{A}_i}^0 \neq \mathbf{0}$ and $\boldsymbol{\beta}_{i,\mathcal{A}_i^c}^0 = \mathbf{0}$.

The proposed objective function (7) consists of a loss function $L_{N,m}(\cdot)$ and a penalty function $S_{\lambda_{N,m}}(\cdot)$, where the squared loss function $L_{N,m}(\boldsymbol{\theta})$ in (8) can accommodate diverging N and m. Both the oracle estimator and the subject-wise least squares estimator are obtained by minimizing $L_{N,m}(\boldsymbol{\theta})$, but with different design matrices, where the corresponding quasi-likelihood estimating equation is

$$\boldsymbol{G}_{N,m}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{\theta}) = \sum_{i=1}^{N} \boldsymbol{U}_i(\boldsymbol{\theta})^T \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})) = 0, \tag{9}$$

with $\boldsymbol{U}_i(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$. Due to the linear mean function, $\boldsymbol{U}_i(\boldsymbol{\theta})$ does not depend on unknown parameters and thus is suppressed as $\boldsymbol{U}_i$ in the following, and we also denote $\boldsymbol{G}_{N,m} = \boldsymbol{G}_{N,m}(\boldsymbol{\theta}^0)$ for ease of notation. In addition, let

$$\boldsymbol{D}_{N,m} = -\frac{\partial \boldsymbol{G}_{N,m}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \sum_{i=1}^{N} \boldsymbol{U}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{U}_i,$$

$$\boldsymbol{H}_{N,m} = \mathrm{Cov}(\boldsymbol{G}_{N,m}(\boldsymbol{\theta})) = \sum_{i=1}^{N} \boldsymbol{U}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{\Sigma}_i \boldsymbol{V}_i^{-1} \boldsymbol{U}_i,$$

where $\boldsymbol{\Sigma}_i = \mathrm{Cov}(\boldsymbol{y}_i) = \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}_i^0 \boldsymbol{A}_i^{\frac{1}{2}}$ and $\boldsymbol{R}_i^0$ is the true correlation matrix. Note that $\boldsymbol{D}_{N,m}$ and $\boldsymbol{H}_{N,m}$ do not depend on unknown mean regression parameter $\boldsymbol{\theta}$. We require some common regularity conditions

(A1) The unknown parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ belongs to a compact subset $\mathcal{B} \subseteq \mathbf{R}^{p_{\boldsymbol{\theta}}}$ and its true value $\boldsymbol{\theta}^0$ lies in the interior of $\mathcal{B}$;

(A2) $\boldsymbol{D}_{N,m}$ and $\boldsymbol{H}_{N,m}$ are positive definite when N or m is large.

Note that the standard assumptions of $\boldsymbol{R}_i$ such as converging to a constant positive definite matrix with eigenvalues bounded away from zero and infinity ([32]) might not be valid in the proposed framework, since the dimension of $\boldsymbol{R}_i$ increases as m increases. Here we only require the following general regularity condition for $\boldsymbol{R}_i$ and $\boldsymbol{R}_i^0$:

(A3) There exist $\nu_l > 0, \nu_l' > 0$, such that $\lambda_{min}(\boldsymbol{R}_i^0) > \nu_l$ and $\lambda_{min}(\boldsymbol{R}_i) > \nu_l'$ for all i and m.

The estimating equation $\boldsymbol{G}_{N,m}(\boldsymbol{\theta})$ contains double summations with the sample size N and the cluster size m, which both can diverge. Consequently, the standard asymptotic results for M-estimators are not applicable here even with a fixed number of parameters ([33]). In general, for an estimator $\hat{\boldsymbol{\theta}}$ obtained by solving the estimating equation (9), under regularity conditions (A1)-(A2), by Taylor's expansion, $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = -\boldsymbol{D}_{N,m}^{-1}\boldsymbol{G}_{N,m}$. This implies that the consistency of $\hat{\boldsymbol{\theta}}$ relies on the following condition on the information matrix $\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m}$,

($\mathcal{C}_a$) $\quad \lambda_{min}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m}) \to \infty$.

In the independent model ($\boldsymbol{R}_i^0 = \boldsymbol{R}_i = \boldsymbol{I}_m$), $\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m}$ reduces to $\boldsymbol{D}_{N,m}$ as $\boldsymbol{H}_{N,m} = \boldsymbol{D}_{N,m}$.

The condition $\mathcal{C}_a$ is a standard condition analogous to [33] condition to establish the weak consistency of a fixed-dimensional GEE estimator. However, in contrast to [33]'s setting, the proposed method results in a diverging dimension of the information matrix which is more complicated. In addition, to utilize subpopulation information, the convergence rates for estimators of different parameters are of great importance and interest in this paper. The following lemma provides a convergence property for the estimating equation estimator from (9).

LEMMA 1. *Under regularity condition (A2), for any $\delta > 0$, there exists a solution $\hat{\boldsymbol{\theta}}$ of (9) such that*

$$P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{D}_{N,m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\|_2 > \delta\right) < \tfrac{1}{\delta^2},$$

where $p_{\boldsymbol{\theta}}$ is the dimension of $\boldsymbol{\theta}$. Moreover, if condition ($\mathcal{C}_a$) holds, we have

$$P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\|_2 > \delta\right) \to 0.$$

Lemma 1 presents the consistency result under all settings. It indicates that the estimator's convergence rate depends on the divergence rate of $\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m}$'s eigenvalues.

REMARK 1. Note that Lemma 1 provides consistency under the spectral norm (L_2-norm). For any fixed-dimensional estimator, for example, the oracle estimator and the subject-wise estimator when N is bounded, the consistency in Lemma 1 is equivalent to $P\left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\|_\infty > \delta\right) \to 0$. However, if $p_{\boldsymbol{\theta}}$ is diverging, we need additional conditions to ensure the stronger consistency under the L_∞-norm. More discussion will be provided later regarding the proposed estimator when $N \to \infty$.

In addition, we assume that a few general regularity conditions hold for the design matrix,

(A4) $\tilde{\boldsymbol{X}}_{ij} = (\boldsymbol{X}_{ij}', \boldsymbol{Z}_{ij}')'_{(p+q)\times 1}$ belongs to a compact set $\mathcal{X} \subset \mathbf{R}^{p+q}$ for $1 \le i \le N$ and $1 \le j \le m$;

(A5) Let $\tilde{\boldsymbol{X}}_{i,k}$ denote the kth column of $\tilde{\boldsymbol{X}}_i$, assume $\|\tilde{\boldsymbol{X}}_{i\cdot,k}\|_2^2 = O_p(m)$ and $\sum_{i=1}^N m^{-1}\|\tilde{\boldsymbol{X}}_{i\cdot,k}\|_2^2 = O_p(N)$, for $1 \le k \le p+q$;

(A6) $m^{-1}\lambda_{min}(\boldsymbol{X}_i^T\boldsymbol{X}_i) > c_3$ for any i and $\frac{1}{Nm}\lambda_{min}\left(\sum_{i=1}^N \boldsymbol{Z}_i^T(\boldsymbol{I}_m - \boldsymbol{H}_{\boldsymbol{X}_i})\boldsymbol{Z}_i\right) > c_4$, where $\boldsymbol{H}_{\boldsymbol{X}_i} = \boldsymbol{X}_i(\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^T$, for some constants $0 < c_3 < \infty$, $0 < c_4 < \infty$.

Conditions (A4)-(A6) are regularity conditions which are typically required for the bounded regressors. However, these are less restrictive than other assumptions, e.g., $\frac{1}{m}\boldsymbol{X}_i^T\boldsymbol{X}_i$ converges to a positive constant matrix. Note that condition (A6) allows within-subject invariant covariates, and is less restrictive since it does not require $\tilde{\boldsymbol{X}}_i^T\tilde{\boldsymbol{X}}_i$ to be positive definite.

The regularity conditions (A1)- (A6) are assumed to hold in this section. In Condition (A2) and Lemma 1, matrices $\boldsymbol{D}_{N,m}$ and $\boldsymbol{H}_{N,m}$ represent a general form according to the estimating equation (9). For different estimators using the same data, for example, the oracle estimator or the subject-wise estimator, $\boldsymbol{D}_{N,m}$ and $\boldsymbol{H}_{N,m}$ can be different due to their different formulating.

3.1. *Asymptotic results for the oracle estimator with group effects.* In the proposed framework, the oracle estimator assumes that all subpopulation information ($\mathcal{G}_k$, $1 \le k \le p$) with respect to the individualized predictors is known. This is equivalent to assuming that the true signal sets $\mathcal{A}_i$'s ($1 \le i \le N$) for all subjects are known.

The individualized parameter $\boldsymbol{\beta_i}$ for each subject is linked to the sub-homogeneous parameter $\boldsymbol{\gamma}$ as $\boldsymbol{\omega_i} \circ \boldsymbol{\gamma} = \boldsymbol{\beta_i}$ through an indicator vector $\boldsymbol{\omega_i} = (\omega_{i1}, \ldots, \omega_{ip})' \in \mathbf{R}^p$, where $\omega_{ik} = I_{\{i\in\mathcal{G}_k\}} = I_{\{k\in\mathcal{A}_i\}}$. Hence there exists a mapping linking two parameter spaces, which is $\mathbf{R}^p \mapsto \mathbf{R}^{Np} : \boldsymbol{\Omega\gamma} = \boldsymbol{\beta}$, where $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_N)'$ is a $Np \times p$ matrix and $\boldsymbol{\Omega}_i = \text{diag}(\boldsymbol{\omega}_i)$ is a diagonal matrix. We define $L_{N,m}^{or}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = L_{Nm}(\boldsymbol{\alpha}, \boldsymbol{\beta}(\boldsymbol{\gamma}))$. By noting that $S_{\lambda_{N,m}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = 0$ with $\boldsymbol{\beta} = \boldsymbol{\Omega\gamma}$ and $\boldsymbol{\Omega}$ is known, the oracle estimator can be obtained by minimizing $L_{N,m}^{or}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ as

$$\left((\hat{\boldsymbol{\gamma}}^{or})', \hat{\boldsymbol{\alpha}}^{or})'\right)' = \underset{\boldsymbol{\alpha},\boldsymbol{\gamma}}{\text{argmin}} \sum_{i=1}^N \left(\boldsymbol{y}_i - \boldsymbol{X_i}(\boldsymbol{\omega}_i \circ \boldsymbol{\gamma}) - \boldsymbol{Z_i}\boldsymbol{\alpha}\right)^T \boldsymbol{V}_i^{-1} \left(\boldsymbol{y}_i - \boldsymbol{X_i}(\boldsymbol{\omega}_i \circ \boldsymbol{\gamma}) - \boldsymbol{Z_i}\boldsymbol{\alpha}\right).$$

The oracle individualized estimator for each subject is obtained by $\hat{\boldsymbol{\beta}}_i^{or} = \boldsymbol{\omega}_i \circ \hat{\boldsymbol{\gamma}}^{or}$.

Let $\tilde{\boldsymbol{X}}_i = (\boldsymbol{X}_i, \boldsymbol{Z}_i)$ and $\tilde{\boldsymbol{\omega}}_i = (\boldsymbol{\omega}_i', \mathbf{1}_q')'$, and $\tilde{\boldsymbol{X}}_i^{or} = \tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{\Omega}}_i$ where $\tilde{\boldsymbol{\Omega}}_i = \text{diag}(\tilde{\boldsymbol{\omega}}_i)$. We denote $\boldsymbol{H}_{N,m}^{or} = \sum_{i=1}^N (\tilde{\boldsymbol{X}}_i^{or})^T \boldsymbol{V}_i^{-1}\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1}\tilde{\boldsymbol{X}}_i^{or}$, $\boldsymbol{D}_{N,m}^{or} = \sum_{i=1}^N (\tilde{\boldsymbol{X}}_i^{or})^T \boldsymbol{V}_i^{-1}\tilde{\boldsymbol{X}}_i^{or}$, and Lemma 1 directly applies for the oracle estimator by replacing $\boldsymbol{H}_{N,m}$ and $\boldsymbol{D}_{N,m}$ with $\boldsymbol{H}_{N,m}^{or}$ and $\boldsymbol{D}_{N,m}^{or}$, respectively. Let $\hat{\boldsymbol{\theta}}^{or} = \left((\hat{\boldsymbol{\gamma}}^{or})', (\hat{\boldsymbol{\alpha}}^{or})'\right)'$ and $\tilde{\boldsymbol{\theta}}^0 = \left((\boldsymbol{\gamma}^0)', (\boldsymbol{\alpha}^0)'\right)'$, according to Lemma 1 we have

$$(\boldsymbol{H}_{N,m}^{or})^{-\frac{1}{2}}(\boldsymbol{D}_{N,m}^{or})(\hat{\boldsymbol{\theta}}^{or} - \tilde{\boldsymbol{\theta}}^0) = O_p(1). \tag{10}$$

Note that the divergence rates of $\boldsymbol{H}_{N,m}^{or}$ and $\boldsymbol{D}_{N,m}^{or}$ are associated with the subpopulation size $|\mathcal{G}_k|$'s as N goes to infinity. However, in contrast to other clustering approaches based on an entire set of coefficient vector $\boldsymbol{\beta}_i$ (e.g., [21]; [18]), the proposed model allows the subgroup partitions corresponding to different individualized predictors to be different. Therefore the design matrix for the oracle estimator here cannot be formulated as a block diagonal form, which leads to non-trivial subgroup effects on divergence rates.

REMARK 2. A few comments about the eigenvalues of the matrices are worth mentioning. For two square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with the same dimension, $\boldsymbol{AB}$ and $\boldsymbol{BA}$ have the same non-zero eigenvalues. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are non-singular and $\boldsymbol{A} \leq \boldsymbol{B}$, for any matrix $\boldsymbol{C}$ we have $\boldsymbol{C}^T\boldsymbol{AC} \leq \boldsymbol{C}^T\boldsymbol{BC}$, and $\boldsymbol{A}^{-1} \geq \boldsymbol{B}^{-1}$. The proofs of these results are provided in Appendix (A.1).

To get a better understanding of the group effects on the oracle estimator, we reformulate $\boldsymbol{D}^{or}_{N,m} = \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i^T \tilde{\boldsymbol{X}}_i^T \boldsymbol{V}_i^{-1} \tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{\Omega}}_i = \sum_{i=1}^N (\tilde{\boldsymbol{\Omega}}_i \tilde{\boldsymbol{\Omega}}_i^T) \circ (\boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i)$, where $\tilde{\boldsymbol{\Omega}}_i^T \tilde{\boldsymbol{\Omega}}_i^T$ is a symmetric square matrix with entries to be zero or one. Suppose

(R1). $$\kappa_m^l \leq \lambda_{min}(\tilde{\boldsymbol{X}}_i^T \boldsymbol{V}_i^{-1} \tilde{\boldsymbol{X}}_i) \leq \lambda_{max}(\tilde{\boldsymbol{X}}_i^T \boldsymbol{V}_i^{-1} \tilde{\boldsymbol{X}}_i) \leq \kappa_m^u$$

holds uniformly for any subject i with some positive constant sequences $\{\kappa_m^l\}_{m=1}^\infty$ and $\{\kappa_m^u\}_{m=1}^\infty$, then we have $\kappa_m^l \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i \leq \boldsymbol{D}^{or}_{N,m} \leq \kappa_m^u \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i$ by noting $\tilde{\boldsymbol{\Omega}}_i^2 = \tilde{\boldsymbol{\Omega}}_i$. Under a similar condition to (R1), we could show that $\phi_m^l \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i \leq \boldsymbol{H}^{or}_{N,m} \leq \phi_m^u \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i$ for some positive constant sequences $\{\kappa_m^l\}_{m=1}^\infty$ and $\{\kappa_m^u\}_{m=1}^\infty$. If $\sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i$ is non-singular, then

$$(\phi_m^u)^{-1}(\kappa_m^l)^2 \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i \leq \boldsymbol{D}^{or}_{N,m} (\boldsymbol{H}^{or}_{N,m})^{-1} \boldsymbol{D}^{or}_{N,m} \leq (\phi_m^l)^{-1}(\kappa_m^u)^2 \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i. \tag{11}$$

Let $\boldsymbol{\Lambda}_{N,m} = \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i$ and note that $\boldsymbol{\Lambda}_{N,m} = \text{diag}(N\boldsymbol{1}_q', |\mathcal{G}_1|, \ldots, |\mathcal{G}_p|)$ is a diagonal matrix, where $|\mathcal{G}_k|$'s ($1 \leq k \leq p$) are signal-subgroup sizes corresponding to p individualized predictors. It is clear that $\boldsymbol{\Lambda}_{N,m}$ contains the group effects on estimation. In particular, the group size for the population-shared parameter is N.

REMARK 3. The condition (R1) could be relaxed by replacing $\tilde{\boldsymbol{X}}i$ with $\boldsymbol{X}_i$ since we allow within-subject invariant covariates, especially for the population-shared predictors. Moreover, if m is bounded, it is straightforward to show that $c_l m \leq \kappa_m^l \leq \kappa_m^u \leq c_u m$ and $c_l' m \leq \phi_m^l \leq \phi_m^u \leq c_u' m$ hold for some constants $0 < c_l \leq c_m < \infty$, $0 < c_l' \leq c_m' < \infty$, which immediately implies that $\boldsymbol{D}^{or}_{N,m} \asymp m\boldsymbol{\Lambda_{N,m}}$ and $\boldsymbol{H}^{or}_{N,m} \asymp m\boldsymbol{\Lambda_{N,m}}$. This conclusion also holds for the independent model even when m goes to infinity.

Let $N_k = \sum_{i \in \mathcal{G}_k} m_i = m|\mathcal{G}_k|$ denote the number of observations in group $\mathcal{G}_k$ and $N_a = \sum_{i=1}^N m_i = mN$ denote the total number of observations. For the independent error model, we establish asymptotic normality for the oracle estimators with convergence rates associated to the sample size N and the cluster size m.

THEOREM 1. *Under regularity conditions, suppose $\boldsymbol{R}_i^0 = \boldsymbol{R}_i = \boldsymbol{I}_m$ holds for any i, as either $m \to \infty$ or $\min_{1 \leq k \leq p} (|\mathcal{G}_k|) \to \infty$, we have*

$$(\boldsymbol{H}^{or}_{N,m})^{\frac{1}{2}} \Big(\{(\hat{\boldsymbol{\gamma}}^{or})', (\hat{\boldsymbol{\alpha}}^{or})'\}' - \{(\boldsymbol{\gamma}^0)', (\boldsymbol{\alpha}^0)'\}' \Big) \to_d N\Big(\boldsymbol{0}, \boldsymbol{I}_{p+q}\Big),$$

where $\boldsymbol{H}_{N,m}^{or} \asymp \boldsymbol{M}_{N,m}$, *and* $\boldsymbol{M}_{N,m} = diag(\underbrace{N_1, \ldots, N_p}_{p}, \underbrace{N_a, \ldots, N_a}_{q})$ *is a* $(p+q) \times (p+q)$*-dim diagonal matrix.*

Theorem 1 indicates that the convergence rates of the oracle estimator benefit from both increasing N and m, which implies that incorporating subgroup information is able to improve estimation efficiency as we utilize additional number of observations from each subgroup. In addition, Theorem 1 allows both m and N to go to infinity and has no restriction on their divergence rates.

However, in the correlated model with cluster size m diverging, the analysis of the estimator's asymptotic behavior becomes more complicated , since it involves the working correlation matrix $\boldsymbol{R}_i$ and the unknown true correlation matrix $\boldsymbol{R}_i^0$, which makes it difficult to verify the condition $(\mathcal{C}_a)$ and to figure out the estimators' convergence rates.

Similar to [33], we consider a sufficient condition which may simplify the verification and the discussion. Let $\eta_{N,m} = \max_{1 \le i \le N} \{\lambda_{max}(\boldsymbol{R}_i^{-1}\boldsymbol{R}_i^0)\}$, an alternative condition for consistency is

$$(\mathcal{C}_a^*) \qquad \eta_{N,m}^{-1}\lambda_{min}(\boldsymbol{D}_{N,m}) \to \infty.$$

The sufficiency of $(\mathcal{C}_a^*)$ that implies $(\mathcal{C}_a)$ is trivial by noting $\boldsymbol{H}_{N,m} \le \eta_{N,m}\boldsymbol{D}_{N,m}$. Based on (10), we present the asymptotic theory for the oracle estimator with the the condition $\mathcal{C}_a^*$.

THEOREM 2. *Under regularity conditions, for the oracle estimator* $\hat{\boldsymbol{\theta}}^{or} = ((\hat{\boldsymbol{\gamma}}^{or})', (\hat{\boldsymbol{\alpha}}^{or})')'$, *we have*

$$\eta_{N,m}^{-\frac{1}{2}} \|(\boldsymbol{D}_{N,m}^{or})^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{or} - \tilde{\boldsymbol{\theta}}^0)\|_2 \le O_p(1),$$

and if $\eta_{N,m}^{-1}\lambda_{min}(\boldsymbol{D}_{N,m}^{or}) \to \infty$, $\hat{\boldsymbol{\theta}}^{or} \to_p \tilde{\boldsymbol{\theta}}^0$.

The proof of Theorem 2 is straightforward by following (11) and condition $\mathcal{C}_a^*$. Theorem 2 indicates that the convergence of the estimator depends on the divergence rate of $\eta_{N,m}$ and $\boldsymbol{D}_{N,m}^{or}$. Without considering the group effects, the oracle estimator reduces to a fixed-dimensional GEE estimator by [33] and [1]. Therefore, in the following, we only focus on a few common cases and some useful conditions.

REMARK 4. For any N and m, according to regularity condition (A3), note that

$$\eta_{N,m} \le (\min_{1 \le i \le N} \{\lambda_{min}(\boldsymbol{R}_i)\})^{-1} \max_{1 \le i \le N} \{\lambda_{max}(\boldsymbol{R}_i^0)\} \le (\nu_l')^{-1}\text{tr}(\boldsymbol{R}_1^0) \le (\nu_l')^{-1}m.$$

If m is bounded, then $\eta_{N,m}$ is bounded, which implies that the condition $\mathcal{C}_a^*$ does not depend on unknown true correlation structure $\boldsymbol{R}_i^0$. As $N \to \infty$, we have $\lambda_{min}(\boldsymbol{D}_{N,m}^{or}) \to \infty$ regardless of the choice of working correlation $\boldsymbol{R}_i$. Hence, similar to standard results for the GEE estimator, the oracle estimator $\hat{\boldsymbol{\theta}}^{or}$ has asymptotic normality, although it may not achieve optimal efficiency if $\boldsymbol{R}_i \ne \boldsymbol{R}_i^0$.

REMARK 5. If $m \to \infty$, $\eta_{N,m}$ is not always bounded. For example, if $\boldsymbol{R}_i^0$ admits an exchangeable correlation structure and we choose working correlation $\boldsymbol{R}_i$ as an identity matrix, we

have $\eta_{N,m} = O(m)$. For any bounded N, $\boldsymbol{D}^{or}_{N,m} = O(m)$, which implies that the condition $(\mathcal{C}^*_a)$ fails. Although the condition $(\mathcal{C}_a)$ may still hold with some constraints on the design matrix to ensure consistency (see following Example 1), the convergence rate could be slower than the optimal rate $\sqrt{m}$ and it may not converge to a normal distribution asymptotically ([33]).

We use the following example of a simple linear regression to illustrate some details about the conditions $\mathcal{C}_a$ and $\mathcal{C}^*_a$ with specific covariates design.

EXAMPLE 1. Consider a subject-wise model with homogeneous effect,

$$y_{ij} = x_{ij}\beta + \varepsilon_{ij}, \quad i = 1, \ldots, N; j = 1, \ldots, m,$$

where $\boldsymbol{\varepsilon_i} = (\varepsilon_{i1}, \ldots, \varepsilon_{im})' \sim N(\mathbf{0}, \sigma^2 \boldsymbol{R^0})$ and $\boldsymbol{R^0}$ admits an exchangeable structure with parameter $\rho > 0$, x_{ij}'s are iid $N(\mu, 1)$. For the case of bounded N, without loss of generality, we assume $N = 1$. By using an independent working correlation $\boldsymbol{R}_i = \boldsymbol{I}_m$, we have $\boldsymbol{D}_m = \boldsymbol{x}_1^T \boldsymbol{x}_1 = O(m)$ and $\eta_m = \lambda_{max}(\boldsymbol{R}^0) = m\rho + 1 - \rho$, where $\boldsymbol{x}_1 = (x_{11}, \ldots, x_{1m})'$. Thus condition $\mathcal{C}^*_a$ fails. However, note that $\boldsymbol{R}^0(\rho) = (1-\rho)\boldsymbol{I}_m + \rho \mathbf{1}_m \mathbf{1}_m^T$. We have $\boldsymbol{H}_m = \sigma^2 \boldsymbol{x}_1^T \boldsymbol{R}^0 \boldsymbol{x}_1 = \sigma^2 \boldsymbol{x}_1^T((1-\rho)\boldsymbol{I}_m + \rho \mathbf{1}_m \mathbf{1}_m^T)\boldsymbol{x}_1^T = \sigma^2(1-\rho)\boldsymbol{x}_1^T \boldsymbol{x}_1 + m\rho(m^{-\frac{1}{2}} \sum_{i=1}^m x_{1j})^2 = O(m) + O(m)$ if $\mu = 0$, and thus $\lambda_{min}(\boldsymbol{D}_m \boldsymbol{H}_m^{-1} \boldsymbol{D}_m) = O(m) \to \infty$ as $m \to \infty$. But if $\mu > 0$, it is clear that $m\rho(m^{-\frac{1}{2}} \sum_{i=1}^m x_{1j})^2 = O(m^2)$ and thus $\lambda_{min}(\boldsymbol{D}_m \boldsymbol{H}_m^{-1} \boldsymbol{D}_m) = O(1)$.

COROLLARY 1. *Suppose $\eta_{N,m} \le C_1$ holds uniformly for some constant $0 < C_1 < \infty$, under regularity conditions, we have*

$$\|\boldsymbol{M}^{\frac{1}{2}}_{N,m}(\hat{\boldsymbol{\theta}}^{or} - \tilde{\boldsymbol{\theta}}^0)\|_2 \le O_p(1),$$

where $\boldsymbol{M}_{N,m}$ is defined in Theorem 1.

The condition of uniformly bounded $\eta_{N,m}$ in Corollary 1 naturally holds when m is bounded or for the independent model. However, as m goes to infinity, it implies that either we choose a working correlation matrix $\boldsymbol{R}_i$ close to the true one, or the correlation is not too strong. The first case involves a consistent and efficient estimator of the correlation structure, which has been discussed in [2], [13] and [10]. For the second case, a variety of conditions can be imposed on the correlation structures to ensure a "weak" dependency.

In the following, we provide a sufficient condition which can be verified easily in practice. For an arbitrary correlation matrix $\boldsymbol{R}_{m\times m}(\rho_{ij})$, assume

$(\mathcal{R}_a)$ $\quad |\rho_{ij}| \le \rho_{|i-j|}$ for $i \ne j$ and $\sum_{k=1}^{\infty} \rho_k < \infty$.

We show in the Appendix that if condition $(\mathcal{R}_a)$ holds for the true correlation matrix $\boldsymbol{R}^0_i$, then $\eta_{N,m}$ is bounded uniformly for any working correlation structures. This indicates that $\boldsymbol{R}^0_i$ is bounded as the within-subject correlation decays rapidly as m increases. In practice, a wide family of correlation structures satisfy the conditions $(\mathcal{R}_a)$ including the AR-1 and the M-dependent correlation matrices.

3.2. *Asymptotic results for the proposed estimator.* In general, the least squares estimator plays an important intermediate role in investigating the large sample theory of the penalized estimator. Hence, prior to presenting the theoretical results for the proposed estimator, we provide the asymptotic theory for the subject-wise least squares estimator $\hat{\boldsymbol{\theta}}^{Sub} = ((\hat{\boldsymbol{\beta}}^{Sub})', (\hat{\boldsymbol{\alpha}}^{Sub})')'$ obtained by minimizing $L_{N,m}(\boldsymbol{\theta})$.

Note that, for the proposed estimator and the subject-wise least squares estimator, each term of $\boldsymbol{U}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{U}_i$ in $\boldsymbol{D}_{N,m}$ does not equal to $\boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i$, but is a block sparse, matrix as $\boldsymbol{\mu}_i$ does not contain any other individualized parameter $\boldsymbol{\beta}_j$ for $j \neq i$. We denote

$$\boldsymbol{D}^s_{N,m} = \begin{pmatrix} \boldsymbol{D}^s_{xx}(Np \times Np) & \boldsymbol{D}^s_{xz}(Np \times q) \\ \boldsymbol{D}^s_{zx}(q \times Np) & \boldsymbol{D}^s_{zz}(q \times q) \end{pmatrix},$$

for the subject-wise estimator, where $\boldsymbol{D}^s_{xx} = \text{bdiag}\Big(\{\boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\}_{i=1}^N\Big)$ and bdiag$(\cdot)$ denotes a block-diagonal matrix. Similarly, we have $\boldsymbol{H}^s_{xx} = \text{bdiag}\Big(\{\boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{\Sigma}_i \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\}_{i=1}^N\Big)$ in $\boldsymbol{H}^s_{N,m}$ (see Appendix for details), and both $\boldsymbol{D}^s_{xx}$ and $\boldsymbol{H}^s_{xx}$ will expand as N increases. Following Lemma 1, we obtain the following result:

LEMMA 2. *Under regularity conditions, for any $\delta > 0$ and $\boldsymbol{a} \in \mathbf{R}^{Np+q}$, we have*

$$P\Big(|\boldsymbol{a}^T(\hat{\boldsymbol{\theta}}^{Sub} - \boldsymbol{\theta}^0)|^2 > \delta\Big) \leq \delta^{-2} \boldsymbol{a}^T (\boldsymbol{D}^s_{N,m} (\boldsymbol{H}^s_{N,m})^{-1} \boldsymbol{D}^s_{N,m})^{-1} \boldsymbol{a}.$$

If we choose $\boldsymbol{a}$ as a coordinate indicator for $\boldsymbol{\beta}_i$ in $\boldsymbol{\theta}$, that is, $\boldsymbol{a} = (\mathbf{0}_q', \boldsymbol{a}_1', \ldots, \boldsymbol{a}_N')'$, where $\boldsymbol{a}_j \in \mathbf{R}^p$, $1 \leq j \leq N$, $\boldsymbol{a}_j = \mathbf{1}_p$ if $j = i$ or $\boldsymbol{a}_j = \mathbf{0}_p$ if $j \neq i$, Lemma 2 implies the following corollary, which provides a detailed view of the convergence property for each subject-wise estimator $\hat{\boldsymbol{\beta}}_i^{Sub}$ and the population-shared estimator $\hat{\boldsymbol{\alpha}}^{Sub}$.

COROLLARY 2. *Under regularity conditions, for any $\delta > 0$ and individualized estimator $\hat{\boldsymbol{\beta}}_i^{Sub}$,*

$$P\Big(\|\hat{\boldsymbol{\beta}}_i^{Sub} - \boldsymbol{\beta}_i^0\|_2 > \delta\Big) \leq p\delta^{-2} \eta_{Nm} \lambda_{min}(\boldsymbol{D}^s_{\boldsymbol{X}_i})^{-1},$$

where $\boldsymbol{D}^s_{\boldsymbol{X}_i} = \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i$, $i = 1, \ldots, N$, and for the population-shared estimator $\hat{\boldsymbol{\alpha}}^{Sub}$,

$$P\Big(\|(\hat{\boldsymbol{\alpha}}^{Sub} - \boldsymbol{\alpha}^0)\|_2 > \delta\Big) \leq q\delta^{-2} \eta_{Nm} \lambda_{min}(\boldsymbol{D}^s_{\boldsymbol{Z}})^{-1},$$

where $\boldsymbol{D}^s_{\boldsymbol{Z}} = \sum_{i=1}^N \boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{Z}_i$.

Note that the condition $((C)_a)$ requires that $m \to \infty$. In the case of bounded m and diverging N, it is straightforward that the consistency of any individualized parameter cannot be achieved

since $\lambda_{min}(\boldsymbol{D}^s_{\boldsymbol{X}_i})$ does not diverge. Intuitively, the increasing number of subjects does not accumulate additional information for the subject-wise parameters. However, the estimator of population-shared parameter $\hat{\boldsymbol{\alpha}}$ could still be consistent as $N \to \infty$ by noting that η_{Nm} is bounded and $\lambda_{min}(\boldsymbol{D}^s_{\boldsymbol{Z}}) \to \infty$.

Lemma 2 and Corollary 2 provide consistent estimations under the L_2-norm, which depend on the dimension of parameters. Furthermore, we pursue a stronger uniform consistency with additional conditions on either the random errors' distributions or the divergence rates of N and m. In addition to the basic assumptions of zero mean and finite second moment σ^2 for random error ε_{ij}'s, let $\boldsymbol{\varepsilon}^*_{\boldsymbol{i}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\varepsilon_i}$ and denote $\tau^s_{N,m} = \lambda_{min}(\boldsymbol{D}^s_{N,m}(\boldsymbol{H}^s_{N,m})^{-1}\boldsymbol{D}^s_{N,m})$

($\mathcal{I}_a$) $\quad N = o(\tau^s_{N,m})$,

($\mathcal{I}_b$) (i) $\boldsymbol{\varepsilon}^*_{\boldsymbol{i}}$ is a sub-Gaussian vector, that is, $\mathrm{P}(|\boldsymbol{a}^T\boldsymbol{\varepsilon}^*_{\boldsymbol{i}}| > t) < 2\exp(-\frac{t^2}{c^2_\sigma \|\boldsymbol{a}\|^2_2})$ for any $\boldsymbol{a} \in \mathbf{R}^m$ and $t > 0$, where c_σ is a positive constant; and (ii) $\log(N) = o(\tau^s_{N,m})$.

In the independent model where $\boldsymbol{\Sigma} = \boldsymbol{I}_m$, the condition (i) in $\mathcal{I}_b$ is equivalent to assuming marginal sub-Gaussian tails for ε_{ij}'s, which is a standard assumption in high-dimensional model. Alternatively, if the random errors are assumed to be normally distributed, then the condition (i) in $\mathcal{I}_b$ holds naturally for both independent and correlated models.

Under condition ($\mathcal{I}_a$) or ($\mathcal{I}_b$), we achieve a stronger uniform consistency for the diverging number of parameters when $N \to \infty$ as $m \to \infty$.

LEMMA 3. *Under regularity conditions, if either condition ($\mathcal{I}_a$) or ($\mathcal{I}_b$) is satisfied, for any $\delta > 0$, as $m \to \infty$, we have*

$$P\left(\|\hat{\boldsymbol{\theta}}^{Sub} - \boldsymbol{\theta}^0\|_\infty > \delta\right) \to 0.$$

Theorem 3 indicates that if N diverges at a limited rate compared to m, or the tails of the random errors' distribution decay fast enough, we are able to achieve a stronger consistency under the L_∞ norm. Note that the $\tau^s_{N,m}$ in conditions ($\mathcal{I}_a$) and ($\mathcal{I}_b$) could also be replaced with $\eta^{-1}_{N,m}\lambda_{min}(\boldsymbol{D}^s_{N,m})$ analogous to the above discussion, which leads to a sufficient condition.

Based on the above conditions and results, we establish the large sample theory for the proposed estimator. We first provide insight into the proposed multi-directional separation penalty. Consider a simple independent linear regression model for one subject with the objective function

$$Q_{i,m}(\boldsymbol{\beta_i}|\hat{\boldsymbol{\gamma}}) = \frac{1}{2} \parallel \boldsymbol{y}_i - \boldsymbol{X_i}\boldsymbol{\beta_i} - \boldsymbol{Z_i}\boldsymbol{\alpha} \parallel^2_2 + \lambda_m \sum_{k=1}^{p} s(\beta_{ik}, \hat{\gamma}_k), \tag{12}$$

given an estimator of the sub-homogeneous effects $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \cdots, \hat{\gamma}_p)'$. Therefore, the proposed penalty function $s(\cdot, \hat{\gamma}^k)$ provides an alternative shrinking direction besides zero. The following theorem presents the asymptotic property for the individualized estimator obtained by minimizing (12).

THEOREM 3. *Under regularity conditions, there exists a local minimizer* $\hat{\boldsymbol{\beta}}_i = (\hat{\boldsymbol{\beta}}'_{i,\mathcal{A}_i}, \hat{\boldsymbol{\beta}}'_{i,\mathcal{A}_i^c})'$ *of (12), if* $\lambda_m \to 0$, *as* $m \to 0$, *we have* $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} \to_p \boldsymbol{\beta}_{\boldsymbol{i}}^{\mathbf{0}}$. *In addition, if* $\lambda_m/\sqrt{m} \to \infty$, *suppose* $\sqrt{m}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0) = O_p(1)$, *then we have*

$$P(\hat{\boldsymbol{\beta}}_{i,\mathcal{A}_i^c} = 0) \to 1 \qquad \textit{and} \qquad P(\hat{\boldsymbol{\beta}}_{i,\mathcal{A}_i} = \hat{\boldsymbol{\gamma}}_{\mathcal{A}_i}) \to 1.$$

It is worth noting that the condition of consistency on $\hat{\boldsymbol{\gamma}}$ can be relaxed. The proof of Theorem 3 shows that both estimation consistency and selection consistency still hold even if $\hat{\boldsymbol{\gamma}}$ is not consistent. However, if $a_m(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0) = O_p(1)$ and $a_m/\sqrt{m} \to \infty$ hold for some a_m, then the estimator $\hat{\boldsymbol{\beta}}_{i\mathcal{A}_i}$ can achieve a faster convergence rate than $\sqrt{m}$, which is optimal for any subject-wise model. In the proposed model, $\hat{\boldsymbol{\gamma}}$ is estimated over different subjects via the subgrouping and gains efficiency from increasing number of subjects N.

In another perspective, we investigate group separation as both N and m go to infinity. Denote $\boldsymbol{B}_{\boldsymbol{\beta}_i^0}(r)$ as a ball in $\mathbf{R}^p$ centered at $\boldsymbol{\beta}_i^0$ with a radius $r > 0$.

LEMMA 4. *Suppose either condition* $(\mathcal{I}_a)$ *or* $(\mathcal{I}_b)$ *holds. Under regularity conditions, for any constant* $r > 0$, *as* $\tau_{N,m}^s \to \infty$, *there exists a local minimizer* $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ *of* $Q_{N,m}$ *in (7) such that*

$$P\Bigg(\bigcap_{1 \le i \le N} \{\hat{\boldsymbol{\beta}}_i \in \boldsymbol{B}_{\boldsymbol{\beta}_i^0}(r)\} \bigcap \{\hat{\boldsymbol{\alpha}} \in \boldsymbol{B}_{\boldsymbol{\alpha}^0}(r)\} \bigcap \{\hat{\boldsymbol{\gamma}} \in \boldsymbol{B}_{\boldsymbol{\gamma}^0}(r)\} \Bigg) \to 1.$$

As both sample size N and cluster size m increase, if N diverges at a limited rate, the speed of separation over subjects dominates the speed of increasing subjects. Lemma 4 essentially implies group identification consistency and thus we obtain more information about the correct direction of the true individualized parameters.

In the spirit of Theorem 3 and Lemma 4, we present the oracle property for the proposed estimator under a general double-divergence setting.

THEOREM 4. *Under regularity conditions, suppose either condition* $(\mathcal{I}_a)$ *or* $(\mathcal{I}_b)$ *holds, assuming* $\frac{\lambda_{N,m}}{\tau_{N,m}^s} \to 0$ *and* $\frac{\lambda_{N,m}}{\sqrt{\tau_{N,m}^s}} \to \infty$, *then there exists a local minimizer* $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ *of* $Q_{N,m}$ *in (7); as* $\tau_{N,m}^s \to \infty$, *we have*

$$P\Big(\{\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T\}^T = \{(\hat{\boldsymbol{\alpha}}^{or})^T, (\hat{\boldsymbol{\beta}}^{or})^T, (\hat{\boldsymbol{\gamma}}^{or})^T\}^T \Big) \to 1.$$

COROLLARY 3 (Uniform selection consistency). *Under the same conditions as in Theorem 4, as* $\tau_{N,m}^s \to \infty$, *we have* $P\Big(\bigcap_{i=1}^N \{\hat{\mathcal{A}}_i = \mathcal{A}_i\} \Big) \to 1$.

Theorem 4 indicates that the proposed estimator is the same as the oracle estimator, which utilizes the most information. In fact, by providing additional shrinking directions, the proposed model enables us to separate the strong signals from the weak ones. Consequently, we achieve the

oracle information about the underlying subpopulation structure, which ensures that the proposed estimator inherits the optimal efficiency from the oracle estimator. From the other perspective, Corollary 3 also implies subgroup identification consistency.

In addition, in the independent error model, by noting $\tau^s_{N,m} = m$, the conditions $(\mathcal{I}_a)$ and $(\mathcal{I}_b)$ can be simplified as follows:

$(\mathcal{I}^*_a)$ $N = o(m)$,

$(\mathcal{I}^*_b)$ ε_{ij} has sub-Gaussian tails and $\log(N) = o(m)$.

Hence, we have a simplified result for the independent model.

COROLLARY 4. *Under regularity conditions, if $\boldsymbol{R}_i = \boldsymbol{R}^0_i = \boldsymbol{I}_m$, suppose either condition $(\mathcal{I}^*_a)$ or $(\mathcal{I}^*_b)$ holds, assuming $\frac{\lambda_{N,m}}{m} \to 0$ and $\frac{\lambda_{N,m}}{\sqrt{m}} \to \infty$, then there exists a local minimizer $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ of $Q_{N,m}$ in (7); as $m \to \infty$, we have*

$$P\left(\{\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T\}^T = \{(\hat{\boldsymbol{\alpha}}^{or})^T, (\hat{\boldsymbol{\beta}}^{or})^T, (\hat{\boldsymbol{\gamma}}^{or})^T\}^T\right) \to 1.$$

Combining Theorem 1 and Corollary 4, we have the asymptotic normality of the independent estimator with the optimal efficiency.

The proofs of the theorems and associated lemmas, corollaries and remarks are provided in the Supplement A.

4. Computation. Compared to traditional penalized variable selection methods, the proposed method is more complex to implement since the proposed objective function $Q_{N,m}(\cdot)$ in (7) involves an unknown homogeneous effect $\boldsymbol{\gamma}$ in addition to a non-convex penalty function. We propose an iterative algorithm as follows to simplify the optimization process.

4.1. *Algorithm and convergence property.* Note that the first term of the quadratic loss function in (7) does not involve the subgroup homogeneous effect $\boldsymbol{\gamma}$. Therefore we first fix $\boldsymbol{\gamma}$ to minimize (7) with respect to $\boldsymbol{\beta}, \boldsymbol{\alpha}$. Next, given an estimator of $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}$, we update estimator of $\boldsymbol{\gamma}$ by minimizing the grouping loss through the separation penalty term in (7). We iterate these two steps until the algorithm converges. The specific algorithm is described as follows:

In Algorithm 1, under the homogeneous variance assumption, the $\boldsymbol{V}_i$ in the quadratic loss could be replaced by a working correlation matrix $\boldsymbol{R}_i$. Specifically, we recommend one-step moment estimation for the $\boldsymbol{R}_i$ using the subject-wise least squares estimator $\boldsymbol{\beta_i}$ from an independent model.

Note that at Step 3 in Algorithm 1, the objective function (13) is a Lasso-type penalized loss function, which is convex. We can solve the optimization problem by using existing algorithms developed for Lasso. In addition, Step 4 can be implemented mimicking K-means algorithm with one subgroup centered at zero.

The following theorem provides the convergence property of Algorithm 1.

THEOREM 5. *For a sequence of estimators $\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\alpha}}^{(n)}, \hat{\boldsymbol{\gamma}}^{(n)}$ obtained in Algorithm 1, the objective function $Q_{N,m}(\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\alpha}}^{(n)}, \hat{\boldsymbol{\gamma}}^{(n)})$ in (7) is non-increasing as the number of iterations m increases, which leads to the convergence of $\hat{\boldsymbol{\beta}}^{(n)}$, $\hat{\boldsymbol{\alpha}}^{(n)}$ and $\hat{\boldsymbol{\gamma}}^{(n)}$.*

Algorithm 1

Step 1. (Initialization) Start with initial estimators: $\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}$, e.g. the OLS or Lasso estimators.

Step 2. Estimate an initial value of $\boldsymbol{\gamma}$ by $\hat{\boldsymbol{\gamma}}^{(0)} = \text{argmin}_{\boldsymbol{\gamma}} \sum_{i=1}^{N} \sum_{k=1}^{p} \min(|\hat{\beta}_{ik}^{(0)}|, |\hat{\beta}_{ik}^{(0)} - \gamma_k|)$.

Step 3. (Penalized Regression) At the nth iteration, given $\hat{\boldsymbol{\gamma}}^{(n-1)}$, update $\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\alpha}}^{(n)}$ via minimizing the objective function:

$$\frac{1}{2}\sum_{i=1}^{N}\Big(\boldsymbol{y_i} - \boldsymbol{\mu_i}(\boldsymbol{\beta_i}, \boldsymbol{\alpha})\Big)^T \boldsymbol{V}_i^{-1}\Big(\boldsymbol{y_i} - \boldsymbol{\mu_i}(\boldsymbol{\beta_i}, \boldsymbol{\alpha})\Big) + \lambda_{N,m}\sum_{i=1}^{N}\sum_{k=1}^{p} s^{(n-1)}(\beta_{ik}, \hat{\gamma_k}^{(n-1)}), \tag{13}$$

where $s^{(n-1)}(\beta_{ik}, \gamma_k) = (1 - \hat{\xi}_{ik}^{(n-1)})|\beta_{ik}| + \hat{\xi}_{ik}^{(n-1)}|\beta_{ik} - \gamma_k|$, $\hat{\xi}_{ik}^{(n-1)} = I(|\hat{\beta}_{ik}^{(n-1)}| > |\hat{\beta}_{ik}^{(n-1)} - \gamma_k|)$.

Step 4. (Grouping) Given $\hat{\boldsymbol{\alpha}}^{(n)}$, $\hat{\boldsymbol{\beta}}^{(n)}$, update $\hat{\boldsymbol{\gamma}}^{(n)} = \underset{\boldsymbol{\gamma}}{\text{argmin}} \sum_{i=1}^{N} \sum_{k=1}^{p} \min(|\hat{\beta}_{ik}^{(n)}|, |\hat{\beta}_{ik}^{(n)} - \gamma_k|)$.

Step 5. (Stopping Criterion) Iterate Step 3 and Step 4 until $\| \hat{\boldsymbol{\beta}}^{(n)} - \hat{\boldsymbol{\beta}}^{(n-1)} \|_2 + \| \hat{\boldsymbol{\alpha}}^{(n)} - \hat{\boldsymbol{\alpha}}^{(n-1)} \|_2$ is less than a small predetermined threshold value.

However, the iterative estimator may converge to a local minimizer since the objective function is non-convex. Multiple initial values are recommended so that the optimum value can be identified. In fact, the proposed piece-wise convex penalty function produces local minimums corresponding to different subgroups. However, not all individuals are sensitive to initial values except the corresponding coefficients close to boundary. Heuristically, if $\lambda_{N,m}/\gamma_k$ is small, implying that the true effects $\boldsymbol{\gamma}$ are strong, then the coefficient estimators for these individuals are stable. In addition, we recommend a step-wise tuning in practice, that is, we initialize the tuning parameter by a very small value and increase it to the specified value as the number of iterations increases.

4.2. *Tuning parameter and select number of subgroups.* In this paper, we apply the generalized cross-validation (GCV) method to select an appropriate tuning parameter $\lambda_{N,m}$. The GCV can be regarded as an approximation of leave-one-out cross-validation (CV) and thus provides an approximately unbiased estimator of the prediction error ([21]). The GCV is defined as

$$GCV(\text{df}) = \frac{RSS}{(N_0 - \text{df})^2} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{m_i}(y_{ij} - \hat{y}_{ij})^2}{(N_0 - \text{df})^2},$$

where $N_0 = \sum_{i=1}^{N} m_i$ is the total sample size anddf is the degrees of freedom used in estimating the $\hat{y}_{ij}$'s. In our setting, the degrees of freedom cannot be considered as the number of non-zero parameters, since some of the $\hat{\beta}_{ik}$'s are shrunk to the exact sub-homogeneous effect $\hat{\gamma}_k$. [21] suggest the generalized degrees of freedom (GDF) which is computationally costly. Alternatively, we define the df as the number of homogeneous effects plus the number of remaining non-zero coefficient estimators which are not equal to $\hat{\gamma}_k$'s. To select a tuning parameter $\lambda_{N,m}$, we search from a sequence of grid points which minimizes the GCV.

The proposed method allows a multiple subgroups case as defined in (3), and the number of subgroups is usually unknown. In practice, we could specify the number of the subgroups according to known scientific information or a particular target such as exploring the positive effect, the negative effect and no effect.

In practice, we can select the number of subgroups based on a data-driven approach. One approach is to adopt the idea of the jump statistic ([26]) with a K-means clustering based on some pre-estimators, e.g., the subject-wise least squares estimator. This is easy to implement but might not be reliable, as in the two-step procedure, the pre-estimators are treated as observed responses which do not change as the number of subgroups changes.

Here we provide the modified Bayesian Information Criterion (BIC, [31]) for high-dimensional data settings to select the number of subgroups. We use one individualized covariate as an illustration. The number of subgroups G_k is selected by minimizing

$$\text{BIC}(G_k) = \log\left(\sum_{i=1}^{N}\sum_{j=1}^{m}(y_{ij} - \hat{\mu}_{ij}(G_k))^2/(mN)\right) + b_{N,m}\frac{\log(mN)}{mN}(G_k + q), \tag{14}$$

where $b_{N,m}$ is a positive number and depends on N and m. When $b_{N,m} = 1$, the modified BIC reduces to the traditional BIC ([23]). For the high-dimensional setting, we follow [30] with $b_{N,m} = c\log(\log(p_{\boldsymbol{\theta}}))$, where $p_{\boldsymbol{\theta}} = N + q$ and $c = 2$. To extend to more than one individualized covariate, we adopt a strategy of selecting the number of subgroups for one predictor while fixing other individualized coefficients with the subject-wise least squares estimators.

5. Numerical Study. In this section, we provide simulation studies to investigate the numerical performance of the proposed method in finite samples. Specifically, we compare the proposed model with the subject-wise model, the homogeneous model and five other regularization models in Section 5.1. In addition, we demonstrate the benefit of incorporating within-subject correlations. In Section 5.2, we investigate the subgroup number selection of the proposed model and test the robustness against model misspecification.

5.1. *Individualized regression with correct-specified subgroup numbers.* In this simulation study, we simulate two cases to evaluate the proposed model when the number of subgroups is correctly specified. In the first case, we consider a heterogeneous regression model with one individualized variable and two population-shared variables:

$$y_{ij} = \alpha_0 + \alpha_1 z_{ij1} + \alpha_2 z_{ij2} + \beta_i x_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, m. \tag{15}$$

We set the sample size $N = 40, 100$, and the cluster size $m = 10, 20$. The individualized coefficients are set as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)' = (\underbrace{\gamma, \ldots, \gamma}_{N/2}, \underbrace{0, \ldots, 0}_{N/2})'$, where γ is the true subgroup homogeneous effect chosen as 1 or 2, and the population parameters are $\boldsymbol{\alpha}' = (\alpha_0, \alpha_1, \alpha_2) = (1, 1, 1)$. The covariates z_{ij1}, z_{ij2} and x_{ij} are generated from $N(0, 1)$. The random error ε_{ij}'s are independently generated from $N(0, 1)$.

We compare the performance of the proposed model (MDSP) with five regularized variable selection approaches, namely, the Lasso ([27]) implemented by R package *glmnet* (version 2.0-2) ([7]), the adaptive Lasso (AdapL) ([37]) solved by R package *parcor* (version 0.2-6) ([15]), the SCAD ([6]) and the MCP ([35]) implemented by R package *ncvreg* (version 3.5-1) ([4]), and the

fused Lasso (FusedL) ([29]) solved by R package *penalized* (version 0.9-50) ([8]). Note that there are $N+3$ variables and Nm observations for the above five conventional regularization models. For the fused Lasso, we order estimators of the individualized coefficients based on the least squares estimation as the fused Lasso only imposes L_1-penalties on adjacent coefficients. In addition, we also compare two non-variable-selection models, namely, the heterogeneous model (Sub) assuming subject-wise coefficients β_i's, and the homogeneous model (Homo) assuming homogeneous effect $\beta_i = \beta_h$ $(i = 1, \ldots, N)$. Both of them are based on the least squares estimators.

To evaluate the performance of these approaches on individual variable selection and prediction, we calculate the correct variable selection rate (CVSR), sensitivity and specificity, and the root mean square error (RMSE) for coefficient estimators, where the correct variable selection rate (CVSR) of the individualized variable is defined as the rate of correctly classifying β_i's $(i = 1, \ldots, N)$ to be either zero or non-zero among all individuals, and sensitivity and specificity are the true positive rate $P(\hat{\beta}_i \neq 0 | \beta_i \neq 0)$ and the true negative rate $P(\hat{\beta}_i = 0 | \beta_i = 0)$, respectively. The root mean square error is defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2$, where $\boldsymbol{\beta}^0 = (\beta_{i1}^0, \ldots, \beta_{iN}^0)'$ are the true values.

Table 1 provides the mean of root mean square errors (RMSE) based on 100 simulations. Figures 3 and 4 are the boxplots of the RMSE for all approaches. The proposed method has the smallest RMSE in all settings, which has an improvement of at least 20% $(m = 10)$ and 71% $(m = 20)$ compared to other methods for both sample sizes $N = 40, 100$ when $\gamma = 1$. The improvement is more significant and reaches 150% $(m = 10)$ and 250% $(m = 20)$ when subgroups are separated well $(\gamma = 2)$. This is because that the proposed method is able to borrow strength from different individuals within the same subgroup in estimating individualized coefficients.

The CVSR, sensitivity and specificity for the above simulations are summarized in Table 3. The proposed method (MDSP) clearly outperforms the other conventional penalization approaches in terms of the highest CVSR, especially when the subgroup homogeneous effect is large $(\gamma = 2)$. Although all models achieve similar rates on sensitivity, the proposed model leads to higher specificity rates. Figures 5–8 provide the boxplots of CVSR, sensitivity and specificity for all of the variable selection approaches.

In addition, Table 2 summarizes the estimators and the empirical standard errors of the subgroup homogeneous effects γ from the proposed model. Specifically, the estimators $\hat{\gamma}$'s are consistent as the cluster size m increases. The estimators of the population-shared coefficients $\hat{\boldsymbol{\alpha}}$ are quite similar for all methods and thus are omitted here.

In the second simulation case, we consider a subject-wise model of two individualized predictors with serial correlations:

$$y_{ij} = \beta_{i1} x_{ij1} + \beta_{i2} x_{ij2} + \varepsilon_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, m.$$

The individualized coefficients $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{N1})'$ and $\boldsymbol{\beta}_2 = (\beta_{12}, \ldots, \beta_{N2})'$ are

$$\boldsymbol{\beta}_1 = (\underbrace{\gamma_1, \ldots, \gamma_1}_{N/2}, \underbrace{0, \ldots, 0}_{N/2}), \qquad \boldsymbol{\beta}_2 = (\underbrace{0, \ldots, 0}_{N/2}, \underbrace{\gamma_2, \ldots, \gamma_2}_{N/2}),$$

where $\gamma_1 = 1$ and $\gamma_2 = -2$. We choose the sample size $N = 20, 80$ and the cluster size

$m = 10, 20$. The covariates x_{ij1} and x_{ij2} are generated from $N(0, 1)$. The random error $\boldsymbol{\varepsilon}_{\boldsymbol{i}}' = (\varepsilon_{i1}, \ldots, \varepsilon_m)$'s are generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\sigma^2 \boldsymbol{R}(\rho)$, where $\boldsymbol{R}(\rho)$ is the correlation matrix which has either an AR-1 or exchangeable structure. We set $\sigma = 1$ and $\rho = 0.5$.

We compare the performance of the proposed model using different working correlation structures to the independent model. Table 4 summarizes the average root mean square errors (RMSE) based on 100 simulations under various settings. Overall, the proposed model utilizing within-subject correlation information achieves smaller RMSE than the independent model. In particular, if the correct working structure is correctly specified, the RMSE can be reduced at least 40% compared to the one obtained using independent structure.

5.2. *Subgroup number selection and robustness.* In this simulation study, we first investigate the performance of the data-driven method discussed in Section 4 to select the number of shrinkage centers (subgroups). We compared the proposed method (MDSP) based on BIC-type criterion with a two-stage approach (OLSK) which employs the gap statistic ([28]) to choose the number of subgroups for the K-means algorithm based on the least squares estimators of individualized coefficients. The OLSK method is implemented by R package *cluster* (version 2.0.5) ([19]). The number of bootstrap samples in calculating the gap statistic is set as 100.

We generate the data following (15) under various scenarios. Scenario 1 has only a noise individualized variable ($\beta_i = 0, i = 1, \ldots, N$), while Scenarios 2 and 3 have two ($\beta_i = 0, 1$) or three subgroups ($\beta_i = 0, 2, 5$) for one individualized predictor, respectively, and Scenario 4 assumes a model of two individualized predictors with two ($\beta_{i1} = 0, 2$) or three ($\beta_{i2} = 0, -2, 1$) subgroups, respectively. The subgroup size in each scenario is balanced.

Table 5 provides the mean estimated number of subgroups and proportion of selecting the correct number of subgroups based on 100 replications. Overall, the proposed method is able to select the correct number of subgroup with more than 85% probability over all scenarios with different sample sizes ($N = 60, 120$) and cluster sizes ($m = 5, 10, 20$). The chance of selecting the correct number of subgroups increases as the cluster size increases. In addition, the proposed method consistently outperforms the two-stage OLSK method, especially when the cluster size is small ($m = 5$).

Next we test the robustness of the proposed model when the number of subgroups is misspecified. We generate the data as in model (15) under two scenarios: one has a population homogeneous predictor ($\beta_i = \gamma = 2, i = 1, \ldots, N$) and the other generates an individualized variable with three subgroups ($\gamma_0 = 0, \gamma_1 = -3, , \gamma_2 = 1$) with balanced size. We set the sample size $N = 60$ and the cluster size $m = 10$. For both cases, we fit the proposed model assuming two subgroups ($\beta_i = 0, \gamma$).

Table 6 provides the mean of RMSE and CVSR for the proposed method, the subject-wise model and the five other regularized methods described in Section 5.1. In general, the proposed method is robust against the misspecification of subgroup numbers. In the case of homogeneous effect, all models perform similarly in selecting the true variable for all individuals. However, the proposed method has the smallest RMSE among all methods with a 170% reduction. In addition, in the case

when there are fewer assumed subgroups than is true, the proposed method still has the best correct variable selection rate, and reduces the RMSE at least 14% compared to the other methods.

Figure 9 illustrates the estimation of individualized coefficients from the proposed model. In the setting where the true effect is homogeneous with individuals separate from zero, all subjects are identified correctly as one group, and are shrunk towards a non-zero group. In the scenario with three true subgroups, the subgroup with a relatively stronger signal ($\gamma_1 = -3$) is successfully identified, and therefore we gain more estimation efficiency for the individuals in this subgroup. Moreover, the subgroup with the weaker effect ($\gamma_2 = 1$) is shrunk towards zero since it is the only other shrinking direction we provide, where the proposed estimator is equivalent to the Lasso estimator.

6. Real Data Application. In this section, we illustrate the proposed individualized variable selection method using the Harvard longitudinal AIDS clinical trial group (ACTG) data. One of the goals from this study is to test the treatment effect of Zidovudine on CD4 cell counts (e.g., [5]). The 140 patients from this study are repeated measured over 14 time points with a missing rate of 8.5% and maintain CD4 counts above 50 at the baseline measures.

The demographic information includes age and gender for each patient. We denote ZDV=1 if the patient receives the treatment and ZDV=0 if the patient is in the control group. Let y_{it} be the CD4 counts for the ith patient at time t. Each individuals' CD4 measurements are standardized by within-individual standard deviation to achieve a uniform scale. A marginal model to incorporate time, treatment, interaction of time and treatment, age and gender is provided as follows:

$$y_{it} = \beta_0 + \beta_t * Time + \beta_z * ZDV + \beta_{zt} * ZDV * Time + \beta_a * Age + \beta_g * Gender + \varepsilon_{it}. \tag{16}$$

We are particularly interested in the treatment effect of Zidovudine over time. The standard analysis concludes that the marginal treatment effect over time $\hat{\beta}_{zt}$ is not significant with p-value$= 0.113$.

However, if we examine the time trend of CD4 counts from individuals, there exist subgroups for the treatment group. Given the treatment ZDV, some individuals' CD4 counts are more stable over time while some patients' CD4 counts decrease more rapidly than the average of the control group over time. This could be interpreted that some patients respond more positively, while some respond more negatively, and the remaining patients have no effects from receiving ZDV treatment compared to the average effect of the control group.

Clearly, the subgroup differences are washed out if we apply the above marginal model in (16). Therefore, we employ an individualized regression model which accommodates the personalized treatment effects ZDV over time as the following:

$$y_{it} = \beta_0 + \beta_t * Time + \beta_z * ZDV + \beta_{izt} * ZDV * Time + \beta_a * Age + \beta_g * Gender + \varepsilon_{it}.$$

We assume for the β_{izt} coefficient, that it falls into three subgroups ($\beta_{izt} = \gamma^+ > 0$, $\beta_{izt} = \gamma^- < 0$ or $\beta_{izt} = 0$). Note that for patients in the control group, we set $\beta_{izt} = 0$ since their personalized effects corresponding to the treatment are unobserved. Since the treatment variable is constant over time, we compare our proposed method with the subject-wise Lasso model, the standard population homogeneous model, the random-effects model assuming a random slope of ZDV and time interaction and the fused Lasso model.

We choose observations at times $t = 1, \ldots, 12$ as the training set and the remaining observations at $t = 13, 14$ as the testing set. On the testing set, we calculate the root mean square prediction error for each individual at $t = 13, 14$, where the median of the individuals' prediction errors is reported. Table 7 shows that the proposed method has the smallest median prediction error among all methods. For example, the proposed method has 16.0%, 13.9% and 18.1% improvement in prediction accuracy compared to the marginal model, the random-effects model and the Lasso model, respectively.

Furthermore, Figure 10 shows the individuals corresponding to no effect, positive effect and negative effect in the treatment group identified by the Lasso method and the proposed method respectively. The proposed method is able to detect more individuals with significant responses to the treatment than the Lasso method does, as the proposed separation penalty enables us to shrink the estimated coefficients in multiple directions.

To examine whether subgrouping provides more informative treatment effect over time, we refit a marginal regression model in (16) for each subgroup, where each subgroup consists of the corresponding individuals identified in the treatment group and all individuals in the control group. Table 8 illustrates that the treatment effect over time from the positive-effect subgroup selected by the Lasso method is still not significant, while the negative-effect subgroup is significant with p-value of 0.02. In contrast, the proposed method identifies both positive and negative subgroups with significant p-values of 0.02 and 0.00 respectively.

7. Discussion. In this paper, we consider an individualized regression model where both the number of subjects and the number of subject-wise repeated measurements increase. To select different important predictors for different individuals, we propose a novel multi-directional separation penalty to implement individualized variable selection. In addition, by utilizing subpopulation structure, we induce within-subgroup homogeneous effects and borrow cross-subject information to achieve a good balance of parsimonious modeling and heterogeneous interpretation.

In contrast to the conventional penalized variable selection approaches, the proposed method provides multiple shrinking directions to overcome estimation bias from L_1-regularization, where the alternative shrinking directions in addition to zero are automatically selected through grouping of subjects with similar effects from predictors. Consequently, for any subject, the proposed model achieves estimation consistency and selection consistency, even with the L_1-penalty on each shrinking direction.

In addition, compared to subject-wise modeling, the proposed method is able to achieve the population-wise oracle property when the number of the individualized parameters increases along with the sample size. Consequently, the proposed estimator inherits the optimal convergence rate from the oracle estimator due to increasing sizes of within-subject measurements and subgroups. Moreover, by incorporating within-subject serial correlation, the proposed method is able to gain more efficiency than the model assuming independence.

In this paper, the individualized and the population-shared predictors are pre-specified in the model. Therefore it is also essential to develop a method to identify individualized variables from population-shared variables prior to applying the proposed method. One possible solution is to

impose an additional penalty on sub-homogeneous effects. In addition, it is worth investigating the possibility of linking subgroup membership to population-shared covariates, such as demographic information, which could be useful for making predictions for new subjects without much prior information.

SUPPLEMENTARY MATERIAL

Supplement A: Supplement to "Individualized Multi-directional Variable Selection": (doi: COMPLETED BY THE TYPESETTER). Due to space constraints, we relegate technical details of the proofs to the supplement.

REFERENCES

[1] Balan, R. M. and Schiopu-Kratina, I. (2005). Asymptotic results with generalized estimating equations for longitudinal data. *The Annals of Statistics 32*, 522-541.

[2] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics 36*, 199-227.

[3] Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics 64*, 115-123.

[4] Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression with applications to biological feature selection. *Annals of Applied Statistics 5*, 232-253.

[5] Dolin, R., Amato, D. A., Fischl, M. A. et al. (1995). Zidovudine compared with Didanosine in patients with advanced HIV type 1 infection and little or no previous experience with Zidovudine. *Archives of Internal Medicine 155*, 961-74.

[6] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348-1360.

[7] Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33(1)*, 1-22.

[8] Goeman, J., Meijer, R., Chaturvedi, N. and Lueder, M. (2017). Penalized: L1 (Lasso and fused Lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model. R package version 0.9-50.

[9] Guo, F. J., Levina, E., Michailidis, G. and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics 66*, 793-804.

[10] Han, F. and Liu, H. (2017). Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. *Bernoulli 23(1)*, 23-57.

[11] Hocking, T., Joulin, A., Bach, F. and Vert, J.-P. (2011). Clusterpath: An algorithm for clustering using convex fusion penalties. In L. Getoor and T. Scheffer (eds.), *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*, 745-752.

[12] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Comp. 3*, 79-87.

[13] Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics 37*, 4104-4130.

[14] Ke, T., Fan, J. and Wu, Y. (2010). Homogeneity in regression. *Journal of the American Statistical Association 110*, 175-194.

[15] Kraemer, N., Schaefer, J. and Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene regulatory networks with Gaussian graphical models. *BMC Bioinformatics 10*, 384.

[16] Liang, K. -Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*, 13-22.

[17] Lindsten, F., Ohlsson, H. and Ljung, L. (2011). Clustering using sum-of-norms regularization: With application to particle filter output computation. *2011 IEEE Statistical Signal Processing Workshop (SSP)*, 201-204.

[18] Ma, S., and Huang, J. (2016) A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, in press.

[19] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P. and Gonzalez, J. (2016). Cluster: Finding groups in data. R package version 2.0.5.

[20] Pan, W. and Shen, X. (2006). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research 8*, 1145-1164.

[21] Pan, W., Shen, X. and Liu, B. (2013). Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *Journal of Machine Learning Research 14*, 1865-1889.

[22] Raftery, A. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Asscociation 101*, 168-178.

[23] Schwarz, C. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461-464.

[24] Shen, X. and Huang, H.-C. (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association 105*, 727-739.

[25] Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association 107*, 223-232.

[26] Sugar, C. A. and James, G. M. (2003). Finding the number of clusters in a data set: an information theoretic approach. *Journal of the American Statistical Association 98*, 750-763.

[27] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Ser. B 58*, 267-288.

[28] Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of data clusters via the gap statistic. *Journal of the Royal Statistical Society: Ser. B 63*, 411-423.

[29] Tibshirani, S., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of Royal Statistical Society: Ser. B 67*, 91-108.

[30] Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of Royal Statistical Society, Ser.B 71*, 671-683.

[31] Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*, 553-568.

[32] Wang, L., Zhou, J. and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics 68(2)*, 353-360.

[33] Xie, M. and Yang, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *The Annals of Statistics 31*, 310-347.

[34] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Ser. B 68*, 49-67.

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

[36] Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Ser. B 67*, 301-320.

[37] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418-1429.

TABLE 1

The average root mean square error (RMSE) of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size $N = 40, 100$, cluster size $m = 10, 20$, and subgroup homogeneous effect $\gamma = 1, 2$, where Sub, Homo, FusedL, Lasso, AdapL, SCAD and MCP stand for subject-wise model, homogeneous model, the fused Lasso ([29]), the Lasso ([27]), the adaptive Lasso ([37]), the SCAD([6]) and the MCP ([35]) regularization models, respectively. The number of subgroups (two) is correctly specified in the proposed model.

Sample Size (N)	Cluster Size(m)	Methods							
		MDSP	Sub	Homo	FusedL	Lasso	AdapL	SCAD	MCP
					$\gamma = 1$				
40	10	0.267	0.349	0.504	0.323	0.439	0.339	0.344	0.350
	20	**0.120**	0.232	0.502	0.206	0.298	0.207	0.201	0.201
100	10	0.262	0.350	0.501	0.319	0.394	0.334	0.335	0.345
	20	**0.119**	0.233	0.501	0.210	0.271	0.208	0.205	0.206
					$\gamma = 2$				
40	10	**0.122**	0.349	1.004	0.317	0.408	0.309	0.311	0.309
	20	**0.048**	0.232	1.002	0.204	0.293	0.181	0.168	0.167
100	10	**0.113**	0.350	1.001	0.318	0.387	0.305	0.300	0.299
	20	**0.037**	0.233	1.001	0.210	0.274	0.208	0.206	0.206

TABLE 2

The average RMSE of the estimated subgroup homogeneous effect $\hat{\gamma}$ from the proposed model based on 100 simulations (empirical standard errors in parenthesis), with sample size $N = 40, 100$, cluster size $m = 10, 20$.

Homogeneous Effect	N=40		N=100	
	$T = 10$	$T = 20$	$T = 10$	$T = 20$
$\gamma = 1$	1.03(0.08)	1.00(0.05)	1.02(0.05)	1.00(0.03)
$\gamma = 2$	2.01(0.07)	2.00(0.05)	2.00(0.05)	2.00(0.03)

TABLE 3

The average correct variable selection rate (CVSR), sensitivity and specificity of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size $N = 40, 100$, cluster size $m = 10, 20$, and subgroup homogeneous effect $\gamma = 1, 2$, where Sub, Homo, FusedL, Lasso, AdapL, SCAD and MCP stand for subject-wise model, homogeneous model, the fused Lasso ([29]), the Lasso ([27]), the adaptive Lasso ([37]), the SCAD([6]) and the MCP ([35]) regularization models, respectively. The number of subgroups (two) is correctly specified in the proposed model.

Variable Selection	Sample Size (N)	Cluster Size(m)	Methods **MDSP**	FusedL	Lasso	AdapL	SCAD	MCP
					$\gamma = 1$			
CVSR	40	10	0.916	0.692	0.876	0.820	0.717	0.741
		20	0.970	0.678	0.924	0.869	0.778	0.829
	100	10	0.909	0.673	0.862	0.840	0.718	0.754
		20	0.963	0.682	0.890	0.888	0.773	0.833
Sensitivity	40	10	0.942	0.978	0.898	0.943	0.975	0.966
		20	0.985	1.000	0.990	0.997	0.999	0.999
	100	10	0.946	0.986	0.917	0.941	0.974	0.967
		20	0.990	0.999	0.993	0.994	0.999	0.997
Specificity	40	10	0.909	0.406	0.853	0.696	0.460	0.517
		20	0.956	0.356	0.857	0.742	0.557	0.659
	100	10	0.886	0.360	0.807	0.739	0.462	0.542
		20	0.942	0.364	0.787	0.782	0.547	0.669
					$\gamma = 2$			
CVSR	40	10	0.959	0.639	0.886	0.884	0.800	0.852
		20	0.972	0.670	0.928	0.940	0.908	0.953
	100	10	0.940	0.648	0.868	0.898	0.809	0.871
		20	0.965	0.682	0.890	0.888	0.773	0.832
Sensitivity	40	10	0.997	0.996	0.997	0.998	1.000	0.998
		20	1.000	1.000	1.000	1.000	1.000	1.000
	100	10	0.998	0.997	0.998	0.998	0.999	0.999
		20	1.000	0.999	0.993	0.994	0.999	0.997
Specificity	40	10	0.922	0.282	0.774	0.771	0.602	0.705
		20	0.945	0.340	0.856	0.880	0.816	0.906
	100	10	0.882	0.299	0.738	0.797	0.620	0.744
		20	0.930	0.365	0.787	0.782	0.546	0.668

TABLE 4

The average root mean square error (RMSE) of the proposed MDSP model with different working correlation structures based on 100 simulations, including AR-1 ($\boldsymbol{\beta}_{AR1}$), exchangeable ($\boldsymbol{\beta}_{Ex}$) and independent ($\boldsymbol{\beta}_{Ind}$) models. The true structures for the within-subject serial correlation are AR-1 or exchangeable, and correlation parameter $\rho = 0.5$, sample size $N = 20, 80$, cluster size $m = 10, 20$.

True Correlation	Cluster size (m)	$N = 20$			$N = 80$		
		$\boldsymbol{\beta}_{AR1}$	$\boldsymbol{\beta}_{Ex}$	$\boldsymbol{\beta}_{Ind}$	$\boldsymbol{\beta}_{AR1}$	$\boldsymbol{\beta}_{Ex}$	$\boldsymbol{\beta}_{Ind}$
Exch	10	0.209	**0.165**	0.265	0.193	**0.110**	0.258
	20	0.072	**0.053**	0.078	0.067	**0.051**	0.076
AR-1	10	**0.182**	0.230	0.258	**0.183**	0.205	0.256
	20	**0.091**	0.121	0.132	**0.089**	0.112	0.130

TABLE 5

The mean of identified subgroup numbers of the proposed model compared with the two-stage OLSK method based on 100 simulations, with sample size $N = 60, 120$, cluster size $m = 5, 10, 20$. The first three scenarios contain one individualized predictor ($p = 1$) of one, two and three groups, respectively. The last scenario contains two individualized predictors ($p = 2$), one with two groups and the other with three groups. The subgroup sizes are equal in each scenario. The subgroup homogeneous effects are listed as possible values for β_i in the table.

Number of individualized variables		$p = 1$						$p = 2$			
Sample Size (N)	Cluster Size(m)	$\beta_i = 0$		$\beta_i = 0, 1$		$\beta_i = 0, 2, 5$		$\beta_{1i} = 0, 2$		$\beta_{2i} = -2, 0, 1$	
		MDSP	OLSK	MDSP	OLSK	MDSP	OLSK	MDSP	OLSK	MDSP	OLSK
60	5	1.0(100)	1.0(100)	2.0(95)	1.0(2)	2.9(88)	2.5(68)	2.0(100)	1.5(52)	3.2(85)	1.2(0)
	10	1.0(100)	1.0(100)	2.0(100)	1.3(26)	3.1(90)	2.7(74)	2.0(100)	2.0(100)	3.1(90)	2.4(44)
	20	1.0(100)	1.0(100)	2.0(100)	2.0(100)	3.1(92)	2.8(78)	2.0(100)	2.0(100)	3.0(100)	2.8(80)
120	5	1.0(100)	1.0(100)	2.0(96)	1.0(2)	3.2(86)	2.8(82)	2.0(100)	1.7(72)	3.1(90)	1.4(0)
	10	1.0(100)	1.0(100)	2.0(100)	1.2(24)	3.1(92)	2.9(86)	2.0(100)	2.0(100)	3.1(90)	2.6(64)
	20	1.0(100)	1.0(100)	2.0(100)	2.0(100)	3.0(98)	2.9(96)	2.0(100)	2.0(100)	3.1(92)	2.78(78)

TABLE 6

The average RMSE and CVSR of the proposed MDSP model compared to the subject-wise model (Sub), the fused Lasso (FusedL), the Lasso, the adaptive Lasso (Adapl), the SCAD and the MCP penalization models, based on 100 simulations with sample size $N = 60$ and cluster size $m = 10$. The first case contains a population homogeneous effect ($G_k = 1$) and the second case contains an individualized predictor of three subgroups ($G_k = 3$) with equal subgroup size. In both cases the proposed model assumes two subgroups. The estimated subgroup homogeneous effects from the proposed model are $\hat{\gamma} = 2.01(0.06)$ and $\hat{\gamma} = -2.99(0.06)$ in these two cases (with empirical standard errors in parenthesis), respectively.

Case		**MDSP**	Sub	FusedL	Lasso	AdapL	SCAD	MCP
$G_k = 1$	RMSE	0.115	0.346	0.319	0.414	0.373	0.346	0.345
$(\beta_i = 2)$	CVSR	0.996	-	0.993	0.994	0.992	0.995	0.996
$G_k = 3$	RMSE	0.277	0.349	0.315	0.410	0.335	0.337	0.338
$(\beta_i = -3, 0, 1)$	CVSR	0.901	-	0.748	0.877	0.902	0.816	0.817

TABLE 7

The estimated coefficients of the population model, the random-effects model, the L_1-penalty model and the proposed model with corresponding median prediction errors (MPE) for the ACTG data. The individualized coefficient estimators $\hat{\beta}_{izt}$'s in the Lasso model, the fused Lasso (fusedL) model and the proposed (MDSP) model are not listed.

Model	$\hat{\beta}_0$	$\hat{\beta}_t$	$\hat{\beta}_z$	$\hat{\beta}_a$	$\hat{\beta}_g$	$\hat{\beta}_{zt}$	$\hat{\gamma}^+$	$\hat{\gamma}^-$	MPE
Population	3.09	-0.68	-0.54	0.01	-0.01	-0.24	-	-	1.67
Random-effects	2.56	-0.68	-0.57	0.02	-0.01	-0.29	-	-	1.70
Lasso	3.09	-0.76	-0.54	0.01	-0.01	-	-	-	1.64
fusedL	3.05	-0.72	-0.52	0.01	-0.01	-	-	-	1.62
MDSP	3.10	-0.68	-0.56	0.01	-0.01	-	0.62	-0.60	**1.44**

TABLE 8

The treatment effect estimators within each subgroup model (zero-effect group: β^0_{zt}, negative-effect group: β^-_{zt} and positive-effect group β^+_{zt}) as well as the standard errors (s.e.) and the p-values. Each subgroup consists of the corresponding individuals in the treatment group identified by the Lasso model or the proposed model (MDSP) as well as all the individuals in the control group. The proportion of individuals with the treatment classified into each subgroup is provided.

Model		Estimates	s.e.	p-value	Proportion
Lasso	$\hat{\beta}^0_{zt}$	-0.24	0.17	0.14	0.75
	$\hat{\beta}^-_{zt}$	-0.73	0.31	0.02	0.18
	$\hat{\beta}^+_{zt}$	0.82	0.48	0.10	0.07
MDSP	$\hat{\beta}^0_{zt}$	-0.04	0.30	0.89	0.20
	$\hat{\beta}^-_{zt}$	-0.68	0.08	**0.00**	0.64
	$\hat{\beta}^+_{zt}$	0.72	0.33	**0.02**	0.16

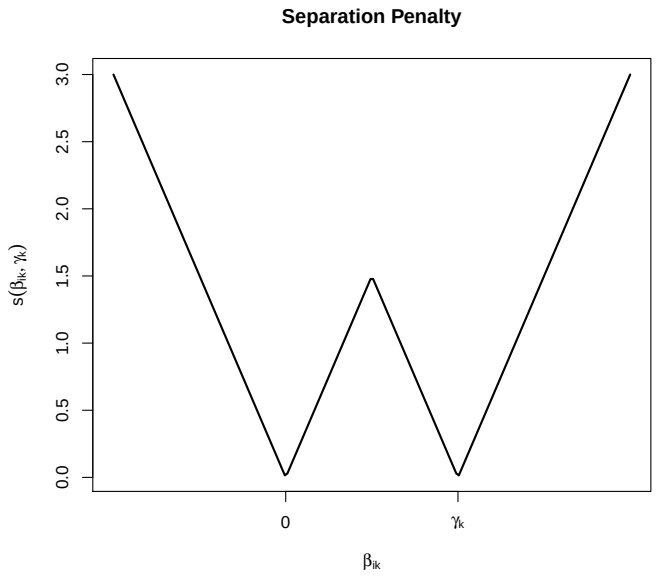

Figure 1: The separation penalty $s(\beta_{ik}, \gamma_k)$ given γ_k.

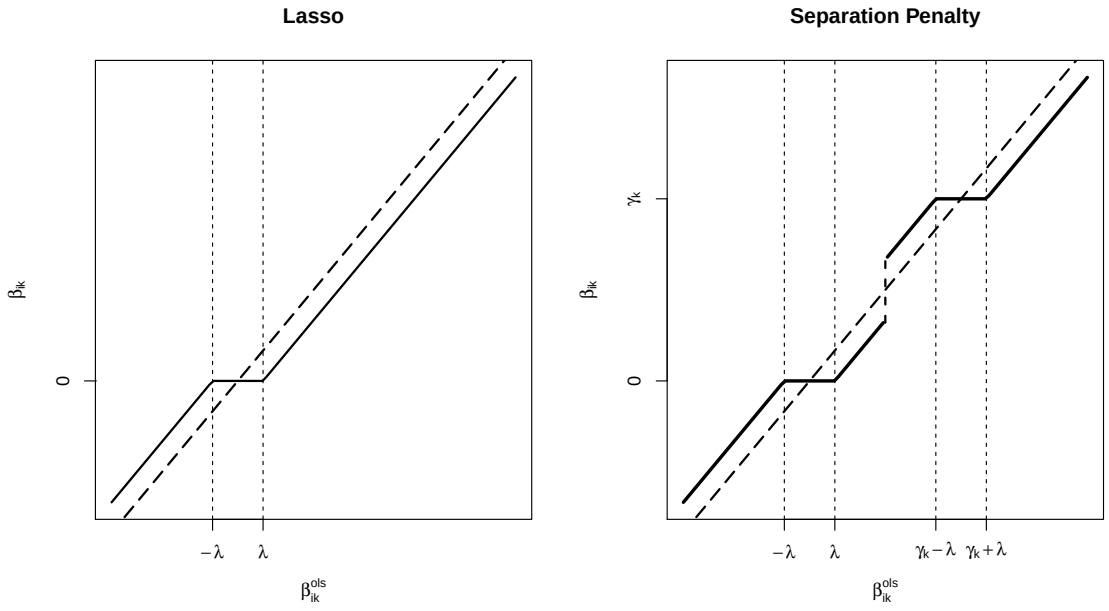

Figure 2: Thresholding functions for the Lasso and the proposed separation penalty.

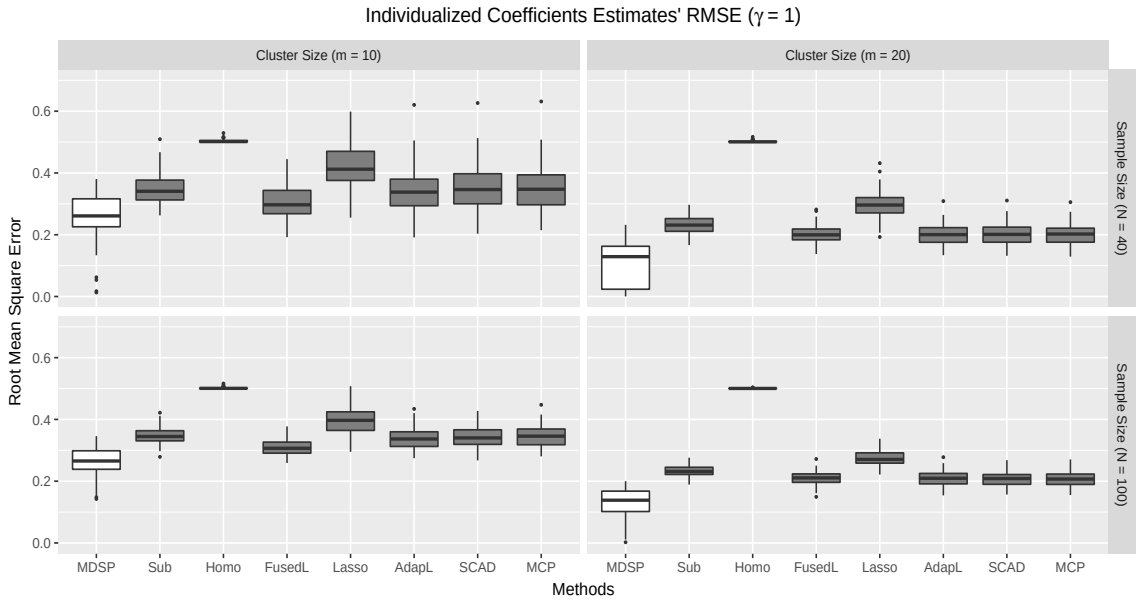

Figure 3: The boxplot of RMSE of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size $N = 40, 100$, cluster size $m = 10, 20$, where homogeneous effect $\gamma = 1$.

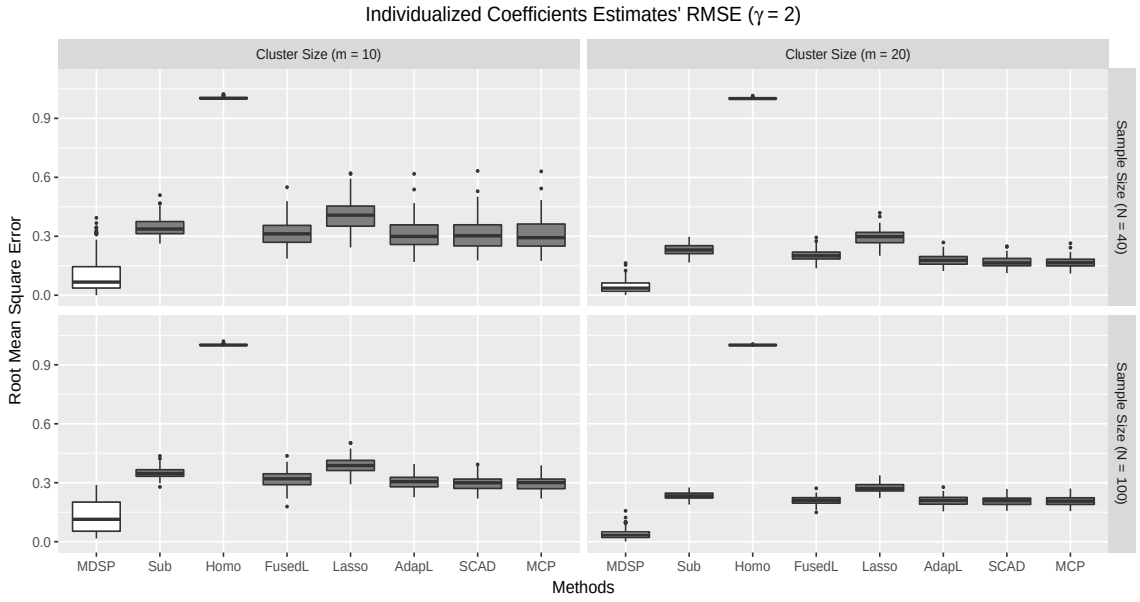

Figure 4: The boxplot of RMSE of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size $N = 40, 100$, cluster size $m = 10, 20$, where homogeneous effect $\gamma = 2$.

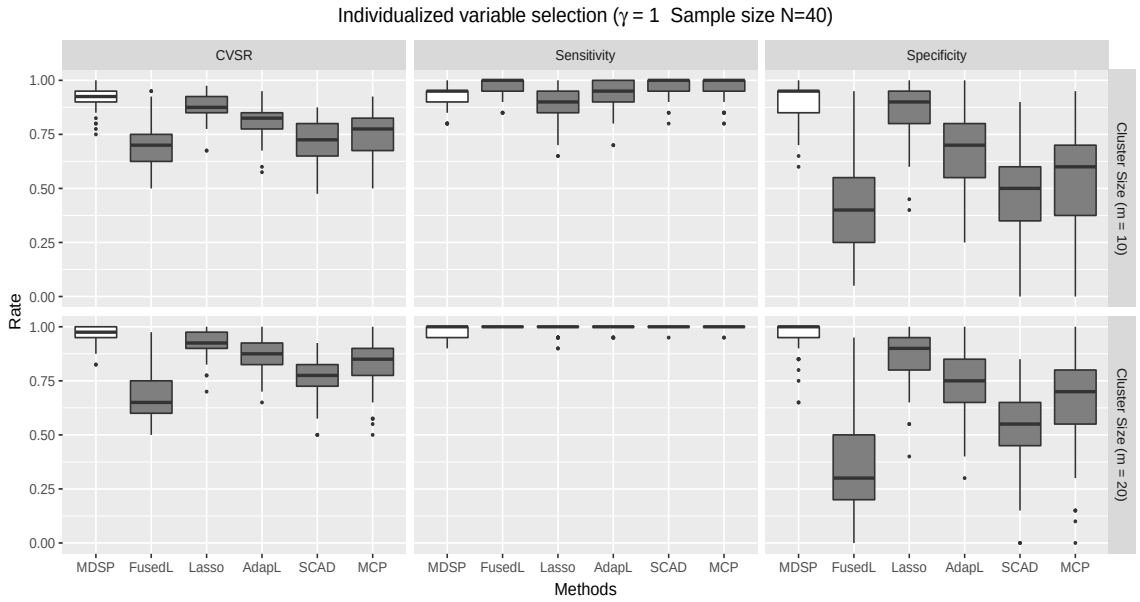

Figure 5: The boxplots of CVSR, sensitivity and specificity for all regularization approaches based on 100 simulations, with cluster size $m = 10, 20$, where homogeneous effect $\gamma = 1$ and sample size $N = 40$.

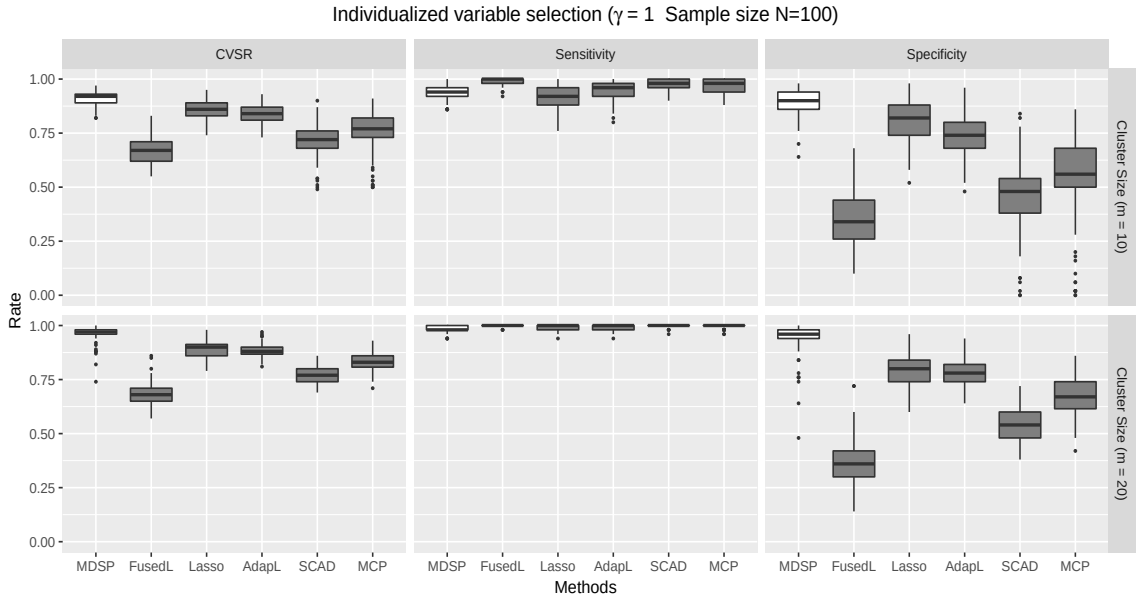

Figure 6: The boxplots of CVSR, sensitivity and specificity for all regularization approaches based on 100 simulations, with cluster size $m = 10, 20$, where homogeneous effect $\gamma = 1$ and sample size $N = 100$.

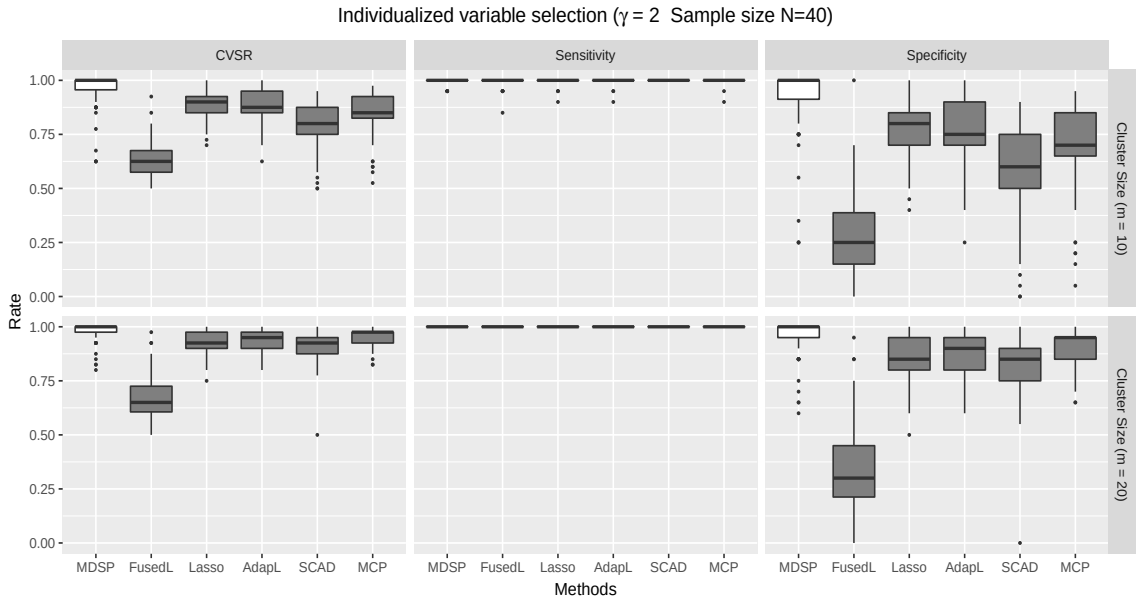

Figure 7: The boxplots of CVSR, sensitivity and specificity for all regularization approaches based on 100 simulations, with cluster size $m = 10, 20$, where homogeneous effect $\gamma = 2$ and sample size $N = 40$.

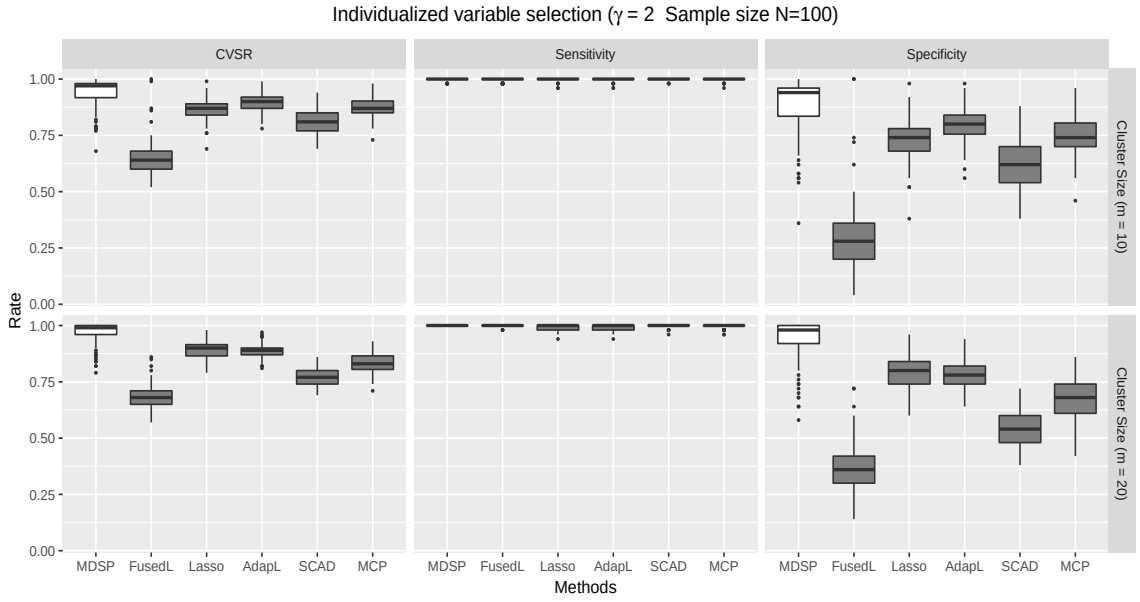

Figure 8: The boxplots of CVSR, sensitivity and specificity for all regularization approaches based on 100 simulations, with cluster size $m = 10, 20$, where homogeneous effect $\gamma = 2$ and sample size $N = 100$.

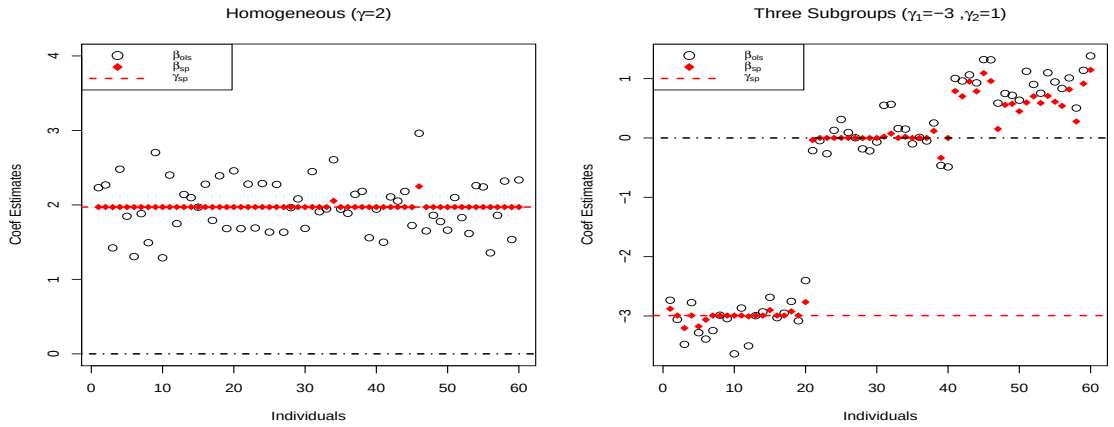

Figure 9: The subject-wise least squares estimator and the proposed estimator assuming two subgroups (including a zero group) for individualized parameters in two scenarios: a homogeneous group, and three subgroups, where the sample size $N = 60$ and cluster size $m = 10$.

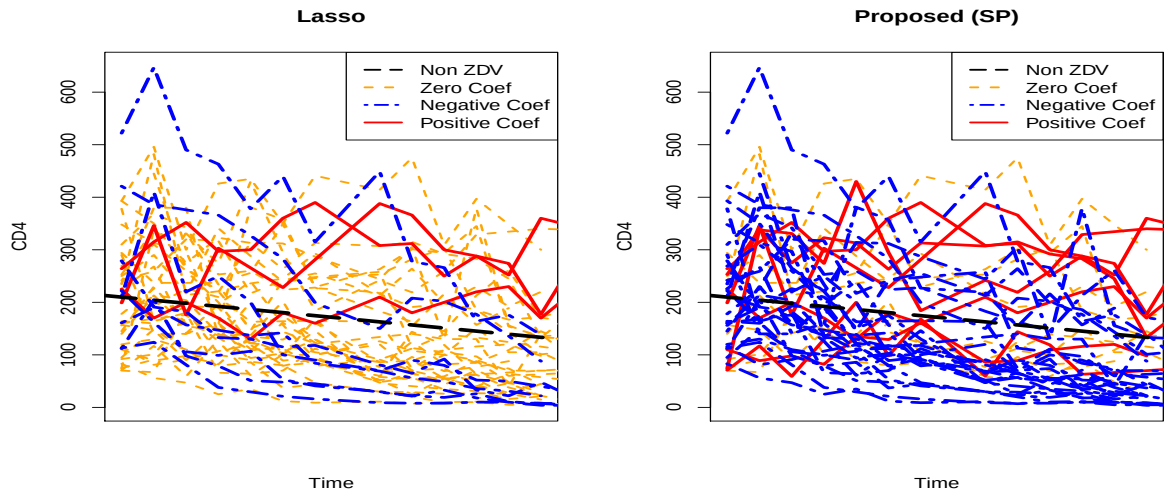

Figure 10: The different individuals corresponding to no effect, positive effect and negative effect in the treatment group selected by the Lasso model and the proposed method.

X. TANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
CHAMPAIGN, ILLINOIS, 61820, USA
E-MAIL: xtang14@illinois.edu

A. QU
DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
CHAMPAIGN, ILLINOIS, 61820, USA
E-MAIL: anniequ@illinois.edu