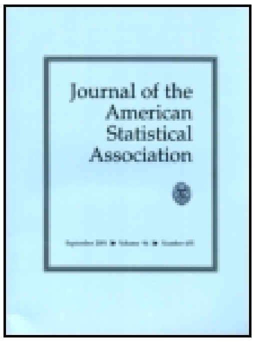


Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

# Individualized Multi-directional Variable Selection

Xiwei Tang, Fei Xue & Annie Qu

**To cite this article:** Xiwei Tang, Fei Xue & Annie Qu (2020): Individualized Multi-directional Variable Selection, Journal of the American Statistical Association, DOI: 10.1080/01621459.2019.1705308

**To link to this article:** https://doi.org/10.1080/01621459.2019.1705308

# Individualized Multi-directional Variable Selection

Xiwei Tang[1], Fei Xue[2], Annie Qu[3*]

[1]Xiwei Tang is Assistant Professor, Department of Statistics, University of Virginia, VA 22903 (E-mail: xt4yj@virginia.edu).

[2]Fei Xue is Postdoctoral Researcher, Department of Biostatistics, University of Pennsylvania, PA 19104 (Email: feixue4@illinois.edu).

[3]Annie Qu is Professor, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820 (E-mail: anniequ@illinois.edu).

Received: 28 May 2019; Revised: 15 September 2019; Accepted: 22 November 2019

***Abstract***

In this paper we propose a heterogeneous modeling framework which achieves individual-wise feature selection and heterogeneous covariates' effects subgrouping simultaneously. In contrast to conventional model selection approaches, the new approach constructs a separation penalty with multi-directional shrinkages, which facilitates individualized modeling to distinguish strong signals from noisy ones and selects different relevant variables for different individuals. Meanwhile, the proposed model identifies subgroups among which individuals share similar covariates' effects, and thus improves individualized estimation efficiency and feature selection accuracy. Moreover, the proposed model also incorporates within-individual correlation for longitudinal data to gain extra efficiency. We provide a general theoretical foundation under a double-divergence modeling framework where the number of individuals and the number of individual-wise measurements can both diverge, which enables inference on both an individual level and a population level. In particular, we establish a strong oracle property for the individualized estimator to ensure its optimal large sample property under various conditions. An efficient ADMM algorithm is developed for computational scalability. Simulation studies and applications to post-trauma mental disorder analysis with genetic variation and an HIV longitudinal treatment study are illustrated to compare the new approach to existing methods.

**Key words**: double-divergence, heterogeneous treatment effects, individualized inference, longitudinal data, multi-directional penalty, personalized prediction, subgroup analysis

## 1 Introduction

In recent years there has been a growing demand for exploring individualized modeling to account for data heterogeneity, 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 essential heterogeneous patterns and make accurate personalized predictions for specific individuals. For example, in a genetic study to identify biomarkers associated with a certain disease, one gene could be a relevant biomarker for a subgroup of individuals in the population, but not for the other individuals. In addition, the rise of precision medicine and personalized marketing strategies also motivate us to develop more effective personalized treatment and recommendation by selecting unique features for each individual. Therefore it is urgently needed to develop new statistical methodology and theory for variable selection and estimation for individualized modeling.

In the past two decades, penalized variable selection methods have been developed, e.g., the Lasso (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), the adaptive Lasso (Zou, 2006), the group Lasso (Yuan and Lin, 2006), the minimax concave penalty (MCP) (Zhang, 2010) and the truncated $L_1$-penalty (TLP) (Shen et al., 2012), based upon a homogeneous model assumption. To pursue an individualized model selection assuming different relevant predictors for different individuals, one naive choice is to apply traditional variable selection methods on each individual separately, which essentially requires multiple individual-wise observations, as in longitudinal data settings. However, in practice, the number of measurements for particular individuals could be limited, yielding inefficient model estimation and statistical inference. Therefore, it is more sensible to assume that subpopulations of individuals share common effects on selected predictors, which enables us to integrate individual information within subgroups and thus to enhance the model efficiency.

In order to utilize cross-individual information, we pursue an underlying subpopulation structure depending on unobserved covariates' effects. Existing approaches dealing with clustering on regression coefficients include the mixture-of-experts model (Jacobs et al., 1991; Tang and Qu, 2016), for which the developed variable selection procedures (Raftery and Dean, 2006; Pan and Shen, 2007; Guo et al., 2010) only focus on choosing informative variables to distinguish different subgroups. Alternative approaches to model-based clustering analysis employ grouping penalization. For example, Tibshirani et al. (2005) propose a fused Lasso by adding an $L_1$-penalty to the pair of adjacent coefficients; Bondell and Reich (2008) propose a clustering algorithm for regression by imposing a special octagonal shrinkage penalty on each pair of coefficients; Shen and Huang (2010) develop a grouping pursuit algorithm utilizing the truncated $L_1$-penalty for fusions, and Ke et al. (2013) propose a data-driven segmentation method to explore homogeneous groups with regression. Nevertheless, these approaches all assume parameter homogeneity over individuals and target on grouping similar-effect covariates. For the purpose of subgrouping different individuals, Hocking et al. (2011), Lindsten et al. (2011), Pan et al. (2013) and Ma and Huang (2017) formulate clustering as a penalized regression problem by adopting a fusion-type penalty with either an $L_p$-shrinkage or a non-convex penalty function. However, the fusion-type of penalties emphasizes on subgrouping and feature selection is not incorporated. In addition, the pairwise fusion also leads to estimation bias due to pulling individuals together from different subgroups (Rinaldo, 2009).

In this paper we propose an effective individualized model selection approach utilizing multi-directional shrinkage to select unique relevant features for different individuals and identify subgroups based on heterogeneous covariates' effects simultaneously. To the best of our knowledge, this is a new approach which has not been offered in the existing literature. From the feature selection point of view, the proposed penalty allows multiple possible shrinking directions including the one towards zero, where the best shrinking option is determined by the data itself. This provides a new perspective beyond the scope of conventional penalty functions which shrink towards zero only. One advantage of the proposed method is that, as long as the candidate directions contain one near the truth, the optimal oracle property holds by applying a regular $L_1$-penalty in each direction. Compared to traditional non-convex penalties such as the SCAD, the MCP and the TLP, the proposed approach does not rely on addition tuning parameters for penalization thresholding.

In addition to individual-wise feature selection, our paper considers a new covariate-specific subgrouping framework which is different from traditional subgroup analysis in terms of the following: (1) pursuing subgrouping on heterogeneous covariates' effects and allowing subgroup structure on individuals to vary over different covariates, which has the most flexibility in utilizing essential subgroup information compared to traditional clustering approaches with a uniform subgroup structure assumed for all covariates (Jacobs et al., 1991; Shen et al., 2012; Pan et al., 2013; Zhu et al., 2018); and (2) identifying subgroups including a specific null effect using a center-based scheme, which naturally embeds feature selection into subgrouping pursuit. Note that it is crucial to achieve simultaneous feature selection and subgrouping, as post-subgrouping inference could suffer from potential estimation bias (Desai et al., 2014; Foster et al., 2011). Through introducing the sub-homogeneous effects, the proposed approach enables individualized modeling to borrow information across individuals effectively and thus gain efficiency.

In theory, we lay out a theoretical framework for the double-divergence heterogeneous model with correlated data, where the number of individuals and the individual-wise measurement size are both increasing, yielding a divergent number of individualized parameters. Furthermore, we develop asymptotic theory for the proposed estimator under a variety of conditions and establish the optimal strong oracle property for individualized model estimation and feature selection, and uniform subgroup identification consistency.

The major contributions of theory development in this paper can be outlined as follows. (1) Traditional subgroup analysis mostly establishes theoretical results on the population or subpopulation level, for example, the average effect from a subgroup. In contrast, the theoretical framework established in this paper provides an individual-wise model inference, with a strong oracle property ensuring optimal model selection consistency, estimation efficiency and subgroup identification consistency for each individual. (2) To the best of our knowledge, in order to achieve the desired oracle property for either heterogeneous model estimation or uniform subgroup identification consistency (all individuals correctly classified), most existing penalization-based subgroup analyses (Tang and Song, 2016; Zhu and Qu, 2018) consider the scenario of a fixed number of individuals, *N*, and a divergent number of measurements on each individual, *m*, which could be restrictive in practice. The proposed double-divergence framework allows both *N* and *m* to diverge, which also provides the divergence rate of individualized parameters. (3) We also establish the theoretical results incorporating within-individual correlation under mild conditions, which brings non-trivial theoretical challenges since the dimension of the correlation structure diverges as individual measurement size increases.

The paper is organized as follows. Section 2 introduces the general framework and presents the methodology. Section 3 establishes the theoretical results. Section 4 discusses the computation and proposes an efficient algorithm. Section 5 presents simulation studies. Section 6 illustrates an application on post-trauma mental disorder analysis from the Detroit Neighborhood Health Study. The last section provides concluding remarks and discussion.

## 2 Model Framework and Methodology

### 2.1 Heterogeneous regression model

We formulate the problem under the longitudinal data setting, where each individual can have multiple observations. For the *i*th individual, let $\boldsymbol{y_i} = (y_{i,1}, \ldots, y_{i,m_i})^T$ be an $m_i$-dimensional response variable, $\boldsymbol{X_i} = (\boldsymbol{x_{i1}}, \ldots, \boldsymbol{x_{ip}})$ be an $m_i \times p$ covariate matrix of predictors with heterogeneous effects, and $\boldsymbol{Z_i} = (\boldsymbol{z_{i1}}, \ldots, \boldsymbol{z_{iq}})$ be an $m_i \times q$ covariate matrix of population-shared predictors. We consider a heterogeneous 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 is associated with a unique effect $\boldsymbol{\beta_i} = (\beta_{i1}, \ldots, \beta_{ip})^T_{p\times 1}$ for some targeting variables $\boldsymbol{X}_i$, in addition to a homogeneous effect $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^T_{q\times 1}$ for some control variables $\boldsymbol{Z}_i$. The random errors $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,m})^T_{m\times 1}$ are independent over different individuals, while within an individual, $\varepsilon_{i,t}$'s ($t = 1, \ldots, m$) have mean 0 and variance $\sigma^2$, and could be correlated. For ease of notation, we assume a balanced dataset with $m_i = m$ in this paper.

In general, to identify unique features for different individuals, with an independent error assumption and a squared loss, we could employ a penalization method to select and estimate the regression parameters $\boldsymbol{\beta_i}$'s and $\boldsymbol{\alpha}$ through minimizing the penalized objective function

$$\frac{1}{2}\sum_{i=1}^{N}\|\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta_i} - \boldsymbol{Z_i}\boldsymbol{\alpha}\|_2^2 + \sum_{i=1}^{N}\sum_{k=1}^{p} h_{\lambda_{N,m}}(\beta_{ik}), \quad (1)$$

where $\|\cdot\|_2$ denotes the Euclidean norm, and $h_{\lambda_{N,m}}(\cdot)$ refers to a feature selection penalty function, e.g., Lasso, adaptive Lasso, MCP, SCAD or TLP. Notice that the population-shared predictors $\boldsymbol{Z}_i$ mostly serve as control variables in applications, and thus, in this paper, we focus on individualized variable selection of $\boldsymbol{\beta}_i$'s.

Next, we introduce some notations here. Define $\mathrm{vec}(\boldsymbol{b}_i)_{i=1}^{N} \equiv (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_N^T)^T$ as a vectorization of a sequence of vectors $\{\boldsymbol{b}_i\}_{i=1,\ldots,N}$, and define $\mathrm{bdiag}(\boldsymbol{A}_i)_{i=1}^{N} \equiv \mathrm{diag}(\boldsymbol{A_1}, \ldots, \boldsymbol{A_N})$ as a block-diagonal matrix with a sequence of matrices $\{\boldsymbol{A}_i\}_{i=1,\ldots,N}$ at the diagonal. We let $\boldsymbol{\beta}_{(N)} = \mathrm{vec}(\boldsymbol{\beta}_i)_{i=1}^{N}$ denote the *Np*-by-1 grand vector of individualized coefficients. Furthermore, we denote $\boldsymbol{Y} = \mathrm{vec}(\boldsymbol{y}_i)_{i=1}^{N}$, $\boldsymbol{X} = \mathrm{bdiag}(\boldsymbol{X}_i)_{i=1}^{N}$ and $\boldsymbol{Z} = [\boldsymbol{Z}_1^T\ \boldsymbol{Z}_2^T \cdots \boldsymbol{Z}_N^T]^T$. Without the penalty term in (1), the ordinary least squares (OLS) estimator is obtained as

$$\mathrm{vec}(\boldsymbol{\beta}_{(N)}^{OLS}, \boldsymbol{\alpha}^{OLS}) = ([\boldsymbol{X}\ \boldsymbol{Z}]^T[\boldsymbol{X}\ \boldsymbol{Z}])^{-1}[\boldsymbol{X}\ \boldsymbol{Z}]^T\boldsymbol{Y},$$

where the dimension of parameters (*Np* + *q*) will diverge as sample size *N* increases. It is clear that the model in (1) only utilizes individual-specific information in estimating the heterogenous coefficients $\boldsymbol{\beta}_i$'s, which is hence named individual-wise modeling. As a result, this will lead to inefficient estimation and over-fitting of a model, especially when the individual-specific information is limited, e.g., when the individual-wise measurement size *m* is small.

### 2.2 Multi-directional separation penalty

To achieve more efficient individualized modeling, it is crucial and beneficial to encourage grouping some individuals which share similar treatment (covariates) effects. We propose a novel penalization approach by providing multiple shrinking directions for individualized parameters and further utilizing homogeneity information within the identified subpopulations, which achieves simultaneous parameter estimation, variable selection and individuals' subgrouping.

We propose a general framework which allows different subgrouping with respect to different heterogeneous-effect predictors. Specifically, for the individualized coefficients $\boldsymbol{\beta}_{\cdot k} = (\beta_{1k}, \ldots, \beta_{Nk})^T$ of the *k*th heterogeneous-effect predictor ($k = 1, \ldots, p$), we assume that there are $B_k$ subgroups as

$$\beta_{ik} = \begin{cases} \gamma_k^{(l)}, & \text{if} \quad i \in \mathcal{G}_k^{(l)}, \quad l = 1, \ldots, B_k - 1 \\ 0, & \text{if} \quad i \in \mathcal{G}_k^{(0)} \end{cases}, \quad \text{for } i = 1, \ldots, N, \tag{2}$$

where each $\gamma_k^{(l)}$ ($l = 1, \ldots, B_k - 1$) is an unknown non-zero sub-homogeneous effect shared by individuals within the *l*th subgroup, and the index partition sets $\{\mathcal{G}_k^{(l)}\}_{l=0,1,\ldots,B_k-1}$ represent the corresponding subgroup memberships in terms of the heterogeneous effects of the *k*th predictor. For ease of notation, in the following, we focus on the setting where there are two subgroups with respect to each heterogeneous-effect covariate: the non-zero-effect group ($\beta_{ik} = \gamma_k, i \in \mathcal{G}_k$) and the zero-effect group ($\beta_{ik} = 0, i \in \mathcal{G}_k^c$). This is rather different from conventional subgroup analysis approaches which assume a uniform subgroup structure on individuals over all covariates' effects. Detailed discussion is provided in Section 2.3.

Under the setting (2), in order to achieve simultaneous variable selection and individual subgrouping, we propose a penalized objective function with the sub-homogeneous effect $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ induced in a multi-directional separation penalty (MDSP) $s_\lambda(\cdot,\cdot)$ as

$$Q_{N,m}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{(N)}, \boldsymbol{\gamma}) = \frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_i, \boldsymbol{\alpha})\right)^T \boldsymbol{V}_i^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_i, \boldsymbol{\alpha})\right) + \sum_{i=1}^{N}\sum_{k=1}^{p} s_\lambda(\beta_{ik}, \gamma_k) \tag{3}$$

$$= L_{N,m}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{(N)}) + S_{\lambda_{N,m}}(\boldsymbol{\beta}_{(N)}, \boldsymbol{\gamma}), \tag{4}$$

where $\boldsymbol{\mu}_i(\boldsymbol{\beta}_i, \boldsymbol{\alpha}) = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{Z}_i\boldsymbol{\alpha}$. To obtain more efficient estimation, the within-individual serial correlations are utilized by a weighting matrix $\boldsymbol{V}_i = \boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{R}_i\boldsymbol{A}_i^{\frac{1}{2}}$, where $\boldsymbol{A}_i$ is a diagonal matrix of marginal variance of $\boldsymbol{y}_i$ and $\boldsymbol{R}_i$ is a working correlation matrix (Liang and Zeger, 1986).

The key component of the proposed model is a designed multi-directional separation penalty (MDSP) function $s_\lambda(\beta_{ik}, \gamma_k)$, defined as

$$s_\lambda(\beta_{ik}, \gamma_k) = \lambda_{N,m}\min\left(|\beta_{ik}|, |\beta_{ik} - \gamma_k|\right), \tag{5}$$

taking a selection over multiple marginal penalizations on individualized coefficients, where $\lambda_{N,m}$ is a tuning parameter for penalization level. This MDSP term applies in (3) with a double-summation over both individuals and covariates, essentially providing two perspectives regarding the proposed individualized modeling.

First, from an individual-wise point of view, given $\gamma_k$'s, the penalty term $\sum_{k=1}^{p} s_\lambda(\beta_{ik}, \gamma_k)$ carries feature selection on the *i*th individualized coefficients $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})^T$. Specifically, the constructed MDSP is a piece-wise convex penalization (Figure 1(a)) imposed on $\beta_{ik}$, yielding multiple shrinking directions including the one towards zero (Figure 1(b)). Traditional feature-selection penalties shrink all parameters towards zero, leading to estimation bias on those strong signals, especially with a convex penalty such as the $L_1$-penalty. By contrast, the MDSP function $s_\lambda(\cdot,\gamma_k)$ provides each $\beta_{ik}$ an alternative shrinking direction $\gamma_k$ in addition to zero, which is able to protect the strong signals from being pulled towards zero while shrinking those weak signals for sparsity. Consequently, the MDSP approach can efficiently reduce the estimation bias even with an $L_1$-penalty and does not depend on additional tuning parameters like those in non-convex penalty functions. Although the underlying effects $\gamma_k$'s are also unknown and to be estimated, as illustrated in Figure 1(b), the proposed MDSP-estimator still reduces the bias on the non-zero coefficient estimators, as long as the estimated $\hat{\gamma}_k$ provides a roughly reasonable direction along one dimension.

Essentially, if the potential alternative direction $\gamma_k$'s can be estimated precisely, the MDSP model will gain extra accuracy in model estimation and future prediction from reducing both estimation bias and variance, by pulling individual coefficients to either $\gamma_k$'s or zero. The MDSP term $\sum_{i=1}^{N} s_\lambda(\beta_{ik},\gamma_k)$ in (3) sums over individual effects of the $k$th covariate, serving as a center-based clustering. Analogous to the K-means algorithm, subgroup centers ($\gamma_k$'s) and the memberships captured by individual shrinking directions are updated iteratively in fitting the MDSP model. However, in contrast to traditional clustering algorithms, the MDSP approach carries subgrouping on estimated coefficients, $\hat{\beta}_{ik}$'s, which also change dynamically along with the updates of subgroup centers and subgroup memberships. Figure 2 provides an illustrative example to show the dynamic fitting of the MDSP approach. The updating memberships allow each individual to shrink towards an optimal direction and thus improve the individual-wise model fitting, while the $\hat{\gamma}$ is further estimated adaptively to capture the subgroup pattern of individuals. Indeed, as estimated as the centers of non-zero coefficient subgroups, the sub-homogeneous effects $\gamma_k$'s enable the individualized models to borrow information from other individuals who share similar effects and effectively utilize the subgroup information in individual-wise model estimation.

Compared to the commonly adopted fusion-based clustering models, the MDSP approach enjoys unique properties through utilizing a center-based scheme, which can efficiently integrate feature selection with clustering by fixing one of the subgroup centers as zero. Therefore, the MDSP model does not require an additional penalty term for feature selection. Furthermore, apparently, the fused Lasso suffers from significant estimation bias due to the pairwise fusion, $\sum_{i,j}|\beta_{ik}-\beta_{jk}|$, pulling individuals together even from different subgroups (Rinaldo, 2009). In contrast, the proposed MDSP approach intends to "separate" very different individual coefficients rather than to "combine" them. As long as the number of subgroups is correctly specified, the MDSP penalty does not introduce bias even with an $L_1$-shrinkage along each direction.

Though other non-convex penalties can be employed for fusion to reduce the bias, they all rely on additional thresholding parameters to control the penalization level, where the extra tuning could be challenging and computationally costly unless the subgroups are well-separated, especially in the setting which allows different subgroups for different covariates. Intuitively, to better identify subgroup clustering, a larger penalization threshold for a non-convex penalty is preferred to fuse more pairs of individual coefficients, as otherwise it leads to many local clusters. However, this is in contradiction to the purpose of reducing the bias from merging individuals from different subgroups, which requires a smaller threshold value. However, the MDSP approach is much more robust, as the embedded center-based clustering accounts for relative distances between individual coefficients and subgroup centers, and thus is less affected by the selection of tuning parameters.

The proposed MDSP can be easily generalized to accommodate various settings. For example, the above two-subgroup penalty can be extended to multiple subgroups, even with additional constraints. We illustrate the extension of three subgroups allowing positive and negative effects of individualized treatments as

$$s_{\lambda}(\beta_{ik},\gamma_k^+,\gamma_k^-) = \lambda \min\Big(|\beta_{ik}|,|\beta_{ik}-\gamma_k^+|,|\beta_{ik}-\gamma_k^-|\Big), \quad \text{s.t.} \quad \gamma_{\mathrm{k}}^+ > 0, \quad \gamma_{\mathrm{k}}^- < 0.$$

In addition, the MDSP approach can also handle the case where some of the covariates may share the same subgrouping. For example, the MDSP term can be generalized to a vector-based form: $\min\big\{|\boldsymbol{\beta}_i - \boldsymbol{\gamma}_{\boldsymbol{k}}|\big\}_{k=0,\ldots,B}$, with $\boldsymbol{\gamma}_0 = \boldsymbol{0}_p$ in particular, where $\boldsymbol{\beta}_i = (\beta_{i1},\ldots,\beta_{ip})^T$ corresponds to a group of covariates which share the same subgroup structure. Moreover, we can further generalize the model to incorporate different tuning parameters associated with penalizations on different directions, which can be useful in cases when prior knowledge such as mixing proportions is known. Furthermore, the $L_1$-penalty can be also replaced by a non-convex penalty to accommodate potential outliers.

### 2.3 Comparison with existing subgroup analysis

In this section, we make a few remarks comparing the proposed model with existing subgroup models. In addition to subgrouping on individualized regression coefficients, a key difference compared to the most of the conventional subgrouping approaches (Jacobs et al., 1991; Gunter et al., 2011; Pan et al., 2013; Ma and Huang, 2017; Zhu and Qu, 2018), is that our model in (3) allows different subgroups with respect to heterogeneous coefficients of different predictors (2). We refer to it as a *covariate-specific subgrouping.*

Specifically, we consider a simple example of a heterogeneous model with ten predictors:

$$y_{i,t} = \beta_0 + \beta_{i1}x_{i1,t} + \cdots + \beta_{i10}x_{i10,t} + \varepsilon_{i,t}, \quad i=1,\ldots,N, t=1,\ldots,m, \qquad (6)$$

where each $\beta_{ik}$ ($i=1,\ldots,N, k=1,\ldots,p$) is generated independently from a Bernoulli distribution with a probability of 0.5. Conventional clustering methods target subgrouping the coefficient vectors $\{\boldsymbol{\beta}_i \equiv (\beta_{i1},\ldots,\beta_{i10})^T\}$'s ($i=1,\ldots,N$), yielding subgroups corresponding to individuals sharing the same effects on all covariates. As a result, this limits potential applications, as the inference is still at a population level, but not at an individual level. For instance, if we further perform a variable selection based on the obtained subgroups, a variable will be selected/eliminated for all the individuals within the subgroup.

Furthermore, population-level inference can also be unreliable in many situations. Consider the above example in (6). The coefficient vector $\boldsymbol{\beta}_i$ essentially has $2^{10} = 1,024$ unique (0, 1) combinations leading to a potential 1, 024 underlying subpopulations. However, conventional clustering approaches are very likely to combine some of them as one group, e.g., $(1,\ldots,1,0)^T$ and $(1,\ldots,1,1)^T$ with finite samples, which results in estimation biases. Even under the assumption that all individuals are correctly classified into the true subpopulation, the estimation for each $\beta_{ik}$ is less efficient as it only utilizes approximately $N/1024$ samples in one subgroup, which trades-off small variance for unbiasedness. In contrast, the proposed model with covariate-specific subgrouping is able to utilize almost $N/2$ samples in estimation of each parameter, which can achieve unbiased and efficient estimation simultaneously, while allowing each individual to have a unique coefficient vector.

## 3 Theory

### 3.1 Double-divergence framework and notation

In this section, we lay out a new theoretical framework for individual-wise modeling inference and population-wise subgrouping analysis in a double-divergence structure, which allows both sample size *N* and individual measurements size *m* go to infinity.

We make contributions to two unique challenges under this framework. First, as sample size *N* increases, it is difficult to preserve the desired strong oracle property of the individualized coefficients, which enables each individual to utilize the true subpopulation information and thus to achieve optimal estimation efficiency. This is because the number of individualized parameters is diverging and a strong oracle property essentially requires a subgrouping consistency, that is, classifying all the individuals into the correct subpopulation. We establish theoretical results indicating that the proposed estimator enjoys the strong oracle property and we also outline the optimal divergence rates of *N* with different assumptions. Second, in contrast to traditional longitudinal analysis, as the number of individual measurements *m* increases, the individual-specific correlation can have a significant effect on the convergence rate of the estimator, as the correlation matrices in (3) are also expanding. We provide the convergence rate of the proposed estimator taking unknown correlation structure into account based on a double-divergence estimating equation.

We start by introducing some notation. For a 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 $\boldsymbol{A}_{m\times n}(a_{ij})$, denote $\|\boldsymbol{A}\|_2=\sqrt{\lambda_{max}(\boldsymbol{A}^T\boldsymbol{A})}$ as its $L_2$-norm, $\|\boldsymbol{A}\|_1=\max_{1\le j\le n}(\sum_{i=1}^{m}|a_{ij}|)$ as its $L_1$-norm, $\|\boldsymbol{A}\|_\infty=\max_{1\le i\le m}(\sum_{j=1}^{n}|a_{ij}|)$ as its $L_\infty$-norm, and denote $tr(\boldsymbol{A})$ as its trace. For a vector $\boldsymbol{a}=(a_1,\ldots,a_n)^T$, let $\|\boldsymbol{a}\|_0=\sum_{i=1}^{n}I_{\{a_i\neq 0\}}$. Moreover, let $\boldsymbol{A}\circ\boldsymbol{B}$ denote the entry-wise Hadamard product between two same-dimension matrices and let "⊗" denote the Kronecker product.

In addition, we let $|\mathcal{G}_k|$ denote the cardinal norm of the index set $\mathcal{G}_k\subset\{i:1,\ldots,N\}$ where $\beta_{ik}=\gamma_k$ if $i\in\mathcal{G}_k$, and $\mathcal{G}_k^c$ is its complement ($\beta_{ik}=0$). We denote $\boldsymbol{\theta}=\mathrm{vec}(\boldsymbol{\beta}_{(N)},\boldsymbol{\alpha})$ as a grand coefficients vector and let $\boldsymbol{\theta}^0=\mathrm{vec}(\boldsymbol{\beta}^0_{(N)},\boldsymbol{\alpha}^0)$ be its true value, and let $\boldsymbol{\gamma}^0$ be the true value of $\boldsymbol{\gamma}$. Furthermore, we denote the true value of an individual coefficient $\boldsymbol{\beta}_i$ as $\boldsymbol{\beta}_i^0=\mathrm{vec}(\boldsymbol{\beta}^0_{i,\mathcal{A}_i},\boldsymbol{0})$, where $\mathcal{A}_i\subset\{1,\ldots,p\}$ denotes the signal index sets such that $\boldsymbol{\beta}^0_{ik}=\boldsymbol{\gamma}^0_k$ if $k\in\mathcal{A}_i$.

The individual-wise estimator without subgrouping refers to an unpenalized estimator minimizing the squared loss function $L_{N,m}(\boldsymbol{\theta})$ in (4), which corresponds to solving the quasi-likelihood estimating equation

$$\boldsymbol{G}_{N,m}(\boldsymbol{\theta})=\sum_{i=1}^{N}\boldsymbol{g}_i(\boldsymbol{\theta})=\sum_{i=1}^{N}\boldsymbol{U}_i(\boldsymbol{\theta})^T\boldsymbol{V}_i^{-1}\big(\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\theta})\big)=0, \qquad (7)$$

where $\boldsymbol{U}_i(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$. With a linear mean function, $\boldsymbol{U}_i(\boldsymbol{\theta})$ does not actually depend on unknown parameter $\boldsymbol{\theta}$ and thus is suppressed as $\boldsymbol{U}_i$ for simple notation. In addition, we 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 $\boldsymbol{H}_{N,m}$ are both $(Np+q)$-dimensional symmetric matrices, which do not involve unknown parameter $\boldsymbol{\theta}$. Under the homogeneous variance assumption, $\boldsymbol{A}_i$ can be dropped. In addition, we usually assume $\boldsymbol{R}_i^0 = \boldsymbol{R}^0$ and choose working correlation $\boldsymbol{R}_i = \boldsymbol{R}$ for $i = 1, \ldots, N$. Due to the unknown true correlation $\boldsymbol{R}^0$, $\boldsymbol{D}_{N,m}$ and $\boldsymbol{H}_{N,m}$ are not necessarily equal, unless $\boldsymbol{R}$ is correctly specified.

Section A.2 of the Supplementary Materials lists a set of mild regularity conditions which are assumed in the following discussion. They are all standard assumptions made on regressors in penalized variable selection approaches and longitudinal data models (Xie and Yang, 2003; Balan and Schiopu-Kratina, 2005; Wang et al., 2012), with a small extension to the current individualized model setting. In particular, the standard assumptions of $\boldsymbol{R}_i^0$ converging to a constant positive definite matrix with eigenvalues bounded away from zero and infinity (Wang et al., 2012) might not be valid here, as the dimension of $\boldsymbol{R}_i^0$ diverges as the individual measurement size *m* diverges. We impose a mild regularity condition (A3) instead on the expanding correlation matrices which can be easily verified on a set of common correlation structures such as Exchangeable, AR-1 and Toeplitz.

### 3.2 Oracle estimator and unpenalized individual-wise estimator

In this section, we provide asymptotic results to the individualized estimator without penalization and the oracle estimator with true subgroup information. Both of the two estimators play important roles in understanding the individual-wise model inference and in investigating the large sample property of the proposed MDSP estimator.

The estimating equation $\boldsymbol{G}_{N,m}(\boldsymbol{\theta})$ contains double summations with sample size *N* and individual measurement size *m* that both can diverge. Therefore, the standard asymptotic results for *M*-estimators are not applicable here even with a fixed number of parameters (Xie and Yang, 2003). The following lemma implies that the consistency of the unpenalized estimator $\boldsymbol{\theta}^{u}$ solved from the equation $\boldsymbol{G}_{N,m}(\boldsymbol{\theta}) = \boldsymbol{0}$ in (7) relies on the information matrix $\boldsymbol{D}_{N,m} \boldsymbol{H}_{N,m}^{-1} \boldsymbol{D}_{N,m}$.

Lemma 1. *Under regularity conditions (A1)-(A2) provided in the Supplementary Materials, for any* $\delta > 0$*, there exists a solution* $\boldsymbol{\theta}^{u}$ *of the equation in (7) such that*

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

*where* $p_\theta = Np + q$ *is the dimension of* $\boldsymbol{\theta}$*. Moreover, if condition (*$\mathcal{C}_a$*):* $\lambda_{min}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m}) \to \infty$ *holds, we have*

$$P\Big(p_{\boldsymbol{\theta}}^{-\frac{1}{2}} \| \overset{u}{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \|_2 > \delta\Big) \to 0.$$

Remark 1. The condition ($\mathcal{C}_a$) is a standard condition analogous to the one in Xie and Yang (2003) for the weak consistency of a fixed-dimensional generalized estimating equation (GEE) estimator. In an independent model where $\boldsymbol{R}^0 = \boldsymbol{R} = \boldsymbol{I}_m$ or the working correlation $\boldsymbol{R}$ is correctly specified, the information $\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m}$ reduces to $\boldsymbol{D}_{N,m}$. Notice that, in the individualized model setting, the divergence rate of the smallest eigenvalue of $\boldsymbol{D}_{N,m}$ (the same as $\boldsymbol{H}_{N,m}$) only depends on the number of individual measurements *m*. Therefore, the condition ($\mathcal{C}_a$) essentially implies the divergence of *m*, that is, we need cumulative individual information to ensure consistent estimation.

Lemma 1 provides the consistency result under an $L_2$ norm (spectral norm), which actually requires a limited sample size *N*, otherwise the parameter dimension $p_{\boldsymbol{\theta}}$ will diverge as *N* increases. However, the proof of Lemma 1 shows that, as *m* diverges, the consistency of $\overset{u}{\boldsymbol{\theta}}$ can be guaranteed as long as *N* diverges with a limited rate. We will have more discussion regarding this point later.

Next, we provide the theoretical results for the oracle estimator, which assumes being given the true subpopulation information ($\mathcal{G}_k, 1 \le k \le p$) with respect to all individualized predictors. This is equivalent to assuming that all individualized true signal sets $\mathcal{A}_i$'s ($1 \le i \le N$) are known. Consequently, each individualized oracle parameter $\boldsymbol{\beta}_i^{or}$ is linked to the sub-homogeneous effect $\boldsymbol{\gamma}$ as $\boldsymbol{\omega_i} \circ \boldsymbol{\gamma} = \boldsymbol{\beta}_i^{or}$ through an indicator vector $\boldsymbol{\omega_i} = (\omega_{i1}, \ldots, \omega_{ip})^T$, where $\omega_{ik} = \mathbf{1}_{\{i \in \mathcal{G}_k\}} = \mathbf{1}_{\{k \in \mathcal{A}_i\}}$ and $\mathbf{1}_{\{\cdot\}}$ denotes an indicator function. Hence there exists a mapping linking two parameter spaces: $\mathbf{R}^p(\boldsymbol{\gamma}) \mapsto \mathbf{R}^{Np}(\boldsymbol{\beta}_{(N)}^{or}) : \boldsymbol{\Omega\gamma} = \boldsymbol{\beta}_{(N)}^{or}$, where $\boldsymbol{\Omega}_{Np\times p} \equiv [\boldsymbol{\Omega}_1 \cdots \boldsymbol{\Omega}_N]^T$ and $\boldsymbol{\Omega}_i = \mathrm{diag}(\boldsymbol{\omega}_i)$ is a diagonal matrix. Therefore, by noting that $S_{\lambda_{N,m}}(\boldsymbol{\beta}_{(N)}^{or}, \boldsymbol{\gamma}) = 0$, the oracle estimator is obtained as

$$\mathrm{vec}(\hat{\boldsymbol{\gamma}}^{or}, \boldsymbol{\alpha}^{or}) = \underset{\boldsymbol{\alpha},\boldsymbol{\gamma}}{\mathrm{argmin}} \sum_{i=1}^N \big(\boldsymbol{y}_i - \boldsymbol{X_i}(\boldsymbol{\omega}_i \circ \boldsymbol{\gamma}) - \boldsymbol{Z_i\alpha}\big)^T \boldsymbol{V}_i^{-1} \big(\boldsymbol{y}_i - \boldsymbol{X_i}(\boldsymbol{\omega}_i \circ \boldsymbol{\gamma}) - \boldsymbol{Z_i\alpha}\big), \tag{8}$$

and the oracle individualized estimator is $\boldsymbol{\beta}_i^{or} = \boldsymbol{\omega}_i \circ \hat{\boldsymbol{\gamma}}^{or}$. We first establish the asymptotic result for the oracle estimator with an independent model to reveal the subpopulation effect on estimation.

Theorem 1. *Under regularity conditions (A4)-(A6) provided in the Supplementary Materials, suppose* $\mathrm{vec}(\hat{\boldsymbol{\gamma}}^{or}, \boldsymbol{\alpha}^{or})$ *is the oracle estimator of an independent model obtained in (8), where* $\boldsymbol{R}^0 = \boldsymbol{R} = \boldsymbol{I}_m$*; as either* $m \to \infty$ *or* $\min_{1\le k\le p}(|\mathcal{G}_k|) \to \infty$*, we have*

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

*where* $\boldsymbol{H}_{N,m}^{or} \asymp \boldsymbol{M}_{N,m}$*, and* $\boldsymbol{M}_{N,m} = \mathrm{diag}(\underbrace{N_1,\ldots,N_p}_{p},\underbrace{N_a,\ldots,N_a}_{q})$ *is a* $(p+q)$*-dimensional diagonal matrix, in which,* $N_k = m|\mathcal{G}_k|, k=1,\ldots,p$*, and* $N_a = mN$*. The operator "*$\asymp$*" denotes that the matrix* $\boldsymbol{H}_{N,m}^{or}$ *has the same order as* $\boldsymbol{M}_{N,m}$*. The rigorous definition of "*$\asymp$*" and the explicit form of* $\boldsymbol{H}_{N,m}^{or}$ *are provided in Section A.4 of the Supplementary Materials.*

Theorem 1 indicates that the convergence rates of the oracle estimator benefit from increasing both $N$ and $m$, as it fully utilizes the subpopulation information and thus achieves optimal estimation efficiency. In particular, the convergence rates of the sub-homogeneous-effect estimator $\hat{\gamma}_k$'s are covariate-specific, corresponding to $\sqrt{N_k}$ ($1 \le k \le p$), respectively. The asymptotic result for the oracle estimator with correlated data is further discussed in the next subsection.

### 3.3 Multi-directional separation penalty estimator with correlated data

In this section, we establish the large sample results for the proposed MDSP estimator with correlated data. In addition, we provide the optimal divergence rate of $N$ that can be achieved while ensuring the oracle property of the proposed estimator.

Incorporating correlations on individual-wise measurements brings additional theoretical challenges to the double-divergence framework, as it involves divergent-dimensional correlation matrices $\boldsymbol{R}_i$ and $\boldsymbol{R}_i^0$. This makes it difficult to figure out the estimators' convergence rates. In addition to condition ($\mathcal{C}_a$), we provide an alternative sufficient condition in the following theorem, which could simplify the verification and discussion similar to Xie and Yang (2003).

Theorem 2. *Let* $\eta_m = \max_{1\le i\le N}\{\lambda_{max}(\boldsymbol{R}_i^{-1}\boldsymbol{R}_i^0)\}$*. Under regularity conditions (A3)-(A6) provided in the Supplementary Materials, for the oracle estimator* $\boldsymbol{\theta}^{or} = \mathrm{vec}(\hat{\boldsymbol{\gamma}}^{or}, \boldsymbol{\alpha}^{or})$ *obtained in (8), we have*

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

*where* $\boldsymbol{\theta}^0 = \mathrm{vec}(\boldsymbol{\gamma}^0, \boldsymbol{\alpha}^0)$*, and* $\boldsymbol{D}_{N,m}^{or}$ *is the second-order derivative matrix for the objective function in (8). The explicit form of* $\boldsymbol{D}_{N,m}^{or}$ *is provided in Section A.4 of the Supplementary Materials; Furthermore, if condition* $(\mathcal{C}_a^*)$*:* $\eta_m^{-1}\lambda_{min}(\boldsymbol{D}_{N,m}^{or}) \to \infty$ *holds, then* $\boldsymbol{\theta}^{or} \to_p \boldsymbol{\theta}^0$ *under an* $L_2$ *norm.*

Theorem 2 indicates that the convergence of the estimator depends on the divergence rate of $\eta_m$ and $\boldsymbol{D}_{N,m}^{or}$, where $\eta_m$ measures the "deviation" between the working correlation structure $\boldsymbol{R}_i$ and the true correlation structure $\boldsymbol{R}_i^0$. It is clear that if an appropriate working correlation matrix $\boldsymbol{R}_i$ is specified, we gain extra estimation efficiency by reducing $\eta_m$. However, in general, as $m \to \infty$, the value of $\eta_m$ is not always bounded. Therefore, the convergence rate of the estimator could be slower than the optimal rate $\sqrt{m}$ and it may not converge to a normal distribution asymptotically (Xie and Yang, 2003). We provide more discussion with a few common cases and some useful conditions in Section A.6 of the Supplementary Materials.

To finally establish the large sample theory for the MDSP estimator, as well as providing the divergence rate of sample size *N*, we consider two sets of assumptions on random error $\boldsymbol{\varepsilon_i}$'s:

($\mathcal{I}_a$): Assume that $\boldsymbol{\varepsilon_i} = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,m})^T$ is independent and identically generated with mean zero and the covariance matrix $\Sigma_m = \sigma^2 \boldsymbol{R}^0$, where $\sigma < \infty$, for $i = 1, \ldots, N$;

($\mathcal{I}_b$): In addition to ($\mathcal{I}_a$), let $\boldsymbol{\varepsilon_i}^* = \boldsymbol{\Sigma}_m^{-\frac{1}{2}} \boldsymbol{\varepsilon_i}$, assuming that $\boldsymbol{\varepsilon_i}^*$ is a sub-Gaussian vector, that is, $\mathrm{P}(|\boldsymbol{a}^T \boldsymbol{\varepsilon_i}^*| > t) < 2\exp(-\frac{t^2}{c_\sigma^2 \|\boldsymbol{a}\|_2^2})$ for any $\boldsymbol{a} \in \mathbf{R}^m$ and *t* > 0, where $c_\sigma$ is a positive constant.

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

Based on the above conditions and results, we establish the large sample theory for the proposed estimator under a double-divergence setting.

Theorem 3. *Let* $\tau_m = \lambda_{min}(\boldsymbol{D}_{N,m}(\boldsymbol{H}_{N,m})^{-1}\boldsymbol{D}_{N,m})$. *Under regularity conditions (A1)-(A6) provided in the Supplementary Materials, suppose* $\frac{\lambda_{N,m}}{\tau_m} \to 0$ *and* $\frac{\lambda_{N,m}}{\sqrt{\tau_m}} \to \infty$ *holds, there exists a local minimizer* $\mathrm{vec}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{(N)}, \hat{\boldsymbol{\gamma}})$ *of the MDSP objective function in (3); as* $\tau_m \to \infty$, *we have*

$$\mathrm{P}\left\{\mathrm{vec}\left(\boldsymbol{\alpha}, \boldsymbol{\beta}_{(N)}, \hat{\boldsymbol{\gamma}}\right) = \mathrm{vec}\left(\boldsymbol{\alpha}^{or}, \boldsymbol{\beta}_{(N)}^{or}, \hat{\boldsymbol{\gamma}}^{or}\right)\right\} \to 1,$$

*with*

(i). $N = o(\tau_m)$, if Assumption ($\mathcal{I}_a$) holds, or

(ii). $\log(N) = o(\tau_m)$, if Assumption ($\mathcal{I}_b$) holds.

The explicit forms of $\boldsymbol{D}_{N,m}$ and $\boldsymbol{H}_{N,m}$ are provided in Section A.7.2 of the Supplementary Materials. If the working correlation is correctly specified $\boldsymbol{R}_i = \boldsymbol{R}_i^0, 1 \le i \le N$, we have $\tau_m = \lambda_{min}(\boldsymbol{D}_{N,m})$.

Theorem 3 indicates that the proposed estimator is the same as the oracle estimator, which utilizes most of the information of the underlying subpopulation structure, ensuring that the proposed estimator inherits optimal efficiency from the oracle estimator and that the effects for each individualized predictor are correctly classified. To summarize, we achieve both individual-wise variable selection consistency and covariate-wise subgroup identification consistency as follows.

**Corollary 1** (Uniform variable selection consistency). *Under the same conditions as in Theorem 3, as* $\tau_m \to \infty$, *we have* $P\left(\bigcap_{i=1}^{N} \{\mathcal{A}_i = \mathcal{A}_i\}\right) \to 1$.

**Corollary 2** (Uniform subgroup identification consistency). *Under the same conditions as in Theorem 3, as* $\tau_m \to \infty$*, we have* $P\left(\bigcap_{k=1}^{p}\{\mathcal{G}_k = \mathcal{G}_k\}\right) \to 1$.

Theorem 3 also provides the optimal divergence rates of *N*, which depends on the order of $\tau_m$, to ensure the oracle property for the proposed estimator given different assumptions on random errors. It is apparent that $\tau_m \to \infty$ as $m \to \infty$, while the explicit order of $\tau_m$ is not easy to obtain in general as it involves unknown divergent-dimension correlation structures. Under additional assumptions or given specific structures on the correlation matrices, we are able to establish it as discussed in Section A.6 of the Supplementary Materials. In particular, with an independent error-model, by noting $\tau_m = m$, we have a simplified result as stated in the following corollary.

**Corollary 3** (Oracle property in independent model). *Under the same conditions as in Theorem 3, suppose* $\boldsymbol{R}_i = \boldsymbol{R}_i^0 = \boldsymbol{I}_m$*, for* $1 \le i \le N$*, if* $\frac{\lambda_{N,m}}{m} \to 0$ *and* $\frac{\lambda_{N,m}}{\sqrt{m}} \to \infty$*, there exists a local minimizer* $\mathrm{vec}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{(N)}, \hat{\boldsymbol{\gamma}})$ *of the MDSP objective function in (3); as* $m \to \infty$*, we have*

$$\mathrm{P}\left\{\mathrm{vec}\left(\boldsymbol{\alpha}, \boldsymbol{\beta}_{(N)}, \hat{\boldsymbol{\gamma}}\right) = \mathrm{vec}\left(\boldsymbol{\alpha}^{or}, \boldsymbol{\beta}_{(N)}^{or}, \hat{\boldsymbol{\gamma}}^{or}\right)\right\} \to 1,$$

with (i) $N = o(m)$ if Assumption ($\mathcal{I}_a$) holds, or (ii) $\log(N) = o(m)$ if Assumption ($\mathcal{I}_b$) holds.

Lastly, we consider applying the MDSP model to a new dataset such as a new individual which is usually challenging but also crucial for subgroup analysis. Since this framework focuses on unobservable predictor effects, we assume to have a semi-new individual which has initial observations $\boldsymbol{y}_i^*$ with independent errors. Given a pre-estimated sub-homogeneous effect $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)^T$ from a training dataset, we fit the model on a semi-new individual as

$$Q_{i,m^*}(\boldsymbol{\beta}_i^*, \boldsymbol{\alpha}^* \mid \hat{\boldsymbol{\gamma}}) = \frac{1}{2}\| \boldsymbol{y}_i^* - \boldsymbol{X}_i^* \boldsymbol{\beta}_i^* - \boldsymbol{Z}_i^* \boldsymbol{\alpha}^* \|_2^2 + \sum_{k=1}^{p} s_{\lambda_{m^*}}(\beta_{ik}^*, \hat{\gamma}_k). \qquad (9)$$

Theorem 4. *Suppose* $\sqrt{m^*}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0) \le O_p(1)$*. Under regularity conditions (A1)-(A6) provided in the Supplementary Materials, there exists a minimizer* $\overset{*}{\boldsymbol{\beta}}_i = \mathrm{vec}(\overset{*}{\boldsymbol{\beta}}_{i,\mathcal{A}_i}, \overset{*}{\boldsymbol{\beta}}_{i,\mathcal{A}_i^c})$ *of (9), if* $\lambda_{m^*} \to 0$ *and* $\lambda_{m^*}/\sqrt{m^*} \to \infty$*, as* $m^* \to \infty$*, we have*

$$\mathrm{P}(\overset{*}{\boldsymbol{\beta}}_{i,\mathcal{A}_i^c} = 0) \to 1 \quad \text{and} \quad \mathrm{P}(\overset{*}{\boldsymbol{\beta}}_{i,\mathcal{A}_i} = \hat{\boldsymbol{\gamma}}_{\mathcal{A}_i}) \to 1,$$

*where* $\mathcal{A}_i$ *denotes the true signal index set for the ith semi-new individual.*

Theorem 4 provides an insight from an individual-wise perspective about how the MDSP enhances individualized model inference on variable selection and model estimation. As a given $\hat{\gamma}$ provides a reasonably good direction towards sub-homogeneous effects, the individualized estimator for the semi-new individual is able to achieve selection consistency even with a limited number of observations. The theorem does not require that the given estimator $\hat{\gamma}$ is more efficient than the individualized estimator which is based on new observations only (with an order of $\sqrt{m^*}$ ). However, if $\hat{\gamma}$ is obtained from a larger training sample with a convergence rate beyond $\sqrt{m^*}$ , a single-individual based model can achieve a faster convergence rate inherited from the given $\hat{\gamma}$ .

The proofs of all of the theoretical results are provided in Appendix A of the Supplementary Materials.

## 4 Computation

### 4.1 ADMM Algorithm

The optimization problem of the objective function in (3) is challenging as it involves the non-convex penalty function with an unknown sub-homogeneous-effect parameter, yielding non-separable parameters in estimation. To achieve computational scalability, we propose an efficient ADMM-based algorithm (Boyd et al., 2011), which decomposes the original optimization into several smaller pieces that can be solved more easily.

To minimize the objective function in (3), we introduce a set of constraints $\beta_{ij} = \nu_{ij}, 1 \le i \le N, 1 \le j \le p$ , and consider a new constraint optimization problem

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\nu},\boldsymbol{\gamma}} L_{N,m}(\boldsymbol{\alpha},\boldsymbol{\beta}) + S_{\lambda_{N,m}}(\boldsymbol{\nu},\boldsymbol{\gamma}), \quad s.t. \quad \boldsymbol{\beta} = \boldsymbol{\nu}, \qquad (10)$$

where $\boldsymbol{\beta}_{Np\times 1} \equiv (\beta_{ij})_{1\le i\le N, 1\le j\le p}$ and $\boldsymbol{\nu}_{Np\times 1} \equiv (\nu_{ij})_{1\le i\le N, 1\le j\le p}$. To solve (10), we take the ADMM algorithm with the augmented Lagrangian function as

$$\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\nu},\boldsymbol{\gamma}) = L_{N,m}(\boldsymbol{\alpha},\boldsymbol{\beta}) + S_{\lambda_{N,m}}(\boldsymbol{\nu},\boldsymbol{\gamma}) + \boldsymbol{\Lambda}^T(\boldsymbol{\beta}-\boldsymbol{\nu}) + \frac{\kappa}{2}\|\boldsymbol{\beta}-\boldsymbol{\nu}\|_2^2, \qquad (11)$$

where $\boldsymbol{\Lambda}_{Np\times 1} \equiv (\Lambda_{ij})_{1\le i\le N, 1\le j\le p}$ is the Lagrangian multiplier, and $\kappa$ is a fixed augmented parameter. We update $\{\boldsymbol{\alpha},\boldsymbol{\beta}\}, \{\boldsymbol{\nu},\boldsymbol{\gamma}\}$ and $\boldsymbol{\Lambda}$ alternately at the $(l+1)$th iteration as follows:

$$\{\boldsymbol{\alpha}^{(l+1)},\boldsymbol{\beta}^{(l+1)}\} = \operatorname*{argmin}_{\boldsymbol{\alpha},\boldsymbol{\beta}} L_{N,m}(\boldsymbol{\alpha},\boldsymbol{\beta}) + \frac{\kappa}{2}\|\boldsymbol{\beta} - \boldsymbol{\nu}^{(l)} + \kappa^{-1}\boldsymbol{\Lambda}^{(l)}\|_2^2, \qquad (12)$$

$$\begin{aligned} \{\boldsymbol{\nu}^{(l+1)},\boldsymbol{\gamma}^{(l+1)}\} &= \operatorname*{argmin}_{\boldsymbol{\nu},\boldsymbol{\gamma}} S_{\lambda_{N,m}}(\boldsymbol{\nu},\boldsymbol{\gamma}) + \frac{\kappa}{2}\|\boldsymbol{\beta}^{(l+1)} - \boldsymbol{\nu} + \kappa^{-1}\boldsymbol{\Lambda}^{(l)}\|_2^2, \\ \boldsymbol{\Lambda}^{(l+1)} &= \boldsymbol{\Lambda}^{(l)} + \kappa(\boldsymbol{\beta}^{(l+1)} - \boldsymbol{\nu}^{(l+1)}). \end{aligned} \qquad (13)$$

The optimization in (12) turns to be a quadratic minimization problem given a specified working correlation structure, which leads to an explicit solution. We recommend a one-step moment estimation for the correlation structure $\boldsymbol{R}_i$ using the individual-wise estimator from an independent model. The objective function in the second optimization can be split into *p* parallel pieces based on different heterogeneous covariates as

$$\underset{\boldsymbol{\nu}_{\cdot j}}{\operatorname{argmin}} \sum_{i=1}^{N} \Big\{ \frac{\kappa}{2} (\nu_{ij} - \beta_{ij}^{(l+1)} - \kappa^{-1} \Lambda_{ij}^{(l)})^2 + \lambda_{N,m} \min(|\nu_{ij}|, |\nu_{ij} - \gamma_j|) \Big\}, \qquad (14)$$

for $j = 1, \ldots, p$, where $\boldsymbol{\nu}_{\cdot j} = (\nu_{1j}, \ldots, \nu_{Nj})'$. Along the *j*th heterogeneous covariate, we iteratively estimate $\boldsymbol{\nu}_{\cdot j}$ and $\gamma_j$ with fixed $\boldsymbol{\beta}^{(l+1)}$ and $\boldsymbol{\Lambda}^{(l)}$. Specifically, given $\gamma_j$, the $\nu_{ij}$'s ($i = 1, \ldots, N$) in (14) can be estimated separately with explicit solutions, and given $\nu_{ij}$'s, the $\gamma_j$ can be estimated via a one-dimensional exhaustive grid-search. Since all those pieces only involve univariate optimization, the minimization of (14) can be solved easily. More implementation details in (12), (13) and (14) are provided in Section B.4 of the Supplementary Materials. The proposed algorithm is outlined in Algorithm 1.

**Algorithm 1** ADMM algorithm with parallel computing

```
Initialization. Initialize $\boldsymbol{\nu}^{(0)}, \boldsymbol{\gamma}^{(0)}$. Set $\lambda_{N,m}$ and $\kappa$. Set $\boldsymbol{\Lambda} = \boldsymbol{0}$. Set stopping tolerance levels $\epsilon_1$ and $\epsilon_2$.

For $l = 0, 1, 2, \ldots$

    Step 2. Update $\{\boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}\}$ via (12).

    Step 3. Update $\{\boldsymbol{\nu}_{\cdot j}^{(l+1)}, \gamma_j^{(l+1)}\}$ via (14) with parallel computing over $j = 1, \ldots, p$.

    Step 4. Update $\boldsymbol{\Lambda}^{(l+1)} = \boldsymbol{\Lambda}^{(l)} + \kappa(\boldsymbol{\beta}^{(l+1)} - \boldsymbol{\nu}^{(l+1)})$.

    Step 5. (Stopping Criterion) Iterate Steps 2-4 until
    $\big\{\|\boldsymbol{\beta}^{(l+1)} - \boldsymbol{\beta}^{(l)}\|_2/(Np) + \|\boldsymbol{\alpha}^{(l+1)} - \boldsymbol{\alpha}^{(l)}\|_2/q + \|\boldsymbol{\gamma}^{(l+1)} - \boldsymbol{\gamma}^{(l)}\|_2/p\big\} < \epsilon_1$ and $\|\boldsymbol{r}^{(l+1)} - \boldsymbol{r}^{(l)}\|_2 < \epsilon_2$, where
    $\boldsymbol{r}^{(l)} = \boldsymbol{\beta}^{(l)} - \boldsymbol{\nu}^{(l)}$.
```

Proposition 1. *For the objective function in (3), with a sufficiently large κ, the estimator sequence generated by the proposed ADMM Algorithm 1 converges to a stationary point of (3) subsequently.*

The proof of Proposition 1 can be shown by verifying the conditions R1-R3 in Proposition 1 of Zhu et al. (2018). In practice, the iterative estimators may converge to a local minimizer due to the non-convex objective function. Multiple initial values can be applied to identify the optimum value. In fact, most individuals are not sensitive to initial values except the ones close to the boundaries of subgroups. Heuristically, if $\lambda_{N,m} / \gamma_k$ is small, implying that the true effects $\boldsymbol{\gamma}$ are strong, then the coefficient estimators are likely consistent. Therefore, we recommend using a warm-start for initialization, which can be obtained by using the individual-wise least square estimator or the proposed MDSP estimator with a very small value of $\lambda_{N,m}$ and a random initialization.

### 4.2 Tuning and subgroup number selection

In this paper, we tune the shrinkage parameter $\lambda_{N,m}$ based on the generalized cross-validation (GCV) method as suggested by Pan et al. (2013), which can be regarded as an approximation of leave-one-out cross-validation. Specifically, the GCV is defined as

$$GCV(\text{df}) = \frac{RSS}{(mN-\text{df})^2} = \frac{\|\boldsymbol{Y}-\boldsymbol{Y}\|_2^2}{(mN-\text{df})^2},$$

where $\text{df}$ is the degree of freedom used in estimating the $\boldsymbol{Y}$, and the tuning parameter $\lambda_{N,m}$ is thus selected by a grid-based search to minimize the GCV. In this setting, the degree of freedom cannot simply be treated as the total number of non-zero parameters, since some of the coefficient estimator $\hat{\beta}_{ik}$'s are shrunk to the exact sub-homogeneous effect $\hat{\gamma}_k$. Pan et al. (2013) suggests a generalized degree of freedom (GDF) which provides a more accurate estimation to the degrees of freedom. However, this procedure is computationally costly, as it depends on Monte Carlo samplings. Alternatively, we approximate the degrees of freedom ($\text{df}$) as the total number of unique non-zero coefficient estimators. This approximation can be regarded as a calculation of the df based on a grand linear model including all individuals with *Np* + *q* parameters and a series of subgroup constraints.

In general, the proposed method allows a multi-subgroup setting as defined in (2), while the number of subgroups is usually unknown and its selection is always challenging. In practice, we could specify the subgroup numbers according to known scientific information or a particular target such as exploring the positive and negative treatment effects. Alternatively, we can select the number of subgroups based on a data-driven approach. One option is to adopt the idea of the jump statistic (Sugar and James, 2003) or the gap statistic (Tibshirani et al., 2001) based on the warm-start estimators. In addition, Ma and Huang (2017) provides a subgroup number selection strategy based on the modified Bayesian Information Criterion (Wang et al., 2007a). Specifically, for the *k*th predictor, the number of subgroups $B_k$ is selected by minimizing

$$\text{BIC}(B_k) = \log\Big(\sum_{i=1}^{N}\sum_{t=1}^{m}\{y_{i,t} - \hat{\mu}_{i,t}(B_k)\}^2/(mN)\Big) + b_N\frac{\log(mN)}{mN}(B_k + q - 1),$$

where $b_N$ is a positive number depending on *N* and *m*. When $b_N$ = 1, the modified BIC reduces to the traditional BIC (Schwarz, 1978). For the high-dimensional setting, we follow Wang et al. (2009) to take $b_N = c\log(\log(p_{\boldsymbol{\theta}}))$, where $p_{\boldsymbol{\theta}} = Np + q$ and *c* = 2. To extend to multivariate individualized predictors, we select the number of subgroups one-at-a-time with penalizations only on the target predictor.

## 5 Numerical Study

### 5.1 Individualized Regression and Model Robustness

In this section, we provide simulation studies to investigate the numerical performance of the proposed method in finite samples. In the first simulation study, we consider a heterogeneous regression model with two population-shared variables and one individualized variable which, for example, can be an interested treatment effect:

$$y_{i,t} = \alpha_0 + \alpha_1 z_{i1,t} + \alpha_2 z_{i2,t} + \beta_i x_{i,t} + \varepsilon_{i,t}, \quad i = 1,\ldots,N, \quad t = 1,\ldots,m. \tag{15}$$

We set the sample size *N* = 40, 100, and the individual measurement 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 sub-homogeneous effect chosen as 1 or 2, and the population parameters are $\alpha_0 = \alpha_1 = \alpha_2 = 1$. The covariates $z_{i1,t}$, $z_{i2,t}$ and $x_{i,t}$ are generated from *N*(0, 1). The random error $\varepsilon_{i,t}$'s are independently generated from *N*(0, 1).

We compare the performance of the proposed model (MDSP) with four individual-wise regularized variable selection approaches, namely, the Lasso (Tibshirani, 1996) implemented by R package *glmnet* (version 2.0-2) (Friedman et al., 2010), the adaptive Lasso (AdapL) (Zou, 2006) solved by R package *parcor* (version 0.2-6) (Krämer et al., 2009), the SCAD (Fan and Li, 2001) and the MCP (Zhang, 2010) implemented by R package *ncvreg* (version 3.5-1) (Breheny and Huang, 2011). Moreover, we also compare two non-variable-selection models: the individual-wise least-squares model (Sub) and the homogeneous least-squares model (Homo) assuming $\beta_i = \beta$ for $i = 1,\ldots,N$.

In addition, we compare three existing subgrouping-based feature selection approaches: (1) the pairwise fused Lasso (FLPa) with an $L_1$ penalty: $\lambda_1\sum_{k=1}^{p}\sum_{i<j}|\beta_{ik}-\beta_{jk}|+\lambda_2\sum_{i=1}^{N}||\boldsymbol{\beta_i}||_1$; (2) fusion and feature selection with a truncated $L_1$-penalty (FTLP) (Shen et al., 2012): $\lambda_1\sum_{k=1}^{p}\sum_{i<j}J_\tau(|\beta_{ik}-\beta_{jk}|)+\lambda_2\sum_{k=1}^{p}\sum_{i=1}^{N}J_\tau(|\beta_{ik}|)$, where $J_\tau(a)=\min(\frac{a}{\tau},1)$; and (3) the fused Lasso (FuseL) (Tibshirani et al., 2005) with an adjacent fusion: $\lambda_1\sum_{k=1}^{p}\sum_{i'=i+1}|\beta_{ik}-\beta_{i'k}|+\lambda_2\sum_{i=1}^{N}||\boldsymbol{\beta_i}||_1$. The first two methods are both implemented by the R package *FGSG* (version 1.0.2) (Shen et al., 2012), and the last one is implemented by the R package *penalized* (version 0.9-50) (Goeman et al., 2018), where the least-squares estimators are used as initials to order the coefficients analogous to the strategy used in Tang and Song (2016).

Table 1 summarizes the average root mean square errors (RMSE) of the individualized coefficient estimator, $(Np)^{-\frac{1}{2}}||\boldsymbol{\beta}-\boldsymbol{\beta}||_2$, based on 100 simulations, while Figures 3 and 4 provide the corresponding boxplots. 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 the other methods for both sample sizes *N* = 40, 100 when *γ* = 1. The improvement is more significant reaching 150% (*m* = 10) and 250% (*m* = 20) when the subgroups are separated well (*γ* = 2). This is because the proposed method is able to borrow strength from different individuals within the same subgroup in estimating individualized coefficients, while successfully shrinking weak signals to be zero. The three fusion-based methods have similar performances, which are all better than the other non-subgrouping approaches. However, the additional parameter tuning and inefficient pairwise fusion result in greater estimation errors, and are effective than the MDSP approach.

Figures 5 and 6 provide the boxplots of correct variable identification rate (CVSR: correct rate of classifying $\beta_i$'s to be either zero or non-zero), sensitivity and specificity for all variable selection approaches. The three fusion-type approaches have very similar performances and thus the fused Lasso (FusedL) is displayed as a representative. The MDSP approach clearly outperforms the other feature selection approaches in terms of the highest CVSR and the specificity rates. Additional tables and boxplots summarizing the estimation of sub-homogeneous effects, the CVSR, sensitivity and specificity in other settings, are provided in Section B.3 of the Supplementary Materials.

In unsupervised subgrouping analysis, determining the number of subgroups is always challenging. Here we adopt the modified-BIC-based strategy introduced in Section 4.2. In the interest of space, an additional simulation study investigating the selection of subgroup numbers is reported in Section B.1 of the Supplementary Materials. In general, the proposed method is able to obtain an accurate estimation on the number of subgroups, with a probability of more than 85% to identify the correct number of subgroups under different subgrouping scenarios with various sample sizes (*N* = 60, 120), and individual repeated measurement sizes (*m* = 5, 10, 20). In addition, the proposed approach also outperforms the alternative two-stage strategy based on the individualized least-squares estimates and the Gap statistics (Tibshirani et al., 2001).

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 individualized coefficients with three subgroups ($\gamma_0 = 0, \gamma_1 = -3,, \gamma_2 = 1$) with balanced size. For both scenarios, we fit the proposed model assuming two subgroups ($\beta_i = 0, \gamma$).

Table 3 provides the average RMSEs and CVSRs for the proposed method, the individual-wise model and the five other regularized methods described in Section 5.1. Figure 7 illustrates the estimation of individualized coefficients from the proposed model. In general, the proposed method is robust against the misspecification of subgroup numbers in terms of the consistently smallest RMSE and the highest CVSR among all methods. Specifically, the MDSP model does not suffer from the homogeneous-effect setting, as all individuals are essentially shrunk towards a unique non-zero group effect. In the scenario with three true subgroups, the subgroup with a relatively stronger signal ($\gamma_1 = -3$) is successfully identified which gains more estimation efficiency, while the subgroup with the weaker effect ($\gamma_2 = 1$) is shrunk towards zero which does not have extra loss as it is just equivalent to the Lasso estimator.

**5.2 Correlated data and application on semi-new individual**

In this subsection, we investigate the performance of the proposed model utilizing within-individual correlation and its application on newly observed individuals. We consider an individual-wise model of two individualized predictors with serial correlations:

$$y_{i,t} = \alpha_0 + \alpha_1 z_{i1,t} + \alpha_2 z_{i2,t} + \beta_{i1} x_{i1,t} + \beta_{i2} x_{i2,t} + \varepsilon_{i,t}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, m. \qquad (16)$$

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

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

where $\gamma_1 = 1$ and $\gamma_2 = -2$. The covariates $z_{i1,t}, z_{i2,t}, x_{i1,t}$ and $x_{i2,t}$ are generated from *N*(0, 1). The random error $\boldsymbol{\varepsilon_i} = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,m})^T$ is generated from a multivariate normal distribution with mean **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 with $\sigma$ = 1 and $\rho = 0.5$.

Table 2 summarizes the average RMSEs of the MDSP model using different working correlation structures compared to the independent model. In general, the proposed model utilizing within-individual correlation information achieves smaller RMSE than the independent model. In particular, if the correct working structure is specified, the RMSE can be reduced at least 40% compared to the one obtained using independent structure.

As an unsupervised learning, subgrouping analysis has a great challenge in dealing with the new individuals unless additional assumptions are imposed, as in subgroup membership depending on some other observable variables. However, these assumptions are essentially difficult to validate in practice. Since this paper targets non-observable covariates effects, following the existing literature about individualized dosage (Zhu and Qu, 2016; Diaz et al., 2012), here we consider a semi-new individual with a limited number of initial individual observations. Specifically, we generate a semi-new individual with $m^*$ initial observations $\boldsymbol{y}_i^* = (y_{i1}^*, \ldots, y_{im^*}^*)^T$ with covariates $\boldsymbol{x}_{ik}^*$'s and $\boldsymbol{z}_{ik}^*$'s (*k* = 1, 2) following (16), for $i = 1, \ldots, N^*$, with independent errors, where the coefficients $\beta_{i1}^*$ and $\beta_{i2}^*$ are generated from a Bernoulli distribution with a probability of 0.5. We first estimate the sub-homogenous effects $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ by fitting an MDSP model on a training set of 100 individuals, each individual with 20 individual measurements. For the *i*th semi-new individual, we apply the MDSP model given $(\tilde{\gamma}_1, \tilde{\gamma}_2)$:

$$\min_{\alpha^*, \beta_{i1}^*, \beta_{i2}^*} \| \boldsymbol{y}_i^* - \alpha_0^* - \alpha_1^* \boldsymbol{z}_{i1}^* - \alpha_2^* \boldsymbol{z}_{i2}^* - \beta_{i1} \boldsymbol{x}_{i1}^* - \beta_{i2} \boldsymbol{x}_{i2}^* \|_2^2 + \sum_{k=1}^{2} s_{\lambda^*}(\beta_{ik}^*, \tilde{\gamma}_k).$$

We investigate the parameter estimation (RMSE) and the variable selection (for $\beta_1$ and $\beta_2$) on a semi-new individual using the MDSP model, the individual-specific linear model, and the individual-specific Lasso model. For the linear model, the variable selection is based on the marginal p-value with a significance level of 0.05. All results are evaluated based on $N^* = 100$ semi-new individuals with $m^*$ varying from 6 to 20. We add a homogeneous model estimator from the training as a reference.

Figure 8 shows that the MDSP model consistently achieves the smallest RMSE values, indicating the most efficient prediction accuracy, and also has the best accuracy in predictor selection/elimination. The improvement of the MDSP model is more significant as the semi-new individual has fewer initial observations, e.g., when $m^* = 6$, the MDSP model reduces the RMSE value by 476% and 62% compared to the OLS model and the Lasso model, respectively. In addition, the MDSP model also consistently outperforms the homogeneous model with an improvement of at least 34% (and up to 250% as *m* increases) in the RMSE value.

## 6 Real Data Application

In this section, we apply the proposed individualized variable selection method to the Detroit Neighborhood Health Study (DNHS) (https://dnhs.unc.edu/), which is a representative longitudinal study investigating genetic variation or traumatic events effects on mental disorders of African American adults in Detroit, Michigan.

The DNHS contains blood samples and five-wave surveys which ask questions about demographics, traumas, stressful events, and post-traumatic stress disorder (PTSD). The survey at each wave includes a post-traumatic checklist (PCL) based on incident trauma exposures, which is a 17-item self-reported measure of PTSD symptoms. We treat the average of 17 PCL scores as the response variable with a logarithm transformation. Studies (Rusiecki et al., 2013; Chen et al., 2016) show that pathophysiology of PTSD is associated with DNA methylation (DNAm) in glucocorticoid receptor regulatory network (GRRN) genes, since the process is intrinsically linked to gene regulation. To identify cytosine-phosphate-guanine (CpG) sites in GRRN genes which are significantly associated with PTSD, we use DNAm values at 1648 CpG sites as potential predictors.

Specifically, we target investigating the potential heterogeneous effects of the CpG predictors on the PCL scores. In addition, we incorporate the numbers of traumas and stressful events as homogeneous control variables. The DNHS has 126 individuals with traumas whose average PCL scores in the first and second waves are completely observed. Since missing rates of average PCL scores from the third to fifth waves are higher than 50% and our sample size is limited, we impute the missing response values $y_{it}^{*}$ (for the *i*th individual at the *t*th wave) from $N(\mu_i, 0.35^2)$, where $\mu_i$ is the individual mean calculated based on previous observed $y_{it}$'s, while 0.35 is determined based on the sample standard deviation of all complete responses. We split the data into training and testing sets with three waves and two waves, respectively.

Given the limited number of individual-wise repeated measurements (three waves for training) and the ultrahigh-dimensional covariates (1,648 CpG sites), we carry out a screening process to identify potential covariates with significant heterogeneous effects. We fit a marginal homogenous model for each CpG predictor and filter out the CpG cites with p-values greater than 0.4, which are unlikely to have significant effects for any reasonably large subgroup. For the remaining 376 covariates, we fit a marginal MDSP model to each of them and estimate the number of subgroups based on the gap statistic (Tibshirani et al., 2001). We are able to identify three CpG sites (*cg03256465, cg03762702* and *cg06473843*) which have significant heterogeneous effects.

For illustration, we compare the proposed MDSP model with the homogeneous regression model and the mixture-of-regression model (McLachlan and Peel, 2004). Notice that all DNAm values at the CpG sites are measured only once, thus there is no variation on those covariates within an individual over longitudinal waves. Therefore, any individual-wise models such as the individual-wise OLS model and the Lasso model as well as the random-effects model are inapplicable. We implement the mixture of regression model by the R package *"mixtools"* (version 1.1.0) where the number of the mixture components is selected as two by bootstrap sequential testing (McLachlan and Peel, 2004).

To evaluate the model performance, we calculate the average prediction RMSE of the response PCL scores on the testing dataset. In addition, to examine whether subgrouping (the MDSP model and the mixture model) provides more informative data structure, we refit a homogeneous model within each identified subgroup, and report the marginal p-values for CpG predictors, respectively.

Table 4 summarizes the RMSE values and the p-values of the estimated CpG coefficients. The MDSP model reduces the RMSE by 15% and 32% compared to the mixture model and the homogeneous model, respectively. For variable selection, the homogeneous model does not provide any significant results. However, the MDSP model successfully obtains significant p-values corresponding to three CpG sites with identified non-zero-effect subgroups, while the p-values in the zero-effect subgroups are clearly insignificant. In contrast, only one CpG site (*cg0647384*) presents significance in one subgroup of the mixture model (Component 1). This indicates that the MDSP model provides more informative subgrouping structure as it achieves individualized variable selection and subgrouping simultaneously. Additionally, we note that the non-zero-effect subgroups identified by the MDSP model have reasonably large sizes, consisting of 36.5%, 34.2% and 40.4% of sample size with respect to CpG sites *cg03256465, cg03762702* and *cg06473843*.

In Section B.2 of the Supplementary Materials, we provide another illustration of the proposed method analyzing the Harvard longitudinal AIDS clinical trial group data to investigate the heterogeneous treatment effects of Zidovudine on CD4 cell counts.

## 7 Discussion

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

The proposed multi-directional separation penalty naturally embeds feature selection into subgrouping pursuit by leveraging a center-based clustering scheme with a subgroup center of zero. The alternative-directional shrinkage provides a new perspective beyond the scope of traditional penalization approaches, where the oracle properties can be achieved even with a convex penalty along each direction. Moreover, by incorporating within-individual serial correlation, the proposed method is able to gain more efficiency than the model assuming independence.

In subgroup analysis, to access heterogeneous covariates' effects, the existing literature (Shuster and van Eys, 1983; Gail and Simon, 1985; Yusuf et al., 1991; Lagakos, 2006; Wang et al., 2007b; Gunter et al., 2011; Rendle, 2012) proposes adding more interaction terms under a homogeneous model setting, which relies on pre-specified model assumptions such as linear relationships (Gunter et al., 2011; Rendle, 2012). However, these assumptions are usually difficult to verify in applications. The covariates' heterogeneity could be more complex due to, for example, unobserved factors rather than observed covariates. By contrast, the proposed method detects heterogeneous structures on individual covariates' effects without relying on additional model assumptions on subgroup mechanisms.

To provide individual-wise model inference, we lay out a double-divergence theoretical framework which allows both sample size and individual-wise measurement size to diverge, and also incorporates a divergent longitudinal correlation structure. The established large sample results indicate that the proposed method achieves a strong oracle property and thus inherits the optimal convergence rate with true subpopulation information. In addition, we also provide the optimal divergence rate of the dimension of individualized parameters as the sample size increases.

In this paper, we mainly consider a fixed dimension of individualized covariates, as the number of typical individualized predictors for heterogeneous modeling is usually limited in practice, such as the treatments in personalized medicine (Yusuf et al., 1991; Gunter et al., 2011) or biomarkers in personalized cancer genomics (Tursz et al., 2011; Simon and Roychowdhury, 2013), due to the limited size of individual-wise observations. It would be of great interest to extend the theory to a diverging number of individualized covariates, which could follow the standard results for a high-dimensional setting applying an individual-wise Lasso model, and then incorporating grouping effects through a similar strategy as in Theorem 3 of this paper.

## Supplementary Materials

The online supplement contains all technical proofs, additional numerical results and computation details.

## Acknowledgments

The authors would like to acknowledge support for this project from the National Science Foundation grants DMS-1613190, DMS-1821198, the National Institute of Health grant 7R01MD011728-03, and Dr. Monica Uddin to provide the PTSD data. The authors are grateful to reviewers, the Associate Editor and Editor for their insightful comments and suggestions which have improved the manuscript significantly.

## References

Balan, R. M. and Schiopu-Kratina, I. (2005). Asymptotic results with generalized estimating equations for longitudinal data. *The Annals of Statistics*, 33(2):522–541.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123.

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

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232.

Chen, Y., Li, X., Kobayashi, I., Tsao, D., and Mellman, T. A. (2016). Expression and methylation in posttraumatic stress disorder and resilience; evidence of a role for odorant receptors. *Psychiatry research*, 245:36–44.

Desai, M., Pieper, K. S., and Mahaffey, K. (2014). Challenges and solutions to pre-and post-randomization subgroup analyses. *Current cardiology reports*, 16(10):531.

Diaz, F. J., Cogollo, M. R., Spina, E., Santoro, V., Rendon, D. M., and de Leon, J. (2012). Drug dosage individualization based on a random-effects linear model. *Journal of biopharmaceutical statistics*, 22(3):463–484.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, pages 361–372.

Goeman, J., Meijer, R., Chaturvedi, N., Lueder, M., Goeman, M. J., Rcpp, I., and Rcpp, L. (2018). Package ‘penalized’. *R package version*.

Gunter, L., Zhu, J., and Murphy, S. (2011). Variable selection for qualitative interactions. *Statistical methodology*, 8(1):42–55.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, 66(3):793–804.

Hocking, T. D., Joulin, A., Bach, F., and Vert, J.-P. (2011). Clusterpath an algorithm for clustering using convex fusion penalties.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., Hinton, G. E., et al. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Ke, T., Fan, J., and Wu, Y. (2013). Homogeneity in regression. *arXiv preprint arXiv:1303.7409*.

Krämer, N., Schäfer, J., and Boulesteix, A.-L. (2009). Regularized estimation of large scale gene regulatory networks.

Lagakos, S. W. (2006). The challenge of subgroup analyses-reporting without distorting. *New England Journal of Medicine*, 354(16):1667.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

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

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

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8(May):1145–1164.

Pan, W., Shen, X., and Liu, B. (2013). Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty. *The Journal of Machine Learning Research*, 14(1):1865–1889.

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178.

Rendle, S. (2012). Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):57.

Rinaldo, A. (2009). Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5B):2922–2952.

Rusiecki, J. A., Byrne, C., Galdzicki, Z., Srikantan, V., Chen, L., Poulin, M., Yan, L., and Baccarelli, A. (2013). Ptsd and dna methylation in select immune function gene promoter regions: a repeated measures case-control study of us military service members. *Frontiers in Psychiatry*, 4:56.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Shen, X. and Huang, H.-C. (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, 105(490):727–739.

Shen, X., Huang, H.-C., and Pan, W. (2012). Simultaneous supervised clustering and feature selection over a graph. *Biometrika*, 99(4):899–914.

Shuster, J. and van Eys, J. (1983). Interaction between prognostic factors and treatment. *Controlled clinical trials*, 4(1-2):209–214.

Simon, R. and Roychowdhury, S. (2013). Implementing personalized cancer genomics in clinical trials. *Nature reviews Drug discovery*, 12(5):358–369.

Sugar, C. A. and James, G. M. (2003). Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association*, 98(463):750–763.

Tang, L. and Song, P. X. (2016). Fused lasso approach in regression coefficients clustering: learning parameter heterogeneity in data integration. *The Journal of Machine Learning Research*, 17(1):3915–3937.

Tang, X. and Qu, A. (2016). Mixture modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, 25(4):1117–1137.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Tursz, T., Andre, F., Lazar, V., Lacroix, L., and Soria, J.-C. (2011). Implications of personalized medicine–perspective from a cancer center. *Nature Reviews Clinical Oncology*, 8(3):177.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.

Wang, H., Li, R., and Tsai, C.-L. (2007a). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568.

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

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007b). Statistics in medicine–reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21):2189–2194.

Xie, M. and Yang, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *The Annals of Statistics*, 31(1):310–347.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Yusuf, S., Wittes, J., Probstfield, J., and Tyroler, H. A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *Jama*, 266(1):93–98.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhu, X. and Qu, A. (2016). Individualizing drug dosage with longitudinal data. *Statistics in medicine*, 35(24):4474–4488.

Zhu, X. and Qu, A. (2018). Cluster analysis of longitudinal profiles with subgroups. *Electronic Journal of Statistics*, 12(1):171–193.

Zhu, X., Tang, X., and Qu, A. (2018). Longitudinal clustering for heterogeneous binary data. *Statistica Sinica*.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

**Fig. 1** An illustration of the MDSP function in an individual-wise variable selection. (a) The MDSP function $s(\cdot,\gamma_k)$ given $\gamma_k$; (b) The $L_1$-penalized estimator and the MDSP-penalized estimator (only on $\beta_2$: $s(\boldsymbol{\beta},\gamma_2)=|\beta_1|+\min(|\beta_1|,|\beta_2-\gamma_2|)$) for an individual model, where $\gamma_2$ is given and $\boldsymbol{\beta}^{LS}$ denotes the least-squares estimator.

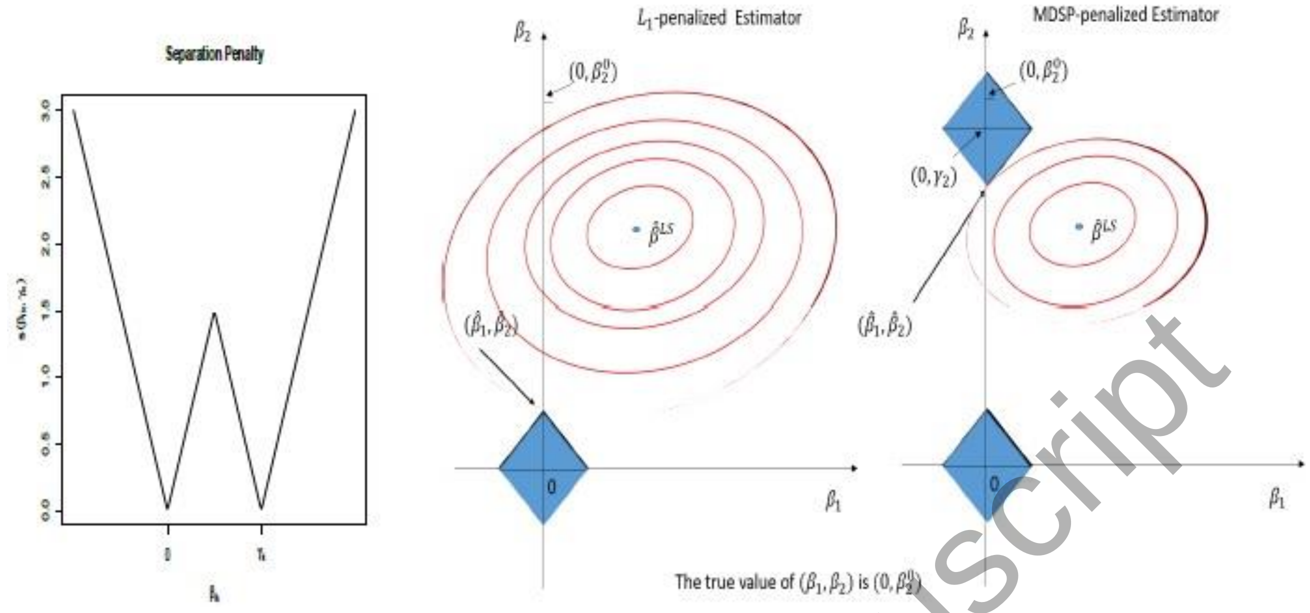
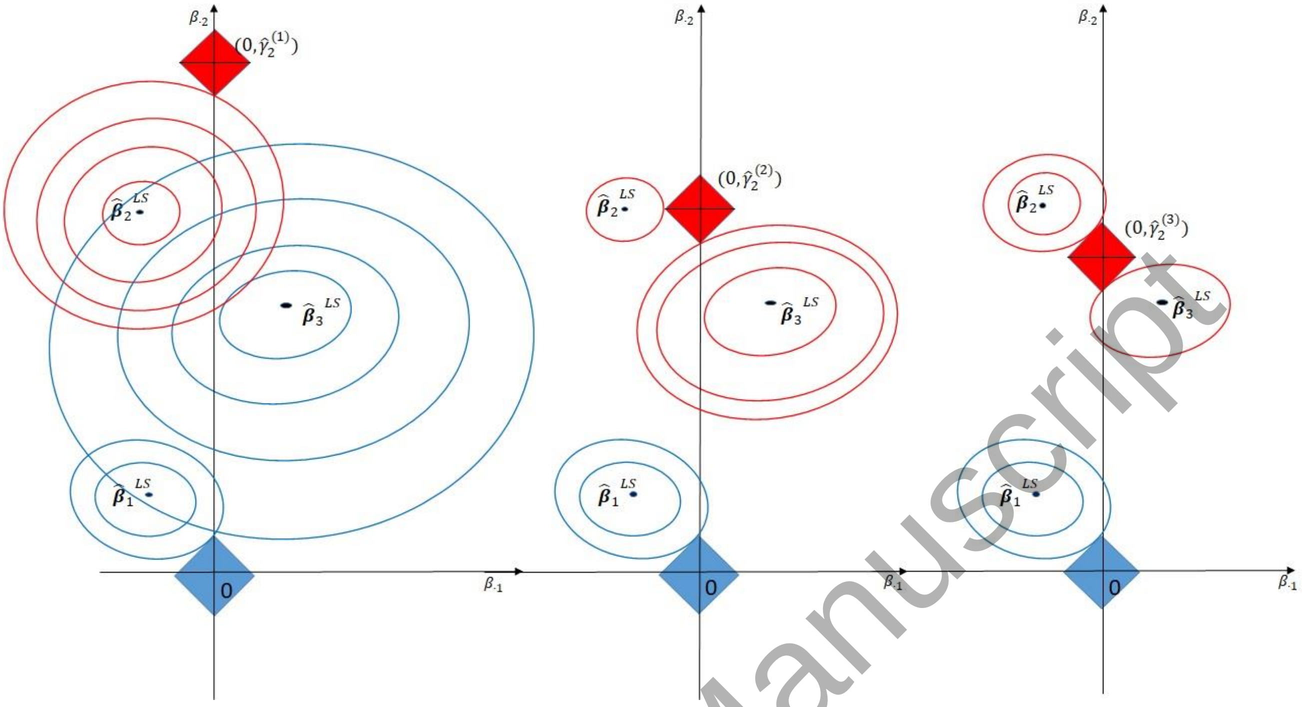


**Fig. 2** An illustration of the MDSP modeling with dynamic updating of $\gamma$ and adaptive individual estimations. A marginal MDSP is applied to the second covariate effect, $\beta_{i2}$, where the penalty function is $s(\boldsymbol{\beta},\gamma_2)=\sum_{i=1}^{3}\left\{|\beta_{i1}|+\min(|\beta_{i2}|,|\beta_{i2}-\gamma_2|)\right\}$, and $\boldsymbol{\beta}_i^{LS}$ denotes the least-squares estimator for $\boldsymbol{\beta}_i=(\beta_{i1},\beta_{i2})$.



$\beta_{\cdot 2}$
$(0,\hat{\gamma}_2^{(1)})$
$\hat{\boldsymbol{\beta}}_2^{LS}$
$\hat{\boldsymbol{\beta}}_3^{LS}$
$\hat{\boldsymbol{\beta}}_1^{LS}$
0
$\beta_{\cdot 1}$
$\beta_{\cdot 2}$
$(0,\hat{\gamma}_2^{(2)})$
$\hat{\boldsymbol{\beta}}_2^{LS}$
$\hat{\boldsymbol{\beta}}_3^{LS}$
$\hat{\boldsymbol{\beta}}_1^{LS}$
0
$\beta_{\cdot 1}$
$\beta_{\cdot 2}$
$(0,\hat{\gamma}_2^{(3)})$
$\hat{\boldsymbol{\beta}}_2^{LS}$
$\hat{\boldsymbol{\beta}}_3^{LS}$
$\hat{\boldsymbol{\beta}}_1^{LS}$
0
$\beta_{\cdot 1}$

**Fig. 3** The boxplot of RMSE of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size *N* = 40, 100, individual measurement size (cluster size) *m* = 10, 20, where homogeneous effect *γ* = 1.

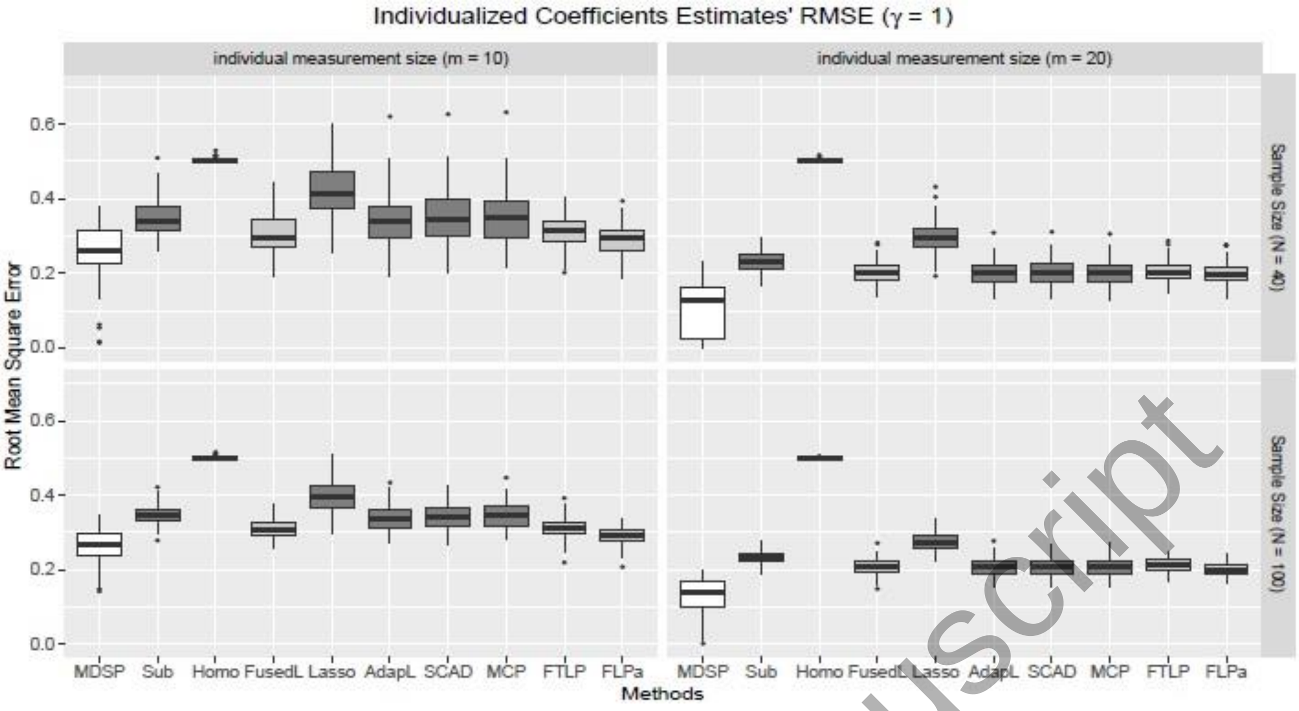


**Fig. 4** The boxplot of RMSE of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size *N* = 40, 100, individual measurement size (cluster size) *m* = 10, 20, where homogeneous effect *γ* = 2.

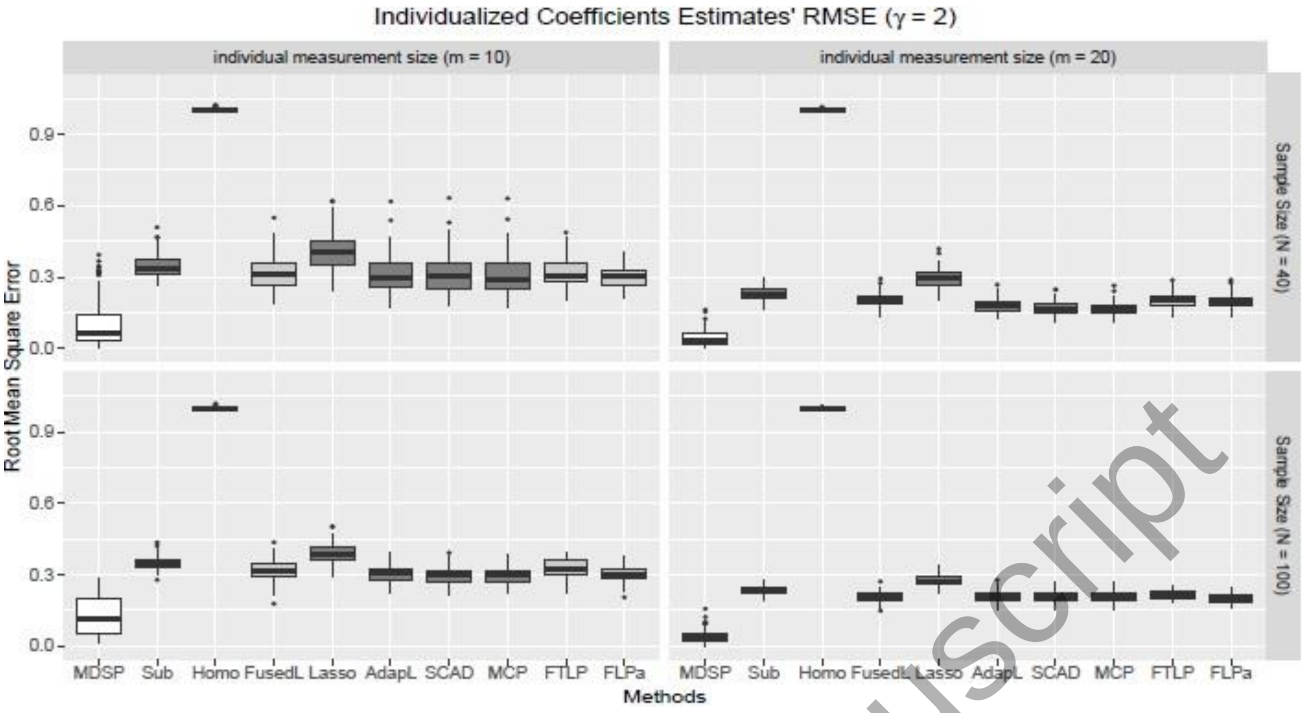


**Fig. 5** The boxplots of CVSR, sensitivity and specificity for all regularization approaches based on 100 simulations, with individual measurement size (cluster size) $m$ = 10, 20, where homogeneous effect $\gamma$ = 1 and sample size $N$ = 100.

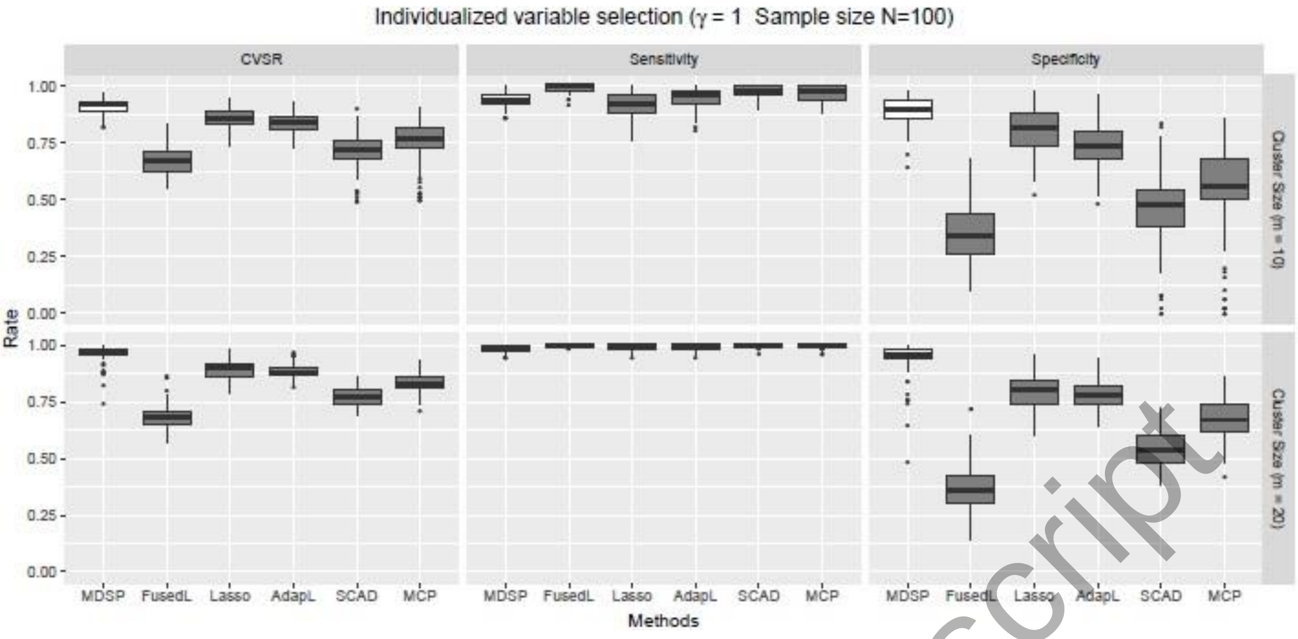


**Fig. 6** The boxplots of CVSR, sensitivity and specificity for all regularization approaches based on 100 simulations, with individual measurement size (cluster size) *m* = 10, 20, where homogeneous effect *γ* = 2 and sample size *N* = 100.

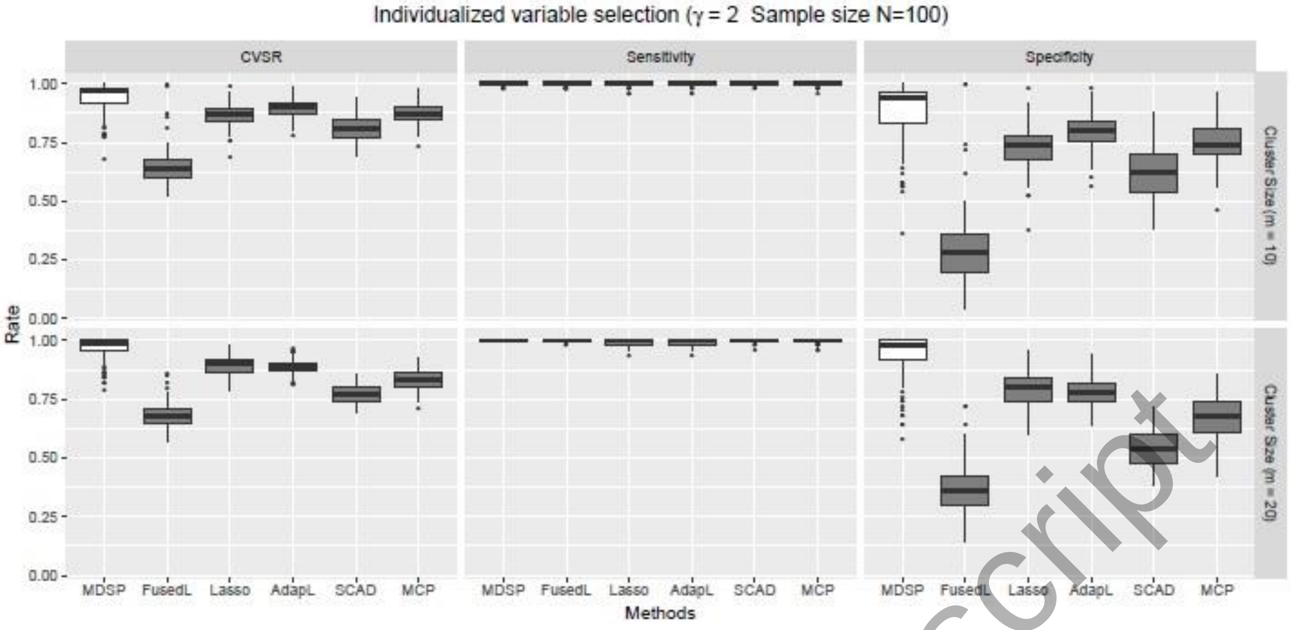


**Fig. 7** The individual-wise least squares estimator and the proposed estimator assuming two subgroups (including a zero group) for individualized parameters in two scenarios: a homogeneous group, and three subgroups, where the sample size *N* = 60 and individual measurement size *m* = 10.

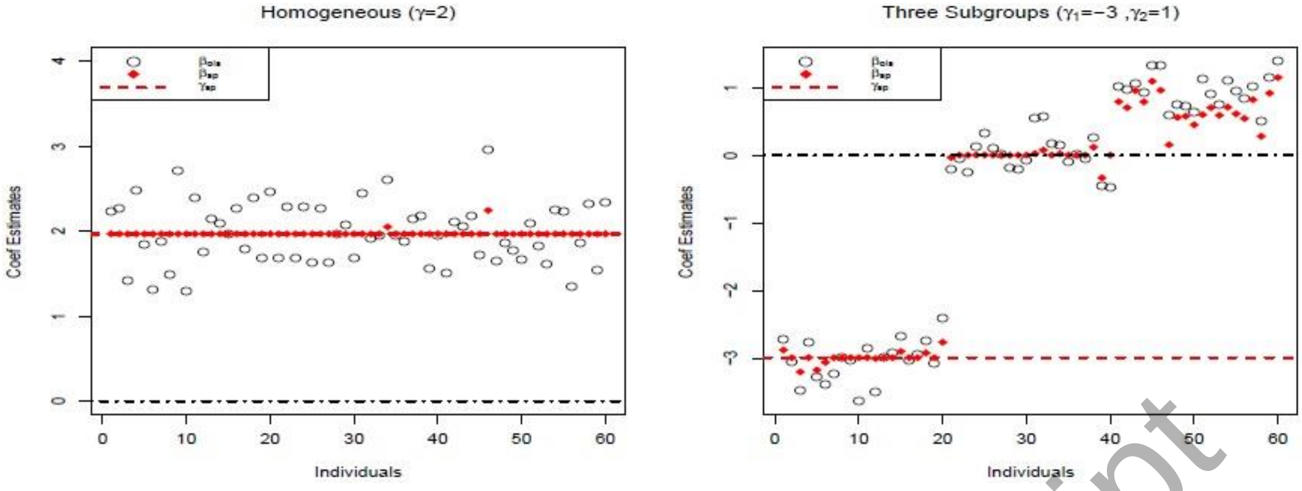


**Fig. 8** The left figure provides the average RMSE values of the coefficients estimations ($(\hat{\beta}_1, \hat{\beta}_2)$ for the MDSP model, the individual-wise OLS model, the individual-wise Lasso (L1) model and the homogeneous model estimated on the training set. The right two figures report the correct variable selection/elimination rates for $\beta_1$ and $\beta_2$, respectively. All results are evaluated based on 5 replications of $N^* = 100$ semi-new individuals over different numbers of individual measurements ranging from 6 to 20.

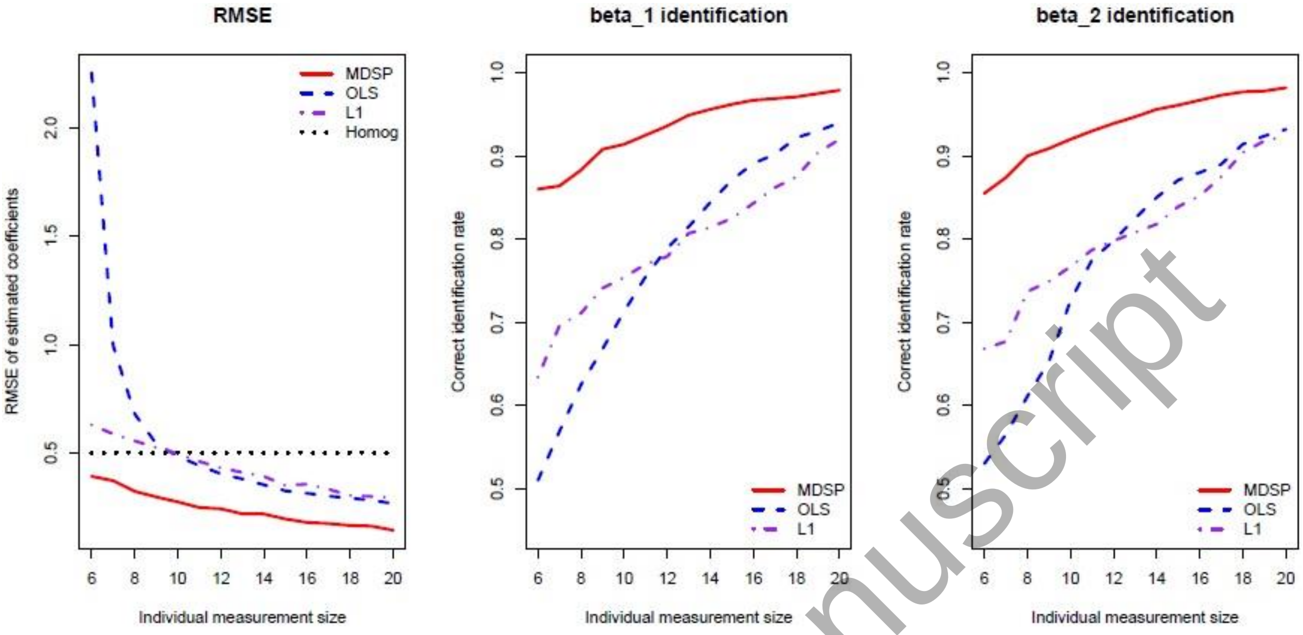


**Table 1** The average RMSE of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size (Sp size) $N$ = 40, 100, individual measurement size (Ind size) $m$ = 10, 20 where Sub, Homo, FusedL, Lasso, AdapL, SCAD and MCP stand for individual-wise model, homogeneous model, the fused Lasso, the Lasso, the adaptive Lasso, the SCAD and the MCP regularization models, respectively.

| Sp | Ind | Methods | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Size | Size | **MDSP** | Sub | Homo | FusedL | FLPa | FTLP | Lasso | AdapL | SCAD | MCP |
| $N$ | $m$ | $\gamma = 1$ | | | | | | | | | |
| 40 | 10 | **0.267** | 0.349 | 0.504 | 0.306 | 0.296 | 0.312 | 0.439 | 0.339 | 0.344 | 0.350 |
| | 20 | **0.120** | 0.232 | 0.502 | 0.206 | 0.196 | 0.207 | 0.298 | 0.207 | 0.201 | 0.201 |
| 100 | 10 | **0.262** | 0.350 | 0.501 | 0.319 | 0.290 | 0.311 | 0.394 | 0.334 | 0.335 | 0.345 |
| | 20 | **0.119** | 0.233 | 0.501 | 0.210 | 0.200 | 0.212 | 0.271 | 0.208 | 0.205 | 0.206 |
| $N$ | $m$ | $\gamma = 2$ | | | | | | | | | |
| 40 | 10 | **0.122** | 0.348 | 1.004 | 0.317 | 0.303 | 0.299 | 0.408 | 0.309 | 0.311 | 0.309 |
| | 20 | **0.048** | 0.230 | 1.002 | 0.204 | 0.197 | 0.196 | 0.293 | 0.181 | 0.168 | 0.167 |
| 100 | 10 | **0.113** | 0.351 | 1.001 | 0.318 | 0.301 | 0.315 | 0.387 | 0.305 | 0.300 | 0.299 |
| | 20 | **0.037** | 0.235 | 1.001 | 0.210 | 0.201 | 0.205 | 0.274 | 0.208 | 0.206 | 0.206 |
| | | | | | | | | | | | |

**Table 2** 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-individual serial correlation are AR-1 or exchangeable, and correlation parameter $\rho = 0.5$, sample size $N = 20, 80$, cluster size (individual measurement size) $m = 10, 20$.

| True | Cluster | $N = 20$ | | | $N = 80$ | | |
|---|---|---|---|---|---|---|---|
| Correlation | size (m) | $\boldsymbol{\beta}_{AR1}$ | $\boldsymbol{\beta}_{Ex}$ | $\boldsymbol{\beta}_{Ind}$ | $\boldsymbol{\beta}_{AR1}$ | $\boldsymbol{\beta}_{Ex}$ | $\boldsymbol{\beta}_{Ind}$ |
| 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 3** The average RMSE and CVSR of the proposed MDSP model compared to the individual-wise model (Sub), the fused Lasso (FusedL), the Lasso, the adaptive Lasso (Adapl), the SCAD and the MCP penalization models, with sample size $N = 60$ and cluster size (individual measurement size) $m = 10$. Scenario 1 contains a population homogeneous effect ($B_k = 1$) and Scenario 2 contains an individualized predictor of three subgroups ($B_k = 3$) with equal subgroup size. In both cases the MDSP model assumes two subgroups, where the estimated sub-homogeneous effects are $\hat{\gamma} = 2.01(0.06)$ and $\hat{\gamma} = -2.99(0.06)$ (with empirical standard errors in parenthesis), respectively.

| Scenario | | **MDSP** | Sub | FusedL | Lasso | AdapL | SCAD | MCP |
|---|---|---|---|---|---|---|---|---|
| $B_k = 1$ | 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 |
| | | | | | | | | |
| $B_k = 3$ | 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 4** The p-values of the estimated CpG coefficients in DNHS study from the homogeneous model, the refitted model within subgroups identified by the MDSP model ($\mathcal{G}^{(0)}$ and $\mathcal{G}^{(\gamma)}$), and by the mixture model (Comp 1 and Comp 2), and the prediction RMSE of PCL scores on testing set.

| | P-values of the coefficients | | | | |
|---|---|---|---|---|---|
| CpG sites | Homogeneous | MDSP | | MixReg | |
| | | $\mathcal{G}^{(0)}$ | $\mathcal{G}^{(\gamma)}$ (Proportion) | Comp1 | Comp2 |
| cg03256465 | 0.189 | 0.708 | **0.001** (36.5%) | 0.783 | 0.228 |
| cg03762702 | 0.396 | 0.468 | **0.029** (34.1%) | 0.189 | 0.223 |
| cg06473843 | 0.376 | 0.156 | **0.001** (40.4%) | 0.007 | 0.082 |
| Prediction RMSE | 0.385 | **0.292** | | 0.336 | |
| | | | | | |