HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2022 April 15.

Published in final edited form as:

Stat Med. 2021 April 15; 40(8): 1901–1916. doi:10.1002/sim.8878.

Capturing Heterogeneity in Repeated Measures Data by Fusion Penalty

Lili Liu[1], **Mae Gordon**[2], **J. Philip Miller**[3], **Michael Kass**[2], **Lu Lin**[4], **Shujie Ma**[5], **Lei Liu**[3]

[1]Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, China

[2]Department of Ophthalmology, Washington University in St. Louis, St. Louis, U.S.A

[3]Division of Biostatistics, Washington University in St. Louis, St. Louis, U.S.A.

[4]Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan, China

[5]Department of Statistics, University of California, Riverside, Riverside, U.S.A.

Summary

In this paper we are interested in capturing heterogeneity in clustered or longitudinal data. Traditionally such heterogeneity is modeled by either fixed effects or random effects. In fixed effects models, the degree of freedom for the heterogeneity equals the number of clusters/subjects minus 1, which could result in less efficiency. In random effects models, the heterogeneity across different clusters/subjects is described by e.g., a random intercept with 1 parameter (for the variance of the random intercept), which could lead to oversimplification and biases (for the estimates of subject-specific effects). Our "fused effects" model stands in between these two approaches: we assume that there are unknown number of distinct levels of heterogeneity, and use the fusion penalty approach for estimation and inference. We evaluate and compare the performance of our method to the fixed and random effects models by simulation studies. We apply our method to the Ocular Hypertension Treatment Study (OHTS) to capture the heterogeneity in the progression rate of primary open-angle glaucoma of left and right eyes of different subjects.

Keywords

Variable selection; high dimensional data; precision medicine; fusion penalty

1 | INTRODUCTION

Longitudinal or clustered data are commonly encountered in biomedical studies. For example, biomarkers are measured over time in longitudinal studies. The repeated measures of a biomarker on the same subject tend to be correlated. In clustered studies, health

*Correspondence Lei Liu, Division of Biostatistics, Washington University in St. Louis, St. Louis, U.S.A, lei.liu@wustl.edu.

SUPPLEMENTAL MATERIALS

The computer code is available at https://github.com/lililius/Fused-effect

outcomes of subjects within the same cluster (e.g., twins, families, or communities) are more alike due to shared genetic and/or environmental characteristics. In this paper, for ease of illustration, we will use the term “repeated measures” in a general sense to denote either the measures from multiple units within a cluster (repeated over space, e.g., left and right eyes of the same person), or those on the same marker across time (repeated over time, e.g., longitudinal measures of blood pressure of the same subject). The correlation of repeated measures from the same subject or cluster needs to be accounted for to yield more accurate and efficient estimates.

Two statistical models - fixed effects models and random effects models - are widely utilized to model repeated measures data[1,2]. However, both methods have their limitations. In fixed effects models, each subject has its own intercept, which leads to a large degree of freedom (df) for estimation, resulting in low efficiency in parameter estimates. On the other hand, random effects models use e.g., a random intercept, to capture the heterogeneity across different subjects to improve efficiency. However, it leads to shrinkage in the estimates of the heterogeneity, i.e., the values of random effects. Furthermore, the distributional assumption (e.g., normal) for the random intercept may be oversimplified in the presence of e.g., outliers, which might affect the bias and efficiency of the regression coefficient estimates.

To achieve an appropriate balance between accuracy and efficiency, we propose a new approach in between the fixed effects and random effects models. In our model, we assume that the heterogeneity for each subject belongs to different groups. By penalizing the fused effect (the difference between two subject-specific effects), we automatically group the subject-specific effects without knowing the group membership of the subjects in advance. We thus term our method as the “fused effects” model. Our model is along the lines of Ma and Huang[3], adapting their method to the repeated measures data. Computationally, we use an alternating direction method of multipliers algorithm (ADMM[4,5]) to implement the estimating procedure, which has been used for solving a large class of convex optimization problems. We use concave penalties on the pairwise differences of the parameters. Such penalties include the smoothly clipped absolute deviations penalty (SCAD[6]) and the minimax concave penalty (MCP[7]), which enjoy the consistency property.

Of note, related models have been considered in Wang and Zhu[8] and Wang et al.[9] in spatial areal data. Zhu and Qu[10]adapted the Ma and Huang’s method to the cluster analysis of longitudinal profiles. However, our work originates more naturally from the ordinary clustered data. We also investigate the performance of our method vs. the traditional random effects and fixed effects models, in particular when there exist outliers in data.

Our motivating example is the Ocular Hypertension Treatment Study (OHTS[11,12,13]). Ocular Hypertension (OH) is a common condition occurring in 3 to 8% of US population over age 40. People with OH have a higher risk of developing glaucoma, with the most common form being Primary Open Angle Glaucoma (POAG). In the first and second phases of OHTS, 1,636 participants were followed from 1994 to 2009. We would investigate the risk factors for the slope of Visual Field Measure (sVFM) - an indicator of POAG, among subjects having developed the POAG endpoint in at least one eye. The onset date of POAG is

determined by the first abnormal test (on the "first suspicious date") that is confirmed by at least 2 subsequent, consecutive abnormal tests. In this dataset, 67 subjects had both eyes, and 178 subjects had only one eye, reach the POAG endpoint. Our dataset includes a total of 312 eyes from these 245 subjects. The sVFMs of the clustered (paired) eyes are correlated within the same subject. We will apply our method to capture the heterogeneity between eyes across different subjects while investigating the risk factors of sVFM.

The rest of the paper is organized as follows. In Section 2, we describe our method. In Section 3 we assess the performance of our method via the Monte Carlo simulation studies. Section 4 illustrates the proposed method through the OHTS study. We summarize our method and present some future directions in Section 5.

2 | MODEL AND ESTIMATION

2.1 | Model

Let y_{ij} denote the jth response for subject i, and $\mathbf{x}_{ij}$ denote a $p \times 1$ vector of predictors, where $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$. We consider the linear model:

$$y_{ij} = a_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij}, \; i = 1, \cdots, m, j = 1, \cdots, n_i, \tag{2.1}$$

where a_i's are unknown subject-specific intercepts; $\boldsymbol{\beta} = (\beta_1 \cdots, \beta_p)^T$ is the vector of unknown covariate coefficients; $\epsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$ is the random error independent of $\mathbf{x}_{ij}$ and a_i.

We assume that a_i's have K distinct values, i.e., they belong to K groups $\mathcal{G}_1, \cdots, \mathcal{G}_K$ which are mutually exclusive partitions of subjects $\{1, \cdots, m\}$. However, the number of the groups K and the groups $\mathcal{G}_k$'s are unknown in advance. Thus, we aim to estimate K, identify the groups, and estimate the unknown parameters.

We further assume that the number of groups is much smaller than the number of subjects, i.e., $K \ll m$. Consider the following criterion

$$\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(y_{ij} - a_i - \mathbf{x}_{ij}^T \boldsymbol{\beta} \right)^2 + \sum_{1 \le i < k \le m} p_{\vartheta}(|a_i - a_k|, \lambda), \tag{2.2}$$

where $p_{\vartheta}(t, \lambda)$ is a given penalty function with the penalty parameter λ and a built-in constant ϑ. Note that the penalty is taken on all the pairwise differences in the subject-specific effects, i.e., $|a_i - a_k|$, the so-called "fusion penalty"[14,15]. Such a penalty promotes both similarity between elements and sparsity of the elements. Following Ma and Huang[3], since it is well known that the Lasso penalty leads to biased parameter estimates[6] and tends to produce too many groups, we consider some concave penalties including the SCAD penalty and the MCP penalty. Specifically, the SCAD penalty is defined as

$$p_{\vartheta}(t, \lambda) = \lambda \int_0^{|t|} \min\{1, \; (\vartheta - x/\lambda)_+ / (\vartheta - 1)\} dx,$$

and the MCP penalty is expressed as

$$p_{\vartheta}(t, \lambda) = \lambda \int_0^{|t|} (-x/(\vartheta\lambda))_+ dx,$$

where $(x)_+ = x$ if $x > 0$ and = 0 otherwise, and ϑ is a parameter that controls the concavity of the penalty function.

2.2 | Estimation procedure

Many machine learning and statistical problems can be formulated as linearly constrained convex programs, which can be efficiently solved by the alternating direction method of multipliers (ADMM). It takes the form of a decomposition-coordination procedure, in which the solutions to small local subproblems are coordinated to find a solution to a large global problem. ADMM blends the benefits of dual decomposition and augmented Lagrangian methods for constrained optimization.

The penalty function $p_{\vartheta}(|a_i - a_k|, \lambda)$ is not separable between a_i and a_k, i.e., it cannot be written in the form of addition of separate terms of $p_{\vartheta}(|a_i|, \lambda)$ and $p_{\vartheta}(|a_k|, \lambda)$ as in LASSO. As a result, it is difficult to compute the estimates directly by minimizing objective function (2.2) through the commonly used coordinate descent algorithm. A new parameter $\theta_{ik} = a_i - a_k$ is introduced and an ADMM algorithm is used to identify the groups in objective function (2.2). Thus, the minimization problem in (2.2) becomes the constraint optimization problem,

$$L_0(\mathbf{a}, \beta, \theta) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(y_{ij} - a_i - \mathbf{x}_{ij}^T \beta\right)^2 + \sum_{i<k} p_{\vartheta}(|\theta_{ik}|, \lambda) \quad \text{subject to } a_i - a_k - \theta_{ik} = 0,$$

where $\boldsymbol{\theta} = \{\theta_{ik}, i < k\}^T$ and $\mathbf{a} = (a_1, \cdots, a_m)^T$. The estimators of the parameters are yielded by the augmented Lagrangian

$$L(\mathbf{a}, \beta, \theta, \upsilon) = L_0(\mathbf{a}, \beta, \theta) + \sum_{i<k} \upsilon_{ik}(\theta_{ik} - a_i + a_k) + \frac{\eta}{2} \sum_{i<k} (\theta_{ik} - a_i + a_k)^2,$$

where $\boldsymbol{\upsilon} = \{\upsilon_{ik}, i < k\}^T$ are Lagrange multipliers, η is the penalty parameter. We use the ADMM to iteratively compute the estimators of $(\mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\upsilon})$. For given $\boldsymbol{\theta}^{(l)}$, $\boldsymbol{\upsilon}^{(l)}$ at step l, we use the following algorithm

$$\left(\mathbf{a}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}\right) = \underset{\mathbf{a}, \boldsymbol{\beta}}{\operatorname{argmin}} \, L\left(\mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\theta}^{(l)}, \boldsymbol{\upsilon}^{(l)}\right), \tag{2.3}$$

$$\boldsymbol{\theta}^{(l+1)} = \underset{\theta}{\operatorname{argmin}} \, L\left(\mathbf{a}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\theta}, \boldsymbol{\upsilon}^{(l)}\right), \tag{2.4}$$

$$\upsilon_{ik}^{(l+1)} = \upsilon_{ik}^{(l)} + \eta\left(a_i^{(l+1)} - a_k^{(l+1)} - \theta_{ik}^{(l+1)}\right). \tag{2.5}$$

To update $\mathbf{a}$, minimization problem (2.3) is equivalent to minimizing

$$F(\mathbf{a}, \boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij} - a_i - \mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^2 + \frac{\eta}{2}\sum_{i<k}\left(a_i - a_k - \theta_{ik}^{(l)} + \eta^{-1}\upsilon_{ik}^{(l)}\right)^2 + C, \tag{2.6}$$

where C is a constant independent of $\mathbf{a}$ and $\boldsymbol{\beta}$. Let $\mathbf{y} = \left(\mathbf{y}_1^T, \cdots, \mathbf{y}_m^T\right)^T$ with $\mathbf{y}_i = \left(y_{i1}, \cdots, y_{in_i}\right)^T$, and $\mathbf{X}_i = \left(\mathbf{x}_{i1}, \cdots, \mathbf{x}_{in_i}\right)^T$, some algebra shows that Equation (2.6) can be rewritten as

$$F(\mathbf{a}, \boldsymbol{\beta}) = \frac{1}{2}\left\|\mathbf{y} - \mathbf{Z}\mathbf{a} - \mathbf{X}\boldsymbol{\beta}\right\|^2 + \frac{\eta}{2}\left\|\boldsymbol{\Delta}\mathbf{a} - \theta^{(l)} + \eta^{-1}\boldsymbol{\upsilon}^{(l)}\right\|^2 + C.$$

where $Z = \begin{pmatrix} 1_{n1} & & \\ & \ddots & \\ & & 1_{nm} \end{pmatrix}_{N \times m}$ with $\mathbf{1}_{n_i} = (1, \cdots, 1)^T$ being the vector with n_I ones, $\mathbf{X} = \left(\mathbf{X}_1^T, \cdots, \mathbf{X}_m^T\right)^T$, and $= \{(\mathbf{e}_I - \mathbf{e}_j),\ i < j)\}^T$ with $\mathbf{e}_I$ being the ith unit $m \times 1$ vector whose ith element is 1 and the remaining elements are 0.

For given $\boldsymbol{\theta}^{(l)}$, $\boldsymbol{\upsilon}^{(l)}$ at the lth step, we set the derivatives $F(\mathbf{a}, \boldsymbol{\beta})/\ \mathbf{a} = 0$ and $F(\mathbf{a}, \boldsymbol{\beta})/\ \boldsymbol{\beta} = 0$ to obtain the following updates $\mathbf{a}^{(l+1)}$ and $\boldsymbol{\beta}^{(l+1)}$:

$$\mathbf{a}^{(l+1)} = \left(\mathbf{Z}^T\mathbf{Q}_x\mathbf{Z} + \eta\boldsymbol{\Delta}^T\boldsymbol{\Delta}\right)^{-1}\left[\mathbf{Z}^T\mathbf{Q}_x\mathbf{y} + \eta\boldsymbol{\Delta}^T\left(\theta^{(l)} - \eta^{-1}\boldsymbol{\upsilon}^{(l)}\right)\right], \tag{2.7}$$

where $\mathbf{Q}_x = \mathbf{I}_N - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ with $N = \sum_{i=1}^{m} n_i$, and

$$\boldsymbol{\beta}^{(l+1)} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{y} - \mathbf{Z}\mathbf{a}^{(l+1)}\right). \tag{2.8}$$

To update $\boldsymbol{\theta}$, we need to minimize the function

$$\frac{\eta}{2}\left(\theta_{ik} - \pi_{ik}^{(l)}\right)^2 + \sum_{i<k} p_{\vartheta}(|\theta_{ik}|, \lambda),$$

where $\pi_{ik}^{(l)} = a_i^{(l)} - a_k^{(l)} + \eta^{-1}\upsilon_{ik}^{(l)}$. It is worth noting that by using the concave penalties, the objective function $L(\mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\upsilon})$ is not a convex function, but it is convex with respect to each θ_{ik} when $\vartheta > 1/\eta + 1$ for the SCAD penalty and $\vartheta > 1/\eta$ for the MCP penalty. Moreover, for given $(\mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\upsilon})$, the minimizer of $L(\mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\upsilon})$ with respect to θ_{ik} is unique with a closed-form expression. Thus, for the MCP penalty with $\vartheta > 1/\eta$, the update $\theta_{ik}^{(l+1)}$ is

$$\theta_{ik}^{(l+1)} = \begin{cases} \dfrac{S\left(\pi_{ik}^{(l)}, \lambda/\eta\right)}{1 - 1/(\vartheta\eta)} & \left|\pi_{ik}^{(l)}\right| \le \lambda\vartheta, \\ \pi_{ik}^{(l)} & \left|\pi_{ik}^{(l)}\right| > \lambda\vartheta. \end{cases} \tag{2.9}$$

For the SCAD penalty with $\vartheta > 1/\eta + 1$, the solution is

$$\theta_{ik}^{(l+1)} = \begin{cases} S\left(\pi_{ik}^{(l)}, \lambda/\eta\right) & \left|\pi_{ik}^{(l)}\right| \le \lambda + \lambda/\eta, \\ \dfrac{S\left(\pi_{ik}^{(l)}, \vartheta\lambda/((\vartheta-1)\eta)\right)}{1 - 1/((\vartheta-1)\eta)} & \lambda + \lambda/\eta < \left|\pi_{ik}^{(l)}\right| \le \lambda\vartheta, \\ \pi_{ik}^{(l)} & \left|\pi_{ik}^{(l)}\right| > \lambda\vartheta, \end{cases} \tag{2.10}$$

where $S(x, t) = (1 - t/|x|)_+ x$ is a groupwise soft thresholding operator.

Following Ma and Huang[3], we fix $\vartheta = 3$ and $\eta = 1$ for both MCP and SCAD penalties in the simulation and application studies, which satisfies the conditions of (2.9) and (2.10).

Finally, the Lagrange multiplier υ_{ik} is updated by (2.5). It is worth noting that subjects i and k are classified into the same group if $\hat{\theta}_{ik} = 0$. After the group $\mathcal{G}_k$ is identified, we obtain the estimated number of groups $\hat{K}$, the estimated groups $\widehat{\mathcal{G}_1}, \cdots, \widehat{\mathcal{G}_{\hat{K}}}$, and the estimated common value for $\hat{a}_i$'s from group $\widehat{\mathcal{G}}_k$: $\hat{\alpha}_k = \left|\widehat{\mathcal{G}}_k\right|^{-1} \sum_{i \in \widehat{\mathcal{G}}_k} \hat{a}_i$, where $\left|\widehat{\mathcal{G}}_k\right|$ is the cardinality of $\widehat{\mathcal{G}}_k$.

We apply the modified Bayesian Information Criterion (BIC)[16] to select the tuning parameter λ, which is defined as the value that minimizes

$$\mathrm{BIC}(\lambda) = \log\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij} - a_i - \mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^2/N\right] + C_N(\hat{K}(\lambda) + p)\frac{\log N}{N},$$

where C_N is a positive number depending on the total number of observations $N = \sum_{i=1}^{m} n_i$. When $C_N = 1$, it corresponds to the traditional BIC. $\hat{K}(\lambda)$ is the estimated number of groups based on the tuning parameter λ, and p is the dimension of the parameter $\boldsymbol{\beta}$. Following Wang et al[17] and Ma and Huang[3], it is chosen as $C_N = c\log(\log(N + p))$ with $c = 5$. The tuning parameter λ is selected by minimizing the modified BIC with a grid search.

When there exists true group structure and $K \ll m$, we can calculate the standard errors of both α and $\boldsymbol{\beta}$ in a traditional regression model setting after we identify the group structure. However, when there is no group structure, e.g., in Simulation Setting 3 when the random effects model is correct, this approach yields less reliable results. Therefore, we rely on the bootstrap estimates of the standard errors.[18]. By our previous experiences, Bootstrap sampling can yield reasonable coverage probabilities in variable selection of sophisticated random effects models, e.g., Han et al[19,20]. Efron and Tibshirani[18] showed that 50–100 bootstrap replications are generally sufficient for standard error estimation. In this paper, we will take 100 samples of cluster bootstrap data for each dataset[21], and estimate the standard

errors of α and β by the empirical standard deviations of the bootstrap replications. Our simulation studies show that such estimation is satisfactory.

2.3 | Algorithm

It is important to find appropriate initial values for the ADMM algorithm. In this paper, the initial values $\mathbf{a}^{(0)}$ are obtained from the best linear unbiased predictors (BLUPs) of the random effects model. We can then set $\theta_{ik}^{(0)} = a_i^{(0)} - a_k^{(0)}$ and $\boldsymbol{v}^{(0)} = \mathbf{0}$. The convergence of the ADMM algorithm is evaluated based on the primal residual $\mathbf{r}^{(l+1)} = \mathbf{a}^{(l+1)} - \boldsymbol{\theta}^{(l+1)}$. The algorithm terminates when $\mathbf{r}^{(l+1)}$ is close to zero, i.e., $\|\mathbf{r}^{(l+1)}\| < \epsilon$ for some small value ϵ. If $\hat{\theta}_{ik} = 0$ for some λ, then a_i and a_j belong to the same group. As a result, we obtain $\hat{K}$ estimated groups $\hat{G}_1, \cdots, \hat{G}_k$. The subject-specific intercept for the kth group is estimated as $\hat{\alpha}_k = \left|\hat{G}_k\right|^{-1} \sum_{i \in \hat{G}_k} \hat{a}_i$, where $\left|\hat{G}_k\right|$ is the cardinality of $\hat{G}_k$.

The algorithm consists of the following steps:

Algorithm 1

ADMM for concave penalty

Require: Initialize $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{v}^{(0)}$
  for $l = 0, 1, 2, \cdots$ **do**
    Compute $\mathbf{a}^{(l+1)}$ using (2.7)
    Compute $\boldsymbol{\beta}^{(l+1)}$ using (2.8)
    Compute $\boldsymbol{\theta}^{(l+1)}$ using (2.9) or (2.10)
    Compute $\boldsymbol{v}^{(l+1)}$ using (2.5)
    if the convergence criterion is met, **then**
      Stop and denote the last iteration by $\hat{\mathbf{a}}$
    else
      $l = l + 1$
    end if
  end for
Ensure: Output

3 | SIMULATION

In this section, we will examine the finite sample behavior of our method by simulation studies. Traditionally, in Model (2.1), the intercept a_i is either taken as a random effect (RE), often assumed to be independent and identically distributed as $N(0, \sigma^2)$; or a_i is treated as a fixed effect (FE), which is a fixed non-random quantity for each i. We use the best linear unbiased predictor (BLUP) of random effects implemented in the function *lme* of the R package *nlme*. We compare the performance of our estimators and the RE and FE approaches. Three different sample sizes $m = 50$, 100 and 200 are considered, and all the simulation results are obtained via 100 replicates.

Setting 1.

We generate data from the following linear model:

$$y_{ij} = a_i + x_{ij}\beta + \epsilon_{ij}, \quad i = 1, \cdots, m, j = 1, \cdots, n_i, \tag{3.1}$$

where the covariates x_{ij}'s are sampled from the normal distribution $N(0, 1)$; the error terms ϵ_{ij}'s are independent and identically distributed as $N(0, 0.4^2)$. Let $\beta = 2$. We randomly assign a_i to one of the three groups $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ with equal probabilities 1/3, so that $a_i = -1.5$ for $i \in \mathcal{G}_1$, $a_i = 0$ for $i \in \mathcal{G}_2$, $a_i = 1.5$ for $i \in \mathcal{G}_3$. We consider unbalanced data, where for each i there may be some y-values missing. The number of observation of each subject n_i is generated from the distribution: $P(n_i = 1) = 0.5$, $P(n_i = 2) = 0.5$. Thus the number of observations in $\mathbf{y}_i$ varies for different subjects.

To compare the estimated partitions to the true partitions, we use the Rand Index[22] to evaluate the accuracy of the clustering results. Each pair of observations a_i and a_j can be fit to one of four categories: a true positive (TP): a_i and a_j from the same group are assigned to the same cluster; true negative (TN): a_i and a_j from different groups are assigned to different clusters; false negative (FN): a_i and a_j from different groups are assigned to the same cluster; false positive (FP): a_i and a_j from the same group are assigned to different clusters. Thus, the Rand Index is given by

$$\mathrm{RI} = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}} = \frac{\mathrm{TP+TN}}{\binom{\mathrm{N}}{2}}.$$

Intuitively, TP and TN indicate agreement between the true group and the estimated cluster, while FP and FN indicate disagreement between the true group and the estimated cluster. The Rand index has a range of [0,1]: higher values indicate better performance of the clustering methods.

We minimize the modified BIC to select the tuning parameter λ. The top panel of Table 1 reports the mean, median, and standard deviation (sd) of the estimated number of groups $\widehat{K}$, the percentage of $\widehat{K}$ equal to the true number of groups (per), and the Rand Index (RI) for measuring clustering accuracy. It can be seen that the medians of $\widehat{K}$ are equal to 3 - the true number of groups for all cases, indicating that our method can correctly identify the groups. The Rand Index values and the percentages of correctly identifying the groups are close to 1, implying the clustering accuracy. Moreover, as the subject size m increases, the standard deviation becomes smaller and the mean becomes closer to the true value 3; the Rand Index and the percentage of correctly selecting the number of subgroups get closer to 1.

To assess the estimators $\hat{\mathbf{a}} = (\hat{a}_1, \cdots, \hat{a}_m)^T$, we calculate the square root of the mean squared error (SRMSE) for the estimator $\hat{\mathbf{a}}$ by using the formula $\|\hat{\mathbf{a}} - \mathbf{a}\|/\sqrt{m}$ for each replicated dataset. In the top panel of Table 2 we present the mean SRMSE for $\hat{\mathbf{a}}$. The Oracle estimator is obtained with *a priori* knowledge of the true grouping information. For SCAD and MCP, we can see that the SRMSEs of $\hat{\mathbf{a}}$ are smaller than those from the random effects and fixed

effects models. To graphically depict the numerical results of Table 2, the boxplots of the SRMSEs for $\hat{\mathbf{a}}$ with $m = 200$ are presented in Figure 1.

We also present the bias, the empirical standard error (SEE), the sampling mean of the standard error estimate (SEM), and the coverage probability of the 95% confidence intervals of the estimator $\hat{\beta}$ in the right of Table 2. For the estimator $\hat{\beta}$, we observe that the biases, SEEs, and SEMs by our method are close to those of the Oracle. In contrast, the fixed effects and random effects models yield larger SEEs and SEMs. All the performance measures generally improve with increased subject sizes.

Let $\hat{\alpha}_k$ be the common value for the estimator $\hat{a}_i$'s from group $\widehat{\mathcal{G}}_k$ with $k = 1, 2, 3$. Table 3 presents the mean, the empirical standard error (SEE), the sampling mean of the standard error estimate (SEM), and the coverage probability of the 95% confidence intervals (CP) of the estimators $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$ by the SCAD and MCP methods, which are calculated based on replicates with the estimated number of groups equal to three. The Oracle estimators are calculated based on all 100 replicates. We observe that the SCAD and MCP methods perform very close to the Oracle. Clearly, our estimators $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$ agree well with the corresponding true values on average for all cases. There is good agreement between SEE and SEM values, and the coverage probabilities are acceptably close to the nominal level 0.95.

Setting 2.

In this setting, we generate data from a linear model given by:

$$y_{ij} = a_i + x_{ij}^T \beta + \epsilon_{ij}, \quad i = 1, \cdots, m, j = 1, \cdots, n_i, \tag{3.2}$$

where x_{ij}, β, n_i, and ϵ_{ij} are generated from the same distributions as given in Setting 1. $m - 1$ of all the intercept a_i is divided into three groups as in the Setting 1. Mimicking the OHTS data, we also add a fourth group with only one subject: an outlier at −10. Thus, the true group number is 4.

From the second panel of Table 1, we can see that the medians of $\widehat{K}$ over the 100 replicates are 4, the true number of subgroups, and the mean values are very close to 4 for both the MCP and SCAD methods. Moreover, the standard deviation becomes smaller and the mean gets closer to the true value of 4, and the percentage of correctly selecting the number of subgroups increases with the sample size m.

From the second panel of Table 2, we observe that the SRMSE values of $\hat{\mathbf{a}}$ by SCAD and MCP are smaller than those of the random effects and fixed effects models. Moreover, the SRMSE decreases as m increases for both MCP and SCAD. For the estimators $\hat{\beta}$, the MCP and SCAD methods perform better than the random effects and fixed effects models in terms of smaller biases and smaller SEEs. Also, our estimates of β have CPs close to the nominal level. These results indicate that the proposed method can obtain relatively robust results even with an outlier in the data. The boxplots of the SRMSEs of $\hat{\mathbf{a}}$ with $m = 200$ are shown in Figure 1.

From Table 4, it can be seen that the means of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$ are close to the true values and the Oracle estimators. We also observe that the SEMs of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$ are close to the corresponding SEEs, leading to valid coverage probabilities. To ensure that α_4 can be estimated, we always include a single subject from the fourth group in all bootstrap samples. For $\hat{\alpha}_4$, the biases are small but the SEMs are substantially below the SEEs, leading to poor coverage probabilities. Of note, this also happens for the Oracle model when we know the group membership of each subject.

Setting 3.

In this setting, we generate data from a linear model given by:

$$y_{ij} = a_i + x_{ij}^T \beta + \epsilon_{ij}, \quad i = 1, \cdots, m, j = 1, \cdots, n_i, \tag{3.3}$$

where x_{ij}, β, n_i, and ϵ_{ij} are generated from the same distributions as given in Setting 1. To show the robustness of our method, a_i is simulated from the normal distribution $N(0, 0.5^2)$, therefore, the random effects model is the correct model.

The grouping results of $\hat{a}_i$ by the MCP and SCAD methods are presented in the third panel of Table 1. For SCAD and MCP, the median of the estimated number of groups $\widehat{K}$ is 4 with $m = 50$; 5 with $m = 100$; and 6 with $m = 200$. It can be seen that the fusion penalty tends to select less groups for $m = 50$ and more groups for $m = 100,200$ in general. Since the number of parameters grows with sample size m, the median and the standard deviation of $\widehat{K}$ increase as well.

From the third panel of Table 2, we observe that the SRMSE values of $\hat{\mathbf{a}}$ by SCAD and MCP perform just slightly worse than the random effects model. Moreover, the SRMSE value decreases as m increases for both MCP and SCAD. The boxplots of the SRMSEs of $\hat{\mathbf{a}}$ with $m = 200$ are shown in Figure 1. For the estimators $\hat{\beta}$, the standard deviations from MCP and SCAD are smaller than the fixed effects model, and slightly larger than the random effects models (as sample size increases). Also, our estimates of β have CPs close to the nominal level.

Setting 4.

In this setting, we consider a linear model with relatively large number of repeated measures as suggested from a reviewer:

$$y_{ij} = a_i + x_{ij}^T \beta + \epsilon_{ij}, \quad i = 1, \cdots, m, j = 1, \cdots, n_i, \tag{3.4}$$

where x_{ij}, β, a_i and ϵ_{ij} are generated from the same distributions as given in Setting 1. To show the performance of our method with a relatively large number of repeated measures, n_i is set to 10.

From the bottom panel of Table 1, it can be seen that the mean and median of $\widehat{K}$ are the same as the true value 3 for all cases. Moreover, the Rand Index (RI) values and the percentages of

correctly selecting the number of subgroups reach 1. Therefore, if the number of repeated measures is relatively large, our method can recover the true group structure very well.

The bottom panel of Table 2 shows that, when the number of repeated measures is large enough, the performance of our method is equal to that of Oracle model. Moreover, the SRMSE values from our method are much smaller than those from the RE and FE methods. In this setting, our method is superior to the competing models, and there is not much difference between the FE model and RE model.

From Table 5, since SCAD and MCP can correctly recover the subgroup structure, the means of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$ are close to the corresponding true values, and equal to the Oracle estimators.

In summary, the fused effects approach performs well if the clusters are well separated and enough observations are available. Nevertheless, even if the true individual-specific effects are a sample from a Gaussian population (when the normal random effects model is true), our method's performance is still satisfactory.

4 | APPLICATIONS

In this section, we apply our method to investigate the risk factors for the slope of the visual field measures (sVFM) after POAG conversion in the OHTS study. Out of 1,636 participants, we include 245 individuals who had developed POAG in at least one eye in this analysis. Among these 245 subjects, 67 had both eyes develop POAG, and 178 subjects had only one eye reach the POAG endpoint. We thus have clustered (paired) sVFM from both eyes for 67 of these subjects. We will apply our method to capture heterogeneity across different subjects while investigating the risk factors of sVFM.

We consider the following baseline risk factors in the model: x_1: the patient's randomization assignment (RA, 1=Medication, 0=Observation); x_2: vertical cup-to-disc ratio (VCD); x_3: gender; x_4: stroke; x_5: race; x_6: age; x_7: central corneal thickness (CCT); x_8: intraocular pressure (IOP); x_9: visual field pattern standard deviation (PSD).

We use the sVFM as the response y_{ij}. The following linear model is considered:

$$y_{ij} = a_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij}, \quad i = 1, \cdots, 245, \ j = 1 \text{ or } 2, \tag{4.1}$$

where $\mathbf{x}_{ij}$ is a 9-dimensional covariate vector, and predictors are centered and standardized except for the binary variables before applying the regularization method. $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_9)^T$ is the unknown parameter; $j = 1, 2$ indicates eye (we do not distinguish left and right eyes in this study).

We note that some of the covariates considered, including the patient's randomization assignment, gender, stroke, race, age, are the same for both eyes. Therefore, they are confounded with the fixed effects, making the fixed effects model not identifiable for this dataset. We thus only compare the performance of our method to that of the random effects model.

We use the Akaike Information Criteria (AIC, smaller is better) to assess the performance of each method:$\mathrm{AIC} = N \log\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij} - a_i - \mathbf{x}_{ij}^T\boldsymbol{\beta}\right)^2/N\right] + 2(k+p)$, where $N = \sum_{i=1}^{m} n_i = 312$ is the total number of eyes and $p = 9$ is the dimension of the parameter $\boldsymbol{\beta}$. If a_i is treated as a random intercept to capture the correlation between the 2 eyes, we obtain AIC=−670.86 with $k = 1$. If a_i is estimated by our concave pairwise fused approach, the 245 subjects are classified into 12 groups by SCAD and 13 groups by MCP, respectively. For SCAD, the AIC value is −799.63 with $k = 12$, and for MCP, the AIC value is −808.55 with $k = 13$. We see that our method leads to a notable improvement of the model fitting.

Let $\hat{\alpha}_k$ be the common value for the estimator $\hat{a}_i$'s from group $\widehat{\mathcal{G}}_k$ with $k = 1,\cdots, 13$. Table 6 reports the estimators $\hat{\alpha}_k$ (Est.) and the number of elements (num.) in each subgroup of $\hat{a}_i$ by the SCAD and MCP methods. Both methods yield almost the same grouping results. The estimator $\hat{\alpha}_1$ deviates greatly from the other estimators, which is considered as an outlier. In Table 7 we report the estimate (Est.), standard error (s.e.), and p-value of $\hat{\beta}$ for testing the significance of the coefficients by the MCP, SCAD, random effects model, and random effects model without the outlier, respectively. The standard errors for the MCP and SCAD methods are calculated by cluster bootstrap. We can see that Age and PSD have p-values less than 0.05 by SCAD and MCP. When the dataset contains the large outlier, race, age, and PSD have statistically significant effects by the random effects model. However, only age and PSD are statistically significant by the random effects model when the outlier is removed, which is consistent with the result of our methods.

Figure 2 displays the histograms of the estimator $\hat{\mathbf{a}} = (\hat{a}_1, \cdots, \hat{a}_{245})^T$ and the kernel density plots of the residuals $y_{ij} - \hat{a}_i - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}$ by SCAD, MCP and the random effects model (RE), respectively. We can see that the distribution of the estimator $\hat{\mathbf{a}}$ by our method deviates from the normal distribution, especially with a large outlier at the far left end. Even after removing this outlier, the remaining plot is left skewed. This may be the reason to the poor performance of the random effects model which assumes a_i to follow a normal distribution. For the kernel density plot of the residuals, it can be seen that the distributions by our method appear to be normal and more smooth than that by the RE model.

5 | DISCUSSION

In this paper we proposed a "fused effects" model by applying fusion penalties to heterogeneity in repeated measures. Our approach stands in between the traditional fixed effects and random effects models. Compared to the fixed effects model, our method is more efficient in using fewer parameters to capture the heterogeneity. Compared to the random effects model, our method is more flexible, e.g., when the common normality assumption does not hold for random effects as in the Application study. By extensive simulation studies, we showed that our method has satisfactory results in a variety of settings.

In principle, our approach is similar to a latent class (finite mixture) model with an unknown number of classes for the "fused effects". However, that method usually needs to fit models with different numbers of latent groups for a_i, and then choose the best one by comparing

these models through e.g., BIC. Our own simulation experience shows that the estimation of such latent class models with a large number of groups (e.g., > 5) is often subject to convergence issues. Our method avoids such issues and it is straightforward in implementation and interpretation.

Our method can be applied to many different research areas, e.g., provider profiling of health care facilities[23]. It can be used when there exist heterogeneities for different factors, e.g., treatment, that is, the same treatment can have different effects on different patients[24,25,26,27,28].

Our method can be extended in several directions. First, we can used other loss function in Equation (2.2), for example, weighted least square loss when the outcome (e.g., the slope of VFM in the Application Study) has unequal variance. Second, it is natural to extend this method to other types of repeated measures outcomes, e.g., binary outcomes (usually fitted by generalized linear mixed models) and survival outcomes (usually fitted by frailty Cox proportional hazards models). Finally, extensions to more sophisticated hierarchical data (e.g., longitudinal biomarkers of subjects clustered within families) form another topic for future research.

ACKNOWLEDGEMENTS

This research is partly supported by NIH grants R21 EY031884, UG1 EY025180, UG1 EY025182, UL1 TR002345, and by the China Scholarship Council (201806220145). The data that support the findings of this study are available from the OHTS Coordinating Center. Restrictions apply to the availability of these data, which were used under license for this study. Data request can be made at https://ohts.wustl.edu/ with the permission of the OHTS Coordinating Center.

References

1. Laird N, Ware J. Random-effects models for longitudinal data. Biometrics. 1982;38:963–974. [PubMed: 7168798]
2. Diggle P, Heagert P, Liang K, Zeger S. Anal Longitudinal Da. 2nd ed. Oxford: Oxford University Press; 2002.
3. Ma S, Huang J. A concave pairwise fusion approach to subgroup analysis. J Am Stat Assoc. 2017;112:410–423.
4. Boyd S, Parikh N, Chu E, Peleato B, and Eckstein JDistributed optimization and statistical learning via the alternating direction method of multipliers. Foundations Trends Machine L. 2011;3:1–122.
5. Chi EC and Lange K Splitting methods for convex clustering. J Comput Graph Stat. 2011;24:994–1013.
6. Fan J, Li R,. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96:1348–1360.
7. Zhang CNearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010;38:894–942.
8. Wang X, Zhu Z, and Zhang HHSpatial automatic subgroup analysis for areal data with repeated measures. 2019; arXiv:1906.01853.
9. Wang X and Zhu Z Small area estimation with subgroup analysis. Stat Theory Related FIeld. 2019;3:129–135.
10. Zhu X, and Qu ACluster analysis of longitudinal profiles with subgroups. Electron J Stat. 2018;12:171–193.

11. Gordon MO, Kass MA, for the Ocular Hypertension Treatment Study Group. The ocular hypertension treatment study: design and baseline description of the participants. Arch Ophthalmol-Chic. 1999;117:573–583.
12. Kass MA, Gordon MO, Gao F, et al.Delaying treatment of ocular hypertension: the ocular hypertension treatment study. Arch Ophthalmol-Chic. 2010;128:276–87.
13. Kass MA, Heuer DK, Higginbotham EJ, et al.The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. Arch Ophthalmol-Chic. 2002;120:701–13.
14. Land S and Friedman J Variable fusion: A new adaptive signal regression. Technical report, Department of Statistics, Stanford University. 1996;
15. Tibshirani R, Saunders M, Rosset S, Zhu J, and Knight KSparsity and smoothness via the fused lasso. J R Stat Soc Ser B. 2005;67:91–108.
16. Wang H and Li R and Tsai CL Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika. 2007;94:553–568. [PubMed: 19343105]
17. Wang H, Li B, and Leng CShrinkage tuning parameter selection with a diverging number of parameters. J R Stat Soc Ser B. 2009;71:671–683.
18. Efron B and Tibshirani RJ. An introduction to the Bootstrap. London, UK: Chapman & Hall; 1993.
19. Han D, Liu L, Su X, Johnson B, Sun L. Variable selection for random effects two-part model. Stat Methods Med Res. 2019;28:2697–2709. [PubMed: 30001684]
20. Han D, Su X, Sun L, Zhang Z, Liu L. Variable selection in joint frailty models of recurrent and terminal events. Biometrics. In press.2020;
21. Cameron AC, Gelbach JB, Miller DL. Bootstrap-based improvements for inference with clustered errors. Rev Econ Stat. 2008;90:414–427.
22. Rand WMObjective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66:846–850.
23. Kalbfleisch JD, Wolfe RAOn monitoring outcomes of medical providers. Stat Biosci. 2013;5:286–302.
24. Liu L and Lin L Subgroup analysis for heterogeneous additive partially linear models and its application to car sales data. Comput Stat Data Anal. 2019;138:239–259.
25. Zhang Z, Wang C, Nie L, and Soon GAssessing the heterogeneity of treatment effects via potential outcomes of individual patients. J R Stat Soc Ser C. 2013;62:687–704.
26. Zhang Z, Nie L, Soon G, and Liu AThe use of covariates and random effects in evaluating predictive biomarkers under a potential outcome framework. Ann Appl Stat. 2014;8:2336–2355. [PubMed: 26779295]
27. Ma S, Huang J, and Zhang ZExploration of heterogeneous treatment effects via concave fusion. Int J Biostat. 2020;16.
28. Shen J and He X Inference for Subgroup Analysis with a Structured Logistic-Normal Mixture Model. J Am Stat Assoc. 2015;110:303–312.

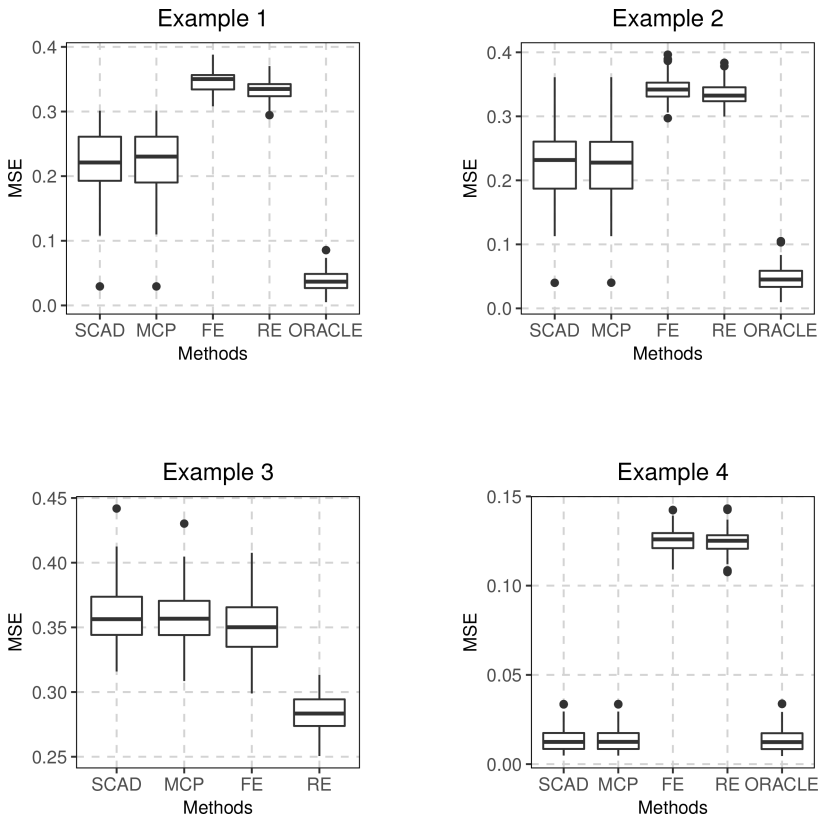

FIGURE 1.

The boxplots of square root of the mean squared error for $\hat{\mathbf{a}}$ by SCAD, MCP, fixed effects, and random effects with $m = 200$ in Settings 1–4, respectively.

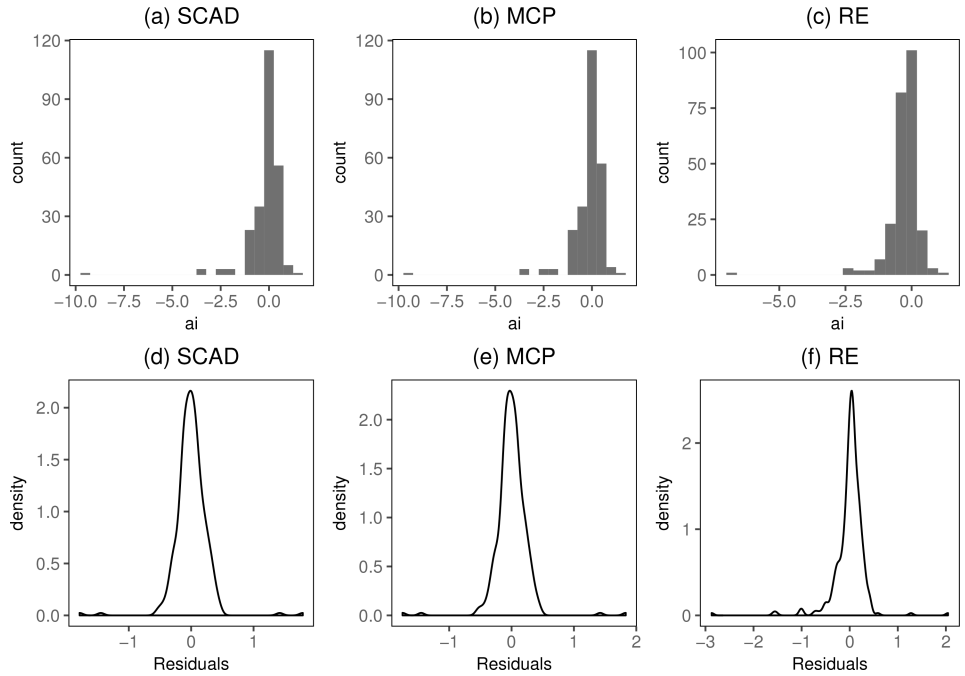

FIGURE 2.

The histograms of $\hat{\mathbf{a}}$ (a,b,c) and the kernel density plots of the residuals (d,e,f) by SCAD, MCP and random effects model (RE), respectively, in the Application.

TABLE 1

The sample mean, median, and standard deviation (s.d.) of $\widehat{K}$, the percentage (per) of $\widehat{K}$ equal to the true number of subgroups, and the Rand Index (RI) value by MCP and SCAD with $m = 50, 100, 200$ in Settings 1–4, respectively

	m	Method	mean	median	sd	per	RI
Setting 1	50	SCAD	3.07	3.00	0.2932	0.94	0.9083
		MCP	3.07	3.00	0.2564	0.93	0.9085
	100	SCAD	3.05	3.00	0.2190	0.95	0.9292
		MCP	3.06	3.00	0.2387	0.94	0.9305
	200	SCAD	3.04	3.00	0.1969	0.96	0.9328
		MCP	3.05	3.00	0.2190	0.95	0.9325
Setting 2	50	SCAD	4.01	4.00	0.1969	0.96	0.9171
		MCP	4.25	4.00	0.2778	0.95	0.9178
	100	SCAD	4.02	4.00	0.1407	0.98	0.9261
		MCP	4.02	4.00	0.1407	0.98	0.9250
	200	SCAD	4.01	4.00	0.1000	0.99	0.9346
		MCP	4.01	4.00	0.1000	0.99	0.9350
Setting 3	50	SCAD	3.63	4.00	1.1160		
		MCP	3.65	4.00	1.1044		
	100	SCAD	4.36	5.00	1.1238		
		MCP	4.41	5.00	1.1290		
	200	SCAD	5.75	6.00	1.4451		
		MCP	5.77	6.00	1.5032		
Setting 4	50	SCAD	3.00	3.00	0.0000	1.00	1.0000
		MCP	3.00	3.00	0.0000	1.00	1.0000
	100	SCAD	3.00	3.00	0.0000	1.00	1.0000
		MCP	3.00	3.00	0.0000	1.00	1.0000
	200	SCAD	3.00	3.00	0.0000	1.00	1.0000
		MCP	3.00	3.00	0.0000	1.00	1.0000

TABLE 2

The mean SRMSEs for $\hat{\mathbf{a}}$; and the bias, the empirical standard error (SEE), the sampling mean of the standard error estimate (SEM), and the coverage probability of the 95% confidence intervals of $\hat{\beta}$ in Settings 1–4, respectively

			$\hat{\mathbf{a}}$		$\hat{\beta}$		
	m	**Methods**	**SRMSE**	**bias**	**SEE**	**SEM**	**CP**
Setting 1	50	SCAD	0.2512	0.0021	0.0532	0.0652	95.0
		MCP	0.2519	0.0011	0.0539	0.0654	95.0
		FE	0.3547	0.0095	0.0739	0.0809	98.0
		RE	0.3383	0.0077	0.0663	0.0751	98.0
		Oracle	0.0784	0.0047	0.0498	0.0471	94.0
	100	SCAD	0.2241	0.0003	0.0383	0.0407	96.0
		MCP	0.2245	0.0007	0.0383	0.0407	96.0
		FE	0.3448	0.0094	0.0565	0.0563	95.0
		RE	0.3312	0.0092	0.0525	0.0523	95.0
		Oracle	0.0505	0.0025	0.0336	0.0320	96.0
	200	SCAD	0.2227	0.0012	0.0265	0.0274	96.0
		MCP	0.2201	0.0012	0.0266	0.0275	96.0
		FE	0.3468	0.0024	0.0424	0.0402	95.0
		RE	0.3335	0.0045	0.0389	0.0375	91.0
		Oracle	0.0389	0.0016	0.0220	0.0213	95.0
Setting 2	50	SCAD	0.2385	0.0006	0.0535	0.0584	97.0
		MCP	0.2375	0.0009	0.0533	0.0581	96.0
		FE	0.3523	0.0015	0.0867	0.0869	96.0
		RE	0.3506	0.0022	0.0826	0.0837	94.0
		Oracle	0.0946	0.0038	0.0470	0.0483	95.0
	100	SCAD	0.2302	0.0075	0.0388	0.0406	97.0
		MCP	0.2317	0.0069	0.0389	0.0407	97.0
		FE	0.3486	0.0084	0.0647	0.0572	94.0
		RE	0.3412	0.0095	0.0609	0.0548	93.0
		Oracle	0.0627	0.0059	0.0314	0.0324	98.0
	200	SCAD	0.2229	0.0023	0.0252	0.0268	97.0
		MCP	0.2221	0.0022	0.0253	0.0269	97.0
		FE	0.3425	0.0031	0.0366	0.0397	98.0
		RE	0.3346	0.0026	0.0363	0.0378	97.0
		Oracle	0.0468	0.0017	0.0238	0.0245	96.0
Setting 3	50	SCAD	0.3829	0.0106	0.0695	0.0712	96.0
		MCP	0.3813	0.0134	0.0672	0.0694	97.0
		FE	0.3525	0.0149	0.0827	0.0824	96.0
		RE	0.2919	0.0083	0.0616	0.0629	94.0

			$\hat{\mathbf{a}}$		$\hat{\beta}$		
	m	**Methods**	**SRMSE**	**bias**	**SEE**	**SEM**	**CP**
	100	SCAD	0.3708	0.0058	0.0460	0.0485	97.0
		MCP	0.3715	0.0051	0.0457	0.0472	98.0
		FE	0.3516	0.0033	0.0568	0.0566	95.0
		RE	0.2902	0.0008	0.0410	0.0434	96.0
	200	SCAD	0.3594	0.0042	0.0323	0.0338	98.0
		MCP	0.3580	0.0052	0.0315	0.0332	97.0
		FE	0.3504	0.0038	0.0381	0.0398	94.0
		RE	0.2833	0.0040	0.0294	0.0313	98.0
Setting 4	50	SCAD	0.0272	0.0025	0.0185	0.0175	91.0
		MCP	0.0272	0.0025	0.0185	0.0175	91.0
		FE	0.1267	0.0024	0.0217	0.0189	92.0
		RE	0.1262	0.0025	0.0217	0.0189	91.0
		Oracle	0.0272	0.0025	0.0185	0.0175	91.0
	100	SCAD	0.0200	0.0003	0.0131	0.0125	94.0
		MCP	0.0200	0.0003	0.0131	0.0125	94.0
		FE	0.1255	0.0009	0.0136	0.0133	94.0
		RE	0.1245	0.0009	0.0136	0.0133	94.0
		Oracle	0.0200	0.0003	0.0131	0.0125	94.0
	200	SCAD	0.0137	0.0017	0.0082	0.0088	97.0
		MCP	0.0137	0.0017	0.0082	0.0088	97.0
		FE	0.1254	0.0017	0.0087	0.0095	98.0
		RE	0.1247	0.0017	0.0087	0.0095	98.0
		Oracle	0.0137	0.0017	0.0082	0.0088	97.0

TABLE 3

The mean, the empirical standard error (SEE), the sampling mean of the standard error estimate (SEM), and the coverage probability of the 95% confidence intervals (CP) of the estimators of the fused effects by MCP and SCAD and Oracle estimator (Oracle) in Setting 1

m		Method	mean	SEE	SEM	CP(%)
50	$\hat{\alpha}_1$	SCAD	−1.4974	0.0911	0.0955	96.8
		MCP	−1.5029	0.0923	0.0960	97.8
		Oracle	−1.4974	0.0838	0.0839	94.0
	$\hat{\alpha}_2$	SCAD	0.0128	0.0934	0.1010	94.7
		MCP	0.0094	0.0925	0.1001	94.6
		Oracle	−0.0014	0.0851	0.0804	92.0
	$\hat{\alpha}_3$	SCAD	1.5328	0.0996	0.0905	90.4
		MCP	1.5322	0.1001	0.0901	90.2
		Oracle	1.5166	0.0889	0.0828	90.0
100	$\hat{\alpha}_1$	SCAD	−1.5098	0.0590	0.0617	94.8
		MCP	−1.5109	0.0591	0.0619	94.7
		Oracle	−1.5052	0.0559	0.0587	96.0
	$\hat{\alpha}_2$	SCAD	0.0025	0.0566	0.0651	97.9
		MCP	0.0015	0.0559	0.0653	97.9
		Oracle	0.0044	0.0519	0.0548	97.0
	$\hat{\alpha}_3$	SCAD	1.5099	0.0617	0.0607	92.7
		MCP	1.5094	0.0615	0.0610	93.6
		Oracle	1.5036	0.0591	0.0562	93.0
200	$\hat{\alpha}_1$	SCAD	−1.5043	0.0465	0.0483	93.8
		MCP	−1.5039	0.0466	0.0488	94.2
		Oracle	−1.5027	0.0406	0.0427	96.0
	$\hat{\alpha}_2$	SCAD	0.0012	0.0492	0.0511	94.2
		MCP	0.0022	0.0484	0.0515	94.5
		Oracle	−0.0012	0.0453	0.0476	96.0
	$\hat{\alpha}_3$	SCAD	1.5016	0.0418	0.0406	92.8
		MCP	1.5012	0.0416	0.0402	93.3
		Oracle	1.4963	0.0399	0.0387	93.0

TABLE 4

The mean, the empirical standard error (SEE), the sampling mean of the standard error estimate (SEM), and the coverage probability of the 95% confidence intervals (CP) of the estimators of the fused effects by MCP and SCAD and Oracle estimator (Oracle) in Setting 2

m		Method	mean	SEE	SEM	CP(%)
50	$\hat{\alpha}_1$	SCAD	−1.5084	0.0833	0.0917	91.9
		MCP	−1.5089	0.0829	0.0923	92.9
		Oracle	−1.5091	0.0800	0.0796	92.0
	$\hat{\alpha}_2$	SCAD	−0.0018	0.0919	0.1044	94.9
		MCP	0.0003	0.0914	0.1035	93.1
		Oracle	0.0071	0.0866	0.0816	94.0
	$\hat{\alpha}_3$	SCAD	1.4896	0.0979	0.0928	92.9
		MCP	1.4896	0.0991	0.0925	92.9
		Oracle	1.4923	0.0857	0.0803	90.0
	$\hat{\alpha}_4$	SCAD	−9.8853	0.4167	0.0573	23.2
		MCP	−9.8862	0.4157	0.0570	21.1
		Oracle	−9.9941	0.4192	0.0395	18.0
100	$\hat{\alpha}_1$	SCAD	−1.4970	0.0613	0.0608	92.8
		MCP	−1.4970	0.0615	0.0612	92.8
		Oracle	−1.5015	0.0558	0.0552	91.0
	$\hat{\alpha}_2$	SCAD	0.0050	0.0546	0.0710	95.9
		MCP	0.0045	0.0551	0.0716	95.9
		Oracle	0.0052	0.0524	0.0585	97.0
	$\hat{\alpha}_3$	SCAD	1.4878	0.0574	0.0627	92.8
		MCP	1.4875	0.0577	0.0630	94.9
		Oracle	1.4932	0.0525	0.0565	94.0
	$\hat{\alpha}_4$	SCAD	−9.9176	0.3980	0.0472	14.3
		MCP	−9.9179	0.3974	0.0475	14.3
		Oracle	−10.0329	0.3976	0.0291	8.0
200	$\hat{\alpha}_1$	SCAD	−1.4877	0.0465	0.0492	93.9
		MCP	−1.4879	0.0468	0.0497	93.9
		Oracle	−1.5015	0.0436	0.0459	94.0
	$\hat{\alpha}_2$	SCAD	−0.0015	0.0472	0.0513	95.9
		MCP	−0.0009	0.0474	0.0516	95.9
		Oracle	0.0044	0.0439	0.0468	95.0

m		Method	mean	SEE	SEM	CP(%)
	$\hat{\alpha}_3$	SCAD	1.4817	0.0446	0.0434	92.9
		MCP	1.4820	0.0445	0.0429	92.9
		Oracle	1.4987	0.0392	0.0376	94.0
	$\hat{\alpha}_4$	SCAD	−9.7332	0.4088	0.0338	13.3
		MCP	−9.7347	0.4048	0.0366	14.2
		Oracle	−9.9543	0.4091	0.0189	7.0

TABLE 5

The mean, the empirical standard error (SEE), the sampling mean of the standard error estimate (SEM), and the coverage probability of the 95% confidence interval (CP) of the estimators of the fused effects by MCP and SCAD and Oracle estimator (Oracle) in Setting 4

m		Method	mean	SEE	SEM	CP(%)
50	$\hat{\alpha}_1$	SCAD	−1.4952	0.0323	0.0313	94.0
		MCP	−1.4952	0.0323	0.0313	94.0
		Oracle	−1.4952	0.0323	0.0313	94.0
	$\hat{\alpha}_2$	SCAD	0.0003	0.0319	0.0302	91.0
		MCP	0.0003	0.0319	0.0302	91.0
		Oracle	0.0003	0.0319	0.0302	91.0
	$\hat{\alpha}_3$	SCAD	1.5011	0.0276	0.0314	96.0
		MCP	1.5011	0.0276	0.0314	96.0
		Oracle	1.5011	0.0276	0.0314	96.0
100	$\hat{\alpha}_1$	SCAD	−1.4993	0.0201	0.0212	97.0
		MCP	−1.4993	0.0201	0.0212	97.0
		Oracle	−1.4993	0.0201	0.0212	97.0
	$\hat{\alpha}_2$	SCAD	0.0052	0.0223	0.0215	92.0
		MCP	0.0052	0.0223	0.0215	92.0
		Oracle	0.0052	0.0223	0.0215	92.0
	$\hat{\alpha}_3$	SCAD	1.4992	0.0231	0.0213	92.0
		MCP	1.4992	0.0231	0.0213	92.0
		Oracle	1.4992	0.0231	0.0213	92.0
200	$\hat{\alpha}_1$	SCAD	−1.5000	0.0167	0.0159	93.0
		MCP	−1.5000	0.0167	0.0159	93.0
		Oracle	−1.5000	0.0167	0.0159	93.0
	$\hat{\alpha}_2$	SCAD	−0.0020	0.0158	0.0151	96.0
		MCP	−0.0020	0.0158	0.0151	96.0
		Oracle	−0.0020	0.0158	0.0151	96.0
	$\hat{\alpha}_3$	SCAD	1.5009	0.0129	0.0136	97.0
		MCP	1.5009	0.0129	0.0136	97.0
		Oracle	1.5009	0.0129	0.0136	97.0

TABLE 6

Results of $\hat{\alpha}_i$ on the sVFM in the OHTS study

Methods		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	$\hat{\alpha}_7$	$\hat{\alpha}_8$	$\hat{\alpha}_9$	$\hat{\alpha}_{10}$	$\hat{\alpha}_{11}$	$\hat{\alpha}_{12}$	$\hat{\alpha}_{13}$
SCAD	Est	−9.507	−3.486	−2.538	−2.424	−1.762	−1.159	−0.621	−0.153	0.303	0.926	1.160	1.463	
	num	1	3	2	1	3	23	35	115	56	4	1	1	
MCP	Est	−9.522	−3.498	−2.587	−2.367	−1.773	−1.171	−0.631	−0.163	0.265	0.649	1.002	1.104	1.450
	num	1	3	2	1	3	23	35	115	51	6	3	1	1

TABLE 7

Results of $\widehat{\beta}$ on the sVFM by SCAD, MCP, random effects with the outlier (RE1) and random effects without the outlier (RE2), respectively

Methods		RA	VCD	gender	stroke	race	age	CCT	IOP	PSD
SCAD	Est	−0.1152	−0.0400	0.0073	−0.0500	−0.2559	−0.1225	0.0601	0.0234	−0.1300
	s.e	0.1256	0.0462	0.1139	0.1697	0.1316	0.0567	0.0467	0.0402	0.0500
	p-value	0.3589	0.3858	0.9488	0.7683	0.0519	0.0307	0.1980	0.5606	0.0093
MCP	Est	−0.1122	−0.0419	0.0143	−0.0601	−0.2544	−0.1206	0.0575	0.0265	−0.1284
	s.e	0.1281	0.0475	0.1165	0.1635	0.1345	0.0577	0.0434	0.0363	0.0524
	p-value	0.3813	0.3779	0.9026	0.7133	0.0585	0.0366	0.1852	0.4655	0.0142
RE1	ESt	−0.1592	−0.0450	0.0119	−0.0408	−0.2619	−0.1480	0.0208	0.0215	−0.1321
	s.e	0.1184	0.0538	0.1191	0.2226	0.1257	0.0584	0.0581	0.0525	0.0484
	p-value	0.1799	0.4064	0.9203	0.8549	0.0383	0.0119	0.7220	0.6835	0.0082
RE2	ESt	−0.0734	−0.0344	0.0968	−0.0779	−0.1572	−0.1096	0.0507	0.0195	−0.1159
	s.e	0.0886	0.0420	0.0890	0.1650	0.0940	0.0440	0.0439	0.0414	0.0399
	p-value	0.4085	0.4160	0.2780	0.6372	0.0959	0.0133	0.2529	0.6397	0.0051