The Annals of Applied Statistics
2022, Vol. 16, No. 1, 80–103
https://doi.org/10.1214/21-AOAS1503

SUBGROUP-EFFECTS MODELS FOR THE ANALYSIS OF PERSONAL TREATMENT EFFECTS

BY LING ZHOU[1,a], SHIQUAN SUN[2,b], HAODA FU[3,c] AND PETER X.-K. SONG[4,d]

[1]*Center for Statistical Research, Southwestern University of Finance and Economics,* [a]*zhouling@swufe.edu.cn*
[2]*School of Public Health, Xi'an Jiaotong University,* [b]*sqsunsph@xjtu.edu.cn*
[3]*Eli Lilly and Company,* [c]*fu_haoda@lilly.com*
[4]*Department of Biostatistics, University of Michigan,* [d]*pxsong@umich.edu*

The emerging field of precision medicine is transforming statistical analysis from the classical paradigm of population-average treatment effects into that of personal treatment effects. This new scientific mission has called for adequate statistical methods to assess heterogeneous covariate effects in regression analysis. This paper focuses on a subgroup analysis that consists of two primary analytic tasks: identification of treatment effect subgroups and individual group memberships, and statistical inference on treatment effects by subgroup. We propose an approach to synergizing supervised clustering analysis via alternating direction method of multipliers (ADMM) algorithm and statistical inference on subgroup effects via expectation-maximization (EM) algorithm. Our proposed procedure, termed as hybrid operation for subgroup analysis (HOSA), enjoys computational speed and numerical stability with interpretability and reproducibility. We establish key theoretical properties for both proposed clustering and inference procedures. Numerical illustration includes extensive simulation studies and analyses of motivating data from two randomized clinical trials to learn subgroup treatment effects.

1. Introduction. Consider a random sample of $(y_i, x_i, \boldsymbol{z}_{0i})$, $i = 1, \ldots, n$, collected from a clinical study, where y_i is the outcome of interest (e.g., fasting glucose level) and x_i is the treatment covariate of interest which may be categorical (e.g., drugs A and B) or continuous (e.g., exposure to toxic agents). In addition, $\boldsymbol{z}_{0i}$ is a q_0-dimensional vector of potential confounders, including the intercept, useful to adjust the assessment of treatment effects. Suppose that the linear model is adopted to study the treatment-response relationship,

$$y_i = x_i\beta_i + \boldsymbol{z}_{0i}^T\boldsymbol{\alpha} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where β_i's represent personal treatment effects, each for one subject, and random errors $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$, $i = 1, \ldots, n$ which are independent of x_i and $\boldsymbol{z}_{0i}$. In this paper these subject-specific parameters β_i's are of the central interest in the analysis. The classical statistical analysis concerns primarily a population-average treatment effect, *say*, β_x under the homogeneity assumption of $\beta_i \equiv \beta_x$, $i = 1, \ldots, n$. In this case the maximum likelihood inference is the method of choice which has been well studied and extensively used in practice. In effect, evaluation of population-average treatment effectiveness is mandatory in a new drug development to fullfil the requirements by the FDA drug approval protocol. According to Wong, Siah and Lo (2019), the success rate of clinical trial, in light of population-average treatment effectiveness, has been unfortunately very low, reportedly being only about 5% during the period of 16 years from 2000 to 2015. To address the issue that a considerably large number of clinical trails failed in their third phase of study, a revolutionary initiative has been proposed to relax the population-average efficacy paradigm. This gives rise to a fundamental

Received February 2021.
Key words and phrases. ADMM algorithm, EM algorithm, maximum likelihood, precision medicine, supervised clustering.

question: whether or not there are some patients in the same treatment arm, such as those in an active drug arm in a randomized placebo-controlled trial, who may experience stronger treatment efficacy than others; and if so, who they are. Identifying and characterizing those patients who benefit more from a therapy would shed light on designing a subsequent confirmatory clinical study that targets at a specific subpopulation of patients, instead of a general population, as potential drug users.

As far as statistical methodology concerns, answers to the aforementioned questions cannot be obtained by a widely used standard random effects modeling approach in which subject-specific random effects β_i's in model (1.1) are assumed to be independent and identically distributed (i.i.d.) from a normal distribution, *say* $\mathcal{N}(\beta_x, \sigma_b^2)$. This is because this random-effects specification does not give any subgroup structures to characterize clustered individual-level treatment benefits, rather treating all patients belonging to one treatment group with the group-level average effect β_x. To allow multiple subgroups of treatment effects, we may consider an explicit formulation of subgroups via a K-component mixture of normals, namely, $\beta_i \overset{\text{iid}}{\sim} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k^2)$, $i = 1, \ldots, n$, where K denotes the number of treatment-effects subgroups. This model-based clustering formulation enables us to perform a supervised clustering analysis on the subject-specific treatment effects, leading to straightforward interpretations. That is, μ_k represents the average treatment effect of subgroup k or the centroid of β_i's in the kth subgroup, σ_k^2 describes the variability or the size of the kth subgroup, and π_k is the probability of β_i belonging to subgroup k. When all $\sigma_k^2 = 0$, this mixture model degenerates to a discrete model with K singletons $\{\mu_1, \ldots, \mu_K\}$, which is the underlying subgroup model assumed by many existing methods of subgroup analyses, such as Ma and Huang (2017). Clearly, when $K = n$, each subject forming one's own group, the individual treatment parameters β_i's are not estimable. Thus, in practice, K should be an integer smaller than n. In this paper we consider a two-level hierarchical model given below:

$$
\begin{aligned}
y_i &= x_i \beta_i + \boldsymbol{z}_{0i}^T \boldsymbol{\alpha} + \varepsilon_i, \quad i = 1, \ldots, n, \\
\beta_i &\overset{\text{iid}}{\sim} \sum_{k=1}^{K} \pi_k \mathcal{N}\big(\mu_k, \sigma_k^2\big), \quad i = 1, \ldots, n
\end{aligned}
\tag{1.2}
$$

which is referred to as *subgroup-effects model (SGEM)*. To emphasize the presence of K subgroups in the model specification (1.1)–(1.2), we may take an abbreviation as SGEM(K).

To further elucidate the interpretation of parameters β_i's in SGEM, let us consider a motivating example of a randomized two-arm placebo-controlled trial conducted by scientists at the University of Michigan Children's Environmental Health Center (CEHC) to assess the effect of maternal calcium supplementation during pregnancy on reducing infant blood lead concentration (Ettinger, Hu and Hernandez-Avila (2007), Ettinger et al. (2009)). This trial is part of the nutritional study with participants from Mexico City (Perng et al. (2019)). A vast literature has unveiled that lead is detrimental on multiple human body systems as well as harmful on human neurobehavioral and cognitive development, particularly in children (Bellinger, Stiles and Needleman (1992), Boivin and Giordani (1995), Hu et al. (2006)). This is an important trial in the emerging field of precision nutrition to develop tailored recommendations of nutrients, based on personal internal and external environmental exposures; see the detail of the data description and data analysis of this trial in Section 6. By coding $x_i = 1$ for daily intake of calcium supplement and 0 for placebo, the parameter β_i is regarded as a differential treatment effect from the placebo effect on patient i who takes the calcium supplementation. Note that for subjects who are randomized into the placebo arm, they can receive a baseline dose of calcium from their food. Here, the objective of clinical interest lies in potential subgroup structures for the subjects on the calcium supplementation arm relative

to the reference effect of placebo dose of calcium intake from food. In this study our collaborators consider specifically two subgroups ($K = 2$) in order to deliver a viable nutritional recommendation to the pregnant women: one being the subgroup of women who experience no difference of treatment effect from the placebo (i.e., calcium from food) and the other being the subgroup of women who experience better treatment benefit. In practice, the number of subgroups, K, may often be specified by practitioners according to their *a priori* clinical hypothesis or objectives of their clinical study. Thus, in the subgroup analysis a more important analytic task is to determine memberships of patients with high precision and reliability than to determine the number of subgroups. After the memberships being determined, the maximum likelihood method can be readily applied to carry out the remaining statistical analysis.

Subgroup labels or memberships as well as associated probabilities of label assignment may be determined by invoking individual latent categorical variables $\delta_i \in \mathcal{D} = \{1, \ldots, K\}$, $i = 1, \ldots, n$, whose probability mass function is denoted as $P(\delta_i = k) = \pi_{ik}$, $k = 1, \ldots, K$. Here, π_{ik} is the probability of patient i belonging to subgroup k which may be modeled as a function of some covariates. It is easy to see that SGEM(K) can be, equivalently, rewritten in the form of a mixture linear model (MLM),

$$y_i = \sum_{k=1}^{K} I(\delta_i = k)\mu_k x_i + z_{0i}^T \boldsymbol{\alpha} + \xi_i, \quad \text{with independent errors } \xi_i \sim \mathcal{N}(0, \sigma_i^2), \tag{1.3}$$

where variances $\sigma_i^2 = \sum_{k=1}^{K} I(\delta_i = k)\sigma_{ik}^2$ and $\sigma_{ik}^2 = x_i^2\sigma_k^2 + \sigma_\varepsilon^2$. The class of MLMs in (1.3) has been well studied and applied in many practical areas; see, for example, Deb and Holmes (2000) and Grün and Leisch (2007); also, see McLachlan, Lee and Rathnayake (2019) for a comprehensive review of finite mixture models. In addition, an MLM may be regarded as an instance of the hierarchical mixture of experts (Jacobs et al. (1991)) where the mixture proportions or gate functions may depend on some covariates (Wei and Kosorok (2013)).

Parameter estimation in an MLM (1.3) is notoriously difficult, due to the nonconvexity of the likelihood function and the existence of local optima. In the literature the maximum likelihood estimation (MLE) method is mostly widely adopted which is often implemented via the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin (1977), Muthén and Shedden (1999), Verbeke and Lesaffre (1996), Viele and Tong (2002), Xu and Hedeker (2001)). It is known that the EM algorithm depends heavily on the quality of initial values, and its convergent values are local optima (Wu (1983)). A vast literature has focused on parameter estimation in the case of two-component mixture models, namely, $\beta_i = \delta_i\mu_{1i} + (1 - \delta_i)\mu_{2i}$ with latent label $\delta_i \in \{0, 1\}$; see, for example, Balakrishnan, Wainwright and Yu (2017), Yi, Caramanis and Sanghavi (2014). Proust and Jacqmin-Gadda (2005) adopted a direct maximization of the likelihood via a Marquard optimization algorithm which has been implemented in the popular R package **lcmm** (Proust-Lima, Philipps and Liquet (2017)). We will compare the performance of this R package thoroughly with our proposed method in this paper. The method of moments has also been studied in the literature to handle general K-mode mixture models, including tensor decomposition (Anandkumar et al. (2014), Chaganty and Liang (2013), Sedghi, Janzamin and Anandkumar (2016)), subspace clustering (Elhamifar and Vidal (2013), Pimentel-Alarcón et al. (2017), Soltanolkotabi, Elhamifar and Candès (2014)), and convex formulation (Chen, Yi and Caramanis (2018)), among others.

Recently, several two-stage algorithms have shown encouraging improvements on both fast convergence rate and desirable statistical efficiency. At the first stage this type of algorithm generates high-quality initial estimates, followed at the second stage by a procedure to yield refined estimation via either the EM algorithm or its variants. One example worth mentioning is that in a multiclass labeling problem, the first stage uses a spectral method

by Zhang et al. (2016) or a tenor method by Zhong, Jain and Dhillon (2016), and the resulting estimation enjoys both faster computational speed and better statistical power. Such success motivates us to develop a new two-stage procedure for SGEM in which the first stage produces high-quality initial results of subgroups. Specifically, to address the technical need of supervised clustering in parameters β_i's, we invoke the alternating direction method of multipliers (ADMM, Boyd et al. (2011)) algorithm to generate initial values which will be used as inputs to the EM algorithm at the second-stage to perform the maximum likelihood estimation and inference. The advantage of ADMM lies in the fact that it provides a fast and direct hard-division in a clustering analysis, as supposed to a soft-division via a posterior probability-voting scheme in the EM algorithm, which makes the algorithm slow to converge. In other words, ADMM can help quickly reach an orbit in the parameter space near the optimal solution. Being an interesting synergy of machine learning and statistical inference, the proposed procedure is named as hybrid operation for subgroup analysis (HOSA). The first stage of HOSA is a key to parameter estimation by running a supervised clustering analysis of parameters via ADMM algorithm (Chi and Lange (2015), Ma and Huang (2017)). Dated back to 1970s (Gabay and Mercier (1976), Glowinski (2014)), ADMM has been shown in multiple occasions to enjoy the power of dual decomposition and augmented Lagrangian methods in constrained optimization. In particular, ADMM algorithm has demonstrated its advantages in the operation of subgroup fusion and the implementation of parallel calculation; refer to Ma and Huang (2017), Mihić, Zhu and Ye (2021), Mota et al. (2013), among others.

This paper makes the following new contributions to the study of personal treatment effects: (i) Incorporating subject-level characterizations into SGEM, we can not only evaluate subgroup-level treatment effects but also predict individual person's subgroup label. (ii) The SGEM specification allows to quantify both within-subgroup variability for personal treatment effects and the amount of uncertainty associated with subgroup assignments. As becoming evident throughout this paper, the SGEM formulation is more flexible and interpretable than existing hard-threshold methods, like the classical K-means clustering analysis. More importantly, (iii) with high quality initial values given by the ADMM algorithm, the EM algorithm-based maximum likelihood method enjoys stable numerical performance and fast convergence at a geometric rate, as shown by Corollary 1 in Section 4. Thus, the final numerical results of the parameter estimation and inference from the EM algorithm are little sensitive to the specification of initial values which is "guarded" by the ADMM algorithm that yields parameter estimates near the optimal solution. These properties are particularly important for the method to be applied in clinical studies. The R package HOSA of the proposed method is available in the Supplementary Material (Zhou et al. (2022)) and on the following GitHub URL: https://github.com/sqsun/HOSA.

The paper is organized as follows. Section 2 presents the SGEM model specification and interpretation. Section 3 concerns the implementation of HOSA, including the selection of the number of mixture components. Large sample properties of HOSA are studied in Section 4. Numerical simulation studies and a real data analysis are provided in Sections 5 and 6, respectively. Some concluding remarks are given in Section 7. All technique details and additional numerical results are given in the Appendix and the Supplementary Material (Zhou et al. (2022)).

2. Formulation.

2.1. *Subgroup-effects model*. The primary objective of precision medicine is twofold: to investigate whether there exist distinct personal treatment effects and, if so, to determine the treatment subgroup membership of each subject. For such purposes it is inevitable to include individual predictors in SGEM useful for personal treatment decision-making. First,

we propose a multilevel logistic model on π_{ik} in (1.2) for subgroup label prediction via a q_2-dimensional vector of covariates $\boldsymbol{z}_{2i} = (z_{2i,1}, \ldots, z_{2i,q_2})^T$, with $z_{2i,1} = 1$ for the intercept term. That is, the (subgroup) membership model, $\log(\pi_{ik}/\pi_{iK}) = \boldsymbol{z}_{2i}^T \boldsymbol{\zeta}_{2k}$, $k = 1, \ldots, K-1$, where subgroup K is set as the reference. Second, we model personal mean effects by a linear model $\mu_{ik} = \boldsymbol{z}_{1i}^T \boldsymbol{\zeta}_{1k}$ with a q_1-element vector of covariates, $\boldsymbol{z}_{1i}$, including $z_{1i,1} = 1$ as the intercept term. The resulting (1.2) may be written as $\beta_i \sim \sum_{k=1}^K \pi_{ik}(\boldsymbol{z}_{i2}) \mathcal{N}(\mu_{ik}(\boldsymbol{z}_{i1}), \sigma_k^2)$. Note that the q_1-dimensional parameter vector $\boldsymbol{\zeta}_{1k}$ quantifies a set of interaction effects between treatment x_i and covariates in $\boldsymbol{z}_{1i}$ for the kth subgroup. Gunter, Zhu and Murphy (2011) refer to such types of covariates involved in these interactions as *prescriptive variables* which are typically different from confounding covariates in $\boldsymbol{z}_0$. For example, in a randomized clinical trial $\boldsymbol{z}_0$ may be chosen as an empty set, due to randomization, but the vector of prescriptive variables $\boldsymbol{z}_1$ is not, due to its role in characterizing group-level treatment heterogeneity.

Following the classical decomposition of fixed- and random-effects, we may write $\beta_i = b_i + a_i$, where b_i and a_i denote fixed-effect and random-effect, respectively. Consequently, we rewrite the SGEM model above as a mixture linear model (LML) with random effects,

$$y_i = x_i b_i + \boldsymbol{z}_{0i}^T \boldsymbol{\alpha} + x_i a_i + \varepsilon_i, \quad \text{and} \quad \log\left(\frac{\pi_{ik}}{\pi_{iK}}\right) = \boldsymbol{z}_{2i}^T \boldsymbol{\zeta}_{2k}, \quad k = 1, \ldots, K-1, \tag{2.1}$$

with $b_i = \sum_{k=1}^K I(\delta_i = k) \boldsymbol{z}_{1i}^T \boldsymbol{\zeta}_{1k}$ and $a_i = \sum_{k=1}^K I(\delta_i = k) v_{ik}$, where errors $v_{ik} \sim \mathcal{N}(0, \sigma_k^2)$ are independent and identically distributed (i.i.d.) within subgroup k. The issue of parameter identifiability in the family of mixture models has been extensively studied in the literature (e.g., Frühwirth-Schnatter (2006), McLachlan, Lee and Rathnayake (2019)). It pertains essentially to three aspects: interchangeability of component labels (Redner and Walker (1984)), potential overfitting (Crawford (1994)), and generic nonidentificability (Teicher (1961)). These may be ruled out through certain formal identifiability constraints (Frühwirth-Schnatter (2006), McLachlan, Lee and Rathnayake (2019)). Specifically, for our model (2.1) in which all normal components are different and the mixing proportions π_{ik}'s are nonzero, the identifiability condition is given as follows. If two mixture density functions are equal for almost every y with two equal sets of nonzero model parameters $(K, \boldsymbol{\alpha}, \sigma_\varepsilon, \boldsymbol{\zeta}_{1k}, \boldsymbol{\zeta}_{2k}, \sigma_k, k = 1, \ldots, K)$ and $(K', \boldsymbol{\alpha}', \sigma_\varepsilon', \boldsymbol{\zeta}_{1k}', \boldsymbol{\zeta}_{2k}', \sigma_k', k = 1, \ldots, K')$, then $K = K'$ and corresponding parameters are equivalent, respectively.

2.2. *Maximum likelihood estimation.* For the proposed parametric model (2.1), we use the maximum likelihood estimation (MLE) method to estimate the model parameters, denoted by $\boldsymbol{\Theta} = (\boldsymbol{\theta}^T, \boldsymbol{\zeta}_2^T)^T$ with $\boldsymbol{\theta} = (\boldsymbol{\zeta}_{11}^T, \ldots, \boldsymbol{\zeta}_{1K}^T, \boldsymbol{\alpha}^T, \sigma_1^2, \ldots, \sigma_K^2, \sigma_\varepsilon^2)^T$ and $\boldsymbol{\zeta}_2 = (\boldsymbol{\zeta}_{21}^T, \ldots, \boldsymbol{\zeta}_{2,K-1}^T)^T$. Let $\boldsymbol{w}_i = (\boldsymbol{\mathcal{Z}}_i^T, \boldsymbol{z}_{0i}^T, \boldsymbol{z}_{2i}^T)^T$ be the combined set of covariates, where the vector of interaction covariates is denoted by $\boldsymbol{\mathcal{Z}}_i = x_i \boldsymbol{z}_{1i} = (\mathcal{Z}_{i,1}, \ldots, \mathcal{Z}_{i,q_1})^T$. Let $\boldsymbol{\theta}^*$ and $\boldsymbol{\zeta}_2^*$ be the true values of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}_2$, respectively. Denote the conditional density function of y_i, given both random effect a_i and label $\delta_i = k$, as $f_k(y_i \mid a_i, \delta_i; \boldsymbol{\theta}) = \sigma_\varepsilon^{-1} \phi(\sigma_\varepsilon^{-1}(y_i - \boldsymbol{\mathcal{Z}}_i^T \boldsymbol{\zeta}_{1k} - \boldsymbol{z}_{0i}^T \boldsymbol{\alpha} - x_i a_i))$ and the density of a_i, given $\delta_i = k$, as $f_k(a_i \mid \delta_i; \boldsymbol{\theta}) = \phi(a_i/\sigma_k)/\sigma_k$, where $\phi(\cdot)$ is the standard normal density function. Then, the argumented likelihood for subject i takes the following form:

$$\mathcal{L}(\boldsymbol{\Theta}; y_i, a_i, \delta_i) = \prod_{k=1}^K \{\pi_{ik} f_k(y_i \mid a_i, \delta_i; \boldsymbol{\theta}) f_k(a_i \mid \delta_i; \boldsymbol{\theta})\}^{I(\delta_i = k)},$$

where $I(\cdot)$ is the indicator function. Note that here both a_i and δ_i are not observed which will be handled using the EM algorithm (Dempster, Laird and Rubin (1977)) in the estimation and inference for the model parameters $\boldsymbol{\Theta}$. The EM algorithm requires iteratively executing E-step and M-step, where, at iteration $t+1$, *say*, E-step is to calculate the so-called Q-function,

$Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)})$, namely, the expected value of the log-likelihood function with respect to the conditional distribution of (a_i, δ_i), given y_i, under the current updates $\boldsymbol{\Theta}^{(t)}$. That is,

$$Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)}) = \mathbb{E}_{y;\boldsymbol{\theta}^*,\boldsymbol{\zeta}_2^*}\big[\mathbb{E}_{a,\delta|y;\boldsymbol{\Theta}^{(t)}}\{\log \mathcal{L}(\boldsymbol{\Theta}; y_i, a_i, \delta_i)\}\big], \tag{2.2}$$

where $\mathbb{E}_{u;\boldsymbol{\theta}}$ and $\mathbb{E}_{u|v,\boldsymbol{\theta}}$ represent the expectations with respect to, respectivley, the distribution of random variable u and the conditional distribution of u, given v, under the parameter $\boldsymbol{\theta}$. On the other hand, M-step is to update parameters $\boldsymbol{\Theta}$ by maximizing the conditional expectation (2.2), namely, $\boldsymbol{\Theta}^{(t+1)} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)})$. The details of the E-step and M-step are given in Section 3.

With certain high-quality initialization, Theorem 2 in Section 4 shows that the EM algorithm converges at a geometric rate to a local maximum close to the maximum likelihood estimate. In addition, Section 5 illustrates through extensive simulation experiments that estimation results from the EM algorithm are sensitive to initial values of interaction effects $\boldsymbol{\zeta}_{1k}$'s but less sensitive to the other model parameters. This is in the agreement with findings from Proust and Jacqmin-Gadda (2005). To address this numerical challenge, we invoke the ADMM algorithm to generate good initial values of $\boldsymbol{\zeta}_{1k}$'s which can produce high-quality initialization for parameter $\boldsymbol{\zeta}_{1k}$'s with a well-estimated number of subgroups. This will positively impact numerical performances of the EM algorithm in terms of reliability and stability. This is because that ADMM offers, with theoretical guarantees, a direct and fast hard-division of interaction effects into subgroups and produces initial values near the optimal solution. As proved in Theorem 1, the theoretical guarantees give rise to a well-behaved initialization procedure that results in desired clustering and membership determination.

Denote by $\mathcal{G} = \mathcal{G}_1 \cup \cdots \cup \mathcal{G}_K$ a group partition of n subjects (precisely, of n treatment effects β_i's). Here, these group memberships are unknown and need to be estimated by the ADMM algorithm. Once a subgroup structure is fixed, parameters $\boldsymbol{\zeta}_{1k}$'s are fixed as the common value of subgroup $k, k = 1, \ldots, K$. To proceed, we first project the subgroup-level interaction effects $\boldsymbol{\zeta}_{1k}, k = 1, \ldots, K$ into higher dimensional vectors, denoted by $\boldsymbol{\eta}_{1i} = (\eta_{1i,1}, \ldots, \eta_{1i,q_1})^T$, where $\eta_{1i,j}$ is the personal interaction effect of prescriptive covariate j for subject i. Clearly, after this project there do not exist any subgroup structures among the parameters $\boldsymbol{\eta}_{1i}$. To learn the underlying group structures and associate subjects' memberships, we then consider minimizing the following penalized objective function with respect to parameters $\boldsymbol{\eta}_{1i}$'s to reconstruct the underlying subgroup structures,

$$\min_{\boldsymbol{\eta}_1, \boldsymbol{\alpha}} \left\{ \sum_{i=1}^{n} \frac{1}{2} (y_i - \boldsymbol{\mathcal{Z}}_i^T \boldsymbol{\eta}_{1i} - \boldsymbol{z}_{0i}^T \boldsymbol{\alpha})^2 + \sum_{i<j} p_\gamma(\|\boldsymbol{\eta}_{1i} - \boldsymbol{\eta}_{1j}\|_2, \lambda) \right\}, \tag{2.3}$$

where $p_\gamma(\cdot, \lambda)$ is a fusion penalty function on the difference between parameters $\boldsymbol{\eta}_{1i}$'s, and both $\lambda > 0$ and $\gamma > 0$ are tuning parameters. The rationale of optimization, given in (2.3), is to fuse similar interaction effects $\boldsymbol{\eta}_{1i}$'s into subgroups, one subgroup with a common value $\boldsymbol{\zeta}_{1k}$ obtained from the fused penalty function $p_\gamma(\cdot, \lambda)$. When $\lambda = \infty$, all $\boldsymbol{\eta}_{i1}$'s will be fused into one value, while $\lambda = 0$ all $\boldsymbol{\eta}_{1i}$'s will form their own clusters (i.e., n subgroups). With a suitable choice of λ, $\boldsymbol{\eta}_{1i}$'s are forced to form some subgroups in the above penalized estimation so that certain clusters of personal effects can be identified. Moreover, this fused regularization method produces subgroup memberships for all subjects. Several types of penalty functions are available in the literature, for example, the LASSO penalty (Tibshirani et al. (2005)), MCP penalty (Zhang (2010)), SCAD penalty (Fan and Li (2001)), among others. In this paper we choose MCP penalty function, due to its proven advantages of low bias and stable numerical performance (Ma and Huang (2017)). The MCP penalty takes the form, $p_a(x; \lambda) = \lambda \int_0^x \{1 - s/(a\lambda)\}_+ ds$ for $a > 1$, where $(x)_+ = x$ if $x > 0$ and $(x)_+ = 0$, otherwise. Note that the dimension of the entire vector $\boldsymbol{\eta}_1 = (\boldsymbol{\eta}_{11}^T, \ldots, \boldsymbol{\eta}_{1n}^T)^T$ is nq_1 which linearly increases along the sample size n. Therefore, we invoke a coordinate (blockwise) ADMM algorithm to overcome the computational burden in the fused MCP regularization with related details given in Section 3.

3. Implementation. The maximum likelihood estimation is implemented by the EM algorithm with initial values from the ADMM algorithm.

3.1. *EM algorithm.* Given a value of K and updates $\boldsymbol{\Theta}^{(t)}$ from step t, both E-step and M-step at iteration $t+1$ are given as follows. The detailed derivations of the EM algorithm can be found in Section 3 of the Supplementary Material (Zhou et al. (2022)). For the ease of exposition, denote $f_k(y, a \mid \delta; \boldsymbol{\theta}) = f(y, a \mid \delta = k; \boldsymbol{\theta})$ and $f_k(a \mid \delta; \boldsymbol{\theta}) = f(a \mid \delta = k; \boldsymbol{\theta})$. At the E-step we evaluate the Q-function by integrating out random effects a_i's, while initial δ_i's are passed from the ADMM algorithm,

$$
\begin{aligned}
&Q_n\big(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)}\big) \\
&\quad = \mathbb{E}_{a,\delta \mid y;\boldsymbol{\Theta}^{(t)}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log \mathcal{L}(\boldsymbol{\Theta}; y_i, a_i, \delta_i) \right\} \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\gamma_{ik}^{(t)}}{f_{ik}(\boldsymbol{\theta}^{(t)})} \int_{\mathcal{R}} \log\big\{ f_k(y_i \mid a_i, \delta_i; \boldsymbol{\theta})\, f_k(a_i \mid \delta_i; \boldsymbol{\theta}) \big\} f_k\big(y_i, a_i \mid \delta_i; \boldsymbol{\theta}^{(t)}\big)\, da_i \\
&\qquad + \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \log \pi_{ik} \gamma_{ik}^{(t)},
\end{aligned} \tag{3.1}
$$

where $f_{ik}(\boldsymbol{\theta}) = \int_{\mathcal{R}} f_k(y_i, a_i \mid \delta_i; \boldsymbol{\theta})\, da_i$ and $\gamma_{ik}^{(t)}$ is the updated posterior probability of subject i belonging to subgroup k, given as follows:

$$
\gamma_{ik}^{(t)} = \frac{\pi_{ik}^{(t)} f_{ik}(\boldsymbol{\theta}^{(t)})}{\sum_{k=1}^{K} \pi_{ik}^{(t)} f_{ik}(\boldsymbol{\theta}^{(t)})}, \quad \text{with } \pi_{ik}^{(t)} = \pi_{iK}^{(t)} \exp\big(z_{2i}^T \boldsymbol{\zeta}_{2k}^{(t)}\big),\ k = 1, \ldots, K-1, \tag{3.2}
$$

where $\pi_{iK}^{(t)} = \{1 + \sum_{k=1}^{K-1} \exp(z_{2i}^T \boldsymbol{\zeta}_{2k}^{(t)})\}^{-1}$. At the M-step we update the parameters $\boldsymbol{\theta}$ of the outcome model by the following closed-form expressions: for $k = 1, \ldots, K$,

$$
\begin{aligned}
\boldsymbol{\zeta}_{1k}^{(t+1)} &= \left(\sum_{i=1}^{n} \gamma_{ik}^{(t)} \mathcal{Z}_i \mathcal{Z}_i^T \right)^{-1} \left\{ \sum_{i=1}^{n} \gamma_{ik}^{(t)} \big(y_i - z_{0i}^T \boldsymbol{\alpha}^{(t)} - x_i A_{ik}^{-1}(\boldsymbol{\theta}^{(t)}) B_{ik}(\boldsymbol{\theta}^{(t)})\big) \mathcal{Z}_i \right\}, \\
\boldsymbol{\alpha}^{(t+1)} &= \left(\sum_{i=1}^{n} z_{0i} z_{0i}^T \right)^{-1} \left\{ \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik}^{(t)} \big(y_i - \mathcal{Z}_i^T \boldsymbol{\zeta}_{1k}^{(t)} - x_i A_{ik}^{-1}(\boldsymbol{\theta}^{(t)}) B_{ik}(\boldsymbol{\theta}^{(t)})\big) z_{0i} \right\}, \\
\sigma_{\varepsilon}^{2(t+1)} &= n^{-1} \Bigg[\sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik}^{(t)} \big\{ x_i A_{ik}^{-1}(\boldsymbol{\theta}^{(t)}) \big(1 + B_{ik}^2(\boldsymbol{\theta}^{(t)}) A_{ik}^{-1}(\boldsymbol{\theta}^{(t)})\big) x_i - 2 x_i A_{ik}^{-1}(\boldsymbol{\theta}^{(t)}) \\
&\qquad \times B_{ik}(\boldsymbol{\theta}^{(t)}) \big(y_i - z_{0i}^T \boldsymbol{\alpha}^{(t)} - \mathcal{Z}_i^T \boldsymbol{\zeta}_{1k}^{(t)}\big) + \big(y_i - z_{0i}^T \boldsymbol{\alpha}^{(t)} - \mathcal{Z}_i^T \boldsymbol{\zeta}_{1k}^{(t)}\big)^2 \big\} \Bigg], \\
\sigma_k^{2(t+1)} &= \left(\sum_{i=1}^{n} \gamma_{ik}^{(t)} \right)^{-1} \left\{ \sum_{i=1}^{n} \gamma_{ik}^{(t)} A_{ik}^{-1}(\boldsymbol{\theta}^{(t)}) \big(1 + B_{ik}^2(\boldsymbol{\theta}^{(t)}) A_{ik}^{-1}(\boldsymbol{\theta}^{(t)})\big) \right\},
\end{aligned}
$$

where $A_{ik}(\boldsymbol{\theta}) = (\sigma_{\varepsilon}^2 \sigma_k^2)^{-1}(x_i^2 \sigma_k^2 + \sigma_{\varepsilon}^2)$ and $B_{ik}(\boldsymbol{\theta}) = \sigma_{\varepsilon}^{-2}(y_i - z_{0i}^T \boldsymbol{\alpha} - \mathcal{Z}_i^T \boldsymbol{\zeta}_{1k}) x_i$. To update parameters $\boldsymbol{\zeta}_{2k}^{(t+1)}$ of the membership model, we run the multilevel logistic model by maximizing the log-likelihood, $\sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik}^{(t)} [z_{2i}^T \boldsymbol{\zeta}_{2k} - \log\{1 + \sum_{k=1}^{K-1} \exp(z_{2i}^T \boldsymbol{\zeta}_{2k})\}]$, with respect to $\boldsymbol{\zeta}_{2k}^{(t+1)}$, $k = 1, \ldots, K-1$.

3.2. *ADMM algorithm.* The key idea behind the ADMM algorithm is the augmentation of the parameter space, which enables to divide the original optimization problem into several simpler optimization subproblems, and each subproblem has a closed form solution for a certain target parameter. This strategy can significantly speed up the iterative search for the optimal solution. Let $\boldsymbol{\vartheta}_{ij} = (\vartheta_{ij,1}, \ldots, \vartheta_{ij,q_1})^T$ and $\boldsymbol{t}_{ij} = (t_{ij,1}, \ldots, t_{ij,q_1})^T$, $i, j = 1, \ldots, n$, be individual-level auxiliary parameters under the parameter augmentation; $\boldsymbol{\vartheta}_{ij}$ is the collection of all pairwise differences between $\boldsymbol{\eta}_{1i}$ and $\boldsymbol{\eta}_{1j}$ for all $i \neq j$, while $\boldsymbol{t}_{ij}$ is a vector of multipliers monitoring parameter fusion. For the interaction effect of $\mathcal{Z}_r = xz_{1r}$ (z_{1r} being the rth perspective covariate, $r = 1, \ldots, q_1$), $[\cdot]_r$ denotes a column vector corresponding to z_{1r}; for example, $[\boldsymbol{\eta}_1]_r = (\eta_{11,r}, \ldots, \eta_{1n,r})^T$, $[\boldsymbol{\vartheta}]_r = (\vartheta_{12,r}, \vartheta_{13,r}, \ldots, \vartheta_{(n-1)n,r})^T$, and $[\boldsymbol{t}]_r = (t_{12,r}, \ldots, t_{(n-1)n,r})^T$. These are subvectors of the $\boldsymbol{\eta}_1$, $\boldsymbol{\vartheta}$, and $\boldsymbol{t}$ corresponding to $\mathcal{Z}_r$. Also, denote $[\boldsymbol{\mathcal{Z}}^d]_r = \operatorname{diag}(\mathcal{Z}_{1,r}, \ldots, \mathcal{Z}_{n,r})$ where superscript d means diagonal matrix. Using the ADMM algorithm to minimize the objective function (2.3), after some simple calculations, we have the following closed-form expressions:

$$
\begin{aligned}
\hat{\boldsymbol{\eta}}_1 = \arg\min_{\boldsymbol{\eta}_1} \Bigg\{ & 0.5 \Bigg(\boldsymbol{y} - \sum_{r=1}^{q_1} [\boldsymbol{\mathcal{Z}}^d]_r [\boldsymbol{\eta}_1]_r \Bigg)^T (I - Q) \Bigg(\boldsymbol{y} - \sum_{r=1}^{q_1} [\boldsymbol{\mathcal{Z}}^d]_r [\boldsymbol{\eta}_1]_r \Bigg); \\
& + 0.5\rho \sum_{r=1}^{q_1} \big([\boldsymbol{\vartheta}]_r - \boldsymbol{v}[\boldsymbol{\eta}_1]_r + [\boldsymbol{t}]_r/\rho \big)^T \big([\boldsymbol{\vartheta}]_r - \boldsymbol{v}[\boldsymbol{\eta}_1]_r + [\boldsymbol{t}]_r/\rho \big) \Bigg\}. \qquad (3.3) \\
\hat{\boldsymbol{\vartheta}} = \arg\min_{\boldsymbol{\vartheta}} \Bigg[& \sum_{i<j} p_\gamma(|\boldsymbol{\vartheta}_{ij}|, \lambda) + 0.5\rho (\boldsymbol{\vartheta} - \boldsymbol{v}\boldsymbol{\eta} + \boldsymbol{t}/\rho)^T (\boldsymbol{\vartheta} - \boldsymbol{v}\boldsymbol{\eta} + \boldsymbol{t}/\rho) \Bigg],
\end{aligned}
$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{Z}_0 = (z_{01}, \ldots, z_{0n})^T$, and $Q = \boldsymbol{Z}_0 (\boldsymbol{Z}_0^T \boldsymbol{Z}_0)^{-1} \boldsymbol{Z}_0^T$. In addition, $\boldsymbol{v} = \{\boldsymbol{e}_i - \boldsymbol{e}_j, i < j\}$ is defined by the operation $\{\boldsymbol{a}_{ij}, i < j\} = (a_{12}, \ldots, a_{1n}, \ldots, a_{(n-1)n})^T$, and $\boldsymbol{e}_i$ is an n-dimensional vector of all zeros, except the ith element equal to 1. Here, $\tau > 0$ is a tuning parameter that determines the convergence rate of the ADMM algorithm which is usually set at a fixed value; for example, $\tau = 1$ for simplicity. Then, the $(s+1)$th iteration in the ADMM algorithm proceeds with following updates:

$$
\begin{aligned}
[\boldsymbol{\eta}_1^{(s+1)}]_r = {} & \big\{ \big([\boldsymbol{\mathcal{Z}}^d]_r \big)^T (I - Q) [\boldsymbol{\mathcal{Z}}^d]_r + \rho \boldsymbol{v}^T \boldsymbol{v} \big\}^{-1} \big\{ \big([\boldsymbol{\mathcal{Z}}^d]_r \big)^T (I - Q) \boldsymbol{y}_r^{*(s)} + \boldsymbol{v}^T [\boldsymbol{t}^{(s)}]_r \\
& + \rho \boldsymbol{v}^T [\boldsymbol{\vartheta}^{(s)}]_r \big\}, \quad \text{with } \boldsymbol{y}_r^{*(s)} = \boldsymbol{y} - \sum_{k \neq r}^{q_1} [\boldsymbol{\mathcal{Z}}^d]_k [\boldsymbol{\eta}_1^{(s)}]_k, \text{ for } r = 1, \ldots, q_1; \\
\boldsymbol{\vartheta}_{ij}^{(s+1)} = {} & \begin{cases} \dfrac{S(\boldsymbol{\eta}_{1i}^{(s+1)} - \boldsymbol{\eta}_{1j}^{(s+1)} - \boldsymbol{t}_{ij}^{(s)}/\rho; \lambda/\rho)}{1 - 1/(\gamma\rho)}, & \text{if } \big\| \boldsymbol{\eta}_{1i}^{(s+1)} - \boldsymbol{\eta}_{1j}^{(s+1)} - \boldsymbol{t}_{ij}^{(s)}/\rho \big\|_2 \leq \gamma\lambda; \\ \boldsymbol{\eta}_{1i}^{(s+1)} - \boldsymbol{\eta}_{1j}^{(s+1)} - \boldsymbol{t}_{ij}^{(s)}/\rho, & \text{else}; \end{cases} \\
\boldsymbol{t}_{ij}^{(s+1)} = {} & \boldsymbol{t}_{ij}^{(s)} + \rho \big(\boldsymbol{\vartheta}_{ij}^{(s+1)} - \boldsymbol{\eta}_{1i}^{(s+1)} + \boldsymbol{\eta}_{1j}^{(s+1)} \big), \quad 1 \leq i < j \leq n,
\end{aligned}
$$

where the soft-thresholding operator is $S(\boldsymbol{\alpha}; \kappa) = (1 - \frac{\kappa}{\|\boldsymbol{\alpha}\|_2})_+ \boldsymbol{\alpha}$. With a prefixed K, the ADMM estimates of the subgroup-level effects $\tilde{\boldsymbol{\zeta}}_{1k}, k = 1, \ldots, K$ may be obtained via a classical clustering method, such as the K-means, from the convergent values of $\boldsymbol{\eta}_1^{(s)}$ when the ADMM algorithm stops under a convergence criterion. The detailed derivations of these updates above are given in Section 4 of the Supplementary Material (Zhou et al. (2022)). The algorithmic convergence of the ADMM algorithm can be proved using similar arguments as those given in Sun, Luo and Ye (2015) and Mihić, Zhu and Ye (2021). According to Theorem 1 in Section 4, with a proper selection of the tuning parameter λ the ADMM algorithm provides good initial values with desired theoretical guarantees.

3.3. *Selection of the number of mixture components.* In the context of personal treatment evaluation, the number of mixture components, K, is often chosen by clinicians according to specific scientific hypotheses or objectives. From a fully data-driven perspective, a vast literature provides various methods for the selection of K. According to McLachlan and Peel (2000), empirical criteria may not be able to exhibit differences between a mixture density with K components and one with either fewer than K components or more than K components. In McLachlan and Rathnayake (2014), the "true" order K^* is defined to be the smallest value of K such that the mixture model is compatible with the data. In Theorem 1 we show that, with a proper selection of λ, the number of clusters obtained by the parameter fusion is close to the true value K^* in high probability. Such high-quality initial estimate of K^* can greatly reduce the complexity of final decision on K^* in the subsequent EM algorithm that uses another selection criterion to improve the initially estimated K^*. This double-tuning scheme allows a certain margin of error in the initial estimate of K^* in the ADMM algorithm.

In the EM algorithm we consider two popular ways to determine K, including: (i) Bayesian information criteria (BIC, Schwarz (1978)) or integrated classification criterion (ICL, Biernacki, Celeux and Govaert (2000)) and (ii) hypothesis test, such as likelihood ratio test (LRT). The latter requires the resampling method to obtain the null distribution of LRT statistic which can be computationally burdensome. Li and Chen (2012) proposed an EM based-test for the null hypothesis $H_0 : K = K_0$ vs. $H_A : K > K_0$, for some given positive integer K_0, under a finite normal mixture model. In this paper we choose and implement BIC. The likelihood is given by

$$\ell_n(\boldsymbol{\Theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log \ell(\boldsymbol{\Theta}; y_i), \quad \text{with } \ell(\boldsymbol{\Theta}; y_i) = \sum_{k=1}^{K} \pi_{ik} \phi\big(y_i; \boldsymbol{\mathcal{Z}}_i^T \boldsymbol{\zeta}_{1k} + \boldsymbol{z}_{0i}^T \boldsymbol{\alpha}, \sigma_{ik}^2\big) \tag{3.4}$$

with $\phi(\cdot; \mu, \sigma^2)$ representing a normal density function of mean μ and variance σ^2. Then, the BIC takes the following form: $-2\ell_n(\hat{\boldsymbol{\Theta}}; \boldsymbol{y}) + p \log n$, where $\hat{\boldsymbol{\Theta}}$ is the convergent value from the EM algorithm and $p = q_1 K + q_2(K-1) + q_0 + K + 1$ is the total number of parameters in the model. Theoretical guarantees for the use of BIC are well documented in the literature; Roeder and Wasserman (1997) and Keribin (2000) showed the selection consistency for BIC in determining the true number of components in a mixture model; see also Dasgupta and Raftery (1998) on confirmatory results of selection consistency for BIC in mixture models.

For the sake of comparison, we also consider the ICL criterion, proposed by Biernacki, Celeux and Govaert (2000), for the mixture model, $-2\ell_n(\hat{\boldsymbol{\Theta}}; \boldsymbol{Y}) + p \log n + EN(\hat{\boldsymbol{\gamma}})$, where $EN(\hat{\boldsymbol{\gamma}}) = -\sum_{k=1}^{K} \sum_{i=1}^{n} \hat{\gamma}_{ik} \log \hat{\gamma}_{ik}$, with $\hat{\gamma}_{ik}$ being the convergent posterior probability in (3.2) from the EM algorithm. The performances of BIC and ICL criteria are reported through some additional simulation studies in Table 9 in the Supplementary Material (Zhou et al. (2022)).

4. Theoretical guarantees. In this section we establish the algorithmic convergence of the proposed hybrid algorithm HOSA, estimation consistency of the ADMM estimator, and asymptotic normality of the HOSA estimator $\hat{\boldsymbol{\Theta}}^{\text{HOSA}}$, given by the convergent values from the HOSA algorithm. We begin with regularity conditions in which true values are denoted by superscript $*$; for example, K^* denotes the true number of subgroups, and so on. Let $S_i(\boldsymbol{\Theta}; y)$ be the score function of subject i given by

$$S_i(\boldsymbol{\Theta}; y_i) = \partial \log \ell(\boldsymbol{\Theta}; y_i) / \partial \boldsymbol{\Theta}, \tag{4.1}$$

where $\ell(\boldsymbol{\Theta}, y)$ is the marginal likelihood for subject i given in (3.4). The dimension of the score function is $p^* = K^* q_1 + q_0 + (K^* - 1) q_2 + K^* + 1$:

(C1) All errors, within-set or cross-set, $\{\varepsilon_i, i = 1, \ldots, n\}$ and $\{v_{ik}, i = 1, \ldots, n\}$, $k = 1, \ldots, K^*$ are independent.

(C2) Confounding and prescriptive covariates (i.e., z_0 and z_1) satisfy $\|\mathbb{E}(z_0^T z_1 | x = 1)\|_\infty \le 1/(\pi_1\sqrt{K^*})$, where $\pi_1 = P(x = 1)$, and $\|a\|_\infty := \max_j |a_j|$ for $a = (a_1, a_2, \ldots, a_p)^T$.

(C3) There exist two constants c_1 and c^1 such that $0 < c_1 \le \min_k \mathbb{E}(\gamma_{ik}^*) < \max_k \mathbb{E}(\gamma_{ik}^*) \le c^1 < 1$, where $\gamma_{ik}^* = \pi_{ik}^* f_{ik}(\theta^*)/(\sum_{k=1}^{K^*} \pi_{ik}^* f_{ik}(\theta^*))$ and $\pi_{ik}^* = \pi_{iK}^* \exp(z_{2i}^T \zeta_{2k}^*)$, $k = 1, \ldots, K^* - 1$.

(C4) Θ^* is the true value that maximizes the Q-funcion $Q(\Theta \mid \Theta^*)$, given in (3.1).

(C5) The information matrix $i(\Theta) = \mathbb{E}\{S_i(\Theta; y) S_i^T(\Theta; y)\}$ is positive definite for $\Theta \in B(\Theta^*, c_2) := \{\Theta : \|\Theta - \Theta^*\|_2 \le c_2\}$, a compact neighborhood of Θ^* with some radius c_2, where $S_i(\Theta; y)$ is the score function of subject i.

Conditions (C1) and (C3)–(C5) are routinely assumed in the literature of MLMs. It is easy to see that the score function is unbiased, namely, $\mathbb{E}[S_i(\Theta^*; y)] = 0$, and bounded over the compact set centered at the true value Θ^*; that is, there exists a positive constant c_3 such that $\|\mathbb{E} S_i(\Theta; y)\|_2 < c_3$ for $\Theta \in B(\Theta^*, c_2)$.

Condition (C2) is a very mild technical condition required for the convergence rate of the EM algorithm. Essentially, it requires a limited overlap between the two covariate vectors z_0 and z_1 among subjects receiving treatment $x = 1$. In a randomized trial with $\pi_1 = 0.5$, when both z_0 and z_1 are normalized with mean 0 and variance 1, condition (C2) becomes $\max_{s \ne t} \operatorname{corr}(z_{0s}, z_{1t} | x = 1) \le 2/\sqrt{K^*}$. If there is at least one covariate appearing in both z_0 and z_1, the maximum correlation equals to 1. In this case, condition (C2) allows to divide the subjects in the treatment "$x = 1$ arm" into four subgroups or less ($K^* \le 4$). In most of clinical trials, two to three subgroups are of interest, where condition (C2) automatically holds.

Theorem 1 below shows that ADMM algorithm can help consistently estimate interaction effects $\zeta_1 = (\zeta_{11}, \ldots, \zeta_{1K})$. Let $|\mathcal{G}_k|$ be the cardinality of $\mathcal{G}_k$ and $|\mathcal{G}_{\min}| = \min_k |\mathcal{G}_k|$. Let $\tau = \min_{k \ne k'} \|\zeta_{1k}^* - \zeta_{1k'}^*\|_2$, for $k, k' = 1, \ldots, K^*$, which represents the minimum difference of distinct true values ζ_{1k}^* between subgroups. Denote $\|\alpha\|_\infty \equiv \max_j |\alpha_j|$, and $a \gg b$ represents $a^{-1} b = o(1)$. Throughout this paper we denote two positive constants c_0 and $a_0 > 1$ such that $\int_0^x (1 - s/(a_0\lambda))_+ \, ds$ is a constant for all $x \ge c_0\lambda$.

THEOREM 1. *If the minimal difference τ, the tuning parameter λ, and the minimal subgroup size $|\mathcal{G}_{\min}|$ are bounded below, respectively, by $\tau > c_0\lambda$, $\lambda \gg n^{1/2}/|\mathcal{G}_{\min}|$ and $\sqrt{n \log n} \ll |\mathcal{G}_{\min}|$, then there exists a local minimizer $\tilde{\eta}_1$ of the objective function in (2.3) satisfying*

$$P\big(\|\tilde{\eta}_1 - \eta_1^*\|_\infty \le \psi_n\big) \to 1, \quad \text{as } n \to \infty,$$

where $\psi_n = c_4\sqrt{n \log n}/|\mathcal{G}_{\min}|$ and c_4 is a certain positive constant.

According to Theorem 1, if the minimal subgroup size diverges at a polynomial rate of the form, $|\mathcal{G}_{\min}| = O(n^\nu)$ with $0.5 < \nu \le 1$, then $\|\tilde{\eta}_1 - \eta_1^*\|_\infty = O_p(n^{-\nu+0.5}) = o_p(1)$. Technically speaking, to ensure $\tilde{\eta}_1$ to be a consistent estimator, the size of every subgroup needs to increase as the total sample size n increases. In this way, estimates from the ADMM algorithm would get closer to their true subgroup-level values when the sample size increases. It is worth pointing out that these initial values from the ADMM algorithm are not the final solutions, and thus in practice some marginal errors are allowed and further overcome by the subsequent operation of the EM algorithm.

Next, Theorem 2 shows the convergence rate of the EM algorithm which depends on the rate of the initial values. It is worth noting that the ADMM algorithm provides fusion-regularized estimates of η_1, from which subgroups $\mathcal{G}_k$ as well as estimated subgroup labels

δ_i may be directly extracted from the ADMM estimates $\tilde{\boldsymbol{\eta}}_1$. Other model parameters σ_k^2, σ_ε^2, $\boldsymbol{\alpha}$ can be consequently estimated once the estimated group structures are given.

THEOREM 2. *If conditions* (*C*1)–(*C*4) *hold, then there exists a constant* $0 < \kappa_1 < 1$ *such that, for any initial* $\boldsymbol{\Theta}^{(0)}$ *with* $\|\boldsymbol{\Theta}^{(0)} - \boldsymbol{\Theta}^*\|_2 = o_p(1)$, *updates* $\boldsymbol{\Theta}^{(t)}$ *from the* t*-th iteration of the EM algorithm satisfies the following inequality*:

$$\left\|\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^*\right\|_2 \le (\kappa_1)^t \left\|\boldsymbol{\Theta}^{(0)} - \boldsymbol{\Theta}^*\right\|_2 + c_5 n^{-1/2}, \quad \text{for } t = 1, 2, \ldots,$$

where $c_5 > 0$ *is a constant*.

It is clear that the ADMM initial estimates have a lower rate (i.e., $O_p(n^{-\nu+0.5})$) than the parametric rate $n^{-1/2}$ which is, however, boosted by the EM algorithm via the factor $(\kappa_1)^t$. As $t \to \infty$, the EM estimates eventually reach the rate of the second term, that is, $n^{-1/2}$, leading to an efficient parametric inference. For ease of illustration, let us consider an example with $n = 200$. If $\kappa_1 = 0.8$ and $|\mathcal{G}_{\min}| = n^{3/4}$, then the first term in the upper bound of Theorem 2 satisfies $(\kappa_1)^t \|\boldsymbol{\Theta}^{(0)} - \boldsymbol{\Theta}^*\|_2 \le 10^{-5}$ after 52 iterations or so, reaching the desired parametric rate $\|\boldsymbol{\Theta}^{(52)} - \boldsymbol{\Theta}^*\|_2 = O_p(n^{-1/2})$. If a smaller $\kappa_1 = 0.6$ appears, on average, only 23 iterations are needed to find the desirable solution. In the case of a larger minimal size of subgroup $|\mathcal{G}_{\min}| = n/2$, about 50 iterations are needed. Clearly, the smaller κ_1 the smaller number of iterations is needed to reach the $n^{-1/2}$ rate. Overall, the computational time depends more on the boosting factor κ_1 than the minimal size of subgroup $|\mathcal{G}_{\min}|$. This is the theoretical basis for the use of the EM algorithm to improve the ADMM estimates as far as statistical inference concerns. This boosting phenomenon is also illustrated numerically in the simulation studies in Section 5. Thus, combining Theorems 1 and 2, we have the convergence rate of the proposed HOSA algorithm in Corollary 1.

COROLLARY 1. *Under the conditions of Theorem* 1, *if the minimal subgroup size is of order* $|\mathcal{G}_{\min}| = O(n^\nu)$ *with* $0.5 < \nu \le 1$, *the proposed* $\hat{\boldsymbol{\Theta}}^{\text{HOSA}}$ *satisfies the upper bound in Theorem* 2 *with initial* $\boldsymbol{\Theta}^{(0)}$ *from the ADMM algorithm*.

To perform statistical inference, we establish the asymptotic normality for the HOSA estimator in Theorem 3.

THEOREM 3. *Under the conditions* (*C*1–*C*5) *the proposed HOSA estimator is asymptotically normally distributed, namely,*

$$n^{1/2}(\hat{\boldsymbol{\Theta}}^{\text{HOSA}} - \boldsymbol{\Theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{i}^{-1}(\boldsymbol{\Theta}^*)), \quad \text{as } n \to \infty, \tag{4.2}$$

where the Fisher information $\boldsymbol{i}(\boldsymbol{\Theta}^*) = \mathbb{E}\{S_i(\boldsymbol{\Theta}^*; y_i) S_i^T(\boldsymbol{\Theta}^*; y_i)\}$ *with the score* $S_i(\boldsymbol{\Theta}; y_i)$ *given in* (4.1).

All proofs of Theorems 1–3 are given in Section 5 of the Supplementary Material (Zhou et al. (2022)).

5. Simulation results. We conduct extensive simulation experiments to assess the performance of the proposed HOSA method in the following categories: (i) effectiveness of initial values from the ADMM algorithm, (ii) convergence rate of the EM algorithm with the initial values from the ADMM algorithm, (iii) robustness of HOSA estimation against misspecified membership model or absence of membership model, and (iv) sensitivity of BIC and ICL on the selection of the number of subgroups. We compare our HOSA method to

three existing methods, including (a) ADMM method alone from which we assess any further improvement from the use of the EM algorithm, (b) the EM algorithm with all initial values being set at their true values (EM.T), a gold standard with the better initialization than that by the ADMM algorithm, and (c) the LCMM estimation obtained from the R package **lcmm**. In Section 1 of the Supplementary Material (Zhou et al. (2022)), we further compare our HOSA method to the EM algorithm initialized by three additional types of initial values of interaction effects $\boldsymbol{\zeta}_{1k}$ in the hope to evaluate the stability of the HOSA method.

The performance evaluation concerns estimation bias (BIAS), empirical standard error (ESE), and square root of mean square error (RMSE) of $\hat{\boldsymbol{\zeta}}_{1k}$, $\hat{\boldsymbol{\zeta}}_{2k}$ and $\hat{\boldsymbol{\alpha}}$. In addition, to compare the clustering accuracy we report random index (RI) that measures the performance of pairwise fusion, regularized in the ADMM, as well as the probability of correct clustering (PCC), defined in a similar spirit to the probability of correct classification (Dobbin and Simon (2007), Sanchez et al. (2016)). That is,

$$\text{PCC}(k) = P(\hat{\delta} = k \mid \delta = k)P(\delta = k) + P(\hat{\delta} \neq k \mid \delta \neq k)P(\delta \neq k), \tag{5.1}$$

where $\hat{\delta}$ is an estimated cluster membership and the first and second terms correspond, respectively, to the sensitivity and specificity. An empirical PCC is calculated as follows. First, when the EM algorithm stops, we obtain the estimated posterior probability of subject i belonging to subgroup k, $\hat{\gamma}_{ik}$, from equation (3.2). Second, we estimate a subgroup membership for subject i as the one with the largest probability, namely, $\hat{\delta}_i = \arg\max_k \hat{\gamma}_{ik}$. Third, we estimate the probabilities in (5.1) by the corresponding sample proportions; for example, the estimated sensitivity is

$$\widehat{P}(\hat{\delta} = k \mid \delta = k) = \frac{\sum_{i=1}^{n} I[\text{subject } i \text{ assigned to cluster } k \text{ by } \hat{\delta}_i] \times I[\text{subject } i \text{ is in cluster } k]}{\sum_{i=1}^{n} I[\text{subject } i \text{ is in cluster } k]},$$

where $I[\cdot]$ is the indicator function. All simulation experiments are based on sample size $n = 200$ and $B = 200$ replicates.

Simulation experiment 1. The first simulation experiments is designed to examine the first two performance categories (i) and (ii). Data are simulated from the following SGEM model with three subgroups ($K^* = 3$):

$$\begin{aligned} y_i &= z_{0i}\alpha + \sum_{k=1}^{3} I(\delta_i = k)x_i(\zeta_{1k,1} + z_{1i}\zeta_{1k,2} + \upsilon_{ik}) \\ &\quad + \sum_{k=1}^{3} I(\delta_i = k)x_i\upsilon_{ik} + \varepsilon_i, \\ \log P(\delta_i = k)/P(\delta_i = 3) &= \zeta_{2k,1} + z_{2i}\zeta_{2k,2}, \quad k = 1, 2, \end{aligned}$$

where covariates $(z_{0i}, z_{1i}, z_{2i}) \overset{\text{iid}}{\sim} \mathcal{N}_3(0, I)$, variance components $\upsilon_{ik} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_k^2)$, $k = 1, 2, 3$ with $\sigma_k \equiv \sigma \in \{0.01, 0.1, 1, 2, 4\}$, and errors $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon \in \{0.001, 0.01, 0.1, 0.5, 1\}$. Here, we consider a continuous exposure $x_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and a binary treatment $x_i \overset{\text{iid}}{\sim} B(0.5)$, respectively. Clearly, condition (C2) is satisfied at $K^* = 3$. The true parameter values are set at $\alpha = 0.4$, $\boldsymbol{\zeta}_{11} = (-1, -1)^T$ (strongly negative), $\boldsymbol{\zeta}_{12} = (1, 1)^T$ (somewhat positive), $\boldsymbol{\zeta}_{13} = (3, 3)^T$ (strongly positive), $\boldsymbol{\zeta}_{21} = (-1, 1)^T$, and $\boldsymbol{\zeta}_{22} = (-1, -1)^T$, with the respective parameter dimensions equal to $q_0 = 1$, $q_1 = q_2 = 2$. As the subgroup variances σ_k^2 increase, these three subgroups of parameters become less separable, and, consequently, it is harder

to estimate the subgroup labels. To avoid redundancy, we select to report part of numerical results to illustrate the aforementioned comparisons in the simulation experiment.

Tables 1–3 list the simulation results for continuous and binary x_i, respectively. It is evident that our HOSA method has very comparable performances in all performance categories to the gold standard EM.T in all cases considered in both simulation experiments; both methods are much better than the existing ADMM and LCMM methods. In particular, LCMM has showed poor performances on the estimation of the parameters in the subgroup membership models. Clearly, the degree of separability σ_k^2 is a key driving factor to the performances of the four competing methods, and the higher separability (or smaller σ_k^2) the better quality of fit. Overall, HOSA and EM.T outperform ADMM and LCMM in the subgroup membership prediction, judged by the values of RI, sensitivity, specificity, and PCC, especially in the cases of low separability. It is interesting to note that, although LCMM has unstable estimation of the parameters $\zeta_{2k,1}$ and $\zeta_{2k,2}$ in the membership models, its estimation of the cluster memberships appears reasonable, leading to fair PCC, especially in the cases of high separability; see additional results of the clustering accuracy for the other two separability cases, $\sigma_k = 0.01, \sigma_\varepsilon = 1$ and $\sigma_k = 1, \sigma_\varepsilon = 0.1$ in Tables 10 and 11 in Section 1 of the Supplementary Material (Zhou et al. (2022)).

To further demonstrate the performances of HOSA method over more challenging scenarios with large σ_k^2 (or low separability) and with large σ_ε^2 (or high noise), Figures 1 and 2 display various boxplots of HOSA estimates over 200 rounds of simulation with binary x_i. In the case of increasing noise σ_e in errors, the point estimates of the regression parameters are reasonably close to their true values (e.g., the middle row of Figure 1), indicating ignorable estimation biases, while the quality of cluster fusion measured by RI worsens. In the case of decreasing separability, all regression parameters are estimated well, except the parameters of interaction, $\zeta_{11,2}$, one randomly selected from $\zeta_{11,k}$, $k = 1, 2, 3$ for display. HOSA estimate $\hat{\zeta}_{11,2}$ gets more estimation bias as the degree of subgroup separability drops substantially, and, in the meanwhile, the pairwise fusion accuracy (or RI) decreases noticeably. These patterns in Figures 1 and 2 are representative to all parameter estimates, including those that are not shown in the two figures.

Simulation experiment 2. The second simulation study concerns the sensitivity of HOSA to misspecified membership models or the absence of the membership models. We consider the following models for the mixture proportions:

$$\pi_{ik} = \left\{ \frac{\exp(\boldsymbol{z}_{2i}^T \boldsymbol{\zeta}_{2k})}{1 + \sum_{k=1}^{K-1} \exp(\boldsymbol{z}_{2i}^T \boldsymbol{\zeta}_{2k})} \right\}^2, \quad k = 1, 2,$$

and the membership probabilities are then $P(\delta_i = k|\boldsymbol{z}_{2i}) = \pi_{ik} / \sum_{k=1}^{K} \pi_{ik}$, $k = 1, 2, 3$ which are used in the data simulation. In the actual analysis we treat the square function as a linear function, namely, $\text{logit}(\pi_{ik}) = \boldsymbol{z}_{2i}^T \boldsymbol{\zeta}_{2k}$, leading to a situation of misspecified models (MS). Another situation of misspecified models is the case of $\boldsymbol{z}_{2i} = \varnothing$; that is, no covariates enter in the membership models, resulting in a situation referred to as the absence of membership models (WP). In the analyses we fix the variance of errors at $\sigma_\varepsilon = 1$, while we vary the separability with high separability $\sigma_k^2 = 0.1$ (S1) and low separability $\sigma_k^2 = 4$ (S2). This design results in six types of scenarios, as shown in Figure 3, with a binary treatment variable x_i. For example, scenario WP.S2 refers to the HOSA-based analysis with the absence of membership model under the low separability, and the HOSA estimates under CS.S1 and CS.S2 are deemed the best in the comparisons because the correct membership model is used. It is interesting to note from the (2,1)th panel that HOSA estimates of the interaction effects, that is. $\zeta_{11,2}$, the key parameters of interest in the analyses, are not sensitive to the specifications of the membership models, while with no surprise the parameter in the (2,2)th panel, $\zeta_{21,2}$,

TABLE 1

Summary results of BIAS, ESE, RMSE, and RI under a continuous exposure x_i over 200 replicates obtained from the four competing methods

		$\sigma_k = 0.01, \sigma_\varepsilon = 0.1$				$\sigma_k = 0.01, \sigma_\varepsilon = 1$			
Para.		HOSA	ADMM	EM.T	LCMM	HOSA	ADMM	EM.T	LCMM
α	BIAS	0.000	0.002	0.000	0.001	0.003	−0.003	0.004	0.004
	ESE	0.008	0.048	0.007	0.009	0.080	0.117	0.079	0.081
	RMSE	0.008	0.048	0.007	0.009	0.080	0.117	0.079	0.081
$\zeta_{11,2}$	BIAS	0.003	0.941	0.000	0.032	−0.008	1.037	0.002	0.138
	ESE	0.019	0.347	0.016	0.256	0.186	0.524	0.177	0.456
	RMSE	0.019	1.003	0.016	0.258	0.186	1.162	0.177	0.477
$\zeta_{12,2}$	BIAS	0.000	0.384	−0.002	−0.000	0.005	0.417	0.016	0.127
	ESE	0.019	0.188	0.018	0.030	0.231	0.261	0.228	0.403
	RMSE	0.019	0.428	0.018	0.030	0.231	0.492	0.229	0.423
$\zeta_{13,2}$	BIAS	−0.001	−0.661	0.000	−0.018	0.005	−0.770	0.001	−0.055
	ESE	0.012	0.195	0.010	0.146	0.115	0.339	0.113	0.249
	RMSE	0.012	0.689	0.010	0.147	0.115	0.841	0.113	0.255
$\zeta_{21,2}$	BIAS	0.004	–	0.009	2.016	0.074	–	0.056	5.459
	ESE	0.269	–	0.271	26.817	0.432	–	0.426	48.824
	RMSE	0.269	–	0.271	26.893	0.439	–	0.430	49.128
$\zeta_{22,2}$	BIAS	−0.007	–	−0.008	2.031	−0.169	–	−0.119	−15.384
	ESE	0.273	–	0.272	26.960	0.649	–	0.578	108.912
	RMSE	0.273	–	0.272	27.037	0.671	–	0.590	109.993
$\zeta_{23,2}$	BIAS	0.000	–	0.000	−1.043	0.000	–	0.000	−5.388
	ESE	0.000	–	0.000	16.120	0.000	–	0.000	38.555
	RMSE	0.000	–	0.000	16.154	0.000	–	0.000	38.930
RI		0.923	0.586	0.922	0.919	0.702	0.572	0.702	0.668
		$\sigma_k = 1, \sigma_\varepsilon = 0.1$				$\sigma_k = 1, \sigma_\varepsilon = 1$			
Para.		HOSA	ADMM	EM.T	LCMM	HOSA	ADMM	EM.T	LCMM
α	BIAS	−0.003	−0.000	−0.016	−0.003	−0.001	−0.005	−0.000	−0.001
	ESE	0.023	0.054	0.031	0.026	0.093	0.117	0.092	0.095
	RMSE	0.024	0.054	0.035	0.026	0.093	0.117	0.092	0.095
$\zeta_{11,2}$	BIAS	0.251	1.073	0.001	0.132	0.040	1.116	−0.010	0.334
	ESE	0.310	0.417	0.086	0.464	0.425	0.559	0.365	0.710
	RMSE	0.399	1.151	0.086	0.482	0.427	1.248	0.365	0.785
$\zeta_{12,2}$	BIAS	0.170	0.413	0.009	0.118	0.080	0.434	0.015	0.279
	ESE	0.269	0.219	0.082	0.450	0.564	0.284	0.458	0.607
	RMSE	0.318	0.467	0.083	0.466	0.570	0.519	0.458	0.668
$\zeta_{13,2}$	BIAS	−0.072	−0.649	−0.003	−0.106	−0.007	−0.797	−0.016	−0.152
	ESE	0.141	0.252	0.053	0.323	0.267	0.379	0.244	0.468
	RMSE	0.158	0.696	0.053	0.340	0.267	0.882	0.245	0.492
$\zeta_{21,2}$	BIAS	−0.043	–	0.037	4.854	0.079	–	0.120	10.304
	ESE	0.377	–	0.360	46.683	0.609	–	0.562	51.785
	RMSE	0.379	–	0.362	46.935	0.614	–	0.575	52.800
$\zeta_{22,2}$	BIAS	−0.161	–	−0.069	0.479	−2.889	–	−2.283	−33.984
	ESE	0.720	–	0.471	34.802	21.118	–	18.716	145.145
	RMSE	0.738	–	0.476	34.805	21.315	–	18.855	149.070
$\zeta_{23,2}$	BIAS	0.000	–	0.000	0.032	0.011	–	0.000	−2.677
	ESE	0.000	–	0.000	0.591	0.166	–	0.000	48.611
	RMSE	0.000	–	0.000	0.592	0.167	–	0.000	48.684
RI		0.759	0.567	0.766	0.738	0.658	0.562	0.658	0.595

TABLE 2

Empirical sensitivity, specificity, and PCC of the four competing methods for continuous and binary exposure x_i in the most $\sigma_k = 0.01$, $\sigma_\varepsilon = 0.1$) and least ($\sigma_k = 1$, $\sigma_\varepsilon = 1$) separability scenarios over 200 replicates

		$\sigma_k = 0.01, \sigma_\varepsilon = 0.1$				$\sigma_k = 1, \sigma_\varepsilon = 1$			
True δ	Method	$\hat{\delta} = 1$	$\hat{\delta} = 2$	$\hat{\delta} = 3$	PCC	$\hat{\delta} = 1$	$\hat{\delta} = 2$	$\hat{\delta} = 3$	PCC
				Continuous x_i					
1	HOSA	0.945	0.017	0.038	0.973	0.655	0.111	0.234	0.852
	ADMM	0.468	0.532	0.000	0.862	0.437	0.524	0.039	0.814
	EM.T	0.944	0.017	0.039	0.973	0.655	0.094	0.250	0.854
	LCMM	0.939	0.022	0.039	0.971	0.645	0.179	0.176	0.770
2	HOSA	0.022	0.904	0.074	0.960	0.087	0.554	0.358	0.783
	ADMM	0.000	1.000	0.000	0.617	0.121	0.779	0.100	0.540
	EM.T	0.021	0.904	0.075	0.960	0.081	0.528	0.391	0.791
	LCMM	0.021	0.904	0.075	0.956	0.189	0.531	0.280	0.722
3	HOSA	0.015	0.023	0.961	0.953	0.081	0.157	0.762	0.733
	ADMM	0.000	0.509	0.491	0.755	0.018	0.553	0.429	0.688
	EM.T	0.015	0.024	0.961	0.952	0.078	0.132	0.790	0.735
	LCMM	0.015	0.029	0.957	0.951	0.191	0.243	0.566	0.677
				Binary x_i					
1	HOSA	0.978	0.010	0.011	0.991	0.744	0.107	0.148	0.881
	ADMM	0.567	0.433	0.000	0.891	0.526	0.465	0.009	0.849
	EM.T	0.981	0.008	0.011	0.992	0.737	0.103	0.159	0.881
	LCMM	0.976	0.013	0.011	0.990	0.693	0.201	0.106	0.766
2	HOSA	0.007	0.963	0.030	0.984	0.096	0.602	0.302	0.798
	ADMM	0.000	1.000	0.000	0.685	0.111	0.830	0.060	0.582
	EM.T	0.007	0.965	0.028	0.984	0.092	0.596	0.312	0.806
	LCMM	0.008	0.965	0.028	0.981	0.233	0.545	0.222	0.716
3	HOSA	0.004	0.010	0.986	0.983	0.065	0.156	0.779	0.782
	ADMM	0.000	0.421	0.579	0.794	0.005	0.522	0.473	0.725
	EM.T	0.004	0.010	0.986	0.984	0.064	0.137	0.799	0.785
	LCMM	0.004	0.015	0.981	0.981	0.204	0.254	0.542	0.700

can only be estimated accurately under the properly specified membership models. In this panel, estimation biases are evident from MS.S1 and MS.S2 because of misspecified membership models in the analyses. In addition, with the low degree of separability all standard errors appear larger than those obtained under the high degree of separability.

Simulation experiment 3. As pointed out above, in the context of personal medicine the number of mixture components is typically chosen a priori by practitioners, based on a scientific hypothesis or analysis objective. However, in other settings with the absence of such prior knowledge, certain data-driven methods may be invoked, such as BIC and ICL, as suggested in Section 3.3. Table 9 in Section 1 of the Supplementary Material (Zhou et al. (2022)) reports the proportion of correctly selected number of mixture components, that is, $K^* = 3$, via criteria BIC and ICL. The simulation results show that BIC and ICL perform reasonably well when separability is not too low, and that overall BIC outperforms ICL.

6. Application. We illustrate our SGEM modeling methodology via two real-world data examples. In this section we present the analysis of the motivating data collected from a randomized calcium supplementation trial; the second analysis, concerning a randomized trial on treatments for type 2 diabetics, is included in Section 2 of the Supplementary Material (Zhou et al. (2022)). The aim of the calcium supplementation trial is to study if calcium

TABLE 3

Summary results of BIAS, ESE, RMSE, and RI under a binary treatment x_i over 200 replicates obtained from the four competing methods

		$\sigma_k = 0.01, \sigma_\varepsilon = 0.1$				$\sigma_k = 0.01, \sigma_\varepsilon = 1$			
Para.		HOSA	ADMM	EM.T	LCMM	HOSA	ADMM	EM.T	LCMM
α	BIAS	0.001	0.001	0.001	0.001	0.005	0.006	0.007	0.006
	ESE	0.007	0.010	0.007	0.007	0.075	0.096	0.074	0.075
	RMSE	0.007	0.010	0.007	0.007	0.075	0.096	0.074	0.075
$\zeta_{11,2}$	BIAS	0.001	0.990	0.000	0.007	0.050	1.089	0.025	0.213
	ESE	0.021	0.523	0.020	0.101	0.302	0.577	0.269	0.552
	RMSE	0.021	1.120	0.020	0.101	0.306	1.233	0.270	0.592
$\zeta_{12,2}$	BIAS	0.000	0.451	−0.000	0.001	0.002	0.462	0.009	0.163
	ESE	0.024	0.270	0.023	0.031	0.363	0.300	0.323	0.477
	RMSE	0.024	0.526	0.023	0.031	0.363	0.551	0.323	0.504
$\zeta_{13,2}$	BIAS	−0.002	−0.609	−0.002	−0.007	−0.025	−0.621	−0.025	−0.092
	ESE	0.016	0.273	0.016	0.079	0.182	0.346	0.185	0.332
	RMSE	0.016	0.667	0.016	0.080	0.184	0.711	0.186	0.344
$\zeta_{21,2}$	BIAS	0.049	–	0.049	0.184	0.157	–	0.108	7.250
	ESE	0.356	–	0.350	0.531	0.544	–	0.479	44.072
	RMSE	0.359	–	0.354	0.562	0.566	–	0.491	44.664
$\zeta_{22,2}$	BIAS	−0.052	–	−0.052	1.491	−0.310	–	−0.203	−18.671
	ESE	0.375	–	0.362	19.973	1.009	–	0.810	113.304
	RMSE	0.379	–	0.366	20.028	1.055	–	0.835	114.833
$\zeta_{23,2}$	BIAS	0.000	–	0.000	−0.173	0.000	–	0.000	−4.980
	ESE	0.000	–	0.000	4.385	0.000	–	0.000	46.675
	RMSE	0.000	–	0.000	4.389	0.000	–	0.000	46.940
RI		0.971	0.645	0.972	0.969	0.767	0.613	0.770	0.713
		$\sigma_k = 1, \sigma_\varepsilon = 0.1$				$\sigma_k = 1, \sigma_\varepsilon = 1$			
Para.		HOSA	ADMM	EM.T	LCMM	HOSA	ADMM	EM.T	LCMM
α	BIAS	0.001	0.001	0.001	0.001	0.005	0.006	0.005	0.006
	ESE	0.010	0.016	0.010	0.010	0.088	0.097	0.088	0.088
	RMSE	0.010	0.016	0.010	0.010	0.088	0.097	0.088	0.088
$\zeta_{11,2}$	BIAS	0.467	1.115	−0.006	0.404	0.066	1.136	−0.000	0.433
	ESE	0.481	0.597	0.108	0.718	0.512	0.672	0.427	0.780
	RMSE	0.670	1.265	0.109	0.824	0.516	1.320	0.427	0.892
$\zeta_{12,2}$	BIAS	0.307	0.480	0.005	0.199	0.161	0.481	0.070	0.291
	ESE	0.338	0.315	0.100	0.629	0.593	0.338	0.497	0.672
	RMSE	0.457	0.574	0.100	0.660	0.614	0.588	0.502	0.732
$\zeta_{13,2}$	BIAS	−0.141	−0.605	−0.003	−0.135	−0.024	−0.629	−0.010	−0.122
	ESE	0.193	0.328	0.069	0.424	0.301	0.409	0.276	0.497
	RMSE	0.239	0.688	0.070	0.445	0.302	0.750	0.277	0.512
$\zeta_{21,2}$	BIAS	0.010	–	0.204	3.974	0.112	–	0.319	12.868
	ESE	0.515	–	0.806	52.250	0.835	–	2.346	72.479
	RMSE	0.515	–	0.831	52.401	0.843	–	2.368	73.613
$\zeta_{22,2}$	BIAS	−3.234	–	−0.256	−1.695	−6.397	–	−4.297	−39.868
	ESE	20.778	–	1.008	78.565	31.710	–	25.062	157.761
	RMSE	21.029	–	1.041	78.583	32.349	–	25.427	162.721
$\zeta_{23,2}$	BIAS	0.000	–	0.000	−2.369	−0.012	–	0.000	−2.110
	ESE	0.000	–	0.000	35.252	0.095	–	0.000	25.759
	RMSE	0.000	–	0.000	35.332	0.096	–	0.000	25.845
RI		0.757	0.613	0.776	0.702	0.703	0.592	0.705	0.609

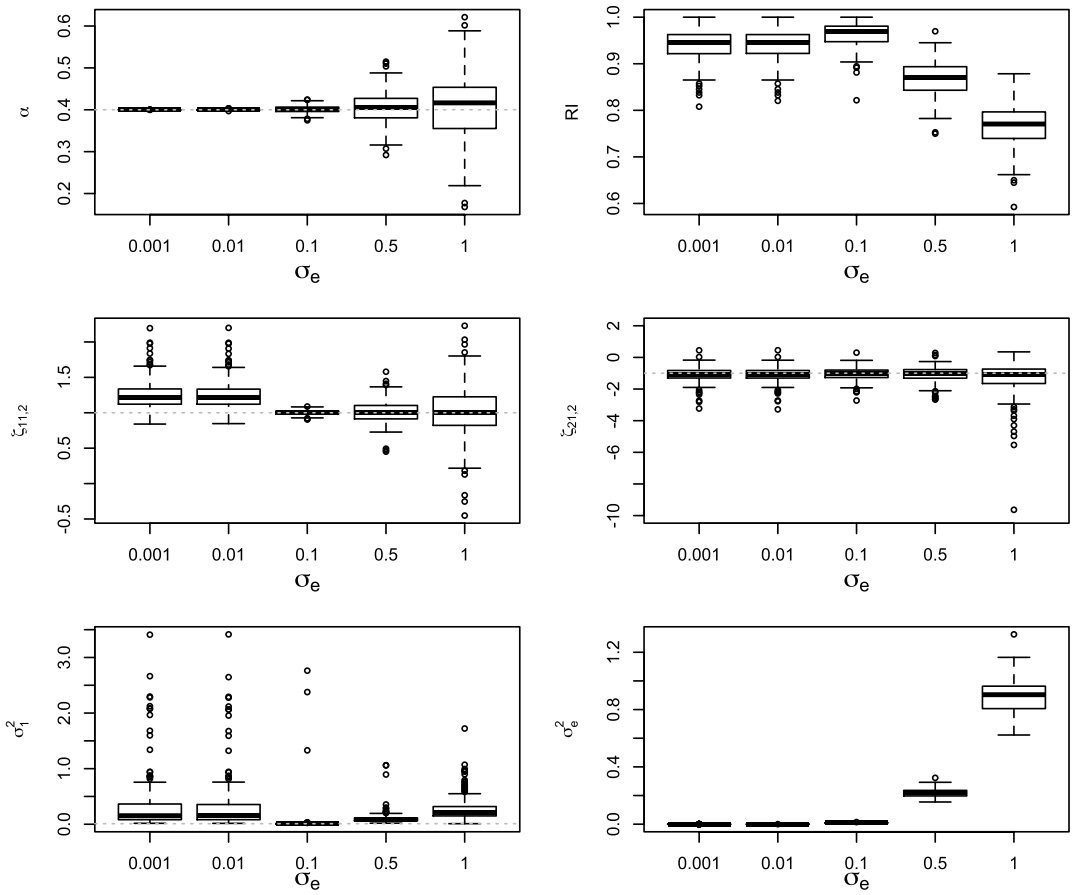

FIG. 1. *Boxplots of HOSA estimates of randomly selected regression parameters* α, $\zeta_{11,2}$, $\zeta_{21,2}$, *and variance parameters* σ_1^2 *and* σ_ε^2 *as well as RI with increasing noise in errors at* $\sigma_\varepsilon = 0.001, 0.01, 0.1, 0.5, 1$ *and fixed separability at* $\sigma_k = 0.1$, *respectively, and with a binary treatment* x_i.

supplement in daily diet is effective to alleviate maternal blood lead (PB) concentration so to benefit children's neurobehavioral and cognitive development (Ettinger, Hu and Hernandez-Avila (2007), Ettinger et al. (2009)). Here, we focus on the outcome of PB concentration at the third trimester (i.e., y), as this late period of pregnancy is brimming with fast development of neurons and wiring in the brain (Ackerman (1992)). In particular, the weight of baby's brain roughly triples, and its formerly once smooth surface becomes increasingly grooved and indented.

We use the data from a total of 351 (n) Mexican women in the analysis, after removing a handful of subjects with missing data. The treatment variable x (CA) is coded as 1 for calcium supplementation and 0 for placebo (i.e., calcium intake from food only). Consulting our collaborators, we choose a set of mother's baseline characterizations in the analysis, including age (in year), weight (in kg), education (the number of years in school), marriage (1 yes, 0 no), parity (number of pregnancies), baseline dietary calcium intake (bDCA), and baseline PB concentration at the first semester (bPB). These covariates are normalized to be mean zero and variance 1. Choosing $z_0 = (1, \text{age, weight, education, marriage, parity, bDCA})^T$, we begin with a linear model with no subgroups of the following form:

$$y_i = z_{0i}^T \boldsymbol{\alpha} + \beta_1 x_i + \beta_2 (x_i \times bDCA_i) + \varepsilon_i, \tag{6.1}$$

where errors $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Table 4 reports the results of this analysis, where the estimated population-average effect of calcium supplementation, $\hat{\beta}_1$, appears marginally insignificant

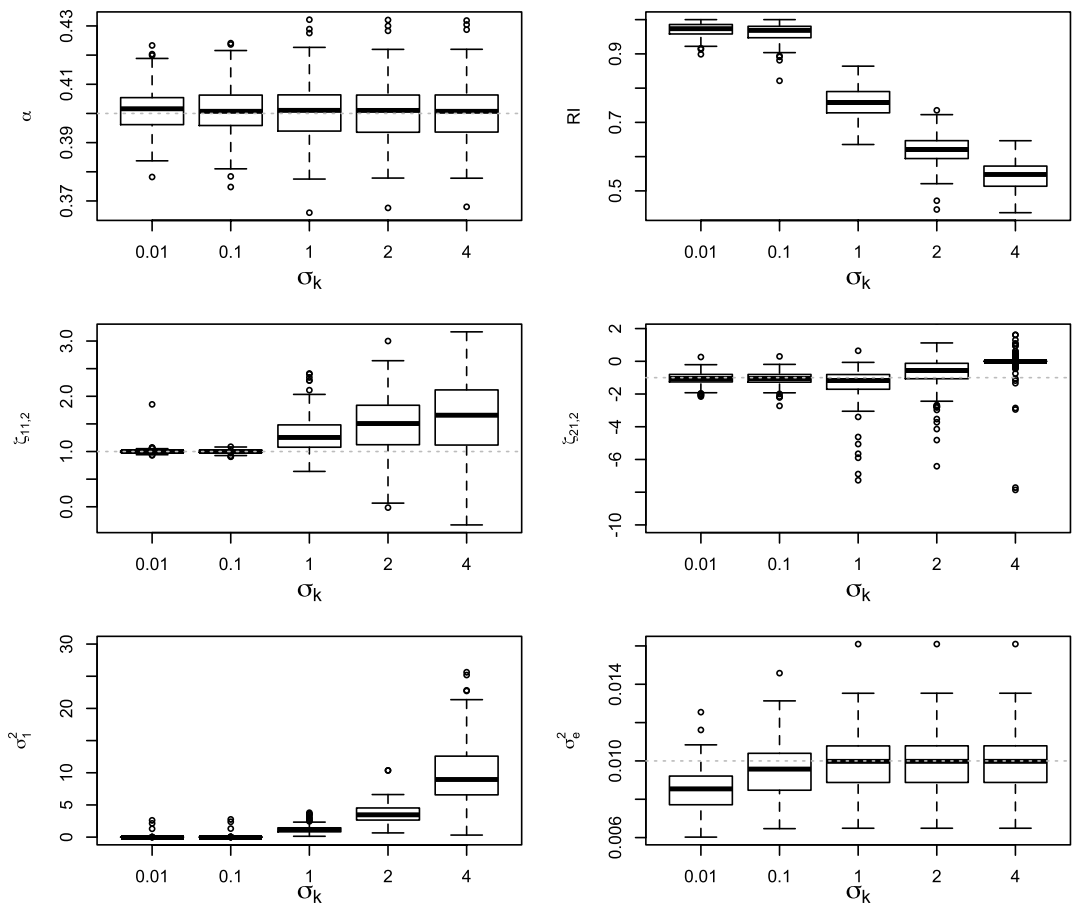

FIG. 2. *Boxplots of HOSA estimates of randomly selected regression parameters* α, $\zeta_{11,2}$, $\zeta_{21,2}$, *and variance parameters* σ_1^2 *and* σ_ε^2 *as well as RI with decreasing subgroup separability at* $\sigma_k = 0.01, 0.1, 1, 2, 4$ *and fixed* $\sigma_\varepsilon = 0.1$, *respectively, and with a binary treatment* x_i.

with $p = 0.075$ in reducing the PB concentration during the third trimester of pregnancy. The baseline dietary calcium intake (bDCA) is not significant in both main effect (α_6) and interaction effect with the treatment (β_2). Mother's age (α_1) is the only baseline covariate that is significantly positively correlated with the PH concentration.

From a perspective of precision nutrition, we would like to ask a question of clinical importance: although the population-average effect of calcium supplementation is marginally insignificant, whether or not, is there a subgroup of mothers who may experience significant benefit from calcium supplementation to reduce their third trimester PB concentration? This

TABLE 4
Estimated population average effects of calcium supplementation and baseline covariates with no subgroups

		Estimate	p-value			Estimate	p-value
int	α_0	5.049	0.000				
age	α_1	0.507	0.010	marriage	α_4	0.218	0.215
weight	α_2	−0.152	0.384	parity	α_5	−0.371	0.064
education	α_3	−0.285	0.132	bDCA	α_6	−0.188	0.404
CA	β_1	−0.609	0.075	CA × bDCA	β_2	0.315	0.365

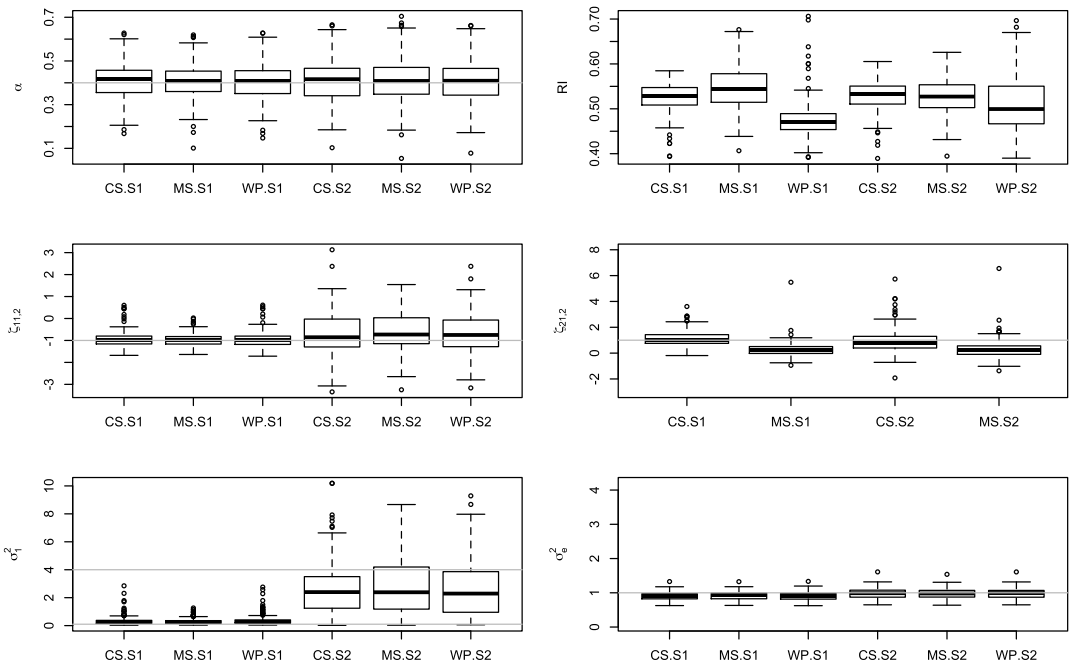

FIG. 3. *Boxplots of the HOSA estimates of randomly selected regression parameters* α, $\zeta_{11,2}$, $\zeta_{21,2}$, *and the variance parameters* σ_ε^2 *and* σ_k^2 *as well as RI, respectively, with the membership models being correctly specified* (*CS*), *misspecified* (*MS*), *or absent* (*WP*), *and with variance of errors* $\sigma_\varepsilon = 1$ *and degree of separability* $\sigma_k^2 = 0.1$ (*S*1) *or* 4 (*S*2), *and with a binary treatment* x_i.

question may be answered using our proposed SGEM toolbox. Among several subgroup-effects models used in the analysis, below we present two representative ones to illustrate the usefulness of the SGEM methodology. We also used model selection via hypothesis testing to reach parsimonious models for the subgroup means.

First, we present a SGEM with only intercept term in z_1, namely, the subgroup-level effect μ_k does not dependent on any covariates, $\mu_k = \zeta_{1k,1}$, $k = 1, \ldots, K$. In this case, because $\mathrm{Cov}(z_0, z_1) = 0$, condition (C2) is satisfied. Let $z_2 = (1, \text{age}, \text{bPB})^T$. The outcome model is $y_i = z_{0i}^T \boldsymbol{\alpha} + \beta_i x_i + \varepsilon_i$, with $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$, where

$$\beta_i \overset{\text{ind}}{\sim} \sum_{k=1}^{K} \pi_{ik} \mathcal{N}(\mu_k, \sigma_k^2), \quad \text{and} \quad \log(\pi_{ik}/\pi_{iK}) = z_{2i}^T \boldsymbol{\zeta}_{2k}, \quad k = 1, \ldots, K-1. \tag{6.2}$$

Using the BIC in the EM algorithm, $K = 2$ is chosen, implying that there exist two subgroups of women taking the calcium supplementation. Consequently, the membership model in (6.2) becomes a classical logistic model. The left block of Table 5 lists the parameter estimates, standard errors (SE) and p-values calculated via 1000 bootstrap samples, using our HOSA method, as well as the estimates obtained by the R package **lcmm**.

According to Table 5, the estimated group-level effects of subgroup 1, $\hat{\mu}_1 = \hat{\zeta}_{11,1} = -1.782$ (p-value < 0.001) and of subgroup 2, $\hat{\mu}_2 = \hat{\zeta}_{12,1} = 2.424$ (p-value < 0.001), calcium supplementation shows a significant benefit to the first subgroup but no benefit to the second subgroup. It is known in the literature that overdosed calcium during pregnancy is associated with some adverse health outcomes, such as pregnancy loss (Norman, Politz and Politz (2009)). Thus, it is of clinical interest to determine subgroup memberships of the study participants for benefit or for harm. Through variable selection, two covariates, age and bPB, are found to be predictive to the group membership, as shown by a scatterplot of these subjects

TABLE 5

Results from two subgroup-effects models obtained by HOSA and lcmm methods, where p-values are obtained by 1000 *bootstrap samples*

		No modeling of μ_k				Modeling of μ_k with bDCA			
		Estimate	SE	p-value	lcmm.est	Estimate	SE	p-value	lcmm.est
int	α_0	5.065	0.213	0.000	2.417	5.055	0.205	0.000	5.064
age	α_1	0.648	0.178	0.000	0.507	0.538	0.168	0.001	0.649
weight	α_2	−0.096	0.150	0.522	−0.147	−0.084	0.147	0.569	−0.107
education	α_3	−0.221	0.161	0.169	−0.276	−0.217	0.156	0.164	−0.213
marriage	α_4	0.137	0.147	0.352	0.220	0.121	0.142	0.394	0.145
parity	α_5	−0.299	0.173	0.084	−0.354	−0.329	0.165	0.047	−0.302
bDCA	α_6	−0.135	0.152	0.376	−0.051	−0.172	0.197	0.383	−0.168
CA	$\zeta_{11,1}$	−1.782	0.295	0.000	−139.799	−1.453	0.262	0.000	−1.816
CA×bDCA	$\zeta_{11,2}$	–	–	–	−−	0.217	0.307	0.480	0.048
	$\zeta_{12,1}$	2.424	0.583	0.000	145.049	4.468	0.752	0.000	2.123
	$\zeta_{12,2}$	–	–	–	−−	−0.778	0.742	0.294	0.179
int	$\zeta_{21,1}$	1.673	0.833	0.045	–	2.431	0.449	0.000	1.504
age	$\zeta_{21,2}$	1.562	0.778	0.045	–	0.230	0.284	0.418	1.701
bPB	$\zeta_{21,3}$	−4.950	2.139	0.021	–	−2.248	0.439	0.000	−5.391

in Figure 4 with estimated subgroup labels colored in blue (benefit subgroup) and red (harm subgroup), respectively. It is interesting to see that subjects with lower PB concentration at the first trimester would benefit the calcium supplementation to reduce the PB concentration in the third trimester. Also, calcium supplementation is slightly more beneficial for older women. The results, obtained from the R package **lcmm,** are included in Table 5. The estimates $\hat{\boldsymbol{\alpha}}$ in the outcome model are comparable to those obtained from our HOSA method; the estimates of μ_k have the same directions but are too large to be trustful in comparison to the HOSA estimates. It is worth pointing out that age is found to be significant by HOSA, in agreement with the result in Table 4.

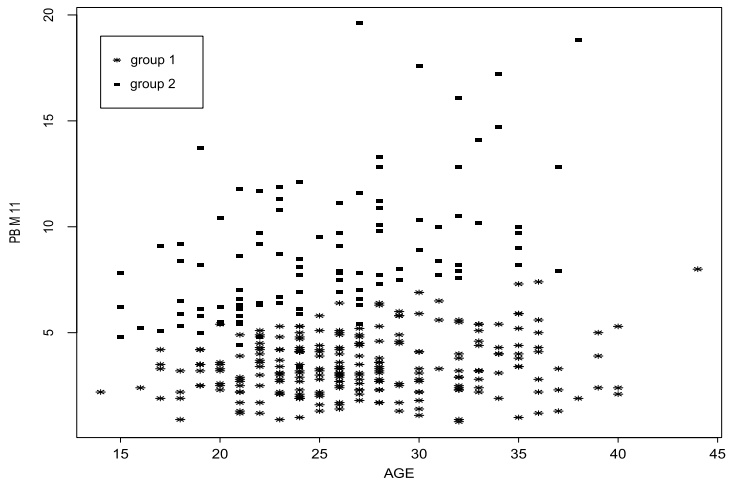

FIG. 4. *Scatterplot of PB concentration at the first trimester* (*bPB*) *vs. age of subjects whose memberships are determined by the higher posterior probability* $\hat{\gamma}_{ik}$ *from the EM algorithm. Blue stars denote the members of the benefit subgroup, and red squares denote the members of harm subgroup.*

Second, we present a SGEM with the subgroup-level treatment effects being modeled by covariates. We choose the scenario of μ_k being a function of the baseline calcium intake from food (bDCA) because an interaction term of CA and bDCA is considered in the previous model (6.1). This SGEM provides more details than the results seen in Table 4, due to the inclusion of $z_1 = (1, \text{bDCA})^T$ in the group mean models $\mu_{ik} = z_1^T \zeta_{1k}$. Once again, condition (C2) is satisfied because z_0 and z_1 share only one common covariate, bDCA. The results by both HOSA and lcmm methods are reported in the right block of Table 5. The HOSA results from the left block and the right block appear similar, so are the lcmm results. Clearly, the two interaction parameters, $\zeta_{11,2}$ (p-value $= 0.48$) and $\zeta_{12,2}$ (p-value $= 0.294$), are not significant, so the results from the left block are used to draw conclusions. Finally, the R package **lcmm** gives comparable estimates to the HOSA estimates; include those in the logistic model for the subgroup memberships.

7. Discussion. Although the population-average treatment effect is of primary interest in clinical trials, patients treated by a drug can experience different treatment efficacy. Profiling those patients who have strong treatment benefit from those having weak, or even no benefit, would provide a tailored therapeutic protocol in clinical practice. Relevant statistical methods for such an analytic task are of great importance. In this paper we have achieved two goals for the subgroup analysis: First, we introduced a class of subgroup-effects models with clear and direct clinical interpretations; second, we developed a new hybrid statistical algorithm, termed HOSA, that generates more stable and reliable numerical results in both estimation and inference. These technical advances, given in the paper, are of great importance to deliver a better statistical toolbox for the evaluation of personal treatment effects. As demonstrated by both extensive simulation studies and real-world data analyses, our proposed HOSA method has shown satisfactory performances in various easy and hard situations, and has outperformed some popular existing methods, including exiting ADMM and R package **lcmm**. In the case of low subgroup separability, neither method seems to work well in the prediction of subgroup labels. A natural solution to address this issue in practice is that we may simply merge those subgroups with significant overlap and then conduct subgroup analyses with a reduced number of subgroups. We will develop an effective procedure for subgroup merging in the future work to overcome this challenge, including a certain hypothesis testing based approach.

The key methodological contribution in this paper is to utilize the power of supervised clustering analysis via the ADMM algorithm to yield high-quality initial values that are desired inputs to begin the EM algorithm. The step of refinement through the EM algorithm is necessary because the resulting parameter estimation achieves the parametric rate, leading to the same efficiency of the maximum likelihood estimation. In contrast, the ADMM estimation has not yet reached the parametric rate and suffers from potential estimation bias, so the ADMM method alone is not ready to make statistical inference. Note that conducting inference, *say* obtaining p-value for treatment effect, is of highly clinical interest. Thus, the proposed HOSA method, which upgrades the ADMM solution by the EM algorithm, enjoys both numerical stability and validity of statistical inference with clear theoretical guarantees shown in this paper. In comparison to the existing LCMM method that has shown unstable estimation for the parameters in the membership model, the HOSA is our recommended method to be used in clinical studies.

This proposed HOSA method may be extended in several directions. Developing a systematic procedure to determine which covariates enter the set of confounding covariates or the set of prescriptive covariates is of importance in the model specification. Two models of clinical importance are the logistic outcome model for categorical outcomes and the Cox proportional hazards model for time-to-event outcomes, and developing HOSA on these nonlinear regression models is certainly practically useful. These are worth further exploration in the future work.

Acknowledgments. Part of the research was done when Zhou and Sun were postdoctoral research fellows at the Department of Biostatistics, University of Michigan. They appreciate the computing and other logistic support provided by the University of Michigan.

Funding. Zhou's research was partially supported by Fund of National Natural Science (Nos. 11901470, 11931014, 11571282, 11829101) and by Fundamental Research Funds for the Central Universities (Nos. JBK190904 and JBK1806002). Sun's research was partially supported by the National Natural Science Foundation of China (No. 61902319 and No.82122061). Song's research was partially supported by a National Institutes of Health grant R01ES024732 and National Science Foundation grants DMS1811734 and DMS2113564.

SUPPLEMENTARY MATERIAL

Supplement to "Subgroup-effects models for the analysis of personal treatment effects" (DOI: 10.1214/21-AOAS1503SUPPA; .pdf). We provide additional simulation results, the analysis of a diabetes clinical trial data, derivations of the ADMM and EM algorithms, and details of technical proofs.

Supplement to "Subgroup-effects models for the analysis of personal treatment effects" (DOI: 10.1214/21-AOAS1503SUPPB; .zip). The R package is also available on GitHub https://github.com/sqsun/HOSA.

REFERENCES

ACKERMAN, S. (1992). *Discovering the Brain*. National Academies Press, Washington.

ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15** 2773–2832. MR3270750

BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.* **45** 77–120. MR3611487 https://doi.org/10.1214/16-AOS1435

BELLINGER, D., STILES, K. and NEEDLEMAN, H. (1992). Low-level lead exposure, intelligence and academic achievement: A long-term follow-up study. *Pediatrics* **90** 855–861.

BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 719–725.

BOIVIN, M. and GIORDANI, B. (1995). A risk evaluation of the neuropsychological effects of childhood lead toxicity. *Dev. Neuropsychol.* **11** 157–180.

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.

CHAGANTY, A. T. and LIANG, P. (2013). Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning* 1040–1048.

CHEN, J., LI, P. and FU, Y. (2012). Inference on the order of a normal mixture. *J. Amer. Statist. Assoc.* **107** 1096–1105.

CHEN, Y., YI, X. and CARAMANIS, C. (2018). Convex and nonconvex formulations for mixed regression with two components: Minimax optimal rates. *IEEE Trans. Inf. Theory* **64** 1738–1766. MR3766312 https://doi.org/10.1109/TIT.2017.2773474

CHI, E. C. and LANGE, K. (2015). Splitting methods for convex clustering. *J. Comput. Graph. Statist.* **24** 994–1013. MR3432926 https://doi.org/10.1080/10618600.2014.948181

CRAWFORD, S. L. (1994). An application of the Laplace method to finite mixture distributions. *J. Amer. Statist. Assoc.* **89** 259–267. MR1266298

DASGUPTA, A. and RAFTERY, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *J. Amer. Statist. Assoc.* **93** 294–302.

DEB, P. and HOLMES, A. M. (2000). Estimates of use and costs of behavioural health care: A comparison of standard and finite mixture models. *Health Econ.* **9** 475–489.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DOBBIN, K. and SIMON, R. (2007). Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics* **8** 101–117.

ELHAMIFAR, E. and VIDAL, R. (2013). Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 2765–2781.

ETTINGER, A., HU, H. and HERNANDEZ-AVILA, M. (2007). Dietary calcium supplementation to lower blood lead levels in pregnancy and lactation. *J. Nutr. Biochem* **18** 172–178.

ETTINGER, A., LAMADRID-FIGUEROA, H., TELLEZ-ROJO, M., MERCADO-GARCIA, A., PETERSON, K., SCHWARTZ, J., HU, H. and HERNANDEZ-AVILA, M. (2009). Effect of calcium supplementation on blood lead levels in pregnancy: A randomized placebo-controlled trial. *Environ. Health Perspect.* **117** 26–31.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models. Springer Series in Statistics*. Springer, New York. MR2265601

GABAY, D. and MERCIER, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.* **2** 17–40.

GLOWINSKI, R. (2014). On alternating direction methods of multipliers: A historical perspective. In *Modeling, Simulation and Optimization for Science and Technology. Comput. Methods Appl. Sci.* **34** 59–82. Springer, Dordrecht. MR3330832 https://doi.org/10.1007/978-94-017-9054-3_4

GRÜN, B. and LEISCH, F. (2007). Applications of finite mixtures of regression models. Available at http://cran.r-project.org/web/packages/flexmix/vignettes/regression-examples.pdf.

GUNTER, L., ZHU, J. and MURPHY, S. A. (2011). Variable selection for qualitative interactions. *Stat. Methodol.* **8** 42–55. MR2741508 https://doi.org/10.1016/j.stamet.2009.05.003

HU, H., TÉLLEZ-ROJO, M., BELLINGER, D., SMITH, D., ETTINGER, A., LAMADRID-FIGUEROA, H., SCHWARTZ, J., SCHNAAS, L., MERCADO-GARCIA, A. et al. (2006). Fetal lead exposure at each stage of pregnancy as a predictor of infant mental development. *Environ. Health Perspect.* **114** 1730–1735.

JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. and HINTON, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput.* **3** 79–87. https://doi.org/10.1162/neco.1991.3.1.79

KERIBIN, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A* **62** 49–66. MR1769735

MA, S. and HUANG, J. (2017). A concave pairwise fusion approach to subgroup analysis. *J. Amer. Statist. Assoc.* **112** 410–423. MR3646581 https://doi.org/10.1080/01621459.2016.1148039

MCLACHLAN, G. J., LEE, S. X. and RATHNAYAKE, S. I. (2019). Finite mixture models. *Annu. Rev. Stat. Appl.* **6** 355–378. MR3939525 https://doi.org/10.1146/annurev-statistics-031017-100325

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models. Wiley Series in Probability and Statistics: Applied Probability and Statistics*. Wiley Interscience, New York. MR1789474 https://doi.org/10.1002/0471721182

MCLACHLAN, G. J. and RATHNAYAKE, S. (2014). On the number of components in a Gaussian mixture model. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **4** 341–355.

MIHIĆ, K., ZHU, M. and YE, Y. (2021). Managing randomization in the multi-block alternating direction method of multipliers for quadratic optimization. *Math. Program. Comput.* **13** 339–413. MR4266928 https://doi.org/10.1007/s12532-020-00192-5

MOTA, J. F. C., XAVIER, J. M. F., AGUIAR, P. M. Q. and PÜSCHEL, M. (2013). D-ADMM: A communication-efficient distributed algorithm for separable optimization. *IEEE Trans. Signal Process.* **61** 2718–2723. MR3053838 https://doi.org/10.1109/TSP.2013.2254478

MUTHÉN, B. and SHEDDEN, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55** 463–469.

NORMAN, J., POLITZ, D. and POLITZ, L. (2009). Hyperparathyroidism during pregnancy and the effect of rising calcium on pregnancy loss: A call for earlier intervention. *Clin. Endocrinol.* **71** 104–109.

PERNG, W., TAMAYO-ORTIZ, M., TANG, L., SANCHEZ, B., CANTORAL, A., MEEKER, J., DOLINOY, D., ROBERTS, E., MIER, A. et al. (2019). The Early Life Exposure in Mexico to Environmental Toxicants (ELEMENT) Project. *British Med. J. Open* **9** e030427.

PIMENTEL-ALARCÓN, D., BALZANO, L., MARCIA, R., NOWAK, R. and WILLETT, R. (2017). Mixture regression as subspace clustering. In *Sampling Theory and Applications* (*SampTA*), 2017 *International Conference on* 456–459. IEEE, New York.

PROUST, C. and JACQMIN-GADDA, H. (2005). Estimation of linear mixed models with a mixture of distribution for the random effects. *Comput. Methods Programs Biomed.* **78** 165–173.

PROUST-LIMA, C., PHILIPPS, V. and LIQUET, B. (2017). Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *J. Stat. Softw.* **78** 1–56.

REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. MR0738930 https://doi.org/10.1137/1026034

ROEDER, K. and WASSERMAN, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92** 894–902. MR1482121 https://doi.org/10.2307/2965553

SANCHEZ, B., WU, M., SONG, P. and WANG, W. (2016). Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics* **17** 722–736.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SEDGHI, H., JANZAMIN, M. and ANANDKUMAR, A. (2016). Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics* 1223–1231.

SOLTANOLKOTABI, M., ELHAMIFAR, E. and CANDÈS, E. J. (2014). Robust subspace clustering. *Ann. Statist.* **42** 669–699. MR3210983 https://doi.org/10.1214/13-AOS1199

SUN, R., LUO, Z.-Q. and YE, Y. (2015). On the expected convergence of randomly permuted ADMM. Preprint. Available at arXiv:1503.06387.

TEICHER, H. (1961). Identifiability of mixtures. *Ann. Math. Stat.* **32** 244–248. MR0120677 https://doi.org/10.1214/aoms/1177705155

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641 https://doi.org/10.1111/j.1467-9868.2005.00490.x

VERBEKE, G. and LESAFFRE, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *J. Amer. Statist. Assoc.* **91** 217–221.

VIELE, K. and TONG, B. (2002). Modeling with mixtures of linear regressions. *Stat. Comput.* **12** 315–330. MR1951705 https://doi.org/10.1023/A:1020779827503

WEI, S. and KOSOROK, M. R. (2013). Latent supervised learning. *J. Amer. Statist. Assoc.* **108** 957–970. MR3174676 https://doi.org/10.1080/01621459.2013.789695

WONG, C. H., SIAH, K. W. and LO, A. W. (2019). Estimation of clinical trial success rates and related parameters. *Biostatistics* **20** 273–286. MR3922133 https://doi.org/10.1093/biostatistics/kxx069

WU, C.-F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103. MR0684867 https://doi.org/10.1214/aos/1176346060

XU, W. and HEDEKER, D. (2001). A random-effects mixture model for classifying treatment response in longitudinal clinical trials. *J. Biopharm. Statist.* **11** 253–273.

YI, X., CARAMANIS, C. and SANGHAVI, S. (2014). Alternating minimization for mixed linear regression. In *International Conference on Machine Learning* 613–621.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHANG, Y., CHEN, X., ZHOU, D. and JORDAN, M. I. (2016). Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *J. Mach. Learn. Res.* **17** Paper No. 102, 44. MR3543508

ZHONG, K., JAIN, P. and DHILLON, I. S. (2016). Mixed linear regression with multiple components. In *Advances in Neural Information Processing Systems* 2190–2198.

ZHOU, L., SUN, S., FU, H. and SONG, P. X. (2022). Supplement to "Subgroup-Effects Models for the Analysis of Personal Treatment Effects." https://doi.org/10.1214/21-AOAS1503SUPPA, https://doi.org/10.1214/21-AOAS1503SUPPB