Statistica Sinica **27** (2017), 711-731
doi: https://doi.org/10.5705/ss.202015.0371

PENALIZED LIKELIHOOD FOR LOGISTIC-NORMAL MIXTURE MODELS WITH UNEQUAL VARIANCES

Juan Shen[1], Yingchuan Wang[2] and Xuming He[2]

[1]*Fudan University and* [2]*University of Michigan*

Abstract: Subgroup analysis with unspecified subgroup memberships has received increasing attention in recent years. In Shen and He (2015), a structured logistic-normal mixture model was proposed to characterize the subgroup distributions and the subgroup membership simultaneously, but under the assumption that the subgroups differ only in the means. In this paper, we consider a penalized likelihood approach for more general cases with heterogeneous subgroup variances. Despite substantial technical complications in the development of the statistical theory, we show that the penalized likelihood inference for the existence of subgroups and for the estimation of subgroup membership can be carried out in the existing framework. Empirical results with a simulation study and two data examples demonstrate the usefulness of the proposed method.

Key words and phrases: EM algorithm; heterogeneous components; homogeneity test; likelihood ratio test; mixture models; subgroup identification.

1. Introduction

Subgroup analysis is important in clinical trials and market segmentation. In recent years, the extraction of unknown subgroups with distinct response patterns to a treatment has gained increasing popularity. Much of the early research in subgroup analysis has focused on pre-specified subgroups (Simon (2002), Song and Chi (2007), and Altstein, Li and Elashoff (2011), among others). Su et al. (2009) introduced an interaction tree procedure to obtain subpopulations with heterogeneous treatment effects across subpopulations. Foster, Taylor and Ruberg (2011) proposed the "Virtual Twins" method to identify a subgroup for the binary response in a randomized clinical trial. A parametric scoring system based on multiple covariates, Cai et al. (2011) and Zhao et al. (2013), assists in assigning treatments to new patients. Lipkovich et al. (2011) and Lipkovicha and Dmitrienkoa (2014) provided a recursive partitioning method for treatment assignments to patient subpopulations. Berger, Wang and Shen (2014) proposed a Bayesian method for subgroup analysis of multiple subgroups defined by a

binary predictive variable. Kang, Janes and Huang (2014) relied on a novel boosting algorithm to choose an optimal treatment. Besides interaction models, methods based on mixture models have been proposed in Shen and He (2015) and Van Horn et al. (2015). They showed that regression mixture models can be effective in evaluating differential treatment effects.

A critical concern with various subgroup identification methods is that they tend to identify a subgroup even when no meaningful subgroup exists. Sleight (2000) described subgroup analyses as "fun to look at, but don't believe them". Shen and He (2015) and Fan, Lu and Song (2016) have advocated the use of hypothesis testing for the existence of subgroups.

Shen and He (2015) proposed a structured logistic-mixture model and considered a test based on the iterations in an Expectation-Maximization algorithm for mixture models for the existence of subgroups. The model-based approach has two distinctive features. First, it models simultaneously the subgroup membership and the treatment outcomes within each subgroup as functions of the covariates. Second, it provides a model-based test for the existence of the subgroups with differential treatment effects. Such a test is not generally available with other methods, making false discovery a real risk in subgroup identification.

The results of Shen and He (2015) assume the homogeneity in the subgroup variances, which does not always hold in practice. When the equal subgroup variance assumption is violated, it is unclear whether the EM test loses power and whether the model estimation is biased. The purpose of this paper is to relax the equal variance assumption in the logistic-normal mixture model. Allowing unequal subgroup variances is highly valuable in practice, but brings technical challenges in the theoretical development.

The first difficulty with mixture of normal models with possibly unequal variances is that the likelihood is unbounded and the maximum likelihood estimator (MLE) does not exist. To illustrate the issue, let $Y_1, \cdots, Y_n$ be i.i.d. from a simple normal mixture model

$$\pi_1 N(\theta_1, \sigma_1^2) + (1-\pi_1) N(\theta_2, \sigma_2^2). \tag{1.1}$$

Here the likelood

$$\prod_{i=1}^{n} \Big\{ \frac{\pi_1}{\sqrt{2\pi\sigma_1^2}} \exp\Big(-\frac{(Y_i-\theta_1)^2}{2\sigma_1^2}\Big) + \frac{1-\pi_1}{\sqrt{2\pi\sigma_2^2}} \exp\Big(-\frac{(Y_i-\theta_2)^2}{2\sigma_2^2}\Big)\Big\}$$

goes to infinity by taking $\theta_1 = Y_1$ and letting σ_1 go to zero. To overcome this difficulty, we work with a penalized likelihood following Chen, Tan and Zhang (2008). We propose a data-driven strategy to select the penalty parameter

that maximizes the potential *overall* subgroup effect and provide the asymptotic theory for the penalized likelihood estimator and its associated EM tests.

The rest of this paper is organized as follows. We formulate the structured logistic-normal model with unequal variances in Section 2.1. The penalized likelihood estimation and the corresponding EM tests are proposed and studied from Section 2.2 to Section 2.5, together with a discussion concerning the selection of the penalty parameter in Section 2.6. Simulation studies and case studies are reported in Section 3 and Section 4, respectively, with concluding remarks in Section 5. Proofs for the theorems and additional information in the case studies are presented in the supplemental materials.

2. Methodology

2.1. Structured logistic-normal mixture model with unequal variances

We consider a logistic-normal mixture model that allows unequal variances in each component. For $i = 1, \ldots, n$,

$$\begin{aligned} Y_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, \delta_i &= \boldsymbol{Z}_i^T(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2\delta_i) + \varepsilon_i(\delta_i\sigma_1 + (1-\delta_i)\sigma_2), \\ P(\delta_i = 1 \mid \boldsymbol{X}_i, \boldsymbol{Z}_i) &= \pi(\boldsymbol{X}_i^T\boldsymbol{\gamma}) \equiv \frac{\exp(\boldsymbol{X}_i^T\boldsymbol{\gamma})}{1+\exp(\boldsymbol{X}_i^T\boldsymbol{\gamma})}, \\ P(\delta_i = 0 \mid \boldsymbol{X}_i, \boldsymbol{Z}_i) &= 1 - P(\delta_i = 1|\boldsymbol{X}_i), \end{aligned} \tag{2.1}$$

where n is the sample size, $Y_i \in \mathbb{R}$ is the outcome, $\delta_i \in \{0, 1\}$ is the latent subgroup indicator, $\boldsymbol{Z}_i \in \mathbb{R}^{q_1}$ is the covariate associated with the subgroup mean, $\boldsymbol{X}_i \in \mathbb{R}^{q_2}$ is the baseline covariate associated with the group membership, $\boldsymbol{\beta}_1 \in \mathbb{R}^{q_1}, \boldsymbol{\beta}_2 \in \mathbb{R}^{q_1}, \boldsymbol{\gamma} \in \mathbb{R}^{q_2}$ are the corresponding coefficients, $\varepsilon_i \sim N(0, 1)$ are independent of $\boldsymbol{Z}_i$, $\boldsymbol{X}_i$, and δ_i, and σ_1 and σ_2 are the standard derivations within each subgroup. The first element of $\boldsymbol{X}_i$ and the first component of $\boldsymbol{Z}_i$ are taken to be 1 to allow intercepts in the model, and the second element of $\boldsymbol{Z}_i$ is the treatment indicator. We can have overlapping variables in the random vectors of $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$.

In the two-component model, the overall model parameter is $\boldsymbol{\eta}^T = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma_1, \sigma_2)$. We use $\boldsymbol{\theta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma_1, \sigma_2)$ as the parameters except for $\boldsymbol{\gamma}$. We observe a random sample $\{\boldsymbol{W}_i = (Y_i, \boldsymbol{Z}_i^T, \boldsymbol{X}_i^T), i = 1, \ldots n\}$, but δ_i's are latent variables.

Remark 1. In the model formulation, we assume that the first nonzero component of $\boldsymbol{\beta}_2$ is positive, and in the case of $\boldsymbol{\beta}_2 = 0$ we assume that $\sigma_1 > \sigma_2$, to ensure parameter identifiability. The model is degenerate if $\boldsymbol{\beta}_2 = 0$ and $\sigma_1 = \sigma_2$. In our implementation, we identify the subgroups by taking the second

component of $\boldsymbol{\beta}_2$ (the treatment effect difference) to be positive. We are not concerned with the special case where the treatment effect difference is zero, in which case the identification of subgroups is not practically important.

In the case of $\boldsymbol{X}_i = \boldsymbol{Z}_i$, we can think of the proposed model as a special case of the mixture-of-experts models (Jordan and Jacobs (1994)). This is well studied in machine learning. Here we have distinct and clear interpretations of the variables $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$. In particular, the covariates in $\boldsymbol{X}_i$ are baseline measurements that are available prior to the treatment and can be used to predict subgroup membership for future subjects, while the covariates in $\boldsymbol{Z}_i$ include any variables relevant to the treatment effects within subgroups. The existence of meaningful subgroups with differential treatment effects is our focus.

The proposed model with $\boldsymbol{\gamma} = 0$ has been quite well studied in the literature; see, for instance, Goeffinet, Loisel and Laurent (1992), Chen, Chen and Kalbfleisch (2001), and Chen and Li (2009). In subgroup analysis, the case of $\boldsymbol{\gamma} = 0$ is rather uninteresting, because even if subgroups exist, no covariates are informative for predicting the subgroup membership; an important feature of the proposed model is to characterize subgroup membership given the baseline covariates $\boldsymbol{X}$.

2.2. The Penalized likelihood

Unlike the equal variance case, even when Model (2.1) is well-defined, the maximum likelihood estimator does not exist, and the key is to restrict the two variances from being too close to zero. We consider penalty terms $p_n(\sigma_1)$ and $p_n(\sigma_2)$ on the scale parameters. In particular, we use

$$p_n(\sigma) = -\lambda[\frac{S_n^2}{\sigma^2} + \log(\frac{\sigma^2}{S_n^2})], \tag{2.2}$$

where S_n^2 is the estimator of σ^2 from the equal variance model, and λ is a tuning parameter. Given any positive λ, $p_n(\sigma)$ achieves its maximum at $\sigma^2 = S_n^2$, and goes to negative infinity as σ approaches zero or infinity.

The penalized log-likelihood is

$$\begin{aligned} pl(\boldsymbol{\eta}; \boldsymbol{W}) = \sum_{i=1}^{n} \log[\sum_{j=0}^{1} f(Y_i|\boldsymbol{Z}_i, \boldsymbol{X}_i, \delta_i = j; \boldsymbol{\beta}_j, \sigma_j) P(\delta_i = j|\boldsymbol{X}_i; \boldsymbol{\gamma})] \\ + p_n(\sigma_1) + p_n(\sigma_2), \end{aligned} \tag{2.3}$$

which is also written as $pl(\boldsymbol{\eta})$ without $\boldsymbol{W}$ later in the paper, and the maximum penalized likelihood estimator of $\boldsymbol{\eta}$ is given by

$$\underset{\boldsymbol{\eta}}{\operatorname{argmax}}\, pl(\boldsymbol{\eta}; \boldsymbol{W}). \tag{2.4}$$

To maximize the penalized likelihood, a slightly modified EM algorithm of Dempster, Laird and Rubin (1977) is described as follows.

At the kth step of the EM iteration, the objective function is

$$Q(\boldsymbol{\eta}|\boldsymbol{\eta}^{(k)}) = \sum_{i=1}^{n} \mathbb{E}_{\delta_i|w_i,\boldsymbol{\eta}^{(k)}} \left\{ I_{(\delta_i=1)} \log \left(\frac{\pi(\boldsymbol{X}_i^T\boldsymbol{\gamma})}{\sigma_1} \exp\left(-\frac{\left(Y_i - \boldsymbol{Z}_i^T(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)\right)^2}{2\sigma_1^2} \right) \right) \right.$$
$$\left. + I_{(\delta_i=0)} \log \left(\frac{1-\pi(\boldsymbol{X}_i^T\boldsymbol{\gamma})}{\sigma_2} \exp\left(-\frac{(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}_2)^2}{2\sigma_2^2} \right) \right) \right\} + p_n(\sigma_1) + p_n(\sigma_2).$$

To evaluate it, the E step involves the calculation of

$$a_i^{(k)} = P(\delta_i = 1 | Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i; \eta^{(k)})$$
$$= \frac{f(Y_i|\delta_i = 1, \boldsymbol{Z}_i; \theta^{(k)}) P(\delta_i=1|\boldsymbol{X}_i; \boldsymbol{\gamma}^{(k)})}{f(Y_i|\delta_i=1, \boldsymbol{Z}_i; \boldsymbol{\theta}^{(k)}) P(\delta_i=1|\boldsymbol{X}_i; \boldsymbol{\gamma}^{(k)}) + f(Y_i|\delta_i=0, \boldsymbol{Z}_i; \boldsymbol{\theta}^{(k)}) P(\delta_i=0|\boldsymbol{X}_i; \boldsymbol{\gamma}^{(k)})},$$

and $b_i^{(k)} = 1 - a_i^{(k)}$, and the M step gives

$$\boldsymbol{\gamma}^{(k+1)} = \underset{\boldsymbol{\gamma}}{\operatorname{argmax}} \sum_{i=1}^{n} a_i^{(k)} \log \pi(\boldsymbol{X}_i^T\boldsymbol{\gamma}) + b_i^{(k)} \log\Big(1 - \pi(\boldsymbol{X}_i^T\boldsymbol{\gamma})\Big),$$

$$(\boldsymbol{\beta}_{12}^{(k+1)}, \sigma_1^{(k+1)}) = \underset{\boldsymbol{\beta}_{12},\sigma}{\operatorname{argmax}} \sum_{i=1}^{n} a_i^{(k)} \log \left(\sigma^{-1} \exp\Big(-\frac{(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}_{12})^2}{2\sigma^2} \Big) \right) + p_n(\sigma),$$

$$(\boldsymbol{\beta}_1^{(k+1)}, \sigma_2^{(k+1)}) = \underset{\boldsymbol{\beta}_1,\sigma}{\operatorname{argmax}} \sum_{i=1}^{n} b_i^{(k)} \log \left(\sigma^{-1} \exp\Big(-\frac{(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}_1)^2}{2\sigma^2} \Big) \right) + p_n(\sigma),$$

and $\boldsymbol{\beta}_2^{(k+1)} = \boldsymbol{\beta}_{12}^{(k+1)} - \boldsymbol{\beta}_1^{(k+1)}$.

In the M step, the updating formula for $\boldsymbol{\theta}^{(k+1)}$ is a weighted least squares solution. For the particular penalty (2.2), the calculations for $\sigma_1^{(k+1)}$ and $\sigma_2^{(k+1)}$ given $\boldsymbol{\beta}_1^{(k+1)}$ and $\boldsymbol{\beta}_2^{(k+1)}$ yield

$$\sigma_1^{(k+1)} = \left(\frac{\sum_{i=1}^n a_i^{(k)} \Big(Y_i - \boldsymbol{Z}_i^T(\boldsymbol{\beta}_1^{(k+1)} + \boldsymbol{\beta}_2^{(k+1)})\Big)^2/2 + \lambda S_n^2}{\sum_{i=1}^n a_i^{(k)}/2 + \lambda} \right)^{1/2},$$

$$\sigma_2^{(k+1)} = \left(\frac{\sum_{i=1}^n b_i^{(k)} (Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}_1^{(k+1)})^2/2 + \lambda S_n^2}{\sum_{i=1}^n b_i^{(k)}/2 + \lambda} \right)^{1/2}.$$

We see that the two variances from the penalized likelihood are weighted sums of S_n^2 and the corresponding estimators without the penalty.

Following Shen and He (2015), we consider testing the null hypothesis of $H_0 : \boldsymbol{\beta}_2 = 0$ and $\sigma_1 = \sigma_2$ against $H_a = H_0^c$ based on the penalized version of

the EM test, hereinafter called the pEM test. If this null hypothesis is rejected, we then proceed to examine the treatment effect as measured by the second component of $\boldsymbol{\beta}_2$. The construction of a confidence interval on the parameter, for example, can be carried out with the standard large-sample approximations. It is possible that we have a mixture of two groups in the model but there are no differential treatment effects. On the other hand, if the null hypothesis H_0 is not rejected, we suggest no further subgroup identification to avoid false discoveries.

Because the null hypothesis represents a singular point of the model space, the standard asymptotic theory does not apply to the likelihood ratio test. The proposed pEM test uses the penalized likelihood evaluated at a finite set of $\boldsymbol{\gamma}$ values.

Given a compact set $\tilde{\Gamma}$ of $\boldsymbol{\gamma}$ whose intercept parameters are bounded away from 0, we randomly choose $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_J \in \tilde{\Gamma}$ for a small number J to form the sets of starting values $\Gamma = \{\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_J\}$. For each $\boldsymbol{\gamma}_j$, we obtain the maximum penalized likelihood estimator of $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_j^{(0)} = \text{argmax}_{\boldsymbol{\theta}} pl(\boldsymbol{\theta}, \boldsymbol{\gamma}_j). \tag{2.5}$$

With the starting value $\boldsymbol{\eta}_j^{(0)} = (\boldsymbol{\theta}_j^{(0)}, \boldsymbol{\gamma}_j)$, we perform the EM iterations described for K times to obtain $\boldsymbol{\eta}_j^{(1)}, \cdots, \boldsymbol{\eta}_j^{(K)}$, where K is a finite integer. If in the EM process, the $\boldsymbol{\gamma}$ value goes beyond $\tilde{\Gamma}$, we stop the iteration so the effective number of iterations might be less than K.

At each $j = 1, 2, \cdots, J$, let

$$\hat{\boldsymbol{\theta}}_0 = \text{argmax}_{\boldsymbol{\theta} \in H_0} pl(\boldsymbol{\theta}, \boldsymbol{\gamma}_j) \tag{2.6}$$

be the maximum penalized likelihood estimator of $\boldsymbol{\theta}$ under the null hypothesis with $\boldsymbol{\gamma} = \boldsymbol{\gamma}_j$, and let

$$pEM_j^{(K)} = 2\Big(pl(\boldsymbol{\eta}_j^{(K)}) - pl(\hat{\boldsymbol{\theta}}_0, \boldsymbol{\gamma}_j)\Big),$$

in which $pl(\cdot)$ is as defined in (2.3). Then we define the penalized EM test statistic as

$$pEM^{(K)} = \max\Big\{pEM_j^{(K)} : j = 1, \ldots, J\Big\}. \tag{2.7}$$

For the choice of J, K, and $\tilde{\Gamma}$, we follow the suggestions made by Shen and He (2015). If q_2 is not very large, a typical choice is $J = \min\{2^{q_2}, 16\}$ and $K = 9$ to limit the computation time without sacrificing quality.

2.3. Convergence of the penalized likelihood estimator

We study the properties of the estimators $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ from (2.4) and (2.6) under

the alternative hypothesis and the null hypotheses, respectively. The consistency of the estimator for $\boldsymbol{\eta}$ under the two-component model assures the validity for further applications of the model, while the consistency of the estimator for $\boldsymbol{\theta}$ under the null hypothesis is needed for developing the limiting distribution of the *pEM* test statistic. We give sufficient conditions on the penalty and on the covariates in C1-C3 and C4-C5, respectively.

Conditions on the penalty.

C1. The penalty $p_n(\sigma) < 0$ almost surely.

C2. For any given constant C, for almost all sample $\omega \in \Omega$, there exists $n_0(\omega)$, such that when $n \geq n_0(\omega)$,

$$\inf \left\{ p_n(\sigma)(\log n)^{-2}(\log \sigma)^{-1} : 0 < \sigma \leq n^{-1} \right\} \geq C.$$

C3. If $\boldsymbol{\beta}_2 = 0$ and $\sigma_1 = \sigma_2 = \sigma_0$, we have $p_n(\sigma_0) = o(n)$ almost surely; otherwise (under the alternative model), $p_n(\sigma_1) = o(n)$ and $p_n(\sigma_2) = o(n)$ almost surely.

Remark 2. Condition C2 basically requires that the penalty should be small when σ is small, and Condition C3 requires that the penalty should not dominate the likelihood function evaluated at the true parameters. These two conditions together guarantee that the penalized likelihood does not attain its maximum when σ is near zero, and, therefore, the estimator of σ stays away from zero. The conditions allow the penalties to depend on the data, which is quite useful in numerical analysis in practice. We discuss how to choose the tuning parameter λ for the penalty in (2.2) in Section 2.6.

Conditions on the covariates. We partition the covariate vector $\boldsymbol{Z}$ as $\boldsymbol{Z}^T = (1, \boldsymbol{U}^T, \boldsymbol{V}^T)$, where 1 corresponds to the intercept, $\boldsymbol{U}$ consists of only discrete variables, and $\boldsymbol{V}$ consists of only continuous variables.

C4. The sample space of $\boldsymbol{U}$ is finite. For any unit vector $\boldsymbol{\alpha}$ of the same dimension as the vector $\boldsymbol{V}$, the conditional distribution of $\boldsymbol{V}^T\boldsymbol{\alpha}|\boldsymbol{U}$ is continuous and the maximum of its density is uniformly bounded from above.

C5. The expectation $\mathbb{E}(\|\boldsymbol{V}\|_1 \mid \boldsymbol{U} = \boldsymbol{u}) < \infty$ uniformly in $\boldsymbol{u}$, where $\|\cdot\|_1$ is the L_1 norm.

Theorem 1. *If Conditions C1-C5 hold, then*
(i) under the null model that $\boldsymbol{\beta}_2 = 0$ and $\sigma_1 = \sigma_2 = \sigma_0$ where σ_0 is unknown, for any fixed $\boldsymbol{\gamma}$ with nonzero slope the maximum penalized likelihood estimator

of $\boldsymbol{\theta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma_1, \sigma_2)$ *is consistent;*
(ii) under the alternative model that $\boldsymbol{\beta}_2 \neq 0$ *or* $\sigma_1 \neq \sigma_2$*, the maximum penalized likelihood estimator of* $\boldsymbol{\eta}^T = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma_1, \sigma_2)$ *from* (2.4) *is consistent.*

To illustrate the idea used in the proof for Theorem 1, we consider any sequence of positive numbers σ_n and let $W_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n 1\Big(|Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}| \le |\sigma_n \log \sigma_n|\Big)$, $A_n(C) = \Big\{\sup_{\boldsymbol{\beta}\in R^{q_1}} W_n(\boldsymbol{\beta}) > C|\sigma_n \log \sigma_n|\Big\}$, and $B_n = \Big\{\sup_{\boldsymbol{\beta}\in R^{q_1}} W_n(\boldsymbol{\beta}) > 4n^{-1}(\log n)^2\Big\}$. We show

S1. There exist $C_1, C_2 > 0$ such that uniformly for $\sigma_n \in \big(n^{-1}, e^{-1}\big)$, $P\Big(A_n(C_1)\Big) \le C_2 n^{-2}$;

S2. There exists $C_3 > 0$ such that uniformly in $\sigma_n \in (0, n^{-1})$, $P(B_n) \le C_3 n^{-2}$.

As the log-likelihood of the normal mixture model becomes unbounded when some sample points are close to one of the estimated component means and when the corresponding variance estimator goes to zero, **S1** and **S2** actually give an upper bound of the number of points that fall into such *trouble regions*. The upper bound is approximately limited to the order of $O((\log n)^2)$. The penalty that satisfies C1-C3 ensures that the variance estimators stay away from zero.

Here **S1** and **S2** play the same role as **Lemma 1** of Chen, Tan and Zhang (2008) in a somewhat simpler setting. By **Lemma 2** of Chen, Tan and Zhang (2008), we can show that the number of sample points that fall within the range of $|\sigma \log \sigma|$ to either one of the estimated component means is in the order of $O((\log n)^2)$. As a consequence, we can show that the estimates of σ_1 and σ_2 stay away from zero. Standard techniques in the large sample theory can then be applied to show the consistency of the maximum penalized likelihood estimators. Since proving **S2** is essentially the same as proving **S1**, we only provide the details of the proof for **S1** in the supplementary file.

2.4. Distribution of the penalized EM test statistic

The *Fisher information matrix* for $\boldsymbol{\theta}$ given $\boldsymbol{\gamma}$ based on the penalized likelihood is

$$I_{\boldsymbol{\gamma}}^*(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\frac{pl(\boldsymbol{\theta}^T, \boldsymbol{\gamma}^T)}{n}\right].$$

By direct calculations, for a given $\boldsymbol{\gamma}$ under the null hypothesis of $\boldsymbol{\beta}_2 = 0, \sigma_1 = \sigma_2 = \sigma$,

$$I_{\boldsymbol{\gamma}}^*(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{pmatrix} I_1 & 0_{2\times 2} \\ 0_{2\times 2} & I_2 \end{pmatrix}, \tag{2.8}$$

where

$$I_1 = \begin{pmatrix} \mathbb{E}\big(\boldsymbol{Z}\boldsymbol{Z}^T\big) & \mathbb{E}\big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})\boldsymbol{Z}\boldsymbol{Z}^T\big) \\ \mathbb{E}\big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})\boldsymbol{Z}\boldsymbol{Z}^T\big) & \mathbb{E}\big(\pi^2(\boldsymbol{X}^T\boldsymbol{\gamma})\boldsymbol{Z}\boldsymbol{Z}^T\big) \end{pmatrix},$$

$$I_2 = \begin{pmatrix} 2\mathbb{E}\big(\pi^2(\boldsymbol{X}^T\boldsymbol{\gamma})\big) - n^{-1}\sigma^2\mathbb{E}\big(p_n''(\sigma)\big) & 2\mathbb{E}\big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})[1-\pi(\boldsymbol{X}^T\boldsymbol{\gamma})]\big) \\ 2\mathbb{E}\big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})[1-\pi(\boldsymbol{X}^T\boldsymbol{\gamma})]\big) & 2\mathbb{E}\big([1-\pi(\boldsymbol{X}^T\boldsymbol{\gamma})]^2\big) - n^{-1}\sigma^2\mathbb{E}\big(p_n''(\sigma)\big) \end{pmatrix}.$$

Here I^* is positive definite if the variable vectors $\boldsymbol{X}$ and $\boldsymbol{Z}$ are non-degenerate, $\mathbb{E}\big(p_n''(\sigma)\big) < 0$, and $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_{-X}, \boldsymbol{\gamma}_X^T)$ where $\boldsymbol{\gamma}_{-X}$ refers to the intercept, and the slope coefficient $\boldsymbol{\gamma}_X^T \neq 0$. If the slope $\boldsymbol{\gamma}_X$ is zero, the matrix is degenerate, so we use only $\boldsymbol{\gamma}$ with nonzero slopes.

C6. Under the null model of $(\boldsymbol{\beta}_2 = 0, \sigma_1 = \sigma_2 = \sigma > 0)$, we have $\mathbb{E}p_n''(\sigma) < 0$, $\mathbb{E}\big(p_n''(\sigma)\big) = o_p(n)$, and $p_n'(\sigma) = o_p(n^{-1/2})$.

Under C6, we have,

$$I_{\boldsymbol{\gamma}}^*(\boldsymbol{\theta}) = \mathbb{E}\bigg[\frac{\partial}{\partial\boldsymbol{\theta}}\frac{pl(\boldsymbol{\theta}^T,\boldsymbol{\gamma}^T)}{n}\frac{\partial}{\partial\boldsymbol{\theta}^T}\frac{pl(\boldsymbol{\theta}^T,\boldsymbol{\gamma}^T)}{n}\bigg] + o_p(1),$$

and $I_{\boldsymbol{\gamma}}^*(\boldsymbol{\theta})$ works just like the usual Fisher information matrix for deriving the limiting distribution of the likelihood ratio statistic.

Given $\boldsymbol{\gamma}$ with nonzero slope, with $\hat{\boldsymbol{\theta}}_n = \operatorname{argmax}_{\boldsymbol{\theta}} pl(\boldsymbol{\theta},\boldsymbol{\gamma})$, and $\hat{\boldsymbol{\theta}}_0 = \operatorname{argmax}_{\boldsymbol{\theta}\in H_0} pl(\boldsymbol{\theta},\boldsymbol{\gamma})$, we have a quadratic approximation of the penalized likelihood ratio statistic $T^*(\boldsymbol{\gamma})$ and

$$T^*(\boldsymbol{\gamma}) = 2[pl(\hat{\boldsymbol{\theta}}_n,\boldsymbol{\gamma}) - pl(\hat{\boldsymbol{\theta}}_0,\boldsymbol{\gamma})] = \|n^{-1/2}\psi^*(Y_i,\boldsymbol{Z}_i,\boldsymbol{X}_i;\boldsymbol{\gamma})\|^2 + o_p(1), \tag{2.9}$$

where $\psi^*(Y_i,\boldsymbol{Z}_i,\boldsymbol{X}_i;\boldsymbol{\gamma}) = (\psi(Y_i,\boldsymbol{Z}_i,\boldsymbol{X}_i;\boldsymbol{\gamma})^T, \psi_0(Y_i,\boldsymbol{Z}_i,\boldsymbol{X}_i;\boldsymbol{\gamma}))$, in which

$$\boldsymbol{\psi}(Y_i,\boldsymbol{Z}_i,\boldsymbol{X}_i;\boldsymbol{\gamma}) = \sigma_0^{-1}\boldsymbol{D}(\boldsymbol{\gamma})^{-1/2}\{\pi(\boldsymbol{X}_i^T\boldsymbol{\gamma})\boldsymbol{I}_{q_2} - \boldsymbol{B}(\boldsymbol{\gamma})\boldsymbol{A}^{-1}\}(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}_0)\boldsymbol{Z}_i, \tag{2.10}$$

and

$$\psi_0(Y_i,\boldsymbol{Z}_i,\boldsymbol{X}_i;\boldsymbol{\gamma}) = \Big\{2[\mathbb{E}\Big(\pi^2(\boldsymbol{X}^T\boldsymbol{\gamma})\Big) - \mathbb{E}^2\Big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})\Big)]\Big\}^{-1/2} \Big\{\mathbb{E}\Big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})\Big) - \pi(\boldsymbol{X}^T\boldsymbol{\gamma})\Big\}\Big\{\sigma^{-2}(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}_1)^2 - 1\Big\},$$

with $A = \mathbb{E}\big(\boldsymbol{Z}\boldsymbol{Z}^T\big), B(\boldsymbol{\gamma}) = \mathbb{E}\big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})\boldsymbol{Z}\boldsymbol{Z}^T\big), C(\boldsymbol{\gamma}) = \mathbb{E}\big(\pi^2(\boldsymbol{X}^T\boldsymbol{\gamma})\boldsymbol{Z}\boldsymbol{Z}^T\big)$, and $D(\boldsymbol{\gamma}) = C(\boldsymbol{\gamma}) - B(\boldsymbol{\gamma})A^{-1}B(\boldsymbol{\gamma})$. Direct calculations show that both $\psi(Y_i,\boldsymbol{Z}_i,\boldsymbol{X}_i;\boldsymbol{\gamma})$ and $\psi_0(Y_i,\boldsymbol{Z}_i,\boldsymbol{X}_i;\boldsymbol{\gamma})$ have mean zero, and the covariance matrix of ψ^* is I_{q_1+1}. Therefore, $T^*(\boldsymbol{\gamma})$ has a χ^2 limiting distribution with the degrees of freedom $q_1 + 1$. We have not updated the estimates through the EM iterations, so $T^*(\boldsymbol{\gamma}) = pEM^{(0)}$ with only one starting value, $\boldsymbol{\gamma}$.

Following similar arguments to those used for Theorem 1 of Shen and He (2015), we see that the representation in (2.9) holds uniformly in $\boldsymbol{\gamma} \in \tilde{\Gamma}$.

Theorem 2. *Under the null hypothesis and C1-C6, for any finite integers $J > 0$ and $K \geq 0$, the penalized EM test statistic $pEM^{(K)}$ converges in distribution as $n \to \infty$. For $J = 1$ and $K = 0$, the limiting distribution is $\chi^2_{q_1+1}$, where q_1 is the dimension of $\boldsymbol{\beta}_2$.*

If the null hypothesis H_0 is rejected, the model parameter estimator is consistent from the penalized likelihood due to Theorem 1. Furthermore, from (2.9) it follows that the bootstrap method can be used to compute the p value of the pEM test. The limiting distribution of the test statistic under the null hypothesis is not a simple chi-square distribution when $J > 1$ and $K \geq 1$ and, moreover, the convergence to the limiting distribution is very slow for the test statistic even without covariates (Goeffinet, Loisel and Laurent (1992)). Therefore, we recommend use of the bootstrap method for computing the p values.

2.5. Local power

We investigate the local power of the pEM test by considering hypothesis testing of

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}, \sigma_1 = \sigma_2 = \sigma_0 \text{ vs.} H_a^* : \boldsymbol{\beta}_2 = n^{-1/2}\boldsymbol{h}, \sigma_1 = \sigma_2 + n^{-1/2}h_1 = \sigma_0 + n^{-1/2}h_1,$$

for some fixed quantities h_1 and $\boldsymbol{h}$.

Theorem 3. *Under H_a^*, the test statistic $T_K^*(\boldsymbol{\gamma})$, with any value $\boldsymbol{\gamma} \in \tilde{\boldsymbol{\Gamma}}$ and for any positive integer K, converges to a noncentral chi-square distribution with $q_1 + 1$ degrees of freedom and the noncentrality parameter*

$$\lambda^*(\boldsymbol{\gamma}) = \lambda(\boldsymbol{\gamma}) + \lambda_1(\boldsymbol{\gamma}), \tag{2.11}$$

in which

$$\lambda(\boldsymbol{\gamma}) = \sigma_0^{-2} \big\| \boldsymbol{I}_{\boldsymbol{\gamma}22\cdot1}^{-1/2} \Big\{ \mathbb{E}[\pi(\boldsymbol{X}^T\boldsymbol{\gamma})\pi(\boldsymbol{X}^T\boldsymbol{\gamma}_0)\boldsymbol{Z}\boldsymbol{Z}^T] - B(\boldsymbol{\gamma})A^{-1}B(\boldsymbol{\gamma}_0) \Big\} \boldsymbol{h} \big\|^2, \tag{2.12}$$

$$\lambda_1(\boldsymbol{\gamma}) = \sigma_0^{-2} 2h_1^2 \Big\{ \mathbb{E}\Big(\pi^2(\boldsymbol{X}^T\boldsymbol{\gamma})\Big) - \mathbb{E}^2\Big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})\Big) \Big\}. \tag{2.13}$$

Remark 3. The local power of the test under the unequal variance model can be compared with that under the equal variance model. Here $\lambda(\boldsymbol{\gamma})$ is the same as (15) of Shen and He (2015), the noncentral parameter of the noncentral chi-square distribution of the test statistic under the local alternative for the equal variance model. If the equal valiance model is actually true and $h_1 = 0$, the noncentral parameters are the same under two models, but the degree of freedom is higher by one under the unequal variance model. Consequently the test under the unequal variance model loses some power. The comparison reverses in direction when h_1 is sufficiently large.

2.6. Choice of the tuning parameter

The choice of the tuning parameter λ is practically important. Our conditions allow data-dependent penalties. For the specific penalty in (2.2), we show that C1-C3 and C6 are satisfied with any choice of λ in the interval

$$\left[n^{-2+c_1}(\log n)^3,\, n^{c_2}\right], \tag{2.14}$$

where $c_1, c_2 \in (0,1)$ are any constants. In our empirical studies, we choose $c_1 = c_2 = 0.9$.

Under the null model, C1 and C3 are satisfied by the choice of c_2. Also, $n^{c_2} = o(n)$ and $n^{c_2-1/2} = o(n^{1/2})$ imply $\mathbb{E}\big(p_n''(\sigma_0)\big) = -4\lambda/\sigma_0^2 = o(n)$ and $p_n'(\sigma_0) = 2(n^{-1/2}\lambda)\{n^{1/2}(S_n^2 - \sigma_0^2)\}/\sigma_0^3 = 2(n^{-1/2}\lambda)O_p(1) = o_p(n^{1/2})$; hence C6 is satisfied.

For C2, under the null hypothesis $S_n^2 \to \sigma_0^2$ almost surely. Write $S_n^2 = \sigma_0^2 + \epsilon_n$, where $\epsilon_n \to 0$ almost surely. Then for any $\sigma \in (0, n^{-1})$ and sufficiently large n, we have

$$\begin{aligned} p_n(\sigma)[(\log n)^2 \log\sigma]^{-1} &= -\lambda[\frac{\sigma_0^2+\epsilon_n}{\sigma^2} + \log(\frac{\sigma^2}{\sigma_0^2+\epsilon_n})][(\log n)^2\log\sigma]^{-1} \\ &\geq -\frac{\lambda}{2}\frac{\sigma_0^2}{\sigma^2}[(\log n)^2\log\sigma]^{-1}. \end{aligned} \tag{2.15}$$

If $f_n(\sigma) = (-\lambda/2)(\sigma_0^2/\sigma^2)[(\log n)^2\log\sigma]^{-1}$, then

$$\inf\Big\{p_n(\sigma)[(\log n)^2\log\sigma]^{-1} : 0 < \sigma \leq n^{-1}\Big\} \geq \inf\Big\{f_n(\sigma) : 0 < \sigma \leq n^{-1}\Big\}.$$

Because $f_n(\sigma)$ is decreasing in $\sigma \in (0, n^{-1})$ for large n, we have

$$\inf\Big\{p_n(\sigma)[(\log n)^2\log\sigma]^{-1} : 0 < \sigma \leq n^{-1}\Big\} \geq f_n(n^{-1}) = O\Big(\lambda n^2(\log n)^{-3}\Big).$$

By the choice of λ, for sufficiently large n, $\inf\{p_n(\sigma)\{(\log n)^2\log\sigma\}^{-1} : 0 < \sigma \leq n^{-1}\} \geq O(n^{c_1}) > C$, for any given constant C. Then, C2 is satisfied under the null hypothesis. The same results can be obtained under the alternative hypothesis due to the fact that S_n^2 is almost surely bounded.

In practice, we can choose λ from (2.14) for a specific purpose. For subgroup analysis, we suggest a 5-fold cross validation, where roughly 4/5 of the data are used as the training set.

For each λ, from the training set, we obtain the estimator and the resultant mixture model:

$$\pi(X^T\hat\gamma)N\Big(Z^T(\hat\beta_1+\hat\beta_2), \hat\sigma_1^2\Big) + [1-\pi(X^T\hat\gamma)]N\Big(Z^T\hat\beta_1, \hat\sigma_2^2\Big),$$

where we suppose without loss of generality that the second component of $\hat\beta_2$, denoted as $\hat\beta_2(trt)$ (as a measure of the treatment effect difference between two

subgroups), is positive, so the treatment effect is higher for subjects with higher π values.

For each quantile level $q \in (0,1)$, we rank the subjects in the testing set by their $\hat{\pi}$ (referred to as *membership scores*, $S(\boldsymbol{X};\lambda)$), and then choose the subjects whose *membership scores* are above the top q-th quantile $Q_q(S)$ to form a target subpopulation. Then, we evaluate the treatment effect difference in the selected subpopulation as

$$TT(\lambda;q) = \mathbb{E}\Big\{Y \mid [trt=1, S(\boldsymbol{X};\lambda) > Q_q(S)]\Big\} - \mathbb{E}\Big\{Y \mid [trt=0, S(\boldsymbol{X};\lambda) > Q_q(S)]\Big\}. \tag{2.16}$$

We choose the tuning parameter λ by maximizing the overall treatment effect difference as given by

$$TT(\lambda) = \int_{q_l}^{q_u} TT(\lambda;q)dq \tag{2.17}$$

for pre-specified values q_l and q_u. The intuition behind it is that, for each q corresponding to the size of the subgroup, the larger $TT(\lambda;q)$ is, the larger the treatment effect difference is between the selected subgroup and the rest. Subgroups that are associated with larger $TT(\lambda)$ are more practically useful to identify.

In practice, we use the empirical analogs for both (2.16) and (2.17). For the data analysis of Section 4, we evaluate the integral in (2.17) by an average over N equally spaced q values using $N = 12, q_l = 0.2, q_u = 0.8$, and the q's are $0.2, 0.25, \cdots, 0.8$.

3. Simulations

We studied the performance of the proposed methods through Monte Carlo simulations. We compared the parameter estimates from the structured logistic-normal mixture model with equal and unequal variances, and the performance of the proposed pEM test versus the EM test of Shen and He (2015). We used $q_2 = 2$ and $\tilde{\Gamma} = [-5,5] \times ([0.2,5] \bigcup [-5,-0.2])$, and other parameters are given below. The bootstrap method was used to compute the p values of the tests for the empirical studies.

3.1. Estimation

We evaluated the parameter estimates when the mixture model parameters were all well defined. Data as random samples of sizes $n = 400$ were generated

from

$$Y_i \mid (X_i, Z_i, \delta_i) = \beta_{11} + \beta_{12}T_i + \beta_{13}Z_i + (\beta_{21} + \beta_{22}T_i + \beta_{23}Z_i)\delta_i + \varepsilon_{1i}\delta_i + \varepsilon_{2i}(1-\delta_i),$$
$$P(\delta_i = 1 \mid X_i) = \pi(\gamma_{11} + \gamma_{12}X_i),$$

for $i = 1, \ldots, n$, where $\varepsilon_{1i} \sim N(0, \sigma_1^2)$ and $\varepsilon_{2i} \sim N(0, \sigma_2^2)$, independent of X_i, Z_i and T_i. We generated $X_i = Z_i$ from Uniform $(0, 4)$, and $T_i \in \{0, 1\}$, used to mimic the treatment indicator, was generated from the Bernoulli distribution with $P(T_i = 1) = 0.5$. We fixed $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13}) = (2, 0, 2), \boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23}) = (1, 2, 0), \boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}) = (2, -1), \sigma_1 = 1.5$, and $\sigma_2 = 0.5$. In the computations, we adopted the constraint $\beta_{22} > 0$ to guarantee the uniqueness of the parameters. We show in Figure 1 the boxplots of the absolute bias of the parameter estimates based on 100 data sets. Not surprisingly, the estimates from the equal variance model have larger biases than those from the unequal variances model, so it is helpful to take the heterogeneity in the variances into consideration.

3.2. Type I errors

To evaluate the validity of the pEM test, we generated data from Model (2.1) with $q_1 = 3, q_2 = 2$, $\boldsymbol{\beta}_1 = (1, 0, 2)^T$, $\boldsymbol{\beta}_2 = (0, 0, 0)^T$, $\boldsymbol{Z} = (1, t, x)^T$, $\boldsymbol{X} = (1, x)^T$, where t was a treatment indicator distributed as Bernoulli(0.5), x was independent of t with the distribution $N(-1, 1)$, and the error ε was $N(0, 0.5^2)$. The pEM test used $\boldsymbol{\Gamma} = \{(1, -2)^T, (1, 2)^T\}$. The resulting type I errors at $n = 60$ and 100 are summarized in Table 1, from which we can see the type I errors are quite close to the nominal levels for $K = 0, 3$, and 9, even for relatively small sample sizes. The tuning parameter was set as $\lambda = 1$ here, but the results are similar for other choices of λ. For instance, the results for $\lambda = 50$ are given in Table S1 in the supplement.

3.3. Power comparison

We used the same model and the same pEM test as in the previous subsection, except that $\boldsymbol{\beta}_2 = (1, a, b)^T$, $\boldsymbol{\gamma} = (1, 1)^T$ for some non-negative values of a and b to be given in the tables and for different sets of σ values. In particular, we considered $(\sigma_1 = 0.5, \sigma_2 = 0.5), (\sigma_1 = 0.4, \sigma_2 = 0.6), (\sigma_1 = 0.5, \sigma_2 = 1.0)$, and $(\sigma_1 = 0.5, \sigma_2 = 1.5)$ in Table 2 to represent different levels of heterogeneity. The power was obtained from the EM or pEM test under the equal or unequal variance model. We only show the comparisons at the iterations times $K = 9$ as

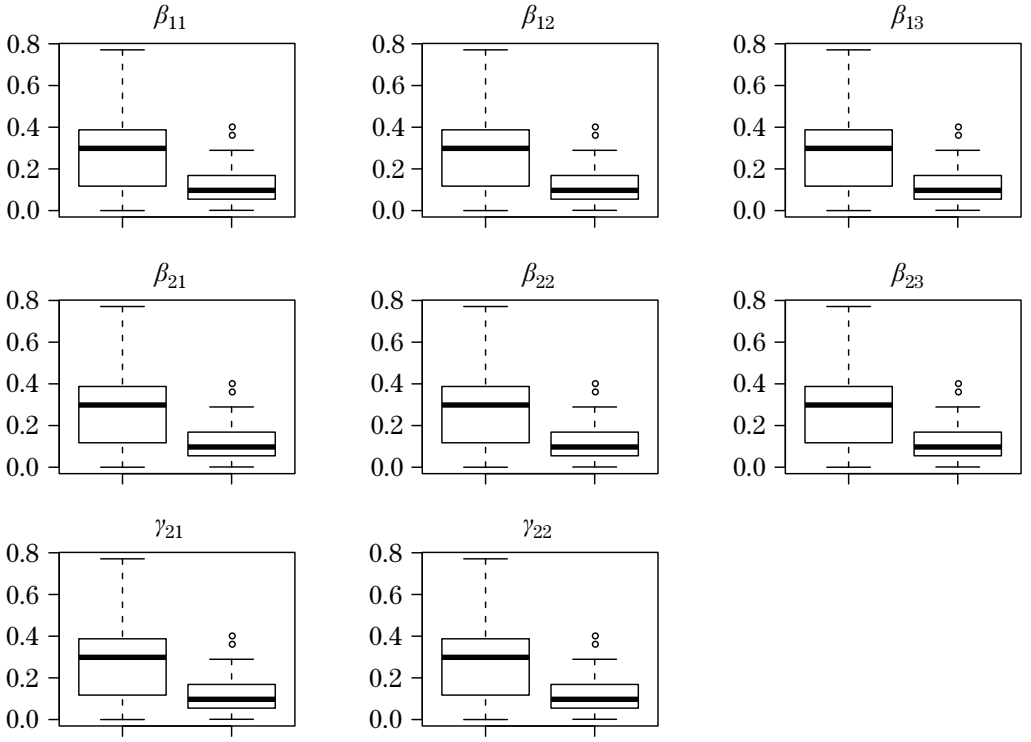

Figure 1. The boxplots of the absolute biases in 100 experiments discussed in Section 3.1. In each sub-panel, the left and the right boxes are for the estimates under the equal and the unequal variance models, respectively.

Table 1. Type I errors of the pEM tests with bootstrap approximations in 1,000 data sets with standard errors in the parenthesis, with $\lambda = 1$.

n	Nominal level α	$pEM^{(0)}$	$pEM^{(3)}$	$pEM^{(9)}$
$n = 60$	0.01	0.012(0.003)	0.011(0.003)	0.011(0.003)
	0.05	0.055(0.007)	0.055(0.007)	0.050(0.007)
	0.10	0.102(0.010)	0.103(0.010)	0.106(0.010)
$n = 100$	0.01	0.010(0.003)	0.011(0.003)	0.010(0.003)
	0.05	0.049(0.007)	0.051(0.007)	0.050(0.007)
	0.10	0.102(0.010)	0.099(0.009)	0.104(0.010)

this is our recommended choice. When the two component variances are close, the EM test based on the equal variance assumption is slightly more powerful, but when the two σ's differ with their ratios equal to 2 and 3, the pEM test under the unequal variance model is significantly more powerful.

Table 2. Power (%) of the (penalized) EM tests at the 5% level. The (penalized) EM test used $\mathbf{\Gamma} = \{(1,2)^T, (1,-2)^T\}$, with $K = 9$ iterations. The parameters of Model (2.1) were $\boldsymbol{\beta}_1 = (1,0,2)^T$, $\boldsymbol{\beta}_2 = (1,a,b)^T$, $\boldsymbol{\gamma} = (1,1)^T$, and the tuning parameter was $\lambda = 1.0$.

n	a	b	$pEM^{(9)}$	$EM^{(9)}$	$pEM^{(9)}$	$EM^{(9)}$
			($\sigma_1 = 0.5$, $\sigma_2 = 0.5$)		($\sigma_1 = 0.4$, $\sigma_2 = 0.6$)	
60	0.5	1	71.2	77.8	73.4	73.6
60	0.5	0	35.6	36.0	42.2	37.6
60	1.0	1	85.2	87.8	86.6	87.8
60	1.0	0	81.4	84.8	82.8	82.0
100	0.5	1	92.0	96.8	92.8	94.8
100	0.5	0	57.8	54.8	74.6	49.6
100	1.0	1	96.8	99.4	97.8	98.8
100	1.0	0	95.8	97.6	97.2	96.0
			($\sigma_1 = 0.5$, $\sigma_2 = 1.0$)		($\sigma_1 = 0.5$, $\sigma_2 = 1.5$)	
60	0.5	1	49.0	38.4	53.8	31.0
60	0.5	0	36.8	27.2	55.0	40.8
60	1.0	1	63.4	47.2	63.8	39.6
60	1.0	0	63.0	44.8	70.8	47.8
100	0.5	1	77.6	60.6	81.2	42.0
100	0.5	0	65.8	34.2	81.8	51.2
100	1.0	1	87.6	75.4	86.6	51.8
100	1.0	0	89.8	58.6	90.0	58.6

4. Applications

We applied the proposed model and the pEM test to two studies and discuss our findings in comparison with what have been known from earlier investigations.

4.1. NSW data

The National Supported Work (NSW) study was used to examine whether the job training program was beneficial to certain disadvantaged workers from 1975 to 1978 in the United States. We focused on the same subset of the data as previously used in Imai and Ratkovic (2013) where the treatment and control groups were randomly selected. As in Imai and Ratkovic (2013), we used the earning difference from 1975 to 1978 (log scale) as the response variable Y. More specifically, we considered $Y = \log(\text{RE78}+1) - \log(\text{RE75}+1)$, in which RE78 and RE75 are the earning of each individual in 1975 and in 1978, respectively. From preliminary studies, we narrowed down to these variables: *trt*: the treatment indicator whether the subject joins the training program; X_1: education in years; X_2: race indicator (1 for Black and 0 otherwise); X_3: binary indicator for whether

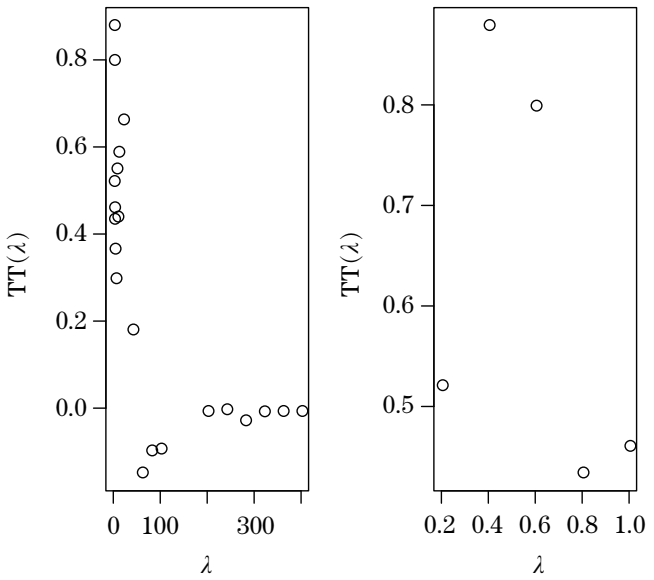

Figure 2. NSW data. The overall treatment effect in (2.17) with different choices of λ's. In the left figure, we plot the values $TT(\lambda)$ against $\lambda \in [0.2, 374]$; in the right figure, we narrow down to a small region with $\lambda = 0.2, 0.4, \cdots, 1$ to show that $TT(\lambda)$ achieves the maximum at $\lambda = 0.4$ among the selected λ values.

the baseline income is zero; X_4: binary indicator for whether the baseline income is above the median of the nonzero income group.

They are summarized in Table S2 in the supplement. If we fit a linear regression model to all the covariates and their two-way interactions, the treatment effect is estimated at -0.17 with a p-value 0.94. We ask whether there exists a subgroup for which the training program is effective, using the structured logistic-normal mixture model, with $\boldsymbol{X} = (1, X_1, X_2, X_3, X_4)$ and $\boldsymbol{Z}$ consists of trt and $\boldsymbol{X}$.

The EM and pEM tests rejected the null hypothesis of no subgroup, with p-values< 0.001. The parameter estimates from the equal variance and the unequal variance models are shown in Table 3. For the tuning parameter in the penalized likelihood, we used the range of λ of $[0.2, 374]$ by setting $c_1 = 0.9$ and $c_2 = 0.9$ in (2.14). Using the proposed criterion in Section 2.6, we plot $TT(\lambda)$ in Figure 2 from which we choose $\lambda = 0.4$. Although the EM test assuming equal variances rejects the null hypothesis of no subgroups, the coefficient estimates and their standard errors in Table 3 suggest that the treatment effect difference, $\beta_2(trt)$, with the value of 0.11 and standard error of 0.15, is not significant. In contrast, the estimate based on the unequal variance model gives the treatment

Table 3. Parameter estimates and their standard errors when the equal/unequal variance structured logistic-normal mixture model was used to fit the NSW data.

From **equal** variance structured logistic-normal mixture model						
	$\boldsymbol{\beta}_1(1)$	$\boldsymbol{\beta}_1(\text{trt})$	$\boldsymbol{\beta}_1(X_1)$	$\boldsymbol{\beta}_1(X_2)$	$\boldsymbol{\beta}_1(X_3)$	$\boldsymbol{\beta}_1(X_4)$
est	1.25	−0.05	0.04	−8.59	7.11	−1.85
se	0.39	0.12	0.04	0.12	0.14	0.15
	$\boldsymbol{\beta}_2(1)$	$\boldsymbol{\beta}_2(\text{trt})$	$\boldsymbol{\beta}_2(X_1)$	$\boldsymbol{\beta}_2(X_2)$	$\boldsymbol{\beta}_2(X_3)$	$\boldsymbol{\beta}_2(X_4)$
est	−8.53	0.11	0.00	16.83	0.05	0.25
se	0.54	0.15	0.05	0.28	0.18	0.19
	$\boldsymbol{\gamma}(1)$		$\boldsymbol{\gamma}(X_1)$	$\boldsymbol{\gamma}(X_2)$	$\boldsymbol{\gamma}(X_3)$	$\boldsymbol{\gamma}(X_4)$
est	−1.70		−0.02	2.75	−0.26	0.08
se	0.58		0.05	0.27	0.20	0.22
	σ					
est	0.98					
se	0.03					
From **unequal** variance structured logistic−normal mixture model						
	$\boldsymbol{\beta}_1(1)$	$\boldsymbol{\beta}_1(\text{trt})$	$\boldsymbol{\beta}_1(X_1)$	$\boldsymbol{\beta}_1(X_2)$	$\boldsymbol{\beta}_1(X_3)$	$\boldsymbol{\beta}_1(X_4)$
est	1.13	0.01	0.04	−0.03	7.20	−1.64
se	1.71	0.47	0.15	0.76	0.54	0.61
	$\boldsymbol{\beta}_2(1)$	$\boldsymbol{\beta}_2(\text{trt})$	$\boldsymbol{\beta}_2(X_1)$	$\boldsymbol{\beta}_2(X_2)$	$\boldsymbol{\beta}_2(X_3)$	$\boldsymbol{\beta}_2(X_4)$
est	−3.45	1.15	0.03	−1.37	−1.46	−0.30
se	2.04	0.56	0.17	0.86	0.65	0.73
	$\boldsymbol{\gamma}(1)$		$\boldsymbol{\gamma}(X_1)$	$\boldsymbol{\gamma}(X_2)$	$\boldsymbol{\gamma}(X_3)$	$\boldsymbol{\gamma}(X_4)$
est	−1.45		0.04	1.14	−0.24	−0.6
se	0.52		0.05	0.22	0.19	0.2
	σ_1	σ_2				
est	3.92	0.77				
se	0.19	0.03				

effects difference, $\beta_2(trt)$, with a value of 1.15 and standard error 0.56, is indeed significant. Furthermore, the estimated $\boldsymbol{\gamma}$ coefficients indicate that those who are black, more educated, and have low (but nonzero) baseline salary are more likely to benefit from the program, which is different from the conclusion of Imai and Ratkovic (2013) that "unemployed Hispanics and highly educated, low-earning non-Hispanics are predicted to benefit from the program". From each model, we can score individuals based on the estimated probabilities π of belonging to the subgroup with better treatment effects. Based on the equal variance and the unequal variance models, the *membership scores* are

$$S_1(\boldsymbol{X}) = \pi\Big(-1.70 - 0.02X_1 + 2.75X_2 - 0.26X_3 + 0.08X_4\Big); \tag{4.1}$$

$$S_2(\boldsymbol{X}) = \pi\Big(-1.45 + 0.04X_1 + 1.14X_2 - 0.24X_3 - 0.6X_4\Big). \tag{4.2}$$

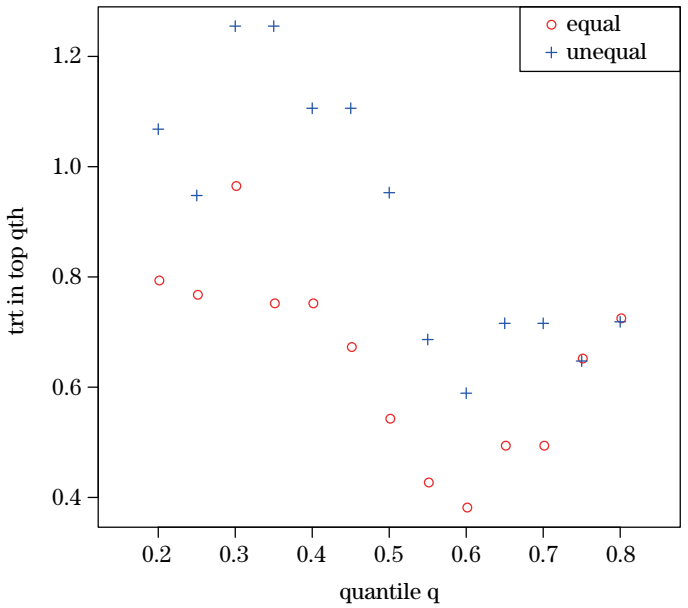

Figure 3. NSW data. Treatment effects in selected subgroups determined by *membership score* from the equal and the unequal variance models. The circles denote the results from the equal variance model, and the + signs denote the results from the unequal variance model.

If we select the subjects whose *membership scores* are in the top $q \times 100\%$ to form a subgroup for $q = 0.2, 0.25, \ldots, 0.8$, respectively, we then show the average treatment effects in the selected subgroup in Figure 3. It is clear from the figure that, across different subgroup sizes, the subgroup assignment based on the *membership scores* from the unequal variance model leads to high treatment effects for the selected subgroup. This is another piece of evidence that subgroup identification from the unequal variance model did better in this example.

4.2. AIDS data

We revisited the ACTG 320 clinical trial for AIDS patients that has been analyzed in Zhao et al. (2013) and Shen and He (2015). The AIDS patients were randomly assigned to a standard two drug combination and a new three drug combination in this randomized trial. We used the CD4 count changes at the 24th week (cd4.24) as the response variable, following the previous studies. Three baseline variables, age: in years; $\log(cd4.0)$: baseline CD4 counts on the log scale; $\log_{10}(rna.0)$, baseline RNA concentration on the $\log_{10}$ scale were used with the treatment indicator *trt*. Since there was a concern regarding the side

effect of the newly-added component in the three drug combination, we were interested in examining whether there exists a subgroup in which the treatment effect (compared to the standard two drug combination) is high enough to compensate for the side effect. In addition, if the subgroup exists, we would predict the subgroup membership based on the baseline variables for individual patients.

In the logistic-normal mixture model (2.1) with $\boldsymbol{Z} = (1,\ trt,\ \log(cd4.0),\ \log_{10}(rna.0),\ Age)$ for the normal component and $\boldsymbol{X} = (1, \log(cd4.0), \log_{10}(rna.0),\ Age)$ for the logistic component, we used the criterion of the preceding section to select $\lambda = 400$ for the penalized likelihood. The same choices of Γ and $\tilde{\Gamma}$ were used as in the analysis of Shen and He (2015). Then, the *pEM* test rejected the null hypothesis of one component only (p-value < 0.001). The parameter estimates are given in Table S3 of the supplement.

The difference of the treatment effects (51.74) is smaller than that (112.98) under the equal variance model. The mean probability of falling into the subgroup of higher treatment is around 0.42. The ratio of the two σ's is around 1.2. Just as what we did in the NSW example, we computed the *membership scores* from the equal variance model,

$$S_3(\boldsymbol{X}) = \pi\Big(-7.89 + 0.44\log(cd4.0) + 1.10\log_{10}(rna.0) - 0.02\text{Age}\Big) \qquad (4.3)$$

and the *membership scores* from the unequal variance model

$$S_4(\boldsymbol{X}) = \pi\Big(-9.16 + 0.67\log(cd4.0) + 1.40\log_{10}(rna.0) - 0.02\text{Age}\Big). \qquad (4.4)$$

In Figure S1 in the supplement, we present the *membership scores* for all the subjects estimates by (4.3) and (4.4). Compared to the estimated $S_3(\boldsymbol{X})$ values, the estimated $S_4(\boldsymbol{X})$ values are more spread out on both sides of 0.5, which give more meaningful interpretations in subgroup applications.

In this case, the two sets of *membership scores* have linear and rank correlations around 0.99, so the two models lead to very similar subgroups. The unequal variance model however allows a more relaxed condition on the model, and results in more interpretable π values for subgroup assignments.

5. Discussions

As with any model-based method, if the underlying data generating mechanism differs much from our proposed logistic-normal mixture model, the estimators can be biased and the test results untrustworthy. Model diagnostics tools are certainly worth developing. Another useful extension of our work is to allow the number of covariates to be large, in which case feature selection is needed.

Here we only considered modeling two subgroups, but large studies may require more than two subgroups.

Supplementary Materials

In the Supplementary Material, we provide the proofs for **Theorem 1** and **Theorem 2**, and give more details for the analysis of the data in Section 4.

Acknowledgment

Shen's research was partially supported by National Natural Science Foundation of China (Grant No. 11501123 and Grant No. 11571081). He's research was supported by USA National Science Foundation Awards DMS-1307566 and DMS-1607840), and National Natural Science Foundation of China (Grant No. 11129101).

References

Altstein, L. L., Li, G. and Elashoff, R. M. (2011). A method to estimate treatment efficacy among latent subgroups of a randomized clinical trial. *Statistics in Medicine* **30**, 709–717.

Berger, J. O., Wang, X. and Shen, L. (2014). A Bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics* **24**, 110–129.

Cai, T., Tian, L., Wong, P. H. and Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* **12**, 270–282.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society B* **63**, 19–29.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: the EM approach. *The Annals of Statistics* **37**, 2523–2542.

Chen, J., Tan, X. and Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica* **18**, 443–465.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* **39**, 1–38.

Fan, A., Lu, W. and Song, R. (2016). Change-plane analysis for subgroup detection and sample size calculation. *Journal of the American Statistical Association, accepted, DOI: 10.1080/01621459.2016.1166115* .

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

Goeffinet, B., Loisel, P. and Laurent, B. (1992). Testing in normal mixture models when the proportions are known. *Biometrika* **79**, 842–846.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* **7**, 443–470.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm.

Neural Computation **6**, 181–214.

Kang, C., Janes, H. and Huang, Y. (2014). Combining biomarkers to optimize patient treatment recommendations. *Biometrics* **70**, 695–707.

Lipkovich, I., Dmitrienko, A., Denne, J. and Enas, G. (2011). Subgroup identification based on differential effect search–a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine* **30**, 2601–2621.

Lipkovicha, I. and Dmitrienkoa, A. (2014). Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. *Journal of Biopharmaceutical Statistics* **24**, 130–153.

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

Simon, R. (2002). Bayesian subset snalysis: spplication to studying treatment-by-gender interactions. *Statistics in Medicine* **21**, 2909–2916.

Sleight, P. (2000). Debate: Subgroup analyses in clinical trials: fun to look at - but don't believe them! *Current Controlled Trials in Cardiovascular Medicine* **1**, 25–27.

Song, Y. and Chi, G. Y. (2007). A method for testing a pre-specified subgroup in clinical trials. *Statistics in Medicine* **26**, 3535–3549.

Su, X., Tsai, C. L., Wang, H., Nickerson, D. M. and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* **10**, 141–158.

Van Horn, M. L., Jaki, T., Masyn, K., Howe, G., Feaster, D. J., Lamont, A. E., George, M. R. W. and Kim, M. (2015). Evaluating differential effects using regression interactions and regression mixture models. *Educational and Psychological Measurement* **75**, 677–714.

Zhao, L., Tian, L., Cai, T., Claggett, B. and Wei, L. J. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association* **108**, 527–539.

Departments of Statistics, Fudan University, Shanghai 200433, China.

E-mail: shenjuan@fudan.edu.cn

Department of Statistics, University of Michigan, 1085 South University, Ann Arbor, MI 48109, USA.

E-mail: yingcw@umich.edu

Department of Statistics, University of Michigan, 1085 South University, Ann Arbor, MI 48109, USA.

E-mail: xmhe@umich.edu

(Received October 2015; accepted March 2016)