Generalized Linear Mixed Models with Gaussian Mixture Random Effects: Inference and Application

Lanfeng Pan[a], Yehua Li[b,*], Kevin He[c], Yanming Li[c], Yi Li[c]

[a]*Amazon.com, Inc., Seattle, WA 98109, U.S.A.*
[b]*Department of Statistics, University of California at Riverside, 900 University Ave. Riverside, CA 92521, U.S.A.*
[c]*School of Public Health & Kidney Epidemiology and Cost Center, University of Michigan, Ann Arbor, MI 48109, U.S.A.*

Abstract

We propose a new class of generalized linear mixed models with Gaussian mixture random effects for clustered data. To overcome the weak identifiability issues, we fit the model using a penalized Expectation Maximization (EM) algorithm, and develop sequential locally restricted likelihood ratio tests to determine the number of components in the Gaussian mixture. Our work is motivated by an application to nationwide kidney transplant center evaluation in the United States, where the patient-level post-surgery outcomes are repeated measures of the care quality of the transplant centers. By taking into account patient-level risk factors and modeling the center effects by a finite Gaussian mixture model, the proposed model provides a convenient framework to study the heterogeneity among the transplant centers and controls the false discovery rate when screening for transplant centers with non-standard performance.

Keywords: Clustering, False discovery rate, Latent variables, Locally restricted likelihood ratio test, Penalized EM algorithm, Repeated measure
AMS 2010 subject classifications: Primary 62H30, Secondary 62H15

1. Introduction

The generalized linear mixed model (GLMM) is the most widely used framework for repeatedly measured non-Gaussian data, where the vast majority of literature assumes the distribution of the random effect to be Gaussian. Most papers focus on estimating the fixed effects while treating the random effects as nuisance [4, 26]. Even though GLMM's are typically robust against deviations from the Gaussian random effect assumption [29], many authors have documented various drawbacks when the Gaussian assumption is violated, including loss of estimation efficiency [10] and reduced power for statistical tests [27]. Even though the predicted random effects are relatively robust in terms of mean squared error, the distribution for the predicted random effect is highly sensitive and mostly reflects the shape of the assumed random effect distribution [29]. Many authors have tried to relax the Gaussian assumption and model the random effect with more flexible distributions, such as the semi-nonparamatric distribution [10]. Caffo et al. [5] considered modeling the random effect with a Gaussian mixture model, but limited their investigation to binary probit models, focusing on numerical performance rather than theoretical justifications.

Finite Gaussian mixture models [30] are intuitively appealing for modeling non-homogeneous populations and detecting subgroups. There has been a recent surge in applications of Gaussian mixture models, including clustering analysis [16], false discovery rate control [11, 25] and genetic imprinting [23]. Statistical inference for Gaussian mixture models is well-known to be difficult, because many regularity conditions in parametric inference are violated in these models [6, 7, 13]. There has been much recent work in hypothesis testing on the order of finite Gaussian mixture models [8, 19]. However, none of the existing methods are directly applicable to generalized linear mixed models.

We investigate a new class of generalized linear mixed models with Gaussian mixture random effects, propose a penalized EM algorithm to fit the proposed model, and develop sequential locally restricted likelihood ratio tests to decide the number of components in the mixture model. Our work is motivated by an application on kidney transplant center evaluation, using the U.S. Organ Procurement and Transplantation Network (OPTN) database. We model the

*Corresponding author. Email address: yehuali@ucr.edu

patient level outcome, e.g., 5-year post-transplant survival status, using a GLMM, where the random effect for a transplant center follows a finite Gaussian mixture distribution. We then propose an empirical Bayes approach to classify the transplant centers using the fitted Gaussian mixture model, while controlling the false discovery rate. The results may have a strong impact on health-policy making and on the patients' choice of transplant centers.

The main advantage of the proposed method is its ability to fulfill multiple tasks under the same framework: it offers flexible modeling on the distribution of the random effect in GLMM; the Gaussian mixture structure enables a model based clustering on the units (i.e., the transplant centers in the OPTN data); the proposed inference procedure can be used to test if there are any clusters among the units and thus examine the goodness-of-fit of the classic GLMM assuming homogeneous Gaussian random effects; when the proposed tests are used sequentially, the procedure can automatic determine the number of mixture components; furthermore, the proposed framework provides an intuitive way to detect transplant centers with non-standard performance while controlling the false discovery rate. The proposed procedures are computationally intensive when analyzing large medical data sets, but can be done efficiently using parallel computing. We have developed a software package written in Julia (http://julialang.org/), which is a high-level, high-performance dynamic programming language. Our package is based on open source math libraries, supports parallel computing, and will be made available on the corresponding author's website.

The rest of the paper is organized as follows. In Section 2, we introduce the model, propose an EM-based estimation procedure and establish the consistency of the procedure. To decide the number of mixture components, we propose sequential locally restricted likelihood ratio tests in Section 3. In Section 4, we propose a false discovery rate control procedure to evaluate the care qualities of the transplant centers. We conduct simulations in Section 5 and report the analysis of the OPTN kidney transplant data in Section 6. Concluding remarks are provided in Section 7, and technical proofs are collected in Section 8. Detailed algorithms are relegated to the supplementary material.

2. Model and parameter estimation

2.1. Model and assumptions

The data consist n independent units (e.g., transplant centers), each with N_i subunits (patients), which brings the total number of measurements to $N = \sum_{i=1}^{n} N_i$. Let Y_{ik} be the outcome variable of the kth subunit in the ith unit and let $\boldsymbol{X}_{ik} \in \mathbb{R}^p$ be the subunit level covariate, $k \in \{1, \ldots, N_i\}$, $i \in \{1, \ldots, n\}$. Denote by $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iN_i})^\top$, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iN_i})^\top$, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^\top$ where γ_i is a random effect shared by all entries in $\boldsymbol{Y}_i$. In our motivating example, γ_i represents the quality of care delivered by the ith transplant center. The conditional density of Y_{ik}, given $\boldsymbol{X}_{ik}$ and γ_i, belongs to the canonical exponential family:

$$f(Y_{ik} \mid \boldsymbol{X}_{ik}, \gamma_i; \boldsymbol{\beta}, \varphi) = \exp\left\{ \frac{Y_{ik}\xi_{ik} - b(\xi_{ik})}{a(\varphi)} + d(Y_{ik}, \varphi) \right\}, \tag{1}$$

where $a(\cdot)$, $b(\cdot)$ and $d(\cdot)$ are known functions, $\xi_{ik} = \boldsymbol{X}_{ik}^\top \boldsymbol{\beta} + \gamma_i$ is the canonical parameter with $\mathrm{E}(Y_{ik} \mid \boldsymbol{X}_{ik}, \gamma_i) = b'(\xi_{ik})$, and φ is a nuisance parameter. Here $\boldsymbol{X}_{ik}$ does not contain the intercept and γ_i is allowed to have a nonzero mean. We also assume that Y_{ik} and $Y_{ik'}$ are independent given γ_i, for any $k \neq k'$. In our transplant center evaluation application, we consider a binary response variable: $Y_{ik} = 1$ if the patient died within 5 years after transplant; -1 otherwise. In the dataset, there was essentially no censoring within the first 5 years since the transplant patients' survival information had been closely monitored and tracked. With that, model (1) becomes $f(Y_{ik} \mid \boldsymbol{X}_{ik}, \gamma_i; \boldsymbol{\beta}) = \{1 + \exp(-\xi_{ik} Y_{ik})\}^{-1}$.

Assume that the units belong to C subpopulations and the cth subpopulation can be described by a Gaussian distribution with mean μ_c and variance σ_c^2, $c \in \{1, \ldots, C\}$. The density of γ_i is $g(\gamma \mid \boldsymbol{\theta}_\gamma) = \sum_{c=1}^{C} \pi_c f_c(\gamma \mid \mu_c, \sigma_c)$, where $f_c(\gamma \mid \mu_c, \sigma_c) = \sigma_c^{-1}\phi\{(\gamma - \mu_c)/\sigma_c\}$, $\phi(\cdot)$ is the standard Gaussian density, $\pi_c \in [0, 1]$ is the weight for subpopulation c, $\sum_{c=1}^{C} \pi_c = 1$, and $\boldsymbol{\theta}_\gamma = (\mu_1, \ldots, \mu_C, \sigma_1^2, \ldots, \sigma_C^2, \pi_1, \ldots, \pi_C)^\top$ collects the parameters in $g(\gamma)$. Here, C represents the number of clusters or subpopulations, and hence should be a positive number. When $C = 1$, the model is the classic GLMM with Gaussian random effects; when $C > 1$, the model is a GLMM with Gaussian mixture random effects.

2.2. Model fitting

Though conceptually appealing, Gaussian mixture models possess some undesirable properties, including a slower convergence rate for parameter estimation when the number of components is unknown [6], unbounded likelihood when any of the component variance parameters σ_c^2 goes to 0 [13], and infinite Fisher information on some boundary

points of the parameter space [7]. The solution to these problems in the literature is to either restrict the value of the parameters away from the boundaries [13] or include a penalty function to prevent any σ_c from converging to 0 [7, 9]. We adopt the latter strategy by maximizing a penalized likelihood

$$\ell_{pen}(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{X}) + \textstyle\sum_{c=1}^{C} p_n(\sigma_c^2), \tag{2}$$

where $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_n^\top)^\top$, $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top)^\top$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_y^\top, \boldsymbol{\theta}_\gamma^\top)^\top$, $\boldsymbol{\theta}_y = (\boldsymbol{\beta}^\top, \varphi)^\top$, and

$$\ell_n(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{X}) = \sum_{i=1}^{n} \ln \int \left\{ \prod_{k=1}^{N_i} f(Y_{ik} \mid \boldsymbol{X}_{ik}, \gamma; \boldsymbol{\theta}_y) \right\} g(\gamma \mid \boldsymbol{\theta}_\gamma) d\gamma. \tag{3}$$

In all of our numerical studies, we use the following penalty proposed by Chen and Li [7]

$$p_n(\sigma^2; \widehat{\sigma}^2_{pilot}) = -a_n \{ \widehat{\sigma}^2_{pilot}/\sigma^2 + \ln(\sigma^2/\widehat{\sigma}^2_{pilot}) - 1 \}, \tag{4}$$

where $\widehat{\sigma}^2_{pilot}$ is a pilot estimate for the variance of γ. One possible choice of $\widehat{\sigma}^2_{pilot}$ is the variance estimator assuming the γ_i are i.i.d. Gaussian variables. When $a_n = o_p(n^{1/4})$, the penalty function in (4) satisfies the assumptions for our asymptotic theory. A similar requirement on a_n is made by Chen et al. [8]. In all of our numerical studies, we choose a_n using the empirical formula (23) in Kasahara and Shimotsu [19].

To derive an EM algorithm, define $\boldsymbol{L}_i = (L_{i1}, \ldots, L_{iC})^\top$ follows multinomial distribution $\mathcal{M}(1; \pi_1, \ldots, \pi_C)$ as a latent random vector of subpopulation memberships, where $L_{ic} = 1$ if γ_i belongs to component c and $L_{ic} = 0$ otherwise. Then the likelihood function for the complete data, comprising of both observed and latent variables, is $\ell_{comp}(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{L}) = \sum_{i=1}^{n} \ell_{i,comp}(\boldsymbol{\theta}; \boldsymbol{Y}_i, \boldsymbol{X}_i, \gamma_i, \boldsymbol{L}_i)$, where $\ell_{i,comp}(\boldsymbol{\theta}; \boldsymbol{Y}_i, \boldsymbol{X}_i, \gamma_i, \boldsymbol{L}_i) = \ln f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma_i; \boldsymbol{\theta}_y) + \sum_{c=1}^{C} L_{ic}[\ln \pi_c - \frac{1}{2}\ln(\sigma_c^2) + \ln \phi\{(\gamma_i - \mu_c)/\sigma_c\}]$ and $f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma_i; \boldsymbol{\theta}_y) = \prod_{k=1}^{N_i} f(Y_{ik} \mid \boldsymbol{X}_{ik}, \gamma_i; \boldsymbol{\theta}_y)$. We estimate the parameters by maximizing the penalized likelihood while treating $\boldsymbol{\gamma}$ and $\boldsymbol{L}$ as missing data. The detailed algorithm is provided in the supplementary material.

2.3. Consistency of the estimator

The parameter space for a model with exactly C mixture components is $\Theta_C = \{\boldsymbol{\theta} \mid \boldsymbol{\beta} \in \mathbb{R}^p,\ \sum_{c=1}^{C} \pi_c = 1,\ 0 < \pi_c < 1,\ \mu_1 < \cdots < \mu_C,\ \sigma_c > 0,\ c \in \{1, \ldots, C\}\}$. The closure of Θ_C is $\bar{\Theta}_C = \{\boldsymbol{\theta} \mid \boldsymbol{\beta} \in \mathbb{R}^p,\ \sum_{c=1}^{C} \pi_c = 1,\ 0 \le \pi_c \le 1,\ \mu_1 \le \cdots \le \mu_C,\ \sigma_c \ge 0,\ c \in \{1, \ldots, C\}\}$, which also includes the over-fitted models. In other words, $\bar{\Theta}_C$ admits models with the true number of components C_0 strictly less than C, in which case a redundant component c can be parameterized in $\bar{\Theta}_C$ in multiple ways, such as setting either $\pi_c = 0$ or $(\mu_c, \sigma_c) = (\mu_{c'}, \sigma_{c'})$ for some $c' \neq c$. Let $\boldsymbol{\theta}_0 \in \bar{\Theta}_C$ be one parameterization for the true density of γ, and $f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta})$ be the joint distribution function of $(\boldsymbol{X}, \boldsymbol{Y})$ associated with the likelihood in (3). Following Hathaway [13], define

$$\mathcal{F}_{\boldsymbol{\theta}_0} = \left\{ \boldsymbol{\theta} \in \bar{\Theta}_C : \int_{-\infty}^{(\boldsymbol{x}', \boldsymbol{y}')} f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}) d\mu(\boldsymbol{x}, \boldsymbol{y}) = \int_{-\infty}^{(\boldsymbol{x}', \boldsymbol{y}')} f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}_0) d\mu(\boldsymbol{x}, \boldsymbol{y}) \text{ for any } (\boldsymbol{x}', \boldsymbol{y}') \right\}.$$

All parameters in $\mathcal{F}_{\boldsymbol{\theta}_0}$ lead to the same mixture density for γ, stressing the lack of identifiability in finite Gaussian mixture models and their fundamental difference from other commonly used parametric models. Denote the maximum penalized likelihood estimator under a C-component mixture model by $\widehat{\boldsymbol{\theta}}_C = \arg\max_{\boldsymbol{\theta} \in \bar{\Theta}_C} \ell_{pen}(\boldsymbol{\theta})$.

Proposition 1. *Under Assumptions 1-6 in Section 8, $\widehat{\boldsymbol{\theta}}_C$ is consistent in the sense that $\inf_{\boldsymbol{\theta}^* \in \mathcal{F}_{\boldsymbol{\theta}_0}} \|\widehat{\boldsymbol{\theta}}_C - \boldsymbol{\theta}^*\| \to 0$ in probability.*

The proof of Proposition 1 is provided in Section 8, which extends the arguments in Chen et al. [9] to the GLMM framework. Proposition 1 implies that we can estimate the mixture density consistently, but this is not necessarily true for the parameters since $\boldsymbol{\theta}_\gamma$ is not unique if we over fit the model by including more mixture components.

3. Deciding the Number of Mixture Components

3.1. Hypothesis tests on the order of the latent Gaussian mixture model

Deciding the number of mixture components is key in addressing the heterogeneity across units. In the context of transplant center evaluation, this is about detecting whether there are subgroups of transplant centers that are under-performing or out-performing the rest. There are two commonly used approaches, the model selection approach

[17, 34] and the hypothesis testing approach [8]. The model selection approach seeks a model to adequately describe the data, while the hypothesis testing approach is used to validate scientific claims. In this paper, we focus on the hypothesis testing approach because it quantifies the confidence of our decisions by providing p-values.

Among the many hypotheses that we can test, the most important one is $H_0: C_0 = 1$ vs $H_1: C_0 = 2$, where C_0 is the true number of components. This test is also referred to as the homogeneity test, since the null hypothesis means all transplant centers are from the same homogeneous population with no anomalies. Chen et al. [8] provided more examples where different orders of mixture models have different scientific interpretations that require testing.

Even though hypothesis tests are not designed for model selection, they can nevertheless be used for such a purpose in an exploratory study. One can determine the order of the latent Gaussian mixture model by sequentially testing $H_{01}: C_0 = 1$, $H_{02}: C_0 = 2$, $H_{03}: C_0 = 3, \ldots$ at level α respectively, and declare $C_0 = \tilde{C}$ if $H_{0\tilde{C}}$ is the first null hypothesis in the sequence that is not rejected. With $\lim_{n\to\infty} \Pr(H_{0C} \text{ being rejected}) = 1$ for any $C < C_0$ and $\lim_{n\to\infty} \Pr(H_{0C_0} \text{ being rejected}) = \alpha$, the sequential test procedure chooses the correct component number with a probability tending to $1-\alpha$ and over-selects the component number with a probability tending to α. This is obviously not a consistent model selection procedure, since we have a positive chance of falsely reject a hypothesis if α is fixed. On the other hand, one can also argue that many widely used model selection procedures are not consistent, such as the Akaike information criterion (AIC) [1]. In our simulation studies, we show that the sequential test procedure that we propose can vastly outperform the Bayesian information criterion (BIC) [31] in model selection when the sample size is moderate. There has been some work on model selection in linear mixed models (LMM) using AIC type of criteria [28, 35], however these methods are not applicable to GLMM and are not designed for selecting the number of components in Gaussian Mixture Models.

Due to the loss of strong identifiability for finite Gaussian mixture models, regular asymptotic theory for likelihood ratio tests (LRT) does not hold. Instead, Chen et al. [8] and Kasahara and Shimotsu [19] proposed locally restricted likelihood ratio tests that confine the parameter space in local alternative models to ensure the existence of asymptotic distributions for the test statistics. We extend such tests to the proposed latent Gaussian mixture models.

3.2. Homogeneity test

We first consider $H_0: C_0 = 1$ vs $H_1: C_0 = 2$. We refer to the model under the null hypothesis as the reduced model and the one under the alternative as the full model. When the null hypothesis is true, γ_i are i.i.d. random variables following $\mathcal{N}(\mu_0, \sigma_0^2)$. However, this model is not uniquely parameterized in the full model, for example, π_1 can be any value between 0 and 1 when $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$. Following Chen et al. [8], we restrict the parameter space under the full model to $\bar{\Theta}_2(\tau) = \{\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi_1, \pi_2)^\top;\ \mu_1, \mu_2 \in \mathbb{R}, \sigma_1, \sigma_2 \geq 0, \pi_1 = \tau, \pi_2 = 1-\tau\}$, for a fixed $\tau \in (0, 0.5]$. By doing so, we do not impose any constraints on the order between μ_1 and μ_2. In $\bar{\Theta}_2(\tau)$, the null model is uniquely parameterized by $\boldsymbol{\theta}_0(\tau) = \{\boldsymbol{\theta}_{y,0}^\top, \boldsymbol{\theta}_{\gamma,0}^\top(\tau)\}^\top$, where $\boldsymbol{\theta}_{\gamma,0}(\tau) = (\mu_0, \mu_0, \sigma_0^2, \sigma_0^2, \tau, 1-\tau)^\top$.

Let $\bar{\Theta}_1 = \{\boldsymbol{\theta} = (\mu, \sigma^2)^\top;\ \mu \in \mathbb{R}, \sigma \geq 0\}$ be the parameter space under the null hypothesis $C_0 = 1$, which is nested in $\bar{\Theta}_2$. Denote the reduced model estimator as $\widehat{\boldsymbol{\theta}}_{red} = \arg\max_{\boldsymbol{\theta}\in\bar{\Theta}_1} \ell_{pen}(\boldsymbol{\theta})$, which is the usual maximum likelihood estimator for a GLMM under the Gaussian random effect assumption. Under the full model, the estimator with a fixed τ is $\widehat{\boldsymbol{\theta}}_{full}(\tau) = \arg\max_{\boldsymbol{\theta}\in\bar{\Theta}_2(\tau)} \ell_{pen}(\boldsymbol{\theta})$. This estimator can be obtained using the EM algorithm described in the supplementary material without the step for updating π_c's. The following proposition provides the convergence rate of $\widehat{\boldsymbol{\theta}}_{full}(\tau)$ when the null hypothesis is true, the proof of which is provided in Section 8.

Proposition 2. *Under $H_0: C_0 = 1$ and Assumptions 1-7 in Section 8, for any fixed $\tau \in (0, 0.5]$, $\widehat{\boldsymbol{\beta}}_{full}(\tau) - \boldsymbol{\beta}_0 = O_p(n^{-1/2})$, $\widehat{\mu}_{c,full}(\tau) - \mu_0 = O_p(n^{-1/8})$ and $\widehat{\sigma}^2_{c,full}(\tau) - \sigma_0^2 = O_p(n^{-1/4})$ for $c \in \{1, 2\}$ where $\widehat{\boldsymbol{\beta}}_{full}$, $\widehat{\mu}_{c,full}$ and $\widehat{\sigma}^2_{c,full}$ are components in $\widehat{\boldsymbol{\theta}}_{full}(\tau)$ while $\boldsymbol{\beta}_0$, μ_0 and σ_0^2 are the true parameters.*

Remark 1. We use a reparameterization similar to that of Kasahara and Shimotsu [19] in the proof of Proposition 2. As shown in the proof, many derivatives of the log likelihood are either exactly zero or have mean zero, and it takes a ninth order Taylor expansion to get a local quadratic approximation for the penalized likelihood. The convergence rate in the proposition means that, for an over-fitted mixture model, the regression coefficient $\boldsymbol{\beta}$ still enjoys the root-n convergence rate, while the parameters of the latent Gaussian mixture model converge much slower. This slow convergence rate also stresses a fundamental difference between our latent Gaussian mixture model and the common parametric models. The $O_p(n^{-1/8})$ convergence rate in $\widehat{\mu}_{c,full}$ is in agreement with the minimax lower bound established in Ho and Nguyen [15] for finite Gaussian mixture models with one redundant component.

Let $\mathcal{T}$ be a finite subset of numbers in $(0, 0.5]$, define the test statistic

$$\widetilde{T}_1 = \max_{\tau \in \mathcal{T}} T_1(\tau), \qquad T_1(\tau) = 2[\ell_n\{\widehat{\boldsymbol{\theta}}_{full}(\tau)\} - \ell_n(\widehat{\boldsymbol{\theta}}_{red})]. \tag{5}$$

The following proposition provides the asymptotic distribution of $\widetilde{T}_1$, the proof of which is provided in Section 8.

Proposition 3. *Under $H_0 : C_0 = 1$ and Assumptions 1-7, $\widetilde{T}_1 \to \chi^2(2)$ in distribution as $n \to \infty$.*

Remark 2. Our proof of Proposition 3 shows that, under $H_0 : C_0 = 1$, $T_1(\tau) \to \chi^2(2)$ in distribution for any fixed τ. In fact, if there is only one true component, no matter how we choose to split that component, the leading term in the asymptotic expansion of $T_1(\tau)$ remains the same. We define $\widetilde{T}_1$ as the maximum of $T_1(\tau)$ over $\mathcal{T}$ to increase the power: if H_1 is true, the more values of τ we try, the better chance we have to detect an extra component. Intuitively, bigger $\mathcal{T}$ leads to higher power, but also a heavier burden in computation. Empirical studies in Chen et al. [8] suggest that $\mathcal{T} = \{0.1, 0.3, 0.5\}$ provides a good balance between statistical power and computational cost. We follow this recommendation in all numerical studies in this paper. The condition $a_n = o_p(n^{1/4})$ guarantees that the asymptotic distribution of test statistic is not affected by penalty (4) in estimation.

The detailed test procedure is as follows.

Step 0. Obtain $\widehat{\boldsymbol{\theta}}_{red}$ and $\ell_n(\widehat{\boldsymbol{\theta}}_{red})$.

Step 1. For a fixed τ, obtain $\widehat{\boldsymbol{\theta}}_{full}(\tau)$. To increase the chance of reaching the global maximum of the penalized likelihood, try 100 randomly selected initial values.

Step 2. (Optional) Using $\widehat{\boldsymbol{\theta}}_{full}(\tau)$ obtained in Step 1 as the starting value, perform two more EM iterations without fixing τ, and use the resulting estimator to evaluate $T_1(\tau)$.

Step 3. Repeat Steps 1 and 2 for each $\tau \in \mathcal{T}$ to obtain $\widetilde{T}_1$, where $\mathcal{T}$ is set to be $\{0.1, 0.3, 0.5\}$ following the recommendation of Chen et al. [8].

Step 4. For a size α test, reject $H_0 : C_0 = 1$ if $\widetilde{T}_1 > \chi^2_\alpha(2)$.

In Step 2, we perform two more EM iterations without fixing τ to increase the power of the test, as recommended by Chen et al. [8].

3.3. Testing for C greater than 1

Next, we consider a test $H_0 : C_0 = C$ vs $H_1 : C_0 = C + 1$ for a $C \geq 2$. We now refer to the model with C components as the reduced model and the one with $C + 1$ components as the full model. We first compute the reduced model estimator $\widehat{\boldsymbol{\theta}}_{red} = \arg\max_{\boldsymbol{\theta} \in \bar{\Theta}_C} \ell_{pen}(\boldsymbol{\theta})$. Assuming H_0 is true, denote the true value of the parameter by $\boldsymbol{\theta}_0$ and order the true mean parameters by $\mu_{1,0} < \cdots < \mu_{C,0}$. This parameter is not uniquely identified in the full model: if any $\pi_c = 0$ or $(\mu_c, \sigma_c) = (\mu_{c+1}, \sigma_{c+1})$ for some $c \in \{1, \ldots, C\}$, the full model degenerates to the reduced model. In order to make the reduced model identifiable in $\bar{\Theta}_{C+1}$, we will impose constraints that $\pi_c > 0$ for all $c \in \{1, \ldots, C + 1\}$ and $\pi_c/(\pi_c + \pi_{c+1}) = \tau$ for some c and a fixed $\tau \in (0, 0.5]$ like we did in Section 3.2.

To test if a $(C+1)$-component mixture model fits the data better, we will test to see if any one of the C components in the reduced model can be further split into two. Define non-overlapping intervals $D_1, \ldots, D_C$ such that $\mu_{c,0} \in D_c$. For a fixed $\tau \in (0, 0.5]$ and $c \in \{1, \ldots, C\}$, define neighborhoods in the parameter space $\bar{\Theta}_{C+1}$:

$$\begin{aligned} \mathcal{N}_{C+1}(c, \tau) \quad = \quad & \big\{\boldsymbol{\theta} \in \bar{\Theta}_{C+1} \mid \pi_c/(\pi_c + \pi_{c+1}) = \tau;\ \mu_{c'} \in D_{c'} \text{ for } 1 \leq c' < c; \\ & \qquad \mu_c, \mu_{c+1} \in D_c; \mu_{c'} \in D_{c'-1} \text{ for } c + 1 < c' \leq C\big\}. \end{aligned}$$

The neighborhood $\mathcal{N}_{C+1}(c, \tau)$ collects the parameters that split the cth component into two daughter components with a split proportion τ, while restricting the other mean parameters from changing too much. The definition of $\mathcal{N}_{C+1}(c, \tau)$ requires knowledge about intervals $\{D_1, \ldots, D_C\}$ that contain the true mean parameters. In practice, we already have a consistent estimator of $\mu_{c,0}$ from fitting the reduced model. Replacing $\{D_c\}_{c=1}^C$ with their consistent estimates does not affect the asymptotic behavior of the test we are about to propose. A practical choice for $\{D_c\}_{c=1}^C$ is provided below in the test procedure. Like in Section 3.2, we do not restrict order between μ_c and μ_{c+1} in $\mathcal{N}_{C+1}(c, \tau)$ because τ is restricted to $(0, 0.5]$.

Define the locally restricted full model estimator as

$$\widehat{\boldsymbol{\theta}}_{full}(c,\tau) = \arg\max_{\boldsymbol{\theta}\in\mathcal{N}_{C+1}(c,\tau)} \ell_{pen}(\boldsymbol{\theta}).$$

To obtain this estimator, we need some minor adjustments to the EM algorithm in Section 2.2. First, we update $\pi_c + \pi_{c+1}$ as a single parameter and then assign values for π_c and π_{c+1} proportional to τ. Second, after each M-step, we enforce the restrictions in $\mathcal{N}_{C+1}(c,\tau)$ by forcing any $\mu_{c'}$ stepping out of the boundary back to its predetermined range. A similar scheme was used in Chen et al. [8]. The following convergence rate result echoes Proposition 2. It shows that the component that we are trying to split suffers a slower convergence rate, because it is overfitted in $\mathcal{N}_{C+1}(c,\tau)$ as a mixture of two daughter components, and the rest of the parameters converge in root-n rate.

Proposition 4. *Under $H_0 : C_0 = C$ and Assumptions 1-8 in Section 8, for any fixed $\tau \in (0, 0.5]$,*

$$\widehat{\mu}_{c,full}(c,\tau) - \mu_{c,0} = O_p(n^{-1/8}), \quad \widehat{\mu}_{c+1,full}(c,\tau) - \mu_{c,0} = O_p(n^{-1/8}),$$
$$\widehat{\sigma}^2_{c,full}(c,\tau) - \sigma^2_{c,0} = O_p(n^{-1/4}), \quad \widehat{\sigma}^2_{c+1,full}(c,\tau) - \sigma^2_{c,0} = O_p(n^{-1/4}),$$

and $\widehat{\boldsymbol{\theta}}_{y,full}(c,\tau) - \boldsymbol{\theta}_{y0} = O_p(n^{-1/2})$, $\widehat{\boldsymbol{\theta}}_{\gamma,c',full}(c,\tau) - \boldsymbol{\theta}_{\gamma,c',0} = O_p(n^{-1/2})$ for $c' < c$, $\widehat{\boldsymbol{\theta}}_{\gamma,c',full}(c,\tau) - \boldsymbol{\theta}_{\gamma,c'-1,0} = O_p(n^{-1/2})$ for $c' > c+1$, where $\boldsymbol{\theta}_{\gamma,c'} = (\mu_{c'}, \sigma^2_{c'}, \pi_{c'})^\top$.

The proof of Proposition 4 is relegated to Section 8. To test if any component in the reduced model can be further divided into two, define the test statistic

$$T_C(\tau) = \max_{c\in\{1,\ldots,C\}} T_C(c,\tau), \quad \text{where } T_C(c,\tau) = 2[\ell_n\{\widehat{\boldsymbol{\theta}}_{full}(c,\tau)\} - \ell_n(\widehat{\boldsymbol{\theta}}_{red})]. \tag{6}$$

For any finite subset $\mathcal{T}$ of the interval $(0, 0.5]$, define the test statistic

$$\widetilde{T}_C = \max_{\tau\in\mathcal{T}} T_C(\tau). \tag{7}$$

In order to understand the asymptotic behavior of $T_C(c,\tau)$, we adopt the reparameterization of Kasahara and Shimotsu [19] in $\mathcal{N}_{C+1}(c,\tau)$. Define the new parameter vector as $\boldsymbol{\psi}(c,\tau) = \{\boldsymbol{\theta}_y^\top, \boldsymbol{\delta}(c)^\top, \boldsymbol{\mu}(c)^\top, \boldsymbol{\sigma}^2(c)^\top, \lambda_\mu, \lambda_\sigma\}^\top$ such that

$$\begin{pmatrix} \mu_c \\ \mu_{c+1} \\ \sigma^2_c \\ \sigma^2_{c+1} \end{pmatrix} = \begin{Bmatrix} \nu_\mu + (1-\tau)\lambda_\mu \\ \nu_\mu - \tau\lambda_\mu \\ \nu_\sigma + (1-\tau)(2\lambda_\sigma - \frac{1+\tau}{3}\lambda^2_\mu) \\ \nu_\sigma - \tau(2\lambda_\sigma + \frac{2-\tau}{3}\lambda^2_\mu) \end{Bmatrix}, \tag{8}$$

and

$$\begin{array}{rcl} \boldsymbol{\delta}(c) & = & (\pi_1, \ldots, \pi_{c-1}, \quad \pi_c + \pi_{c+1}, \quad \pi_{c+2}, \ldots, \pi_C)^\top, \\ \boldsymbol{\mu}(c) & = & (\mu_1, \ldots, \mu_{c-1}, \quad \nu_\mu, \quad \mu_{c+2}, \ldots, \mu_C, \ \mu_{C+1})^\top, \\ \boldsymbol{\sigma}^2(c) & = & (\sigma^2_1, \ldots, \sigma^2_{c-1}, \quad \nu_\sigma, \quad \sigma^2_{c+2}, \ldots, \sigma^2_C, \ \sigma^2_{C+1})^\top. \end{array} \tag{9}$$

Denote the new parameter space as $\bar{\Theta}_{\psi,C+1}$ and partition $\boldsymbol{\psi}$ into $(\boldsymbol{\eta}^\top, \boldsymbol{\lambda}^\top)^\top$ where

$$\boldsymbol{\eta} = \{\boldsymbol{\theta}_y^\top, \boldsymbol{\delta}(c)^\top, \boldsymbol{\mu}(c)^\top, \boldsymbol{\sigma}^2(c)^\top\}^\top, \quad \boldsymbol{\lambda} = (\lambda_\mu, \lambda_\sigma)^\top.$$

The reduced model is uniquely parameterized by $\boldsymbol{\theta}^* \in \mathcal{N}_{C+1}(c,\tau)$, and it is reparameterized as $\boldsymbol{\psi}^* = \{(\boldsymbol{\eta}^*)^\top, 0, 0\}^\top$, or more specifically $\boldsymbol{\theta}_y = \boldsymbol{\theta}_{y,0}$, $\boldsymbol{\lambda}^* = 0$ and $\boldsymbol{\delta}^*(c) = (\pi_{1,0}, \ldots, \pi_{C-1,0})^\top$, $\boldsymbol{\mu}^*(c) = (\mu_{1,0}, \ldots, \mu_{C,0})^\top$, $\boldsymbol{\sigma}^{2*}(c) = (\sigma^2_{1,0}, \ldots, \sigma^2_{C,0})^\top$. The reparameterization in (8) is beneficial because, to test if the cth component can be further split, we can equivalently test if $\boldsymbol{\lambda} = 0$.

Define

$$\boldsymbol{s}_i^{(c)} = \left\{\boldsymbol{s}_{\boldsymbol{\eta},i}^\top, (\boldsymbol{s}_{\boldsymbol{\lambda},i}^{(c)})^\top\right\}^\top, \tag{10}$$

where $\boldsymbol{s}_{\boldsymbol{\eta},i} = (\boldsymbol{s}_{\boldsymbol{\theta}_y,i}^\top, \boldsymbol{s}_{\boldsymbol{\delta},i}^\top, \boldsymbol{s}_{\mu,i}^\top, \boldsymbol{s}_{\sigma,i}^\top)^\top$, $\boldsymbol{s}_{\lambda,i}^{(c)} = \left(\int \zeta_i \pi_c f_{c,i}^* H_{ci}^{3*} / \int \zeta_i g^*, \int \zeta_i \pi_c f_{c,i}^* H_{ci}^{4*} / \int \zeta_i g^*\right)^\top$, $\boldsymbol{s}_{\boldsymbol{\theta}_y,i} = \int (\partial \zeta_i / \partial \boldsymbol{\theta}_y) g^* / \int \zeta_i g^*$, $\boldsymbol{s}_{\boldsymbol{\delta},i} = \{\int \zeta_i (f_{1,i}^* - f_{C,i}^*) / \int \zeta_i g_i^*, \ldots, \int \zeta_i (f_{C-1,i}^* - f_{C,i}^*) / \int \zeta_i g_i^*\}^\top$, $\boldsymbol{s}_{\mu,i} = (\int \zeta_i \pi_1 f_{1,i}^* H_{1i}^{1*} / \int \zeta_i g^*, \ldots, \int \zeta_i \pi_C f_{C,i}^* H_{Ci}^{1*} / \int \zeta_i g^*)^\top$ and $\boldsymbol{s}_{\sigma,i} = (\int \zeta_i \pi_1 f_{1,i}^* H_{1i}^{2*} / \int \zeta_i g^*, \ldots, \int \zeta_i \pi_C f_{C,i}^* H_{Ci}^{2*} / \int \zeta_i g^*)^\top$. Here, we use the short hand notation $\zeta_i = \prod_{k=1}^{N_i} f(y_{ik} \mid \boldsymbol{x}_{ik}, \gamma_i; \boldsymbol{\theta}_y)$, $f_{c,i}^* = f_c(\gamma_i \mid \mu_{c,0}, \sigma_{c,0})$, $g_i^* = g(\gamma_i \mid \boldsymbol{\theta}_\gamma^*)$ and $H_{ci}^{k*} = H^k\{(\gamma_i - \mu_{c,0})/\sigma_{c,0}\}/(k! \sigma_{c,0}^k)$, where $H^k(\cdot)$ is the kth Hermite Polynomial. The following proposition provides the asymptotic distribution of $\widetilde{T}_C$, the proof of which is provided in Section 8.

Proposition 5. *Under $H_0 : C_0 = C$ and Assumptions 1-8 listed in Section 8,*

$$\widetilde{T}_C \to \max\left\{(\boldsymbol{S}_{\lambda|\eta,n}^{(c)})^\top (\boldsymbol{\mathcal{I}}_{\lambda|\eta}^{(c)})^{-1} \boldsymbol{S}_{\lambda|\eta,n}^{(c)}, c \in \{1, \ldots, C\}\right\} \text{ in distribution,}$$

where $\boldsymbol{S}_{\lambda|\eta,n}^{(c)} = \boldsymbol{S}_{\lambda,n}^{(c)} - \boldsymbol{\mathcal{I}}_{\lambda\eta}^{(c)} \boldsymbol{\mathcal{I}}_\eta^{-1} \boldsymbol{S}_{\eta,n}$, $\boldsymbol{\mathcal{I}}_{\lambda|\eta}^{(c)} = \boldsymbol{\mathcal{I}}_\lambda^{(c)} - \boldsymbol{\mathcal{I}}_{\lambda\eta}^{(c)} \boldsymbol{\mathcal{I}}_\eta^{-1} (\boldsymbol{\mathcal{I}}_{\lambda\eta}^{(c)})^\top$, $\boldsymbol{S}_{\eta,n} = n^{-1/2} \sum_{i=1}^n \boldsymbol{s}_{\eta,i}$, $\boldsymbol{S}_{\lambda,n}^{(c)} = n^{-1/2} \sum_{i=1}^n \boldsymbol{s}_{\lambda,i}^{(c)}$, $\boldsymbol{\mathcal{I}}_{\lambda\eta}^{(c)} = \mathrm{E}\{\boldsymbol{s}_{\boldsymbol{\lambda},i}^{(c)} \boldsymbol{s}_{\boldsymbol{\eta},i}^\top\}$, $\boldsymbol{\mathcal{I}}_\eta = \mathrm{E}(\boldsymbol{s}_{\eta,n} \boldsymbol{s}_{\eta,n}^\top)$, *and* $\boldsymbol{\mathcal{I}}_\lambda^{(c)} = E\{\boldsymbol{s}_{\boldsymbol{\lambda},i}^{(c)} (\boldsymbol{s}_{\boldsymbol{\lambda},i}^{(c)})^\top\}$.

One can show $(\boldsymbol{S}_{\lambda|\eta,n}^{(c)})^\top (\boldsymbol{\mathcal{I}}_{\lambda|\eta}^{(c)})^{-1} \boldsymbol{S}_{\lambda|\eta,n}^{(c)} \to \chi^2(2)$ in distribution for each c, but the score vectors $\boldsymbol{S}_{\lambda|\eta,n}^{(c)}$ are correlated across different c's and hence the distribution of $\widetilde{T}_C$ in Proposition 5 is that of the maximum of a few correlated $\chi^2(2)$ random variables. In S2 of the supplementary material, we describe a simulation method to evaluate this asymptotic distribution. This procedure only requires estimating the covariance matrix of $\{\boldsymbol{S}_{\lambda|\eta,n}^{(c)}, c \in \{1, \ldots, C\}\}$ and simulating Gaussian random variables. It is extremely fast and fundamentally different from bootstrap, which requires fitting the model a large number of times to the bootstrap samples.

For any $C \geq 2$, our test procedure for $H_0 : C_0 = C$ is as follows.

Step 0. Obtain $\widehat{\boldsymbol{\theta}}_{red}$ and evaluate $\ell_n(\widehat{\boldsymbol{\theta}}_{red})$. Define subintervals $D_1 = [\widehat{\gamma}_{min},\ \widehat{\mu}_{1,red}/2 + \widehat{\mu}_{2,red}/2]$, $D_2 = (\widehat{\mu}_{1,red}/2 + \widehat{\mu}_{2,red}/2,\ \widehat{\mu}_{3,red}/2 + \widehat{\mu}_{2,red}/2], \ldots, D_C = (\widehat{\mu}_{C-1,red}/2 + \widehat{\mu}_{C,red}/2,\ \widehat{\gamma}_{max}]$, where $\widehat{\gamma}_{min}$ and $\widehat{\gamma}_{max}$ are the minimum and maximum of the predicted γ's under the reduced model.

Step 1. Obtain $\widehat{\boldsymbol{\theta}}_{full}(c, \tau)$ by maximizing the penalized likelihood in the restricted parameter neighborhood $\mathcal{N}_{C+1}(c, \tau)$ using the subintervals $\{D_k\}_{k=1}^C$ defined in Step 0. The penalty on σ_k^2 is $p_n(\sigma_k^2, \widehat{\sigma}_{c',red}^2)$ if μ_k is restricted in $D_{c'}$, $k \in \{1, \ldots, C+1\}$, and a_n is chosen according equation (23) in Kasahara and Shimotsu [19]. If a μ_k steps outside of its range $D_{c'}$ specified by $\mathcal{N}_{C+1}(c, \tau)$ during the EM iterations, we simply set it back to the nearest boundary of $D_{c'}$. To ensure that the maximum of ℓ_{pen} is reached, we repeat the EM algorithm 100 times using randomly selected initial values within $\mathcal{N}_{C+1}(c, \tau)$.

Step 2. Using $\widehat{\boldsymbol{\theta}}_{full}(c, \tau)$ as the starting value, conduct two more EM iterations without fixing τ. Use the resulting estimator to evaluate $T_C(c, \tau)$ in (6).

Step 3. Repeat Steps 1 and 2 for each $c \in \{1, \ldots, C\}$ and $\tau \in \mathcal{T} = \{0.1, 0.3, 0.5\}$, and evaluate $\widetilde{T}_C$ in (7).

Step 4. Evaluate the null distribution in Proposition 5 using the procedure described in S2 of the supplementary material and compare $\widetilde{T}_C$ with the null distribution to get the p value.

4. Use False Discovery Rate Control to Classify Units

A practical utility of model (1) is to classify units based on γ_i. To ease understanding, we frame the ensuing development in the context of the aforementioned transplant center evaluation. That is, different components in the mixture density $g(\gamma)$ represent different clusters in health care quality delivered by transplant centers, and we want to classify the transplant centers into these clusters. However, these clusters are not considered to be equal: usually a subset of clusters, denoted as $\mathcal{C}_0$, represent the norm of care quality, consisting of centers with average performances; those out of $\mathcal{C}_0$ are centers either underperforming or outperforming the industrial standard. Following Efron's "empirical null" idea [11], $\mathcal{C}_0 \subset \{1, \ldots, C\}$ can be identified as one or more components in the fitted mixture model, usually those in the middle of $g(\gamma)$ with high weights π_c's.

With $\mathcal{C}_0$ representing the distribution of normal care quality, one should classify an individual center into clusters outside of $\mathcal{C}_0$ with extreme care, since it declares that center as an anomaly, and the false discovery rate needs to be controlled. As pointed out in Sun et al. [33], classification problems with unequal losses in different classes

are naturally connected with multiple hypothesis tests. In our context, this classification problem is equivalent to performing a test for each center on whether the center is in the empirical null $\mathcal{C}_0$. In other words, we test a sequence of hypotheses $H_{i0} : \sum_{c\in\mathcal{C}_0} L_{ic} = 1$, $i \in \{1,\ldots,n\}$. Since $\mathcal{C}_0$ represents the average quality of care, center i is considered "interesting" (either outperforming or underperforming) if H_{i0} is rejected.

For a given subset of components $\mathcal{C}_0$, identify the "empirical null" distribution of γ as

$$g_0(\gamma \mid \boldsymbol{\theta}_\gamma) = \textstyle\sum_{c\in\mathcal{C}_0} \pi_c f_c(\gamma \mid \mu_c, \sigma_c) / \sum_{c\in\mathcal{C}_0} \pi_c.$$

Since γ_i is not directly observed, our decision rule for H_{i0} is based on the observed data $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$, denoted as $\delta_i = \delta(\boldsymbol{X}_i, \boldsymbol{Y}_i; \boldsymbol{\theta})$, where $\delta_i = 1$ means center i is "interesting" and $\delta_i = 0$ otherwise. The false discovery rate is defined as

$$FDR = \mathrm{E}\left\{ \frac{\sum_i^n I(\delta_i = 1, \sum_{c\in\mathcal{C}_0} L_{ic} = 1)}{\sum_i^n I(\delta_i = 1)} \mid \sum_i^n I(\delta_i = 1) > 0 \right\} \Pr\left\{ \sum_i^n I(\delta_i = 1) > 0 \right\}.$$

When γ_i's are observed, Sun and Cai [32] show that the oracle decision rule is based on the local FDR, $T_{\mathrm{OR}}(\gamma_i) = \Pr(\sum_{c\in\mathcal{C}_0} L_{ic} = 1 \mid \gamma_i) = \sum_{c\in\mathcal{C}_0} \pi_c f_c(\gamma_i)/g(\gamma_i)$. In our case, γ_i is not observed, and the local FDR is defined as the posterior probability given the observed data

$$lFDR_i = \Pr(\textstyle\sum_{c\in\mathcal{C}_0} L_{ic} = 1 \mid \boldsymbol{X}_i, \boldsymbol{Y}_i) = \frac{\sum_{c\in\mathcal{C}_0} \pi_c \int f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma; \boldsymbol{\beta}) f_c(\gamma \mid \mu_c, \sigma_c) d\gamma}{\int f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \gamma; \boldsymbol{\beta}) g(\gamma \mid \boldsymbol{\theta}_\gamma) d\gamma}. \tag{11}$$

It is easy to show $lFDR_i = \mathrm{E}\{T_{\mathrm{OR}}(\gamma_i) \mid \boldsymbol{X}_i, \boldsymbol{Y}_i\}$. Following Sun et al. [33], the multiple hypothesis testing problem is related to a classification problem with the loss function

$$\mathscr{L}(\boldsymbol{L}, \boldsymbol{\delta}) = \lambda \textstyle\sum_i \delta_i (\sum_{c\in\mathcal{C}_0} L_{ic}) + \sum_i (1 - \delta_i)(1 - \sum_{c\in\mathcal{C}_0} L_{ic}),$$

where λ is a penalty for false positives. Let $\mathscr{R} = \mathrm{E}\{\mathscr{L}(\boldsymbol{L}, \boldsymbol{\delta})\}$ be the risk of the classification problem. By Theorem 1 of Sun et al. [33], the optimal decision rule that minimizes this risk is $\delta_i = I(lFDR_i < t)$ for some threshold t.

Let $lFDR_{(1)} \le \cdots \le lFDR_{(n)}$ be the ranked lFDR values. For any $\alpha' > 0$, let $k = \max_i\{\frac{1}{i}\sum_{j=1}^i lFDR_{(j)} \le \alpha'\}$ and our FDR control procedure is to reject all H_{i0} with the rank of $lFDR_i$ less or equal to k.

Proposition 6. *Under the model in (1), the above procedure controls FDR at level α'.*

A sketch proof of Proposition 6 is provided in Section 8. In practice, $lFDR$ is estimated by substituting $\boldsymbol{\theta}$ with its estimator and the integrals in (11) are evaluated using Gaussian quadrature as described in Section S1 of the supplementary material.

To elucidate the connection between our FDR approach with [18] as suggested by a reviewer, we note that [18] compared genes of healthy subjects with those of cancer patients, with a pre-determined group structure among genes based on the prior biological knowledge. The group memberships among the genes were fixed, and randomness or error in grouping was ignored. In contrast, our framework is unsupervised learning, wherein the groups or clusters among the transplant centers are unknown and detected using a mixture model.

5. Simulation Studies

5.1. Simulation 1: Estimation and random effect prediction

We simulate data for $n = 282$ transplant centers, which is the number of kidney transplant centers in the Organ Procurement and Transplantation Network in the year 2008. The number of patients per center has a highly skewed distribution in the real data. To mimic such a distribution, we generate N_i as the floor of the sum of Poisson distribution $\mathcal{P}(55)$ and exponential distribution $\mathcal{E}(95)$. The response Y_{ik} is a binary variable generated using (1) with $\Pr(Y_{ik} = 1) = \{1 + \exp(-\xi_{ik})\}^{-1}$, where $\xi_{ik} = \boldsymbol{X}_{ik}^\top \boldsymbol{\beta} + \gamma_i$, $\boldsymbol{X}$ is generated from a bivariate standard normal distribution and $\boldsymbol{\beta} = (1, 1)^\top$. We generate γ_i's from the following Gaussian mixture models

$$\text{Model 1: } 0.5\,\mathcal{N}(-3.26, 1.2^2) + 0.5\,\mathcal{N}(0.74, 0.8^2), \tag{12}$$

$$\text{Model 2: } 0.3\,\mathcal{N}(-5.26, 1.2^2) + 0.4\,\mathcal{N}(-0.26, 0.8^2) + 0.3\,\mathcal{N}(2.74, 0.9^2). \tag{13}$$

The parameters in these models are selected such that the marginal probability of $\{Y_{ik} = 1\}$ for each model is roughly the same as for the real data and the overall mean of the normal mixture is fixed at -1.26. We repeat the simulation 200 times under each model and apply the estimation procedure in Section 2.2 to each simulated data set. Estimation results for Model 1 and Model 2 in Simulation 1, under correctly specified number of components, are summarized in Table 1 and 2 respectively. The mixture components in the estimated model are ranked according to the value $\widehat{\mu}_c$ to avoid the cluster label switching problem. We can see that the estimation results are quite reasonable: all biases are virtually zero; the standard errors for component means (μ_c) and component standard deviations (σ_c) are slightly inflated compared with Table 1, which is understandable since we are fitting a more complicated mixture model; the standard errors for $\boldsymbol{\beta}$ are not affected by the increased complicity of the latent mixture model.

Table 1
Summary of parameter estimation under Simulation Model 1, which is a two-component latent Gaussian mixture model defined in (12). The table contains the true values of the parameters and the means, biases and standard deviations of the estimators. The results are based on 200 repetitions.

Parameter	Truth	Mean	Bias	Std
π_1	0.50	0.50	-0.00	0.03
π_2	0.50	0.50	0.00	0.03
μ_1	-3.26	-3.26	0.00	0.13
μ_2	0.74	0.74	-0.00	0.08
σ_1	1.20	1.20	-0.00	0.13
σ_2	0.80	0.80	-0.00	0.06
β_1	1.00	1.00	0.00	0.02
β_2	1.00	1.00	0.00	0.02

Table 2
Summary of parameter estimation under Simulation Model 2, which is a three-component latent Gaussian mixture model define in (13). The table contains the true values of the parameters and the means, biases and standard deviations of the estimators. The results are based on 200 repetitions.

Parameter	Truth	Mean	Bias	Std
π_1	0.30	0.30	0.00	0.02
π_2	0.40	0.39	-0.01	0.06
π_3	0.30	0.31	0.01	0.06
μ_1	-5.26	-5.28	-0.02	0.22
μ_2	-0.26	-0.27	-0.00	0.35
μ_3	2.74	2.69	-0.05	0.34
σ_1	1.20	1.18	-0.02	0.27
σ_2	0.80	0.80	0.00	0.19
σ_3	0.90	0.93	0.03	0.25
β_1	1.00	1.00	0.00	0.02
β_2	1.00	1.00	0.00	0.02

Fig. 1 illustrates the effect of model misspecification on random effect prediction. The data are generated in a typical simulation run under simulation Model 1. The left panel shows the prediction results of a common generalized linear mixed model under Gaussian random effect assumption, and the right panel shows the results of the proposed model. In both panels, we compare the true density of γ with the estimated density using the fitted model and the kernel density of the predicted γ using the fitted model. As we can see from the left panel, prediction under the misspecified Gaussian random effect assumption suffers from a shrinkage effect that the values of $\widehat{\gamma}$ are pushed towards

the center of the distribution so that the posterior distribution resembles the shape of a Gaussian distribution. The right panel shows that prediction under our proposed model does not suffer from such a shrinkage effect. Our model recovers the shape of the latent variable distribution and produces better predictions.

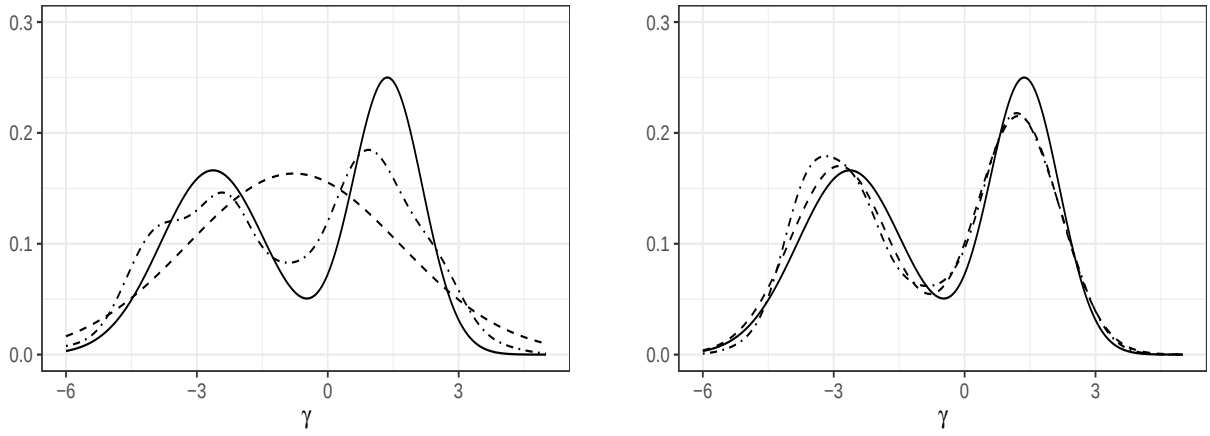


Fig. 1. Simulation Model 1: Impact of random effect assumption. Panel (a) shows results from a common generalized linear mixed model with a mis-specified Gaussian random effect assumption; Panel (b) shows results of the proposed latent Gaussian mixture model with a correctly specified number of components. In both panels, the solid curve is the true density for γ, the dashed curve is the estimated density of γ using the fitted model, and the dot-dash curve is the kernel density of the predicted random effects.

In Table 3, we also present the mean square prediction error of the proposed model averaged over 200 simulation runs, Monte Carlo standard deviation of the prediction error, and the same quantities under GLMM with Gaussian random effects. As we can see the prediction error under the common GLMM with Gaussian assumption has much bigger prediction error than the proposed model. The gap between the prediction errors from the two models is even bigger than for Model 1, because Model 2 is even more heterogeneous.

Table 3
Mean squared prediction errors for the random effects under Simulation Models 1 and 2, defined in (12) and (13), respectively. The fitted models are Gaussian (GLMM with Gaussian random effects) and Gaussian Mixture (the proposed model). Mean: Mean Squared Prediction Error averaged over 200 replicates; Std: standard deviation of the prediction error.

Simulation Model	Fitted Model	Mean	Std
Model 1	Gaussian	0.42	0.04
	Gaussian Mixture	0.36	0.04
Model 2	Gaussian	0.70	0.07
	Gaussian Mixture	0.54	0.06

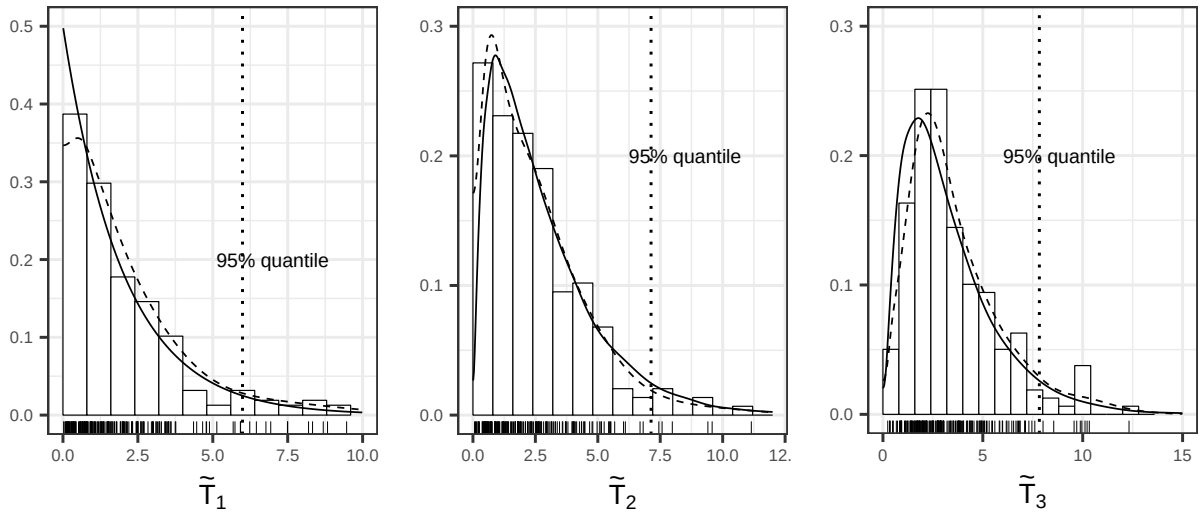
5.2. Simulation 2: Hypothesis tests

Next, we investigate the validity and power for the proposed tests in Section 3. We generate simulated data under similar settings as in Simulation 1, while γ_i's are generated from three models: Model 1, Model 2 and

$$\text{Model 0: } \mathcal{N}(-1.26, 0.5^2).$$

The three models represent latent Gaussian mixture models with orders 1 to 3. We generate 200 simulated data sets under each of the three models, and compute $\widetilde{T}_1$ in data under Model 0, $\widetilde{T}_2$ under Model 1 and $\widetilde{T}_3$ under Model 2. The empirical distributions of the three quantities represent the null distribution for the test statistics under the null hypotheses $C_0 = 1, 2$ and 3 respectively. These empirical distributions are provided in Fig. 2 and compared with the asymptotic distributions provided in Section 3. In each panel of Fig. 2, the dash curve is the kernel density based on

200 replicates of the test statistic and the solid curve is the asymptotic distribution. The asymptotic distributions for $\widetilde{T}_2$ and $\widetilde{T}_3$ are based on 10,000 simulations using the procedure described in Section S2 of the supplementary material. As we can see, the empirical distributions of the test statistics are remarkably close to the asymptotic distribution, which also shows the validity of the proposed tests. We use $\widetilde{T}_1$ - $\widetilde{T}_3$ to test the three null hypotheses, and the empirical sizes of these tests are 0.06, 0.03 and 0.05 respectively, which are close to the nominal level 0.05.



(a) Null distribution of $\widetilde{T}_1$ (b) Null distribution of $\widetilde{T}_2$ (c) Null distribution of $\widetilde{T}_3$

Fig. 2. Empirical (dash) and asymptotic (solid) distributions of the test statistics $\widetilde{T}_1$, $\widetilde{T}_2$ and $\widetilde{T}_3$ (defined in (5) and (7)) under the null hypotheses. The vertical dotted line marks the 95% quantile of the asymptotic distribution. The empirical distributions are obtained based on 200 simulations and the asymptotic distributions are described in Propositions 3 and 5.

Next, we illustrate the power of the tests. The response Y is generated the same way as in Section 5.1, while γ is generated from the following two models:

$$\text{Model 3: } 0.6\,\mathcal{N}(-2.26, 1.2^2) + 0.4\,\mathcal{N}(-0.46, 0.8^2), \tag{14}$$

$$\text{Model 4: } 0.3\,\mathcal{N}(-3.26, 1.2^2) + 0.4\,\mathcal{N}(-0.26, 0.8^2) + 0.3\,\mathcal{N}(2.34, 0.9^2). \tag{15}$$

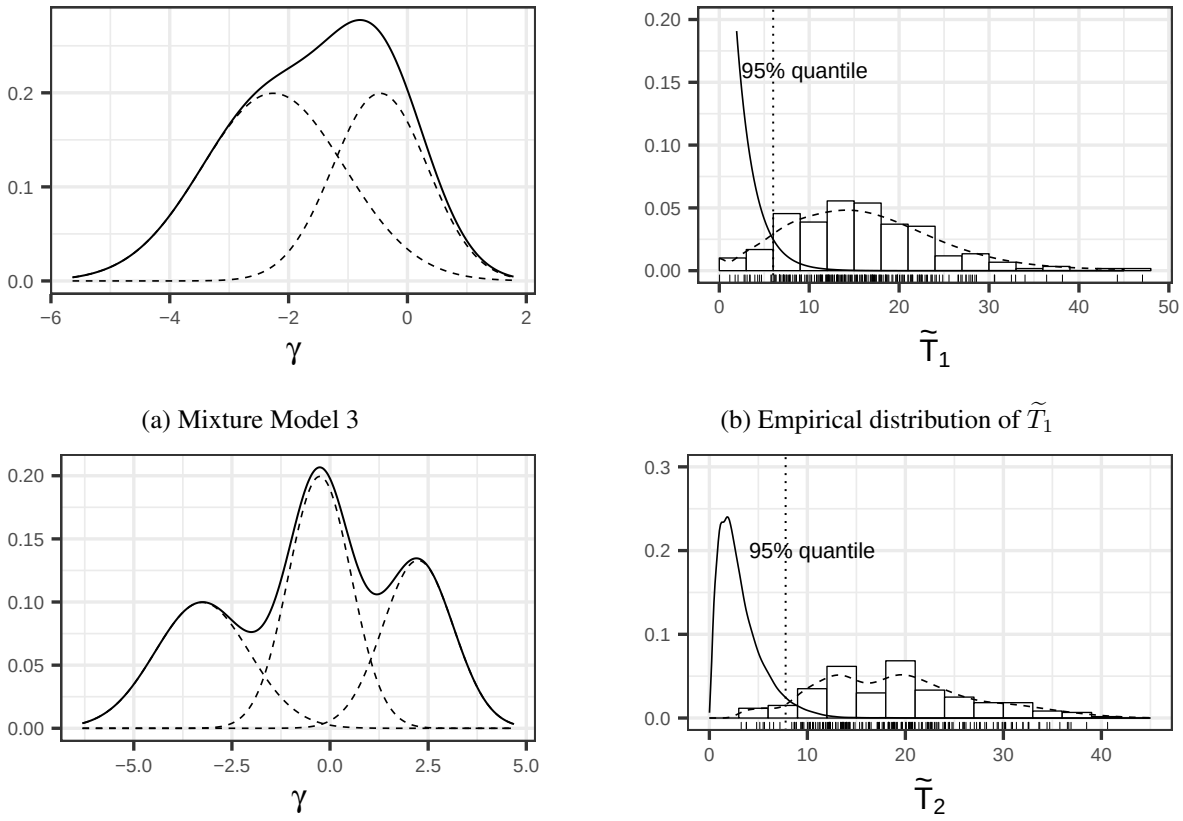
Compared with the Models 1 and 2 considered in Section 5.1, the individual components in Models 3 and 4 are less separated, making it harder to detect the real order of these models, especially when γ is an unobserved latent variable.

To examine the power of the proposed locally restricted likelihood ratio tests in Section 3, we test $H_0: C_0 = 1$ when the data are generated from Model 3, and test $H_0: C_0 = 2$ when the data are generated from Model 4. In Fig. 3, we present the empirical distributions of the test statistics based on 200 simulation runs. When performing 5% tests, the empirical powers of the proposed tests are 91% under Model 3 and 95.5% under Model 4. We have also examined the power of the homogeneity test when γ_i's are simulated from Model 1 and the power of the test on $H_0: C_0 = 2$ when γ_i's are generated from Model 2. The powers under both of these cases virtually equal to 1.

Since a sequential test can be used for model selection purposes, it is of interest to compare the test based procedure with other model selection procedures such as the Bayesian information criterion (BIC) or the Akaike information criterion (AIC), which are the negative log likelihood for the observed data plus a penalty on the number of free parameters in the model. Specifically, BIC for a C component latent Gaussian mixture model is

$$BIC(C) = -2\ell_n(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{X}) + (3C - 1)\ln(n),$$

where $\ell_n(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{X})$ is the marginal likelihood defined as (3) and n is the number of transplant centers. AIC is similarly defined replacing the $\ln(n)$ factor in the BIC by 2. In Table 4, we show the frequency of correctly choosing the number of mixture components for Models 3 and 4 using various model selection methods, including our sequential test procedure with significant levels 0.01, 0.05 and 0.1 and the AIC and BIC. The reported frequencies are based on 200 stimulation runs. As we can see, the sequential test procedure outperforms the two information criteria, especially the BIC, which by classic wisdom is a consistent model selection criterion. Among the three significant levels for the



(a) Mixture Model 3

(b) Empirical distribution of $\widetilde{T}_1$

(c) Mixture Model 4

(d) Empirical distribution of $\widetilde{T}_2$

Fig. 3. Power of the locally restricted likelihood ratio tests. Panels (a) and (c) illustrate the true density (solid) of γ under Model 3 and 4 respectively. The dashed lines represent the individual components. Panels (b) and (d) illustrate the empirical distributions (dash) of $\widetilde{T}_1$ and $\widetilde{T}_2$ comparing to the corresponding null distributions (solid). The vertical dotted line marks the 95% quantile of the null distribution.

sequential test procedure, $\alpha = 0.05$ provides the best results. As shown in Fig. 3, the mixture components in Models 3 and 4 overlap a lot and are hence hard to separate, which may explain the miserable failure of the BIC. The BIC puts a higher penalty on model complexity and therefore tends to choose a lower number of components than the truth. The AIC is more liberal and hence behaves more competitively in these examples, however we do see in other settings AIC overestimates the number of mixture components.

Table 4
Empirical frequency of choosing the correct number of mixture components for various model selection methods, including the sequential test procedure with different levels of α and the AIC and BIC. The data are generated from Simulation Models 3 and 4 (define in (14) and (15), respectively) and the results are based on 200 repetitions.

Simulation Model	α=0.01	α=0.05	α=0.10	BIC	AIC
Model 3	75.5%	88%	88%	39%	78.5%
Model 4	81%	86%	80.5%	50.5%	73%

6. Data Analysis

6.1. Background

Renal failure is one of the most common and severe diseases in the United States. In 2013, a total of 117,162 new cases were reported (www.USRDS.org). Kidney transplantation, a primary therapy for end stage renal disease, is a

complicated procedure typically involving transplant surgeons and physicians, coordinators, social workers, financial counselors, nutritionists, psychologists and referring physicians. The quality of care delivered by a transplant center is often assessed by patient survival, such as the 5 year post-transplant survival rate.

To provide a fair assessment of each transplant center, both patient level risk factors and an effect representing the quality of care of the transplant center are often included in the risk adjustment model. Many statisticians and health policy researchers model the transplant center effects as random effects that follow a Gaussian distribution [21, 22, 24]. This approach ignores the heterogeneity among the transplant centers, and the assumption of a common Gaussian distribution induces a shrinkage effect that makes the predicted random effects similar in value. He et al. [14] argue that borrowing information from other transplant centers is not fair when the goal of the study is to evaluate the centers and advocate modeling the transplant center effects as fixed effects. However, in such a fixed effects model, the number of parameters is large, making statistical inference numerically unstable, especially when the center size varies substantially. A comprehensive critique of these two approaches can be found in a report prepared by the Committee of Presidents of Statistical Societies (COPSS) through a contract with Centers for Medicare and Medicaid Services [2].

Our proposed latent Gaussian mixture model bridges the gap between the existing approaches and has two advantages. First, the model allows the presence of heterogeneities (e.g., the existence of clusters or subpopulations) among the transplant centers, making it a natural framework to identify centers with anomaly performance. Second, the mixture model can be considered as a compromise between the random effects model and the fixed effects model: it reduces to the random effects model when there is only one component in the mixture distribution and it becomes the fixed effects model if each transplant center forms a cluster of its own.

Our motivating data are obtained from the Organ Procurement and Transplantation Network, administered by the U.S. Department of Health and Human Services. The data system includes data on all donors, wait-listed candidates, and transplant recipients in the U.S. Included in the analysis are adult renal failure patients ($\geq$ 18 years of age) who underwent deceased donor kidney transplantation between January 1987 and December 2008. This cohort includes $N = 269,386$ patients receiving kidney transplants from a total of $n = 296$ centers. The number of transplants performed by a center, N_i, has a highly skewed distribution. Most centers performed a few hundred cases of kidney transplantation, but there are centers that took over 5000 cases. The patient level response is the 5-year survival status (1=death and -1=survival) and there is no censoring due to routine and rigorous tracking of the patients. The overall 5-year failure rate is 27.59%.

An important patient level covariate that is directly related to the success of kidney transplants is x_1 = cold ischemic time, which is the time that the donor kidney was kept in a refrigerator before being received by the patient. Other patient level covariates include x_2 = age at transplantation and x_3 = sex of the patient (1 = male, 0 = female), while x_4–x_6 are indicators for Body Mass Index (BMI) in the intervals (22, 25], (25, 30] and 30+ respectively. Since the data were collected over a time span of two decades, it is possible that the technology used in transplant surgeries has improved over time, which also affects the patient level outcome. Therefore, in addition to the other covariates described above, we also include time effects into the model. Using cases before 1990 as the baseline, covariates x_7–x_{10} are indicators for cases performed in 1990–1994, 1995–1999, 2000–2003 and 2004–2008, respectively.

6.2. Model fitting

We fit the proposed model to the data, using a random effect following a Gaussian mixture distribution to represent the care quality of a center. Using the proposed test procedure to decide the order of the latent Gaussian mixture model, the p-value is 0.0016 for $H_0 : C_0 = 1$ vs. $H_1 : C_0 = 2$; and 0.4076 for $H_0 : C_0 = 2$ vs. $H_1 : C_0 = 3$. We conclude that the care quality among the kidney transplant centers is not homogeneous and the distribution of the random effect is adequately described by a two-component Gaussian mixture. In this particular dataset, BIC happens to agree with the sequential test procedure and selects a two-component model as well. The estimated fixed effects under our final model are summarized in Table 5, where the standard errors are obtained using the asymptotic expansion (24) in the supplementary material. As we can see, all covariates considered are significant. Since we code $Y = 1$ as death, the results in Table 5 imply that having longer donor kidney delivery times, being older, being male, and having higher BMI all lead to increased risk of patient death. The coefficients for x_7–x_{10} are negative and decreasing, confirming that the overall death rate is decreasing over time.

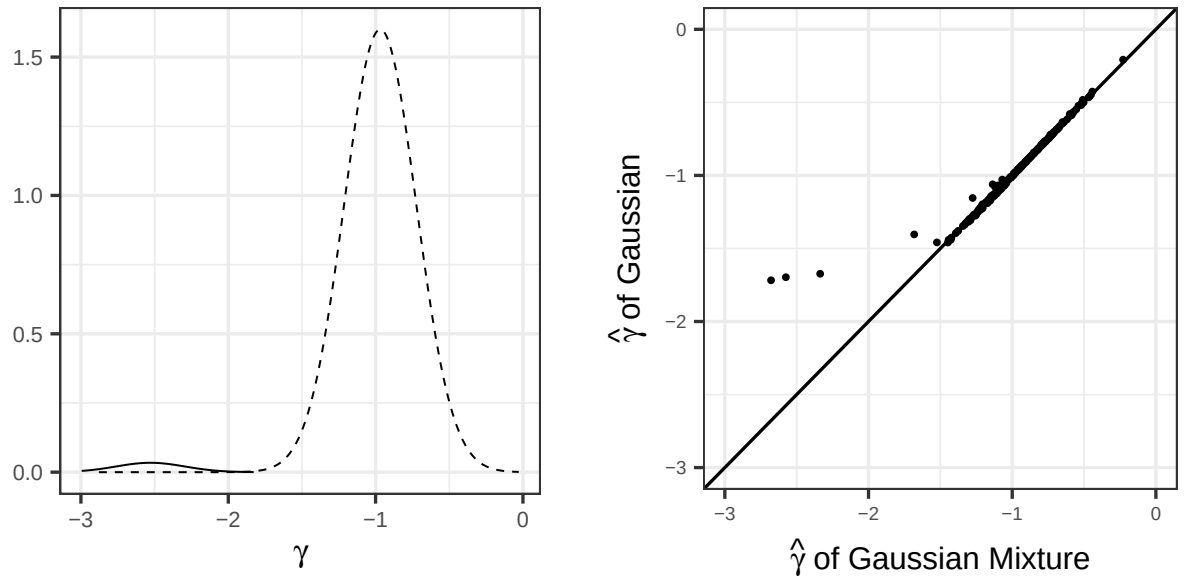
The estimated Gaussian mixture model for the random effect γ is

$$0.98\,\mathcal{N}(-0.969, 0.244^2) + 0.02\,\mathcal{N}(-2.528, 0.234^2).$$

Table 5
U.S. Organ Procurement and Transplantation Network data analysis: estimated fixed effects, standard errors, z-values and p-values. The covariates are x_1 = cold ischemic time, x_2 = age, x_3 = sex (1 = male, 0 = female); x_4–x_6 are indicators for BMI in the intervals (22, 25], (25, 30] and 30+ respectively; x_7–x_{10} are indicators for cases performed in 1990–1994, 1995–1999, 2000–2003 and 2004–2008, respectively. .

Covariate	Estimate	Std. Error	z-value	p-value
x_1	0.02	0.00	63.80	0.00
x_2	0.01	0.00	33.58	0.00
x_3	0.03	0.01	3.27	0.00
x_4	0.08	0.02	5.03	0.00
x_5	0.12	0.01	9.31	0.00
x_6	0.23	0.01	15.26	0.00
x_7	-0.27	0.01	-18.60	0.00
x_8	-0.53	0.01	-41.61	0.00
x_9	-0.63	0.01	-45.93	0.00
x_{10}	-0.80	0.01	-62.07	0.00

The mixture density $g(\gamma)$, as well as its individual components, are illustrated in Fig. 4 (a). The majority of the centers have rather similar care quality, but there is also a small cluster of transplant centers that have lower death rates after taking into account all the patient level covariates and these are the centers that are out-performing the others. In Fig. 4 (b), we also compare the predicted random effects under the standard GLMM with those under our latent Gaussian mixture model. While the predicted γ is almost the same under both models for the majority of the centers, the care quality effects for the a few centers in the left tail are severely shrunken towards the mean if we assume the random effects follow a homogeneous Gaussian distribution.



(a) (b)

Fig. 4. (a) Estimated latent Gaussian mixture model for the kidney transplant data. The solid line and dashed line represent two components. (b) Comparison of the predicted random effects under Gaussian and Gaussian mixture model assumptions.

Since the second component is small, we also run additional simulations to confirm that our methodology really works under such situations. To mimic the real data, we simulate binary Y_{ik} from a logistic GLMM using the covariates

from the real data, set $\boldsymbol{\beta}$ as the estimated values in Table 5 and generate $\boldsymbol{\gamma}$ from the following mixture model:

$$(1-\pi_2)\mathcal{N}(-0.969, 0.244^2) + \pi_2\mathcal{N}(-2.528, 0.234^2).$$

We set π_2 to be 0.005, 0.01, 0.02 or 0.05, and simulate 200 data sets under each setting. The empirical powers for testing $H_0 : C_0 = 1$ are 47%, 78.5%, 97.5% and 100% respectively. These results show that our method can detect a small component under the sample size of the real data and our discovery is likely to be true.

6.3. Performance evaluation

Based on the fitted model for γ in Fig. 4 (a), the majority of the centers provide similar care for their patients. However, the smaller mixture component consists of transplant centers with lower adjusted mortality rates, and these centers outperform the rest. We let the empirical null distribution be the bigger component of the fitted mixture model. Using the evaluation procedure described in Section 4 with the false discovery rate controlled at 5%, we find three transplant centers that outperform the rest. In Table 6, we list the IDs of the three outperforming centers, as well as their $lFDR$, $\widehat{\gamma}$, number of cases treated, and average 5-year survival rate.

Table 6
The out-performing centers detected using local false discovery rate in the kidney transplant data: ID of the transplant center, value of the lFDR defined in (11), predicted random effect $\hat{\gamma}$, number of patients treated in the transplant center and 5-year patient survival rate.

Center ID	lFDR	$\hat{\gamma}$	Num. Patient	Survival Rate
#287	0.00	-2.68	114	0.97
#10	0.01	-2.58	125	0.94
#28	0.07	-2.34	120	0.84

7. Summary

We propose a GLMM model with latent Gaussian mixture random effects that provides a natural framework to model the non-homogeneity among transplant centers and to rank their care quality. We demonstrate that the predicted random effects can be severely shrunken toward the mean if the distribution of the random effect is mis-specified as Gaussian. This shrinkage effect is quite prominent for the centers in the tails of the population. The latent Gaussian mixture model is not strongly identifiable and suffers from a slow convergence rate when the number of mixture components is larger than the truth. We develop test procedures to decide the number of mixture components. Even though the proposed tests are designed mainly for testing scientific claims and providing uncertainty assessments, they can also be used for model selection and our simulation results in Section 5.2 suggest that the sequential test procedure outperforms a naive Bayesian information criterion. We leave development of a consistent model selection procedure for the latent Gaussian mixture model for future work. The proposed test procedures are computationally intensive, especially when analyzing large medical data sets like the OPTN data, since we have to try hundreds of initial values to find the biggest likelihood ratio. These computations are best handled using parallel computing. Our open source software package `LatentGaussianMixtureModel` written in Julia will be made available on the corresponding author's website. Even though comparing transplant centers using the five-year survival rates of the patients has been the standard in the health policy literature, we acknowledge the fact that survival time is a more informative response variable. We intend to explore extending the latent Gaussian mixture model to survival outcomes in future research.

8. Technical Proofs

Assumptions

For simplicity, assume $N_i = n_0$ for $i \in \{1, \ldots, n\}$. Let $(\boldsymbol{X}, \boldsymbol{Y})$ be a generic copy of $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ and have a density

$$f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}) = f(\boldsymbol{x}) \int \left\{ \prod_{k=1}^{n_0} f(y_k \mid \boldsymbol{x}_k, \gamma; \boldsymbol{\beta}) g(\gamma \mid \boldsymbol{\theta}_\gamma) \right\} d\gamma, \tag{16}$$

where $\boldsymbol{y} = (y_1, \ldots, y_{n_0})^\top$, $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_0})^\top$ and $f(\boldsymbol{x})$ is the joint density of $\boldsymbol{X}$. Define metric

$$\delta(\boldsymbol{\theta}', \boldsymbol{\theta}) = \sum_l |\arctan \theta_l' - \arctan \theta_l|,$$

where θ_l is the l-th entry of $\boldsymbol{\theta}$. All convergences in the parameter space are defined with respect to the metric δ defined above.

Assumptions 1- 5 below are equivalent to those in Kiefer and Wolfowitz [20] and Hathaway [13] for the consistency result. Assumption 6 is a regularity assumption on the penalty function used in Chen et al. [9] and [19]. Assumption 7 and 8 are additional assumptions for Propositions 2 and 4 respectively.

Assumption 1. $f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta})$ is a density (the Radon-Nikodym derivative of a probability measure) with respect to a σ-finite measure μ on the space of $(\boldsymbol{x}, \boldsymbol{y})$.

Assumption 2 (Continuity Assumption). The definition of $f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta})$ can be extended to the closure of the parameter space $\bar{\Theta}_C$ such that, for any $\boldsymbol{\theta}^*$ in $\bar{\Theta}_C$ and any Cauchy sequence $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots\} \subset \bar{\Theta}_C$, $f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}_i) \to f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}^*)$ if $\boldsymbol{\theta}_i \to \boldsymbol{\theta}^*$.

Assumption 3. For any $\boldsymbol{\theta} \in \bar{\Theta}_C$ and any $\rho > 0$, $\omega(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}, \rho)$ is a measurable function of $(\boldsymbol{x}, \boldsymbol{y})$, where

$$\omega(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}, \rho) = \sup f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}'),$$

the supreme being taken over all $\boldsymbol{\theta}'$ in $\bar{\Theta}_C$ for which $\delta(\boldsymbol{\theta}', \boldsymbol{\theta}) < \rho$.

Assumption 4 (Identifiability Assumption). Identify $\bar{\Theta}_C$ as the quotient topological space such that

$$\mathcal{F}_{\boldsymbol{\theta}_0} = \left\{ \boldsymbol{\theta} \in \bar{\Theta}_C : \int_{-\infty}^{(\boldsymbol{x}', \boldsymbol{y}')} f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}) d\mu(\boldsymbol{x}, \boldsymbol{y}) = \int_{-\infty}^{(\boldsymbol{x}', \boldsymbol{y}')} f(\boldsymbol{x}, \boldsymbol{y}, \mid \boldsymbol{\theta}_0) d\mu(\boldsymbol{x}, \boldsymbol{y}) \text{ for any } (\boldsymbol{x}', \boldsymbol{y}') \right\}$$

is identified as a single point.

Assumption 5. For any $\boldsymbol{\theta}'$ in $\bar{\Theta}_C$,

$$\lim_{\rho \downarrow 0} \mathrm{E}_{\boldsymbol{\theta}} \left[\ln \frac{\omega(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}', \rho)}{f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta})} \right]^+ < \infty,$$

where $\mathrm{E}_{\boldsymbol{\theta}}$ is the expectation under $f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta})$ and $[x]^+$ equals x if $x > 0$ and 0 otherwise.

Assumption 6. The penalty function satisfies, (a) $\sup_{\sigma^2 > 0} \max\{0, p_n(\sigma^2)\} = o(n)$, $p_n(\sigma^2) = o(n)$ for any fixed σ^2; (b) for any $\sigma \in (0, 8/(nM)]$, $p_n(\sigma^2) \leq 5\{ln(n)\}^2 \ln(\sigma)$ for sufficient large n, where $M = \sup_{\boldsymbol{x}, \boldsymbol{y}} f(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}_0)$; (c) $p_n'(\sigma^2) = o_p(n^{1/4})$ for any fixed σ^2.

Assumption 7. When the true number of component is $C_0 = 1$, assume that $\boldsymbol{\mathcal{I}} = \mathrm{E}(\boldsymbol{I}_n)$ is a finite, positive definite matrix, where $\boldsymbol{I}_n$ is defined in (18).

Assumption 8. When $\boldsymbol{\theta} \in \Theta_C$, assume that $\boldsymbol{\mathcal{I}}^{(c)}$ defined in (23) is positive definite, for $c \in \{1, \ldots, C\}$.

Remark 3. The continuity assumption (Assumption 2) is not satisfied by the finite Gaussian mixture model on the boundary of the parameters space, since the likelihood diverges ∞ if any $\sigma_c^2 \to 0$. That is the reason that Hathaway [13] restricted the estimation in the interior of the parameter space. However, in our problem, the finite Gaussian mixture density $g(\gamma)$ is convoluted with proper density $f(\boldsymbol{y} \mid \boldsymbol{x}, \gamma)$ in (16). Since the integral is bounded, unbounded likelihood is no longer a concern and the condition is satisfied even on boundary points of $\bar{\Theta}_C$.

Remark 4. Assumption 4 is a modified version of the identifiability assumption in Kiefer and Wolfowitz [20]. The same assumption is used in Hathaway [13]. The consistency result in Proposition 1 means consistently estimating the mixture density rather than the parameters.

Proof of Proposition 1

Using similar arguments as in Chen et al. [9] one can show, as long as the penalty function satisfies Assumption 6, the maximizer of (2) is restricted in an interior region of the parameter space $\bar{\Theta}(\epsilon) = \{\boldsymbol{\theta} \in \bar{\Theta}; \min_c \sigma_c^2 \geq \epsilon\}$ for some positive constant ϵ. Since the penalty term is of order $o(n)$, which is much smaller than the likelihood function, the maximum penalized likelihood estimator $\widehat{\boldsymbol{\theta}}$ in the restricted parameter space belong to the class of modified maximum likelihood estimator in Kiefer and Wolfowitz [20] and the strong consistency of $\widehat{\boldsymbol{\theta}}$ follows from their theory.

Proof of Proposition 2

Denote for convenience $\zeta_i = \prod_{k=1}^{n_0} f(y_{ik} \mid \boldsymbol{x}_{ik}, \gamma; \boldsymbol{\theta}_y)$. After fixing $\pi_1 = \tau$, the log likelihood is

$$\ell_n(\boldsymbol{\theta}) \quad = \quad \sum_{i=1}^{n} \ln \int \zeta_i \{\tau f_1(\gamma \mid \mu_1, \sigma_1) + (1-\tau) f_2(\gamma \mid \mu_2, \sigma_2)\} d\gamma.$$

We adopt the re-parameterization of Kasahara and Shimotsu [19],

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \sigma_1^2 \\ \sigma_2^2 \end{pmatrix} = \begin{Bmatrix} \nu_\mu + (1-\tau)\lambda_\mu \\ \nu_\mu - \tau\lambda_\mu \\ \nu_\sigma + (1-\tau)(2\lambda_\sigma - \frac{1+\tau}{3}\lambda_\mu^2) \\ \nu_\sigma - \tau(2\lambda_\sigma + \frac{2-\tau}{3}\lambda_\mu^2) \end{Bmatrix}, \tag{17}$$

collect all parameters except τ into $\boldsymbol{\psi}(\tau) = (\boldsymbol{\eta}^\top, \boldsymbol{\lambda}^\top)^\top$, where $\boldsymbol{\eta} = (\boldsymbol{\theta}_y^\top, \nu_\mu, \nu_\sigma)^\top$ and $\boldsymbol{\lambda} = (\lambda_\mu, \lambda_\sigma)^\top$. Denote $\bar{\Theta}_\psi(\tau)$ as the parameter space of $\boldsymbol{\psi}$ corresponding to $\bar{\Theta}_2(\tau)$. Sometimes we suppress the dependence of $\boldsymbol{\psi}(\tau)$ on τ. Under the null hypothesis $C_0 = 1$, $\lambda_\mu = \lambda_\sigma = 0$ and the true parameter vector is $\boldsymbol{\psi}^* = \{(\boldsymbol{\eta}^*)^\top, 0, 0\}^\top$.

For any multivariate function $f(\boldsymbol{x})$, denote $\nabla_{\boldsymbol{x}^k} f$ as its k-th derivative, which is a multidimensional array. By similar calculations as in Proposition C and equation (29) in the supplementary appendix of Kasahara and Shimotsu [19], we can show

$$\begin{aligned}
&\nabla_{\lambda_\mu^k, \boldsymbol{\eta}^\ell} \ell_n(\boldsymbol{\psi}^*, \tau) = 0, \quad k \in \{1,2,3\}, \quad \ell \in \{0,1,\ldots\}; \\
&\nabla_{\lambda_\mu^k} \ell_n(\boldsymbol{\psi}^*, \tau) = O_p(n^{1/2}), \quad k \in \{4,5,6,7\}; \\
&\nabla_{\lambda_\sigma \boldsymbol{\eta}^\ell, \tau} \ell_n(\boldsymbol{\psi}^*) = 0, \quad \ell \in \{0,1,\ldots\}; \\
&\nabla_{\lambda_\sigma^k} \ell_n(\boldsymbol{\psi}^*, \tau) = O_p(n^{1/2}), \quad k \in \{2,3\}; \\
&\nabla_{\lambda_\mu \lambda_\sigma^2} \ell_n(\boldsymbol{\psi}^*, \tau) = O_p(n^{1/2}); \\
&\nabla_{\lambda_\mu^k \lambda_\sigma} \ell_n(\boldsymbol{\psi}^*, \tau) = O_p(n^{1/2}), \quad k \in \{1,\ldots,4\}.
\end{aligned}$$

Denote $g^*(\gamma) = g(\gamma; \boldsymbol{\psi}^*)$ as the true density of γ under the null hypothesis. Using a ninth order Taylor expansion of ℓ_{pen} around $\boldsymbol{\psi}^*$ as in Kasahara and Shimotsu [19], we get the following local quadratic approximation to the penalized likelihood

$$\begin{aligned}
\ell_{pen}(\boldsymbol{\psi}, \tau) - \ell_{pen}(\boldsymbol{\psi}^*, \tau) \quad &= \quad \boldsymbol{t}_n(\boldsymbol{\psi}, \tau)^\top \boldsymbol{S}_n - \frac{1}{2} \boldsymbol{t}_n(\boldsymbol{\psi}, \tau)^\top \boldsymbol{I}_n \boldsymbol{t}_n(\boldsymbol{\psi}, \tau) + R_n(\boldsymbol{\psi}, \tau) \\
&\quad + \sum_{c=1}^{2} [p_n\{\sigma_c^2(\boldsymbol{\psi}, \tau)\} - p_n\{\sigma_c^2(\boldsymbol{\psi}^*, \tau)\}],
\end{aligned} \tag{18}$$

where $\boldsymbol{t}_n(\boldsymbol{\psi}, \tau) = (\boldsymbol{t}_{\boldsymbol{\eta},n}, \boldsymbol{t}_{\boldsymbol{\lambda},n})^\top$, $\boldsymbol{S}_n = \sum_{i=1}^n \boldsymbol{s}_i / \sqrt{n}$, $\boldsymbol{I}_n = \frac{1}{n} \sum_{i=1}^n \boldsymbol{s}_i \boldsymbol{s}_i^\top$, $\boldsymbol{s}_i = (\boldsymbol{s}_{\boldsymbol{\eta},i}^\top, \boldsymbol{s}_{\boldsymbol{\lambda},i}^\top)^\top$, $\sigma_c^2(\boldsymbol{\psi}, \tau)$ is the variance as a function of $\boldsymbol{\psi}$ defined by the reparameterization in (17),

$$\begin{aligned}
&\boldsymbol{t}_{\boldsymbol{\eta},n} = \sqrt{n}(\boldsymbol{\eta} - \boldsymbol{\eta}*), \quad \boldsymbol{t}_{\boldsymbol{\lambda},n} = \begin{Bmatrix} 6\sqrt{n}\tau(1-\tau)\lambda_\mu\lambda_\sigma \\ \sqrt{n}\tau(1-\tau)(12\lambda_\sigma^2 - \frac{2}{3}(\tau^2 - \tau + 1)\lambda_\mu^4) \end{Bmatrix}, \\
&\boldsymbol{s}_{\boldsymbol{\eta},i} = \begin{pmatrix} \boldsymbol{s}_{\boldsymbol{\theta}_y,i} \\ s_{\nu_\mu,i} \\ s_{\nu_\sigma,i} \end{pmatrix} = \begin{Bmatrix} \frac{\int (\partial \zeta_i / \partial \boldsymbol{\theta}_y) g^*}{\int \zeta_i g^*} \\ \frac{\int \zeta_i g^* H_i^{1*}}{\int \zeta_i g^*} \\ \frac{\int \zeta_i g^* H_i^{2*}}{\int \zeta_i g^*} \end{Bmatrix}, \quad \boldsymbol{s}_{\boldsymbol{\lambda},i} = \begin{pmatrix} \frac{\int \zeta_i g^* H_i^{3*}}{\int \zeta_i g^*} \\ \frac{\int \zeta_i g^* H_i^{4*}}{\int \zeta_i g^*} \end{pmatrix}, \\
&R_n(\boldsymbol{\psi}, \tau) = [O(\|\boldsymbol{\psi} - \boldsymbol{\psi}^*\|) + o(1)] \times O_p[\{1 + \|\boldsymbol{t}_n(\boldsymbol{\psi}, \tau)\|^2\}].
\end{aligned}$$

Here,

$$H_i^{k*} = H^k(\frac{\gamma_i - \mu_\gamma^*}{\sigma_\gamma^*}) / \{k!(\sigma_\gamma^*)^k\},$$

where $H^k(x)$ is the kth order Hermite polynomial, e.g., $H^0(x) = 1$, $H^1(x) = x$, $H^2(x) = x^2 - 1$, $H^3(x) = x^3 - 3x$ and $H^4(x) = x^4 - 6x^2 + 3$.

By consistency of the estimator, we can focus on $\boldsymbol{\psi}$ such that $\|\boldsymbol{\psi} - \boldsymbol{\psi}^*\| = o_p(1)$ and hence $R_n(\boldsymbol{\psi}, \tau) = o_p(\|\boldsymbol{t}_n(\boldsymbol{\psi}, \tau)\|^2)$. By Assumption 6, $p_n'(\sigma^2) = o_p(n^{1/4})$, and by (17)

$$p_n\{\sigma_c^2(\boldsymbol{\psi}, \tau)\} - p_n\{\sigma_c^2(\boldsymbol{\psi}^*, \tau)\} = o_p(n^{1/4})(|\lambda_\sigma| + \lambda_\mu^2) = o_p\{\|\boldsymbol{t}_n(\boldsymbol{\psi}, \tau)\|\}.$$

Therefore, $\ell_{pen}(\boldsymbol{\psi}, \tau) - \ell_{pen}(\boldsymbol{\psi}^*, \tau)$ is dominated by the quadratic function defined by the first two terms on the right hand side of (18). It is then easy to see $\widehat{\boldsymbol{t}}_n = \boldsymbol{t}_n\{\widehat{\boldsymbol{\psi}}(\tau), \tau\}$ that maximizes $\ell_{pen}(\boldsymbol{\psi}, \tau) - \ell_{pen}(\boldsymbol{\psi}^*, \tau)$ is

$$\widehat{\boldsymbol{t}}_n = \boldsymbol{I}_n^{-1}\boldsymbol{S}_n + o_p(1). \tag{19}$$

Under Assumption 7, $\boldsymbol{\mathcal{I}} = \mathrm{E}(\boldsymbol{I}_n)$ is a positive definite matrix. By the law of large numbers, $\boldsymbol{I}_n \to \boldsymbol{\mathcal{I}}$ in probability. On the other hand, by the central limit theorem, $\boldsymbol{S}_n \to N(0, \boldsymbol{\mathcal{I}})$ in distribution. Therefore, $\widehat{\boldsymbol{t}}_n \to N(0, \boldsymbol{\mathcal{I}}^{-1})$ in distribution, which also implies

$$\widehat{\boldsymbol{\beta}}_{full}(\tau) - \boldsymbol{\beta}_0 = O_p(n^{-1/2}), \quad \widehat{\lambda}_\mu = O_p(n^{-1/8}), \quad \widehat{\lambda}_\sigma = O_p(n^{-1/4}).$$

The convergence rate of $\widehat{\boldsymbol{\theta}}_{\gamma, full}(\tau)$ is determined by those of $\widehat{\lambda}_\mu$ and $\widehat{\lambda}_\sigma$.

Proof of Proposition 3

Following arguments in Section 8, we have $\boldsymbol{S}_n \to N(0, \boldsymbol{\mathcal{I}})$ in distribution, where $\boldsymbol{\mathcal{I}} = \mathrm{E}(\boldsymbol{I}_n)$. Under the full model, for any $\boldsymbol{\psi}$ such that $\boldsymbol{t}_n = O_p(1)$, using the local quadratic approximation (18) we have

$$2\{\ell_n(\boldsymbol{\psi}, \tau) - \ell_n(\boldsymbol{\psi}^*, \tau)\} = 2\boldsymbol{t}_n^\top \boldsymbol{S}_n - \boldsymbol{t}_n^\top \boldsymbol{I}_n \boldsymbol{t}_n + o_p(1) = 2\boldsymbol{t}_n^\top \boldsymbol{S}_n - \boldsymbol{t}_n^\top \boldsymbol{\mathcal{I}} \boldsymbol{t}_n + o_p(1).$$

Let $\widehat{\boldsymbol{\psi}}_{full}(\tau)$ be maximizer of (18) under the full model with 2 components, and it is the reparameterized version of $\widehat{\boldsymbol{\theta}}_{full}(\tau)$. By (19), $\boldsymbol{t}_n\{\widehat{\boldsymbol{\psi}}_{full}(\tau)\} = \boldsymbol{\mathcal{I}}^{-1}\boldsymbol{S}_n + o_p(1)$ and hence

$$2[\ell_n\{\widehat{\boldsymbol{\psi}}_{full}(\tau), \tau\} - \ell_n(\boldsymbol{\psi}^*, \tau)] = \boldsymbol{S}_n^\top \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{S}_n + o_p(1). \tag{20}$$

Partition $\boldsymbol{S}_n$ into $\begin{pmatrix} \boldsymbol{S}_{\eta,n} \\ \boldsymbol{S}_{\lambda,n} \end{pmatrix}$ according to the partition of $\boldsymbol{\psi}$. With a similar partition to $\boldsymbol{\mathcal{I}}$, we have

$$\boldsymbol{\mathcal{I}}^{-1} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_\eta & \boldsymbol{\mathcal{I}}_{\eta\lambda} \\ \boldsymbol{\mathcal{I}}_{\lambda\eta} & \boldsymbol{\mathcal{I}}_\lambda \end{pmatrix}^{-1} = \begin{Bmatrix} \boldsymbol{\mathcal{I}}_\eta^{-1} + \boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{\mathcal{I}}_{\eta\lambda}\boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1}\boldsymbol{\mathcal{I}}_{\lambda\eta}\boldsymbol{\mathcal{I}}_\eta^{-1} & -\boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{\mathcal{I}}_{\eta\lambda}\boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1} \\ (-\boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{\mathcal{I}}_{\eta\lambda}\boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1})^\top & \boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1} \end{Bmatrix},$$

where $\boldsymbol{\mathcal{I}}_{\lambda|\eta} = \boldsymbol{\mathcal{I}}_\lambda - \boldsymbol{\mathcal{I}}_{\lambda\eta}\boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{\mathcal{I}}_{\eta\lambda}$. Define

$$\boldsymbol{S}_{\lambda|\eta,n} = \boldsymbol{S}_{\lambda,n} - \boldsymbol{\mathcal{I}}_{\lambda\eta}\boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{S}_{\eta,n},$$

and by simple algebra

$$\boldsymbol{S}_n^\top \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{S}_n = \boldsymbol{S}_{\eta,n}^\top \boldsymbol{\mathcal{I}}_\eta^{-1} \boldsymbol{S}_{\eta,n} + \boldsymbol{S}_{\lambda|\eta,n}^\top \boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1} \boldsymbol{S}_{\lambda|\eta,n}. \tag{21}$$

Under the reduced model, $\boldsymbol{\lambda} = 0$, and hence $\boldsymbol{t}_{\lambda n} = \boldsymbol{S}_{\lambda n} = 0$. Using the same local quadratic approximation, for a parameter vector $\boldsymbol{\psi}_{red}$ in the reduced model,

$$2\{\ell_n(\boldsymbol{\psi}_{red}, \tau) - \ell_n(\boldsymbol{\psi}^*, \tau)\} = 2\boldsymbol{t}_{\eta n}^\top \boldsymbol{S}_{\eta n} - \boldsymbol{t}_{\eta n}^\top \boldsymbol{I}_\eta \boldsymbol{t}_{\eta n} + o_p(1).$$

Let $\widehat{\boldsymbol{\psi}}_{red}$ be the estimator that maximizes the reduced model penalized likelihood, then $\boldsymbol{t}_{\eta n}(\widehat{\boldsymbol{\psi}}_{red}) = \boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{S}_{\eta n} + o_p(1)$, and

$$2\{\ell_n(\widehat{\boldsymbol{\psi}}_{red}, \tau) - \ell_n(\boldsymbol{\psi}^*, \tau)\} = \boldsymbol{S}_{\eta,n}^\top \boldsymbol{\mathcal{I}}_\eta^{-1} \boldsymbol{S}_{\eta,n} + o_p(1). \tag{22}$$

Combining (20), (21) and (22),

$$T_1(\tau) = 2[\ell_n\{\widehat{\boldsymbol{\psi}}_{full}(\tau), \tau\} - \ell_n(\widehat{\boldsymbol{\psi}}_{red}, \tau)] = \boldsymbol{S}_{\lambda|\eta,n}^\top \boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1} \boldsymbol{S}_{\lambda|\eta,n} + o_p(1) \to \chi^2(2) \text{ in distribution.}$$

Because $\boldsymbol{S}_{\lambda|\eta,n}$ and $\boldsymbol{\mathcal{I}}_{\lambda|\eta}$ do not depend on τ and $\boldsymbol{\mathcal{T}}$ is a finite set,

$$\widetilde{T}_1 = \max_{\tau \in \boldsymbol{\mathcal{T}}} T_1(\tau) = \boldsymbol{S}_{\lambda|\eta,n}^\top \boldsymbol{\mathcal{I}}_{\lambda|\eta}^{-1} \boldsymbol{S}_{\lambda|\eta,n} + o_p(1) \to \chi^2(2) \text{ in distribution.}$$

Proof of Proposition 4

Denote $\zeta_i = \prod_{k=1}^{n_0} f(y_{ik} \mid \boldsymbol{x}_{ik}, \gamma; \boldsymbol{\theta}_y)$ as in Section 8. Under the local reparameterization in $\mathcal{N}_{C+1}(c,\tau)$ defined in (3.9) and (3.10) in Section 3.2, the log likelihood is

$$\ell_n(\boldsymbol{\theta}) \quad = \quad \sum_{i=1}^{n} \ln \int \zeta_i g_{c,\tau}(\gamma) d\gamma,$$

where

$$\begin{aligned} g_{c,\tau}(\gamma) \quad &= \quad (\pi_c + \pi_{c+1})\tau f(\gamma \mid \mu_c, \sigma_c) + (\pi_c + \pi_{c+1})(1-\tau) f(\gamma \mid \mu_{c+1}, \sigma_{c+1}) + \sum_{c' \neq c} \pi_{c'} f_{c'}(\gamma \mid \mu_{c'}, \sigma_{c'}) \\ &= \quad (\pi_c + \pi_{c+1})\tau f\left\{\gamma \mid \nu_\mu + (1-\tau)\lambda_\mu, \nu_\sigma + (1-\tau)(2\lambda_\sigma - \frac{1+\tau}{3}\lambda_\mu^2)\right\} \\ &\quad +(\pi_c + \pi_{c+1})(1-\tau) f\left\{\gamma \mid \nu_\mu - \tau\lambda_\mu, \nu_\sigma - \tau(2\lambda_\sigma + \frac{2-\tau}{3}\lambda_\mu^2)\right\} + \sum_{c' \neq c} \pi_{c'} f_{c'}(\gamma \mid \mu_{c'}, \sigma_{c'}). \end{aligned}$$

The score function with respect to $\boldsymbol{\psi}(c,\tau)$ is $\boldsymbol{s}_i^{(c)} = (\boldsymbol{s}_{\boldsymbol{\eta},i}^\top, (\boldsymbol{s}_{\boldsymbol{\lambda},i}^{(c)})^\top)^\top$, which is defined in (3.11). Define $\boldsymbol{S}_n^{(c)} = n^{-1/2}\sum_{i=1}^n \boldsymbol{s}_i^{(c)}$, $\boldsymbol{I}_n^{(c)} = n^{-1}\sum_{i=1}^n \boldsymbol{s}_i^{(c)}(\boldsymbol{s}_i^{(c)})^\top$ and $\boldsymbol{t}_n\{\boldsymbol{\psi}(c,\tau),\tau\} = (\boldsymbol{t}_{\boldsymbol{\eta},n}, \boldsymbol{t}_{\boldsymbol{\lambda},n})^\top$ where

$$\boldsymbol{t}_{\boldsymbol{\eta},n} = \sqrt{n}(\boldsymbol{\eta} - \boldsymbol{\eta}*), \quad \boldsymbol{t}_{\boldsymbol{\lambda},n} = \left\{ \begin{array}{c} 6\sqrt{n}\tau(1-\tau)\lambda_\mu\lambda_\sigma \\ \sqrt{n}\tau(1-\tau)(12\lambda_\sigma^2 - \frac{2}{3}(\tau^2 - \tau + 1)\lambda_\mu^4) \end{array} \right\}.$$

Similar to (18), we can derive a local quadratic approximation to the likelihood

$$\ell_n\{\boldsymbol{\psi}(c,\tau),\tau\} - \ell_n(\boldsymbol{\psi}^*) \quad = \quad \boldsymbol{t}_n\{\boldsymbol{\psi}(c,\tau),\tau\}^\top \boldsymbol{S}_n^{(c)} - \frac{1}{2}\boldsymbol{t}_n\{\boldsymbol{\psi}(c,\tau),\tau\}^\top \boldsymbol{I}_n^{(c)} \boldsymbol{t}_n\{\boldsymbol{\psi}(c,\tau),\tau\} + R_{n,c}\{\boldsymbol{\psi}(c,\tau),\tau\},$$

where $R_n(\boldsymbol{\psi},\tau) = [O(\|\boldsymbol{\psi} - \boldsymbol{\psi}^*\|) + o(1)] \times O_p[\{1 + \|\boldsymbol{t}_n(\boldsymbol{\psi},\tau)\|^2\}]$.

Put $\widehat{\boldsymbol{\psi}}_{full}(c,\tau) = \arg\max_{\boldsymbol{\psi}(c,\tau) \in \Theta_\psi(c,\tau)} \ell_{pen}\{\boldsymbol{\psi}(c,\tau),\tau\}$ and $\widehat{\boldsymbol{t}}_n = \boldsymbol{t}_n\left\{\widehat{\boldsymbol{\psi}}_{full}(c,\tau),\tau\right\}$. Using similar arguments as in Section 8, we can show that the penalty function is asymptotically negligible when $\boldsymbol{\psi}(c,\tau)$ is in a consistent neighborhood of $\boldsymbol{\psi}^*$. Define

$$\boldsymbol{\mathcal{I}}^{(c)} = \mathrm{E}(\boldsymbol{I}_n^{(c)}) = \mathrm{var}(\boldsymbol{s}_i^{(c)}), \tag{23}$$

which is positive definite under Assumption 8. It is then easy to see that

$$\widehat{\boldsymbol{t}}_n = (\boldsymbol{\mathcal{I}}^{(c)})^{-1}\boldsymbol{S}_n^{(c)} + o_p(1) \to \mathcal{N}\{0, (\boldsymbol{\mathcal{I}}^{(c)})^{-1}\} \text{ in distribution.} \tag{24}$$

By the definition of $\boldsymbol{t}_n\{\boldsymbol{\psi}(c,\tau),\tau\}$, we obtain $\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}^* = O_p(n^{-1/2})$, $\widehat{\lambda}_\mu = O_p(n^{-1/8})$ and $\widehat{\lambda}_\sigma = O_p(n^{-1/4})$. Clearly $\widehat{\mu}_{c,full}(c,\tau)$ and $\widehat{\mu}_{c+1,full}(c,\tau)$ converge to the true parameter $\mu_{c,0}$ at a $O_p(n^{-1/8})$ rate. Since the convergence rates for $\widehat{\sigma}^2_{c,full}(c,\tau)$ and $\widehat{\sigma}^2_{c+1,full}(c,\tau)$ are determined by $\widehat{\lambda}_\mu^2$ and $\widehat{\lambda}_\sigma$, they converge to the true parameter $\sigma^2_{c,0}$ in $O_p(n^{-1/4})$ rate. The rest of the parameters in $\widehat{\boldsymbol{\theta}}_{full}(c,\tau)$ converge in a $O_p(n^{-1/2})$ rate.

Proof of Proposition 5

We first derive the asymptotic properties for $T_C(c,\tau)$. By (23) and (24),

$$2[\ell_n\{\widehat{\boldsymbol{\psi}}_{full}(c,\tau),\tau\} - \ell_n(\boldsymbol{\psi}^*)] = (\boldsymbol{S}_n^{(c)})^\top (\boldsymbol{\mathcal{I}}^{(c)})^{-1}\boldsymbol{S}_n^{(c)} + o_p(1),$$

where $\boldsymbol{S}_n^{(c)} \to \mathcal{N}(0, \boldsymbol{\mathcal{I}}^{(c)})$ in distribution by the central limit theorem.

The reduced model estimator $\widehat{\boldsymbol{\psi}}_{red}(c,\tau)$ is obtained by minimizing the penalized likelihood while restricting $\lambda_\mu = \lambda_\sigma = 0$. by similar derivations under the full model, we get

$$2[\ell_n\{\widehat{\boldsymbol{\psi}}_{red}(c,\tau),\tau\} - \ell_n(\boldsymbol{\psi}^*)] = \boldsymbol{S}_{\eta,n}^\top \boldsymbol{\mathcal{I}}_\eta^{-1}\boldsymbol{S}_{\eta,n} + o_p(1),$$

where $\boldsymbol{S}_{\eta,n}$ and $\boldsymbol{\mathcal{I}}_{\eta}$ are sub-vector or sub-matrix of $\boldsymbol{S}_n^{(c)}$ and $\boldsymbol{\mathcal{I}}^{(c)}$ as defined in Proposition 5.

Using algebra similar to that in Section 8, we get

$$T_C(c,\tau) \;=\; 2[\ell_n\{\widehat{\boldsymbol{\psi}}_{full}(c,\tau),\tau\} - \ell_n\{\widehat{\boldsymbol{\psi}}_{red}(c,\tau),\tau\}] = (\boldsymbol{S}^{(c)}_{\lambda|\eta,n})^\top(\boldsymbol{\mathcal{I}}^{(c)}_{\lambda|\eta})^{-1}\boldsymbol{S}^{(c)}_{\lambda|\eta,n} + o_p(1) \to \chi^2(2) \text{ in distribution.}$$

Therefore,

$$T_C(\tau) = \max_c T_C(c,\tau) \to \max_{c\in\{1,\ldots,C\}}\{(\boldsymbol{S}^{(c)}_{\lambda|\eta,n})^\top(\boldsymbol{\mathcal{I}}^{(c)}_{\lambda|\eta})^{-1}\boldsymbol{S}^{(c)}_{\lambda|\eta,n}\} \text{ in distribution.}$$

Since none of the quantities $(\boldsymbol{S}^{(c)}_{\lambda|\eta,n})^\top(\boldsymbol{\mathcal{I}}^{(c)}_{\lambda|\eta})^{-1}\boldsymbol{S}^{(c)}_{\lambda|\eta,n}$ depends on τ, $\widetilde{T}_C$ that maximizes $T_C(\tau)$ over any set $\mathcal{T}$ has the same limiting distribution.

Proof of Proposition 6

The FDR for the described procedure is

$$\begin{aligned} FDR &= \mathrm{E}\left\{\frac{\sum_i^n I(\delta_i=1,\sum_{c\in\mathcal{C}_0}L_{ic}=1)}{\sum_i^n I(\delta_i=1)} \mid \sum_i^n I(\delta_i=1)>0\right\}\Pr\left\{\sum_i^n I(\delta_i=1)>0\right\} \\ &= \mathrm{E}\left\{\frac{\sum_i^n \delta_i\left(\sum_{c\in\mathcal{C}_0}L_{ic}\right)}{\sum_i^n \delta_i\vee 1}\right\} = \mathrm{E}\left\{\frac{\sum_i^n \delta_i \mathrm{E}\left(\sum_{c\in\mathcal{C}_0}L_{ic}=1 \mid \boldsymbol{X}_i,\boldsymbol{Y}_i\right)}{\sum_i^n \delta_i\vee 1}\right\} \\ &= \mathrm{E}\left(\frac{\sum_i^n \delta_i lFDR_i}{\sum_i^n \delta_i\vee 1}\right) = \mathrm{E}\left(\frac{\sum_i^k lFDR_{(i)}}{k}\right) \le \alpha. \end{aligned}$$

Supplementary Material

The online supplementary material contains details of the model fitting algorithm.

Acknowledgement

The authors thank the Editor, the Associate Editor and two anonymous referees for their many helpful comments and constructive suggestions, which lead to significant improvement in the quality of the paper. This research was supported in part by National Institutes of Health grant 5R21AG058198.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: Selected Papers of Hirotugu Akaike, Springer, 1998, pp. 199–213.

[2] A. S. Ash, S. E. Fienberg, T. A. Louis, S.-L. T. Normand, T. A. Stukel, J. Utts, Statistical issues in assessing hospital performance, Report, Committee of Presidents of Statistical Societies, 2012.

[3] J. G. Booth, J. P. Hobert, Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, JJ. R. Stat. Soc. Ser. B. Stat. Methodol. 61 (1999) 265–285.

[4] N. E. Breslow, D. G. Clayton, Approximate inference in generalized linear mixed models, J. Amer. Statist. Assoc. 88 (1993) 9–25.

[5] B. Caffo, M.-W. An, C. Rohde, Flexible random intercept models for binary outcomes using mixtures of normals, Comput. Statist. Data Anal. 51 (2007) 5220–5235.

[6] J. Chen, Optimal rate of convergence for finite mixture models, Ann. Statist. 23 (1995) 221–233.

[7] J. Chen, P. Li, Hypothesis test for normal mixture models: The EM approach, Ann. Statist. 37 (2009) 2523–2542.

[8] J. Chen, P. Li, Y. Fu, Inference on the order of a normal mixture, J. Amer. Statist. Assoc. 107 (2012) 1096–1105.

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

[10] J. Chen, D. Zhang, M. Davidian, A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution, Biostatistics 3 (2002) 347–360.

[11] B. Efron, Large-scale simultaneous hypothesis testing, J. Amer. Statist. Assoc. 99 (2004) 96–104.

[12] J. J. Goeman, A. Solari, The sequential rejection principle of familywise error control, Ann. Statist. 38 (2010) 3782–3810.

[13] R. J. Hathaway, A constrained formulation of maximum likelihood estimation for normal mixture distributions, Ann. Statist. 13 (1985) 795–800.

[14] K. He, J. D. Kalbfleisch, Y. Li, Y. Li, Evaluating hospital readmission rates in dialysis facilities; adjusting for hospital effects, Lifetime Data Anal. 19 (2013) 490–512.

[15] N. Ho, X. Nguyen, Convergence rates of parameter estimation for some weakly identifiable finite mixtures, Ann. Statist. 44 (2016) 2726–2755.

[16] H. Huang, Y. Li, Y. Guan, Joint modeling and clustering paired generalized longitudinal trajectories with application to cocaine abuse treatment data, J. Amer. Statist. Assoc. 109 (2014) 1412–1424.

[17] H. Ishwaran, L. F. James, J. Sun, Bayesian model selection in finite mixtures by marginal density decompositions, J. Amer. Statist. Assoc. 96 (2001) 1316–1332.

[18] H. J.X., H. Zhao, H. Zhou, False discovery rate control with groups, Journal of the American Statistical Association 105 (2010) 1215–1227.

[19] H. Kasahara, K. Shimotsu, Testing the number of components in normal mixture regression models, J. Amer. Statist. Assoc. 110 (2015) 1632–1645.
[20] J. Kiefer, J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, Ann. Math. Statist. 27 (1956) 887–906.
[21] H. Krumholz, S.-L. T. Normand, D. Galusha, J. Mattera, A. Rich, Y. Wang, M. Ward, Risk-adjustment models for AMI and HF: 30-Day mortality, Report, Centers for Medicare and Medicaid Services, 2006.
[22] H. M. Krumholz, Y. Wang, J. A. Mattera, Y. Wang, L. F. Han, M. J. Ingber, S. Roman, S.-L. T. Normand, An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction, Circulation 113 (2006) 1683–1692.
[23] S. Li, J. Chen, J. Guo, B.-Y. Jing, S.-Y. Tsang, H. Xue, Likelihood ratio test for multi-sample mixture model and its application to genetic imprinting, J. Amer. Statist. Assoc. 110 (2015) 867–877.
[24] Y. Li, X. Cai, L. G. Glance, W. D. Spector, D. B. Mukamel, National release of the nursing home quality report cards: implications of statistical methodology for risk adjustment, Health Serv. Res. 44 (2009) 79–102.
[25] F. Liang, J. Zhang, Estimating the false discovery rate using the stochastic approximation algorithm, Biometrika 95 (2008) 961–977.
[26] X. Lin, N. E. Breslow, Bias correction in generalized linear mixed models with multiple components of dispersion, J. Amer. Statist. Assoc. 91 (1996) 1007–1016.
[27] S. Litière, A. Alonso, G. Molenberghs, Type I and type II error under random-effects misspecification in generalized linear mixed models, Biometrics 63 (2007) 1038–1044.
[28] M. J. Lombardía, E. López-Vizcaíno, C. Rueda, Mixed generalized akaike information criterion for small area models, JJ. R. Stat. Soc. Ser. A. Stat. Soc.) 180 (2017) 1229–1252.
[29] C. E. McCulloch, J. M. Neuhaus, Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter, Statist. Sci. 26 (2011) 388–402.
[30] G. McLachlan, D. Peel, Finite mixture models, John Wiley & Sons, New York, 2004.
[31] G. Schwarz, et al., Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[32] W. Sun, T. T. Cai, Oracle and adaptive compound decision rules for false discovery rate control, J. Amer. Statist. Assoc. 102 (2007) 901–912.
[33] W. Sun, B. J. Reich, T. T. Cai, M. Guindani, A. Schwartzman, False discovery control in large-scale spatial multiple testing, JJ. R. Stat. Soc. Ser. B. Stat. Methodol. 77 (2015) 59–83.
[34] M.-J. Woo, T. N. Sriram, Robust estimation of mixture complexity, J. Amer. Statist. Assoc. 101 (2006) 1475–1486.
[35] C. You, S. Müller, J. T. Ormerod, On generalized degrees of freedom with application in linear mixed models selection, Statist. Comput. 26 (2016) 199–210.