Testing the Number of Components in Normal Mixture Regression Models

Hiroyuki Kasahara[*]
Vancouver School of Economics
University of British Columbia
hkasahar@mail.ubc.ca

Katsumi Shimotsu
Faculty of Economics
University of Tokyo
shimotsu@e.u-tokyo.ac.jp

October 30, 2014

Abstract

Testing the number of components in finite normal mixture models is a long-standing challenge because of its non-regularity. This paper studies likelihood-based testing of the number of components in normal mixture regression models with heteroscedastic components. We construct a likelihood-based test of the null hypothesis of m_0 components against the alternative hypothesis of $m_0 + 1$ components for any m_0. The null asymptotic distribution of the proposed modified EM test statistic is the maximum of m_0 random variables that can be easily simulated. The simulations show that the proposed test has very good finite sample size and power properties.

Key words: asymptotic distribution; modified EM test; likelihood ratio test; local MLE; normal mixture models; number of components

[*] Address for correspondence: Hiroyuki Kasahara, Vancouver School of Economics, University of British Columbia, 997-1873 East Mall, Vancouver, BC V6T 1Z1, Canada. The authors are grateful to the editor, the associated editor, and two anonymous referees whose comments greatly improved the paper. This research is support by the Natural Science and Engineering Research Council of Canada and JSPS Grant-in-Aid for Scientific Research (C) No. 26380267.

1 Introduction

Finite mixtures of normal distributions and regressions have been used in numerous empirical applications in diverse fields such as biological, physical, and social sciences, including economics and finance (see, e.g., Kon, 1984; Tucker, 1992; Venkataraman, 1997; Quandt and Ramsey, 1978; Kon and Jen, 1978; Conway and Deb, 2005). Mixture-of-expert models with normal component distribution (see, e.g., Jacobs et al., 1991) can also be viewed as finite mixture of normal regression models. Comprehensive theoretical accounts and examples of applications have been provided by several authors, including Lindsay (1995), Titterington et al. (1985), and McLachlan and Peel (2000).

The number of components is an important parameter in applications of finite mixture models. In economics, the number of components often represents the number of unobservable types or abilities. In other applications, the number of components signifies the number of clusters or latent classes in the data. Despite its importance, testing for the number of components in normal mixture regression models has been a long-standing unsolved problem because the standard asymptotic analysis of the likelihood ratio test (LRT) statistic breaks down due to problems such as non-identifiable parameters and the true parameter being on the boundary of the parameter space. Numerous papers have been written on the subject of the likelihood ratio test for the number of components (see, e.g., Ghosh and Sen, 1985; Chernoff and Lander, 1995; Lemdani and Pons, 1997; Chen and Chen, 2001, 2003; Chen et al., 2004; Garel, 2001, 2005), and the asymptotic distribution of the LRT statistic for general finite mixture models has been derived as a functional of the Gaussian process (Dacunha-Castelle and Gassiat, 1999; Azaïs et al., 2009; Liu and Shao, 2003; Zhu and Zhang, 2004).

In normal mixtures with heteroscedastic components, however, the asymptotic distribution of the LRT statistic remains an open question because, as discussed in Chen et al. (2012), normal mixtures have an additional undesirable mathematical property that invalidates key assumptions in these works. In particular, the normal density with mean μ and variance σ^2, $f(y;\mu,\sigma^2)$, has the property $\frac{\partial^2}{\partial\mu\partial\mu}f(y;\mu,\sigma^2) = 2\frac{\partial}{\partial\sigma^2}f(y;\mu,\sigma^2)$. This leads to the loss of "strong identifiability" condition introduced by Chen (1995). As a result, neither Assumption (P1) of Dacunha-Castelle and Gassiat (1999) nor Assumption 7 of Azaïs et al. (2009) holds, and Assumption 3 of Zhu and Zhang (2004) is violated, while Corollary 4.1 of Liu and Shao (2003) does not hold in heteroscedastic normal mixtures.

This paper develops a likelihood-based testing procedure of the null hypothesis of m_0 components against the alternative hypothesis of $m_0 + 1$ components for a general $m_0 \geq 1$ in heteroscedastic normal mixture regression models. To this end, we introduce a new reparameterization that substantially simplifies the analysis. Under this reparameterization, the log-likelihood function is locally approximated by a quadratic function of polynomials of parameters, and a standard analysis goes through with some adjustment using the results of Andrews (1999) and Zhu and Zhang (2006), who generalize Andrews (1999). We propose a modified EM test by building on this local quadratic representation and extending the EM approach pioneered by Li et al. (2009) and Li and Chen (2010). The asymptotic null distribution of the proposed modified EM test statistic is the maximum of m_0 random variables, which can be easily simulated. In particular, when no regressor is present, the asymptotic null distribution is the maximum of m_0 chi-squared random variables with two degrees of freedom. Furthermore, the modified EM test does not suffer from the infinite Fisher information problem.

To the best of our knowledge, no likelihood-based test has yet been developed for testing the null hypothesis $H_0 : m = m_0$ with $m_0 \geq 1$ against the alternative hypothesis $H_A : m = m_0 + 1$ in normal regression mixtures. Chen and Li (2009) develop an EM test for $m_0 = 1$ in heteroscedastic normal mixtures, and Chen et al. (2012) develop an EM test for testing $H_0 : m = m_0$ against $H_A : m > m_0$ by splitting each component into two, thereby in effect testing against $H_A : m = 2m_0$, but neither Chen and Li (2009) nor Chen et al. (2012) accommodates regressors. Shen and He (2014) develop an EM test for testing

$H_0 : m_0 = 1$ in normal regression mixtures with covariate-dependent mixing proportions, in which the Fisher information matrix is regular. The test of Shen and He (2014) has a simpler limiting distribution, but its power is low when the mixing proportion does not depend on covariates. Our test is designed for such models, and as such, our test and theirs are mutually complementary.

Model selection procedures have been proposed for estimating the number of components in mixture models (see, for example, Henna, 1985; Lindsay and Roeder, 1992; Windham and Cutler, 1992; Roeder, 1994; Chen and Kalbfleisch, 1996; Keribin, 2000; James et al., 2001; Miloslavsky and van der Laan, 2003; Woo and Sriram, 2006; Chen and Khalili, 2008). As discussed in Chen et al. (2012), while model selection procedures seek to find a parsimonious model that adequately describes the observed data, the number of components is often linked to scientific propositions, and a hypothesis test can be used to check their validity.

The remainder of this paper is organized as follows. Section 2 introduces finite normal mixture regression models. Sections 3 and 4 establish the local quadratic approximation in testing the null hypothesis of m_0 components against the alternative of $m_0 + 1$ components. Section 5 introduces the modified EM test. Section 6 reports the simulation results, and empirical examples are provided in Section 7. The supplementary appendix contains proofs, auxiliary results, and additional empirical examples. All limits below are taken as $n \to \infty$, unless stated otherwise. Let := denote "equals by definition." For a $k \times 1$ vector $\boldsymbol{a}$ and a function $f(\boldsymbol{a})$, let $\nabla_{\boldsymbol{a}} f(\boldsymbol{a})$ denote the $k \times 1$ vector of the derivative $(\partial/\partial\boldsymbol{a})f(\boldsymbol{a})$, and let $\nabla_{\boldsymbol{a}\boldsymbol{a}^\top} f(\boldsymbol{a})$ denote the $k \times k$ vector of the derivative $(\partial/\partial\boldsymbol{a}\partial\boldsymbol{a}^\top)f(\boldsymbol{a})$.

2 Finite normal mixture regression models

Denote the density of a normal distribution with mean $\mu + \boldsymbol{x}^\top\boldsymbol{\beta} + \boldsymbol{z}^\top\boldsymbol{\gamma}$ and variance σ^2 by

$$f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}, \boldsymbol{\theta}, \sigma^2) := \frac{1}{\sigma}\phi\left(\frac{y - \mu - \boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{z}^\top\boldsymbol{\gamma}}{\sigma}\right),$$

where $\boldsymbol{\theta} := (\mu, \boldsymbol{\beta}^\top)^\top$ is $(q+1)\times 1$, μ is scalar, $\boldsymbol{x} = (x_1, \ldots, x_q)^\top$ and $\boldsymbol{\beta}$ are $q \times 1$, $\boldsymbol{z}$ and $\boldsymbol{\gamma}$ are $p \times 1$, and $\phi(t) := (2\pi)^{-1/2} \exp(-t^2/2)$. Let $\Theta_{\boldsymbol{\gamma}} \subset \mathbb{R}^p$, $\Theta_{\boldsymbol{\theta}} = \Theta_\mu \times \Theta_{\boldsymbol{\beta}} \subset \mathbb{R}^{q+1}$, and $\Theta_\sigma \subset \mathbb{R}_{++}$ denote the space of $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$, and σ^2, respectively. We consider an m-component finite mixture density:

$$f_m(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\vartheta}_m) := \sum_{j=1}^{m} \alpha_j f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}, \boldsymbol{\theta}_j, \sigma_j^2), \tag{1}$$

where $\boldsymbol{\vartheta}_m := (\boldsymbol{\alpha}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_m^\top, \sigma_1^2, \ldots, \sigma_m^2)^\top$ with $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_{m-1})^\top$, and α_m being determined by $\alpha_m := 1 - \sum_{j=1}^{m-1} \alpha_j$. $\boldsymbol{\theta}_j$ and σ_j^2 are mixing parameters that characterize the j-th component, $\boldsymbol{\gamma}$ is a structural parameter that is common to all the components, and α_js are mixing probabilities. Define the set of admissible values of $\boldsymbol{\alpha}$ by $\Theta_{\boldsymbol{\alpha}} := \{\boldsymbol{\alpha} : \alpha_j \geq 0, \sum_{j=1}^{m-1} \alpha_j \in [0, 1]\}$, and let the space of $\boldsymbol{\vartheta}_m$ be $\Theta_{\boldsymbol{\vartheta}_m} := \Theta_{\boldsymbol{\alpha}} \times \Theta_{\boldsymbol{\gamma}} \times \Theta_{\boldsymbol{\theta}}^m \times \Theta_\sigma^m$.

We define the number of components m by the smallest number such that the data density admits the representation (1). Our objective is to test

$$H_0 : \; m = m_0 \quad \text{against} \quad H_A : m = m_0 + 1.$$

3 Local quadratic approximation for testing $H_0 : m = 1$ against $H_A : m = 2$

In this section, we develop a local quadratic approximation for testing the null hypothesis $H_0 : m = 1$ against $H_A : m = 2$ when the data are from H_0. We consider a random sample of n independent observations $\{Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i\}_{i=1}^n$ from the true one-component density $f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*, \sigma^{2*})$. Here, the superscript $*$ denotes the true population value. Let a two-component mixture density function with $\boldsymbol{\vartheta}_2 = (\alpha, \boldsymbol{\gamma}^\top, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1^2, \sigma_2^2)^\top \in \Theta_{\boldsymbol{\vartheta}_2}$ be

$$f_2(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\vartheta}_2) := \alpha f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}, \boldsymbol{\theta}_1, \sigma_1^2) + (1-\alpha) f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}, \boldsymbol{\theta}_2, \sigma_2^2). \tag{2}$$

The model (2) yields the true density $f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*, \sigma^{2*})$ if $\boldsymbol{\vartheta}_2$ lies in the set $\Theta_2^* := \{\boldsymbol{\vartheta}_2 \in \Theta_{\boldsymbol{\vartheta}_2} : \{(\boldsymbol{\theta}_1, \sigma_1^2) = (\boldsymbol{\theta}_2, \sigma_2^2) = (\boldsymbol{\theta}^*, \sigma^{2*}), \boldsymbol{\gamma} = \boldsymbol{\gamma}^*\}$ or $\{\alpha(1-\alpha) = 0, \boldsymbol{\gamma} = \boldsymbol{\gamma}^*\}\}$.

We partition the null hypothesis $H_0 : m = 1$ into two as follows:

$$H_{01} : (\boldsymbol{\theta}_1, \sigma_1^2) = (\boldsymbol{\theta}_2, \sigma_2^2) \text{ and } H_{02} : \alpha(1-\alpha) = 0.$$

The regularity conditions for a standard asymptotic analysis fails in finite mixture models because (i) under H_{01}, α is not identified, and the Fisher information matrix for the other parameters becomes singular; (ii) under H_{02}, α is on the boundary of the parameter space, and either $\boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_2$ is not identified.

In addition to the failure of regularity conditions that is common to all finite mixture models, the normal mixture model (2) has additional undesirable mathematical properties, as discussed in Chen and Li (2009): (a) The Fisher information for testing H_{02} is not finite unless the range of σ_1^2/σ_2^2 is restricted. (b) The derivatives of $f_2(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\vartheta}_2)$ of different orders are linearly dependent because $\nabla_{\mu\mu} f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}, \boldsymbol{\theta}, \sigma^2) = 2\nabla_{\sigma^2} f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}, \boldsymbol{\theta}, \sigma^2)$ (loss of strong identifiability). (c) The log-likelihood function is unbounded and the maximum likelihood estimate fails to exist (Hartigan, 1985; Kiefer and Wolfowitz, 1956).

In view of problem (a), we focus on testing $H_{01} : (\boldsymbol{\theta}_1, \sigma_1^2) = (\boldsymbol{\theta}_2, \sigma_2^2)$ in the following. We handle problem (c) by considering a maximum likelihood estimator (MLE) in the constrained parameter space $\Theta_{\boldsymbol{\vartheta}_2}(\epsilon_\sigma) := \{\boldsymbol{\vartheta}_2 \in \Theta_{\boldsymbol{\vartheta}_2} : \min\{\sigma_1/\sigma_2, \sigma_2/\sigma_1\} \geq \epsilon_\sigma\}$ for some $\epsilon_\sigma > 0$, as in Hathaway (1985). Let $\hat{\boldsymbol{\vartheta}}_2$ denote the constrained MLE that maximizes $L_n(\boldsymbol{\vartheta}_2) := \sum_{i=1}^n f_2(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\vartheta}_2)$ under the constraint $\boldsymbol{\vartheta}_2 \in \Theta_{\boldsymbol{\vartheta}_2}(\epsilon_\sigma)$. The following proposition shows the consistency of $\hat{\boldsymbol{\vartheta}}_2$ by extending the result of Hathaway (1985) to accommodate covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$.

Assumption 1. *$\boldsymbol{X}$ and $\boldsymbol{Z}$ have finite second moment, and* $\Pr(\boldsymbol{X}_i^\top \boldsymbol{\beta}_j + \boldsymbol{Z}_i^\top \boldsymbol{\gamma} \neq \boldsymbol{X}_i^\top \boldsymbol{\beta}_j^* + \boldsymbol{Z}_i^\top \boldsymbol{\gamma}^*) > 0$ *for any* $(\boldsymbol{\beta}_j^\top, \boldsymbol{\gamma}^\top)^\top \neq ((\boldsymbol{\beta}_j^*)^\top, (\boldsymbol{\gamma}^*)^\top)^\top$ *and* $j = 1, \ldots, m$.

Proposition 1. *Suppose that Assumption 1 holds. Then, under the null hypothesis* H_0 : $m = 1$, $\inf_{\boldsymbol{\vartheta}_2 \in \Theta_2^*} ||\hat{\boldsymbol{\vartheta}}_2 - \boldsymbol{\vartheta}_2|| \to 0$ *almost surely.*

Let $l(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\vartheta}_2) := \ln \left(\alpha f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}, \boldsymbol{\theta}_1, \sigma_1^2) + (1-\alpha) f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}, \boldsymbol{\theta}_2, \sigma_2^2)\right)$ denote the log-density of the two-component model. For any $\bar{\boldsymbol{\vartheta}}_2$ such that $(\boldsymbol{\theta}_1, \sigma_1^2) = (\boldsymbol{\theta}_2, \sigma_2^2)$, the derivatives of the log-density are linearly dependent as

$$\nabla_{\boldsymbol{\theta}_1} l(y|\boldsymbol{x}, \boldsymbol{z}; \bar{\boldsymbol{\vartheta}}_2) = \frac{\alpha}{1-\alpha} \nabla_{\boldsymbol{\theta}_2} l(y|\boldsymbol{x}, \boldsymbol{z}; \bar{\boldsymbol{\vartheta}}_2), \; \nabla_{\sigma_1^2} l(y|\boldsymbol{x}, \boldsymbol{z}; \bar{\boldsymbol{\vartheta}}_2) = \frac{\alpha}{1-\alpha} \nabla_{\sigma_2^2} l(y|\boldsymbol{x}, \boldsymbol{z}; \bar{\boldsymbol{\vartheta}}_2), \tag{3}$$

$$\nabla_{\mu_j \mu_j} l(y|\boldsymbol{x}, \boldsymbol{z}; \bar{\boldsymbol{\vartheta}}_2) = 2 \nabla_{\sigma_j^2} l(y|\boldsymbol{x}, \boldsymbol{z}; \bar{\boldsymbol{\vartheta}}_2) \text{ for } j = 1, 2. \tag{4}$$

Consequently, the Fisher information matrix is degenerate, which invalidates the standard second-order quadratic approximation analysis. In particular, dependence (4) causes substantial difficulties in existing literature.

We analyze the LRT statistic for testing $H_{01} : (\boldsymbol{\theta}_1, \sigma_1) = (\boldsymbol{\theta}_2, \sigma_2)$ by developing a higher-order approximation of the log-likelihood function that can be expressed in a quadratic form, when $\alpha \in (0, 1)$, through a judiciously designed reparameterization by extending the result of Rotnitzky et al. (2000). Consider the following one-to-one mapping between $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1^2, \sigma_2^2) = (\mu_1, \boldsymbol{\beta}_1, \mu_2, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2)$ and the reparameterized parameter $(\boldsymbol{\lambda}_{\boldsymbol{\theta}}, \boldsymbol{\nu}_{\boldsymbol{\theta}}, \lambda_\sigma, \nu_\sigma) = (\lambda_\mu, \boldsymbol{\lambda}_{\boldsymbol{\beta}}, \nu_\mu, \boldsymbol{\nu}_{\boldsymbol{\beta}}, \lambda_\sigma, \nu_\sigma)$:

$$\begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ \sigma_1^2 \\ \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\nu}_{\boldsymbol{\theta}} + (1-\alpha)\boldsymbol{\lambda}_{\boldsymbol{\theta}} \\ \boldsymbol{\nu}_{\boldsymbol{\theta}} - \alpha \boldsymbol{\lambda}_{\boldsymbol{\theta}} \\ \nu_\sigma + (1-\alpha)(2\lambda_\sigma + C_1 \lambda_\mu^2) \\ \nu_\sigma - \alpha(2\lambda_\sigma + C_2 \lambda_\mu^2) \end{pmatrix}, \tag{5}$$

where $\boldsymbol{\nu}_{\boldsymbol{\theta}} = (\nu_\mu, \boldsymbol{\nu}_{\boldsymbol{\beta}}^\top)^\top$, $\boldsymbol{\lambda}_{\boldsymbol{\theta}} = (\lambda_\mu, \boldsymbol{\lambda}_{\boldsymbol{\beta}}^\top)^\top$, $C_1 := -(1/3)(1+\alpha)$, and $C_2 := (1/3)(2-\alpha)$.

Collect the reparameterized parameters, except for α, into one vector $\boldsymbol{\psi}_\alpha$ defined as

$$\boldsymbol{\psi}_\alpha := (\boldsymbol{\gamma}^\top, \boldsymbol{\nu_\theta}^\top, \nu_\sigma, \boldsymbol{\lambda_\theta}^\top, \lambda_\sigma)^\top = (\boldsymbol{\gamma}^\top, \nu_\mu, \boldsymbol{\nu_\beta}^\top, \nu_\sigma, \lambda_\mu, \boldsymbol{\lambda}_{\boldsymbol{\beta}}^\top, \lambda_\sigma)^\top \in \Theta_{\boldsymbol{\psi}_\alpha}. \tag{6}$$

In the reparameterized model, the null hypothesis of $H_{01} : (\boldsymbol{\theta}_1^\top, \sigma_1^2) = (\boldsymbol{\theta}_2^\top, \sigma_2^2)$ is written as $H_{01} : (\boldsymbol{\lambda_\theta}^\top, \lambda_\sigma) = (0, \ldots, 0)$, and the density and its logarithm are given by

$$\begin{aligned} g(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\psi}_\alpha, \alpha) = {} & \alpha f\left(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}, \boldsymbol{\nu_\theta} + (1-\alpha)\boldsymbol{\lambda_\theta}, \nu_\sigma + (1-\alpha)(2\lambda_\sigma + C_1\lambda_\mu^2)\right) \\ & + (1-\alpha) f\left(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}, \boldsymbol{\nu_\theta} - \alpha\boldsymbol{\lambda_\theta}, \nu_\sigma - \alpha(2\lambda_\sigma + C_2\lambda_\mu^2)\right), \end{aligned} \tag{7}$$

and $l(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\psi}_\alpha, \alpha) = \ln[g(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\psi}_\alpha, \alpha)]$.

Partition $\boldsymbol{\psi}_\alpha$ as $\boldsymbol{\psi}_\alpha = (\boldsymbol{\eta}^\top, \boldsymbol{\lambda}^\top)^\top$, where $\boldsymbol{\eta} := (\boldsymbol{\gamma}^\top, \boldsymbol{\nu_\theta}^\top, \nu_\sigma)^\top \in \Theta_{\boldsymbol{\eta}}$ and $\boldsymbol{\lambda} := (\boldsymbol{\lambda_\theta}, \lambda_\sigma)^\top \in \Theta_{\boldsymbol{\lambda}}$. Denote the true values of $\boldsymbol{\eta}$, $\boldsymbol{\lambda}$, and ψ by $\boldsymbol{\eta}^* := ((\boldsymbol{\gamma}^*)^\top, (\boldsymbol{\theta}^*)^\top, \sigma^{2*})^\top$, $\boldsymbol{\lambda}^* := (0, \ldots, 0)^\top$, and $\boldsymbol{\psi}_\alpha^* = ((\boldsymbol{\eta}^*)^\top, 0, \ldots, 0)^\top$, respectively. The first derivative of (7) w.r.t. $\boldsymbol{\eta}$ under $\boldsymbol{\psi}_\alpha = \boldsymbol{\psi}_\alpha^*$ is identical to the score of the one-component model:

$$\nabla_{\boldsymbol{\eta}} l(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\psi}_\alpha^*, \alpha) = \frac{\nabla_{(\boldsymbol{\gamma}^\top, \boldsymbol{\theta}^\top, \sigma^2)^\top} f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*, \sigma^{2*})}{f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*, \sigma^{2*})}. \tag{8}$$

On the other hand, Proposition C in the supplementary appendix shows that the first, second, and third derivatives of the reparameterized log-density (7) w.r.t. λ_μ, and the first derivative w.r.t. λ_σ or $\boldsymbol{\lambda_\beta}$ become zero when evaluated at $\boldsymbol{\psi}_\alpha = \boldsymbol{\psi}_\alpha^*$:

$$\begin{aligned} & \nabla_{\lambda_\mu} l(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\psi}_\alpha^*, \alpha) = 0, \quad \nabla_{\lambda_\mu^2} l(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\psi}_\alpha^*, \alpha) = 0, \quad \nabla_{\lambda_\mu^3} l(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\psi}_\alpha^*, \alpha) = 0, \\ & \nabla_{\lambda_\sigma} l(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\psi}_\alpha^*, \alpha) = 0, \quad \nabla_{\boldsymbol{\lambda_\beta}} l(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\psi}_\alpha^*, \alpha) = 0. \end{aligned} \tag{9}$$

Consequently, the information on λ_μ, λ_σ, and $\boldsymbol{\lambda_\beta}$ is provided by the derivatives w.r.t. λ_μ^4, λ_σ^2, $\lambda_\mu\lambda_\sigma$, $\boldsymbol{\lambda_\beta}\lambda_\mu$, $\boldsymbol{\lambda_\beta}\lambda_\sigma$, and $\text{vech}(\boldsymbol{\lambda_\beta}\boldsymbol{\lambda}_{\boldsymbol{\beta}}^\top)$. For $\boldsymbol{\lambda_\beta} := (\lambda_1, \ldots, \lambda_q)^\top \in \mathbb{R}^q$, collect the elements of $\text{vech}(\boldsymbol{\lambda_\beta}\boldsymbol{\lambda}_{\boldsymbol{\beta}}^\top)$ into a $q(q{+}1)/2{\times}1$ vector as $v_{\boldsymbol{\beta}}(\boldsymbol{\lambda_\beta}) = (v_{11}, \ldots, v_{qq}, v_{12}, \ldots, v_{1q}, v_{23}, \ldots, v_{2q}, \ldots, v_{q-1,q})^\top := (\lambda_1^2, \ldots, \lambda_q^2, \lambda_1\lambda_2, \ldots, \lambda_1\lambda_q, \lambda_2\lambda_3, \ldots, \lambda_2\lambda_q, \ldots, \lambda_{q-1}\lambda_q)^\top$. Collect the relevant parameters as

$$\boldsymbol{t}_n(\boldsymbol{\psi}_\alpha, \alpha) := \left(\boldsymbol{t}_{\boldsymbol{\eta}n}^\top, \boldsymbol{t}_{\boldsymbol{\lambda}n}^\top \right)^\top, \tag{10}$$

where $\boldsymbol{t}_{\boldsymbol{\eta}n} := n^{1/2}(\boldsymbol{\eta} - \boldsymbol{\eta}^*)$ and

$$\boldsymbol{t}_{\boldsymbol{\lambda}n} := \begin{pmatrix} t_{\mu\sigma n} \\ t_{\mu^4 n} \\ \boldsymbol{t}_{\boldsymbol{\beta}\mu n} \\ \boldsymbol{t}_{\boldsymbol{\beta}\sigma n} \\ \boldsymbol{t}_{\boldsymbol{\beta}^2 n} \end{pmatrix} = \begin{pmatrix} n^{1/2}\alpha(1-\alpha)6\lambda_\mu\lambda_\sigma \\ n^{1/2}\alpha(1-\alpha)[12\lambda_\sigma^2 + b(\alpha)\lambda_\mu^4] \\ n^{1/2}\alpha(1-\alpha)2\boldsymbol{\lambda}_{\boldsymbol{\beta}}\lambda_\mu \\ n^{1/2}\alpha(1-\alpha)6\boldsymbol{\lambda}_{\boldsymbol{\beta}}\lambda_\sigma \\ n^{1/2}\alpha(1-\alpha)\boldsymbol{v}_{\boldsymbol{\beta}}(\boldsymbol{\lambda}_{\boldsymbol{\beta}}) \end{pmatrix}, \tag{11}$$

with $b(\alpha) := -(2/3)(\alpha^2 - \alpha + 1) < 0$.

Let $L_n(\boldsymbol{\psi}_\alpha, \alpha) := \sum_{i=1}^n l(Y_i | \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\psi}_\alpha, \alpha)$ denote the reparameterized log-likelihood function. Define the normalized score as $\boldsymbol{S}_n := n^{-1/2} \sum_{i=1}^n \boldsymbol{s}_i$ and $\boldsymbol{\mathcal{I}}_n := n^{-1} \sum_{i=1}^n \boldsymbol{s}_i \boldsymbol{s}_i^\top$, where $\boldsymbol{s}_i := (\boldsymbol{s}_{\boldsymbol{\eta}i}^\top, \boldsymbol{s}_{\boldsymbol{\lambda}i}^\top)^\top = (\boldsymbol{s}_{\boldsymbol{\eta}i}^\top, \boldsymbol{s}_{\boldsymbol{\lambda}_{\mu\sigma}i}^\top, \boldsymbol{s}_{\boldsymbol{\lambda}_{\boldsymbol{\beta}}i}^\top)^\top$ and

$$\boldsymbol{s}_{\boldsymbol{\eta}i} := \begin{pmatrix} H_i^{1*}\boldsymbol{Z}_i \\ H_i^{1*} \\ H_i^{1*}\boldsymbol{X}_i \\ H_i^{2*} \end{pmatrix}, \quad \boldsymbol{s}_{\boldsymbol{\lambda}_{\mu\sigma}i} := \begin{pmatrix} H_i^{3*} \\ H_i^{4*} \\ H_i^{2*}\boldsymbol{X}_i \\ H_i^{3*}\boldsymbol{X}_i \end{pmatrix}, \quad \boldsymbol{s}_{\boldsymbol{\lambda}_{\boldsymbol{\beta}}i} := \begin{pmatrix} H_i^{2*}(X_{1i})^2 \\ \vdots \\ H_i^{2*}(X_{qi})^2 \\ 2H_i^{2*}X_{1i}X_{2i} \\ \vdots \\ 2H_i^{2*}X_{q-1,i}X_{qi} \end{pmatrix}, \tag{12}$$

and $H_i^{k*} := H^k\left((Y_i - \mu^* - \boldsymbol{X}_i^\top\boldsymbol{\beta}^* - \boldsymbol{Z}_i^\top\boldsymbol{\gamma}^*)/\sigma^*\right) / \left((\sigma^*)^k k!\right)$, where $H^k(z)$ is the Hermite polynomial of order k given by $H^1(z) = z$, $H^2(z) = z^2 - 1$, $H^3(z) = z^3 - 3z$, and $H^4(z) = z^4 - 6z^2 + 3$.

With these notations, expanding $L_n(\boldsymbol{\psi}_\alpha, \alpha)$ nine times around $(\boldsymbol{\psi}_\alpha^*, \alpha)$, we can write

$L_n(\boldsymbol{\psi}_\alpha,\alpha) - L_n(\boldsymbol{\psi}^*_\alpha,\alpha)$ as a quadratic function of $\boldsymbol{t}_n(\boldsymbol{\psi}_\alpha,\alpha)$ as

$$L_n(\boldsymbol{\psi}_\alpha,\alpha) - L_n(\boldsymbol{\psi}^*_\alpha,\alpha) = \boldsymbol{t}_n(\boldsymbol{\psi}_\alpha,\alpha)^\top \boldsymbol{S}_n - \frac{1}{2}\boldsymbol{t}_n(\boldsymbol{\psi}_\alpha,\alpha)^\top \boldsymbol{\mathcal{I}}_n \boldsymbol{t}_n(\boldsymbol{\psi}_\alpha,\alpha) + R_n(\boldsymbol{\psi}_\alpha,\alpha). \tag{13}$$

Define the variance of the score as $\boldsymbol{\mathcal{I}} := E[\boldsymbol{s}_i\boldsymbol{s}_i^\top] = E[\boldsymbol{\mathcal{I}}_n]$.

Assumption 2. *(a) $\boldsymbol{X}$ and $\boldsymbol{Z}$ have finite ninth moments. (b) $E[\boldsymbol{U}_1\boldsymbol{U}_1^\top]$ and $E[\boldsymbol{U}_2\boldsymbol{U}_2^\top]$ are nonsingular, where $\boldsymbol{U}_1 = (1,\boldsymbol{X}^\top,\boldsymbol{Z}^\top)^\top$ and $\boldsymbol{U}_2 = (1,\boldsymbol{X}^\top,X_1^2,\ldots,X_q^2,2X_1X_2,\ldots,2X_{q-1}X_q)^\top$.*

Proposition 2. *Suppose that Assumptions 1 and 2 hold. Then, under the null hypothesis $H_0: m=1$, for $\alpha\in(0,1)$ and $\epsilon_\sigma\in(0,1)$, we have (a) for any $\delta>0$, $\limsup_{n\to\infty}\Pr(\sup_{\boldsymbol{\psi}_\alpha\in\Theta_{\boldsymbol{\psi}_\alpha}:||\boldsymbol{\psi}_\alpha-\boldsymbol{\psi}^*_\alpha||\le\kappa}|R_n(\boldsymbol{\psi}_\alpha,\alpha)| > \delta(1+||\boldsymbol{t}_n(\boldsymbol{\psi}_\alpha,\alpha)||)^2)\to 0$ as $\kappa\to 0$, (b) $\boldsymbol{S}_n \to_d \boldsymbol{S}\sim N(0,\boldsymbol{\mathcal{I}})$, and (c) $\boldsymbol{\mathcal{I}}_n\to_p\boldsymbol{\mathcal{I}}$, where $\boldsymbol{\mathcal{I}}$ is nonsingular.*

Let $\hat{\boldsymbol{\psi}}_\alpha := \arg\max_{\boldsymbol{\psi}_\alpha\in\Theta_{\boldsymbol{\psi}_\alpha}(\epsilon_\sigma)} L_n(\boldsymbol{\psi}_\alpha,\alpha)$ denote the (constrained) MLE of $\boldsymbol{\psi}_\alpha$, where $\Theta_{\boldsymbol{\psi}_\alpha}(\epsilon_\sigma)$ is defined so that the value of $\boldsymbol{\vartheta}_2$ implied by $\boldsymbol{\psi}_\alpha$ is in $\Theta_{\boldsymbol{\vartheta}_2}(\epsilon_\sigma)$. Let $(\hat{\gamma}_0,\hat{\boldsymbol{\theta}}_0,\hat{\sigma}_0^2)$ denote the one-component MLE that maximizes the one-component log-likelihood function $L_{0,n}(\boldsymbol{\gamma},\boldsymbol{\theta},\sigma^2) := \sum_{i=1}^n \ln f(Y_i|\boldsymbol{X}_i,\boldsymbol{Z}_i;\boldsymbol{\gamma},\boldsymbol{\theta},\sigma^2)$. Define the LRT statistic for testing H_{01} as $LR_n(\epsilon_1) := \max_{\alpha\in[\epsilon_1,1-\epsilon_1]} 2\{L_n(\hat{\boldsymbol{\psi}}_\alpha,\alpha) - L_{0,n}(\hat{\gamma}_0,\hat{\boldsymbol{\theta}}_0,\hat{\sigma}_0^2)\}$ with $\epsilon_1\in(0,1/2)$.

Let Λ_n be the set of admissible values of $\boldsymbol{t}_n(\boldsymbol{\psi}_\alpha,\alpha) = (\boldsymbol{t}_{\boldsymbol{\eta}n}^\top,\boldsymbol{t}_{\boldsymbol{\lambda}n}^\top)^\top$ defined in (10). The asymptotic null distribution of $LR_n(\epsilon_1)$ is characterized by the supremum of the quadratic-form representation of the log-likelihood ratio under the constraint implied by the limit of Λ_n as $n\to\infty$. As shown in the proof of Proposition 3, the limit of Λ_n is given by the union of $\Lambda^1_{\boldsymbol{\lambda}}$ and $\Lambda^2_{\boldsymbol{\lambda}}$, where $q_{\boldsymbol{\lambda}} := 2+2q+q(q+1)/2$ and

$$\begin{aligned}\Lambda^1_{\boldsymbol{\lambda}} &:= \{\boldsymbol{t}_{\boldsymbol{\lambda}} = (t_{\mu\sigma},t_{\mu^4},\boldsymbol{t}_{\boldsymbol{\beta}\mu}^\top,\boldsymbol{t}_{\boldsymbol{\beta}\sigma}^\top,\boldsymbol{t}_{\boldsymbol{\beta}^2}^\top)^\top\in\mathbb{R}^{q_{\boldsymbol{\lambda}}} : (t_{\mu\sigma},t_{\mu^4},\boldsymbol{t}_{\boldsymbol{\beta}\mu}^\top)^\top\in\mathbb{R}\times\mathbb{R}_-\times\mathbb{R}^q, \boldsymbol{t}_{\boldsymbol{\beta}\sigma} = \boldsymbol{t}_{\boldsymbol{\beta}^2} = 0\} \quad \text{and}\\ \Lambda^2_{\boldsymbol{\lambda}} &:= \{\boldsymbol{t}_{\boldsymbol{\lambda}} = (t_{\mu\sigma},t_{\mu^4},\boldsymbol{t}_{\boldsymbol{\beta}\mu}^\top,\boldsymbol{t}_{\boldsymbol{\beta}\sigma}^\top,\boldsymbol{t}_{\boldsymbol{\beta}^2}^\top)^\top\in\mathbb{R}^{q_{\boldsymbol{\lambda}}} : t_{\mu\sigma} = 6\lambda_\mu\lambda_\sigma, t_{\mu^4} = 12\lambda_\sigma^2, \boldsymbol{t}_{\boldsymbol{\beta}\mu} = 2\boldsymbol{\lambda}_{\boldsymbol{\beta}}\lambda_\mu,\\ &\qquad \boldsymbol{t}_{\boldsymbol{\beta}\sigma} = 6\boldsymbol{\lambda}_{\boldsymbol{\beta}}\lambda_\sigma, \boldsymbol{t}_{\boldsymbol{\beta}^2} = v_{\boldsymbol{\beta}}(\boldsymbol{\lambda}_{\boldsymbol{\beta}}) \text{ for some } \boldsymbol{\lambda}\in\mathbb{R}^{2+q}\}.\end{aligned} \tag{14}$$

Partition $\boldsymbol{S}$ and $\boldsymbol{W} = \boldsymbol{\mathcal{I}}^{-1}\boldsymbol{S}$ as $\boldsymbol{S} = (\boldsymbol{S}_{\boldsymbol{\eta}}, \boldsymbol{S}_{\boldsymbol{\lambda}})^\top$, $\boldsymbol{W} = (\boldsymbol{W}_{\boldsymbol{\eta}}^\top, \boldsymbol{W}_{\boldsymbol{\lambda}}^\top)$, where $\boldsymbol{S}_{\boldsymbol{\eta}}$ and $\boldsymbol{W}_{\boldsymbol{\eta}}$ are $(p+q+2)\times 1$, and $\boldsymbol{S}_{\boldsymbol{\lambda}}$ and $\boldsymbol{W}_{\boldsymbol{\lambda}}$ are $q_{\boldsymbol{\lambda}} \times 1$. Define $\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}} := E(\boldsymbol{S}_{\boldsymbol{\eta}}\boldsymbol{S}_{\boldsymbol{\eta}}^\top)$, $\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\eta}} := E(\boldsymbol{S}_{\boldsymbol{\lambda}}\boldsymbol{S}_{\boldsymbol{\eta}}^\top)$, $\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta\lambda}} := \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\eta}}^\top$, and $\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}} := E(\boldsymbol{S}_{\boldsymbol{\lambda}}\boldsymbol{S}_{\boldsymbol{\lambda}}^\top)$. Note that $\boldsymbol{W}_{\boldsymbol{\lambda}} = \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^{-1}\boldsymbol{S}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}$, where $\boldsymbol{S}_{\boldsymbol{\lambda}.\boldsymbol{\eta}} := \boldsymbol{S}_{\boldsymbol{\lambda}} - \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\eta}}\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}}^{-1}\boldsymbol{S}_{\boldsymbol{\eta}}$ and $\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}} := \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}} - \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\eta}}\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}}^{-1}\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta\lambda}}$. For $j = 1, 2$, define $\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^j$ by

$$r_{\boldsymbol{\lambda}}(\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^j) = \inf_{\boldsymbol{t}_{\boldsymbol{\lambda}} \in \Lambda_{\boldsymbol{\lambda}}^j} r_{\boldsymbol{\lambda}}(\boldsymbol{t}_{\boldsymbol{\lambda}}), \quad r_{\boldsymbol{\lambda}}(\boldsymbol{t}_{\boldsymbol{\lambda}}) := (\boldsymbol{t}_{\boldsymbol{\lambda}} - \boldsymbol{W}_{\boldsymbol{\lambda}})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}(\boldsymbol{t}_{\boldsymbol{\lambda}} - \boldsymbol{W}_{\boldsymbol{\lambda}}). \tag{15}$$

The following proposition establishes the asymptotic null distribution of the LRT statistic. When the model does not have a conditioning variable $\boldsymbol{X}$, the set of admissible values of $\boldsymbol{t}_{\boldsymbol{\lambda}n}$ converges to $\mathbb{R}^2$, and $LR_n(\epsilon_1)$ converges to $\chi^2(2)$ in distribution.

Proposition 3. *Suppose that Assumptions 1 and 2 hold, $\epsilon_\sigma \in (0,1)$, and $\epsilon_1 \in (0, 1/2)$. Then, under the null hypothesis $H_0 : m = 1$, (a) $\boldsymbol{t}_n(\hat{\boldsymbol{\psi}}_\alpha, \alpha) = O_p(1)$ for any $\alpha \in (0,1)$, (b) $LR_n(\epsilon_1) \rightarrow_d \chi^2(2)$ if the model does not have a conditioning variable $\boldsymbol{X}$, and (c) $LR_n(\epsilon_1) \rightarrow_d \max\{(\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^1)^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^1, (\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^2)^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^2\}$.*

Under the alternative $H_A : m = 2$, the constrained MLE $\hat{\boldsymbol{\psi}} := \arg\max_{\alpha \in [\epsilon_1, 1-\epsilon_1]} L_n(\hat{\boldsymbol{\psi}}_\alpha, \alpha)$ is consistent if the true parameter value $\boldsymbol{\vartheta}_2^*$ lies in the set $\Theta_{\boldsymbol{\vartheta}_2}(\epsilon_\sigma, \epsilon_1) := \{\boldsymbol{\vartheta}_2 \in \Theta_{\boldsymbol{\vartheta}_2} : \min\{\sigma_1/\sigma_2, \sigma_2/\sigma_1\} \geq \epsilon_\sigma$ and $\alpha \in [\epsilon_1, 1-\epsilon_1]\}$ that involve ad hoc constants ϵ_σ and ϵ_1. If $\boldsymbol{\vartheta}_2^*$ does not lie in $\Theta_{\boldsymbol{\vartheta}_2}(\epsilon_\sigma, \epsilon_1)$, the constrained MLE converges to the value of ψ_α that minimizes the Kullback–Leibler divergence between the true density $f_2(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\vartheta}_2^*)$ and the class of density $\{f_2(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\vartheta}_2) : \boldsymbol{\vartheta}_2 \in \Theta_{\boldsymbol{\vartheta}_2}(\epsilon_\sigma, \epsilon_1)\}$ (White, 1982). Consequently, $LR_n(\epsilon_1)$ may have reduced power against alternatives in which $(\alpha, \sigma_1, \sigma_2)$ does not satisfy the constraint $\Theta_{\boldsymbol{\vartheta}_2}(\epsilon_\sigma, \epsilon_1)$.

Liu and Shao (2003) analyze the LRT statistic of mixture models. Their Corollary 4.1 with $(m,t) = (2,2)$ corresponds to testing the null hypothesis $H_{01}^* : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}^*$. According to their Corollary 4.1, the generalized score function of the LRT statistic for testing H_{01}^* is obtained by expanding the log-likelihood function twice. Our Proposition 3 and its proof show that in heteroscedastic normal mixture models, the likelihood function needs to be expanded further.

When $\boldsymbol{X}$ contains dummy variables that take value 0 or 1, Assumption 2(b) fails because a dummy variable and its square in $\boldsymbol{U}_2$ are perfectly correlated with each other. The following proposition handles such cases. Define $\boldsymbol{U}_2^0 := (\boldsymbol{X}^\top, X_1^2, \ldots, X_q^2, 2X_1X_2, \ldots, 2X_{q-1}X_q)^\top$, so that $\boldsymbol{U}_2 = (1, (\boldsymbol{U}_2^0)^\top)^\top$, and let $q_2 := \dim(U_2^0) = q + q(q+1)/2$.

Assumption 3. *(a) There exists a $d \times q_2$ matrix $\boldsymbol{B}$ of rank d such that $\boldsymbol{B}\boldsymbol{U}_2^0 = 0$. (b) $E[\boldsymbol{U}_1\boldsymbol{U}_1^\top]$ and $E[\boldsymbol{B}^\perp \boldsymbol{U}_2^0 (\boldsymbol{B}^\perp \boldsymbol{U}_2^0)^\top]$ are nonsingular, where $\boldsymbol{B}^\perp$ is a $(q_2 - d) \times q_2$ matrix whose rows are the basis of the orthogonal complement of the row space of $\boldsymbol{B}$.*

Proposition 4. *Suppose that Assumptions 1, 2(a), and 3 hold. Then, $LR_n(\epsilon_1) \rightarrow_d$ $\max\{(\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^{B1})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^B \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^{B1}, (\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^{B2})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^B \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^{B2}\}$, where the definitions of $\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^{Bj}$ and $\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^B$ are provided in the proof in the supplementary appendix.*

4 Local quadratic approximation for testing $H_0 : m = m_0$ against $H_A : m = m_0 + 1$ for $m_0 \geq 2$

In this section, we develop a local quadratic approximation for testing the null hypothesis of m_0 components against the alternative of $m_0 + 1$ components for general $m_0 \geq 1$. We consider a random sample of n independent observations $\{Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i\}_{i=1}^n$ generated from the m_0-component normal mixture density with the true parameter value $\boldsymbol{\vartheta}_{m_0}^* := (\alpha_1^*, \ldots, \alpha_{m_0-1}^*, (\boldsymbol{\gamma}^*)^\top, (\boldsymbol{\theta}_1^*)^\top, \ldots, (\boldsymbol{\theta}_m^*)^\top, \sigma_1^{2*}, \ldots., \sigma_m^{2*})^\top$ with $\alpha_j^* > 0$:

$$f_{m_0}(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\vartheta}_{m_0}^*) := \sum_{j=1}^{m_0} \alpha_j^* f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}_j^*, \sigma_j^{2*}). \tag{16}$$

We assume $(\boldsymbol{\theta}_1^*, \sigma_1^{2*}) < \ldots < (\boldsymbol{\theta}_{m_0}^*, \sigma_{m_0}^{2*})$ for identification. Let the density of an (m_0+1)-component mixture model be

$$f_{m_0+1}(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\vartheta}_{m_0+1}) := \sum_{j=1}^{m_0+1} \alpha_j f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}, \boldsymbol{\theta}_j, \sigma_j^2), \tag{17}$$

where $\boldsymbol{\vartheta}_{m_0+1} := (\alpha_1, \ldots, \alpha_{m_0}, \boldsymbol{\gamma}^\top, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_{m_0+1}^\top, \sigma_1^2, \ldots, \sigma_{m_0+1}^2)^\top$. Similar to the case of the test of homogeneity, we partition the null hypothesis into two as $H_0 = H_{01} \cup H_{02}$, where $H_{01} := \cup_{h=1}^{m_0} H_{0,1h}$ and $H_{02} := \cup_{h=1}^{m_0+1} H_{0,2h}$ with

$$H_{0,1h} : (\boldsymbol{\theta}_1, \sigma_1^2) < \cdots < (\boldsymbol{\theta}_h, \sigma_h^2) = (\boldsymbol{\theta}_{h+1}, \sigma_{h+1}^2) < \cdots < (\boldsymbol{\theta}_{m_0+1}, \sigma_{m_0+1}^2) \text{ and } H_{0,2h} : \alpha_h = 0.$$

The inequality constraints are imposed on $(\boldsymbol{\theta}_j, \sigma_j^2)$ for identification.

We focus on testing H_{01} because (i) the LRT statistic for testing H_{02} has infinite Fisher information unless a stringent restriction is imposed on the admissible values of σ_j^2, and (ii) implementing the LRT for H_{02} is practically difficult because the asymptotic null distribution depends on the functional of a multidimensional Gaussian process. Define the set of values of $\boldsymbol{\vartheta}_{m_0+1}$ that yields the true density (16) as $\Upsilon^* := \{\boldsymbol{\vartheta}_{m_0+1} : f_{m_0+1}(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\vartheta}_{m_0+1}) = f_{m_0}(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\vartheta}_{m_0}^*)$ with probability one$\}$. Under $H_{0,1h}$, the (m_0+1)-component model (17) generates the true m_0-component density (16) when $(\boldsymbol{\theta}_h, \sigma_h^2) = (\boldsymbol{\theta}_{h+1}, \sigma_{h+1}^2) = (\boldsymbol{\theta}_h^*, \sigma_h^{2*})$. Define the subset of Υ^* corresponding to $H_{0,1h}$ as

$$\begin{aligned} \Upsilon_{1h}^* := \big\{ & \boldsymbol{\vartheta}_{m_0+1} \in \Theta_{\boldsymbol{\vartheta}_{m_0+1}} : \alpha_j > 0 \text{ for } j = 1, \ldots, m_0+1;\ \alpha_h + \alpha_{h+1} = \alpha_h^* \text{ and} \\ & (\boldsymbol{\theta}_h, \sigma_h^2) = (\boldsymbol{\theta}_{h+1}, \sigma_{h+1}^2) = (\boldsymbol{\theta}_h^*, \sigma_h^{2*});\ \alpha_j = \alpha_j^* \text{ and } (\boldsymbol{\theta}_j, \sigma_j^2) = (\boldsymbol{\theta}_j^*, \sigma_j^{2*}) \text{ for } j < h; \\ & \alpha_j = \alpha_{j-1}^* \text{ and } (\boldsymbol{\theta}_j, \sigma_j^2) = (\boldsymbol{\theta}_{j-1}^*, \sigma_{j-1}^{2*}) \text{ for } j > h+1;\ \boldsymbol{\gamma} = \boldsymbol{\gamma}^* \big\}, \end{aligned}$$

and define $\Upsilon_1^* := \Upsilon_{11}^* \cup \cdots \cup \Upsilon_{1m_0}^*$.

Similar to Section 3, we consider the MLE in the constrained parameter space $\Theta_{\boldsymbol{\vartheta}_m}(\epsilon_\sigma) :=$

$\{\boldsymbol{\vartheta}_m \in \Theta_{\boldsymbol{\vartheta}_m} : \min_{j,k}\{\sigma_j/\sigma_k\} \geq \epsilon_\sigma\}$ for some $\epsilon_\sigma > 0$. For $\epsilon_1 \in (0, 1/2)$, let $\Theta_{\boldsymbol{\vartheta}_{m_0+1}}(\epsilon_\sigma, \epsilon_1)$ be a subset of $\Theta_{\boldsymbol{\vartheta}_{m_0+1}}(\epsilon_\sigma)$ such that $\alpha_j \in [\epsilon_1, 1-\epsilon_1]$ for $j = 1, \ldots, m_0+1$, and define the LRT statistic for testing H_{01} as $LR_n^{m_0}(\epsilon_1) := \max_{\boldsymbol{\vartheta}_{m_0+1} \in \Theta_{\boldsymbol{\vartheta}_{m_0+1}}(\epsilon_\sigma,\epsilon_1)} 2\{L_n(\boldsymbol{\vartheta}_{m_0+1}) - L_{0,n}(\hat{\boldsymbol{\vartheta}}_{m_0})\}$, where $L_n(\boldsymbol{\vartheta}_{m_0+1}) := \sum_{i=1}^n \ln f_{m_0+1}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\vartheta}_{m_0+1})$, $L_{0,n}(\boldsymbol{\vartheta}_{m_0}) := \sum_{i=1}^n \ln f_{m_0}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\vartheta}_{m_0})$, and $\hat{\boldsymbol{\vartheta}}_{m_0} := \arg\max_{\boldsymbol{\vartheta}_{m_0} \in \Theta_{\boldsymbol{\vartheta}_{m_0}}(\epsilon_\sigma)} L_{0,n}(\boldsymbol{\vartheta}_{m_0})$. Collect the score vector for testing $H_{0,11}, \ldots, H_{0,1m_0}$ into one vector as

$$\tilde{\boldsymbol{s}}_i := \begin{pmatrix} \tilde{\boldsymbol{s}}_{\boldsymbol{\eta}i} \\ \tilde{\boldsymbol{s}}_{\boldsymbol{\lambda}i} \end{pmatrix}, \quad \text{where } \tilde{\boldsymbol{s}}_{\boldsymbol{\eta}i} := \begin{pmatrix} \boldsymbol{s}_{\boldsymbol{\alpha}i} \\ \boldsymbol{s}_{\boldsymbol{\gamma}i} \\ \boldsymbol{s}_{(\boldsymbol{\theta},\sigma)i} \end{pmatrix} \text{ and } \tilde{\boldsymbol{s}}_{\boldsymbol{\lambda}i} := \begin{pmatrix} \boldsymbol{s}^1_{\boldsymbol{\lambda}_{\mu\sigma}i} \\ \boldsymbol{s}^1_{\boldsymbol{\lambda}_{\boldsymbol{\beta}}i} \\ \vdots \\ \boldsymbol{s}^{m_0}_{\boldsymbol{\lambda}_{\mu\sigma}i} \\ \boldsymbol{s}^{m_0}_{\boldsymbol{\lambda}_{\boldsymbol{\beta}}i} \end{pmatrix}, \tag{18}$$

with $\boldsymbol{s}_{\boldsymbol{\alpha}i} := (f^*_{1,i} - f^*_{m_0,i}, \ldots, f^*_{m_0-1,i} - f^*_{m_0,i})^\top / f_{m_0}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\vartheta}^*_{m_0})$, $\boldsymbol{s}_{\boldsymbol{\gamma}i} := \boldsymbol{Z}_i \sum_{j=1}^{m_0} w^*_{j,i} H^{1*}_{j,i}$, and

$$\boldsymbol{s}_{(\boldsymbol{\theta},\sigma)i} := \begin{pmatrix} w^*_{1,i}H^{1*}_{1,i} \\ \vdots \\ w^*_{m_0,i}H^{1*}_{m_0,i} \\ w^*_{1,i}H^{1*}_{1,i}\boldsymbol{X}_i \\ \vdots \\ w^*_{m_0,i}H^{1*}_{m_0,i}\boldsymbol{X}_i \\ w^*_{1,i}H^{2*}_{1,i} \\ \vdots \\ w^*_{m_0,i}H^{2*}_{m_0,i} \end{pmatrix}, \quad \begin{array}{l} \boldsymbol{s}^h_{\boldsymbol{\lambda}_{\mu\sigma}i} := \begin{pmatrix} w^*_{h,i}H^{3*}_{h,i} \\ w^*_{h,i}H^{4*}_{h,i} \\ w^*_{h,i}H^{2*}_{h,i}\boldsymbol{X}_i \\ w^*_{h,i}H^{3*}_{h,i}\boldsymbol{X}_i \end{pmatrix}, \\ \boldsymbol{s}^h_{\boldsymbol{\lambda}_{\boldsymbol{\beta}}i} := \begin{pmatrix} w^*_{h,i}H^{2*}_{h,i}(X_{1i})^2 \\ \vdots \\ w^*_{h,i}H^{2*}_{h,i}(X_{qi})^2 \\ 2w^*_{h,i}H^{2*}_{h,i}X_{1i}X_{2i} \\ \vdots \\ 2w^*_{h,i}H^{2*}_{h,i}X_{q-1,i}X_{qi} \end{pmatrix}, \end{array} \tag{19}$$

where $f^*_{j,i} := f(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*_j, \sigma^{2*}_j)$, $H^{k*}_{j,i} := H^k\left((Y_i - \mu^*_j - \boldsymbol{X}_i^\top \beta^*_j - \boldsymbol{Z}_i^\top \boldsymbol{\gamma}^*)/\sigma^*_j\right) / \left((\sigma^*_j)^k k!\right)$, and $w^*_{j,i} := \alpha^*_j f^*_{j,i} / f_{m_0}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\vartheta}^*_{m_0})$. Define $\tilde{\boldsymbol{\mathcal{I}}} := E[\tilde{\boldsymbol{s}}_i \tilde{\boldsymbol{s}}_i^\top]$, $\tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\eta}} := E[\tilde{\boldsymbol{s}}_{\boldsymbol{\eta}i} \tilde{\boldsymbol{s}}_{\boldsymbol{\eta}i}^\top]$, $\tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda\eta}} :=$

$E[\tilde{\boldsymbol{s}}_{\boldsymbol{\lambda}i}\tilde{\boldsymbol{s}}_{\boldsymbol{\eta}i}^{\top}]$, $\tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\eta\lambda}} := \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda\eta}}$, $\tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda}} := E[\tilde{\boldsymbol{s}}_{\boldsymbol{\lambda}i}\tilde{\boldsymbol{s}}_{\boldsymbol{\lambda}i}^{\top}]$, and $\tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}} := \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda\eta}}\tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\eta}}^{-1}\tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\eta\lambda}}$. Let $\tilde{\boldsymbol{S}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}} := ((\boldsymbol{S}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^{1})^{\top}, \ldots, (\boldsymbol{S}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^{m_0})^{\top})^{\top} \sim N(0, \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}})$ be an $\mathbb{R}^{m_0 q_{\boldsymbol{\lambda}}}$-valued random vector, and define $\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^{h} := E[\boldsymbol{S}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^{h}(\boldsymbol{S}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^{h})^{\top}]$ and $\boldsymbol{W}_{\boldsymbol{\lambda},h} := (\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^{h})^{-1}\boldsymbol{S}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^{h}$. Similar to $\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^{j}$ in the test of homogeneity, define $\hat{\boldsymbol{t}}_{\boldsymbol{\lambda},h}^{j}$ by $r_{\boldsymbol{\lambda}}^{h}(\hat{\boldsymbol{t}}_{\boldsymbol{\lambda},h}^{j}) = \inf_{\boldsymbol{t}_{\boldsymbol{\lambda}} \in \Lambda_{\boldsymbol{\lambda}}^{j}} r_{\boldsymbol{\lambda}}^{h}(\boldsymbol{t}_{\boldsymbol{\lambda}})$ for $j = 1, 2$, where $r_{\boldsymbol{\lambda}}^{h}(\boldsymbol{t}_{\boldsymbol{\lambda}}) := (\boldsymbol{t}_{\boldsymbol{\lambda}} - \boldsymbol{W}_{\boldsymbol{\lambda},h})^{\top}\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^{h}(\boldsymbol{t}_{\boldsymbol{\lambda}} - \boldsymbol{W}_{\boldsymbol{\lambda},h})$. The following proposition gives the asymptotic null distribution of the LRT statistic for testing H_{01}. In the neighborhood of Υ_{1h}^{*}, the log-likelihood function permits a similar quadratic approximation to the one we derived in Section 3. Consequently, the LRT statistic is asymptotically distributed as the maximum of m_0 random variables, each of which is the maximum of two random variables.

Assumption 4. *(a)* $\alpha_j^{*} \in [\epsilon_1, 1 - \epsilon_1]$ *for* $j = 1, \ldots, m_0$. *(b)* $\min_{(j,k)\in\{1,\ldots,m_0\}}\{\sigma_j^{*}/\sigma_k^{*}\} > \epsilon_{\sigma}$. *(c)* $\tilde{\boldsymbol{\mathcal{I}}}$ *is finite and nonsingular.*

Proposition 5. *Suppose that Assumptions 1 and 4 hold. Then, under the null hypothesis* $H_0 : m = m_0$, $LR_n^{m_0}(\epsilon_1) \rightarrow_d \max\{v_1, \ldots, v_{m_0}\}$, *where* $v_h := \max\{(\hat{\boldsymbol{t}}_{\boldsymbol{\lambda},h}^{1})^{\top}\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^{h}\hat{\boldsymbol{t}}_{\boldsymbol{\lambda},h}^{1}, (\hat{\boldsymbol{t}}_{\boldsymbol{\lambda},h}^{2})^{\top}\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}}^{h}\hat{\boldsymbol{t}}_{\boldsymbol{\lambda},h}^{2}\}$.

In the remainder of this section, we derive a necessary and sufficient condition under which the LRT statistic for testing H_{02} has finite Fisher information. For brevity, we focus on the case without $(\boldsymbol{X}, \boldsymbol{Z})$. The score for testing $H_{0,2h} : \alpha_h = 0$ takes the form $\nabla_{\alpha_h} \ln f_{m_0+1}(Y_i; \boldsymbol{\vartheta}_{m_0+1}) = [f(Y_i; \mu_h, \sigma_h^2) - f(Y_i; \mu_{m_0}^{*}, \sigma_{m_0}^{2*})]/f_{m_0}(Y_i; \boldsymbol{\vartheta}_{m_0}^{*})$. Define the subset of Υ^{*} corresponding to $H_{0,2h} : \alpha_h = 0$ as $\Upsilon_{2h}^{*} := \{\boldsymbol{\vartheta}_{m_0+1} \in \Theta_{\boldsymbol{\vartheta}_{m_0+1}} : \alpha_h = 0; \alpha_j = \alpha_j^{*}$ and $(\mu_j, \sigma_j^2) = (\mu_j^{*}, \sigma_j^{2*})$ for $j < h$; $\alpha_j = \alpha_{j-1}^{*}$ and $(\mu_j, \sigma_j^2) = (\mu_{j-1}^{*}, \sigma_{j-1}^{2*})$ for $j > h + 1\}$. Because (μ_h, σ_h^2) is not identified when $\alpha_h = 0$, the Fisher information of the LRT for testing $H_{0,2h} : \alpha_h = 0$ depends on the supremum of the variance of $\nabla_{\alpha_h} \ln f_{m_0+1}(Y_i; \boldsymbol{\vartheta}_{m_0+1})$ over $\boldsymbol{\vartheta}_{m_0+1} \in \Upsilon_{2h}^{*}$. As shown in the following proposition, the Fisher information is infinite, unless a stringent restriction is imposed on the admissible values of σ_j^2.

Proposition 6. $\sup_{\boldsymbol{\vartheta}_{m_0+1} \in \Upsilon_{2h}^{*}} E[\{\nabla_{\alpha_h} \ln(f_{m_0+1}(Y_i; \boldsymbol{\vartheta}_{m_0+1}))\}^2] < \infty$ *if and only if* $\max\{\sigma^2 : \sigma^2 \in \Theta_{\sigma}\} < 2\max\{\sigma_1^{2*}, \ldots, \sigma_{m_0}^{2*}\}$.

5 Modified EM test

In this section, we develop a test of $H_0 : m = m_0$ against $H_1 : m = m_0 + 1$ for model (16). First, we develop a modified EM test statistic for testing $H_{0,1h} : (\boldsymbol{\theta}_h, \sigma_h^2) = (\boldsymbol{\theta}_{h+1}, \sigma_{h+1}^2)$. We construct m_0 intervals $\{D_1^*, \cdots, D_{m_0}^*\}$ of admissive values of $(\boldsymbol{\theta}, \sigma^2)$, such that $(\boldsymbol{\theta}_h^*, \sigma_h^{2*}) \in D_h^*$ but $(\boldsymbol{\theta}_j^*, \sigma_j^{2*}) \notin D_h^*$ for any $j \neq h$. For example, as in our simulation, we may assume that μ_h^*s are distinct and set $D_1^* = [\underline{\Theta}_\mu, (\mu_1^* + \mu_2^*)/2] \times \Theta_{\boldsymbol{\beta}} \times \Theta_\sigma$, $D_j^* = [(\mu_{j-1}^* + \mu_j^*)/2, (\mu_j^* + \mu_{j+1}^*)/2] \times \Theta_{\boldsymbol{\beta}} \times \Theta_\sigma$ for $j = 2, \ldots, m_0 - 1$, and $D_{m_0}^* = [(\mu_{m_0-1}^* + \mu_{m_0}^*)/2, \overline{\Theta}_\mu] \times \Theta_{\boldsymbol{\beta}} \times \Theta_\sigma$, where $\underline{\Theta}_\mu$ and $\overline{\Theta}_\mu$ are defined by $\Theta_\mu = [\underline{\Theta}_\mu, \overline{\Theta}_\mu]$ and may take either the value $-\infty$ or ∞.

Collect the mixing parameters of the $(m_0 + 1)$-component model into one vector as $\boldsymbol{\varsigma} := (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_{m_0+1}^\top, \sigma_1^2, \ldots, \sigma_{m_0+1}^2)^\top \in \Theta_{\boldsymbol{\varsigma}} := \Theta_{\boldsymbol{\theta}}^{m_0+1} \times \Theta_\sigma^{m_0+1}$. For $h = 1, \ldots, m_0$, define a restricted parameter space of $\boldsymbol{\varsigma}$ by $\Omega_h^* := \{\boldsymbol{\varsigma} \in \Theta_{\boldsymbol{\varsigma}} : (\boldsymbol{\theta}_j, \sigma_j^2) \in D_j^*$ for $j = 1, \ldots, h - 1$; $(\boldsymbol{\theta}_h, \sigma_h^2)$, $(\boldsymbol{\theta}_{h+1}, \sigma_{h+1}^2) \in D_h^*$; $(\boldsymbol{\theta}_j, \sigma_j^2) \in D_{j-1}^*$ for $j = h + 2, \ldots, m_0 + 1\}$. Let $\hat{\Omega}_h$ and $\hat{D}_h$ be consistent estimates of Ω_h^* and D_h^*, which can be constructed from a consistent estimate of the m_0-component model. We test $H_{0,1h} : (\boldsymbol{\theta}_h, \sigma_h^2) = (\boldsymbol{\theta}_{h+1}, \sigma_{h+1}^2)$ by estimating the $(m_0 + 1)$-component model (17) under the restriction $\boldsymbol{\varsigma} \in \hat{\Omega}_h$. For example, when we test $H_{0,11} : (\boldsymbol{\theta}_1, \sigma_1^2) = (\boldsymbol{\theta}_2, \sigma_2^2)$ in a three-component model, the restriction can be given as $(\boldsymbol{\theta}_1, \sigma_1^2), (\boldsymbol{\theta}_2, \sigma_2^2) \in \hat{D}_1$ and $(\boldsymbol{\theta}_3, \sigma_3^2) \in \hat{D}_2$.

Define the penalized log-likelihood function for the (m_0+1)-component model by $PL_n(\boldsymbol{\vartheta}_{m_0+1}) := L_n(\boldsymbol{\vartheta}_{m_0+1}) + \sum_{j=1}^{m_0+1} p_n(\sigma_j^2)$, where $p_n(\sigma^2)$ is a penalty function that satisfies Assumptions 5 and 6 below. Let $\mathcal{T}$ be a finite set of numbers from $(0, 0.5]$. For each $\tau_0 \in \mathcal{T}$, define the restricted penalized MLE as $\boldsymbol{\vartheta}_{m_0+1}^{h(1)}(\tau_0) := \arg\max_{\boldsymbol{\vartheta}_{m_0+1} \in \Theta^h(\tau_0)} PL_n(\boldsymbol{\vartheta}_{m_0+1})$, where $\Theta^h(\tau_0) := \{\boldsymbol{\vartheta}_{m_0+1} \in \Theta_{\boldsymbol{\vartheta}_{m_0+1}} : \alpha_h/(\alpha_h + \alpha_{h+1}) = \tau_0$ and $\boldsymbol{\varsigma} \in \hat{\Omega}_h\}$.

Starting from $\boldsymbol{\vartheta}_{m_0+1}^{h(1)}(\tau_0)$, we update $\boldsymbol{\vartheta}_{m_0+1}$ by the following generalized EM algorithm. Henceforth, we suppress (τ_0) from $\boldsymbol{\vartheta}_{m_0+1}^{h(k)}(\tau_0)$. Suppose we have already calculated $\boldsymbol{\vartheta}_{m_0+1}^{h(k)}$. For $i = 1, \ldots, n$ and $j = 1, \ldots, m_0 + 1$, define the weights for an E-step as $w_{ij}^{(k)} := \alpha_j^{(k)} f(Y_i | \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\gamma}^{(k)}, \boldsymbol{\theta}_j^{(k)}, \sigma_j^{2(k)}) / f_{m_0+1}(Y_i | \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\vartheta}_{m_0+1}^{h(k)})$.

In an M-step, update $\boldsymbol{\vartheta}_{m_0+1}$ by $\boldsymbol{\gamma}^{(k+1)} := (\sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i^\top)^{-1}[\sum_{i=1}^n \boldsymbol{Z}_i(Y_i - \sum_{j=1}^{m_0+1} w_{ij}^{(k)} \tilde{\boldsymbol{X}}_i^\top \boldsymbol{\theta}_j^{(k)})]$, and for $j = 1, \ldots, m_0 + 1$, by $\alpha_j^{(k+1)} := n^{-1} \sum_{i=1}^n w_{ij}^{(k)}$ and

$$\boldsymbol{\theta}_j^{(k+1)} := \left(\sum_{i=1}^n w_{ij}^{(k)} \tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{X}}_i^\top \right)^{-1} \left[\sum_{i=1}^n w_{ij}^{(k)} \tilde{\boldsymbol{X}}_i \left(Y_i - \boldsymbol{Z}_i^\top \boldsymbol{\gamma}^{(k+1)} \right) \right], \tag{20}$$

$$\sigma_j^{2(k+1)} := \arg\max_{\sigma_j^2} \left\{ \sum_{i=1}^n w_{ij}^{(k)} \ln f(Y_i | \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\gamma}^{(k+1)}, \boldsymbol{\theta}_j^{(k+1)}, \sigma_j^2) + p_n(\sigma_j^2) \right\},$$

where $\tilde{\boldsymbol{X}}_i := (1, \boldsymbol{X}_i^\top)^\top$. The penalized likelihood value never decreases after each generalized EM step (Dempster et al., 1977, Theorem 1). Note that $\boldsymbol{\vartheta}_{m_0+1}^{h(k)}$ for $k \geq 2$ does not use the restriction $\hat{\Omega}_h$. For each $\tau_0 \in \mathcal{T}$ and k, define

$$\mathrm{M}_n^{h(k)}(\tau_0) := 2 \left\{ L_n(\boldsymbol{\vartheta}_{m_0+1}^{h(k)}(\tau_0)) - L_{0,n}(\tilde{\boldsymbol{\vartheta}}_{m_0}) \right\}, \tag{21}$$

where $\tilde{\boldsymbol{\vartheta}}_{m_0} := \arg\max_{\boldsymbol{\vartheta}_{m_0} \in \Theta_{\boldsymbol{\vartheta}_{m_0}}} L_{0,n}(\boldsymbol{\vartheta}_{m_0}) + \sum_{j=1}^{m_0} p_n(\sigma_j^2)$.

Finally, with a pre-specified number K, define the *local modified EM test statistic* for testing $H_{0,1h}$ by taking the maximum of $\mathrm{M}_n^{h(K)}(\tau_0)$ over $\tau_0 \in \mathcal{T}$ as $\mathrm{EM}_n^{h(K)} := \max \left\{ \mathrm{M}_n^{h(K)}(\tau_0) : \tau_0 \in \mathcal{T} \right\}$. The *modified EM test statistic* is defined as the maximum of m_0 local modified EM test statistics: $\mathrm{EM}_n^{(K)} := \max \left\{ \mathrm{EM}_n^{1(K)}, \mathrm{EM}_n^{2(K)}, \ldots, \mathrm{EM}_n^{m_0(K)} \right\}$. The following proposition shows that for any finite K, the modified EM test statistic is asymptotically equivalent to the LRT statistic for testing H_{01}. Assumptions 5 and 6 are adopted from Chen et al. (2008) and Chen and Li (2009).

Assumption 5. *(a)* $\sup_{\sigma^2 > 0} \max\{0, p_n(\sigma^2)\} = o(n)$ *and* $p_n(\sigma^2) = o(n)$ *at any fixed* $\sigma > 0$. *(b) For any* $\sigma \in (0, 8/(nM)]$*, we have* $p_n(\sigma^2) \leq 5(\ln n)^2 \ln \sigma$ *for a sufficiently large* n*, where* $M = \sup_{y,z,x} f_{m_0}(y | \boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\vartheta}_{m_0}^*)$.

Assumption 6. $\nabla_{\sigma^2} p_n(\sigma^2) = o_p(n^{1/4})$.

Proposition 7. *Suppose that Assumptions 1, 4, 5, and 6 hold. Then, under the null hypothesis* $H_0 : m = m_0$*, for any fixed finite* K*, as* $n \to \infty$*,* $EM_n^{(K)} \to_d \max\{v_1, \ldots, v_{m_0}\}$*, where* *the v_hs are given in Proposition 5.*

6 Asymptotic testing power

In this section, we study the asymptotic testing power of the modified EM test for testing $H_0 : m = 1$ against $H_A : m = 2$. Consider the following local alternative to the homogeneous model $f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*, \sigma^{2*})$. For $\alpha^* \in (0, 1)$ and $\boldsymbol{\Delta} \in \mathbb{R}^{q_{\boldsymbol{\lambda}}}$, let $H^n_{(\alpha^*,\boldsymbol{\Delta})} : \alpha = \alpha^*, \boldsymbol{\gamma} = \boldsymbol{\gamma}^*$, $(\boldsymbol{\nu}_{\boldsymbol{\theta}}, \nu_\sigma) = (\boldsymbol{\theta}^*, \sigma^{*2})$, and $\boldsymbol{t}_{\boldsymbol{\lambda}n} = \boldsymbol{\Delta}$. This local alternative is contiguous to the null distribution $f(y|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*, \sigma^{2*})$. Using our quadratic-form approximation, the asymptotic distribution of the modified EM statistic under $H^n_{(\alpha^*,\boldsymbol{\Delta})}$ is derived from LeCam's contiguity theory. The following proposition shows that the modified EM test for testing $H_0 : m = 1$ is locally optimal for this class of alternatives.

Proposition 8. *Suppose that Assumptions 1 and 2 and $H^n_{(\alpha^*,\boldsymbol{\Delta})}$ hold with $\alpha^* \in [\epsilon_1, 1-\epsilon_1]$. Let $\boldsymbol{W}_{\boldsymbol{\lambda}} \sim N(0, \boldsymbol{\mathcal{I}}^{-1}_{\boldsymbol{\lambda}.\boldsymbol{\eta}})$ and let $\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}n}$ be an estimator of $\boldsymbol{t}_{\boldsymbol{\lambda}n}$ in (11). Define $\hat{\boldsymbol{t}}^j_{\boldsymbol{\lambda},\boldsymbol{\Delta}}$ in the same manner as $\hat{\boldsymbol{t}}^j_{\boldsymbol{\lambda}}$ in (15) but using $\boldsymbol{W}_{\boldsymbol{\lambda}} + \boldsymbol{\Delta}$ in place of $\boldsymbol{W}_{\boldsymbol{\lambda}}$. Then, (a) $LR_{1,n}(\epsilon_1), EM^{(K)}_n \rightarrow_d (\boldsymbol{W}_{\boldsymbol{\lambda}} + \boldsymbol{\Delta})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}} (\boldsymbol{W}_{\boldsymbol{\lambda}} + \boldsymbol{\Delta})$ if the model has no conditioning variable $\boldsymbol{X}$, (b) $LR_{1,n}(\epsilon_1), EM^{(K)}_n \rightarrow_d \max\{(\hat{\boldsymbol{t}}^1_{\boldsymbol{\lambda},\boldsymbol{\Delta}})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}} \hat{\boldsymbol{t}}^1_{\boldsymbol{\lambda},\boldsymbol{\Delta}}, (\hat{\boldsymbol{t}}^2_{\boldsymbol{\lambda},\boldsymbol{\Delta}})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}} \hat{\boldsymbol{t}}^2_{\boldsymbol{\lambda},\boldsymbol{\Delta}}\}$, and (c) $\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}n} \rightarrow_d \hat{\boldsymbol{t}}^1_{\boldsymbol{\lambda},\boldsymbol{\Delta}} \hat{\xi} + \hat{\boldsymbol{t}}^2_{\boldsymbol{\lambda},\boldsymbol{\Delta}} (1 - \hat{\xi})$, where $\hat{\xi} = \boldsymbol{1}\{(\hat{\boldsymbol{t}}^1_{\boldsymbol{\lambda},\boldsymbol{\Delta}})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}} \hat{\boldsymbol{t}}^1_{\boldsymbol{\lambda},\boldsymbol{\Delta}} > (\hat{\boldsymbol{t}}^2_{\boldsymbol{\lambda},\boldsymbol{\Delta}})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}.\boldsymbol{\eta}} \hat{\boldsymbol{t}}^2_{\boldsymbol{\lambda},\boldsymbol{\Delta}}\}$.*

As highlighted by the use of constant ϵ_1 in the definition of $LR_{1,n}(\epsilon_1)$, Proposition 8 does not cover a sequence of local alternatives where α approaches 0.

7 Simulation

7.1 Choice of penalty function

To apply our modified EM test, we need to specify the set $\mathcal{T}$, number of iterations K, and penalty function for $p_n(\sigma^2)$. Based on our experience, we recommend $\mathcal{T} = \{0.5\}$ and $K = 2$

or 3; following the recommendation given by Chen et al. (2012), we set

$$p_n(\sigma_j^2; \hat{\sigma}_j^2) = -a_n\{\hat{\sigma}_j^2/\sigma_j^2 + \ln(\sigma_j^2/\hat{\sigma}_j^2) - 1\}, \tag{22}$$

where $\hat{\sigma}_j^2$ is the estimate from the m_0-component model. $p_n(\sigma_j^2; \hat{\sigma}_j^2)$ satisfies Assumptions 5 and 6 if $a_n = o_p(n^{1/4})$. In models with a regressor, we use an additional restriction $\sigma_j \geq 0.01\hat{\sigma}_j$, which does not change the theoretical results but improves finite sample performance.

For the model without a conditioning variable $\boldsymbol{X}$, we set $a_n = 0.25$ for testing $H_0 : m = 1$ while we develop the following data-dependent empirical formula for tuning parameter a_n for the cases where $m_0 = 2$ and $m_0 = 3$:

$$a_n = \begin{cases} 1.8q_n(\omega_{12}, n)/(1 + q_n(\omega_{12}, n)) & \text{for } m_0 = 2, \\ 1.5q_n(\omega_{12}, \omega_{23}, n)/(1 + q_n(\omega_{12}, \omega_{23}, n)) & \text{for } m_0 = 3, \end{cases} \tag{23}$$

with $q_n(\omega_{12}, n) = \exp(-1.645 - 0.435\ln(\omega_{12}/(1-\omega_{12})) - 101.60/n)$ and $q_n(\omega_{12}, \omega_{23}, n) = \exp(-1.679 - 0.232\ln(\omega_{12}\omega_{23}/[(1-\omega_{12})(1-\omega_{23})]) - 175.67/n)$, where ω_{12} and ω_{23} are misclassification rates (Maitra and Melnykov, 2010). In a two-component normal mixture model, $\omega_{1|2} = \Pr(\alpha_1 f(Y; \mu_1, \sigma_1^2) > \alpha_2 f(Y; \mu_2, \sigma_2^2))$ gives the probability that an observation Y from component 2 is misclassified into component 1. Similarly, let $\omega_{2|1}$ denote the opposite misclassification rate. Then, ω_{12} is defined as the average of $\omega_{1|2}$ and $\omega_{2|1}$, and ω_{23} is defined as the average of $\omega_{2|3}$ and $\omega_{3|2}$. The empirical formula in (23) is obtained through computer experiments that are similar to those in Chen and Li (2009) and Chen et al. (2012).

For the model with a conditioning variable $\boldsymbol{X}$, the value of a_n that gives accurate Type I errors is sensitive to the dimension of $\boldsymbol{X}$. For testing $H_0 : m = 1$, we choose the value of a_n depending on the dimension of $\boldsymbol{X}$ so that Type I errors are accurate at $n = 200$ using $10,000$ replications as described in the last column of Table 4. For the cases where $m_0 = 2$ and $m_0 = 3$, developing a data-dependent empirical formula for a_n is difficult, and therefore, we set the value of a_n as in Table 4 and use both asymptotic and bootstrap critical values.

7.2 Simulation results

For the model without conditioning variable $\boldsymbol{X}$, we examine the finite sample performance of our proposed modified EM test by simulations and compare it with that of the EM test of Chen et al. (2012) [CLF, hereafter]. Computation was done using R (R Core Team, 2014). We use $5,000$ replications, and the sample sizes are set to 200 and 400. The m_0-component estimate $\tilde{\boldsymbol{\vartheta}}_{m_0}$ in (21) is computed with the penalty function $p_n(\sigma_j^2)$ defined in (22), where $a_n = 1/n$ and $\hat{\sigma}_j^2$ is set to the sample variance of the data for all j.

We simulated Type I error rates for 12 null models with orders 2 and 3 that are the same as in CLF, as specified in Table 1. The simulation results for $H_0 : m = 2$ are summarized in Figure 1. Overall, the finite sample size properties of the modified EM test are very good and similar to those of the EM test of CLF, even though the size of the modified EM test tends to have more dispersion across models. Figure 2 reports the simulation results for $H_0 : m = 3$. The results are similar to those for $H_0 : m = 2$. We examined the power of the modified EM test by considering 10 alternative models with order 3, and 8 alternative models with order 4. The simulation results are summarized in Tables 2 and 3. In both tables, the first four models are the same as those used in CLF. In many cases, our modified EM test has stronger power than the EM test of CLF. The power difference is the most significant when σ_js are heterogeneous.

We also examine the performance of our proposed modified EM test for the model with conditioning variable $\boldsymbol{X}$. Computation was done using Matlab. We use $10,000$ replications when testing $H_0 : m = 1$, and $1,000$ and 500 replications when testing $H_0 : m = 2$ and $H_0 : m = 3$, respectively. For testing $H_0 : m = 1$, Table 4 reports that our choice of a_n gives accurate Type I errors even at $n = 100$ for $\dim(\boldsymbol{X}) = 1, \ldots, 4$. Table 6 reports the power simulation results under 8 alternative models as specified in Table 5. The power tends to decline with the dimension of $\boldsymbol{X}$ but increases with the sample size.

We simulated Type I error rates for 16 null models as specified in Table 7 with orders 2 and

3 and with a scalar conditioning variable. Figures 3 and 4 summarize the simulation results, where bootstrap critical values with 199 bootstrap replications and asymptotic critical values are used. When compared with Figures 1 and 2, the Type I error rates vary more across the models. When testing $H_0 : m = 2$, the bootstrap test performs well, while the asymptotic test is oversized in some models when $n = 200$. When testing $H_0 : m = 3$, the modified EM test performs fairly well, even though it tends to overreject and underreject when $n = 200$ and $n = 400$, respectively. Tables 8 and 9 report power simulation results when testing $H_0 : m = 2$ and $H_0 : m = 3$, respectively. The modified EM test has a good power under both bootstrap and asymptotic critical values.

8 Empirical Applications

8.1 Analysis of Stock Returns

Using $4,639$ observations of daily returns for 30 stocks in the Dow Jones Industrial Average, Kon (1984) estimated a finite mixture of normal distributions. Kon (1984) selected the number of components using LRT, but based on the invalid chi-squared asymptotic distribution. We re-estimated the number of components by sequentially applying the modified EM test with $K = 2$ to test $m = k$ against the alternative $m = k + 1$ for $k = 1, 2, 3$ at the $(1/3) \times 5\%$ significance level. Table 10 shows the frequency of the number of components selected by Kon (1984), the modified EM test, Akaike Information Criteria (AIC), and Bayesian Information Criteria (BIC). In Kon's dataset, some stocks return data contain a substantial number of observations with $y_i = 0.0$, which leads to degeneracy (i.e, $\hat{\mu}_j = 0$ and $\hat{\sigma}_j \to 0$). To deal with this problem, we discarded estimates such that $\max_j(\hat{\sigma}_j)/\min_j(\hat{\sigma}_j) > 100$. Kon (1984) often chooses a two-component model over other procedures. Compared with the modified EM test, AIC tends to select a larger number of components, whereas BIC selects a model with fewer components.

8.2 Analysis of Differential Gene Expression

A finite normal mixture model provides an approach to finding differentially expressed genes by means of the posterior probability that an individual gene is non-differentially expressed (Lee et al., 2000; Efron et al., 2001). We analyzed the leukemia dataset of Golub et al. (1999), which consists of the quantitative expression levels of $7,129$ genes from 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML). The data can be downloaded from `http://waldo.wi.mit.edu/MPR/data_se\bs{t}_ALL_AML.html`. Using the approach of Efron (2004), we computed the two-sample t-statistic comparing 47 ALL patients with 25 AML patients for each of the $7,129$ genes, and obtained z-values as $z = \Phi^{-1}(1-p)$, in which p is the p-value of the t-statistic and Φ is the $N(0,1)$ distribution function. If the gene is not differentially expressed, the corresponding $\boldsymbol{z}$-value should then follow $N(0,1)$. A large $\boldsymbol{z}$-value implies an overexpressed gene in patients with ALL. We applied normal mixtures to model the $\boldsymbol{z}$-values. Table 11 reports the parameter estimates $\tilde{\boldsymbol{\vartheta}}_{m_0}$, the p-value of the modified EM test using $K = 2$ in testing $H_0 : m = m_0$ against $H_0 : m = m_0 + 1$, AIC, and BIC for $m_0 = 1, 2, 3$. The modified EM test and AIC selected $m = 3$, whereas BIC chose $m = 2$. As shown in Figure 5, the three-component model more clearly captured data density, as compared to the two-component model. The three-component model classified approximately 26.8% of the genes as overexpressed in patients with ALL, whereas the two-component model estimated that α_2 comprised approximately 35.6% of the genes, and classified more genes as overexpressed compared to the three-component model.

8.3 Cross-Country Growth Regression

Mankiw et al. (1992) estimate a cross-country growth regression model to investigate a hypothesis that a country's growth rate is negatively related to its initial per capita gross domestic product (GDP). We examined the possibility of multiple regimes by considering the following mixture of regressions: $\ln(Y/L)_{i,1985} - \ln(Y/L)_{i,1960} = \mu_j + \beta_{1j} \ln(Y/L)_{i,1960} +$

$\beta_{2j} \ln(I/Y)_i + \beta_{3j} \ln(n_i + 0.05) + \epsilon_{ji}$, where $\epsilon_{ji} \sim_{iid} N(0, \sigma_j^2)$, $(Y/L)_{it}$ represents country i's per capita GDP at year t, $(I/Y)_i$ is the average ratio of investment to GDP from 1960 to 1985, and n_i is the average growth rate of working-age population from 1960 to 1985. We used the data of 75 "Intermediate" countries from Mankiw et al. (1992) that exclude small countries and countries with poor measurement of GDP. When testing $H_0 : m = 1$ against $H_0 : m = 2$, the asymptotic p-values of the modified EM test are 0.064 and 0.057 at $K = 2$ and 3, respectively, providing some evidence for two regimes. On the other hand, the bootstrapped p-values of the modified EM test for testing $H_0 : m = 2$ against $H_0 : m = 3$ are 0.311 and 0.315 at $K = 2$ and 3, respectively, providing no evidence for three regimes.

8.4 Spread of a Viral Infection in Potato Plants

Turner (2000) uses normal regression mixtures to model the spread of a viral infection in potato plants by aphids. The data set is from an experiment described in Boiteau et al. (1998). In each experiment, a grid of 81 potato plants was placed on the floor of a flight chamber, and varying number of aphids between 1 and 320 were released near the center of the grid, and the plants were taken out to measure the number of infected plants after one day. A total of 51 such experiments were conducted. The response variable is the number of infected plants, and the explanatory variables are the constant and the number of aphids. When testing $H_0 : m = 1$ against $H_A : m = 2$, the asymptotic p-values of the modified EM test are 0.000 at $K = 2$ and 3, providing strong evidence against the one-component model. On the other hand, when testing $H_0 : m = 2$ against $H_A : m = 3$, the bootstrapped p-values of the modified EM test are 0.142 and 0.154 at $K = 2$ and 3, respectively. Therefore, a two-component model is not rejected at the 10% significance level.

9 Supplemental Materials

Technical Details: The supplementary appendix contains proofs, auxiliary results, details of computer experiments, and additional results from empirical examples.

References

Andrews, D. W. K. (1999), "Estimation When a Parameter is on a Boundary," *Econometrica*, 67, 1341–1383.

Azaïs, J.-M., Gassiat, E., and Mercadier, C. (2009), "The Likelihood Ratio Test for General Mixture Models with or without Structural Parameter," *ESAIM: Probability and Statistics*, 13, 301—327.

Boiteau, G., Singh, M., Singh, R., Tai, G., and Turner, T. (1998), "Rate of spread of PVY^n by alate Myzus Persicae (Sulzer) from infected to healthy plants under laboratory conditions," *Potatl Research*, 41, 335–344.

Chen, H. and Chen, J. (2001), "The Likelihood Ratio Test for Homogeneity in Finite Mixture Models," *Canadian Journal of Statistics*, 29, 201–215.

— (2003), "Tests for Homogeneity in Normal Mixtures in the Presence of a Structural Parameter," *Statistica Sinica*, 13, 351–365.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2004), "Testing for a Finite Mixture Model with Two Components," *Journal of the Royal Statistical Society, Series B*, 66, 95–115.

Chen, J. (1995), "Optimal Rate of Convergence for Finite Mixture Models," *Annals of Statistics*, 23, 221–233.

Chen, J. and Kalbfleisch, J. D. (1996), "Penalized Minimum-Distance Estimates in Finite Mixture Models," *Canadian Journal of Statistics*, 24, 167–175.

Chen, J. and Khalili, A. (2008), "Order Selection in Finite Mixture Models With a Non-Smooth Penalty," *Journal of the American Statistical Association*, 103, 1674–1683.

Chen, J. and Li, P. (2009), "Hypothesis Test for Normal Mixture Models: The EM Approach," *Annals of Statistics*, 37, 2523–2542.

Chen, J., Li, P., and Fu, Y. (2012), "Inference on the Order of a Normal Mixture," *Journal of the American Statistical Association*, 107, 1096–1105.

Chen, J., Tan, X., and Zhang, R. (2008), "Inference for Normal Mixtures in Mean and Variance," *Statistica Sinica*, 18, 443–465.

Chernoff, H. and Lander, E. (1995), "Asymptotic Distribution of the Likelihood Ratio Test that a Mixture of two Binomials is a Single Binomial," *Journal of Statistical Planning and Inference*, 43, 19–40.

Conway, K. and Deb, P. (2005), "Is parental care really ineffective? Or, is the 'deveil' in the distribution?" *Journal of Health Economics*, 24, 489–513.

Dacunha-Castelle, D. and Gassiat, E. (1999), "Testing the Order of a Model using Locally Conic Parametrization: Population Mixtures and Stationary ARMA Processes," *Annals of Statistics*, 27, 1178–1209.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via EM Algorithm (with Discussion)," *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Efron, B. (2004), "Large-scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96—104.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151—1160.

Garel, B. (2001), "Likelihood Ratio Test for Univariate Gaussian Mixture," *Journal of Statistical Planning and Inference*, 96, 325–350.

— (2005), "Asymptotic Theory of the Likelihood Ratio Test for the Identification of a Mixture," *Journal of Statistical Planning and Inference*, 131, 271–296.

Ghosh, J. K. and Sen, P. K. (1985), "On the Asymptotic Performance of the Log-likelihood Ratio Statistic for the Mixture Model and Related Results," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, eds. Le Cam, L. and Olshen, R., Belmont, CA: Wadsworth, vol. 2, pp. 789–806.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537.

Hartigan, J. (1985), "Failure of Log-likelihood Ratio Test," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, eds. Le Cam, L. and Olshen, R., Berkeley: University of California Press, vol. 2, pp. 807–810.

Hathaway, R. J. (1985), "A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions," *Annals of Statistics*, 13, 795–800.

Henna, J. (1985), "On estimating of the number of constituents of a finite mixture of continuous distributions," *The Annals of the Institute of Statistical Mathematics*, 37, 235–240.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991), "Adaptive mixtures of local experts," *Neural Computation*, 3, 79–87.

James, L. F., Priebe, C. E., and Marchette, D. J. (2001), "The Annals of Statistics," *Consistent estimation of mixture complexity*, 29, 1281–1296.

Keribin, C. (2000), "Consistent estimation of the order of mixture models," *Sankhya Series A*, 62, 49–62.

Kiefer, J. and Wolfowitz, J. (1956), "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *Annals of Mathematical Statistics*, 27, 887–906.

Kon, S. (1984), "Models of Stock Returns – A Comparison," *Journal of Finance*, 39, 147–165.

Kon, S. and Jen, F. (1978), "Estimation of time-varying systematic risk and performance for mutual fund portfolios: An application of switching regression," *Journal of Finance*, 33, 457–475.

Lee, M.-L. T., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000), "Importance of Replication in Microarray Gene Expression Studies: Statistical Methods and Evidence from Repetitive cDNA Hybridizations," *Proc. Natl Acad. Sci. USA*, 9834–9838.

Lemdani, M. and Pons, O. (1997), "Likelihood Ratio Tests for Genetic Linkage," *Statistics and Probability Letters*, 33, 15–22.

Li, P. and Chen, J. (2010), "Testing the Order of a Finite Mixture," *Journal of the American Statistical Association*, 105, 1084–1092.

Li, P., Chen, J., and Marriott, P. (2009), "Non-finite Fisher Information and Homogeneity: An EM Approach," *Biometrika*, 96, 411–426.

Lindsay, B. G. (1995), *Mixture Models: Theory, Geometry, and Applications*, Bethesda, MD: Institute of Mathematical Statistics.

Lindsay, B. G. and Roeder, K. (1992), "Residual diagnostics for mixture models," *Journal of the American Statistical Association*, 87, 785–794.

Liu, X. and Shao, Y. (2003), "Asymptotics for Likelihood Ratio Tests under Loss of Identifiability," *Annals of Statistics*, 31, 807–832.

Maitra, R. and Melnykov, V. (2010), "Simulating Data to Study Performance of Finite Mixture Modeling and Model-Based Clustering Algorithms," *Journal of Computational and Graphical Statistics*, 19, 354—376.

Mankiw, N. G., Romer, D., and Weil, D. N. (1992), "A Contribution to the Empirics of Economic Growth," *Quarterly Journal of Economics*, 107, 407–437.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Miloslavsky, M. and van der Laan, M. J. (2003), "Fitting of Mix- tures With Unspecified Number of Components Using Cross-Validation Distance Estimate," *Computational Statistics and Data Analysis*, 41, 413–428.

Quandt, R. E. and Ramsey, J. B. (1978), "Estimating Mixtures of Normal Distributions and Switching Regressions," *Journal of the American Statistical Association*, 73, 730–738.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Roeder, K. (1994), "A graphical technique for detecting the number of components in a mixture of normals," *Journal of the American Statistical Association*, 89, 487–495.

Rotnitzky, A., Cox, D. R., Bottai, M., and Robins, J. (2000), "Likelihood-based Inference with Singular Information Matrix," *Bernoulli*, 6, 243–284.

Shen, J. and He, X. (2014), "Inference for Subgroup Analysis with a Structured Logistic-Normal Mixture Model," *Journal of the American Statistical Association*, forthcoming.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Tucker, A. L. (1992), "A Reexamination of Finite- and Infinite-Variance Distributions as Models of Daily Stock Returns," *Journal of Business & Economic Statistics*, 10, 73–81.

Turner, T. R. (2000), "Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49, 371–384.

Venkataraman, S. (1997), "Value at Risk for a Mixture of Normal Distributions: The Use of Quasi-Bayesian Estimation Techniques," *Economic Perspectives, Federal Reserve Bank of Chicago*, 21, 2–13.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

Windham, M. P. and Cutler, A. (1992), "Information ratios for validating mixture analysis," *Journal of the American Statistical Association*, 87, 1188–1192.

Woo, M.-J. and Sriram, T. N. (2006), "Robust estimation of mixture complexity," *Journal of the American Statistical Association*, 101, 1475–1486.

Zhu, H. and Zhang, H. (2006), "Asymptotics for Estimation and Testing Procedures under Loss of Identifiability," *Journal of Multivariate Analysis*, 97, 19–45.

Zhu, H.-T. and Zhang, H. (2004), "Hypothesis Testing in Mixture Regression Models," *Journal of the Royal Statistical Society, Series B*, 66, 3–16.

Table 1: Parameter specifications for 12 null models with order 2 and order 3

Order 2		Order 3	
$(\alpha_1, \alpha_2) =$	$(0.5, 0.5)$, $(0.2, 0.8)$	$(\alpha_1, \alpha_2, \alpha_3) =$	$(1/3, 1/3, 1/3)$, $(0.25, 0.5, 0.25)$
$(\mu_1, \mu_2) =$	$(-1.25, 1.25)$, $(-1.75, 1.75)$, $(-2.25, 2.25)$	$(\mu_1, \mu_2, \mu_3) =$	$(-3.5, 0, 4.5)$, $(-4.5, 0, 4.5)$
$(\sigma_1, \sigma_2) =$	$(1, 1)$, $(1.2, 0.6)$	$(\sigma_1, \sigma_2, \sigma_3) =$	$(1, 1, 1)$, $(0.6, 1.2, 0.6)$, $(0.6, 0.6, 1.2)$

Table 2: Powers (in %) of the Modified EM test and EM test for testing $H_0 : m = 2$ at the 5% level

			Modified EM test						EM test of CLF	
Alternative models			$n = 200$			$n = 400$			$n = 200$	$n = 400$
(μ_1, μ_2, μ_3)	$(\sigma_1, \sigma_2, \sigma_3)$	α_j	$K = 1$	$K = 2$	$K = 3$	$K = 1$	$K = 2$	$K = 3$	$K = 3$	$K = 3$
$(-2.5, 0, 2.5)$	$(1, 1, 1)$	A	27.2	27.7	27.9	66.3	66.4	66.5	26.3	62.1
$(-2.5, 0, 2.5)$	$(1, 1, 1)$	B	26.0	26.2	26.3	55.3	55.4	55.4	24.3	51.3
$(-2.5, 0, 2.5)$	$(0.6, 1.2, 0.6)$	A	99.5	99.5	99.5	100.0	100.0	100.0	99.1	100.0
$(-2.5, 0, 2.5)$	$(0.6, 1.2, 0.6)$	B	99.5	99.5	99.5	100.0	100.0	100.0	99.2	100.0
$(-2.0, 0, 2.0)$	$(0.6, 1.2, 0.6)$	A	69.7	69.9	70.1	96.8	96.8	96.8	64.6	95.0
$(-2.0, 0, 2.0)$	$(0.6, 1.2, 0.6)$	B	64.8	65.3	65.6	94.2	94.2	94.3	58.8	92.2
$(-2.0, 0, 4.0)$	$(1, 1, 1)$	A	18.0	18.2	18.4	34.9	35.1	35.2	17.1	32.6
$(-2.0, 0, 4.0)$	$(1, 1, 1)$	B	23.3	23.4	23.4	46.7	46.8	46.8	21.1	43.7
$(-1.0, 0, 3.0)$	$(0.6, 1.2, 0.6)$	A	44.4	44.7	44.9	79.1	79.3	79.4	38.8	74.8
$(-1.0, 0, 3.0)$	$(0.6, 1.2, 0.6)$	B	41.2	41.6	42.1	74.5	74.7	74.8	35.9	70.1

In the α_j columns, A refers to $(\alpha_1, \alpha_2, \alpha_3) = (1/3, 1/3, 1/3)$ and B refers to $(\alpha_1, \alpha_2, \alpha_3) = (0.4, 0.2, 0.4)$.

Table 3: Powers (in %) of the Modified EM test and EM test for testing $H_0 : m = 3$ at the 5% level

		Modified EM test						EM test of CLF	
Alternative models: $\alpha_j = 1/4$		$n = 200$			$n = 400$			$n = 200$	$n = 400$
$(\mu_1, \mu_2, \mu_3, \mu_4)$	σ_j	$K = 1$	$K = 2$	$K = 3$	$K = 1$	$K = 2$	$K = 3$	$K = 3$	$K = 3$
$(-4.5, -1.5, 1.5, 4.5)$	A	22.2	23.3	23.8	61.8	62.1	62.3	20.7	54.9
$(-6, -2, 2, 6)$	A	94.4	94.9	95.1	99.7	99.7	99.8	93.8	100.0
$(-4.5, -1.5, 1.5, 4.5)$	B	85.3	86.1	86.3	99.4	99.5	99.6	83.0	99.7
$(-6, -2, 2, 6)$	B	99.7	99.8	99.8	100.0	100.0	100.0	100.0	100.0
$(-4.0, -1.25, 1.25, 4.0)$	B	57.3	58.9	59.2	92.3	92.6	92.7	53.1	91.7
$(-3.5, -1.5, 1.5, 5.0)$	B	83.5	83.8	83.9	99.7	99.7	99.7	76.6	99.1
$(-4.0, -1.25, 1.25, 4.0)$	C	11.5	12.4	12.8	36.7	37.0	37.2	12.8	32.3
$(-3.5, -1.5, 1.5, 5.0)$	C	56.1	57.1	57.5	96.2	96.3	96.3	47.2	91.5
$(-4.0, -1.25, 1.25, 4.0)$	D	23.3	24.4	24.7	60.3	60.8	61.0	22.8	56.3
$(-3.5, -1.5, 1.5, 5.0)$	D	31.9	32.7	33.0	67.0	67.3	67.4	27.4	59.0

In the σ_j columns, A refers to $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (1, 1, 1, 1)$, B refers to $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (0.6, 1.2, 0.6, 1.2)$, C refers to $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (0.6, 0.8, 1.0, 1.2)$, and D refers to $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (0.8, 0.8, 0.8, 1.2)$.

Table 4: Type I errors (in %) of the modified EM test for normal regression mixture models when testing $H_0 : m = 1$

	1 % level				5 % level				the choice
	$n = 100$		$n = 200$		$n = 100$		$n = 200$		of a_n
K	2	3	2	3	2	3	2	3	
dim($\boldsymbol{X}$)=1	0.9	1.0	1.0	1.0	4.5	4.7	5.0	5.0	2.2
dim($\boldsymbol{X}$)=2	0.9	0.9	1.1	1.2	4.7	4.9	5.0	5.1	3.1
dim($\boldsymbol{X}$)=3	0.9	1.0	0.9	0.9	4.9	5.2	5.0	5.0	5.4
dim($\boldsymbol{X}$)=4	1.0	1.1	0.9	1.0	4.5	4.8	5.0	5.1	8.3

Samples are simulated from $Y|\boldsymbol{X} \sim N(\mu + \boldsymbol{X}^\top \boldsymbol{\beta}, 1)$, where $\mu = 0$, $\boldsymbol{\beta} = (0.5, \ldots, 0.5)^\top$, and $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_q)$.

Table 5: Parameter values of normal regression mixture models for power assessment when testing $H_0 : m = 1$

Model	α	μ_1	μ_2	β_1	β_2	σ_1	σ_2
I	0.5	-1.15	1.15	0.5	0.5	1.0	1.0
II	0.25	-1.15	1.15	0.5	0.5	1.0	1.0
III	0.5	-0.75	0.75	0.5	0.5	1.2	0.8
IV	0.25	-0.75	0.75	0.5	0.5	1.2	0.8
V	0.5	-1.15	1.15	-0.25	0.25	1.0	1.0
VI	0.25	-1.15	1.15	-0.25	0.25	1.0	1.0
VII	0.5	-0.75	0.75	-0.25	0.25	1.2	0.8
VIII	0.25	-0.75	0.75	-0.25	0.25	1.2	0.8

We set $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_1)^\top$ and $\boldsymbol{\beta}_2 = (\beta_2, \ldots, \beta_2)^\top$.

Table 6: Powers (in %) of the modified EM test for normal regression mixture models when testing $H_0 : m = 1$ at the 5% level

	dim($\boldsymbol{X}$)=1				dim($\boldsymbol{X}$)=2			
	$n = 100$		$n = 200$		$n = 100$		$n = 200$	
Model	$K = 2$	$K = 3$	$K = 2$	$K = 3$	$K = 2$	$K = 3$	$K = 2$	$K = 3$
I	26.3	26.5	59.4	59.6	16.3	16.5	47.0	47.1
II	34.8	35.2	71.9	72.0	25.6	26.1	61.5	61.7
III	35.3	35.7	73.0	73.1	25.9	26.4	63.5	63.6
IV	55.9	56.6	91.1	91.2	46.5	47.9	86.5	86.7
V	59.1	59.4	93.4	93.4	70.4	70.8	97.9	97.9
VI	56.6	57.2	90.9	91.0	60.8	61.8	94.7	94.8
VII	57.6	57.9	91.9	92.0	63.6	64.1	95.2	95.3
VIII	71.2	71.8	96.7	96.7	71.9	73.1	97.5	97.5

	dim($\boldsymbol{X}$)=3				dim($\boldsymbol{X}$)=4			
	$n = 100$		$n = 200$		$n = 100$		$n = 200$	
Model	$K = 2$	$K = 3$	$K = 2$	$K = 3$	$K = 2$	$K = 3$	$K = 2$	$K = 3$
I	7.6	7.7	30.2	30.4	4.5	4.8	13.9	14.1
II	17.2	17.9	51.8	52.1	10.2	11.1	38.3	38.9
III	16.8	17.6	53.0	53.3	10.5	11.5	40.1	40.7
IV	35.8	37.6	81.1	81.4	23.8	26.5	72.5	73.5
V	73.6	74.0	98.9	99.0	73.5	74.1	99.6	99.7
VI	64.2	65.6	96.5	96.6	62.6	64.6	96.9	97.1
VII	66.3	67.1	97.0	97.0	64.1	65.0	97.9	97.9
VIII	73.1	74.5	98.2	98.3	69.9	72.3	98.4	98.5

Samples are simulated from $Y|\boldsymbol{X} \sim \alpha N(\mu_1 + \boldsymbol{X}^\top \boldsymbol{\beta}_1, \sigma_1) + (1 - \alpha) N(\mu_2 + \boldsymbol{X}^\top \boldsymbol{\beta}_2, \sigma_2)$, where $\boldsymbol{X} = (X_1, \ldots, X_q)^\top$ with $X_j \sim_{iid} N(0, 1)$.

Table 7: Parameter specifications for null models with order 2 and order 3, $\dim(\boldsymbol{X}) = 1$

Order 2		Order 3	
$(\alpha_1, \alpha_2) =$	$(0.5, 0.5)$, $(0.2, 0.8)$	$(\alpha_1, \alpha_2, \alpha_3) =$	$(1/3, 1/3, 1/3)$, $(0.25, 0.5, 0.25)$
$(\mu_1, \mu_2) =$	$(-1.25, 1.25)$, $(-1.75, 1.75)$	$(\mu_1, \mu_2, \mu_3) =$	$(-3.5, 0, 4.5)$, $(-4.5, 0, 4.5)$
$(\beta_1, \beta_2) =$	$(0, 0)$, $(-1, 1)$	$(\beta_1, \beta_2, \beta_3) =$	$(0, 0, 0)$, $(-1, 0, 1)$
$(\sigma_1, \sigma_2) =$	$(1, 1)$, $(1.2, 0.6)$	$(\sigma_1, \sigma_2, \sigma_3) =$	$(1, 1, 1)$, $(1.2, 0.9, 0.6)$

Table 8: Power (in %) of the modified EM test for normal regression mixture models when testing $H_0 : m = 2$ at the 5% level

	Bootstrap				Asymptotic			
	$n = 200$		$n = 400$		$n = 200$		$n = 400$	
(μ_j, α_j)	$K = 2$	$K = 3$	$K = 2$	$K = 3$	$K = 2$	$K = 3$	$K = 2$	$K = 3$
(A, C)	35.5	35.5	86.9	86.7	35.0	35.9	86.3	86.4
(A, D)	40.5	40.1	83.2	82.8	40.0	40.4	83.9	84.0
(B, C)	34.1	33.8	73.8	73.6	35.0	35.4	75.0	75.2
(B, D)	37.3	37.4	73.7	73.6	37.4	38.2	74.9	75.1

Samples are simulated from $Y|X \sim \sum_{j=1}^{3} \alpha_j N(\mu_j + \beta_j X, \sigma_j)$, where $X \sim_{iid} N(0, 1)$. In the column of (μ_j, α_j), A and B refer to $(\mu_1, \mu_2, \mu_3) = (-2.5, 0, 2.5)$ and $(-2.0, 0, 4.0)$, respectively; C and D refer to $(\alpha_1, \alpha_2, \alpha_3) = (1/3, 1/3, 1/3)$ and $(0.4, 0.2, 0.4)$, respectively. We set $(\beta_1, \beta_2, \beta_3) = (-0.5, 0, 0.5)$ and $(\sigma_1, \sigma_2, \sigma_3) = (1, 1, 1)$.

Table 9: Power (in %) of the modified EM test for normal regression mixture models when testing $H_0 : m = 3$ at the 5% level

	Bootstrap				Asymptotic			
	$n = 200$		$n = 400$		$n = 200$		$n = 400$	
$(\mu_j, \beta_j, \sigma_j)$	$K = 2$	$K = 3$	$K = 2$	$K = 3$	$K = 2$	$K = 3$	$K = 2$	$K = 3$
(A, C, E)	15.8	16.4	71.4	71.8	25.4	26.6	74.0	75.8
(A, D, E)	56.8	56.8	98.8	98.6	63.2	65.4	98.8	99.0
(B, C, F)	55.4	55.0	96.8	96.8	63.0	64.0	97.6	97.6
(B, D, F)	84.0	84.4	99.8	99.8	87.8	88.8	99.8	99.8

Samples are simulated from $Y|X \sim \sum_{j=1}^{4} \alpha_j N(\mu_j + \beta_j X, \sigma_j)$, where $X \sim_{iid} N(0, 1)$. In the column of $(\mu_j, \beta_j, \sigma_j)$, A and B refer to $(\mu_1, \mu_2, \mu_3, \mu_4) = (-4.5, -1.5, 1.5, 4.5)$ and $(-4.0, -1.25, 1.25, 4.0)$, respectively; C and D refer to $(\beta_1, \beta_2, \beta_3, \beta_4) = (-0.75, -0.25, 0.25, -0.75)$ and $(-1.5, -0.5, 0.5, 1.5)$, respectively; E and F refer to $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (1, 1, 1, 1)$ and $(0.6, 1.2, 0.6, 1.2)$. We set $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.25$.

Table 10: Frequency of the selected number of components for 30 stocks in the Dow Jones Industrial Average

	Kon (1984)	Modified EM	AIC	BIC
$m = 2$	12	2	0	5
$m = 3$	11	15	2	19
$m \geq 4$	7	13	28	6

Table 11: Estimation results for the leukemia data of Golub et al. (1999)

	$m = 1$	$m = 2$	$m = 3$
α_1	1	0.644	0.044
α_2		0.356	0.688
α_3			0.268
μ_1	0.905	0.354	-0.835
μ_2		1.902	0.547
μ_3			2.107
σ_1	1.472	1.124	0.435
σ_2		1.503	1.165
σ_3			1.519
p-value	0.0%	0.0%	76.4%
AIC	25752	25535	25519
BIC	25766	25569	25574

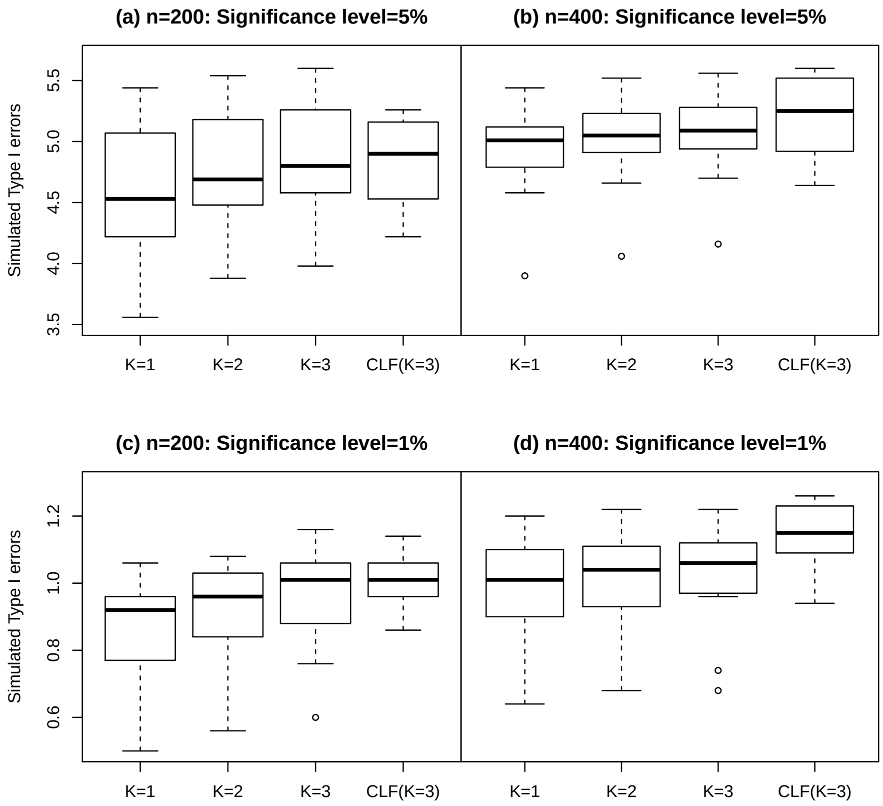

Figure 1: Type I errors of the modified EM test and EM test for testing $H_0 : m = 2$

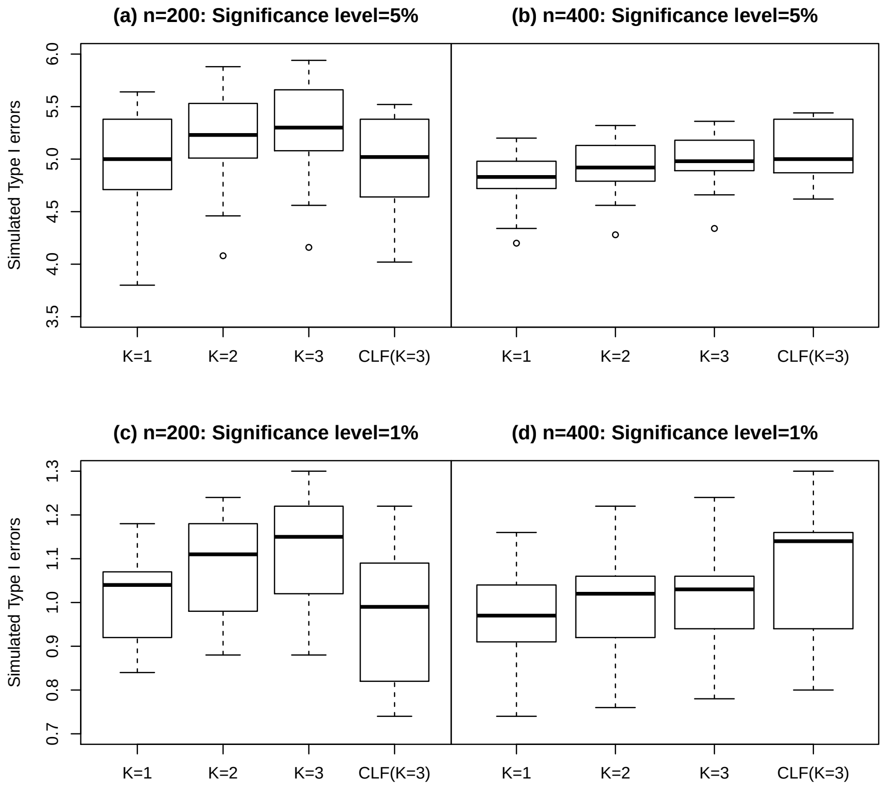

Figure 2: Type I errors of the modified EM test and EM test for testing $H_0 : m = 3$

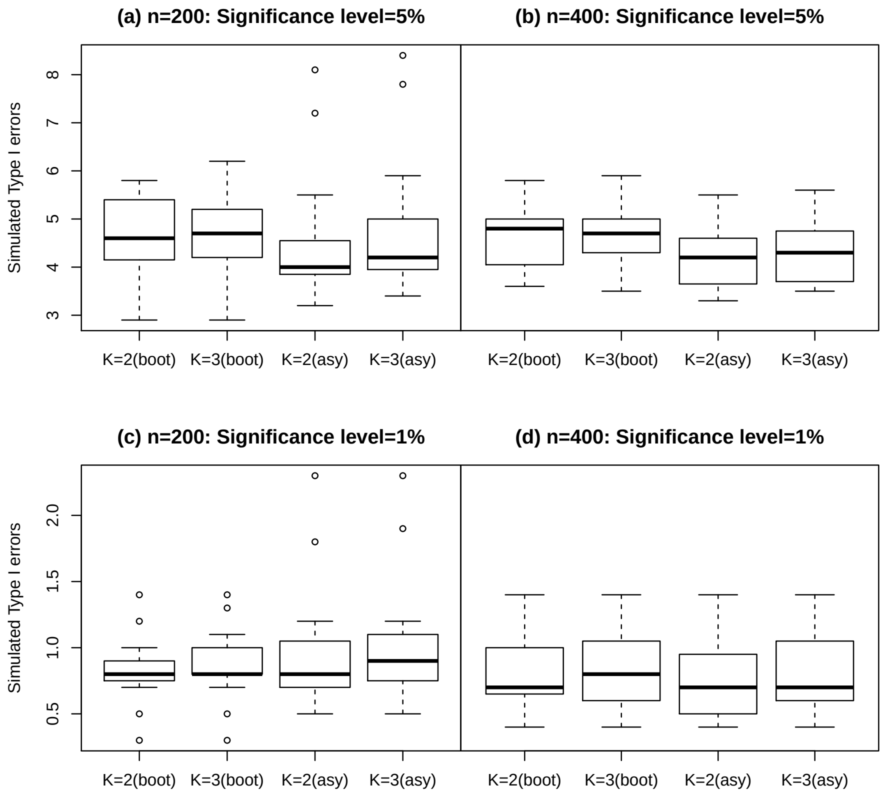

Figure 3: Type I errors of the modified EM test for testing $H_0 : m = 2$ (with regressor)

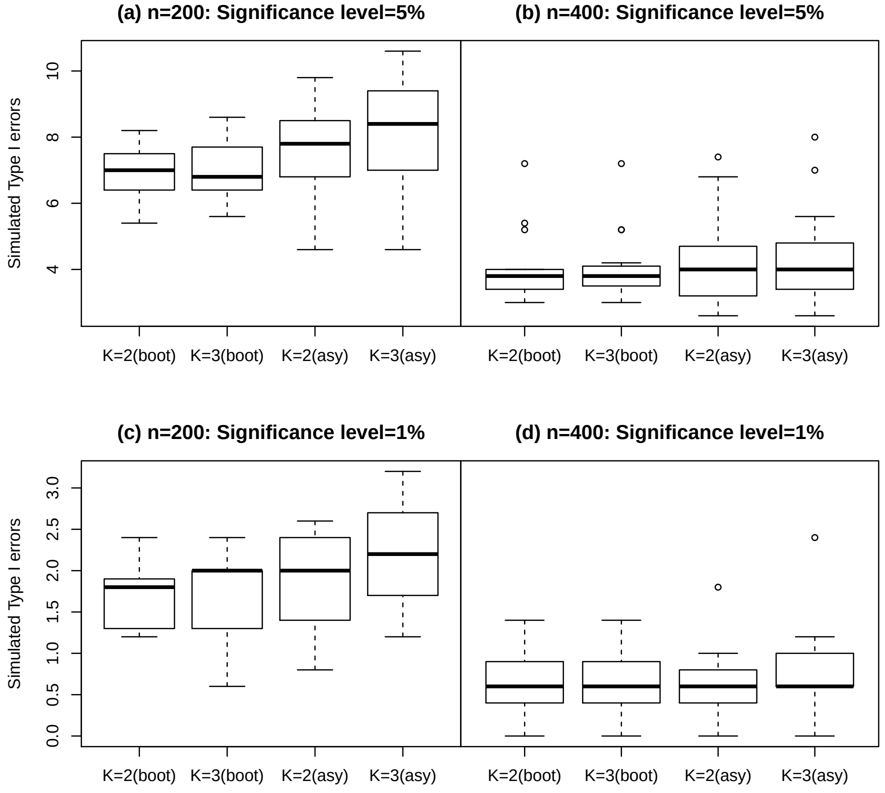

Figure 4: Type I errors of the modified EM test for testing $H_0 : m = 3$ (with regressor)

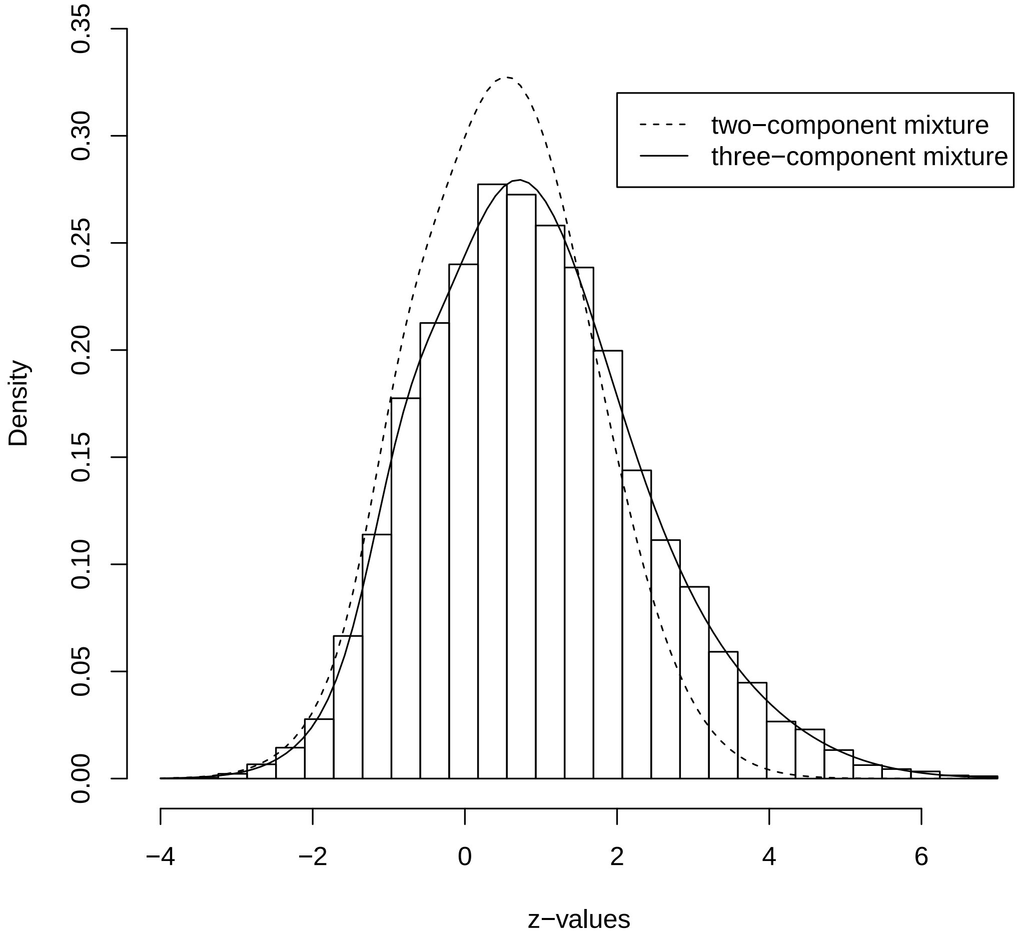

Figure 5: Leukemia data: plot of fitted two- and three-component normal mixture models imposed on a histogram of $\boldsymbol{z}$-values