Testing the Number of Components in Finite Mixture Models[*]

Hiroyuki Kasahara
Department 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 2012

Abstract

This paper considers likelihood-based testing of the null hypothesis of m_0 components against the alternative of $m_0 + 1$ components in a finite mixture model. The number of components is an important parameter in the applications of finite mixture models. Still, testing the number of components has been a long-standing challenging problem because of its non-regularity.

We develop a framework that facilitates the analysis of the likelihood function of finite mixture models and derive the asymptotic distribution of the likelihood ratio test statistic for testing the null hypothesis of m_0 components against the alternative of $m_0 + 1$ components. Furthermore, building on this framework, we propose a likelihood-based testing procedure of the number of components. The proposed test, extending the EM approach of Li et al. (2009), does not use a penalty term and is implementable even when the likelihood ratio test is difficult to implement because of non-regularity and computational complexity.

Keywords and phrases: asymptotic distribution; modified EM test; likelihood ratio test; local MLE; number of components; reparameterization.

1 Introduction

Finite mixture models provide flexible ways to account for unobserved population heterogeneity. Because of their flexibility, finite mixture models have seen numerous applications in diverse fields such as biological, physical, and social sciences. For example, finite mixtures are often used to control unobserved individual-specific effects in labor economics (Heckman and Singer, 1984; Keane

[*] The authors thank Jiahua Chen, Tatsuya Kubokawa, Tong Li and the seminar participants at Vanderbilt University and the University of British Columbia for their helpful comments. This research was supported by the SSHRC and the JSPS Grant-in-Aid for Scientific Research (C) No. 23530249.

and Wolpin, 1997; Cameron and Heckman, 1998), health economics (Deb and Trivedi, 2002), and marketing (Kamakura and Russell, 1989; Andrews and Currim, 2003). 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).

This paper considers likelihood-based testing of the null hypothesis of m_0 components against the alternative of $m_0 + 1$ components in a finite mixture model. The number of components is an important parameter in finite mixture models. In economics applications, the number of components often represents the number of unobservable types or abilities. In many other applications, the number of components signifies the number of clusters or latent classes in the data.

Testing the number of components in finite mixture models has been a long-standing challenging problem because of its non-regularity. When testing the null of m_0 components against the alternative of $m_0 + 1$ components, the true m_0-component density can be described with many elements of the parameter space in the $(m_0 + 1)$-component alternative model. These elements are characterized by the union of the two parameter subsets: A, where two components have the same mixing parameter that takes component-specific values; and B, where one of the components has zero mixing proportion. In both null parameter sets, the regularity conditions for a standard asymptotic analysis fail because of such problems as parameter non-identification, singular Fisher information matrix, and true parameter being on the parameter space boundary. When the parameter space is compact, the asymptotic distribution of the likelihood ratio test (LRT) statistic has been derived as a supremum of the square of a Gaussian process indexed by the closure of the convex cone of directional score functions (Dacunha-Castelle and Gassiat, 1999; Liu and Shao, 2003); however, it is difficult to implement these symbolic results.[1]

This paper makes three main contributions. First, we develop a framework that facilitates the analysis of the likelihood function of finite mixture models. In the null parameter space A discussed above, the standard quadratic expansion of the log-likelihood function is not applicable because of the singular Fisher information matrix. The existing works handle this problem by resorting to nonstandard approaches and tedious manipulations (see, for example, Zhu and Zhang (2004); Cho and White (2007)). We develop an orthogonal parameterization that extracts the direction in which the Fisher information matrix is singular. Under this reparameterization, the log-likelihood function is locally approximated by a quadratic form of squares and cross-products of the reparameterized parameters, leading to a simple characterization of the asymptotic distribution of the LRT statistic.

Second, building on this framework and the results from Andrews (1999, 2001), we derive the asymptotic distribution of the LRT statistic for testing the null hypothesis of m_0 components for a general $m_0 \geq 1$ in a mixture model with a multidimensional mixing parameter and a structural parameter. Under the null parameter set A, the asymptotic distribution is shown to be the maximum of m_0 random variables, each of which is a projection of a Gaussian random vector on a cone. Both

[1] For some specific models, the asymptotic distribution of the LRT statistic has been derived. See, for example, Chernoff and Lander (1995); Lemdani and Pons (1997); Chen and Chen (2003); Garel (2001).

the LRT statistic under the null parameter set B and the (unrestricted) LRT statistic are shown to converge in distribution to the maximum of m_0 random variables, each of which is the supremum of the square of a Gaussian process over the support of the mixing parameter. In contrast to the existing symbolic results, the covariance structure of the Gaussian processes is explicitly presented.

Implementing the LRT has, however, practical difficulties: (i) in some mixture models that are popular in applications (e.g., Weibull duration models), the Fisher information for the null parameter space B is not finite (Li et al., 2009), and the regularity conditions in Dacunha-Castelle and Gassiat (1999) and Liu and Shao (2003) are violated; (ii) the asymptotic distribution depends on the choice of the support of the parameter space, and (iii) simulating the supremum of a Gaussian process is computationally challenging unless the dimension of the parameter space is small, because of the curse of dimensionality.

As our third contribution, we develop a likelihood-based testing procedure of the null hypothesis of m_0 components against the alternative of $m_0 + 1$ components that circumvents these difficulties associated with the null parameter space B. The proposed modified EM test statistic has the same asymptotic distribution as the LRT statistic for testing the null parameter space A, and its asymptotic distribution can be simulated without facing the curse of dimensionality. Furthermore, the modified EM test is implementable even if the Fisher information for the null parameter space B is not finite.

Our modified EM test extends the EM approach pioneered by Li et al. (2009) (henceforth, LCM). In contrast to the original EM approach by LCM, the modified EM test does not use a penalty term to bound the mixing proportion away from 0 and 1. This feature is practically appealing because the choice of the penalty term has an important influence on the finite sample performance of the EM test. Even though a data-driven formula for the penalty term was obtained via simulations for Poisson, binomial, normal, and exponential distributions by Chen and Li (2011b), developing a formula for general mixture models is challenging. Simulations show that the modified EM test has good finite sample size, and power properties comparable to those of the original EM test.

There exist many works on likelihood-based tests of the number of components that either focus on testing homogeneity (i.e., $m_0 = 1$) or assume a scalar mixing parameter, but these existing tests are not applicable to testing the null hypothesis of $m_0 \geq 2$ components in a general mixture model with a vector mixing parameter and a structural parameter. Assuming a scalar mixing parameter, Chen et al. (2001, 2004) developed a modified LRT for the null hypothesis of $m_0 = 1$ and $m_0 = 2$; LCM developed the EM test for testing $m_0 = 1$; Chen and Chen (2001) and Cho and White (2007) derived the asymptotic distribution of the LRT statistic and quasi-LRT statistic for testing $m_0 = 1$, respectively; Li and Chen (2010) and Chen and Li (2011a) constructed EM tests for testing $m_0 \geq 2$. For models with a multidimensional mixing parameter, Zhu and Zhang (2004) analyzed the asymptotics of LRT and Niu et al. (2011) focused on an EM test for testing $m_0 = 1$. Except for Zhu and Zhang (2004) and Cho and White (2007), none of the works discussed above accommodates a structural parameter.

The remainder of this paper is organized as follows. Section 2 introduces finite mixture models and describes examples. Section 3 derives the asymptotic distribution of the LRT statistic of homogeneity as a precursor for the test of general m_0 components. Section 4 establishes the asymptotic distribution of the LRT statistic of the null hypothesis of m_0 components against the alternative of $m_0 + 1$ components. Section 5 introduces the modified EM test and determines its asymptotic distribution. Section 6 reports the simulation results. The appendix contains proofs of results given in the paper, and auxiliary results. All limits below are taken as $n \to \infty$ unless stated otherwise. Let $:=$ denote "equals by definition." For a $k \times 1$ vector a and a function $f(a)$, let $\nabla_a f(a)$ denote the $k \times 1$ vector of the derivative $(\partial/\partial a)f(a)$, and let $\nabla_{aa'} f(a)$ denote the $k \times k$ vector of the derivative $(\partial/\partial a \partial a')f(a)$.

2 Finite mixture models and examples

2.1 Finite mixture models

Given a family of parametric densities $\{f(x;\gamma,\theta) : \gamma \in \Theta_\gamma \subset \mathbb{R}^p, \theta \in \Theta_\theta \subset \mathbb{R}^q\}$ for a random variable $X \in \mathbb{R}^r$, we consider an m-component finite mixture density of the form

$$\sum_{j=1}^{m} \alpha^j f(x;\gamma,\theta^j), \tag{1}$$

where the α^j's are mixing probabilities that satisfy $\alpha^j \in [0,1]$ and $\sum_{j=1}^m \alpha^j = 1$, $\theta^j = (\theta_1^j, \ldots, \theta_q^j)'$ is a mixing parameter that characterizes the j-th component, and $\gamma = (\gamma_1, \ldots, \gamma_p)'$ is a structural parameter that is common to all the components. Here, m is understood as the smallest number such that the data density admits representation (1). In specification (1), each observation may be viewed as a sample from one of the m latent classes, or "types." This specification includes a finite mixture of conditional distributions, in which a component distribution is given by $f(x_1, x_2; \gamma, \theta^j) = f(x_1|x_2;\gamma,\theta^j)f(x_2)$.

We are interested in testing the number of components in a finite mixture model, specifically, in testing

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

2.2 Examples

Example 1 (Duration model with mixed proportional hazard)**.** *Heckman and Singer (1984) proposed a discrete mixture distribution for estimating parametric duration models with unobserved heterogeneity. Consider a finite mixture proportional hazard model of duration $Y \in \mathbb{R}_+$ conditional on observed heterogeneity $X \in \mathbb{R}$, where the hazard rate of the jth component distribution is specified as $\frac{f(y|x;\theta^j)}{1-F(y|x;\theta^j)} = \exp(\theta_1^j)k_1(x;\theta_2^j)k_2(y;\theta_3^j)$, where $\theta^j = (\theta_1^j, (\theta_2^j)', (\theta_3^j)')'$, $k_1(x;\theta_2)$ captures the part of the hazard that is systematically related to observed variable x, and $k_2(y;\theta_3)$ is* *the baseline hazard. Then, the model is written as* $\sum_{j=1}^{m} \alpha^m f(y|x;\theta^j)f(x)$*, where* $f(y|x;\theta^j) = \exp(\theta_1^j)k_1(x;\theta_2^j)k_2(y;\theta_3^j)\exp[-\exp(\theta_1^j)k_1(x;\theta_2^j)\int_0^y k_2(s;\theta_3^j)ds]$ *is the conditional density of* y *given* x *implied by the hazard* $\exp(\theta_1^j)k_1(x;\theta_2^j)k_2(y;\theta_3^j)$.

3 Likelihood ratio test of $H_0 : m = 1$ against $H_A : m = 2$

Before developing the LRT of m_0 components, we analyze a simpler case of 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 $X_1, \ldots, X_n$ from the true density $f(x;\gamma^*,\theta^*)$. Here, the superscript $*$ denotes the true population value. Consider a two-component mixture density function

$$f(x;\alpha,\gamma,\theta^1,\theta^2) := \alpha f(x;\gamma,\theta^1) + (1-\alpha)f(x;\gamma,\theta^2), \tag{2}$$

where $(\alpha,\gamma,\theta^1,\theta^2) \in \Theta := [0,1] \times \Theta_\gamma \times \Theta_\theta^2$. The two-component model (2) gives rise to the true density $f(x;\gamma^*,\theta^*)$ if the parameter $(\alpha,\gamma,\theta^1,\theta^2)$ lies in a subset of the parameter space

$$\Gamma^* := \left\{(\alpha,\gamma,\theta^1,\theta^2) \in \Theta : \ \theta^1 = \theta^2 = \theta^* \text{ and } \gamma = \gamma^*; \text{ or } \alpha(1-\alpha) = 0 \text{ and } \gamma = \gamma^*\right\}.$$

Let $(\hat{\alpha},\hat{\gamma},\hat{\theta}^1,\hat{\theta}^2)$ denote the maximum likelihood estimator (MLE) that maximizes the log-likelihood function $\sum_{i=1}^n \ln f(X_i;\alpha,\gamma,\theta^1,\theta^2)$. The following proposition shows that the MLE is consistent under the standard condition.[2]

Assumption 1. *(a) If* $(\gamma,\theta) \neq (\gamma^*,\theta^*)$*, then* $f(X;\gamma,\theta) \neq f(X;\gamma^*,\theta^*)$ *with a nonzero probability. (b)* Θ_θ *and* Θ_γ *are compact. (c)* $\ln f(X;\gamma,\theta)$ *is continuous at each* $(\gamma,\theta) \in \Theta_\gamma \times \Theta_\theta$ *with probability one.*
(d) $E[\sup_{(\gamma,\theta)\in\Theta_\gamma\times\Theta_\theta} |\ln f(X_i;\gamma,\theta)|] < \infty$.

Proposition 1. *Suppose that Assumption 1 holds. Then,*
we have $\inf_{(\alpha,\gamma,\theta^1,\theta^2)\in\Gamma^*} ||(\hat{\alpha},\hat{\gamma},\hat{\theta}^1,\hat{\theta}^2) - (\alpha,\gamma,\theta^1,\theta^2)|| \to_p 0$.

As in Cho and White (2007), we partition the null hypothesis $H_0 : m = 1$ into two sub-hypotheses:

$$H_{01} : \theta^1 = \theta^2 \quad \text{and} \quad H_{02} : \alpha(1-\alpha) = 0.$$

Under H_{01}, α is not identified, and furthermore, the Fisher information matrix for the other parameters becomes singular. Under H_{02}, α is on the boundary of the parameter space, and either θ^1 or θ^2 is not identified. In the following, we analyze the asymptotic distribution of the LRT statistics for testing H_{01} and H_{02} in turn, and combining these results, derive the asymptotic distribution of the LRT statistics for testing H_0.[3]

[2] Alternatively, we can use the sufficient condition in Redner (1981).

[3] This approach is used by Cho and White (2007) to analyze the quasi-LRT of $H_0 : m = 1$ in a model with a

3.1 Reparameterization

We first develop a reparameterization that substantially simplifies the analysis of the LRT statistic for testing H_{01}. One difficult problem in the analysis of finite mixture models is that the Fisher information matrix is singular. Under the true parameter value $(\gamma^*, \theta^*, \theta^*)$ at $\alpha \in (0,1)$, the first derivative of the log-density

$$l(x; \alpha, \gamma, \theta^1, \theta^2) := \ln \big(\alpha f(x; \gamma, \theta^1) + (1-\alpha) f(x; \gamma, \theta^2)\big)$$

with respect to (w.r.t.) θ^1 is a linear function of the first derivative of $l(x; \alpha, \gamma, \theta^1, \theta^2)$ w.r.t. θ^2:

$$\begin{aligned} &\nabla_{\theta^1} l(x; \alpha, \gamma^*, \theta^*, \theta^*) = \alpha \nabla_\theta f(x; \gamma^*, \theta^*)/f(x; \gamma^*, \theta^*) \quad \text{and} \\ &\nabla_{\theta^2} l(x; \alpha, \gamma^*, \theta^*, \theta^*) = (1-\alpha) \nabla_\theta f(x; \gamma^*, \theta^*)/f(x; \gamma^*, \theta^*). \end{aligned}$$

In addition, the first derivative of $l(x; \alpha, \gamma, \theta^1, \theta^2)$ w.r.t. α evaluated at $(\gamma^*, \theta^*, \theta^*)$ is identically equal to zero. Consequently, the Fisher information matrix is rank deficient by 1+dim(θ), and the log-likelihood function is not amenable to the standard analysis using a second-order Taylor expansion.

We handle the singular Fisher information problem via a reparameterization. Our approach generalizes that of Rotnitzky et al. (2000), who derive the asymptotics of the LRT statistic when the Fisher information matrix is rank deficient by 1. A key insight is that by a particular reparameterization, we can determine the direction in which the Fisher information matrix is singular. Consider the following one-to-one reparameterization:

$$\begin{pmatrix} \lambda \\ \nu \end{pmatrix} := \begin{pmatrix} \theta^1 - \theta^2 \\ \alpha\theta^1 + (1-\alpha)\theta^2 \end{pmatrix}, \quad \text{so that} \quad \begin{pmatrix} \theta^1 \\ \theta^2 \end{pmatrix} = \begin{pmatrix} \nu + (1-\alpha)\lambda \\ \nu - \alpha\lambda \end{pmatrix}, \tag{3}$$

where $\nu = (\nu_1, \nu_2, \dots, \nu_q)'$ and $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_q)'$ are $q \times 1$ reparameterized parameter vectors. Collect the reparameterized parameters except for α into one vector as

$$\psi := (\gamma', \nu', \lambda')' \in \Theta_\psi,$$

where $\Theta_\psi := \{\psi : \gamma \in \Theta_\gamma,\ \nu + (1-\alpha)\lambda \in \Theta_\theta \text{ and } \nu - \alpha\lambda \in \Theta_\theta\}$. The parameter ψ and the parameter space Θ_ψ depend on α, although we do not explicitly indicate their dependence for notational brevity. In the reparameterized model, the null hypothesis of $H_{01} : \theta^1 = \theta^2$ is written as $H_{01} : \lambda = (0, \dots, 0)'$. We denote the true value of ψ by $\psi^* = ((\gamma^*)', (\theta^*)', 0, \dots, 0)'$.

Under the reparameterization (3), the density and its logarithm are expressed as

$$f(x; \psi, \alpha) := \alpha f(x; \gamma, \nu + (1-\alpha)\lambda) + (1-\alpha) f(x; \gamma, \nu - \alpha\lambda), \quad l(x; \psi, \alpha) := \ln[f(x; \psi, \alpha)]. \tag{4}$$

scalar mixing parameter. Their asymptotic analysis is very complex, however, and can only handle the case with a scalar mixing parameter. Cho and White (2007) do not analyze testing $H_0 : m = m_0$ for $m_0 \geq 2$, either.

Evaluated at the true parameter value, the first derivative of the reparameterized log-density (4) w.r.t. λ becomes zero:

$$\nabla_{\lambda} l(x;\psi^*,\alpha) = [(1-\alpha)\alpha\nabla_{\theta} f(x;\gamma^*,\theta^*) - \alpha(1-\alpha)\nabla_{\theta} f(x;\gamma^*,\theta^*)]/f(x;\gamma^*,\theta^*) = 0. \tag{5}$$

On the other hand, the first derivative of (4) w.r.t. γ and ν under the true parameter value is a mean-zero non-degenerate random vector:

$$\begin{aligned}\nabla_{\gamma} l(x;\psi^*,\alpha) &= \nabla_{\gamma} f(x;\gamma^*,\theta^*)/f(x;\gamma^*,\theta^*),\\ \nabla_{\nu} l(x;\psi^*,\alpha) &= \nabla_{\theta} f(x;\gamma^*,\theta^*)/f(x;\gamma^*,\theta^*).\end{aligned} \tag{6}$$

Because $\nabla_{\lambda} l(x;\psi^*,\alpha) = 0$, the information matrix for the reparameterized model is singular, and the standard quadratic approximation of the log-likelihood function fails. Nonetheless, we may characterize the asymptotic distribution of the LRT statistic using the second derivative of $l(x;\psi,\alpha)$ w.r.t. λ in place of its score:

$$\nabla_{\lambda\lambda'} l(x;\psi^*,\alpha) = \alpha(1-\alpha)\frac{\nabla_{\theta\theta'} f(x;\gamma^*,\theta^*)}{f(x;\gamma^*,\theta^*)}. \tag{7}$$

When $\alpha \neq \{0,1\}$ and $\nabla_{\theta\theta'} f(x;\gamma^*,\theta^*)/f(x;\gamma^*,\theta^*) \neq 0$ with positive probability, the elements of $\nabla_{\lambda\lambda'} l(x;\psi^*,\alpha)$ are mean-zero random variables and serve as the scores.

3.2 Approximation of the log-likelihood function in quadratic form

In this section, we analyze the asymptotic behavior of the log-likelihood function. Let $L_n(\psi,\alpha) := \sum_{i=1}^n l(X_i;\psi,\alpha)$ denote the reparameterized log-likelihood function. Define $\eta := (\gamma',\nu')'$ and $\eta^* := ((\gamma^*)',(\theta^*)')'$ so that $\psi = (\eta',\lambda')'$ and $\psi^* = ((\eta^*)',0,\dots,0)'$. Fix the value of $\alpha \in (0,1)$. Then, $L_n(\psi,\alpha)$ has a quartic expansion around (ψ^*,α) as

$$\begin{aligned} L_n(\psi,\alpha) - L_n(\psi^*,\alpha) &= \nabla_{\eta} L_n(\psi^*,\alpha)(\eta-\eta^*) + \frac{1}{2!}(\eta-\eta^*)'\nabla_{\eta\eta'} L_n(\psi^*,\alpha)(\eta-\eta^*)\\ &+ \frac{1}{2!}\sum_{i=1}^q\sum_{j=1}^q \nabla_{\lambda_i\lambda_j} L_n(\psi^*,\alpha)\lambda_i\lambda_j + \frac{3}{3!}\sum_{i=1}^q\sum_{j=1}^q (\eta-\eta^*)'\nabla_{\eta\lambda_i\lambda_j} L_n(\psi^*,\alpha)\lambda_i\lambda_j\\ &+ \frac{1}{4!}\sum_{i=1}^q\sum_{j=1}^q\sum_{k=1}^q\sum_{l=1}^q \nabla_{\lambda_i\lambda_j\lambda_k\lambda_l} L_n(\psi^*,\alpha)\lambda_i\lambda_j\lambda_k\lambda_l + R_n(\psi,\alpha), \end{aligned} \tag{8}$$

where $R_n(\psi,\alpha)$ is a remainder term whose stochastic order is established later.

We introduce some notations to simplify (8). Define $q_\lambda := q(q+1)/2$. For $\lambda \in \mathbb{R}^q$, collect the elements of $\text{vech}(\lambda\lambda')$ into a $q_\lambda \times 1$ vector:

$$\begin{aligned} v(\lambda) &= (v_{11}, \dots, v_{qq}, v_{12}, \dots, v_{1q}, v_{23}, \dots, v_{2q}, \dots, v_{q-1,q})' \\ &:= (\lambda_1^2, \dots, \lambda_q^2, \lambda_1\lambda_2, \dots, \lambda_1\lambda_q, \lambda_2\lambda_3, \dots, \lambda_2\lambda_q, \dots, \lambda_{q-1}\lambda_q)'. \end{aligned}$$

Note that the elements of $v(\lambda)$ must satisfy the restriction $v_{ii} \geq 0$ and $v_{ij}v_{kl} = v_{ik}v_{jl}$ for all $i \leq j \leq k \leq l$. We rewrite the right hand side of (8) as a quadratic function of η and $v(\lambda)$. Combine η and $v(\lambda)$ into a $(p + q + q_\lambda) \times 1$ vector:

$$\zeta := (\eta', v(\lambda)')'.$$

Let $\widetilde{\nabla}_\zeta l(X_i; \psi, \alpha)$ be a $(p + q + q_\lambda) \times 1$ vector defined by

$$\widetilde{\nabla}_\zeta l(X_i; \psi, \alpha) := (\nabla_{\eta'} l(X_i; \psi, \alpha), \widetilde{\nabla}_{v(\lambda)'} l(X_i; \psi, \alpha)/(\alpha(1 - \alpha)))', \tag{9}$$

with $\widetilde{\nabla}_{v(\lambda)'} l(X_i; \psi, \alpha) := (c_{11}\nabla_{\lambda_1\lambda_1} l_i, \dots, c_{qq}\nabla_{\lambda_q\lambda_q} l_i, c_{12}\nabla_{\lambda_1\lambda_2} l_i, \dots, c_{q-1,q}\nabla_{\lambda_{q-1}\lambda_q} l_i)$, where $c_{jk} = 1/2$ if $j = k$ and $c_{jk} = 1$ if $j \neq k$, and $\nabla_{\lambda_j\lambda_k} l_i$ denotes $\nabla_{\lambda_j\lambda_k} l(X_i; \psi, \alpha)$. The coefficients c_{ij}'s are necessary because of the coefficient $(1/2!)$ in front of the third term on the right hand side of (8) and because the $\nabla_{\lambda_j\lambda_k} l_i$'s with $j \neq k$ appear twice in the expansion owing to $\nabla_{\lambda_j\lambda_k} = \nabla_{\lambda_k\lambda_j}$. Define

$$t_n(\psi, \alpha) := \begin{pmatrix} n^{1/2}(\eta - \eta^*) \\ n^{1/2}\alpha(1 - \alpha)v(\lambda) \end{pmatrix}. \tag{10}$$

Define the normalized score and its variance as

$$G_n := n^{-1/2} \sum_{i=1}^n \widetilde{\nabla}_\zeta l(X_i; \psi^*, \alpha) \quad \text{and} \quad \mathcal{I} := E[\widetilde{\nabla}_\zeta l(X_i; \psi^*, \alpha)\widetilde{\nabla}_{\zeta'} l(X_i; \psi^*, \alpha)], \tag{11}$$

respectively, where $\widetilde{\nabla}_\zeta l(X_i; \psi^*, \alpha)$ satisfies

$$\widetilde{\nabla}_\zeta l(X_i; \psi^*, \alpha) = \begin{pmatrix} \nabla_\gamma f(X_i; \gamma^*, \theta^*)/f(X_i; \gamma^*, \theta^*) \\ \nabla_\theta f(X_i; \gamma^*, \theta^*)/f(X_i; \gamma^*, \theta^*) \\ \widetilde{\nabla}_{v(\theta)} f(X_i; \gamma^*, \theta^*)/f(X_i; \gamma^*, \theta^*) \end{pmatrix}$$

with $\widetilde{\nabla}_{v(\theta)} f(X_i; \gamma^*, \theta^*)$ defined similarly to $\widetilde{\nabla}_{v(\lambda)} l(X_i; \psi, \alpha)$. Note that neither G_n nor $\mathcal{I}$ depends on α. With these notations, (8) can be written as a quadratic expansion in terms of $t_n(\psi, \alpha)$:

$$L_n(\psi, \alpha) - L_n(\psi^*, \alpha) = t_n(\psi, \alpha)'G_n - \frac{1}{2}t_n(\psi, \alpha)'\mathcal{I}_n t_n(\psi, \alpha) + R_n(\psi, \alpha), \tag{12}$$

where $\mathcal{I}_n$ corresponds to the negative of the sample Hessian, which converges to $\mathcal{I}$ in probability. For brevity, the formula of $\mathcal{I}_n$ is provided in the proof of Proposition 2. We introduce the following sufficient condition for expanding the log-likelihood function four times:

Assumption 2. *(a) γ^* and θ^* are in the interior of $\Theta_\gamma \times \Theta_\theta$. (b) For every x, $f(x;\gamma,\theta)$ is four times continuously differentiable in a neighborhood of (γ^*,θ^*). (c) For $\alpha \in (0,1)$, $E\sup_{\psi\in\mathcal{N}} ||\nabla^{(k)} \ln f(X;\psi,\alpha)|| < \infty$ for a neighborhood $\mathcal{N}$ of ψ^* and for $k = 1,\ldots,4$, where $\nabla^{(k)}$ denotes the kth derivative w.r.t. ψ. (d) For $\alpha \in (0,1)$, $E||\nabla^{(k)} f(X;\psi^*,\alpha)/f(X;\psi^*,\alpha)||^2 < \infty$ for $k = 1,2,3$.*

The following proposition establishes the asymptotic behavior of $R_n(\psi,\alpha)$, $\mathcal{I}_n$, and G_n.

Proposition 2. *Suppose that Assumption 2 holds. Then, for each $\alpha \in (0,1)$, we have (a) for any $\delta > 0$, $\limsup_{n\to\infty} \Pr(\sup_{\psi\in\Theta_\psi:||\psi-\psi^*||\leq\kappa} |R_n(\psi,\alpha)| > \delta(1 + ||t_n(\psi,\alpha)||)^2) \to 0$ as $\kappa \to 0$. (b) $G_n \to_d G \sim N(0,\mathcal{I})$, (c) $\mathcal{I}_n \to_p \mathcal{I}$.*

3.3 The asymptotic distribution of the LRT statistics for testing H_{01}

In this section, we derive the asymptotics of the LRT statistic for testing H_{01}, building on the representation (12) and Proposition 2. Let us introduce an assumption on the rank of $\mathcal{I}$.

Assumption 3. *$\mathcal{I}$ is finite and positive definite.*

In view of $\nabla_{xy} f(x,y)/f(x,y) = \nabla_{xy} \ln f(x,y) + \nabla_x \ln f(x,y)\nabla_y \ln f(x,y)$, Assumption 3 holds if the covariance matrix of $(\nabla_{(\gamma',\theta')} \ln f(X_i;\gamma^*,\theta^*)$,
$(\text{vech}(\nabla_{\theta\theta'} \ln f(X_i;\gamma^*,\theta^*) + \nabla_\theta \ln f(X_i;\gamma^*,\theta^*)\nabla_{\theta'} \ln f(X_i;\gamma^*,\theta^*))')'$ is finite and nonsingular.

Define $Z_n := \mathcal{I}_n^{-1} G_n$, and rewrite (12) as

$$L_n(\psi,\alpha) - L_n(\psi^*,\alpha) = \frac{1}{2} Z_n' \mathcal{I}_n Z_n - \frac{1}{2}[t_n(\psi,\alpha) - Z_n]' \mathcal{I}_n [t_n(\psi,\alpha) - Z_n] + R_n(\psi,\alpha). \tag{13}$$

Let Θ_η be the parameter space of $\eta = (\gamma',\nu')'$, and let Θ_λ be the parameter space of λ so that $\Theta_\psi = \{\psi = (\eta',\lambda')' : \ \eta \in \Theta_\eta; \ \lambda \in \Theta_\lambda\}$.

The set of feasible values of $t_n(\psi,\alpha)$ is given by the shifted and rescaled parameter space for $(\eta, v(\lambda))$ defined as $\Lambda_n := n^{1/2}(\Theta_\eta - \eta^*) \times n^{1/2}\alpha(1-\alpha)v(\Theta_\lambda)$, where $v(A) := \{x \in \mathbb{R}^{q_\lambda} : x = v(\lambda) \text{ for some } \lambda \in A \subset \mathbb{R}^q\}$. Because $\Lambda_n/n^{1/2}$ is locally approximated by a cone $\Lambda := \mathbb{R}^{p+q} \times v(\mathbb{R}^q)$ (see Andrews (1999) for the definition of "locally approximated by a cone"), the supremum of the left hand side of (13) is approximated as follows (Andrews, 1999, Lemma 2, Theorem 3):

$$\begin{aligned} \sup_{\psi\in\Theta_\psi} 2\{L_n(\psi,\alpha) - L_n(\psi^*,\alpha)\} &= Z_n'\mathcal{I}_n Z_n - \inf_{t\in\Lambda}(t - Z_n)'\mathcal{I}_n(t - Z_n) + o_p(1) \\ &\to_d Z'\mathcal{I}Z - \inf_{t\in\Lambda}(t - Z)'\mathcal{I}(t - Z) = \hat{t}'\mathcal{I}\hat{t}, \end{aligned} \tag{14}$$

where $Z \sim N(0,\mathcal{I}^{-1})$ and $\hat{t}$ is a version of the projection of a Gaussian random vector Z onto the cone Λ w.r.t. the norm $(t'\mathcal{I}t)^{1/2}$ defined by

$$g(\hat{t}) = \inf_{t\in\Lambda} g(t), \quad g(t) := (t - Z)'\mathcal{I}(t - Z). \tag{15}$$

Here, $\hat{t}$ is not necessarily unique because Λ is not necessarily convex. The equality (14) uses the orthogonality condition $\hat{t}'\mathcal{I}(\hat{t} - Z) = 0$; see Andrews (1999, p. 1361) or Lindsay (1995, p. 98). Combining (14) with the asymptotic representation of the log-likelihood function of the one-component model, we obtain the asymptotic distribution of the LRT statistic.

We collect some notations before providing a formal proposition. Partition Z and G as

$$Z = \begin{bmatrix} Z_\eta \\ Z_\lambda \end{bmatrix}, \quad G = \begin{bmatrix} G_\eta \\ G_\lambda \end{bmatrix}, \quad Z_\eta, G_\eta : (p+q) \times 1, \quad Z_\lambda, G_\lambda : q_\lambda \times 1.$$

Define $\mathcal{I}_\eta := E(G_\eta G_\eta')$, $\mathcal{I}_{\lambda\eta} := E(G_\lambda G_\eta')$, $\mathcal{I}_{\eta\lambda} := \mathcal{I}_{\lambda\eta}'$, and $\mathcal{I}_\lambda := E(G_\lambda G_\lambda')$. Note that $Z_\lambda = \mathcal{I}_{\lambda.\eta}^{-1} G_{\lambda.\eta}$, where $G_{\lambda.\eta} := G_\lambda - \mathcal{I}_{\lambda\eta}\mathcal{I}_\eta^{-1}G_\eta$ and $\mathcal{I}_{\lambda.\eta} := \mathcal{I}_{\lambda\lambda} - \mathcal{I}_{\lambda\eta}\mathcal{I}_\eta^{-1}\mathcal{I}_{\eta\lambda} = \text{var}(G_{\eta.\lambda}) = (\text{var}(Z_\lambda))^{-1}$. Similar to $\hat{t}$ in (15), define $\hat{t}_\lambda$ by

$$g_\lambda(\hat{t}_\lambda) = \inf_{t_\lambda \in \Lambda_\lambda} g_\lambda(t_\lambda), \quad g_\lambda(t_\lambda) := (t_\lambda - Z_\lambda)'\mathcal{I}_{\lambda.\eta}(t_\lambda - Z_\lambda), \tag{16}$$

where $\Lambda_\lambda := v(\mathbb{R}^q)$.

The following proposition derives the convergence rate of the MLE and the asymptotic distribution of the LRT statistic. Let $\hat{\psi}_\alpha = (\hat{\eta}_\alpha', \hat{\lambda}_\alpha')'$ denote the MLE that maximizes $L_n(\psi, \alpha)$ for a given α. Let $(\hat{\gamma}_0, \hat{\theta}_0)$ denote the one-component MLE that maximizes the one-component log-likelihood function $L_{0,n}(\gamma, \theta) := \sum_{i=1}^n \ln f(X_i; \gamma, \theta)$. For $\epsilon_1 \in (0, 1/2)$, define the LRT statistic for testing H_{01} as $LR_{n,1}(\epsilon_1) := \max_{\alpha \in [\epsilon_1, 1-\epsilon_1]} 2\{L_n(\hat{\psi}_\alpha, \alpha) - L_{0,n}(\hat{\gamma}_0, \hat{\theta}_0)\}$. As shown in the following proposition, the asymptotic null distribution of the LRT statistic is invariant to α.

Proposition 3. *Suppose Assumptions 1, 2, and 3 hold. Then, for each $\alpha \in (0,1)$, we have (a) $\hat{\eta}_\alpha - \eta^* = O_p(n^{-1/2})$ and $\hat{\lambda}_\alpha = O_p(n^{-1/4})$, (b) $2\{L_n(\hat{\psi}_\alpha, \alpha) - L_n(\psi^*, \alpha)\} \rightarrow_d \hat{t}_\lambda' \mathcal{I}_{\lambda.\eta} \hat{t}_\lambda + G_\eta' \mathcal{I}_\eta^{-1} G_\eta$, (c) $2\{L_n(\hat{\psi}_\alpha, \alpha) - L_{0,n}(\hat{\gamma}_0, \hat{\theta}_0)\} \rightarrow_d \hat{t}_\lambda' \mathcal{I}_{\lambda.\eta} \hat{t}_\lambda$, and (d) $LR_{n,1}(\epsilon_1) \rightarrow_d \hat{t}_\lambda' \mathcal{I}_{\lambda.\eta} \hat{t}_\lambda$.*

When q=1, we have $v(\lambda) = \lambda_1^2$, and the cone Λ becomes convex. Then, $\hat{t}_\lambda$ is uniquely defined as $\hat{t}_\lambda = \arg\inf_{\lambda \geq 0} (\lambda - Z_\lambda)^2 (Var(Z_\lambda))^{-1} = Z_\lambda I\{Z_\lambda \geq 0\}$, and $\hat{t}_\lambda' \mathcal{I}_{\lambda.\eta} \hat{t}_\lambda \sim (\max\{N(0,1), 0\})^2$. Furthermore, it follows from Corollary 1(b) of Andrews (1999) that

$$n^{1/2} v(\hat{\lambda}) \rightarrow_d \hat{t}_\lambda, \quad n^{1/2}(\hat{\eta} - \eta^*) \rightarrow_d \mathcal{I}_\eta^{-1} G_\eta - \mathcal{I}_\eta^{-1} \mathcal{I}_{\eta\lambda} \hat{t}_\lambda. \tag{17}$$

Hence, under the null hypothesis, the MLE of η has a non-standard asymptotic distribution. This is also true when $q \geq 2$.

In a mixture regression model with an intercept and dummy explanatory variables, Assumption 3 fails because some "second-derivative" scores, $\nabla_{\lambda_k \lambda_\ell} l(X_i; \psi^*, \alpha)$'s, are perfectly correlated with the other "second-derivative" scores. The following assumption relaxes Assumption 3 to accommodate such cases.

Assumption 4. *(a) rank$(\mathcal{I}) = p + q + q_\lambda - r$ with $1 \leq r < q_\lambda$, and there exists an $r \times q_\lambda$ matrix B of rank r such that $B \widetilde{\nabla}_{v(\lambda)} l(X; \psi^*, \alpha) = 0$ and $B \nabla^{(k)} \widetilde{\nabla}_{v(\lambda)} l(X; \psi^*, \alpha) = 0$ for $k = 1, 2$ hold almost* *surely. (b) Let $B^{\perp}$ be an $(q_{\lambda}-r)\times q_{\lambda}$ matrix such that $B^{\perp}B'=0$ and $B^{\perp}(B^{\perp})'=I_{q_{\lambda}-r}$, and define*

$$\underset{(p+q+q_{\lambda}-r)\times(p+q+q_{\lambda})}{Q} := \begin{pmatrix} I_{p+q} & 0 \\ 0 & B^{\perp} \end{pmatrix}. \tag{18}$$

Then, the matrix $Q\mathcal{I}Q'$ is finite and positive definite.

The matrix $B^{\perp}$ satisfies the property in Assumption 4(b) when its rows form the basis of the orthogonal complement of the column space of B'. Under Assumption 4, $E[\widetilde{\nabla}_{v(\lambda)}l(X_i;\psi^*,\alpha)\widetilde{\nabla}_{v(\lambda)'}l(X_i;\psi^*,\alpha)]$ can be rank deficient by r, but the non-degenerate linear combinations $B^{\perp}\widetilde{\nabla}_{v(\lambda)}l(X_i;\psi^*,\alpha)$ are not perfectly correlated with $\nabla_{\eta}l(X_i;\psi^*,\alpha)$. Furthermore, the derivatives of $B\widetilde{\nabla}_{v(\lambda)}l(X;\psi^*,\alpha)$ do not provide information for identifying the parameters. As we later illustrate through examples, Assumptions 3 and 4 can be verified for various popular mixture models by computing the first and the second derivatives of the log-likelihood function. When neither Assumption 3 nor Assumption 4 holds, the log-likelihood function needs to be expanded further, up to the sixth or the eighth order, to obtain a valid approximation.

Under Assumption 4, we obtain the following expression from (12) (see the proof of Proposition 4 for the derivation):

$$L_n(\psi,\alpha)-L_n(\psi^*,\alpha)=(Qt_n(\psi,\alpha))'QG_n-\frac{1}{2}(Qt_n(\psi,\alpha))'(Q\mathcal{I}_nQ')Qt_n(\psi,\alpha)+R_n(\psi,\alpha), \tag{19}$$

where $R_n(\psi,\alpha)$ is the remainder term defined in (12). The following proposition extends Proposition 3 under Assumption 4. Define $Z_Q:=[Z_{Q\eta}',Z_{Q\lambda}']'=(Q\mathcal{I}_nQ')^{-1}QG$, where $Z_{Q\lambda}$ is $(q_{\lambda}-r)\times 1$. Define $\hat{t}_{Q\lambda}$ by $\hat{t}_{Q\lambda}\in\arg\inf_{t_{\lambda}\in\Lambda_{\lambda}}(B^{\perp}t_{\lambda}-Z_{Q\lambda})'\mathcal{I}_{Q\lambda.\eta}(B^{\perp}t_{\lambda}-Z_{Q\lambda})$ and $\mathcal{Q}:=(B^{\perp}\hat{t}_{Q\lambda})'\mathcal{I}_{Q\lambda.\eta}B^{\perp}\hat{t}_{Q\lambda}$, where $\mathcal{I}_{Q\lambda.\eta}$ is defined similarly to $\mathcal{I}_{\lambda.\eta}$ using the submatrices of $Q\mathcal{I}Q'$.

Proposition 4. *Suppose Assumptions 1, 2, and 4 hold. Then, for each $\alpha\in(0,1)$, we have (a) for any $\epsilon>0$, $\limsup_{n\to\infty}\Pr(\sup_{\psi\in\Theta_{\psi}:||\psi-\psi^*||\le\kappa}|R_n(\psi,\alpha)|>\epsilon(1+||Qt_n(\psi,\alpha)||)^2)\to 0$ as $\kappa\to 0$, (b) $\hat{\eta}_{\alpha}-\eta^*=O_p(n^{-1/2})$ and $B^{\perp}v(\hat{\lambda}_{\alpha})=O_p(n^{-1/2})$, (c) $2\{L_n(\hat{\psi}_{\alpha},\alpha)-L_n(\psi^*,\alpha)\}\to_d\mathcal{Q}+G_{\eta}'\mathcal{I}_{\eta}G_{\eta}$, (d) $2\{L_n(\hat{\psi}_{\alpha},\alpha)-L_{0,n}(\hat{\gamma}_0,\hat{\theta}_0)\}\to_d\mathcal{Q}$, and (e) $LR_{1,n}(\epsilon_1)\to_d\mathcal{Q}$.*

In Proposition 4, the exact form of Q is model-specific. In the following, we provide some examples, paying close attention to Assumptions 3 and 4. The formula of $\nabla_{v(\lambda)}l(X_i;\psi,\alpha)$ is easily derived using the relation $\nabla_{xy}f(x,y)/f(x,y)=\nabla_x\ln f(x,y)\nabla_y\ln f(x,y)+\nabla_{xy}\ln f(x,y)$.

Example 1 (continued)**.** *(i) Consider the Weibull duration model with the conditional density $f(y|x;\theta^j,\gamma)=\gamma_2y^{\gamma_2-1}\exp\left(\theta^j+\gamma_1'x-\exp(\theta^j+\gamma_1'x)y^{\gamma_2}\right)$ for $j=1,2$, where θ^j is scalar-valued. From (6) and (7), the derivatives of the log-density are given by $\nabla_{\nu}l(y|x;\psi^*,\alpha)=1-\exp(\theta^*+(\gamma_1^*)'x)y^{\gamma_2^*}$, $\nabla_{\gamma_1}l(y|x;\psi^*,\alpha)=x\nabla_{\nu}l(y|x;\psi^*,\alpha)$, $\nabla_{\gamma_2}l(y|x;\psi^*,\alpha)=1/\gamma_2^*+\nabla_{\nu}l(y|x;\psi^*,\alpha)\ln y$, and $\nabla_{\lambda\lambda}l(y|x;\psi^*,\alpha)=\alpha(1-\alpha)\{[\nabla_{\nu}l(y|x;\psi^*,\alpha)]^2-\exp(\theta^*+\gamma_1^*x)y^{\gamma_2^*}\}$. Hence, by inspection, Assumption 3 holds. In view of (17), one should not use the standard asymptotic normal inference on $\hat{\gamma}_2$ when the number of components is over-specified.*

(ii) Consider another Weibull duration model with the conditional density $f(y|x;\theta^j,\gamma) = \gamma y^{\gamma-1}\exp(\theta_1^j+\theta_2^j x - \exp(\theta_1^j+\theta_2^j x)y^\gamma)$. Then, we have $\nabla_{\nu_1} l(y|x;\psi^*,\alpha) = 1-\exp(\theta_1^*+\theta_2^* x)y^{\gamma^*}$, $\nabla_{\nu_2} l(y|x;\psi^*,\alpha) = x\nabla_{\nu_1} l(y|x;\psi^*,\alpha)$, $\nabla_\gamma l(y|x;\psi^*,\alpha) = 1/\gamma^* + \nabla_{\nu_1} l(y|x;\psi^*,\alpha)\ln y$, and

$$\begin{pmatrix} \nabla_{\lambda_1\lambda_1} l(y|x;\psi^*,\alpha)/2 \\ \nabla_{\lambda_2\lambda_2} l(y|x;\psi^*,\alpha)/2 \\ \nabla_{\lambda_1\lambda_2} l(y|x;\psi^*,\alpha) \end{pmatrix} = \begin{pmatrix} \alpha(1-\alpha)\{[\nabla_{\nu_1} l(y|x;\psi^*,\alpha)]^2 - \exp(\theta_1^*+\theta_2^* x)y^{\gamma^*}\}/2 \\ x^2\nabla_{\lambda_1\lambda_1} l(y|x;\psi^*,\alpha)/2 \\ x\nabla_{\lambda_1\lambda_1} l(y|x;\psi^*,\alpha) \end{pmatrix}. \tag{20}$$

When X is neither a constant nor a dummy variable, $\mathcal{I}$ is of full rank and Assumption 3 holds.

(iii) Suppose X is a dummy variable in model (ii). We consider a parameterization such that x_1 and x_2 are dummy variables each taking the value 0 or 1 and satisfying $x_1+x_2=1$.[4] *Let the density be $f(y|x;\gamma,\theta^j) = \gamma y^{\gamma-1}\exp(\theta_1^j x_1+\theta_2^j x_2 - \exp(\theta_1^j x_1+\theta_2^j x_2)y^\gamma)$. Because $x_1x_2=0$, we have $\nabla_{\lambda_1\lambda_2} l(y|x;\psi^*,\alpha)=0$, and Assumption 3 fails. Assumption 4 holds with $B=(0,0,1)$ and $B^\perp = \left(\begin{smallmatrix}1&0&0\\0&1&0\end{smallmatrix}\right)$, and we may apply Proposition 4 with*

$$QG_n = n^{-1/2}\sum_{i=1}^n \begin{pmatrix} \nabla_\eta l(Y_i|X_i;\psi^*,\alpha) \\ X_{1i}\nabla_{\lambda_1\lambda_1} l(Y_i|X_i;\psi^*,\alpha)/2\alpha(1-\alpha) \\ X_{2i}\nabla_{\lambda_2\lambda_2} l(Y_i|X_i;\psi^*,\alpha)/2\alpha(1-\alpha) \end{pmatrix} \text{ and } Qt_n(\psi,\alpha) = \begin{pmatrix} n^{1/2}(\eta-\eta^*) \\ n^{1/2}\alpha(1-\alpha)\lambda_1^2 \\ n^{1/2}\alpha(1-\alpha)\lambda_2^2 \end{pmatrix},$$

where $\nabla_{\lambda_1\lambda_1} l(Y_i|X_i;\psi^*,\alpha)$ and $\nabla_{\lambda_2\lambda_2} l(Y_i|X_i;\psi^*,\alpha)$ are defined in (20).

Assumption 3 does not hold for normal regression model with unknown variance.

Example 2 (Normal mixtures)**.** *Consider a mixture of normal regressions with a common variance $\alpha f(y;\theta^1+w'\beta,\sigma^2)+(1-\alpha)f(y;\theta^2+w'\beta,\sigma^2)$, where $f(y;\mu,\sigma^2)=(1/\sigma)\phi((y-\mu)/\sigma)$ and $\phi(z)$ is the standard normal density. Here, the structural parameter is $\gamma=(\beta',\sigma^2)'$. Because $\nabla_{\mu\mu}f(y;\mu,\sigma^2) = 2\nabla_{\sigma^2}f(y;\mu,\sigma^2)$ holds, $\nabla_{\lambda\lambda}l(y|x;\psi^*,\alpha)$ is perfectly correlated with $\nabla_{\sigma^2}l(y|x;\psi^*,\alpha)$, and Assumption 4 is violated. Similarly, Assumption 4 is violated when the variance is component-specific. Cho and White (2007), Chen and Li (2009), Chen et al. (2012), and Kasahara and Shimotsu (2012) analyze likelihood-based test of the number of components in normal mixture models.*

3.4 The asymptotic distribution of the LRT statistic for testing H_{02}

We now examine the LRT statistic for testing $H_{02}: \alpha(1-\alpha)=0$. We focus on the null hypothesis of $\alpha=0$ below because, by symmetry, the analysis for $\alpha=1$ is the same. Because λ is not identified when H_{02} is true, we follow Andrews (2001, p. 694) and derive the limit of the LRT statistic for each λ in $\Theta_\lambda(\epsilon_2) = \{\lambda\in\Theta_\lambda : ||\lambda|| \geq \epsilon_2\}$ for some $\epsilon_2>0$ and then take its supremum over $\lambda\in\Theta_\lambda(\epsilon_2)$.

To simplify the asymptotic representation and regularity conditions, we use the parameter $\lambda=\theta^1-\theta^2$ defined in (3) and reparameterize (θ^1,θ^2) to (λ,θ^2). Define $\xi := (\gamma',(\theta^2)')' \in \Theta_\xi$ so that

[4] This parameterization gives a simpler representation of QG_n and $Qt_n(\psi,\alpha)$ than that with a constant term and one dummy variable.

the model parameters are (ξ, λ, α), and define the reparameterized log-density as $l(x; \xi, \lambda, \alpha) := \ln(\alpha f(x; \gamma, \theta^2 + \lambda) + (1 - \alpha) f(x; \gamma, \theta^2))$. Collect the partial derivative of $l(x; \xi, \lambda, \alpha)$ w.r.t ξ and its right partial derivative w.r.t. α evaluated at $(\xi^*, \lambda, 0)$ as

$$s(x; \lambda) := \begin{pmatrix} s_\xi(x) \\ s_\alpha(x; \lambda) \end{pmatrix} := \begin{pmatrix} \nabla_\xi l(x; \xi^*, \lambda, 0) \\ \nabla_\alpha l(x; \xi^*, \lambda, 0) \end{pmatrix} = \begin{pmatrix} \dfrac{\nabla_{(\gamma', \theta')'} f(x; \gamma^*, \theta^*)}{f(x; \gamma^*, \theta^*)} \\ \dfrac{f(x; \gamma^*, \theta^* + \lambda) - f(x; \gamma^*, \theta^*)}{f(x; \gamma^*, \theta^*)} \end{pmatrix}. \tag{21}$$

Define $\mathcal{J}(\lambda) := E[s(X_i; \lambda) s(X_i; \lambda)']$, and define its submatrices as $\mathcal{J}_\xi := E[s_\xi(X_i) s_\xi(X_i)']$, $\mathcal{J}_{\xi\alpha}(\lambda) := E[s_\xi(X_i) s_\alpha(X_i; \lambda)]$, $\mathcal{J}_{\alpha\xi}(\lambda) := \mathcal{J}_{\xi\alpha}(\lambda)'$, and $\mathcal{J}_\alpha(\lambda) := E[s_\alpha(X_i; \lambda)^2]$. Let $\{G(\lambda) = (G_\xi', G_\alpha(\lambda))' : \lambda \in \Theta_\lambda(\epsilon_2)\}$ be a mean zero $\mathbb{R}^{(p+q+1)}$-valued Gaussian process such that $E[G(\lambda) G(\lambda)'] = \mathcal{J}(\lambda)$, where G_ξ is $(p+q) \times 1$ and independent of λ, and $G_\alpha(\lambda)$ is 1×1. Define $G_{\alpha.\xi}(\lambda) := G_\alpha(\lambda) - \mathcal{J}_{\alpha\xi}(\lambda) \mathcal{J}_\xi^{-1}(\lambda) G_\xi$ and $\mathcal{J}_{\alpha.\xi}(\lambda) := \mathcal{J}_\alpha(\lambda) - \mathcal{J}_{\alpha\xi}(\lambda) \mathcal{J}_\xi^{-1} \mathcal{J}_{\xi\alpha}(\lambda) = E[G_{\alpha.\xi}(\lambda)^2]$. Define the LRT statistic for testing H_{02} as $LR_{n,2}(\epsilon_2) := 2\{\max_{(\xi, \lambda, \alpha) \in \Theta_\xi \times \Theta_\lambda(\epsilon_2) \times [0,1/2]} L_n(\xi, \lambda, \alpha) - L_{0,n}(\hat{\gamma}_0, \hat{\theta}_0)\}$.

Assumption 5. *(a) γ^* and θ^* are in the interior of $\Theta_\gamma \times \Theta_\theta$. (b) $f(x; \gamma, \theta)$ is twice continuously differentiable on $\Theta_\gamma \times \Theta_\theta$. (c) $\mathcal{J}(\lambda)$ satisfies $0 < \inf_{\lambda \in \Theta_\lambda(\epsilon_2)} \rho_{\min}(\mathcal{J}(\lambda)) \leq \sup_{\lambda \in \Theta_\lambda(\epsilon_2)} \rho_{\max}(\mathcal{J}(\lambda)) < \infty$, where $\rho_{\min}(A)$ and $\rho_{\max}(A)$ are the smallest and the largest eigenvalues of matrix A, respectively.*

Proposition 5. *Suppose Assumptions 1 and 5 hold. Then,*
$LR_{n,2}(\epsilon_2) \to_d \sup_{\lambda \in \Theta_\lambda(\epsilon_2)} (\max\{0, \mathcal{J}_{\alpha.\xi}(\lambda)^{-1/2} G_{\alpha.\xi}(\lambda)\})^2$.

A necessary condition for Assumption 5(c) is $\sup_{\lambda \in \Theta_\lambda(\epsilon_2)} E[\nabla_\alpha l(X_i; \xi^*, \lambda, 0)]^2 < \infty$. This condition is violated in many models including the following Weibull duration model. Furthermore, LCM and Chen and Li (2009) show that a mixture of exponentials and a mixture of normals, respectively, have the same infinite variance problem. The asymptotic distribution of the LRT statistic in such cases remains an open question. In Section 4, we develop a test of H_0 that does not rely on this assumption.

Example 1 (continued)**.** *Consider the Weibull duration model (ii) with the density function $f(y|x; \gamma, \theta^j) = \gamma y^{\gamma-1} \exp[\theta_1^j + \theta_2^j x - \exp(\theta_1^j + \theta_2^j x) y^\gamma]$ for $j = 1, 2$. The score w.r.t. α at $(\xi^*, \lambda, 0)$ is given by $\nabla_\alpha l(y|x; \xi^*, \lambda, 0) = \exp\{\lambda_1 + \lambda_2 x - [\exp(\theta_1^* + \lambda_1 + (\theta_2^* + \lambda_2)x) - \exp(\theta_1^* + \theta_2^* x)] y^\gamma\} - 1$. The conditional variance of $\nabla_\alpha l(Y|X; \xi^*, \lambda, 0)$ given X is*

$$E\left[(\nabla_\alpha l(Y|X; \xi^*, \lambda, 0))^2 | X\right] = \begin{cases} -1 + \frac{\exp(\lambda_1 + \lambda_2 X)}{2 - \exp(-\lambda_1 - \lambda_2 X)} & \text{if } \lambda_1 + \lambda_2 X > -\ln 2, \\ \infty & \text{if } \lambda_1 + \lambda_2 X \leq -\ln 2. \end{cases}$$

Hence, the score has an infinite variance when $\Pr(\lambda_1 + \lambda_2 X \leq -\ln 2) > 0$.

3.5 The asymptotic distribution of the LRT statistic for testing H_0

In this section, we complete the analysis of the LRT statistic for testing $H_0 : (\theta^1 - \theta^2)\alpha(1 - \alpha)=0$ by analyzing the asymptotic behavior of $L_n(\xi, \lambda, \alpha)$ when λ is small. Define the complement of

$\Theta_\lambda(\epsilon_2)$ as $\overline{\Theta}_\lambda(\epsilon_2) = \{\lambda \in \Theta_\lambda : ||\lambda|| < \epsilon_2\}$, and define the LRT statistic when $\lambda \in \overline{\Theta}_\lambda(\epsilon_2)$ as $\overline{LR}_{n,2}(\epsilon_2) := 2\{\sup_{(\xi,\lambda,\alpha)\in\Theta_\xi\times\overline{\Theta}_\lambda(\epsilon_2)\times[0,1/2]} L_n(\xi,\lambda,\alpha) - L_{0,n}(\hat{\gamma}_0,\hat{\theta}_0)\}$. Define the LRT statistic for testing H_0 as $LR_n := 2\{L_n(\hat{\alpha},\hat{\gamma},\hat{\theta}^1,\hat{\theta}^2) - L_{0,n}(\hat{\gamma}_0,\hat{\theta}_0)\}$.

Proposition 6. *(a) Suppose Assumptions 1, 2, and 3 hold. Then,* $\overline{LR}_{n,2}(\epsilon_2) = LR_{n,1}(\epsilon_1) + R_n(\epsilon_2)$, *where* $\limsup_{n\to\infty} \Pr(|R_n(\epsilon_2)| > \delta) \to 0$ *as* $\epsilon_2 \to 0$ *for any* $\delta > 0$. *(b) Suppose Assumptions 1, 2, 3, and 5 hold. Then* $LR_n \to_d \sup_{\lambda\in\Theta_\lambda}(\max\{0, \mathcal{J}_{\alpha.\xi}(\lambda)^{-1/2} G_{\alpha.\xi}(\lambda)\})^2$.

Proposition 6(b) shows that the asymptotic distribution of the LRT statistic for testing H_0 is the supremum of the square of a Gaussian process over Θ_λ, thus generalizing the results of Chen and Chen (2001) and Cho and White (2007, 2010) to the case with a vector mixing parameter. Here, both the compactness of the parameter space Θ_θ and the finiteness of Fisher information under H_{02} are crucial.

Proposition 6 does not apply to testing the homogeneity in the normal mixture with a common variance because neither Assumption 3 nor Assumption 4 holds (see Example 2). Chen and Chen (2003) and Cho and White (2007) derive the asymptotic distribution of the LRT statistic in such a case.

4 Likelihood ratio test of $H_0 : m = m_0$ against $H_0 : m = m_0 + 1$

In this section, we derive the asymptotic distribution of the LRT statistic for testing m_0 against $m_0 + 1$ components for any $m_0 \geq 1$. When $m_0 \geq 2$, there are many ways to generate the m_0-component true model from the (m_0+1)-component model. We develop a partition of the parameter space, where each subset corresponds to a specific way of generating the true model. We then derive the asymptotic distribution of the LRT statistic for each subset, and characterize the asymptotic distribution of the LRT statistic by their maximum.

Consider the mixture pdf with m_0 components $f_0(x;\varphi_0) = \sum_{j=1}^{m_0} \alpha_0^j f(x;\gamma_0,\theta_0^j)$, where $\varphi_0 := (\alpha_0', \gamma_0', \vartheta_0')'$, $\alpha_0 := (\alpha_0^1,\ldots,\alpha_0^{m_0-1})' \in \Theta_{\alpha_0} := \{(\alpha^1,\ldots,\alpha^{m_0-1})' : \alpha^j \geq 0, \sum_{j=1}^{m_0-1}\alpha^j \in (0,1)\}$, and $\vartheta_0 := ((\theta_0^1)',\ldots,(\theta_0^{m_0})')' \in \Theta_{\vartheta_0} := \Theta_\theta^{m_0}$ with $\Theta_\theta \subset \mathbb{R}^q$. Here, the subscript "0" signifies the parameter of the m_0–component model. The parameter $\alpha_0^{m_0}$ is omitted from α_0 and is determined by $\alpha_0^{m_0} = 1 - \sum_{j=1}^{m_0-1}\alpha_0^j$. We define $\Theta_{\varphi_0} := \Theta_{\alpha_0} \times \Theta_\gamma \times \Theta_{\vartheta_0}$.

We assume that a random sample $X_1,\ldots,X_n$ of size n is generated from this m_0-component mixture density with the true parameter value $\varphi_0^* = ((\alpha_0^*)', (\gamma^*)', (\vartheta_0^*)')'$, where $\alpha_0^{j*} > 0$ for $j = 1,\ldots,m_0$:

$$f_0(x;\varphi_0^*) := \sum_{j=1}^{m_0} \alpha_0^{j*} f(x;\gamma_0^*,\theta_0^{j*}). \qquad (22)$$

Finite mixture models are identified only up to label switching. Thus, for identification, we assume that $\theta_0^{1*} < \ldots < \theta_0^{m_0*}$ using the lexicographic order.

We are interested in testing the number of components in a finite mixture model:

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

Let the density of an (m_0+1)–component mixture model be

$$f(x;\varphi) := \sum_{j=1}^{m_0+1} \alpha^j f(x;\gamma,\theta^j), \tag{23}$$

where $\varphi := (\alpha', \gamma', \vartheta')'$, $\alpha := (\alpha^1, \ldots, \alpha^{m_0})'$ with $\alpha^{m_0+1} = 1 - \sum_{j=1}^{m_0} \alpha^j$, and $\vartheta := ((\theta^1)', \ldots, (\theta^{m_0+1})')' \in \Theta_\vartheta := \Theta_\theta^{m_0+1}$. Define the set of admissible values of α by $\Theta_\alpha := \{(\alpha^1, \ldots, \alpha^{m_0})' : \alpha^j \geq 0, \sum_{j=1}^{m_0} \alpha^j \in [0,1]\}$, and let $\Theta_\varphi := \Theta_\alpha \times \Theta_\gamma \times \Theta_\vartheta$. Define a subset of Θ_φ that excludes α on the boundary of Θ_α as $\Theta_{\varphi+} := \{\varphi \in \Theta_\varphi : \alpha^j > 0, \sum_{j=1}^{m_0} \alpha^j \in (0,1)\}$. Define the set of the values of φ that gives rise to the true density (22) as $\Upsilon^* := \{\varphi : f(X;\varphi) = f_0(X;\varphi_0^*)$ with probability one$\}$.

Define the unrestricted ((m_0+1)-component) and the restricted (m_0-component) MLE as

$$\hat{\varphi} = \arg\max_{\varphi \in \Theta_\varphi} L_n(\varphi) \quad \text{and} \quad \hat{\varphi}_0 = \arg\max_{\varphi_0 \in \Theta_{\varphi_0}} L_{0,n}(\varphi_0), \tag{24}$$

respectively, where $L_n(\varphi) := \sum_{i=1}^n \ln f(X_i;\varphi)$ and $L_{0,n}(\varphi_0) := \sum_{i=1}^n \ln f_0(X_i;\varphi_0)$. As the following proposition shows, the unrestricted MLE is consistent in the sense that the distance between $\hat{\varphi}$ and Υ^* tends to 0 in probability. Its proof is essentially the same as the proof of Proposition 1 and hence is omitted. Assumption 6 extends Assumption 1 to the (m_0+1)-component model.

Assumption 6. *(a) If $\varphi \notin \Upsilon^*$, then $f(X;\varphi) \neq f_0(X;\varphi_0^*)$ with a nonzero probability. (b) Assumption 1(b)-(d) hold.*

Proposition 7. *Suppose Assumption 6 holds. Then, we have $\inf_{\varphi \in \Upsilon^*} ||\hat{\varphi} - \varphi|| \rightarrow_p 0$.*

The model (23) generates the true density (22) in two different cases: (i) two components have the same mixing parameter so that $\theta^h = \theta^{h+1} = \theta_0^{h*}$ for some h, and (ii) one component has zero mixing proportion so that $\alpha^h = 0$ for some h. Accordingly, we define the subsets of the parameter space Θ_φ corresponding to (i) and (ii) as, for $h = 1, \ldots, m_0$,

$$\begin{aligned} \Upsilon_{1h}^* := \Big\{ &\varphi \in \Theta_{\varphi+} : \alpha^h + \alpha^{h+1} = \alpha_0^{h*} \text{ and } \theta^h = \theta^{h+1} = \theta_0^{h*}; \\ &\alpha^j = \alpha_0^{j*} \text{ and } \theta^j = \theta_0^{j*} \text{ for } j < h; \\ &\alpha^j = \alpha_0^{j-1*} \text{ and } \theta^j = \theta_0^{j-1*} \text{ for } j > h+1; \ \gamma = \gamma^* \Big\}, \end{aligned}$$

and for $h = 1, \ldots, m_0 + 1$,

$$\Upsilon^*_{2h} := \Big\{ \varphi \in \Theta_\varphi : \alpha^h = 0;\ \alpha^j = \alpha_0^{j*} \text{ and } \theta^j = \theta_0^{j*} \text{ for } j < h; \\ \alpha^j = \alpha_0^{j-1*} \text{ and } \theta^j = \theta_0^{j-1*} \text{ for } j > h;\ \gamma = \gamma^* \Big\}.$$

Because one can always permute the component labels on the (α^j, θ^j)'s, we define Υ^*_{kh} to be the set such that the equalities in braces hold for some permutations of the component labels. Define the union of the Υ^*_{kh}'s as $\Upsilon^*_1 := \{\Upsilon^*_{11} \cup \cdots \cup \Upsilon^*_{1m_0}\}$, $\Upsilon^*_2 := \{\Upsilon^*_{21} \cup \cdots \cup \Upsilon^*_{2,m_0+1}\}$; then, Υ^* is expressed as $\Upsilon^* = \Upsilon^*_1 \cup \Upsilon^*_2$.

Similar to the case of the test of homogeneity, we partition the null hypothesis H_0. Define $H_{01} = \cup_{h=1}^{m_0} H_{0,1h}$ and $H_{02} := \cup_{h=1}^{m_0+1} H_{0,2h}$, where

$$H_{0,1h} : \theta^1 < \cdots < \theta^{h-1} < \theta^h = \theta^{h+1} < \theta^{h+2} < \cdots < \theta^{m_0+1}$$

and

$$H_{0,2h} : \alpha^h = 0$$

so that $H_0 = H_{01} \cup H_{02}$. The inequality constraints are imposed on the θ^j's for identification.

In the following, we analyze the LRT statistics of H_{01}, H_{02}, and H_0 in turn.

4.1 Reparameterization and the LRT statistics for testing H_{01}

In this section, we analyze the behavior of the LRT statistic for testing $H_{01} = \cup_{h=1}^{m_0} H_{0,1h}$. Similar to the case of the test of homogeneity, we approximate the log-likelihood function by expanding it around the true parameter value. Unlike in the homogeneous case, however, the true m_0-component density (22) can be described by many different elements of the parameter space of the (m_0+1)-component model (23). A key observation here is that if we assume $\alpha^h, \alpha^{h+1} > 0$, only Υ^*_{1h} is compatible with $H_{0,1h}$ because $H_{0,1h}$ requires that the hth largest θ^j and the $(h+1)$th largest θ^j take the same value. Therefore, if we assume $\alpha^j > 0$ for all j's, the LRT statistic for testing H_{01} is obtained by maximizing the log-likelihood function locally in a neighborhood of Υ^*_{1h} for each h and then taking the maximum of the maximized values. Furthermore, the local quadratic approximation of the log-likelihood function around Υ^*_{1h} is structurally identical to the approximation we derived in Section 3 in testing H_{01} in the test of homogeneity.

Consider a sufficiently small neighborhood of Υ^*_{1h} such that $\theta^1 < \cdots < \theta^{h-1} < \theta^h, \theta^{h+1} < \theta^{h+2} < \cdots < \theta^{m_0+1}$ holds, and introduce the following one-to-one reparameterization from $(\alpha^1, \ldots, \alpha^{m_0}, \theta^h, \theta^{h+1})$ to $(\beta^1, \ldots, \beta^{m_0-1}, \tau, \nu, \lambda)$ similar to (3):

$$\beta^h := \alpha^h + \alpha^{h+1}, \quad \tau := \frac{\alpha^h}{\alpha^h + \alpha^{h+1}}, \quad \lambda := \theta^h - \theta^{h+1}, \quad \nu := \tau\theta^h + (1-\tau)\theta^{h+1}, \\ (\beta^1, \ldots, \beta^{h-1}, \beta^{h+1} \ldots, \beta^{m_0-1})' := (\alpha^1, \ldots, \alpha^{h-1}, \alpha^{h+2}, \ldots, \alpha^{m_0})', \tag{25}$$

so that $\theta^h = \nu + (1-\tau)\lambda$ and $\theta^{h+1} = \nu - \tau\lambda$. In the reparameterized model, the null restriction $\theta^h = \theta^{h+1}$ implied by $H_{0,1h}$ holds if and only if $\lambda = 0$.

For $h < m_0$, collect the reparameterized model parameters other than τ and λ into

$$\eta^h := (\beta^1, \dots, \beta^{m_0-1}, \gamma', (\theta^1)', \dots, (\theta^{h-1})', \nu', (\theta^{h+2})', \dots, (\theta^{m_0+1})')'$$

and denote its true value by[5]

$$\eta^{h*} := (\alpha_0^{1*}, \dots, \alpha_0^{m_0-1*}, (\gamma^*)', (\theta_0^{1*})', \dots, (\theta_0^{h-1*})', (\theta_0^{h*})', (\theta_0^{h+1*})', \dots, (\theta_0^{m_0*})')'. \tag{26}$$

We also define $\psi^h := ((\eta^h)', \lambda')'$, $\psi^{h*} := ((\eta^{h*})', 0, \dots, 0)'$, and define the parameter space Θ_{ψ^h} similarly to Θ_ψ.

Define the density of X, $f(x;\varphi)$, in (23) in terms of the reparameterized parameters as

$$\begin{aligned} f^h(x;\psi^h,\tau) := \beta^h\left[\tau f(x;\gamma,\nu+(1-\tau)\lambda) + (1-\tau) f(x;\gamma,\nu-\tau\lambda)\right] \\ + \sum_{j=1}^{h-1} \beta^j f(x;\gamma,\theta^j) + \sum_{j=h+1}^{m_0} \beta^j f(x;\gamma,\theta^{j+1}), \end{aligned}$$

with $\beta^{m_0} = 1 - \sum_{j=1}^{m_0-1} \beta^j$. As in (5), the derivative of the reparameterized density w.r.t. λ at ψ^{h*} is zero by $\nabla_\lambda f^h(x;\psi^{h*},\tau) = \beta^h[(1-\tau)\tau f(x;\gamma^*,\theta_0^{h*}) - \tau(1-\tau) f(x;\gamma^*,\theta_0^{h*})] = 0$, whereas its derivative w.r.t. γ and ν at ψ^{h*} are proportional to $\sum_{j=1}^{m_0} \alpha_0^{j*} \nabla_\gamma f(x;\gamma^*,\theta_0^{j*})$ and $\nabla_\theta f(x;\gamma^*,\theta_0^{h*})$.

Define the reparameterized log-likelihood function by

$$L_n^h(\psi^h,\tau) := \sum_{i=1}^n l^h(X_i;\psi^h,\tau), \quad \text{where } l^h(x;\psi^h,\tau) := \ln f^h(x;\psi^h,\tau). \tag{27}$$

Then, $L_n^h(\psi^h,\tau) - L_n^h(\psi^{h*},\tau)$ admits the same expansion (12) as $L_n(\psi,\alpha) - L_n(\psi^*,\alpha)$ with $(t_n(\psi,\alpha), G_n, \mathcal{I}_n, R_n(\psi,\tau))$ on the right of (12) replaced with $(t_n^h(\psi^h,\tau), G_n^h, \mathcal{I}_n^h, R_n^h(\psi^h,\tau))$, where

$$t_n^h(\psi^h,\tau) := \begin{pmatrix} n^{1/2}(\eta^h - \eta^{h*}) \\ n^{1/2}\tau(1-\tau)v(\lambda) \end{pmatrix}, \quad G_n^h := n^{-1/2} \sum_{i=1}^n \widetilde{\nabla}_{\zeta^h} l^h(X_i;\psi^{h*},\tau), \tag{28}$$

where $\zeta^h = ((\eta^h)', v(\lambda)')'$, and $\widetilde{\nabla}_{\zeta^h} l^h(X_i;\psi^{h*},\tau)$ is defined similarly to $\widetilde{\nabla}_\zeta l(X_i;\psi^*,\alpha)$ in (9) and

[5] When $h = m_0$, we need to redefine α as $\alpha = (\alpha_2, \dots, \alpha_{m_0+1})'$ by dropping α_1 from α and redefine η^h and η^{h*} accordingly; however, the essence of our argument remains unchanged.

takes the form $\widetilde{\nabla}_{\zeta^h} l^h(X_i;\psi^{h*},\tau) = (s_{\alpha i}', s_{\gamma i}', s_{\theta i}', (s_{v(\theta)i}^h)')'$, where

$$
\begin{aligned}
s_{\alpha i} &:= \begin{pmatrix} f(X_i;\gamma^*,\theta_0^{1*}) - f(X_i;\gamma^*,\theta_0^{m_0*}) \\ \vdots \\ f(X_i;\gamma^*,\theta_0^{m_0-1*}) - f(X_i;\gamma^*,\theta_0^{m_0*}) \end{pmatrix} \Big/ f_0(X_i;\varphi_0^*), \\
s_{\gamma i} &:= \sum_{j=1}^{m_0} \alpha_0^{j*} \nabla_\gamma f(X_i;\gamma^*,\theta_0^{j*}) / f_0(X_i;\varphi_0^*), \\
s_{\theta i} &:= \begin{pmatrix} \alpha_0^{1*} \nabla_\theta f(X_i;\gamma^*,\theta_0^{1*}) \\ \vdots \\ \alpha_0^{m_0*} \nabla_\theta f(X_i;\gamma^*,\theta_0^{m_0*}) \end{pmatrix} \Big/ f_0(X_i;\varphi_0^*), \\
s_{v(\theta)i}^h &:= \alpha_0^{h*} \nabla_{v(\theta)} f(X_i;\gamma^*,\theta_0^{h*}) / f_0(X_i;\varphi_0^*).
\end{aligned}
\tag{29}
$$

$\mathcal{I}_n^h$ is a matrix that converges to $\mathcal{I}^h := E[\widetilde{\nabla}_{\zeta^h} l^h(X_i;\psi^{h*},\tau)(\widetilde{\nabla}_{\zeta^h} l^h(X_i;\psi^{h*},\tau))']$ in probability.

For $\tau \in (0,1)$ and $h = 1,\ldots,m_0$, define the local MLE of ψ^h by $\hat{\psi}_\tau^h := ((\hat{\eta}_\tau^h)', \hat{\lambda}_\tau)' = \arg\max_{\psi^h \in \mathcal{N}_h^*} L_n^h(\psi^h,\tau)$, where $\mathcal{N}_h^*$ is a closed neighborhood of ψ^{h*} such that ψ^{h*} is in its interior and $\psi^{k*} \notin \mathcal{N}_h^*$ for any $k \neq h$. Because $||\theta_0^{j*} - \theta_0^{k*}|| > 0$ for any $j \neq k$, it is possible to construct such $\mathcal{N}_h^*$'s by making them sufficiently small. Define the local LRT statistic for testing $H_{0,1h}$ as $LR_{n,1h}^\tau := 2\{L_n(\hat{\psi}_\tau^h,\tau) - L_{0,n}(\hat{\varphi}_0)\}$. For $\epsilon_1 \in (0,1/2)$, let $\Theta_\alpha(\epsilon_1) := \{\alpha \in \Theta_\alpha : \alpha^1,\ldots,\alpha^{m_0} \in [\epsilon_1, 1-\epsilon_1]\}$, and define the LRT statistic for testing H_{01} subject to $\alpha \in \Theta_\alpha(\epsilon_1)$ as $LR_{n,1}^{m_0}(\epsilon_1) := \max_{\varphi\in\Theta_\varphi,\alpha\in\Theta_\alpha(\epsilon_1)} 2\{L_n(\varphi) - L_{0,n}(\hat{\varphi}_0)\}$.

The following proposition derives the asymptotic distribution of these LRT statistics. Collect the unique elements of the $\widetilde{\nabla}_{\zeta^h} l^h(X_i;\psi^{h*},\tau)$'s into s_{1i} and s_{vi} defined as $s_{1i} := (s_{\alpha i}', s_{\gamma i}', s_{\theta i}')'$ and $s_{vi} := ((s_{v(\theta)i}^1)',\ldots,(s_{v(\theta)i}^{m_0})')'$. Define $\mathcal{I}_{11} := E[s_{1i}s_{1i}']$, define $\mathcal{I}_{1v}$, $\mathcal{I}_{v1}$, and $\mathcal{I}_{vv}$ similarly, and define $\mathcal{I}_{v.1} := \mathcal{I}_{vv} - \mathcal{I}_{v1}\mathcal{I}_{11}^{-1}\mathcal{I}_{1v}$. Let

$$
\tilde{G}_{\lambda.\eta} = ((G_{\lambda.\eta}^1)',\ldots,(G_{\lambda.\eta}^{m_0})')' \sim N(0,\mathcal{I}_{v.1}),
\tag{30}
$$

be an $\mathbb{R}^{m_0 q_\lambda}$–valued random vector, and define $\mathcal{I}_{\lambda.\eta}^h := \text{var}(G_{\lambda.\eta}^h)$ and $Z_\lambda^h := (\mathcal{I}_{\lambda.\eta}^h)^{-1} G_{\lambda.\eta}^h$. Similar to $\hat{t}_\lambda$ in the test of homogeneity, define $\hat{t}_\lambda^h$ by $g_\lambda^h(\hat{t}_\lambda^h) = \inf_{t_\lambda\in\Lambda_\lambda} g^h(t_\lambda)$, where $g_\lambda^h(t_\lambda) := (t_\lambda - Z_\lambda^h)'\mathcal{I}_{\lambda.\eta}^h(t_\lambda - Z_\lambda^h)$. Assumption 7 corresponds to Assumptions 2 and 3 in the homogeneous case.

Assumption 7. *For $h = 1,\ldots,m_0$, the following holds: (a) γ^* and θ_0^{h*} are in the interior of Θ_γ and Θ_θ. (b) For every x, $\ln f(x;\gamma,\theta)$ is four times continuously differentiable in a neighborhood of (γ^*,θ_0^{h*}). (c) For $\tau \in [0,1]$ and a neighborhood $\mathcal{N}^h$ of ψ^{h*}, $E\sup_{\psi^h\in\mathcal{N}^h} |\nabla^{(k)} \ln f^h(X;\psi^h,\tau)| < \infty$ for $k = 1,\ldots,4$. (d) For $\tau \in [0,1]$, $E||\nabla^{(k)} f^h(X;\psi^{h*},\tau)/f^h(X;\psi^{h*},\tau)||^2 < \infty$ for $k = 1,2,3$. (e) $\mathcal{I} := \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{1v} \\ \mathcal{I}_{v1} & \mathcal{I}_{vv} \end{bmatrix}$ is finite and positive definite.*

Proposition 8. *Suppose Assumptions 6 and 7 hold. Then, for $h = 1,\ldots,m_0$ and for each $\tau \in (0,1)$, (a) $\hat{\eta}_\tau^h - \eta^{h*} = O_p(n^{-1/2})$ and $\hat{\lambda}_\tau = O_p(n^{-1/4})$. (b) $(LR_{n,11}^\tau,\ldots,LR_{n,1m_0}^\tau)' \to_d$*

$[(\hat{t}_\lambda^1)'\mathcal{I}_{\lambda.\eta}^1\hat{t}_\lambda^1,\dots,(\hat{t}_\lambda^{m_0})'\mathcal{I}_{\lambda.\eta}^{m_0}\hat{t}_\lambda^{m_0}]'$. *(c)* $LR_{n,1}^{m_0}(\epsilon_1)\to_d \max\{(\hat{t}_\lambda^1)'\mathcal{I}_{\lambda.\eta}^1\hat{t}_\lambda^1,\dots,(\hat{t}_\lambda^{m_0})'\mathcal{I}_{\lambda.\eta}^{m_0}\hat{t}_\lambda^{m_0}\}$ *if* $\epsilon_1<\min_j \alpha_0^{j*}$.

Proposition 8 can be applied for testing $m_0\geq 2$ in normal mixtures with a common variance.

Example 2. *(continued) Consider testing* $H_0: m=m_0$ *in mixtures of normal regressions with a common variance* $f(x;\varphi)=\sum_{j=1}^{m_0+1}\alpha^j f(y;\theta^j+w'\beta,\sigma^2)$. *When* $m_0\geq 2$, *Assumption 7(e) holds in general because* $\nabla_{\sigma^2}l^h(x;\psi^{h*},\tau)=\sum_{j=1}^{m_0}\alpha^{j*}\nabla_{\sigma^2}f(y;\theta^{j*}+w'\beta^*,\sigma^{2*})$, *which is not perfectly correlated with the* $\nabla_{\mu\mu}f(y;\theta^{j*}+w'\beta,\sigma^{2*})$*'s. Then, applying Proposition 8(c), we have* $LR_{n,1}^{m_0}(\epsilon_1)\to_d\max\{(\xi^{1+})^2,\dots,(\xi^{m_0+})^2\}$, *where* $\xi^{h+}:=\max\{\xi^h,0\}$ *and* $\xi^h:=E[(Z_\lambda^h)^2]^{-1/2}Z_\lambda^h$ *for* $h=1,...,m_0$.

On the other hand, when the variance is component-specific so that $f(x;\varphi)=\sum_{j=1}^{m_0+1}\alpha f(y;\theta_1^j+w'\beta,\theta_2^j)$, *Assumption 7(e) is violated because* $\nabla_{\nu_2}l^h(x;\psi^{h*},\tau)=\nabla_{\sigma^2}f(y;\theta_1^{h*}+w'\beta^*,\theta_2^{h*})$ $=(1/2)\nabla_{\mu\mu}f(y;\theta_1^{h*}+w'\beta^*,\theta_2^{h*})$.

4.2 Testing the null hypotheses H_{02} and H_0

As in Sections 3.4 and 3.5, we consider a testing procedure for the null hypotheses H_{02} and H_0. For $h\in\{1,\dots,m_0\}$, introduce the reparameterized parameter $\lambda^h:=\theta^h-\theta^{h+1}$, and collect all the parameters except for λ^h and α^h into $\xi^h\in\Theta_\xi^h$. Let $l^h(x;\xi^h,\lambda^h,\alpha^h)$ denote the reparameterized log-density, and let $L_n^h(\xi^h,\lambda^h,\alpha^h):=\sum_{i=1}^n l^h(X_i;\xi^h,\lambda^h,\alpha^h)$ denote the reparameterized log-likelihood function. Define the LRT statistic for testing H_{02} as $LR_{n,2}^{m_0}(\epsilon_2):=2\{\max_{h=1,\dots,m_0}\max_{\xi^h\in\Theta_\xi^h,\lambda^h\in\Theta_\lambda^h(\epsilon_2),\alpha^h\in[0,1/2]}L_n^h(\xi^h,\lambda^h,\alpha^h)-L_{0,n}(\hat{\varphi}_0)\}$, where $\Theta_\lambda^h(\epsilon_2):=\{\lambda^h\in\Theta_\lambda:||\lambda^h||\geq\epsilon_2\}$ for some $\epsilon_2>0$.

As in (21), collect the partial derivative of $l^h(x;\xi^h,\lambda^h,\alpha^h)$ w.r.t. ξ^h and its right partial derivative w.r.t. α^h evaluated at $(\xi^{h*},\lambda^h,0)$ as

$$s^h(x;\lambda^h):=\begin{pmatrix}\nabla_\xi l^h(x;\xi^{h*},\lambda^h,0)\\ \nabla_{\alpha^h}l^h(x;\xi^{h*},\lambda^h,0)\end{pmatrix}.$$

Note that $s^h(X_i;\lambda^h)$ is written as $s^h(X_i;\lambda^h)=(s_{\alpha i}',s_{\gamma i}',s_{\theta i}',s_{2i}^h(\lambda^h))'$, where $s_{\alpha i}$, $s_{\gamma i}$, and $s_{\theta i}$ are defined in (29) and $s_{2i}^h(\lambda^h):=[f(X_i;\gamma^*,\theta_0^{h*}+\lambda^h)-f(X_i;\gamma^*,\theta_0^{h*})]/f_0(X_i;\varphi_0^*)$. Collect the unique elements of the $s^h(X_i;\lambda^h)$'s into $\tilde{s}_i(\tilde{\lambda}):=(s_{1i}',s_{2i}^1(\lambda^1),\dots,s_{2i}^{m_0}(\lambda^{m_0}))'$, where $s_{1i}:=(s_{\alpha i}',s_{\gamma i}',s_{\theta i}')'$ and $\tilde{\lambda}=(\lambda^1,\dots,\lambda^{m_0})'$. Define $\tilde{\mathcal{J}}(\tilde{\lambda}):=E[\tilde{s}_i(\tilde{\lambda})\tilde{s}_i(\tilde{\lambda})']$, $\tilde{\mathcal{J}}_{11}:=E[s_{1i}s_{1i}']$, $\tilde{\mathcal{J}}_{12}^h(\lambda^h):=E[s_{1i}s_{2i}^h(\lambda^h)]$, $\tilde{\mathcal{J}}_{21}^h(\lambda^h):=\tilde{\mathcal{J}}_{12}^h(\lambda^h)'$, and $\tilde{\mathcal{J}}_{22}^h(\lambda^h):=E[(s_{2i}^h(\lambda^h))^2]$. Let $\{G(\tilde{\lambda})=(G_1',G_2^1(\lambda^1),\dots,G_2^{m_0}(\lambda^{m_0}))':\tilde{\lambda}\in\tilde{\Theta}_\lambda(\epsilon_2):=\Theta_\lambda^1(\epsilon_2)\times\cdots\times\Theta_\lambda^{m_0}(\epsilon_2)\}$ be a mean zero Gaussian process such that $E[G(\tilde{\lambda})G(\tilde{\lambda})']=\tilde{\mathcal{J}}(\tilde{\lambda})$, where G_1 is $(m_0-1+p+m_0q)\times 1$ and independent of $\tilde{\lambda}$, and $G_2^h(\lambda^h)$ is 1×1. Define $G_{2.1}^h(\lambda^h):=G_2^h(\lambda^h)-\tilde{\mathcal{J}}_{21}^h(\lambda^h)\tilde{\mathcal{J}}_{11}^{-1}G_1$ and $\tilde{\mathcal{J}}_{2.1}^h(\lambda^h):=\tilde{\mathcal{J}}_{22}^h(\lambda^h)-\tilde{\mathcal{J}}_{21}^h(\lambda^h)\tilde{\mathcal{J}}_{11}^{-1}\tilde{\mathcal{J}}_{12}^h(\lambda^h)=E[(G_{2.1}^h(\lambda^h))^2]$.

The following propositions derive the asymptotic distribution of the LRT statistics for testing H_{02} and H_0. Assumption 8 corresponds to Assumption 5 in the homogeneous case. Define $\varrho^h(\epsilon_2):=\sup_{\lambda^h\in\Theta_\lambda^h(\epsilon_2)}(\max\{0,\tilde{\mathcal{J}}_{2.1}^h(\lambda^h)^{-1/2}G_{2.1}^h(\lambda^h)\})^2$.

Assumption 8. *For $h = 1, \ldots, m_0$, the following holds: (a) γ^* and θ_0^{h*} are in the interior of $\Theta_\gamma \times \Theta_\theta$. (b) $f(x; \gamma, \theta)$ is twice continuously differentiable on $\Theta_\gamma \times \Theta_\theta$. (c) $\tilde{\mathcal{J}}(\tilde{\lambda})$ satisfies $0 < \inf_{\tilde{\lambda} \in \tilde{\Theta}_\lambda(\epsilon_2)} \rho_{\min}(\tilde{\mathcal{J}}(\tilde{\lambda})) \leq \sup_{\tilde{\lambda} \in \tilde{\Theta}_\lambda(\epsilon_2)} \rho_{\max}(\tilde{\mathcal{J}}(\tilde{\lambda})) < \infty$.*

Proposition 9. *Suppose Assumptions 6 and 8 hold. Then, (a) $LR_{n,2}^{m_0}(\epsilon_2) \rightarrow_d \max\{\varrho^1(\epsilon_2), \ldots, \varrho^{m_0}(\epsilon_2)\}$. (b) Suppose Assumptions 6, 7, and 8 hold. Then, $2\{L_n(\hat{\varphi}) - L_{0,n}(\hat{\varphi}_0)\} \rightarrow_d \max\{\varrho^1(0), \ldots, \varrho^{m_0}(0)\}$.*

Proposition 9(b) shows that the LRT statistic converges in distribution to the maximum of m_0 random variables, each of which is the supremum of the square of a Gaussian process over Θ_λ and corresponds to the null hypothesis that one component density (for example, $f(x; \gamma^*, \theta_0^{h*})$) has redundancy.

We may obtain the p-value for testing H_{01} by drawing $\tilde{G}_{\lambda.\eta}$ from the multivariate normal distribution in (30) and computing the $(\hat{t}_\lambda^h)' \mathcal{I}_{\lambda.\eta}^h \hat{t}_\lambda^h$'s across different draws of $\tilde{G}_{\lambda.\eta}$. To obtain the p-value for testing H_0, we need to simulate $\sup_{\lambda^h \in \Theta_\lambda^h} (\max\{0, \tilde{\mathcal{J}}_{2.1}^h(\lambda^h)^{-1/2} G_{2.1}^h(\lambda^h)\})^2$. This involves taking a supremum of a stochastic process over Θ_λ^h and is computationally challenging when the dimension of λ is high.[6] On the other hand, simulating the distribution of $LR_{n,1}(\epsilon_1)$ does not involve taking the supremum over unidentified parameters and is thus less costly than simulating the distribution of the LRT statistic in general.

5 Modified EM test

In this section, we develop a test of $H_0 : m = m_0$ against $H_A : m = m_0 + 1$ by extending the EM approach pioneered by LCM. The proposed *modified EM statistic* has the same asymptotic distribution as the LRT statistic for testing H_{01}, and as discussed in the introduction, it has several advantages over the LRT test.

We first develop a (local) modified EM test static for testing $H_{0,1h} : \theta^h = \theta^{h+1}$. Because any of the $\Upsilon_{1\ell}^*$'s is compatible with the true density $f(x; \varphi_0^*)$, we need a device to restrict our estimator to be in a neighborhood of Υ_{1h}^*. To this end, we construct m_0 closed subsets $\{D_1^*, \ldots, D_{m_0}^*\}$ of the parameter space Θ_θ such that $\theta_0^{h*} \in int(D_h)$ and $\theta_0^{k*} \notin D_h^*$ for any $k \neq h$. In practice, we may consider, for $h = 1, \ldots, m_0$,

$$D_h^* := \{\theta \in \Theta_\theta : \ b^{h-1*} \leq \theta_1 \leq b^{h*}\} \tag{31}$$

where θ_1 denotes the first element of θ, b^{0*} and b^{m_0*} are the lower and upper bounds of the support of θ_1, and b^{h*} for $h = 1, \ldots, m_0 - 1$ lies in the open segment $(\theta_{01}^{h*}, \theta_{01}^{h+1*})$ with θ_{01}^{h*} denoting the first element of θ_0^{h*}. When $\theta_{01}^{h*} = \theta_{01}^{h+1*}$, we use the other elements of θ to construct additional cutoff

[6] For instance, for $q = 3$, if we choose 100 discrete grid points for each element of λ to approximate Θ_λ^h, we need to maximize over $(100)^3 = 1000000$ points for each draw.

points. For $h = 1, \ldots, m_0$, define a restricted parameter space

$$\Omega_h^* := \begin{Bmatrix} \vartheta \in \Theta_\vartheta : \theta^j \in D_j^* \text{ for } j = 1, \ldots, h-1; \\ \theta^h, \theta^{h+1} \in D_h^*; \ \theta^j \in D_{j-1}^* \text{ for } j = h+2, \ldots, m_0+1 \end{Bmatrix}.$$

Let $\hat{\Omega}_h$ and $\hat{D}_h$ be consistent estimates of Ω_h^* and D_h^*, which can be obtained from the MLE of the m_0-component model.

We test $H_{0,1h}$ by estimating an (m_0+1)–component model under the restriction $\vartheta \in \hat{\Omega}_h$. For example, when we test a two–component model with $\theta^1 = \theta^2$ against a three–component model, the restriction becomes $\theta^1, \theta^2 \in \hat{D}_1$ and $\theta^3 \in \hat{D}_2$. Because $\hat{\varphi}_0$ is consistent, with probability approaching one, $\Upsilon_{1h}^* \cap (\hat{\Omega}_h \times \Theta_\gamma \times \Theta_\alpha)$ is nonempty while $\Upsilon_{1\ell}^* \cap (\hat{\Omega}_h \times \Theta_\gamma \times \Theta_\alpha)$ is an empty set for all $\ell \neq h$. Therefore, if we maximize $L_n(\alpha, \vartheta, \gamma)$ under the restriction $\{\alpha^j\}_{j=1}^{m_0+1} > 0$ and $\vartheta \in \hat{\Omega}_h$, the resulting estimator approaches a neighborhood of Υ_{1h}^* when the true density is $f(x; \varphi_0^*)$.

In implementing a modified EM test, we consider another reparameterization similar to (25),

$$\begin{aligned} &\beta^h := \alpha^h + \alpha^{h+1}, \quad \tau := \frac{\alpha^h}{\alpha^h + \alpha^{h+1}}, \\ &(\beta^1, \ldots, \beta^{h-1}, \beta^{h+1} \ldots, \beta^{m_0-1})' := (\alpha^1, \ldots, \alpha^{h-1}, \alpha^{h+2}, \ldots, \alpha^{m_0})', \\ &\beta := (\beta^1, \ldots, \beta^{m_0-1})', \quad \beta^* := (\alpha_0^{1*}, \ldots, \alpha_0^{m_0-1*})'. \end{aligned} \tag{32}$$

Let $\phi^h := (\beta', \gamma', \vartheta')'$ with its true value $\phi^{h*} := ((\beta^*)', (\gamma^*)', (\theta_0^{1*})', \ldots,$ $(\theta_0^{h*})', (\theta_0^{h*})', \ldots, (\theta_0^{m_0*})')'$ so that the model parameter is (ϕ^h, τ) and the reparameterized density is $f^h(X_i; \phi^h, \tau)$. Let $L_n^h(\phi^h, \tau) := \sum_{i=1}^n \ln f^h(X_i; \phi^h, \tau)$ denote the log-likelihood function.

Let $\mathcal{T}$ be a finite set of numbers from $(0, 0.5]$. For each $\tau_0 \in \mathcal{T}$, compute

$$\phi^{h(1)}(\tau_0) := \underset{\phi^h : \vartheta \in \hat{\Omega}_h}{\arg\max} \, L_n^h(\phi^h, \tau_0). \tag{33}$$

Note that $\phi^{h(1)}(\tau_0)$ maximizes the log-likelihood function without a penalty term. In the original EM approach by LCM, $\phi^{h(1)}(\tau_0)$ maximizes a penalized log-likelihood function with a penalty term $p(\tau)$ that tends to $-\infty$ as τ approaches to 0 or 1.

Let $\tau^{(1)}(\tau_0) = \tau_0$. Starting from $(\phi^{h(1)}(\tau_0), \tau^{(1)}(\tau_0))$, we update ϕ^h and τ by a generalized EM algorithm. Henceforth, we suppress (τ_0) from $\phi^{h(k)}(\tau_0)$ and $\tau^{(k)}(\tau_0)$. Suppose we have $\phi^{h(k)}$ and $\tau^{(k)}$ calculated. For $i = 1, \ldots, n$ and $j = 1, \ldots, m_0+1$, define the weights for an E-step as

$$\begin{aligned} w_i^{j(k)} &:= \begin{cases} \beta^{j(k)} f(X_i; \gamma^{(k)}, \theta^{j(k)}) / f(X_i; \phi^{h(k)}, \tau^{(k)}) & \text{for } j = 1, \ldots, h-1, \\ \beta^{j-1(k)} f(X_i; \gamma^{(k)}, \theta^{j(k)}) / f(X_i; \phi^{h(k)}, \tau^{(k)}) & \text{for } j = h+2, \ldots, m_0+1, \end{cases} \\ w_i^{h(k)} &:= \frac{\tau^{(k)} \beta^{h(k)} f(X_i; \gamma^{(k)}, \theta^{h(k)})}{f(X_i; \phi^{h(k)}, \tau^{(k)})}, \quad w_i^{h+1(k)} := \frac{(1-\tau^{(k)}) \beta^{h(k)} f(X_i; \gamma^{(k)}, \theta^{h+1(k)})}{f(X_i; \phi^{h(k)}, \tau^{(k)})}. \end{aligned}$$

In an M-step, update τ and β by

$$\tau^{(k+1)} := \frac{\sum_{i=1}^n w_i^{h(k)}}{\sum_{i=1}^n w_i^{h(k)} + \sum_{i=1}^n w_i^{h+1(k)}},$$

$$\beta^{j(k+1)} := \begin{cases} n^{-1}\sum_{i=1}^n w_i^{j(k)} & \text{for } j = 1,\dots,h-1, \\ n^{-1}\sum_{i=1}^n \left(w_i^{h(k)} + w_i^{h+1(k)}\right), & \text{for } j = h, \\ n^{-1}\sum_{i=1}^n w_i^{j+1(k)} & \text{for } j = h+1,\dots,m_0, \end{cases}$$

and update γ and ϑ by

$$\gamma^{(k+1)} := \arg\max_{\gamma\in\Theta_\gamma} \left\{ \sum_{i=1}^n \sum_{j=1}^{m_0+1} w_i^{j(k)} \ln f(X_i;\gamma,\theta^{j(k)}) \right\},$$

$$\theta^{j(k+1)} := \arg\max_{\theta\in\Theta_\theta} \left\{ \sum_{i=1}^n w_i^{j(k)} \ln f(X_i;\gamma^{(k+1)},\theta) \right\}, \quad j = 1,\dots,m_0+1.$$

We update γ and ϑ sequentially to reduce computational burden. Note that $\vartheta^{(k+1)}$ is not restricted to be in $\hat{\Omega}_h$.

For each $\tau_0 \in \mathcal{T}$ and k, define

$$\mathrm{M}_n^{h(k)}(\tau_0) := 2\left\{ L_n^h(\phi^{h(k)}(\tau_0), \tau^{(k)}(\tau_0)) - L_{0,n}(\hat{\varphi}_0) \right\}.$$

Finally, with a pre-specified number K, define the *modified local EM test statistic* by taking the maximum of $\mathrm{M}_n^{h(K)}(\tau_0)$ over $\tau_0 \in \mathcal{T}$ as

$$\mathrm{EM}_n^{h(K)} := \max\left\{ \mathrm{M}_n^{h(K)}(\tau_0) : \tau_0 \in \mathcal{T} \right\}.$$

There are m_0 modified local EM test statistics. If $H_0 : m = m_0$ is correct, then each $\mathrm{EM}_n^{h(K)}$ will have the same asymptotic size. On the other hand, different $\mathrm{EM}_n^{h(K)}$'s have different powers under the alternative depending on the true parameter value. We define the modified EM-test statistic by taking the maximum of m_0 modified local EM test statistics:

$$\mathrm{EM}_n^{(K)} := \max\left\{ \mathrm{EM}_n^{1(K)}, \mathrm{EM}_n^{2(K)}, \dots, \mathrm{EM}_n^{m_0(K)} \right\}.$$

We introduce the following additional regularity condition to derive the asymptotic distribution of the modified EM test statistic.

Assumption 9. *(a)* $E[f(X_i;\gamma^*,\theta_0^{j*})/f(X_i;\phi^{h*},0.5)]^2 < \infty$ *for* $j = 1,\dots,m_0+1$. *(b) For a neighborhood* $\mathcal{N}^h$ *of* ϕ^{h*} *and for an arbitrary small* $\epsilon_1 > 0$*, we have* $E\sup_{(\phi^h,\tau)\in\mathcal{N}^h\times\in[\epsilon_1,1-\epsilon_1]} \left|\nabla_{\phi^h}\left[f(X_i;\gamma,\theta^j)/f(X_i;\phi^h,\tau)\right]\right| < \infty$ *for* $j = 1,\dots,m_0+1$.

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}.

Proposition 10. *Suppose that Assumptions 6, 7, and 9 hold. For any fixed finite K, as $n \to \infty$, $\{EM_n^{h(K)}\}_{h=1}^{m_0} \to_d \{(\hat{t}_\lambda^h)' \mathcal{I}_{\lambda.\eta}^h \hat{t}_\lambda^h\}_{h=1}^{m_0}$ and*

$$EM_n^{(K)} \to_d \max \left\{ (\hat{t}_\lambda^1)' \mathcal{I}_{\lambda.\eta}^1 \hat{t}_\lambda^1, \ldots, (\hat{t}_\lambda^{m_0})' \mathcal{I}_{\lambda.\eta}^{m_0} \hat{t}_\lambda^{m_0} \right\},$$

where the $(\hat{t}_\lambda^h)' \mathcal{I}_{\lambda.\eta}^h \hat{t}_\lambda^h$'s are given in Proposition 8.

One can use simulations or parametric bootstrap to obtain the p-values of the modified EM test. The consistency of the parametric bootstrap follows from the standard argument because the distribution of $(\hat{t}_\lambda^h)' \mathcal{I}_{\lambda.\eta}^h \hat{t}_\lambda^h$ is continuous in ϕ^h.

The modified EM test statistic has the same asymptotic distribution for any finite K, even though it does not use a penalty term. The intuition behind this result is as follows. Note that, given τ_0, $\phi^{h(1)}(\tau_0)$ maximizes the log-likelihood function. When the data are from the m_0-component model, updating τ changes τ only by an $o_p(1)$ amount, because the log-likelihood function is invariant to $\tau \in (0,1)$ up to an negligible term as shown in (13) and (14).

6 Simulation results

This section examines the finite sample performance of the modified EM test for $H_0 : m_0 = 2$ against $H_1 : m_0 = 3$ by Monte Carlo simulation using the Weibull model in Example 1(ii) on page 12, where $X \sim N(0,1)$. Note that as illustrated in Example 1 in Section 3.4, neither the LRT statistic for testing H_{02} nor the LRT statistic for testing H_0 is applicable here because the Fisher information is not finite.

We obtain the critical values for the test statistics by simulation using the result of Proposition 8(c). We set $\mathcal{T} = \{0.5\}$ and consider $K = 1$, $K = 3$, and $K = 5$. We set D_h by (5.1) with $b^{h*} = \kappa \theta_{01}^{h*} + (1-\kappa)\theta_{01}^{h+1*}$ for $h = 1, \ldots, m_0 - 1$ and $\kappa = 0.9$. The sizes and powers are computed from 2000 simulated samples.

Table 1 reports the type I errors of the modified EM test. The data are generated from the Weibull model in Example 1(ii) on page 12 under $\theta^1 = (-1,-1)$, $\theta^2 = (1,1)$, and $\gamma = 1$ with $\alpha = 0.5$ or $\alpha = 0.8$. Across different values of K, the modified EM test has a good size when $n \geq 1000$ if $\alpha = 0.5$ and when $n = 2000$ if $\alpha = 0.8$. The type I errors increase with K. Comparing the upper panel with the lower panel, we notice that the modified EM test has a better size when the mixing proportions are equal across components at $\alpha = 0.5$ than when they are unequal at $\alpha = 0.8$.

The first two panels of Table 2 report the powers of the modified EM test when the data are generated from the Weibull model in Example 1(ii) on page 12 with three components under $\theta^1 = (-1,-1)$, $\theta^2 = (0,0)$, $\theta^3 = (1,1)$, and $\gamma = 1$ with $(\alpha^1, \alpha^2, \alpha^3) = (1/3, 1/3, 1/3)$ or $(\alpha^1, \alpha^2, \alpha^3) = (0.4, 0.4, 0.2)$. The power of the modified EM test increases with sample size. The power also increases with K but not substantially. In view of this result, we recommend using $K = 1$ or $K = 3$. Comparing the first panel with the second panel, the modified EM test has a stronger power when the mixing proportions are equal across components than when they are unequal across components. The last two panels of Table 2 indicate that it is harder to correctly reject the null hypothesis when the values of the coefficients of X are close to each other across components.

We also examine the performance of the original EM test that applies EM steps to a penalized log-likelihood function $PL_n^h(\phi^h, \tau) = L_n^h(\phi^h, \tau) + p(\tau)$, where the penalty term $p(\tau)$ takes the form $p(\tau) = C \ln(2 \min\{\tau, 1 - \tau\})$, as in LCM. The tuning parameter C in the penalty term affects the finite sample performance of the EM test. We experiment with three values, $C = 1$, $C = 2$, and $C = 5$, because no data-driven formula is available for this model. Following LCM, we set $\mathcal{T} = \{0.1, 0.3, 0.5\}$ and $K = 3$. The type I error and powers are examined using the same model as in Tables 1 and 2. The results are reported in Tables 3 and 4. In terms of the type I error, the modified EM test with $K = 1$ and the original EM test with $C = 5$ perform similarly. The EM test with $C = 1$ and $C = 2$ is oversized. In terms of power, the EM test with $C = 5$ performs slightly better than the modified EM test with $K = 1$.

Overall, the performance of the modified EM test and original EM test are similar, although the original EM test is slightly more powerful than the modified EM test. The modified EM test provides a useful alternative to the EM test in applications where it is difficult to find an appropriate value of C for the model at hand.

Table 1: Type I errors (in %) of the modified EM test of $H_0 : m_0 = 2$ against $H_A : m_0 = 3$

nominal level	10%	5%	1%	10%	5%	1%	10%	5%	1%
	$K = 1$			$K = 3$			$K = 5$		
$\theta^1 = (-1, -1)$, $\theta^2 = (1, 1)$, $\gamma = 1$, $\alpha = 0.50$									
$n = 500$	13.7	7.0	1.8	14.8	8.3	2.0	15.2	8.7	2.1
$n = 1000$	10.2	5.4	1.0	10.4	5.9	1.1	10.7	6.2	1.1
$n = 2000$	9.8	5.0	1.5	10.1	5.3	1.5	10.2	5.4	1.5
$\theta^1 = (-1, -1)$, $\theta^2 = (1, 1)$, $\gamma = 1$, $\alpha = 0.80$									
$n = 500$	20.7	12.7	3.9	22.2	13.7	4.3	22.8	14.2	4.8
$n = 1000$	14.7	8.3	2.7	15.1	8.8	2.8	15.4	9.0	2.9
$n = 2000$	13.2	6.9	2.2	13.5	7.1	2.4	13.8	7.2	2.5

Note: Based on 2000 simulated samples. Critical values are obtained by randomly drawing 5000 statistics at the true parameter value. We set $\kappa = 0.9$ and $\mathcal{T} = \{0.5\}$.

Table 2: Powers (in %) of the modified EM test of $H_0 : m_0 = 2$ against $H_A : m_0 = 3$

nominal level	10%	5%	1%	10%	5%	1%	10%	5%	1%
	$K = 1$			$K = 3$			$K = 5$		
$\theta^1 = (-1, -1)$, $\theta^2 = (0, 0)$, $\theta^3 = (1, 1)$, $\gamma = 1$, $(\alpha^1, \alpha^2, \alpha^3) = (1/3, 1/3, 1/3)$									
$n = 500$	94.9	89.8	75.6	95.3	90.3	76.5	95.5	90.7	76.9
$n = 1000$	99.9	99.6	98.3	99.9	99.6	98.4	99.9	99.6	98.5
$n = 2000$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$\theta^1 = (-1, -1)$, $\theta^2 = (0, 0)$, $\theta^2 = (1, 1)$, $\gamma = 1$, $(\alpha^1, \alpha^2, \alpha^3) = (0.4, 0.4, 0.2)$									
$n = 500$	85.2	77.3	57.6	86.6	78.9	60.5	87.2	79.6	61.6
$n = 1000$	98.0	96.5	90.1	98.4	97.2	90.6	98.5	97.7	91.1
$n = 2000$	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7
$\theta^1 = (-1, -0.5)$, $\theta^2 = (0, 0)$, $\theta^3 = (1, 0.5)$, $\gamma = 1$, $(\alpha^1, \alpha^2, \alpha^3) = (1/3, 1/3, 1/3)$									
$n = 500$	42.8	29.9	11.2	44.1	30.8	12.0	44.9	31.4	12.4
$n = 1000$	64.0	53.4	29.1	64.3	53.9	29.6	64.4	54.3	30.3
$n = 2000$	91.5	85.6	68.5	91.5	85.8	68.7	91.6	85.8	68.7
$\theta^1 = (-0.5, -1)$, $\theta^2 = (0, 0)$, $\theta^3 = (0.5, 1)$, $\gamma = 1$, $(\alpha^1, \alpha^2, \alpha^3) = (1/3, 1/3, 1/3)$									
n=500	79.3	69.0	42.2	79.4	69.5	42.8	79.8	70.3	43.3
$n = 1000$	95.9	92.3	80.5	96.0	92.4	80.8	96.1	92.5	80.9
$n = 2000$	100.0	99.9	99.1	100.0	99.9	99.2	100.0	99.9	99.3

Note: Based on 2000 simulated samples. Critical values are obtained by randomly drawing 5000 statistics at the pseudo-true parameter value. We set $\kappa = 0.9$ and $\mathcal{T} = \{0.5\}$.

Table 3: Type I errors (in %) of the original EM test of $H_0 : m_0 = 2$ against $H_A : m_0 = 3$

nominal level	10%	5%	1%	10%	5%	1%	10%	5%	1%
	$C = 1$			$C = 2$			$C = 5$		
$\theta^1 = (-1, -1)$, $\theta^2 = (1, 1)$, $\gamma = 1$, $\alpha = 0.50$									
$n = 500$	19.7	10.8	2.9	15.9	8.8	2.1	13.7	7.5	1.8
$n = 1000$	14.3	8.1	1.8	11.2	6.8	1.6	9.5	5.7	1.1
$n = 2000$	13.7	7.6	1.5	10.8	5.7	1.5	9.8	5.1	1.3
$\theta^1 = (-1, -1)$, $\theta^2 = (1, 1)$, $\gamma = 1$, $\alpha = 0.80$									
$n = 500$	24.2	14.9	4.9	21.3	12.8	4.4	19.4	11.8	3.8
$n = 1000$	17.6	10.2	3.5	15.2	8.8	2.8	13.6	7.8	2.6
$n = 2000$	16.6	9.3	2.9	14.0	7.5	2.5	12.0	6.4	2.2

Note: Based on 2000 simulated samples. Critical values are obtained by randomly drawing 5000 statistics at the true parameter value. We set $\kappa = 0.9$, $K = 3$, and $\mathcal{T} = \{0.1, 0.3, 0.5\}$.

Table 4: Powers (in %) of the original EM test of $H_0 : m_0 = 2$ against $H_A : m_0 = 3$

nominal level	10%	5%	1%	10%	5%	1%	10%	5%	1%
	$C = 1$			$C = 2$			$C = 5$		
$\theta^1 = (-1, -1)$, $\theta^2 = (0, 0)$, $\theta^3 = (1, 1)$, $\gamma = 1$, $(\alpha^1, \alpha^2, \alpha^3) = (1/3, 1/3, 1/3)$									
$n = 500$	95.5	92.2	78.8	95.5	91.7	78.4	95.0	91.1	77.8
$n = 1000$	100.0	99.8	98.8	100.0	99.7	98.7	99.9	99.7	98.5
$n = 2000$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$\theta^1 = (-1, -1)$, $\theta^2 = (0, 0)$, $\theta^2 = (1, 1)$, $\gamma = 1$, $(\alpha^1, \alpha^2, \alpha^3) = (0.4, 0.4, 0.2)$									
$n = 500$	91.0	83.9	66.6	89.6	82.9	64.5	87.5	79.9	62.1
$n = 1000$	99.2	98.6	95.2	99.2	98.5	94.4	98.7	97.8	93.0
$n = 2000$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$\theta^1 = (-1, -0.5)$, $\theta^2 = (0, 0)$, $\theta^3 = (1, 0.5)$, $\gamma = 1$, $(\alpha^1, \alpha^2, \alpha^3) = (1/3, 1/3, 1/3)$									
$n = 500$	49.1	34.6	13.2	47.2	32.7	12.8	45.6	31.6	12.3
$n = 1000$	67.8	56.6	32.6	66.8	55.9	32.1	66.6	55.5	31.6
$n = 2000$	93.4	87.0	70.7	92.8	86.7	70.3	92.7	86.5	70.3
$\theta^1 = (-0.5, -1)$, $\theta^2 = (0, 0)$, $\theta^3 = (0.5, 1)$, $\gamma = 1$, $(\alpha^1, \alpha^2, \alpha^3) = (1/3, 1/3, 1/3)$									
$n = 500$	80.0	69.7	42.5	78.8	69.0	42.1	78.1	68.2	41.4
$n = 1000$	96.0	92.5	80.8	95.7	92.2	80.4	95.7	92.2	80.3
$n = 2000$	100.0	99.9	99.2	100.0	99.9	99.1	100.0	99.9	99.1

Note: Based on 2000 simulated samples. Critical values are obtained by randomly drawing 5000 statistics at the pseudo-true parameter value. We set $\kappa = 0.9$, $K = 3$, and $\mathcal{T} = \{0.1, 0.3, 0.5\}$.

A Proofs

A.1 Proof of Proposition 1

Observe that $1/n$ times the log-likelihood function converges uniformly to $E[\ln f(X_i;\alpha,\gamma,\theta^1,\theta^2)]$ in view of
$\sup_{(\alpha,\gamma,\theta^1,\theta^2)}|\ln[\alpha f(X;\gamma,\theta^1)+(1-\alpha)f(X;\gamma,\theta^2)]|\le\sup_{(\gamma,\theta)}|\ln f(X;\gamma,\theta)|$ and that $E[\ln f(X_i;\alpha,\gamma,\theta^1,\theta^2)]$ is maximized when $(\alpha,\gamma,\theta^1,\theta^2)\in\Gamma^*$. Consequently, the proof follows a standard argument such as Theorem 2.1 of Newey and McFadden (1994) with an adjustment for the fact that the maximizer of $E[\ln f(X_i;\alpha,\gamma,\theta^1,\theta^2)]$ is a set, not a singleton. □

A.2 Proof of Proposition 2

To prove part (a), we first show that

$$\nabla_{\eta\lambda_j}L_n(\psi^*,\alpha)=0,\quad \nabla_{\lambda_i\lambda_j\lambda_k}L_n(\psi^*,\alpha)=O_p(n^{1/2}), \tag{A.1}$$

$$\nabla_{\eta\eta\lambda_i}L_n(\psi^*,\alpha)=O_p(n),\quad \nabla_{\eta\eta\eta}L_n(\psi^*,\alpha)=O_p(n), \tag{A.2}$$

and that for a neighborhood $\mathcal{N}$ of ψ^*,

$$\sup_{\psi\in\Theta_\psi\cap\mathcal{N}}\left|n^{-1}\nabla^{(4)}L_n(\psi,\alpha)-E\nabla^{(4)}\ln f(X_i;\psi,\alpha)\right|=o_p(1), \tag{A.3}$$

$$E\nabla^{(4)}\ln f(X_i;\psi,\alpha)\text{ is continuous in }\psi\in\Theta_\psi\cap\mathcal{N}. \tag{A.4}$$

Equation (A.1) follows from Proposition A(a)(b) and Assumption 2(d). Equation (A.2) is a simple consequence of Assumption 2(d). Equations (A.3) and (A.4) follow from Assumption 2(c) and Lemma 2.4 of Newey and McFadden (1994).

Expanding $L_n(\psi,\alpha)$ four times around (ψ^*,α), noting that $\nabla_\lambda L_n(\psi^*,\alpha)=0$, comparing the expansion with the right hand side of (8), and applying (A.1)–(A.3) gives

$$R_n(\psi,\alpha)=O_p(n^{1/2})\sum_{i=1}^q\sum_{j=1}^q\sum_{k=1}^q\lambda_i\lambda_j\lambda_k+O_p(n)\left(\sum_{i=1}^q||\eta-\eta^*||^2\lambda_i+||\eta-\eta^*||^3\right) \tag{A.5}$$

$$+O_p(n)\sum_{i=1}^q\sum_{j=1}^q\sum_{k=1}^q\left(||\eta-\eta^*||^4+||\eta-\eta^*||^3|\lambda_i|+||\eta-\eta^*||^2|\lambda_i\lambda_j|+||\eta-\eta^*|||\lambda_i\lambda_j\lambda_k|\right) \tag{A.6}$$

$$+\frac{1}{4!}\sum_{i=1}^q\sum_{j=1}^q\sum_{k=1}^q\sum_{\ell=1}^q\{\nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell}L_n(\psi^\dagger,\alpha)-\nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell}L_n(\psi^*,\alpha)\}\lambda_i\lambda_j\lambda_k\lambda_\ell. \tag{A.7}$$

with $\psi^\dagger$ being between ψ and ψ^*. Because $||t_n(\psi,\alpha)||^2=n||\eta-\eta^*||^2+n\sum_{i=1}^q\sum_{j=1}^i\alpha^2(1-\alpha)^2|\lambda_i\lambda_j|^2$, the right hand side of (A.5) and the terms in (A.6) are bounded by $O_p(1)(||t_n(\psi,\alpha)||+||t_n(\psi,\alpha)||^2)(||\eta-\eta^*||+||\lambda||)$. In view of (A.3) and (A.4), (A.7) is bounded by $||t_n(\psi,\alpha)||^2[d(\psi^\dagger)+$

$o_p(1)]$ with $d(\psi^\dagger) \to 0$ as $\psi^\dagger \to \psi^*$, where a function $d(\psi^\dagger)$ corresponds to $n^{-1}E[\nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell}L_n(\psi^\dagger, \alpha) - \nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell}L_n(\psi^*, \alpha)]$. Therefore, $R_n(\psi, \alpha) = (1 + ||t_n(\psi, \alpha)||)^2[d(\psi^\dagger) + o_p(1) + O_p(||\psi - \psi^*||)]$, and part (a) follows.

For part (b), note that $E\nabla_{\lambda\lambda'}l(X; \psi^*, \alpha) = \alpha(1-\alpha)E[\nabla_{\theta\theta'}f(X; \gamma^*, \theta^*)/f(X; \gamma^*, \theta^*)] = 0$ from (7). Therefore, $E[\widetilde{\nabla}_\varsigma l(X; \psi^*, \alpha)] = 0$, and part (b) follows from the Lindeberg–Lévy central limit theorem and the finiteness of $\mathcal{I}$.

For part (c), we first provide the formula of $\mathcal{I}_n$. Partition $\mathcal{I}_n$ as

$$\mathcal{I}_n = \begin{pmatrix} \mathcal{I}_{n\eta} & \mathcal{I}_{n\eta v} \\ \mathcal{I}_{n\eta v}' & \mathcal{I}_{nv} \end{pmatrix}, \quad \mathcal{I}_{n\eta} : (p+q) \times (p+q), \quad \mathcal{I}_{n\eta v} : (p+q) \times q_\lambda, \quad \mathcal{I}_{nv} : q_\lambda \times q_\lambda.$$

$\mathcal{I}_{n\eta}$ is given by $\mathcal{I}_{n\eta} = -n^{-1}\nabla_{\eta\eta'}L_n(\psi^*, \alpha)$. For $\mathcal{I}_{n\eta v}$, define $A_{ij} = n^{-1}\nabla_{\eta\lambda_i\lambda_j}L_n(\psi^*, \alpha)$, so that the fourth term on the right of (8) is written as
$(n/2)\sum_{i=1}^q\sum_{j=1}^q(\eta - \eta^*)'A_{ij}\lambda_i\lambda_j = n\sum_{i=1}^q\sum_{j=1}^i c_{ij}(\eta - \eta^*)'A_{ij}\lambda_i\lambda_j$, where the c_{ij}'s are defined when we introduce $\widetilde{\nabla}_\varsigma l(X; \psi, \alpha)$ in (9). Then, by defining
$\mathcal{I}_{n\eta v} = -(c_{11}A_{11}, \ldots, c_{qq}A_{qq}, c_{12}A_{12}, \ldots, c_{q-1,q}A_{q-1,q})/\alpha(1-\alpha)$, the fourth term on the right of (8) equals $-n(\eta - \eta^*)'\mathcal{I}_{n\eta v}[\alpha(1-\alpha)v(\lambda)]$. For $\mathcal{I}_{nv}$, define $B_{ijk\ell} = n^{-1}(8/4!)\nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell}L_n(\psi^*, \alpha)$ so that the fifth term on the right of (8) is written as $(n/8)\sum_{i=1}^q\sum_{j=1}^q\sum_{k=1}^q\sum_{\ell=1}^q B_{ijk\ell}\lambda_i\lambda_j\lambda_k\lambda_\ell =$ $(n/2)\sum_{i=1}^q\sum_{j=1}^i\sum_{k=1}^q\sum_{\ell=1}^k c_{ij}c_{k\ell}B_{ijk\ell}\lambda_i\lambda_j\lambda_k\lambda_\ell$. Define $\mathcal{I}_{nv}$ such that the $(ij, k\ell)$'s element of $\mathcal{I}_{nv}$ is $-c_{ij}c_{k\ell}B_{ijk\ell}/\alpha^2(1-\alpha)^2$, where the ij's run over $\{(1,1), \ldots, (q,q), (1,2), \ldots, (q-1,q)\}$. Then, the fifth term on the right of (8) equals $-(n/2)[\alpha(1-\alpha)v(\lambda)]'\mathcal{I}_{nv}[\alpha(1-\alpha)v(\lambda)]$. With this definition of $\mathcal{I}_n$, the expansion (8) is written as (12) in terms of $t_n(\psi, \alpha)$.

We complete the proof of part (c) by showing that $\mathcal{I}_n \to_p \mathcal{I}$. $\mathcal{I}_{n\eta} \to_p \mathcal{I}_\eta$ holds trivially. For $\mathcal{I}_{n\eta v}$, it follows from Proposition A(b), Assumption 2(c), and the law of large numbers that $A_{ij} \to_p -E[\nabla_\eta l(X; \psi^*, \alpha)\nabla_{\lambda_i\lambda_j}l(X; \psi^*, \alpha)]$, giving $\mathcal{I}_{n\eta v} \to_p E[\nabla_\eta l(X, \psi^*, \alpha)\widetilde{\nabla}_{v(\lambda)'}l(X, \psi^*, \alpha)/\alpha(1-\alpha)] = \mathcal{I}_{\eta v}$. For $\mathcal{I}_{nv}$, Proposition A(c), Assumption 2(c), and the law of large numbers imply that $\sum_{i=1}^q\sum_{j=1}^q\sum_{k=1}^q\sum_{\ell=1}^q B_{ijk\ell}\lambda_i\lambda_j\lambda_k\lambda_\ell$
$\to_p -\sum_{i=1}^q\sum_{j=1}^q\sum_{k=1}^q\sum_{\ell=1}^q E[\nabla_{\lambda_i\lambda_j}l(X; \psi^*, \alpha)\nabla_{\lambda_k\lambda_\ell}l(X; \psi^*, \alpha)]\lambda_i\lambda_j\lambda_k\lambda_\ell$, where the factor $(8/4!) = 1/3$ in $B_{ijk\ell}$ and the three derivatives on the right hand side of Proposition A(c) cancel each other. Therefore, we have $\mathcal{I}_{nv} \to_p E[\widetilde{\nabla}_{v(\lambda)}l(X, \psi^*, \alpha)\widetilde{\nabla}_{v(\lambda)'}l(X, \psi^*, \alpha)$
$/\alpha^2(1-\alpha)^2] = \mathcal{I}_v$, and $\mathcal{I}_n \to_p \mathcal{I}$ follows. □

A.3 Proof of Proposition 3

We suppress the subscript α from $\hat{\psi}_\alpha$, $\hat{\eta}_\alpha$, and $\hat{\lambda}_\alpha$. The proof of part (a) closely follows the proof of Theorem 1 of Andrews (1999) (A99, hereafter). Let $T_n := \mathcal{I}_n^{1/2} t_n(\hat{\psi}, \alpha)$. Then, in view of (12),

$$
\begin{aligned}
o_p(1) &\leq L_n(\hat{\psi}, \alpha) - L_n(\psi^*, \alpha) \\
&= T_n' \mathcal{I}_n^{1/2} Z_n - \frac{1}{2} ||T_n||^2 + R_n(\hat{\psi}, \alpha) \\
&= O_p(||T_n||) - \frac{1}{2} ||T_n||^2 + (1 + ||\mathcal{I}_n^{-1/2} T_n||)^2 o_p(1) \\
&= ||T_n|| O_p(1) - \frac{1}{2} ||T_n||^2 + o_p(||T_n||) + o_p(||T_n||^2) + o_p(1),
\end{aligned}
\tag{A.8}
$$

where the third equality holds because $\mathcal{I}_n^{1/2} Z_n = O_p(1)$ and $R_n(\hat{\psi}, \alpha) = o_p((1 + ||\mathcal{I}_n^{-1/2} T_n||)^2)$ from Propositions 1 and 2 and Assumption 3. Rearranging this equation gives $||T_n||^2 \leq 2||T_n|| O_p(1) + o_p(1)$. Denote the $O_p(1)$ term by ς_n. Then, $(||T_n|| - \varsigma_n)^2 \leq \varsigma_n^2 + o_p(1) = O_p(1)$, and taking square roots gives $||T_n|| \leq O_p(1)$. In conjunction with $\mathcal{I}_n \to_p \mathcal{I}$, we obtain $t_n(\hat{\psi}, \alpha) = O_p(1)$, giving part (a).

Part (b) follows from Corollary 1(c) of A99. $(B_T, D\ell_T(\theta_0), \mathcal{J}_T, Z_T)$ and $(\mathcal{J}, Z)$ in A99 correspond to our $(n^{1/2}, \sum_{i=1}^n \widetilde{\nabla}_\varsigma l(X_i; \psi^*, \alpha), \mathcal{I}_n, Z_n)$ and $(\mathcal{I}, Z)$. Furthermore, $(H \mathcal{J}_*^{-1} H')^{-1}$ in the statement of Corollary 1(c) of A99 corresponds to our $\mathcal{I}_{\lambda.\eta}$ because ψ in A99 does not exist in our setting. We verify that the conditions of Corollary 1(c) of A99 hold, namely, Assumptions 2-5, 7, and 8 of A99 hold. Assumption 2 holds because Assumption 2* of A99 holds by our Proposition 2(a). Assumption 3 holds by our Proposition 2(b)(c) and our Assumption 3. Assumption 5 follows from Assumption 5* and Lemma 3 of A99 with $b_n = n^{1/2}$ because $(\Theta_\eta - \eta^*) \times v(\Theta_\lambda)$ is locally equal to a cone Λ. Assumption 7(a) does not apply to our problem, and Assumptions 7(b) and 8 hold from our definition of Λ.

For part (c), note that (6) implies that $\nabla_{(\gamma,\theta)} f(x; \gamma^*, \theta^*)$ is identical to $\nabla_\eta f(x; \psi^*, \alpha)$. Therefore, a standard analysis gives $2\{L_{0,n}(\hat{\gamma}_0, \hat{\theta}_0) - L_{0,n}(\gamma^*, \theta^*)\} \to_d G_\eta' \mathcal{I}_\eta^{-1} G_\eta$, where G_η is the same random variable as that in part (b). Hence, part (c) follows from subtracting $2\{L_{0,n}(\hat{\gamma}_0, \hat{\theta}_0) - L_{0,n}(\gamma^*, \theta^*)\}$ from $2\{L_n(\hat{\psi}_\alpha, \alpha) - L_n(\psi^*, \alpha)\}$ and using $L_n(\psi^*, \alpha) = L_{0,n}(\gamma^*, \theta^*)$. Part (d) follows from part (c). □

A.4 Proof of Proposition 4

We prove part (a) by adjusting the proof of Proposition 2(a) to take into account Assumption 4(a). Define a nonsingular matrix $\tilde{Q} := \begin{pmatrix} Q \\ [0\ B] \end{pmatrix}$. Then, under Assumption 4, we obtain the following expression from (12):

$$\begin{aligned}
&L_n(\psi,\alpha) - L_n(\psi^*,\alpha)\\
&= (t_n(\psi,\alpha))'\tilde{Q}^{-1}\tilde{Q}G_n - \frac{1}{2}t_n(\psi,\alpha)'\tilde{Q}^{-1}\tilde{Q}\mathcal{I}_n\tilde{Q}'(\tilde{Q}')^{-1}t_n(\psi,\alpha) + R_n(\psi,\alpha)\\
&= (Qt_n(\psi,\alpha))'QG_n - \frac{1}{2}(Qt_n(\psi,\alpha))'(Q\mathcal{I}_nQ')Qt_n(\psi,\alpha) + R_n(\psi,\alpha)\\
&= \frac{1}{2}Z_{Q,n}'(Q\mathcal{I}_nQ')Z_{Q,n} - \frac{1}{2}[Qt_n(\psi) - Z_{Q,n}]'(Q\mathcal{I}_nQ')[Qt_n(\psi) - Z_{Q,n}] + R_n(\psi,\alpha),
\end{aligned}$$

where $Z_{Q,n} := (Q\mathcal{I}_nQ')^{-1}QG_n$, and the second equality follows from $\tilde{Q}G_n = \left(\begin{smallmatrix} QG_n \\ 0 \end{smallmatrix}\right)$, $(\tilde{Q}^{-1})'t_n(\psi,\alpha) = \left(\begin{smallmatrix} Qt_n(\psi,\alpha) \\ (BB')^{-1}Bn^{1/2}(\eta-\eta^*) \end{smallmatrix}\right)$, and $\tilde{Q}'\mathcal{I}\tilde{Q} = \left(\begin{smallmatrix} Q\mathcal{I}Q' & 0 \\ 0 & 0 \end{smallmatrix}\right)$. Write $R_n(\psi,\alpha)$ in (A.5)-(A.7) as

$$R_n(\psi,\alpha) = R_{1n} + R_{2n} + R_{3n} + (1 + ||n^{1/2}(\eta-\eta^*)||)^2 O_p(||\eta-\eta^*|| + ||\lambda||),$$

where R_{1n}, R_{2n}, and R_{3n} correspond to the first term in the right hand side of (A.5), the fourth term in (A.6), and (A.7), respectively, and are given by
$R_{1n} = O(1)\sum_{j=1}^q \lambda_j v(\lambda)' \sum_{i=1}^n \nabla_{\lambda_j}\widetilde{\nabla}_{v(\lambda)} l(X_i;\psi^*,\alpha)$,
$R_{2n} = O(1)\sum_{j=1}^q \lambda_j v(\lambda)' \sum_{i=1}^n \nabla_{\lambda_j\eta'}\widetilde{\nabla}_{v(\lambda)} l(X_i;\psi^*,\alpha)(\eta-\eta^*)$, and
$R_{3n} = O(1)v(\lambda)' \sum_{i=1}^n \widetilde{\nabla}_{v(\lambda)}\widetilde{\nabla}_{v(\lambda)'}[l(X_i;\psi^\dagger,\alpha) - l(X_i;\psi^*,\alpha)]v(\lambda)$.

Define $P = \left(\begin{smallmatrix} B^\perp \\ B \end{smallmatrix}\right)$ and $B^- = B'(BB')^{-1}$; then, P^{-1} is given by $P^{-1} = [(B^\perp)' \vdots B^-]$. For R_{1n}, note that it follows from Assumption 4(a) that

$$\begin{aligned}
v(\lambda)'\nabla_{\lambda_j}\widetilde{\nabla}_{v(\lambda)} l(X_i;\psi^*,\alpha) &= v(\lambda)'P^{-1}P\nabla_{\lambda_j}\widetilde{\nabla}_{v(\lambda)} l(X_i;\psi^*,\alpha)\\
&= (B^\perp v(\lambda))'B^\perp\nabla_{\lambda_j}\widetilde{\nabla}_{v(\lambda)} l(X_i;\psi^*,\alpha).
\end{aligned}$$

Hence, $R_{1n} = n^{1/2}B^\perp v(\lambda)O_p(||\lambda||)$ holds. A similar argument in view of (A.3) and (A.4) gives $R_{2n} = n^{1/2}(\eta-\eta^*)'n^{1/2}B^\perp v(\lambda)O_p(||\lambda||)$ and $R_{3n} = n^{1/2}B^\perp v(\lambda)'[d(\psi^\dagger) + o_p(1)]n^{1/2}B^\perp v(\lambda)$, where $d(\psi^\dagger)$ is defined similarly to $d(\psi^\dagger)$ in the proof of Proposition 2. Therefore, $R_n(\psi,\alpha) = (1 + ||Qt_n(\psi,\alpha)||)^2$
$[d(\psi^\dagger) + o_p(1) + O_p(||\psi-\psi^*||)]$, giving part (a).

Part (b) follows from applying the proof of Proposition 3(a) to (19). Parts (c)-(e) follow from repeating the proof of Proposition 3(b)-(d). □

A.5 Proof of Proposition 5

The proof is based on Theorem 2(b) of Andrews (2001). Observe that, for each $\lambda \in \Theta_\lambda(\epsilon_2)$, the log-likelihood function $L_n(\xi,\lambda,\alpha)$ can be approximated around $(\xi,\alpha) = (\xi^*,0)$ using the partial derivative w.r.t ξ and the right partial derivative w.r.t. α as (compare it with equation (3.3) of

Andrews (2001, p. 694))

$$\begin{aligned} L_n(\xi,\lambda,\alpha) - L_n(\xi^*,\lambda,0) &= \frac{1}{2} Z_n(\lambda)'\mathcal{J}_n(\lambda)Z_n(\lambda) \\ &- \frac{1}{2}[t_n(\xi,\alpha) - Z_n(\lambda)]'\mathcal{J}_n(\lambda)[t_n(\xi,\alpha) - Z_n(\lambda)] + R_n(\xi,\lambda,\alpha), \end{aligned} \tag{A.9}$$

where $R_n(\xi,\lambda,\alpha)$ is a remainder term, and $\mathcal{J}_n(\lambda)$, $Z_n(\lambda)$, and $t_n(\xi,\alpha)$ are defined as

$$\begin{aligned} &\mathcal{J}_n(\lambda) := \frac{1}{n}\sum_{i=1}^n s(X_i;\lambda)s(X_i;\lambda)', \quad Z_n(\lambda) := \mathcal{J}_n(\lambda)^{-1} n^{-1/2}\sum_{i=1}^n s(X_i;\lambda), \\ &t_n(\xi,\alpha) := \begin{pmatrix} n^{1/2}(\xi-\xi^*) \\ n^{1/2}\alpha \end{pmatrix}, \end{aligned} \tag{A.10}$$

with $s(X_i;\lambda)$ defined in (21). (θ,π) and $(B_T, D\ell_T(\theta_0,\pi), \mathcal{J}_{T\pi}, Z_{T\pi})$ in Andrews (2001) correspond to our $((\xi',\alpha)',\lambda)$ and $(n^{1/2}, \sum_{i=1}^n s(X_i;\lambda), \mathcal{J}_n(\lambda), Z_n(\lambda))$.

We prove the stated result by applying Theorem 2(b) of Andrews (2001) to (A.9). (β,δ,π) and $(B_T, G_\pi, \mathcal{J}_\pi, Z_\pi, Z_{\beta\pi})$ in Andrews (2001, pp. 697-699) correspond to our (α,ξ,λ) and $(n^{1/2}, G(\lambda), \mathcal{J}(\lambda), Z(\lambda), Z_\alpha(\lambda))$, where $Z(\lambda) := \mathcal{J}(\lambda)^{-1}G(\lambda)$, $Z_\alpha(\lambda) := \mathcal{J}_{\alpha.\xi}(\lambda)^{-1}G_{\alpha.\xi}(\lambda)$, and ψ in Andrews (2001, pp. 697-699) does not exist in our setting. The stated result then follows because $s_\xi(x)$ is identical to the score of the one-component model and $\hat{\lambda}'_{\beta\pi}(H\mathcal{J}_{*\pi}^{-1}H')^{-1}\hat{\lambda}_{\beta\pi}$ in Theorem 2(b) of Andrews (2001) is distributed as $(\max\{0, \mathcal{J}_{\alpha.\xi}(\lambda)^{-1/2}G_{\alpha.\xi}(\lambda)\})^2$. We proceed to verify the assumptions of Theorem 2(b) of Andrews (2001) (hereafter, A-Assumptions 2^{2^*}, 3-5, 7, and 8). A-Assumption 2^{2^*}(a)(b) follow from our Assumption 5(a)(b). A-Assumption 2^{2^*}(c) holds because our Assumptions 1 and 5(c) and the uniform law of large numbers imply that $\sup_{\lambda\in\Theta_\lambda(\epsilon_2)} ||\mathcal{J}_n(\lambda) - \mathcal{J}(\lambda)|| \to_p 0$ and $\mathcal{J}(\lambda)$ is continuous. A-Assumption 3 follows from Proposition B(a), $\sup_{\lambda\in\Theta_\lambda(\epsilon_2)} ||\mathcal{J}_n(\lambda) - \mathcal{J}(\lambda)|| \to_p 0$, and our Assumption 5(c). A-Assumption 4 follows from Lemma 1 of Andrews (2001) because, for each $\lambda \in \Theta_\lambda(\epsilon_2)$, $(\tilde{\xi}(\lambda), \tilde{\alpha}(\lambda)) = \arg\max_{(\xi,\alpha)\in\Theta_\xi\times[0,1/2]} L_n(\xi,\lambda,\alpha)$ converges to $(\xi^*,0)$ in probability from the standard consistency proof. A-Assumption 5 holds because (i) the set $[0,1]$ equals a nonnegative half-line locally around 0, and (ii) $\Theta_\xi - \xi^*$ is locally equal to $\mathbb{R}^{p+q}$. A-Assumption 7(a) is not relevant for our problem. A-Assumptions 7(b) and 8 follow from our proof of A-Assumption 5. □

A.6 Proof of Proposition 6

The proof is similar to the proof of Lemma 6 of Cho and White (2007). For brevity, we drop γ from $f(x;\gamma,\theta)$ so that $\xi = \theta^2$, assume λ is scalar, and let f_i^* and ∇f_i^* denote $f(X_i;\theta^*)$ and its derivative, respectively. Define the leading term in the approximation (A.9) of $L_n(\xi,\lambda,\alpha) - L_n(\xi^*,\lambda,0)$ as $D_n(\xi,\alpha,\lambda) := (1/2)Z_n(\lambda)'\mathcal{J}_n(\lambda)Z_n(\lambda) - (1/2)[t_n(\xi,\alpha) - Z_n(\lambda)]'\mathcal{J}_n(\lambda)[t_n(\xi,\alpha) - Z_n(\lambda)]$; then, the stated result follows if we show that the maximum of $D_n(\xi,\alpha,\lambda)$ over (ξ,α,λ) is the same as the maximum of (13) over ψ_α up to an $o_p(1)$ term when λ is small. Note that Assumption 5 is implied by Assumption 2 and 3 when λ is small.

We proceed to obtain an approximation of $D_n(\xi,\alpha,\lambda)$ when λ is small. Expanding $\nabla_\alpha l(x;\xi^*,\lambda,0)$ around $\lambda=0$ twice gives

$$\nabla_\alpha l(x;\xi^*,\lambda,0)=\frac{\nabla_\theta f(x;\theta^*)}{f(x;\theta^*)}\lambda+\frac{1}{2}\frac{\nabla_{\theta\theta} f(x;\theta^*)}{f(x;\theta^*)}\lambda^2+r(x;\lambda^\dagger)\lambda^2, \tag{A.11}$$

where $r(x;\lambda^\dagger):=(1/2)([\nabla_{\theta\theta}f(x;\theta^*+\lambda^\dagger)-\nabla_{\theta\theta}f(x;\theta^*)]/f(x;\theta^*)$ with $\lambda^\dagger\in[0,\lambda]$. Substituting (A.11) into $t_n(\xi,\alpha)'\mathcal{J}_n Z_n(\lambda)$ and $t_n(\xi,\alpha)'\mathcal{J}_n(\lambda)t_n(\xi,\alpha)$ and rearranging terms gives

$$\begin{aligned} t_n(\xi,\alpha)'\mathcal{J}_n Z_n(\lambda)&=\tilde{t}_n(\xi,\alpha)'\tilde{G}_n+r_n(\lambda^\dagger)n^{1/2}\alpha\lambda^2,\\ t_n(\xi,\alpha)'\mathcal{J}_n(\lambda)t_n(\xi,\alpha)&=\tilde{t}_n(\xi,\alpha)'\tilde{\mathcal{I}}_n\tilde{t}_n(\xi,\alpha)+A_n(\lambda^\dagger)O(\|\tilde{t}_n(\xi,\alpha)\|^2),\end{aligned} \tag{A.12}$$

where $\tilde{t}_n(\xi,\alpha):=(n^{1/2}(\xi+\alpha\lambda-\xi^*),n^{1/2}\alpha\lambda^2)'$, $\tilde{G}_n:=n^{-1/2}\sum_{i=1}^n g_i$ and $\tilde{\mathcal{I}}_n:=n^{-1}\sum_{i=1}^n g_i g_i'$ with $g_i:=(\nabla_\theta f_i^*/f_i^*,\nabla_{\theta\theta}f_i^*/2f_i^*)'$, $r_n(\lambda^\dagger):=n^{-1/2}\sum_{i=1}^n r(X_i;\lambda^\dagger)$, and $A_n(\lambda^\dagger):=n^{-1}\sum_{i=1}^n r(X_i,\lambda^\dagger)[\nabla_\theta f_i^*/f_i^*+\nabla_{\theta\theta}f_i^*/2f_i^*+r(X_i,\lambda^\dagger)]$. Note that $\limsup_{n\to\infty}\Pr(|A_n(\lambda^\dagger)|>\delta)\to 0$ as $\lambda^\dagger\to 0$ for any $\delta>0$ and $r_n(\lambda)$ converges to a stochastic process $r(\lambda)$ that is continuous in λ. Moreover, $r(0)$=0 because $E[r(X_i;\lambda^\dagger)]=0$ for any $\lambda^\dagger$ and $r(X_i;0)=0$.

Substituting (A.12) into $D_n(\xi,\alpha,\lambda)$ and defining $\tilde{Z}_n:=\tilde{\mathcal{I}}_n^{-1}\tilde{G}_n$, we obtain $D_n(\xi,\alpha,\lambda)=(1/2)\tilde{Z}_n'\tilde{\mathcal{I}}_n\tilde{Z}_n-(1/2)[\tilde{t}_n(\xi,\alpha)-\tilde{Z}_n]'\tilde{\mathcal{I}}_n[\tilde{t}_n(\xi,\alpha)-\tilde{Z}_n]+R_n(\lambda)$, where $\limsup_{n\to\infty}\Pr(\sup_{\|\lambda\|\le\kappa}|R_n(\lambda)|>\delta(1+\|\tilde{t}_n(\xi,\alpha)\|)^2)\to 0$ as $\kappa\to 0$. Finally, observe that $\tilde{G}_n$ is equal to G_n defined in (11). Therefore, part (a) follows from comparing $D_n(\xi,\alpha,\lambda)$ with (13). Part (b) follows from part (a) and Proposition 5. □

A.7 Proof of Proposition 8

We first prove that $\hat{\psi}_\tau^h-\psi^{h*}=o_p(1)$ for $\tau\in(0,1)$. Because $\psi^{\ell*}\notin\mathcal{N}_h^*$ for any $\ell\neq h$, ψ^{h*} is the only parameter value in $\mathcal{N}_h^*$ that generates the true density. Consequently, $\hat{\psi}_\tau^h-\psi^{h*}=o_p(1)$ follows from a standard consistency proof.

Next, $L_n^h(\psi^h,\tau)-L_n^h(\psi^{h*},\tau)$ admits the same expansion (12) as $L_n(\psi,\alpha)-L_n(\psi^*,\alpha)$ with $(t_n(\psi,\alpha),G_n,\mathcal{I}_n,R_n(\psi,\alpha))$ on the right of (12) replaced with $(t_n^h(\psi^h,\tau),G_n^h,\mathcal{I}_n^h,R_n^h(\psi^h,\tau))$. Hence, part (a) follows from repeating the proof of Proposition 2. Part (b) is proven by extending the proof of Proposition 3 to derive the joint asymptotic distribution of $(LR_{n,11}^\tau,\ldots,LR_{n,1m_0}^\tau)'$, and part (c) follows immediately. □

A.8 Proof of Proposition 9

The proof is essentially the same as the proof of Propositions 5 and 6, except for analyzing the joint asymptotic distribution of m_0 statistics using Proposition B(b), and thus is omitted. □

A.9 Proof of Proposition 10

We suppress (τ_0) from $\phi^{h(k)}(\tau_0)$ and $\tau^k(\tau_0)$. Let $\omega_n^h := (\hat{t}^h_{n\lambda})'\mathcal{I}^h_{n\lambda.\eta}\hat{t}^h_{n\lambda}$ be the sample counterpart of $(\hat{t}^h_{\lambda})'\mathcal{I}^h_{\lambda.\eta}\hat{t}^h_{\lambda}$ such that the local LRT statistic for testing $H_{0,1h}$ satisfies $2\{L_n(\hat{\psi}^h_\tau,\tau) - L_{0,n}(\hat{\varphi}_0)\} = \omega_n^h + o_p(1)$.

We first show $\mathrm{EM}_n^{h(1)} = \omega_n^h + o_p(1)$. Because ϕ^{h*} is the only value of ϕ^h that gives the true density if $\vartheta \in \Omega_h^*$ and $\tau \in (0,1)$, $\phi^{h(1)}$ is also a reparameterized local MLE in a neighborhood of ϕ^{h*}. Therefore, in view of Proposition 8 and its proof, we have $\phi^{h(1)} - \phi^{h*} = O_p(n^{-1/4})$ and $2\{L_n^h(\phi^{h(1)},\tau_0) - L_{0,n}(\hat{\varphi}_0)\} = \omega_n^h + o_p(1)$. Consequently, we have $\mathrm{EM}_n^{h(1)} = \omega_n^h + o_p(1)$.

We proceed to show $\mathrm{EM}_n^{h(k)} = \omega_n^h + o_p(1)$. Because $\phi^{h(1)} - \phi^{h*} = O_p(n^{-1/4})$ and $\tau^{(1)} - \tau_0 = 0$, it follows from Proposition C and induction that $\phi^{h(K)} - \phi^{h*} = O_p(n^{-1/4})$ and $\tau^{(K)} - \tau_0 = O_p(n^{-1/4})$ for all finite K. Because an EM step never decreases the likelihood value (Dempster et al., 1977), we have $L_n^h(\phi^{h(K)},\tau^{(K)}) \geq L_n^h(\phi^{h(1)},\tau_0)$. Let $\tilde{\phi}^h$ be the maximizer of $L_n^h(\phi^h,\tau^{(K)})$ in an arbitrary small closed neighborhood of ϕ^{h*}, then we have $L_n^h(\tilde{\phi}^h,\tau^{(K)}) \geq L_n^h(\phi^{h(K)},\tau^{(K)})$ from the consistency of $\phi^{h(K)}$. Therefore, $2\{L_n^h(\phi^{h(K)},\tau^{(K)}) - L_{0,n}(\hat{\varphi}_0)\} = \omega_n^h + o_p(1)$ holds because $L_n^h(\tilde{\phi}^h,\tau^{(K)}) \geq L_n^h(\phi^{h(K)},\tau^{(K)}) \geq L_n^h(\phi^{h(1)},\tau_0)$ and both $2\{L_n^h(\phi^{h(1)},\tau_0) - L_{0,n}(\hat{\varphi}_0)\}$ and $2\{L_n^h(\tilde{\phi}^h,\tau^{(K)}) - L_{0,n}(\hat{\varphi}_0)\}$ can be written as $\omega_n^h + o_p(1)$ in view of Proposition 8 and its proof. Hence, $\mathrm{EM}_n^{h(K)} = \omega_n^h + o_p(1)$ holds for all h, and the stated result then follows from the definition of $\mathrm{EM}_n^{(K)}$. □

B Auxiliary results and their proofs

Proposition A. *Suppose $f(x;\psi,\alpha)$ is given by (4). Then, for $i,j,k,\ell = 1,2,...,q$,*

$$
\begin{aligned}
&(a)\quad \nabla_{\lambda_i} f(x;\psi^*,\alpha) = 0, \quad \nabla_{\lambda_i}\ln f(x;\psi^*,\alpha) = 0, \quad \nabla_{\eta\lambda_i}\ln f(x;\psi^*,\alpha) = 0,\\
&(b)\quad E[\nabla_{\lambda_i\lambda_j}\ln f^*] = 0, \quad E[\nabla_{\lambda_i\lambda_j\lambda_k}\ln f^*] = 0, \quad E[\nabla_{\eta\lambda_i\lambda_j}\ln f^*] = -E[\nabla_\eta \ln f^* \nabla_{\lambda_k\lambda_\ell}\ln f^*],\\
&(c)\quad E[\nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell}\ln f^*] = -E[\nabla_{\lambda_i\lambda_j}\ln f^* \nabla_{\lambda_k\lambda_\ell}\ln f^*\\
&\qquad\qquad + \nabla_{\lambda_i\lambda_k}\ln f^* \nabla_{\lambda_j\lambda_\ell}\ln f^* + \nabla_{\lambda_i\lambda_\ell}\ln f^* \nabla_{\lambda_j\lambda_k}\ln f^*],
\end{aligned}
$$

where $\nabla^{(k)}\ln f^* = \nabla^{(k)}\ln f(X;\psi^*,\alpha)$ for $k = 1,2,3,4$.

Proof. A direct calculation gives part (a). For parts (b) and (c), observe that $\int \nabla_{\lambda_i}\ln f(x;\psi,\alpha) f(x;\psi,\alpha)dx = 0$ holds for any ψ in the interior of Θ_ψ, and differentiating this equation w.r.t. λ_j gives

$$\int \{\nabla_{\lambda_i\lambda_j}\ln f(x;\psi,\alpha) + \nabla_{\lambda_i}\ln f(x;\psi,\alpha)\nabla_{\lambda_j}\ln f(x;\psi,\alpha)\} f(x;\psi,\alpha)dx = 0. \tag{B.1}$$

Evaluating (B.1) at $\psi = \psi^*$ in conjunction with part (a) gives the first equation in part (b). Differentiating (B.1) w.r.t. λ_k or η and evaluating at $\psi = \psi^*$ give the latter two equations in part (b). Part (c) follows from differentiating (B.1) w.r.t. λ_k and λ_ℓ and evaluating at $\psi = \psi^*$ in conjunction with parts (a)(b). □

Proposition B. *(a) Suppose Assumptions 1 and 5 hold, and let $Z_n(\lambda)$ defined by (A.10) in the proof of Proposition 5. Then $Z_n(\lambda) \Rightarrow Z(\lambda)$ as a stochastic process indexed by $\lambda \in \Theta_\lambda(\epsilon_2)$, where $\{Z(\lambda) : \lambda \in \Theta_\lambda(\epsilon_2)\}$ is a mean zero $\mathbb{R}^q$-valued Gaussian process that has bounded continuous sample paths with probability one and that satisfies $E[Z(\lambda)Z(\lambda)'] = \mathcal{J}(\lambda)^{-1}$. (b) Suppose Assumptions 6 and 8 hold, and define $\tilde{Z}_n(\tilde{\lambda}) := \tilde{\mathcal{J}}(\tilde{\lambda})^{-1}n^{-1/2}\sum_{i=1}^n \tilde{s}_i(\tilde{\lambda})$, where $\tilde{\mathcal{J}}(\tilde{\lambda})$ and $\tilde{s}_i(\tilde{\lambda})$ are defined in Section 4.2. Then $\tilde{Z}_n(\tilde{\lambda}) \Rightarrow \tilde{Z}(\tilde{\lambda})$ as a stochastic process indexed by $\tilde{\lambda} \in \tilde{\Theta}_\lambda(\epsilon_2)$, where $\tilde{Z}(\tilde{\lambda})$ is a mean zero $\mathbb{R}^{(m_0-1+p+m_0q)}$-valued Gaussian process that has bounded continuous sample paths with probability one and that satisfies $E[\tilde{Z}(\tilde{\lambda})\tilde{Z}(\tilde{\lambda})'] = \tilde{\mathcal{J}}(\tilde{\lambda})^{-1}$.*

Proof. Part (a) follows from Theorem 10.2 of Pollard (1990) if (i) $\Theta_\lambda(\epsilon_2)$ is totally bounded, (ii) the finite dimensional distributions of $Z_n(\cdot)$ converge to those of $Z(\cdot)$, and (iii) $\{Z_n(\cdot) : n \geq 1\}$ is stochastically equicontinuous. Condition (i) holds because Θ_θ is compact in the Euclidean space. Condition (ii) follows from Assumption 5(b)(c) and the multivariate CLT. Condition (iii) can be verified by our Assumption 5(b)(c) and Theorem 2 of Andrews (1994) because $\nabla_\xi l(\cdot;\xi^*,\lambda,0)$ and $\nabla_\alpha l(\cdot;\xi^*,\lambda,0)$ are Lipschitz functions indexed by a finite dimensional parameter λ by Assumption 5(b). Part (b) is proven similarly. □

Proposition C. *Suppose Assumptions 6-9 hold. If $\phi^{h(k)}(\tau_0) - \phi^{h*} = O_p(n^{-1/4})$ and $\tau^{(k)}(\tau_0) - \tau_0 = O_p(n^{-1/4})$, then (a) $\tau^{(k+1)}(\tau_0) - \tau_0 = O_p(n^{-1/4})$ and (b) $\phi^{h(k+1)}(\tau_0) - \phi^{h*} = O_p(n^{-1/4})$.*

Proof. We suppress (τ_0) from $\phi^{h(k)}(\tau_0)$ and $\tau^{(k)}(\tau_0)$. The proof of part (a) uses the arguments of the proof of Lemma 3 of Li and Chen (2010). Let $f_i(\gamma,\theta^h)$ and $f_i(\phi^h,\tau)$ denote $f(X_i;\gamma,\theta^h)$ and $f(X_i;\phi^h,\tau)$, respectively. Applying a Taylor expansion to $\sum_{i=1}^n w_i^{h(k)}$ gives

$$\begin{aligned}\sum_{i=1}^n w_i^{h(k)} &= \tau^{(k)}\beta^{h(k)}\sum_{i=1}^n \frac{f_i(\gamma^{(k)},\theta^{h(k)})}{f_i(\phi^{h(k)},\tau^{(k)})} \\ &= \tau^{(k)}\beta^{h(k)}\sum_{i=1}^n \frac{f_i(\gamma^*,\theta_0^{h*})}{f_i(\phi^{h*},\tau_0)} + O_p(n)(\phi^{h(k)}-\phi^{h*}) + O_p(n)(\tau^{(k)}-\tau_0).\end{aligned} \tag{B.2}$$

Because $f_i(\phi^{h*},\tau)$ does not depend on τ, it follows from Assumption 9(a) and $E[f_i(\gamma^*,\theta_0^{h*})/f_i(\phi^{h*},\tau_0)] = 1$ that the right hand side equals $\tau^{(k)}\beta^{h(k)}n(1+O_p(n^{-1/2}))+O_p(n^{3/4}) = \tau^{(k)}\beta^{h(k)}n(1+O_p(n^{-1/4}))$. Similarly, $\sum_{i=1}^n w_i^{h+1(k)} = (1-\tau^{(k)})\beta^{h(k)}n(1+O_p(n^{-1/4}))$. Therefore, we have $\tau^{(k+1)} = \tau^{(k)} + O_p(n^{-1/4}) = \tau_0 + O_p(n^{-1/4})$, giving part (a).

We proceed to show part (b). $\beta^{(k+1)} = \beta^* + o_p(1)$ follows from a similar argument to (B.2). Note that $\gamma^{(k+1)}$ maximizes $Q_n(\gamma) := n^{-1}\sum_{i=1}^n\sum_{j=1}^{m_0+1} w_i^{j(k)} \ln f_i(\gamma,\theta^{j(k)})$. Using a similar argument to (B.2) in conjunction with Assumption 9(b) and $|w_i^{j(k)}| \leq 1$, we have $\sup_{\gamma\in\Theta_\gamma}|Q_n(\gamma)-Q(\gamma)| = o_p(1)$, where $Q(\gamma) := \sum_{j=1}^{m_0}\alpha_0^{j*}E^{j*}[\ln f_i(\gamma,\theta_0^{j*})]$ and E^{j*} denotes the expectation taken under $f(x;\gamma^*,\theta_0^{j*})$, and $\gamma^{(k+1)} \to_p \gamma^*$ follows. Given the consistency of $\gamma^{(k+1)}$, a similar argument gives $\theta^{j(k+1)} \to_p \arg\max_\theta E^{j*}\ln f(X_i;\gamma^*,\theta) = \theta_0^{j*}$ for $j = 1,\ldots,h$ and $\theta^{j(k+1)} \to_p \theta_0^{j-1*}$ for $j = h+1,\ldots,m_0$. This proves $\phi^{h(k+1)} \to_p \phi^{h*}$. Giving the consistency of $\phi^{h(k+1)}$, part (b) follows from repeating the argument in the proof of Proposition 3(a). □

References

Andrews, D. W. K. (1994), “Empirical Process Methods in Econometrics,” in *Handbook of Econometrics*, Amsterdam: North-Holland, vol. 4, pp. 2247–2794.

— (1999), “Estimation When a Parameter is on a Boundary,” *Econometrica*, 67, 1341–1383.

— (2001), “Testing when a Parameter is on the Boundary of the Maintained Hypothesis,” *Econometrica*, 69, 683–734.

Andrews, R. and Currim, I. S. (2003), “A Comparison of Segment Retention Criteria for Finite Mixture Logit Models,” *Journal of Marketing Research*, 40, 235–243.

Cameron, S. V. and Heckman, J. J. (1998), “Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males,” *Journal of Political Economy*, 106, 262–333.

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. (2001), “A Modified Likelihood Ratio Test for Homogeneity in Finite Mixture Models,” *Journal of the Royal Statistical Society, Series B*, 63, 19–29.

— (2004), “Testing for a Finite Mixture Model with Two Components,” *Journal of the Royal Statistical Society, Series B*, 66, 95–115.

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

— (2011a), “The Limiting Distribution of the EM-test of the Order of a Finite Mixture,” in *Mixture Estimation and Applications*, eds. Mengersen, K., Robert, C., and Titterington, D., Hoboken, NJ: Wiley, pp. 55–76.

— (2011b), “Tuning the EM-test for finite mixture models,” *Canadian Journal of Statistics*, 39, 389–404.

Chen, J., Li, P., and Fu, Y. (2012), “Inference on the Order of a Normal Mixture,” *Journal of the American Statistical Association*, Forthcoming.

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.

Cho, J. S. and White, H. (2007), "Testing for Regime Switching," *Econometrica*, 75, 1671–1720.

— (2010), "Testing for Unobserved Heterogeneity in Exponential and Weibull Duration Models," *Journal of Econometrics*, 15, 458–480.

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.

Deb, P. and Trivedi, P. (2002), "The Structure of Demand for Health Care: Latent Class versus Two-part Models." *Journal of Health Economics*, 21, 601–625.

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.

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

Heckman, J. and Singer, B. (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52, 271–320.

Kamakura, W. A. and Russell, G. J. (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structuring," *Journal of Marketing Research*, 26, 379–390.

Kasahara, H. and Shimotsu, K. (2012), "Testing the number of components in normal mixture models," Preprint, University of British Columbia.

Keane, M. P. and Wolpin, K. I. (1997), "The Career Decisions of Young Men," *Journal of Political Economy*, 105, 473–522.

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.

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

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

Newey, W. K. and McFadden, D. L. (1994), “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, Amsterdam: North-Holland, vol. 4, pp. 2111–2245.

Niu, X., Li, P., and Zhang, P. (2011), “Testing Homogeneity in a Multivariate Mixture Model,” *Canadian Journal of Statistics*, 39, 218–238.

Pollard, D. (1990), *Empirical Processes: Theory and Applications*, vol. 2 of *CBMS Conference Series in Probability and Statistics*, Hayward, CA: Institute of Mathematical Statistics.

Redner, R. (1981), “Note on the Consistency of the Maximum Likelihood Estimate for Nonidentifiable Distributions,” *Annals of Statistics*, 9, 225–228.

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

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

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