

CIRJE-F-1044

**Testing the Order of Multivariate Normal Mixture
Models**

Hiroyuki Kasahara
University of British Columbia

Katsumi Shimotsu
The University of Tokyo

March 2017

CIRJE Discussion Papers can be downloaded without charge from:

<http://www.cirje.e.u-tokyo.ac.jp/research/03research02dp.html>

Discussion Papers are a series of manuscripts in their draft form. They are not intended for circulation or distribution except as indicated by the author. For that reason Discussion Papers may not be reproduced or distributed without the written consent of the author.

Testing the Order of Multivariate Normal Mixture Models

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

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

March 2017

Abstract

Testing the number of components in multivariate normal mixture models is a long-standing challenge. This paper develops a likelihood-based test of the null hypothesis of M_0 components against the alternative hypothesis of $M_0 + 1$ components. We derive a local quadratic approximation of the likelihood ratio statistic in terms of the polynomials of the parameters. Based on this quadratic approximation, we propose an EM test of the null hypothesis of M_0 components against the alternative hypothesis of $M_0 + 1$ components, and derive the asymptotic distribution of the proposed test statistic. The simulations show that the proposed test has good finite sample size and power properties.

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

1 Introduction

Finite mixtures of multivariate normal distributions have been widely used in empirical applications in diverse fields such as statistical genetics and statistical finance. Comprehensive surveys on theoretical properties and applications can be found, for example, Lindsay (1995), Titterton et al. (1985), and McLachlan and Peel (2000).

The number of components is an important parameter in applications of finite mixture models. Despite its importance, testing for the number of components in multivariate normal mixture models has been a long-standing unsolved problem because the standard asymptotic analysis of the likelihood ratio test (LRT) statistic breaks down due to problems such as non-identifiable parameters and the true parameter being on the boundary of the parameter space. Numerous

*Address for correspondence: Hiroyuki Kasahara, Vancouver School of Economics, University of British Columbia, 997-1873 East Mall, Vancouver, BC V6T 1Z1, Canada. This research is supported by the Natural Science and Engineering Research Council of Canada and JSPS Grant-in-Aid for Scientific Research (C) No. 26380267. The authors thank the Institute of Statistical Mathematics for the facilities and the use of SGI ICE X.

papers have been written on the subject of the likelihood ratio test for the number of components (see, e.g., Ghosh and Sen, 1985; Chernoff and Lander, 1995; Lemdani and Pons, 1997; Chen and Chen, 2001, 2003; Chen et al., 2004; Garel, 2001, 2005), and the asymptotic distribution of the LRT statistic for general finite mixture models has been derived as a functional of the Gaussian process (Dacunha-Castelle and Gassiat, 1999; Azaïs et al., 2009; Liu and Shao, 2003; Zhu and Zhang, 2004).

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

This paper develops a likelihood-based test of the null hypothesis of M_0 components against the alternative hypothesis of $M_0 + 1$ components for a general $M_0 \geq 1$ in multivariate normal mixtures. We propose an EM test by building on the EM approach pioneered by Li et al. (2009) and Li and Chen (2010). The asymptotic null distribution of the proposed EM test statistic is shown to be the maximum of random variables, each of which is a projection of a Gaussian random variable on a cone.

To the best of our knowledge, no likelihood-based test has yet been developed for testing order of multivariate normal mixtures, even in a simple case of testing the null hypothesis $H_0 : M = 1$ against the alternative hypothesis $H_A : M = 2$. In univariate normal mixtures, Chen and Li (2009) develop an EM test for $M_0 = 1$ against $M_0 = 2$, and Chen et al. (2012) develop an EM test for testing $H_0 : M = M_0$ against $H_A : M > M_0$. Kasahara and Shimotsu (2015) develop an EM test for testing $H_0 : M = M_0$ against $H_A : M = M_0 + 1$ for general $M_0 \geq 1$ in finite normal mixture regression models.

The remainder of this paper is organized as follows. Section 2 introduces multivariate normal mixture models. Section 3 derives a version of LaCam’s differentiable in quadratic mean (DQM) expansion that expands likelihood ratio in terms of a smooth function of parameters. This DQM-type expansion has advantage over the “classical” approach based on the Taylor expansion that expands up to the Hessian term because deriving a higher-order expansion becomes tedious in multivariate normal mixtures. Sections 4 and 5 establish the local quadratic approximation in testing the null hypothesis of M_0 components against the alternative of $M_0 + 1$ components. Section 6 introduces the modified EM test. Section 7 reports the simulation results. The supplementary appendix contains proofs and auxiliary results.

We collect notation. Let $:=$ denote “equals by definition.” Boldface letters denote vectors or matrices. For a matrix \mathbf{B} , let $\lambda_{\min}(\mathbf{B})$ and $\lambda_{\max}(\mathbf{B})$ be the smallest and the largest eigenvalue of \mathbf{B} , respectively. For a k -dimensional vector $\mathbf{x} = (x_1, \dots, x_k)^\top$ and a matrix \mathbf{B} , define $|\mathbf{x}| := (\mathbf{x}^\top \mathbf{x})^{1/2}$

and $|\mathbf{B}| := (\lambda_{\max}(\mathbf{B}^\top \mathbf{B}))^{1/2}$. Let $\mathbf{x}^{\otimes k} := \mathbf{x} \otimes \mathbf{x} \otimes \cdots \otimes \mathbf{x}$ (k times). Let $\mathbb{I}\{A\}$ denote an indicator function that takes value 1 when A is true and 0 otherwise. \mathcal{C} denotes a generic nonnegative finite constant whose value may change from one expression to another. Given a sequence $\{f(\mathbf{Y}_i)\}_{i=1}^n$, let $\nu_n(f(\mathbf{y})) := n^{-1/2} \sum_{i=1}^n [f(\mathbf{Y}_i) - Ef(\mathbf{Y}_i)]$ and $P_n(f(\mathbf{y})) := n^{-1} \sum_{i=1}^n f(\mathbf{Y}_i)$. All the limits are taken as $n \rightarrow \infty$ unless stated otherwise.

2 Multivariate finite normal mixture models

Let $\mathbf{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$ follow the normal distribution with mean $\boldsymbol{\mu} + \mathbf{z}^\top \boldsymbol{\gamma}$ and variance $\boldsymbol{\Sigma}$. The density of \mathbf{x} is

$$f(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) := (2\pi)^{-\frac{d}{2}} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z}^\top \boldsymbol{\gamma})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu} - \mathbf{z}^\top \boldsymbol{\gamma})}{2}\right),$$

where $\boldsymbol{\mu}$ is $d \times 1$, and \mathbf{z} and $\boldsymbol{\gamma}$ are $p \times 1$. Let $\Theta_\gamma \subset \mathbb{R}^p$, $\Theta_\mu \subset \mathbb{R}^d$, and $\Theta_\Sigma \subset \mathbb{S}_+^d$ denote the space of $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$, respectively, where \mathbb{S}_+^d denotes the space of $d \times d$ positive definite matrices. For $M \geq 2$, denote the density of M -component finite normal mixture distribution as:

$$f_M(\mathbf{x}|\mathbf{z}; \boldsymbol{\vartheta}_M) := \sum_{j=1}^M \alpha_j f(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \quad (1)$$

where $\boldsymbol{\vartheta}_M := (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_M, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_M)$ with $\boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_{M-1})^\top$, and α_M being determined by $\alpha_M := 1 - \sum_{j=1}^{M-1} \alpha_j$. $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are mixing parameters that characterize the j -th component, and α_j s are mixing probabilities. $\boldsymbol{\gamma}$ is the coefficient of the covariate \mathbf{z} , and $\boldsymbol{\gamma}$ is assumed to be common to all the components. Define the set of admissible values of $\boldsymbol{\alpha}$ by $\Theta_\alpha := \{\boldsymbol{\alpha} : \alpha_j \geq 0, \sum_{j=1}^{M-1} \alpha_j \in [0, 1]\}$, and let the space of $\boldsymbol{\vartheta}_M$ be $\Theta_{\boldsymbol{\vartheta}_M} := \Theta_\alpha \times \Theta_\gamma \times \Theta_\mu^M \times \Theta_\Sigma^M$.

The number of components M is the smallest number such that the data density admits the representation (1). Our objective is to test

$$H_0 : M = M_0 \quad \text{against} \quad H_A : M = M_0 + 1.$$

3 Quadratic expansion under singular Fisher information matrix

When testing the number of components by the LRT, the Fisher information matrix becomes singular and the log-likelihood function will be approximated by a quadratic function of polynomials of parameters. Further, a part of parameter is not identified under the null hypothesis. We derive a DQM-type expansion for general density that is useful for handling such cases.

Let $\boldsymbol{\vartheta}$ be a parameter vector, and let $f(\mathbf{y}; \boldsymbol{\vartheta})$ denote the density function of \mathbf{y} . Let $L_n(\boldsymbol{\vartheta}) := \sum_{i=1}^n \log f(\mathbf{y}_i; \boldsymbol{\vartheta})$ denote the log-likelihood function. Split $\boldsymbol{\vartheta}$ as $\boldsymbol{\vartheta} = (\boldsymbol{\psi}^\top, \boldsymbol{\pi}^\top)^\top$, and write $L_n(\boldsymbol{\vartheta}) = L_n(\boldsymbol{\psi}, \boldsymbol{\pi})$. $\boldsymbol{\pi}$ corresponds to the part of $\boldsymbol{\vartheta}$ that is not identified under the null. Denote the true

parameter value of $\boldsymbol{\psi}$ by $\boldsymbol{\psi}^*$, and denote the set of $(\boldsymbol{\psi}, \boldsymbol{\pi})$ corresponding to the null hypothesis by $\Gamma^* = \{(\boldsymbol{\psi}, \boldsymbol{\pi}) \in \Theta : \boldsymbol{\psi} = \boldsymbol{\psi}^*\}$. Let $\mathbf{t}(\boldsymbol{\vartheta})$ be a continuous function of $\boldsymbol{\vartheta}$ such that $\mathbf{t}(\boldsymbol{\vartheta}) = 0$ if and only if $\boldsymbol{\psi} = \boldsymbol{\psi}^*$. For $\varepsilon > 0$, define a neighborhood of Γ^* by

$$\mathcal{N}_\varepsilon := \{\boldsymbol{\vartheta} \in \Theta : |\mathbf{t}(\boldsymbol{\vartheta})| < \varepsilon\}.$$

We establish a general quadratic expansion that expresses $L_n(\boldsymbol{\psi}, \boldsymbol{\pi}) - L_n(\boldsymbol{\psi}^*, \boldsymbol{\pi})$ as a quadratic function of $\mathbf{t}(\boldsymbol{\vartheta})$ for $\boldsymbol{\vartheta} \in \mathcal{N}_\varepsilon$. Denote the density ratio by

$$\ell(\mathbf{y}; \boldsymbol{\vartheta}) := \frac{f(\mathbf{y}; \boldsymbol{\psi}, \boldsymbol{\pi})}{f(\mathbf{y}; \boldsymbol{\psi}^*, \boldsymbol{\pi})}, \quad (2)$$

so that $L_n(\boldsymbol{\psi}, \boldsymbol{\pi}) - L_n(\boldsymbol{\psi}^*, \boldsymbol{\pi}) = \sum_{i=1}^n \log \ell(\mathbf{y}_i; \boldsymbol{\vartheta})$. We assume that $\ell(\mathbf{y}; \boldsymbol{\vartheta})$ can be expanded around $\ell(\mathbf{y}; \boldsymbol{\vartheta}^*) = 1$ as follows.

Assumption 1. $\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1$ admits an expansion

$$\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1 = \mathbf{t}(\boldsymbol{\vartheta})^\top \mathbf{s}(\mathbf{y}; \boldsymbol{\pi}) + r(\mathbf{y}; \boldsymbol{\vartheta}), \quad (3)$$

where $\mathbf{s}(\mathbf{y}; \boldsymbol{\pi})$ and $r(\mathbf{y}; \boldsymbol{\vartheta})$ satisfy, for some $C \in (0, \infty)$ and $\varepsilon > 0$, (a) $E \sup_{\boldsymbol{\pi} \in \Theta_\pi} |\mathbf{s}(\mathbf{Y}; \boldsymbol{\pi})|^2 < C$, (b) $\sup_{\boldsymbol{\pi} \in \Theta_\pi} |P_n(\mathbf{s}(\mathbf{y}; \boldsymbol{\pi}) \mathbf{s}(\mathbf{y}; \boldsymbol{\pi})^\top) - \boldsymbol{\mathcal{I}}_\pi| = o_p(1)$ with $\sup_{\boldsymbol{\pi} \in \Theta_\pi} \lambda_{\max}(\boldsymbol{\mathcal{I}}_\pi) < C$, (c) $E[\sup_{\boldsymbol{\vartheta} \in \mathcal{N}_\varepsilon} |r(\mathbf{Y}; \boldsymbol{\vartheta}) / (|\mathbf{t}(\boldsymbol{\vartheta})| |\boldsymbol{\psi} - \boldsymbol{\psi}^*|)^2] < \infty$, (d) $\sup_{\boldsymbol{\vartheta} \in \mathcal{N}_\varepsilon} [\nu_n(r(\mathbf{y}; \boldsymbol{\vartheta})) / (|\mathbf{t}(\boldsymbol{\vartheta})| |\boldsymbol{\psi} - \boldsymbol{\psi}^*|)] = O_p(1)$, (e) $0 < \inf_{\boldsymbol{\pi} \in \Theta_\pi} \lambda_{\min}(\boldsymbol{\mathcal{I}}_\pi)$, (f) $\sup_{\boldsymbol{\pi} \in \Theta_\pi} |\nu_n(\mathbf{s}(\mathbf{y}; \boldsymbol{\pi}))| = O_p(1)$.

We first establish an expansion $L_n(\boldsymbol{\psi}, \boldsymbol{\pi})$ in a neighborhood $\mathcal{N}_{c/\sqrt{n}}$ that holds for any $c > 0$.

Proposition 1. Suppose that Assumption 1(a)–(e) holds. Then, for all $c > 0$,

$$\sup_{\boldsymbol{\vartheta} \in \mathcal{N}_{c/\sqrt{n}}} \left| L_n(\boldsymbol{\psi}, \boldsymbol{\pi}) - L_n(\boldsymbol{\psi}^*, \boldsymbol{\pi}) - \sqrt{n} \mathbf{t}(\boldsymbol{\vartheta})^\top \nu_n(\mathbf{s}(\mathbf{y}; \boldsymbol{\pi})) + n \mathbf{t}(\boldsymbol{\vartheta})^\top \boldsymbol{\mathcal{I}}_\pi \mathbf{t}(\boldsymbol{\vartheta}) / 2 \right| = o_p(1).$$

The next proposition expands $L_n(\boldsymbol{\psi}, \boldsymbol{\pi})$ in $A_n(\delta) := \{\boldsymbol{\vartheta} \in \mathcal{N}_\varepsilon : L_n(\boldsymbol{\psi}, \boldsymbol{\pi}) - L_n(\boldsymbol{\psi}^*, \boldsymbol{\pi}) \geq -\delta\}$ for $\delta \in (0, \infty)$. This proposition is useful for deriving the asymptotic distribution of the LRTS because a consistent MLE is in $A_n(\delta)$ by definition and it is difficult to find a uniform approximation of $L_n(\boldsymbol{\psi}, \boldsymbol{\pi})$ up to an $o_p(1)$ term in \mathcal{N}_ε .

Proposition 2. Suppose that Assumption 1 holds. Then, for any $\delta > 0$, (a) $\sup_{\boldsymbol{\vartheta} \in A_n(\delta)} |\mathbf{t}(\boldsymbol{\vartheta})| = O_p(n^{-1/2})$;

$$(b) \sup_{\boldsymbol{\vartheta} \in A_n(\delta)} \left| L_n(\boldsymbol{\psi}, \boldsymbol{\pi}) - L_n(\boldsymbol{\psi}^*, \boldsymbol{\pi}) - \sqrt{n} \mathbf{t}(\boldsymbol{\vartheta})^\top \nu_n(\mathbf{s}(\mathbf{y}; \boldsymbol{\pi})) + n \mathbf{t}(\boldsymbol{\vartheta})^\top \boldsymbol{\mathcal{I}}_\pi \mathbf{t}(\boldsymbol{\vartheta}) / 2 \right| = o_p(1).$$

4 Local quadratic approximation for testing $H_0 : M = 1$ against $H_A : M = 2$

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

$$f_2(\mathbf{x}|\mathbf{z}; \boldsymbol{\vartheta}_2) := \alpha f(\mathbf{x}|\mathbf{z}; \gamma, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1 - \alpha)f(\mathbf{x}|\mathbf{z}; \gamma, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2). \quad (4)$$

The model (4) yields the true density $f(\mathbf{x}|\mathbf{z}; \gamma^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ if $\boldsymbol{\vartheta}_2$ lies in the set $\Theta_2^* := \{\boldsymbol{\vartheta}_2 \in \Theta_{\boldsymbol{\vartheta}_2} : \{(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = (\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \gamma = \gamma^*\} \text{ or } \{\alpha(1 - \alpha) = 0, \gamma = \gamma^*\}\}$.

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

$$H_{01} : (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \text{ and } H_{02} : \alpha(1 - \alpha) = 0.$$

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

In addition to the failure of regularity conditions that is common to all finite mixture models, the normal mixture model (4) has additional undesirable mathematical properties, as discussed in Chen and Li (2009): (a) The Fisher information for testing H_{02} is not finite unless the range of $\det(\boldsymbol{\Sigma}_1)/\det(\boldsymbol{\Sigma}_2)$ is restricted. (b) The derivatives of $f_2(y|\mathbf{x}, \mathbf{z}; \boldsymbol{\vartheta}_2)$ of different orders are linearly dependent because $\nabla_{\mu_i \mu_j} f(y|\mathbf{x}, \mathbf{z}; \gamma, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = 2\nabla_{\Sigma_{ij}} f(y|\mathbf{x}, \mathbf{z}; \gamma, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ (loss of strong identifiability). (c) The log-likelihood function is unbounded and the maximum likelihood estimate fails to exist (Hartigan, 1985; Kiefer and Wolfowitz, 1956).

In view of problem (a), we focus on testing $H_{01} : (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ in the following. We handle problem (c) by considering a maximum penalized likelihood estimator (PMLE) introduced by Chen and Tan (2009) and Alexandrovich (2014). Let \mathbf{S}_x denote the sample covariance matrix. Similar to Chen and Tan (2009), we use the following penalty function

$$p_n(\boldsymbol{\vartheta}_M) = \sum_{m=1}^M p_{nm}(\boldsymbol{\Sigma}_m) = \sum_{m=1}^M -a_n \{ \text{tr}(\mathbf{S}_x \boldsymbol{\Sigma}_m^{-1}) - 2 \log(\det(\mathbf{S}_x \boldsymbol{\Sigma}_m^{-1})) - d \}, \quad (5)$$

with $M = 2$, where a_n is non-random. Note that $p_n(\boldsymbol{\vartheta}_2) = 0$ if $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{S}_x$. Let $\hat{\boldsymbol{\vartheta}}_2$ denote the PMLE that maximizes $PL_n(\boldsymbol{\vartheta}_2) := \sum_{i=1}^n f_2(\mathbf{X}_i|\mathbf{Z}_i; \boldsymbol{\vartheta}_2) + p_n(\boldsymbol{\vartheta}_2)$. The following assumption on the penalty function is adopted from Chen and Tan (2009), Chen and Li (2009), and Alexandrovich (2014).

Assumption 2. (a) For any fixed Σ such that $\det(\Sigma) > 0$, we have $p_{nm}(\Sigma) = o(n)$, and $\sup_{\Sigma} \max\{0, p_{nm}(\Sigma)\} = o(n)$. (b) $\nabla_{\text{vec}(\Sigma)} p_{nm}(\Sigma) = o(n^{1/6})$ at any fixed Σ such that $\det(\Sigma) > 0$. (c) For sufficiently large n , $p_n(\Sigma) \leq (3/4)\sqrt{n \log \log(n)} \log(\det(\Sigma))$ for $\det(\Sigma) \leq Cn^{-2d}$.

Assumption 3. \mathbf{Z} has finite second moment, and $\Pr(\mathbf{Z}_i^\top \gamma \neq \mathbf{Z}_i^\top \gamma^*) > 0$ for any $\gamma \neq \gamma^*$.

The following proposition shows the consistency of $\hat{\boldsymbol{\vartheta}}_2$.

Proposition 3. Suppose that Assumptions 2 and 3 hold. Then, under the null hypothesis $H_0 : M = 1$, $\inf_{\boldsymbol{\vartheta}_2 \in \Theta_2^*} |\hat{\boldsymbol{\vartheta}}_2 - \boldsymbol{\vartheta}_2| \rightarrow_p 0$.

For any $\bar{\boldsymbol{\vartheta}}_2$ such that $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, the derivatives of the density are linearly dependent as

$$\nabla_{\boldsymbol{\mu}_1} f_2(\mathbf{x}|\mathbf{z}; \bar{\boldsymbol{\vartheta}}_2) = \frac{\alpha}{1-\alpha} \nabla_{\boldsymbol{\mu}_2} f_2(\mathbf{x}|\mathbf{z}; \bar{\boldsymbol{\vartheta}}_2), \quad \nabla_{\boldsymbol{\Sigma}_1} f_2(\mathbf{x}|\mathbf{z}; \bar{\boldsymbol{\vartheta}}_2) = \frac{\alpha}{1-\alpha} \nabla_{\boldsymbol{\Sigma}_2} l(y|\mathbf{x}, \mathbf{z}; \bar{\boldsymbol{\vartheta}}_2), \quad (6)$$

$$\nabla_{\mu_{1i}\mu_{1j}} f_2(\mathbf{x}|\mathbf{z}; \bar{\boldsymbol{\vartheta}}_2) = 2\nabla_{\Sigma_{1,ij}} f_2(\mathbf{x}|\mathbf{z}; \bar{\boldsymbol{\vartheta}}_2), \quad \nabla_{\mu_{2i}\mu_{2j}} f_2(\mathbf{x}|\mathbf{z}; \bar{\boldsymbol{\vartheta}}_2) = 2\nabla_{\Sigma_{2,ij}} f_2(\mathbf{x}|\mathbf{z}; \bar{\boldsymbol{\vartheta}}_2). \quad (7)$$

Consequently, the Fisher information matrix is degenerate, which invalidates the standard second-order quadratic approximation analysis. In particular, dependence (7) leads to the loss of strong identifiability and causes substantial difficulties in existing literature.

We analyze the penalized LRT statistic for testing $H_{01} : (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ by developing a higher-order approximation of the log-likelihood function that can be expressed in a quadratic form, when $\alpha \in (0, 1)$, through a judiciously designed reparameterization. This reparameterization extends the result of Rotnitzky et al. (2000) and Kasahara and Shimotsu (2015). Collect the unique elements in $\boldsymbol{\Sigma}$ into a $d(d+1)/2$ -vector

$$\begin{aligned} \mathbf{v} &= (v_{11}, v_{12}, \dots, v_{1d}, v_{22}, v_{23}, \dots, v_{2d}, \dots, v_{d-1,d-1}, v_{d-1,d}, v_{dd})^\top \\ &:= (\Sigma_{11}, 2\Sigma_{12}, \dots, 2\Sigma_{1d}, \Sigma_{22}, 2\Sigma_{23}, \dots, 2\Sigma_{2d}, \dots, \Sigma_{d-1,d-1}, 2\Sigma_{d-1,d}, \Sigma_{dd})^\top. \end{aligned}$$

Define the density function of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ parameterized in terms of $\boldsymbol{\mu}$ and \mathbf{v} as

$$f_v(\boldsymbol{\mu}, \mathbf{v}) := f(\boldsymbol{\mu}, \mathbf{S}(\mathbf{v})), \quad \text{where } S_{ij}(\mathbf{v}) := \begin{cases} v_{ii} & \text{if } i = j, \\ v_{ij}/2 & \text{if } i \neq j. \end{cases} \quad (8)$$

For a $d \times d$ symmetric matrix \mathbf{A} , define a function $\mathbf{w}(\mathbf{A}) \in \mathbb{R}^{d(d+1)/2}$ as

$$\mathbf{w}(\mathbf{A}) := (A_{11}, 2A_{12}, \dots, 2A_{1d}, A_{22}, 2A_{23}, \dots, 2A_{2d}, \dots, A_{d-1,d-1}, 2A_{d-1,d}, A_{dd})^\top.$$

Then $f_v(\boldsymbol{\mu}, \mathbf{v})$ and $f(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are related as

$$f(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = f_v(\boldsymbol{\mu}, \mathbf{w}(\boldsymbol{\Sigma})).$$

We introduce the following one-to-one mapping between $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{v}_1, \boldsymbol{v}_2)$ and the reparameterized parameter $(\boldsymbol{\lambda}_\mu, \boldsymbol{\nu}_\mu, \boldsymbol{\lambda}_v, \boldsymbol{\nu}_v)$:

$$\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\nu}_\mu + (1 - \alpha)\boldsymbol{\lambda}_\mu \\ \boldsymbol{\nu}_\mu - \alpha\boldsymbol{\lambda}_\mu \\ \boldsymbol{\nu}_v + (1 - \alpha)(2\boldsymbol{\lambda}_v + C_1\boldsymbol{w}(\boldsymbol{\lambda}_\mu\boldsymbol{\lambda}_\mu^\top)) \\ \boldsymbol{\nu}_v - \alpha(2\boldsymbol{\lambda}_v + C_2\boldsymbol{w}(\boldsymbol{\lambda}_\mu\boldsymbol{\lambda}_\mu^\top)) \end{pmatrix}, \quad (9)$$

where $C_1 := -(1/3)(1 + \alpha)$ and $C_2 := (1/3)(2 - \alpha)$. Collect the reparameterized parameters, except for α , into one vector $\boldsymbol{\psi}_\alpha$ defined as

$$\boldsymbol{\psi}_\alpha := (\boldsymbol{\gamma}, \boldsymbol{\nu}_\mu, \boldsymbol{\nu}_v, \boldsymbol{\lambda}_\mu, \boldsymbol{\lambda}_v) \in \Theta_{\boldsymbol{\psi}_\alpha}. \quad (10)$$

In the reparameterized model, the null hypothesis of $H_{01} : (\boldsymbol{\mu}_1, \boldsymbol{v}_1) = (\boldsymbol{\mu}_2, \boldsymbol{v}_2)$ is written as $H_{01} : (\boldsymbol{\lambda}_\mu, \boldsymbol{\lambda}_v) = \mathbf{0}$, and the density and its logarithm are given by

$$\begin{aligned} g(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\psi}_\alpha, \alpha) &= \alpha f_v \left(\boldsymbol{x} \middle| \boldsymbol{z}; \boldsymbol{\gamma}, \boldsymbol{\nu}_\mu + (1 - \alpha)\boldsymbol{\lambda}_\mu, \boldsymbol{\nu}_v + (1 - \alpha)(2\boldsymbol{\lambda}_v + C_1\boldsymbol{w}(\boldsymbol{\lambda}_\mu\boldsymbol{\lambda}_\mu^\top)) \right) \\ &+ (1 - \alpha) f_v \left(\boldsymbol{x} \middle| \boldsymbol{z}; \boldsymbol{\gamma}, \boldsymbol{\nu}_\mu - \alpha\boldsymbol{\lambda}_\mu, \boldsymbol{\nu}_v - \alpha(2\boldsymbol{\lambda}_v + C_2\boldsymbol{w}(\boldsymbol{\lambda}_\mu\boldsymbol{\lambda}_\mu^\top)) \right). \end{aligned} \quad (11)$$

Partition $\boldsymbol{\psi}_\alpha$ as $\boldsymbol{\psi}_\alpha = (\boldsymbol{\eta}^\top, \boldsymbol{\lambda}^\top)^\top$, where $\boldsymbol{\eta} := (\boldsymbol{\gamma}^\top, \boldsymbol{\nu}_\mu^\top, \boldsymbol{\nu}_v^\top)^\top \in \Theta_\eta$ and $\boldsymbol{\lambda} := (\boldsymbol{\lambda}_\mu^\top, \boldsymbol{\lambda}_v^\top)^\top \in \Theta_\lambda$. Denote the true values of $\boldsymbol{\eta}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\psi}$ by $\boldsymbol{\eta}^* := ((\boldsymbol{\gamma}^*)^\top, (\boldsymbol{\mu}^*)^\top, (\boldsymbol{v}^*)^\top)^\top$, $\boldsymbol{\lambda}^* := \mathbf{0}$, and $\boldsymbol{\psi}_\alpha^* = ((\boldsymbol{\eta}^*)^\top, \mathbf{0}^\top)^\top$, respectively. The first derivative of (11) w.r.t. $\boldsymbol{\eta}$ under $\boldsymbol{\psi}_\alpha = \boldsymbol{\psi}_\alpha^*$ is identical to the first derivative of the density of the one-component model:

$$\nabla_{\boldsymbol{\eta}} g(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\psi}_\alpha^*, \alpha) = \nabla_{(\boldsymbol{\gamma}^\top, \boldsymbol{\mu}^\top, \boldsymbol{v}^\top)^\top} f_v(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\gamma}^*, \boldsymbol{\mu}^*, \boldsymbol{v}^*). \quad (12)$$

On the other hand, Lemma 3 in the appendix shows that the first, second, and third derivatives of $g(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\psi}_\alpha, \alpha)$ w.r.t. $\boldsymbol{\lambda}_\mu$ and the first derivative w.r.t. $\boldsymbol{\lambda}_v$ become zero when evaluated at $\boldsymbol{\psi}_\alpha = \boldsymbol{\psi}_\alpha^*$:

$$\begin{aligned} \nabla_{\boldsymbol{\lambda}_\mu} g(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\psi}_\alpha^*, \alpha) &= \mathbf{0}, \quad \nabla_{\boldsymbol{\lambda}_\mu^{\otimes 2}} g(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\psi}_\alpha^*, \alpha) = \mathbf{0}, \quad \nabla_{\boldsymbol{\lambda}_\mu^{\otimes 3}} g(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\psi}_\alpha^*, \alpha) = \mathbf{0}, \\ \nabla_{\boldsymbol{\lambda}_v} g(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\psi}_\alpha^*, \alpha) &= \mathbf{0}. \end{aligned} \quad (13)$$

Consequently, the information on $\boldsymbol{\lambda}_\mu$ and $\boldsymbol{\lambda}_v$ is provided by the fourth derivative w.r.t. $\boldsymbol{\lambda}_\mu$, the cross-derivative w.r.t. $\boldsymbol{\lambda}_\mu$ and $\boldsymbol{\lambda}_v$, and the second derivative w.r.t. $\boldsymbol{\lambda}_v$.

We derive a quadratic approximation of the log-likelihood function by applying Proposition 2. To this end, we collect the relevant score vector and reparameterized parameters that correspond to $\boldsymbol{s}(\boldsymbol{y}; \boldsymbol{\pi})$ and $\boldsymbol{t}(\boldsymbol{\vartheta})$ in Assumption 1. Let f_v^* and ∇f_v^* denote $f_v(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\gamma}^*, \boldsymbol{\mu}^*, \boldsymbol{v}^*)$ and $\nabla f_v(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\gamma}^*, \boldsymbol{\mu}^*, \boldsymbol{v}^*)$, and let $d_\eta := (p + d + d(d + 1)/2)$, $d_{\mu v} := d(d + 1)(d + 2)/6$, and

$d_{\mu^4} := d(d+1)(d+2)(d+3)/24$. Define the score vector $\mathbf{s}(\mathbf{x}, \mathbf{z})$ as

$$\begin{aligned} \mathbf{s}(\mathbf{x}, \mathbf{z}) &:= \begin{pmatrix} \mathbf{s}_\eta \\ \mathbf{s}_{\mu v} \\ \mathbf{s}_{\mu^4} \end{pmatrix} := \begin{pmatrix} \mathbf{s}_\eta \\ \mathbf{s}_{\mu v} \\ \mathbf{s}_{\mu^4} \end{pmatrix} \quad \text{with} \quad \mathbf{s}_\eta \underset{(d_\mu \times 1)}{:=} \frac{\nabla(\boldsymbol{\gamma}^\top, \boldsymbol{\mu}^\top, \mathbf{v}^\top)^\top f_v^*}{f_v^*}, \\ \mathbf{s}_{\mu v} \underset{(d_{\mu v} \times 1)}{:=} &\left\{ \frac{\nabla_{\mu_i \mu_j \mu_k} f_v^*}{f_v^*} \right\}_{1 \leq i \leq j \leq k \leq d}, \quad \mathbf{s}_{\mu^4} \underset{(d_{\mu^4} \times 1)}{:=} \left\{ \frac{\nabla_{\mu_i \mu_j \mu_k \mu_\ell} f_v^*}{f_v^*} \right\}_{1 \leq i \leq j \leq k \leq \ell \leq d}, \end{aligned} \quad (14)$$

where we suppress the dependence of $(\mathbf{s}_\eta, \mathbf{s}_{\mu v}, \mathbf{s}_{\mu^4})$ on (\mathbf{x}, \mathbf{z}) . Collect the relevant reparameterized parameters as

$$\mathbf{t}(\boldsymbol{\psi}_\alpha, \alpha) := \begin{pmatrix} \boldsymbol{\eta} - \boldsymbol{\eta}^* \\ \mathbf{t}(\boldsymbol{\lambda}, \alpha) \end{pmatrix} := \begin{pmatrix} \boldsymbol{\eta} - \boldsymbol{\eta}^* \\ \boldsymbol{\lambda}_{\mu v} \\ \boldsymbol{\lambda}_{\mu^4} \end{pmatrix}, \quad (15)$$

with

$$\begin{aligned} \boldsymbol{\lambda}_{\mu v} \underset{(d_{\mu v} \times 1)}{=} &\{(\boldsymbol{\lambda}_{\mu v})_{ijk}\}_{1 \leq i \leq j \leq k \leq d}, \quad \text{where } (\boldsymbol{\lambda}_{\mu v})_{ijk} := \alpha(1-\alpha) \sum_{(t_1, t_2, t_3) \in p_{12}(i, j, k)} \lambda_{\mu_{t_1}} \lambda_{v_{t_2 t_3}}, \\ \boldsymbol{\lambda}_{\mu^4} \underset{(d_{\mu^4} \times 1)}{=} &\{(\boldsymbol{\lambda}_{\mu^4})_{ijkl}\}_{1 \leq i \leq j \leq k \leq \ell \leq d}, \quad \text{where } (\boldsymbol{\lambda}_{\mu^4})_{ijkl} := \alpha(1-\alpha) \\ &\times \left\{ 12 \sum_{(t_1, t_2, t_3, t_4) \in p_{22}(i, j, k, \ell)} \lambda_{v_{t_1 t_2}} \lambda_{v_{t_3 t_4}} + b(\alpha) \sum_{(t_1, t_2, t_3, t_4) \in p(i, j, k, \ell)} \lambda_{\mu_{t_1}} \lambda_{\mu_{t_2}} \lambda_{\mu_{t_3}} \lambda_{\mu_{t_4}} \right\}, \end{aligned} \quad (16)$$

where $b(\alpha) := -(2/3)(\alpha^2 - \alpha + 1) < 0$, and $\sum_{(t_1, t_2, t_3) \in p_{12}(i, j, k)}$ denotes the sum over all distinct permutations of (i, j, k) to (t_1, t_2, t_3) with $t_2 \leq t_3$, $\sum_{(t_1, t_2, t_3, t_4) \in p_{22}(i, j, k, \ell)}$ denotes the sum over all distinct permutations of (i, j, k, ℓ) to (t_1, t_2, t_3, t_4) with $t_1 \leq t_2$ and $t_3 \leq t_4$, and $\sum_{(t_1, t_2, t_3, t_4) \in p(i, j, k, \ell)}$ denotes the sum over all distinct permutations of (i, j, k, ℓ) to (t_1, t_2, t_3, t_4) .

Let $L_n(\boldsymbol{\psi}_\alpha, \alpha) := \sum_{i=1}^n \log g(\mathbf{X}_i | \mathbf{Z}_i; \boldsymbol{\psi}_\alpha, \alpha)$ denote the reparameterized log-likelihood function. Define $A_{n\alpha}(\delta) := \{\boldsymbol{\vartheta} \in \mathcal{N}_\varepsilon : L_n(\boldsymbol{\psi}_\alpha, \alpha) - L_n(\boldsymbol{\psi}_\alpha^*, \alpha) \geq -\delta\}$.

Assumption 4. (a) \mathbf{Z} has finite tenth moment.

Proposition 4. Suppose that Assumptions 3 and 4 hold. Then, under the null hypothesis $H_0 : m = 1$, for $\alpha \in (0, 1)$ and $\epsilon_\sigma \in (0, 1)$ and any $\delta > 0$, we have (a) $\sup_{\boldsymbol{\vartheta} \in A_{n\alpha}(\delta)} |\mathbf{t}(\boldsymbol{\psi}_\alpha, \alpha)| = O_p(n^{-1/2})$;

(b) $\sup_{\boldsymbol{\vartheta} \in A_{n\alpha}(\delta)} \left| L_n(\boldsymbol{\psi}_\alpha, \alpha) - L_n(\boldsymbol{\psi}_\alpha^*, \alpha) - \sqrt{n} \mathbf{t}(\boldsymbol{\psi}_\alpha, \alpha)^\top \nu_n(\mathbf{s}(\mathbf{x}, \mathbf{z})) + n \mathbf{t}(\boldsymbol{\psi}_\alpha, \alpha)^\top \boldsymbol{\mathcal{I}} \mathbf{t}(\boldsymbol{\psi}_\alpha, \alpha) / 2 \right| = o_p(1)$,

where $\mathbf{t}(\boldsymbol{\psi}_\alpha, \alpha)$ and $\mathbf{s}(\mathbf{x}, \mathbf{z})$ are defined in (15) and (14), and $\boldsymbol{\mathcal{I}} := E[\mathbf{s}(\mathbf{X}, \mathbf{Z})\mathbf{s}(\mathbf{X}, \mathbf{Z})^\top]$.

Let $\widehat{\boldsymbol{\psi}}_\alpha := \arg \max_{\boldsymbol{\psi}_\alpha \in \Theta_{\boldsymbol{\psi}_\alpha}(\epsilon_\sigma)} L_n(\boldsymbol{\psi}_\alpha, \alpha)$ denote the (constrained) MLE of $\boldsymbol{\psi}_\alpha$, where $\Theta_{\boldsymbol{\psi}_\alpha}(\epsilon_\sigma)$ is defined so that the value of $\boldsymbol{\vartheta}_2$ implied by $\boldsymbol{\psi}_\alpha$ is in $\Theta_{\boldsymbol{\vartheta}_2}(\epsilon_\sigma)$. Let $(\widehat{\boldsymbol{\gamma}}_0, \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\Sigma}}_0)$ denote the one-component MLE that maximizes the one-component log-likelihood function $L_{0,n}(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) :=$

$\sum_{i=1}^n \log f(\mathbf{X}_i | \mathbf{Z}_i; \gamma, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Define the penalized LRT statistic for testing H_{01} as $PLR_n(\epsilon_1) := \max_{\alpha \in [\epsilon_1, 1-\epsilon_1]} 2\{PL_n(\widehat{\boldsymbol{\psi}}_\alpha, \alpha) - L_{0,n}(\widehat{\boldsymbol{\gamma}}_0, \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\Sigma}}_0)\}$ with $\epsilon_1 \in (0, 1/2)$.

We proceed to derive the asymptotic distribution of the LRTS. With $(\mathbf{s}_\eta, \mathbf{s}_\lambda)$ defined in (14), define

$$\begin{aligned} \mathcal{I}_\eta &:= E[\mathbf{s}_\eta \mathbf{s}_\eta^\top], & \mathcal{I}_\lambda &:= E[\mathbf{s}_\lambda \mathbf{s}_\lambda^\top], & \mathcal{I}_{\lambda\eta} &:= E[\mathbf{s}_\lambda \mathbf{s}_\eta^\top], \\ \mathcal{I}_{\eta\lambda} &:= \mathcal{I}_{\lambda\eta}^\top, & \mathcal{I}_{\lambda,\eta} &:= \mathcal{I}_\lambda - \mathcal{I}_{\lambda\eta} \mathcal{I}_\eta^{-1} \mathcal{I}_{\eta\lambda}, & \mathbf{Z}_\lambda &:= (\mathcal{I}_{\lambda,\eta})^{-1} \mathbf{G}_{\lambda,\eta}, \end{aligned} \quad (17)$$

where $\mathbf{G}_{\lambda,\eta} \sim N(0, \mathcal{I}_{\lambda,\eta})$. The following set characterizes the limit of possible values of $\sqrt{n}\mathbf{t}(\boldsymbol{\lambda}, \alpha)$ defined in (15) as $n \rightarrow \infty$. For $\mathbf{e} = (e_1, \dots, e_d)^\top \in \{0, 1\}^d$, define

$$\Lambda_\lambda^{\mathbf{e}} := \left(\{(\mathbf{t}_{\boldsymbol{\mu}v}^{\mathbf{e}})_{ijk}\}_{1 \leq i \leq j \leq k \leq d}, \{(\mathbf{t}_{\boldsymbol{\mu}^4}^{\mathbf{e}})_{ijkl}\}_{1 \leq i \leq j \leq k \leq \ell \leq d} \right)^\top \in \mathbb{R}^{d_{\mu v} + d_{\mu^4}}, \quad (18)$$

where $\mathbf{t}_{\boldsymbol{\mu}v}^{\mathbf{e}}$ and $\mathbf{t}_{\boldsymbol{\mu}^4}^{\mathbf{e}}$ satisfy, for some $\boldsymbol{\lambda}_\mu \in \mathbb{R}^d$ and $\{\lambda_{v_{ab}}\}_{1 \leq a \leq b \leq d} \in \mathbb{R}^{d(d+1)/2}$,

$$\begin{aligned} (\mathbf{t}_{\boldsymbol{\mu}v}^{\mathbf{e}})_{ijk} &= \sum_{(t_1, t_2, t_3) \in p_{12}(i, j, k)} \lambda_{\mu_{t_1}} \lambda_{v_{t_2 t_3}} \quad \text{for all } \mathbf{e}, \\ (\mathbf{t}_{\boldsymbol{\mu}^4}^{\mathbf{e}})_{ijkl} &= \begin{cases} \sum_{(t_1, t_2, t_3, t_4) \in p_{22}(i, j, k, \ell)} \lambda_{v_{t_1 t_2}} \lambda_{v_{t_3 t_4}} & \text{if } \mathbf{e} = \mathbf{0}, \\ \sum_{(t_1, t_2, t_3, t_4) \in p(i, j, k, \ell)} \lambda_{\mu_{t_1}} \lambda_{\mu_{t_2}} \lambda_{\mu_{t_3}} \lambda_{\mu_{t_4}} & \text{if } \mathbf{e} \neq \mathbf{0} \text{ and } e_i = e_j = e_k = e_\ell = 1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (19)$$

Define $\widehat{\mathbf{t}}_\lambda^{\mathbf{e}}$ by

$$r(\widehat{\mathbf{t}}_\lambda^{\mathbf{e}}) = \inf_{\mathbf{t}_\lambda \in \Lambda_\lambda^{\mathbf{e}}} r(\mathbf{t}_\lambda), \quad r(\mathbf{t}_\lambda) := (\mathbf{t}_\lambda - \mathbf{Z}_\lambda)^\top \mathcal{I}_{\lambda,\eta} (\mathbf{t}_\lambda - \mathbf{Z}_\lambda). \quad (20)$$

The following proposition establishes the asymptotic null distribution of the penalized LRT statistic. This result follows from the local quadratic approximation established in Proposition 4 and the results in Andrews (1999).

Proposition 5. *Suppose that Assumptions 2, 3, and 4 hold. Then, under the null hypothesis of $M = 1$, $LR_n(\epsilon_1) \rightarrow_d \max_{\mathbf{e} \in \{0,1\}^d} \left(\widehat{\mathbf{t}}_\lambda^{\mathbf{e}} \right)^\top \mathcal{I}_{\lambda,\eta} \widehat{\mathbf{t}}_\lambda^{\mathbf{e}}$.*

For each \mathbf{e} , the random variable $(\widehat{\mathbf{t}}_\lambda^{\mathbf{e}})^\top \mathcal{I}_{\lambda,\eta} \widehat{\mathbf{t}}_\lambda^{\mathbf{e}}$ is a projection of a Gaussian random variable on a cone $\Lambda_\lambda^{\mathbf{e}}$.

5 Local quadratic approximation for testing $H_0 : M = M_0$ against $H_A : M = M_0 + 1$ for $M_0 \geq 2$

In this section, we develop a local quadratic approximation for testing the null hypothesis of M_0 components against the alternative of $M_0 + 1$ components for general $M_0 \geq 1$. We consider a random sample of n independent observations $\{\mathbf{X}_i, \mathbf{Z}_i\}_{i=1}^n$ generated from the M_0 -component d -variate normal mixture density with the true parameter value $\boldsymbol{\vartheta}_{M_0}^* =$

$(\alpha_1^*, \dots, \alpha_{M_0-1}^*, \gamma^*, \boldsymbol{\mu}_1^*, \dots, \boldsymbol{\mu}_M^*, \boldsymbol{\Sigma}_1^*, \dots, \boldsymbol{\Sigma}_M^*)$:

$$f_{M_0}(\mathbf{x}|\mathbf{z}; \boldsymbol{\vartheta}_{M_0}^*) := \sum_{j=1}^{M_0} \alpha_j^* f(\mathbf{x}|\mathbf{z}; \gamma^*, \boldsymbol{\mu}_j^*, \boldsymbol{\Sigma}_j^*), \quad (21)$$

where $\alpha_j^* > 0$. We assume $(\boldsymbol{\mu}_1^*, \boldsymbol{\Sigma}_1^*) < \dots < (\boldsymbol{\mu}_{M_0}^*, \boldsymbol{\Sigma}_{M_0}^*)$ for identification. Let the density of an $(M_0 + 1)$ -component mixture model be

$$f_{M_0+1}(\mathbf{x}|\mathbf{z}; \boldsymbol{\vartheta}_{M_0+1}) := \sum_{j=1}^{M_0+1} \alpha_j f(\mathbf{x}|\mathbf{z}; \gamma, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \quad (22)$$

where $\boldsymbol{\vartheta}_{M_0+1} = (\alpha_1, \dots, \alpha_{M_0}, \gamma, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{M_0+1}, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_{M_0+1})$. Similar to the case of the test of homogeneity, we partition the null hypothesis into two as $H_0 = H_{01} \cup H_{02}$, where $H_{01} := \cup_{m=1}^{M_0} H_{0,1m}$ and $H_{02} := \cup_{m=1}^{M_0+1} H_{0,2m}$ with

$$H_{0,1m} : (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) < \dots < (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = (\boldsymbol{\mu}_{m+1}, \boldsymbol{\Sigma}_{m+1}) < \dots < (\boldsymbol{\mu}_{M_0+1}, \boldsymbol{\Sigma}_{M_0+1}) \text{ and } H_{0,2m} : \alpha_m = 0.$$

The inequality constraints are imposed on $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ for identification.

As discussed in Kasahara and Shimotsu (2015), the LRT statistic for testing H_{02} has infinite Fisher information unless a stringent restriction is imposed on the admissible values of $\boldsymbol{\Sigma}_j$. Therefore, we focus on testing H_{01} . Define the set of values of $\boldsymbol{\vartheta}_{M_0+1}$ that yields the true density (21) as $\Upsilon^* := \{\boldsymbol{\vartheta}_{M_0+1} : f_{M_0+1}(\mathbf{X}|\mathbf{Z}; \boldsymbol{\vartheta}_{M_0+1}) = f_{M_0}(\mathbf{X}|\mathbf{Z}; \boldsymbol{\vartheta}_{M_0}^*) \text{ with probability one}\}$. Under $H_{0,1m}$, the $(M_0 + 1)$ -component model (22) generates the true M_0 -component density (21) when $(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = (\boldsymbol{\mu}_{m+1}, \boldsymbol{\Sigma}_{m+1}) = (\boldsymbol{\mu}_m^*, \boldsymbol{\Sigma}_m^*)$. Define the subset of Υ^* corresponding to $H_{0,1m}$ as

$$\begin{aligned} \Upsilon_{1m}^* := & \left\{ \boldsymbol{\vartheta}_{M_0+1} \in \Theta_{\boldsymbol{\vartheta}_{M_0+1}} : \alpha_j > 0 \text{ for } j = 1, \dots, M_0 + 1; \alpha_m + \alpha_{m+1} = \alpha_m^* \text{ and} \right. \\ & (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = (\boldsymbol{\mu}_{m+1}, \boldsymbol{\Sigma}_{m+1}) = (\boldsymbol{\mu}_m^*, \boldsymbol{\Sigma}_m^*); \alpha_j = \alpha_j^* \text{ and } (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = (\boldsymbol{\mu}_j^*, \boldsymbol{\Sigma}_j^*) \text{ for } j < m; \\ & \left. \alpha_j = \alpha_{j-1}^* \text{ and } (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = (\boldsymbol{\mu}_{j-1}^*, \boldsymbol{\Sigma}_{j-1}^*) \text{ for } j > m + 1; \gamma = \gamma^* \right\}, \end{aligned}$$

and define $\Upsilon_1^* := \Upsilon_{11}^* \cup \dots \cup \Upsilon_{1M_0}^*$.

Similar to Section 4, we consider the penalized MLE. Let $\Theta_{\boldsymbol{\vartheta}_{M_0+1}}(\epsilon_1)$ be a subset of $\Theta_{\boldsymbol{\vartheta}_{M_0+1}}$ such that $\alpha_j \in [\epsilon_1, 1 - \epsilon_1]$ for $j = 1, \dots, M_0 + 1$, and define the penalized LRT statistic for testing H_{01}

$$PLR_n^{M_0}(\epsilon_1) := \max_{\boldsymbol{\vartheta}_{M_0+1} \in \Theta_{\boldsymbol{\vartheta}_{M_0+1}}(\epsilon_1)} 2\{PL_n(\boldsymbol{\vartheta}_{M_0+1}) - PL_{0,n}(\widehat{\boldsymbol{\vartheta}}_{M_0})\},$$

where $PL_n(\boldsymbol{\vartheta}_{M_0+1}) := \sum_{i=1}^n \log f_{M_0+1}(\mathbf{X}_i|\mathbf{Z}_i; \boldsymbol{\vartheta}_{M_0+1}) + p_n(\boldsymbol{\vartheta}_{M_0+1})$, $PL_{0,n}(\boldsymbol{\vartheta}_{M_0}) = \sum_{i=1}^n \log f_{M_0}(\mathbf{X}_i|\mathbf{Z}_i; \boldsymbol{\vartheta}_{M_0}) + p_n(\boldsymbol{\vartheta}_{M_0})$, and $\widehat{\boldsymbol{\vartheta}}_{M_0} = \arg \max_{\boldsymbol{\vartheta}_{M_0} \in \Theta_{\boldsymbol{\vartheta}_{M_0}}} PL_{0,n}(\boldsymbol{\vartheta}_{M_0})$. Collect the score vector for testing

$H_{0,11}, \dots, H_{0,1M_0}$ into one vector as

$$\tilde{\mathbf{s}}(\mathbf{x}, \mathbf{z}) := \begin{pmatrix} \tilde{\mathbf{s}}_\eta \\ \tilde{\mathbf{s}}_\lambda \end{pmatrix}, \quad \text{where } \tilde{\mathbf{s}}_\eta := \begin{pmatrix} \mathbf{s}_\alpha \\ \mathbf{s}_\gamma \\ \mathbf{s}_{(\mu,v)} \end{pmatrix} \text{ and } \tilde{\mathbf{s}}_\lambda := \begin{pmatrix} \mathbf{s}_{\mu v}^1 \\ \mathbf{s}_{\mu^4}^1 \\ \vdots \\ \mathbf{s}_{\mu v}^{M_0} \\ \mathbf{s}_{\mu^4}^{M_0} \end{pmatrix}, \quad (23)$$

where, with $f_0^* := f_{M_0}(\mathbf{x}|\mathbf{z}; \boldsymbol{\nu}_{M_0}^*)$ and for $m = 1, \dots, M_0$,

$$\begin{aligned} \mathbf{s}_\alpha &:= \begin{pmatrix} f_v(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}^*, \boldsymbol{\mu}_1^*, \mathbf{v}_1^*) - f_v(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}^*, \boldsymbol{\mu}_{M_0}^*, \mathbf{v}_{M_0}^*) \\ \vdots \\ f_v(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}^*, \boldsymbol{\mu}_{M_0-1}^*, \mathbf{v}_{M_0-1}^*) - f_v(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}^*, \boldsymbol{\mu}^*, \mathbf{v}_{M_0}^*) \end{pmatrix} / f_0^*, \\ \mathbf{s}_\gamma &:= \sum_{m=1}^{M_0} \alpha_m^* \nabla_\gamma f_v(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}^*, \boldsymbol{\mu}_m^*, \mathbf{v}_m^*) / f_0^*, \\ \mathbf{s}_{\mu v}^m &:= \{ \alpha_m^* \nabla_{\mu_i \mu_j \mu_k} f_v^*(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}^*, \boldsymbol{\mu}_m^*, \mathbf{v}_m^*) / f_0^* \}_{1 \leq i \leq j \leq k \leq d}, \\ \mathbf{s}_{\mu^4}^m &:= \{ \alpha_m^* \nabla_{\mu_i \mu_j \mu_k \mu_\ell} f_v^*(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}^*, \boldsymbol{\mu}_m^*, \mathbf{v}_m^*) / f_0^* \}_{1 \leq i \leq j \leq k \leq \ell \leq d}. \end{aligned} \quad (24)$$

Define $\tilde{\mathbf{I}} := E[\tilde{\mathbf{s}}(\mathbf{X}, \mathbf{Z})\tilde{\mathbf{s}}(\mathbf{X}, \mathbf{Z})^\top]$, $\tilde{\mathbf{I}}_\eta := E[\tilde{\mathbf{s}}_\eta \tilde{\mathbf{s}}_\eta^\top]$, $\tilde{\mathbf{I}}_{\lambda\eta} := E[\tilde{\mathbf{s}}_\lambda \tilde{\mathbf{s}}_\eta^\top]$, $\tilde{\mathbf{I}}_{\eta\lambda} := \tilde{\mathbf{I}}_{\lambda\eta}^\top$, $\tilde{\mathbf{I}}_\lambda := E[\tilde{\mathbf{s}}_\lambda \tilde{\mathbf{s}}_\lambda^\top]$, and $\tilde{\mathbf{I}}_{\lambda,\eta} := \tilde{\mathbf{I}}_\lambda - \tilde{\mathbf{I}}_{\lambda\eta} \tilde{\mathbf{I}}_\eta^{-1} \tilde{\mathbf{I}}_{\eta\lambda}$. Let $\tilde{\mathbf{G}}_{\lambda,\eta} = ((\mathbf{G}_{\lambda,\eta}^1)^\top, \dots, (\mathbf{G}_{\lambda,\eta}^{M_0})^\top)^\top \sim N(0, \tilde{\mathbf{I}}_{\lambda,\eta})$ be an $\mathbb{R}^{M_0(d_{\mu v} + d_{\mu^4})}$ -valued random vector, and define $\mathbf{I}_{\lambda,\eta}^m := E[\mathbf{G}_{\lambda,\eta}^m (\mathbf{G}_{\lambda,\eta}^m)^\top]$ and $\mathbf{Z}_\lambda^m := (\mathbf{I}_{\lambda,\eta}^m)^{-1} \mathbf{G}_{\lambda,\eta}^m$. Similar to $\hat{\mathbf{t}}_\lambda^e$ in the test of homogeneity, define $\hat{\mathbf{t}}_{\lambda,m}^e$ by

$$r^m(\hat{\mathbf{t}}_{\lambda,m}^e) = \inf_{\mathbf{t}_\lambda \in \Lambda_\lambda^e} r^m(\mathbf{t}_\lambda), \quad r^m(\mathbf{t}_\lambda) := (\mathbf{t}_\lambda - \mathbf{Z}_\lambda^m)^\top \mathbf{I}_{\lambda,\eta}^m (\mathbf{t}_\lambda - \mathbf{Z}_\lambda^m).$$

The following proposition gives the asymptotic null distribution of the penalized LRT statistic for testing H_{01} . In the neighborhood of Υ_{1h}^* , the log-likelihood function permits a similar quadratic approximation to the one we derived in Section 4. Consequently, the LRT statistic is asymptotically distributed as the maximum of M_0 random variables.

Assumption 5. (a) $\alpha_j^* \in [\epsilon_1, 1 - \epsilon_1]$ for $j = 1, \dots, M_0$. (b) $\tilde{\mathbf{I}}$ is nonsingular.

Proposition 6. Suppose that Assumptions 3 and 5 hold. Then, under the null hypothesis $H_0 : m = M_0$, $PLR_n^{M_0}(\epsilon_1) \rightarrow_d \max\{v_1, \dots, v_{M_0}\}$, where $v_m := \max_{\mathbf{e} \in \{0,1\}^d} \left((\hat{\mathbf{t}}_{\lambda,m}^e)^\top \mathbf{I}_{\lambda,\eta}^m \hat{\mathbf{t}}_{\lambda,m}^e \right)$.

6 EM test

In this section, we develop an EM test of $H_0 : M = M_0$ against $H_1 : M = M_0 + 1$ for model (21). We drop the covariate \mathbf{Z} in this section. First, we develop an EM test statistic for testing $H_{0,1m} : (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = (\boldsymbol{\mu}_{m+1}, \boldsymbol{\Sigma}_{m+1})$. We construct M_0 intervals $\{D_1^*, \dots, D_{M_0}^*\}$ of admissible values

of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, such that $(\boldsymbol{\mu}_m^*, \boldsymbol{\Sigma}_m^*) \in D_m^*$ but $(\boldsymbol{\mu}_j^*, \boldsymbol{\Sigma}_j^*) \notin D_m^*$ for any $j \neq m$. For example, as in our simulation, we may assume that the first element of $\boldsymbol{\mu}$ are distinct and set, with μ_{1j} denoting the first element of $\boldsymbol{\mu}_j$, $D_1^* = [\underline{\Theta}_{\mu_1}, (\mu_{11}^* + \mu_{12}^*)/2] \times \Theta_{\boldsymbol{\Sigma}}$, $D_j^* = [(\mu_{1,j-1}^* + \mu_{1j}^*)/2, (\mu_{1j}^* + \mu_{1,j+1}^*)/2] \times \Theta_{\boldsymbol{\Sigma}}$ for $j = 2, \dots, M_0 - 1$, and $D_{M_0}^* = [(\mu_{1,M_0-1}^* + \mu_{1M_0}^*)/2, \overline{\Theta}_{\mu_1}] \times \Theta_{\boldsymbol{\Sigma}}$, where $\underline{\Theta}_{\mu_1}$ and $\overline{\Theta}_{\mu_1}$ are defined by $\Theta_{\mu_1} = [\underline{\Theta}_{\mu_1}, \overline{\Theta}_{\mu_1}]$ and may take either the value $-\infty$ or ∞ .

Collect the mixing parameters of the $(M_0 + 1)$ -component model into one vector as $\boldsymbol{\varsigma} := (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{M_0+1}, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_{M_0+1}) \in \Theta_{\boldsymbol{\varsigma}} := \Theta_{\boldsymbol{\mu}}^{M_0+1} \times \Theta_{\boldsymbol{\Sigma}}^{M_0+1}$. For $m = 1, \dots, M_0$, define a restricted parameter space of $\boldsymbol{\varsigma}$ by $\Omega_m^* := \{\boldsymbol{\varsigma} \in \Theta_{\boldsymbol{\varsigma}} : (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \in D_j^* \text{ for } j = 1, \dots, m-1; (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), (\boldsymbol{\mu}_{m+1}, \boldsymbol{\Sigma}_{m+1}) \in D_m^*; (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \in D_{j-1}^* \text{ for } j = m+2, \dots, M_0+1\}$. Let $\hat{\Omega}_m$ and \hat{D}_m be consistent estimates of Ω_m^* and D_m^* , which can be constructed from a consistent estimate of the M_0 -component model. We test $H_{0,1m} : (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = (\boldsymbol{\mu}_{m+1}, \boldsymbol{\Sigma}_{m+1})$ by estimating the $(M_0 + 1)$ -component model (22) under the restriction $\boldsymbol{\varsigma} \in \hat{\Omega}_m$. For example, when we test $H_{0,11} : (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ in a three-component model, the restriction can be given as $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \in \hat{D}_1$ and $(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3) \in \hat{D}_2$.

Let \mathcal{T} be a finite set of numbers from $(0, 0.5]$. We introduce another penalty term $p(\tau)$ that is continuous in τ , $p(0.5) = 0$, and $p(\tau) \rightarrow -\infty$ as τ goes to 0 or 1. For each $\tau_0 \in \mathcal{T}$, define the restricted penalized MLE as $\boldsymbol{\vartheta}_{M_0+1}^{m(1)}(\tau_0) := \arg \max_{\boldsymbol{\vartheta}_{M_0+1} \in \Theta^m(\tau_0)} (PL_n(\boldsymbol{\vartheta}_{M_0+1}) + p(\tau_0))$, where $\Theta^m(\tau_0) := \{\boldsymbol{\vartheta}_{M_0+1} \in \Theta_{\boldsymbol{\vartheta}_{M_0+1}} : \alpha_M / (\alpha_M + \alpha_{m+1}) = \tau_0 \text{ and } \boldsymbol{\varsigma} \in \hat{\Omega}_m\}$. Starting from $\boldsymbol{\vartheta}_{M_0+1}^{m(1)}(\tau_0)$, we update $\boldsymbol{\vartheta}_{M_0+1}$ by the following generalized EM algorithm. Henceforth, we suppress (τ_0) from $\boldsymbol{\vartheta}_{M_0+1}^{m(k)}(\tau_0)$. Suppose we have already calculated $\boldsymbol{\vartheta}_{M_0+1}^{m(k)}$. For $i = 1, \dots, n$ and $j = 1, \dots, M_0 + 1$, define the weights for an E-step as

$$w_{ij}^{(k)} := \begin{cases} \alpha_j^{(k)} f(\mathbf{X}_i; \boldsymbol{\mu}_j^{(k)}, \boldsymbol{\Sigma}_j^{(k)}) / f_{M_0+1}(\mathbf{X}_i; \boldsymbol{\vartheta}_{M_0+1}^{m(k)}) & \text{for } j = 1, \dots, m-1, \\ \alpha_{j-1}^{(k)} f(\mathbf{X}_i; \boldsymbol{\mu}_j^{(k)}, \boldsymbol{\Sigma}_j^{(k)}) / f_{M_0+1}(\mathbf{X}_i; \boldsymbol{\vartheta}_{M_0+1}^{m(k)}) & \text{for } j = m+2, \dots, M_0+1, \end{cases}$$

$$w_{im}^{(k)} := \frac{\tau^{(k)} \alpha_m^{(k)} f(\mathbf{X}_i; \boldsymbol{\mu}_m^{(k)}, \boldsymbol{\Sigma}_m^{(k)})}{f_{M_0+1}(\mathbf{X}_i; \boldsymbol{\vartheta}_{M_0+1}^{m(k)}), \quad w_{i,m+1}^{(k)} := \frac{(1 - \tau^{(k)}) \alpha_m^{(k)} f(\mathbf{X}_i; \boldsymbol{\mu}_{m+1}^{(k)}, \boldsymbol{\Sigma}_{m+1}^{(k)})}{f_{M_0+1}(\mathbf{X}_i; \boldsymbol{\vartheta}_{M_0+1}^{m(k)})}.$$

In an M-step, update τ and $\boldsymbol{\alpha}$ by

$$\tau^{(k+1)} := \arg \max_{\tau} \left\{ \sum_{i=1}^n w_{im}^{(k)} \log(\tau) + \sum_{i=1}^n w_{i,m+1}^{(k)} \log(1 - \tau) + p(\tau) \right\},$$

$$\alpha_j^{(k+1)} := \begin{cases} n^{-1} \sum_{i=1}^n w_{ij}^{(k)} & \text{for } j = 1, \dots, m-1, \\ n^{-1} \sum_{i=1}^n (w_{im}^{(k)} + w_{i,m+1}^{(k)}), & \text{for } j = m, \\ n^{-1} \sum_{i=1}^n w_{i,j+1}^{(k)} & \text{for } j = m+1, \dots, M_0, \end{cases}$$

and update $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ by

$$\boldsymbol{\Sigma}_j^{(k+1)} := \arg \max_{\boldsymbol{\Sigma}_j} \left\{ \sum_{i=1}^n w_{ij}^{(k)} \log f(\mathbf{X}_i; \boldsymbol{\mu}_j^{(k+1)}, \boldsymbol{\Sigma}_j) + p_{nj}(\boldsymbol{\Sigma}_j) \right\}. \quad (25)$$

The penalized likelihood value never decreases after each generalized EM step (Dempster et al., 1977, Theorem 1). Note that $\boldsymbol{\vartheta}_{M_0+1}^{m(k)}$ for $k \geq 2$ does not use the restriction $\hat{\Omega}_m$. For each $\tau_0 \in \mathcal{T}$ and k , define

$$M_n^{m(k)}(\tau_0) := 2 \left\{ PL_n(\boldsymbol{\vartheta}_{M_0+1}^{m(k)}(\tau_0)) - PL_{0,n}(\hat{\boldsymbol{\vartheta}}_{M_0}) \right\}, \quad (26)$$

where $\hat{\boldsymbol{\vartheta}}_{M_0} := \arg \max_{\boldsymbol{\vartheta}_{M_0} \in \Theta_{\boldsymbol{\vartheta}_{M_0}}} PL_{0,n}(\boldsymbol{\vartheta}_{M_0})$.

Finally, with a pre-specified number K , define the *local EM test statistic* for testing $H_{0,1m}$ by taking the maximum of $M_n^{m(K)}(\tau_0)$ over $\tau_0 \in \mathcal{T}$ as $EM_n^{m(K)} := \max \left\{ M_n^{m(K)}(\tau_0) : \tau_0 \in \mathcal{T} \right\}$. The *EM test statistic* is defined as the maximum of M_0 local EM test statistics: $EM_n^{(K)} := \max \left\{ EM_n^{1(K)}, EM_n^{2(K)}, \dots, EM_n^{M_0(K)} \right\}$. The following proposition shows that for any finite K , the EM test statistic is asymptotically equivalent to the penalized LRT statistic for testing H_{01} .

Proposition 7. *Suppose that Assumptions 2, 3, and 5 hold and $\{0.5\} \in \mathcal{T}$. Then, under the null hypothesis $H_0 : m = M_0$, for any fixed finite K , as $n \rightarrow \infty$, $EM_n^{(K)} \rightarrow_d \max\{v_1, \dots, v_{M_0}\}$, where the v_m s are given in Proposition 6.*

7 Simulation

7.1 Choice of penalty function

To apply our EM test, we need to specify the set \mathcal{T} , number of iterations K , and penalty functions for $p_{nm}(\boldsymbol{\Sigma}_m)$ and $p(\tau)$. Based on our experience, we recommend $\mathcal{T} = \{0.1, 0.3, 0.5\}$ and $K = \{1, 2, 3\}$. For $p_{nm}(\boldsymbol{\Sigma}_m)$, we employ a multivariate version of the penalty function used by Chen et al. (2012), namely,

$$p_{nm}(\boldsymbol{\Sigma}_m; \hat{\boldsymbol{\Sigma}}_m) = -a_n \left\{ \text{tr}(\hat{\boldsymbol{\Sigma}}_m \boldsymbol{\Sigma}_m^{-1}) - 2 \log(\det(\hat{\boldsymbol{\Sigma}}_m \boldsymbol{\Sigma}_m^{-1})) - d \right\}, \quad (27)$$

where $\hat{\boldsymbol{\Sigma}}_m$ is the estimate from the M_0 -component model. $p_{nm}(\boldsymbol{\Sigma}_m; \hat{\boldsymbol{\Sigma}}_m)$ satisfies Assumption 2 if $a_n = o_p(n^{1/4})$. We set $p(\tau) = \log(2 \min\{\tau, 1 - \tau\})$ as suggested by Chen and Li (2009). We set $a_n = 1$. When estimating the model under the null hypothesis and computing $L_{0,n}(\hat{\boldsymbol{\vartheta}}_{M_0})$, we use the penalty function (5) and set $a_n = n^{-1/2}$ as recommended by Chen and Tan (2009).

7.2 Simulation results

We examine the type I error rates and powers of the EM test by small simulations using mixtures of bivariate normal distributions. Computation was done using R (R Core Team, 2016). The critical

values are computed by bootstrap with 199 bootstrap replications. We use 1,000 replications, and the sample sizes are set to 200 and 400.

Table 1 reports the type I error rates of the EM test of $H_0 : M = 1$ against the alternative $H_1 : M = 2$ using two models under the null hypothesis. The EM test statistics give accurate type I errors. Table 2 reports the powers of the EM test under two models under the alternative, namely $M = 2$. The EM test shows good power.

Tables 3–5 report the type I error rates and powers of the EM test of $H_0 : M = 2$ against the alternative $H_1 : M = 3$. Overall, EM test performs well under finite sample size, even though the type I error rates are less accurate than in testing $H_0 : M = 1$.

8 Proof of propositions

Proof of Proposition 1. Define $h(\mathbf{y}, \boldsymbol{\vartheta}) := \sqrt{\ell(\mathbf{y}, \boldsymbol{\vartheta})} - 1$. We first show

$$\sup_{\boldsymbol{\vartheta} \in \mathcal{N}_{c/\sqrt{n}}} \left| nP_n(h(\mathbf{y}, \boldsymbol{\vartheta})^2) - n\mathbf{t}(\boldsymbol{\vartheta})^\top \boldsymbol{\mathcal{I}}_\pi \mathbf{t}(\boldsymbol{\vartheta})/4 \right| = o_p(1). \quad (28)$$

To show (28), write $4P_n(h(\mathbf{y}, \boldsymbol{\vartheta})^2)$ as

$$4P_n(h(\mathbf{y}, \boldsymbol{\vartheta})^2) = P_n \left(\frac{4(\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1)^2}{(\sqrt{\ell(\mathbf{y}; \boldsymbol{\vartheta})} + 1)^2} \right) = P_n(\ell(\mathbf{y}, \boldsymbol{\vartheta}) - 1)^2 - P_n \left((\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1)^3 \frac{(\sqrt{\ell(\mathbf{y}; \boldsymbol{\vartheta})} + 3)}{(\sqrt{\ell(\mathbf{y}; \boldsymbol{\vartheta})} + 1)^3} \right). \quad (29)$$

It follows from Assumption 1(a)(b)(c) and $(E|XY|)^2 \leq E|X|^2 E|Y|^2$ that, uniformly for $\boldsymbol{\vartheta} \in \mathcal{N}_\varepsilon$,

$$\begin{aligned} P_n(\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1)^2 &= \mathbf{t}(\boldsymbol{\vartheta})^\top P_n(\mathbf{s}(\mathbf{y}; \boldsymbol{\pi})\mathbf{s}(\mathbf{y}; \boldsymbol{\pi})^\top) \mathbf{t}(\boldsymbol{\vartheta}) + 2\mathbf{t}(\boldsymbol{\vartheta})^\top P_n[\mathbf{s}(\mathbf{y}; \boldsymbol{\pi})r(\mathbf{y}; \boldsymbol{\vartheta})] + P_n(r(\mathbf{y}; \boldsymbol{\vartheta}))^2 \\ &= \mathbf{t}(\boldsymbol{\vartheta})^\top \boldsymbol{\mathcal{I}}_\pi \mathbf{t}(\boldsymbol{\vartheta}) + o_p(|\mathbf{t}(\boldsymbol{\vartheta})|^2) + O_p(|\mathbf{t}(\boldsymbol{\vartheta})|^2 |\boldsymbol{\psi} - \boldsymbol{\psi}^*|). \end{aligned} \quad (30)$$

Note that, if X_1, \dots, X_n are random variables with $\max_{1 \leq i \leq n} \mathbb{E}|X_i|^q < C$ for some $q > 0$ and $C < \infty$, then we have $\max_{1 \leq i \leq n} |X_i| = o_p(n^{1/q})$. Therefore, from Assumption 1(a)(c), we have

$$\max_{1 \leq k \leq n} \sup_{\boldsymbol{\vartheta} \in \mathcal{N}_{c/\sqrt{n}}} |\ell(\mathbf{y}, \boldsymbol{\vartheta}) - 1| = \max_{1 \leq k \leq n} \sup_{\boldsymbol{\vartheta} \in \mathcal{N}_{c/\sqrt{n}}} |\mathbf{t}(\boldsymbol{\vartheta})^\top \mathbf{s}(\mathbf{y}; \boldsymbol{\pi}) + r(\mathbf{y}; \boldsymbol{\vartheta})| = o_p(1).$$

Therefore, the second term on the right of (29) is $o_p(1)P_n(\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1)^2$, and (28) follows from (30).

Consider the following expansion of $h(\mathbf{y}, \boldsymbol{\vartheta})$:

$$h(\mathbf{y}, \boldsymbol{\vartheta}) = (\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1)/2 - h(\mathbf{y}_i, \boldsymbol{\vartheta})^2/2 = (\mathbf{t}(\boldsymbol{\vartheta})^\top \mathbf{s}(\mathbf{y}; \boldsymbol{\pi}) + r(\mathbf{y}; \boldsymbol{\vartheta}))/2 - h(\mathbf{y}, \boldsymbol{\vartheta})^2/2. \quad (31)$$

It follows from (28), (31), and Assumption 1(d) that $nP_n(h(\mathbf{y}, \boldsymbol{\vartheta})) = \sqrt{n}\mathbf{t}(\boldsymbol{\vartheta})^\top \nu_n(\mathbf{s}(\mathbf{y}; \boldsymbol{\pi}))/2 - n\mathbf{t}(\boldsymbol{\vartheta})^\top \boldsymbol{\mathcal{I}}_\pi \mathbf{t}(\boldsymbol{\vartheta})/8 + o_p(1)$ uniformly for $\boldsymbol{\vartheta} \in \mathcal{N}_{c/\sqrt{n}}$. Using the Taylor expansion of $2 \log(1+x) = 2x - x^2(1+o(1))$ for small x , we have, uniformly for $\boldsymbol{\vartheta} \in \mathcal{N}_{c/\sqrt{n}}$,

$$\begin{aligned} L_n(\boldsymbol{\psi}, \boldsymbol{\pi}) - L_n(\boldsymbol{\psi}^*, \boldsymbol{\pi}) &= 2 \sum_{i=1}^n \log(1 + h(\mathbf{y}_i, \boldsymbol{\vartheta})) = nP_n(2h(\mathbf{y}, \boldsymbol{\vartheta}) - [1 + o_p(1)]h(\mathbf{y}, \boldsymbol{\vartheta})^2) \\ &= \sqrt{n}\mathbf{t}(\boldsymbol{\vartheta})^\top \nu_n(\mathbf{s}(\mathbf{y}; \boldsymbol{\pi})) - \mathbf{t}(\boldsymbol{\vartheta})^\top \boldsymbol{\mathcal{I}}_\pi \mathbf{t}(\boldsymbol{\vartheta})/4 + nP_n(h(\mathbf{y}, \boldsymbol{\vartheta})^2) + o_p(1) \\ &= \sqrt{n}\mathbf{t}(\boldsymbol{\vartheta})^\top \nu_n(\mathbf{s}(\mathbf{y}; \boldsymbol{\pi})) - \mathbf{t}(\boldsymbol{\vartheta})^\top \boldsymbol{\mathcal{I}}_\pi \mathbf{t}(\boldsymbol{\vartheta})/2 + o_p(1), \end{aligned}$$

giving the stated result. \square

Proof of Proposition 2. For part (a), applying the inequality $\log(1+x) \leq x$ to the log-likelihood

ratio function and using (31) give

$$L_n(\boldsymbol{\psi}, \boldsymbol{\pi}) - L_n(\boldsymbol{\psi}^*, \boldsymbol{\pi}) = 2 \sum_{i=1}^n \log(1+h(\mathbf{y}_i, \boldsymbol{\vartheta})) \leq 2nP_n(h(\mathbf{y}, \boldsymbol{\vartheta})) = \sqrt{n}\nu_n(\ell(\mathbf{y}; \boldsymbol{\vartheta})-1) - nP_n(h(\mathbf{y}, \boldsymbol{\vartheta})^2). \quad (32)$$

We derive a lower bound on $P_n(h(\mathbf{y}, \boldsymbol{\vartheta})^2)$. From the first equality in (29), for some $\Xi > 0$,

$$\begin{aligned} P_n(h(\mathbf{y}, \boldsymbol{\vartheta})^2) &\geq P_n\left(\frac{(\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1)^2}{\ell(\mathbf{y}; \boldsymbol{\vartheta}) + 1}\right) \\ &\geq \frac{1}{\Xi + 1} P_n(\mathbb{I}\{\ell(\mathbf{y}; \boldsymbol{\vartheta}) \leq \Xi\}(\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1)^2) \\ &\geq \frac{1}{\Xi + 1} [P_n((\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1)^2) - P_n(\mathbb{I}\{\ell(\mathbf{y}; \boldsymbol{\vartheta}) > \Xi\}(\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1)^2)]. \end{aligned}$$

Let $B := \sup_{\boldsymbol{\vartheta} \in \mathcal{N}_\varepsilon} |\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1|$. From Assumption 1(a)(c), we have $EB^2 < \infty$, and hence $\lim_{\Xi \rightarrow \infty} \sup_{\boldsymbol{\vartheta} \in \mathcal{N}_\Xi} P_n(\mathbb{I}\{\ell(\mathbf{y}; \boldsymbol{\vartheta}) > \Xi\}(\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1)^2) \leq \lim_{\Xi \rightarrow \infty} P_n(\mathbb{I}\{B + 1 > \Xi\}B^2) = 0$ almost surely. Let $\kappa = (2(\Xi + 1))^{-1} > 0$. By choosing Ξ sufficiently large, it follows from (30) and Assumption 1(e) that, uniformly for $\boldsymbol{\vartheta} \in \mathcal{N}_\varepsilon$,

$$P_n(h(\mathbf{y}, \boldsymbol{\vartheta})^2) \geq \kappa \mathbf{t}(\boldsymbol{\vartheta})^\top \boldsymbol{\mathcal{I}}_\pi \mathbf{t}(\boldsymbol{\vartheta}) + o_p(|\mathbf{t}(\boldsymbol{\vartheta})|^2) + O_p(|\mathbf{t}(\boldsymbol{\vartheta})|^2 |\boldsymbol{\psi} - \boldsymbol{\psi}^*|). \quad (33)$$

Because $\sqrt{n}\nu_n(\ell(\mathbf{y}; \boldsymbol{\vartheta}) - 1) = \sqrt{n}\mathbf{t}(\boldsymbol{\vartheta})^\top [\nu_n(\mathbf{s}(\mathbf{y}; \boldsymbol{\pi})) + O_p(1)]$ from Assumption 1(d), it follows from (32) and (33) that

$$-\delta \leq L_n(\boldsymbol{\psi}, \boldsymbol{\pi}) - L_n(\boldsymbol{\psi}^*, \boldsymbol{\pi}) \leq \sqrt{n}\mathbf{t}(\boldsymbol{\vartheta})^\top [\nu_n(\mathbf{s}(\mathbf{y}; \boldsymbol{\pi})) + O_p(1)] - \kappa n \mathbf{t}(\boldsymbol{\vartheta})^\top \boldsymbol{\mathcal{I}}_\pi \mathbf{t}(\boldsymbol{\vartheta}) + o_p(n|\mathbf{t}(\boldsymbol{\vartheta})|^2). \quad (34)$$

The rest of the proof is similar to the proof of Theorem 1 of Andrews (1999). Let $\mathbf{T}_n := \boldsymbol{\mathcal{I}}_\pi^{1/2} \sqrt{n}\mathbf{t}(\boldsymbol{\vartheta})$. In view of Assumption 1(e)(f), we can write (34) as $-\delta \leq |\mathbf{T}_n| O_p(1) - \kappa |\mathbf{T}_n|^2 + o_p(|\mathbf{T}_n|^2)$. Rearranging this equation gives $|\mathbf{T}_n|^2 \leq 2|\mathbf{T}_n|s_n + \delta$ with $s_n = O_p(1)$. Then, $(|\mathbf{T}_n| - s_n)^2 \leq s_n^2 + \delta$, and taking the square roots gives $|\mathbf{T}_n| \leq O_p(1)$, giving part (a). Part (b) follows from part (a) and Proposition 1. \square

8.1 Proof of Proposition 3

The stated result follows from Theorem 1 of Chen and Tan (2009) and Corollary 3 of Alexandrovich (2014). \square

8.2 Proof of Proposition 4

We suppress the subscript α from $\boldsymbol{\psi}_\alpha$. We prove the stated result by applying Proposition 2 to $\ell(\mathbf{y}, \boldsymbol{\vartheta})$ with $\ell(\mathbf{y}, \boldsymbol{\vartheta}) = \ell(\mathbf{y}, \boldsymbol{\psi}, \alpha) := g(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi}^*, \alpha)/g(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi}, \alpha)$ as defined in (2). Observe that $\mathbf{t}(\boldsymbol{\vartheta})$ defined in (15) satisfies $\mathbf{t}(\boldsymbol{\vartheta}) = 0$ if and only if $\boldsymbol{\psi} = \boldsymbol{\psi}^*$ because $\boldsymbol{\lambda} = 0$ if and only if $(\boldsymbol{\lambda}_{\boldsymbol{\mu}v})_{iii} = (\boldsymbol{\lambda}_{\boldsymbol{\mu}^4})_{iiii} = 0$ for all $1 \leq i \leq d$. We expand $\ell(\mathbf{y}, \boldsymbol{\vartheta}) - 1$ five times with respect to $\boldsymbol{\psi}$ and show that

the expansion satisfies Assumption 1.

Define

$$\mathbf{v}(\mathbf{y}; \boldsymbol{\vartheta}) := (\nabla_{\boldsymbol{\psi}} g(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi}, \alpha)^\top, \nabla_{\boldsymbol{\psi}^{\otimes 2}} g(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi}, \alpha)^\top, \dots, \nabla_{\boldsymbol{\psi}^{\otimes 5}} g(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi}, \alpha)^\top)^\top / g(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi}^*, \alpha). \quad (35)$$

In order to apply Proposition 2 to $\ell(\mathbf{y}, \boldsymbol{\vartheta}) - 1$, we first show

$$\sup_{\boldsymbol{\vartheta} \in \mathcal{N}_\epsilon} \left| P_n[\mathbf{v}(\mathbf{y}; \boldsymbol{\vartheta}) \mathbf{v}(\mathbf{y}; \boldsymbol{\vartheta})^\top] - E[\mathbf{v}(\mathbf{Y}; \boldsymbol{\vartheta}) \mathbf{v}(\mathbf{Y}; \boldsymbol{\vartheta})^\top] \right| = o_p(1), \quad (36)$$

$$\nu_n(\mathbf{v}(\mathbf{y}; \boldsymbol{\vartheta})) \Rightarrow \mathbf{W}(\boldsymbol{\vartheta}), \quad (37)$$

where $\mathbf{W}(\boldsymbol{\vartheta})$ is a mean-zero continuous Gaussian process with $E[\mathbf{W}(\boldsymbol{\vartheta}_1) \mathbf{W}(\boldsymbol{\vartheta}_2)^\top] = E[\mathbf{v}(\mathbf{Y}; \boldsymbol{\vartheta}_1) \mathbf{v}(\mathbf{Y}; \boldsymbol{\vartheta}_2)^\top]$. (36) holds because $\mathbf{v}(\mathbf{Y}_i; \boldsymbol{\vartheta}) \mathbf{v}(\mathbf{Y}_i; \boldsymbol{\vartheta})^\top$ satisfies a uniform law of large numbers (see, for example, Lemma 2.4 of Newey and McFadden (1994)) because $\mathbf{v}(\mathbf{y}; \boldsymbol{\vartheta})$ is continuous in $\boldsymbol{\vartheta}$ and $E \sup_{\boldsymbol{\vartheta} \in \mathcal{N}_\epsilon} |\mathbf{v}(\mathbf{Y}; \boldsymbol{\vartheta})|^2 < \infty$ from the property of the normal density and Assumption 4. (37) follows from Theorem 10.2 of Pollard (1990) if (i) $\Theta_{\boldsymbol{\vartheta}}$ is totally bounded, (ii) the finite dimensional distributions of $\nu_n(\mathbf{v}(\mathbf{y}; \boldsymbol{\vartheta}))$ converge to those of $\mathbf{W}(\boldsymbol{\vartheta})$, and (iii) $\{\nu_n(\mathbf{v}(\mathbf{y}; \boldsymbol{\vartheta})) : n \geq 1\}$ is stochastically equicontinuous. Condition (i) holds because $\Theta_{\boldsymbol{\vartheta}}$ is compact in the Euclidean space. Condition (ii) follows from Assumption 4 and the multivariate CLT. Condition (iii) holds Theorem 2 of Andrews (1994) because $\mathbf{v}(\mathbf{y}; \boldsymbol{\vartheta})$ is Lipschitz continuous in $\boldsymbol{\vartheta}$.

Note that the $(p+1)$ -th order Taylor expansion of $f(\mathbf{x})$ around $\mathbf{x} = \mathbf{x}^*$ is given by

$$f(\mathbf{x}) = f(\mathbf{x}^*) + \sum_{j=1}^p \frac{1}{j!} \nabla_{(\mathbf{x}^{\otimes j})^\top} f(\mathbf{x}^*) \mathbf{x}^{\otimes j} + \frac{1}{(p+1)!} \nabla_{(\mathbf{x}^{\otimes (p+1)})^\top} f(\bar{\mathbf{x}}) \mathbf{x}^{\otimes (p+1)},$$

where $\bar{\mathbf{x}}$ lies between \mathbf{x} and \mathbf{x}^* , and $\bar{\mathbf{x}}$ may differ from element to element of $\nabla_{(\mathbf{x}^{\otimes (p+1)})^\top} f(\bar{\mathbf{x}})$.

Let g^* and ∇g^* denote $g(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi}^*, \alpha)$ and $\nabla g(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi}^*, \alpha)$, and let $\nabla \bar{g}$ denote $\nabla g(\mathbf{x}|\mathbf{z}; \bar{\boldsymbol{\psi}}, \alpha)$. Let $\dot{\boldsymbol{\eta}} := \boldsymbol{\eta} - \boldsymbol{\eta}^*$. Expanding $\ell(\mathbf{y}; \boldsymbol{\psi}, \alpha)$ five times around $\boldsymbol{\psi}^*$ while fixing α and using Lemma 3, we can write $\ell(\mathbf{y}; \boldsymbol{\psi}, \alpha)$ as follows with $\boldsymbol{\psi}_- := (\dot{\boldsymbol{\eta}}^\top, \boldsymbol{\lambda}_\mu^\top)^\top$:

$$\ell(\mathbf{y}; \boldsymbol{\psi}, \alpha) = \frac{\nabla_{\boldsymbol{\eta}^\top} g^*}{g^*} \dot{\boldsymbol{\eta}} + \frac{1}{2!} \boldsymbol{\lambda}^\top \frac{\nabla_{\boldsymbol{\lambda} \boldsymbol{\lambda}^\top} g^*}{g^*} \boldsymbol{\lambda} + \frac{1}{2!} \dot{\boldsymbol{\eta}}^\top \frac{\nabla_{\boldsymbol{\eta} \boldsymbol{\eta}^\top} g^*}{g^*} \dot{\boldsymbol{\eta}} \quad (38)$$

$$+ \frac{1}{3!} \frac{\nabla_{(\boldsymbol{\psi}^{\otimes 3})^\top} g^*}{g^*} \boldsymbol{\psi}^{\otimes 3} + \frac{1}{4!} \frac{\nabla_{(\boldsymbol{\lambda}_\mu^{\otimes 4})^\top} g^*}{g^*} \boldsymbol{\lambda}_\mu^{\otimes 4} \quad (39)$$

$$+ \sum_{p=0}^3 \frac{1}{p!(4-p)!} \frac{\nabla_{(\boldsymbol{\psi}_-^{\otimes (4-p)} \otimes \boldsymbol{\lambda}_\mu^{\otimes p})^\top} g^*}{g^*} (\boldsymbol{\psi}_-^{\otimes (4-p)} \otimes \boldsymbol{\lambda}_\mu^{\otimes p}) \quad (40)$$

$$+ \frac{1}{5!} \frac{\nabla_{(\boldsymbol{\lambda}_\mu^{\otimes 5})^\top} \bar{g}}{g^*} \boldsymbol{\lambda}_\mu^{\otimes 5} + \sum_{p=0}^4 \frac{1}{p!(5-p)!} \frac{\nabla_{(\boldsymbol{\psi}_-^{\otimes (5-p)} \otimes \boldsymbol{\lambda}_\mu^{\otimes p})^\top} \bar{g}}{g^*} (\boldsymbol{\psi}_-^{\otimes (5-p)} \otimes \boldsymbol{\lambda}_\mu^{\otimes p}). \quad (41)$$

We first analyze the first two terms on the right hand side of (38) and the second term in (39)

because these terms constitute the leading term. Let f^* and ∇f^* denote $f(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ and $\nabla f(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. The first term on the right hand side of (38) is simply $\nabla_{(\boldsymbol{\gamma}^\top, \boldsymbol{\mu}^\top, \mathbf{v}^\top)^\top} f^*/f^*$. Using Lemma 3 and commutativity of partial derivatives, the second term on the right hand side of (38) is written as

$$\begin{aligned}
& \frac{1}{2!} \boldsymbol{\lambda}^\top \frac{\nabla_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top g^*}{g^*} \boldsymbol{\lambda} \\
&= \boldsymbol{\lambda}_\mu^\top \frac{\nabla_{\boldsymbol{\lambda}_\mu} \boldsymbol{\lambda}_\mu^\top g^*}{g^*} \boldsymbol{\lambda}_\mu + \frac{1}{2!} \boldsymbol{\lambda}_v^\top \frac{\nabla_{\boldsymbol{\lambda}_v} \boldsymbol{\lambda}_v^\top g^*}{g^*} \boldsymbol{\lambda}_v \\
&= \alpha(1-\alpha) \sum_{\substack{1 \leq i \leq d \\ 1 \leq j \leq k \leq d}} \frac{\nabla_{\mu_i \mu_j \mu_k} f^*}{f^*} \lambda_{\mu_i} \lambda_{v_{jk}} + \frac{\alpha(1-\alpha)}{2} \sum_{\substack{1 \leq i \leq j \leq d \\ 1 \leq k \leq \ell \leq d}} \frac{\nabla_{\mu_i \mu_j \mu_k \mu_\ell} f^*}{f^*} \lambda_{v_{ij}} \lambda_{v_{k\ell}} \\
&= \alpha(1-\alpha) \sum_{1 \leq i \leq j \leq k \leq d} \frac{\nabla_{\mu_i \mu_j \mu_k} f^*}{f^*} \sum_{(t_1, t_2, t_3) \in p_{12}(i, j, k)} \lambda_{\mu_{t_1}} \lambda_{v_{t_2 t_3}} \\
&\quad + \frac{\alpha(1-\alpha)}{2} \sum_{1 \leq i \leq j \leq k \leq \ell \leq d} \frac{\nabla_{\mu_i \mu_j \mu_k \mu_\ell} f^*}{f^*} \sum_{(t_1, t_2, t_3, t_4) \in p_{22}(i, j, k, \ell)} \lambda_{v_{t_1 t_2}} \lambda_{v_{t_3 t_4}},
\end{aligned}$$

where $\sum_{(t_1, t_2, t_3) \in p_{12}(i, j, k)}$ denotes the sum over all distinct permutations of (i, j, k) to (t_1, t_2, t_3) with $t_2 \leq t_3$, and $\sum_{(t_1, t_2, t_3, t_4) \in p_{22}(i, j, k, \ell)}$ denotes the sum over all distinct permutations of (i, j, k, ℓ) to (t_1, t_2, t_3, t_4) with $t_1 \leq t_2$ and $t_3 \leq t_4$. From Lemma 3, the second term in (39) is written as

$$\frac{1}{4!} \frac{\nabla_{(\boldsymbol{\lambda}_\mu^{\otimes 4})^\top} g^*}{g^*} \boldsymbol{\lambda}_\mu^{\otimes 4} = \frac{\alpha(1-\alpha)}{4!} \sum_{1 \leq i \leq j \leq k \leq \ell \leq d} b(\alpha) \frac{\nabla_{\mu_i \mu_j \mu_k \mu_\ell} f^*}{f^*} \sum_{(t_1, t_2, t_3, t_4) \in p(i, j, k, \ell)} \lambda_{\mu_{t_1}} \lambda_{\mu_{t_2}} \lambda_{\mu_{t_3}} \lambda_{\mu_{t_4}},$$

where $\sum_{(t_1, t_2, t_3, t_4) \in p(i, j, k, \ell)}$ denotes the sum over all distinct permutations of (i, j, k, ℓ) to (t_1, t_2, t_3, t_4) . Combining these results, we obtain the leading term in the expansion

$$\frac{\nabla_{\boldsymbol{\eta}^\top} g^*}{g^*} \dot{\boldsymbol{\eta}} + \frac{1}{2!} \boldsymbol{\lambda}^\top \frac{\nabla_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top g^*}{g^*} \boldsymbol{\lambda} + \frac{1}{4!} \frac{\nabla_{(\boldsymbol{\lambda}_\mu^{\otimes 4})^\top} g^*}{g^*} \boldsymbol{\lambda}_\mu^{\otimes 4} = \mathbf{s}_\eta^\top \dot{\boldsymbol{\eta}} + \mathbf{s}_{\mu v}^\top \boldsymbol{\lambda}_{\mu v} + \mathbf{s}_{\mu^4}^\top \boldsymbol{\lambda}_{\mu^4},$$

with $(\mathbf{s}_\eta, \mathbf{s}_{\mu v}, \mathbf{s}_{\mu^4})$ and $(\boldsymbol{\lambda}_{\mu v}, \boldsymbol{\lambda}_{\mu^4})$ defined in (14) and (16).

$(\mathbf{s}_\eta, \mathbf{s}_{\mu v}, \mathbf{s}_{\mu^4})$ clearly satisfies Assumption 1(a)(b)(e)(f) from Assumption 4, the property of the normal density, (36), and (37). Therefore, the stated result holds if the other terms in (38)–(41) satisfy Assumption 1(c)(d). We proceed to show that these terms can be written as $\mathbf{v}(\mathbf{y}; \boldsymbol{\vartheta})O(|\boldsymbol{\psi}||\mathbf{t}(\boldsymbol{\vartheta})|)$ with $\mathbf{v}(\mathbf{y}; \boldsymbol{\vartheta})$ defined in (35). Then, Assumption 1(c)(d) follows from (36) and (37).

First, the third term on the right hand side of (38) is written as $(\nabla_{\boldsymbol{\eta}^{\otimes 2}} g^*/g^*)O(|\dot{\boldsymbol{\eta}}|^2)$. Second, write the first term in (39) as $\sum_{p=0}^3 (1/p!(3-p)!)(\nabla_{(\boldsymbol{\eta}^{\otimes p} \otimes \boldsymbol{\lambda}^{\otimes (3-p)})^\top} g^*/g^*)(\dot{\boldsymbol{\eta}}^{\otimes p} \otimes \boldsymbol{\lambda}^{\otimes (3-p)})$. The terms with $p \geq 1$ are written as $(\nabla_{\boldsymbol{\psi}^{\otimes 3}} g^*/g^*)O(|\dot{\boldsymbol{\eta}}|)O(|\boldsymbol{\lambda}|)$. The term with $p = 0$ is written as, because

$\nabla_{\lambda_{\mu_i} \lambda_{\mu_j} \lambda_{\mu_k}} g^* = 0$ from Lemma 3,

$$\begin{aligned} & \sum_{i=1}^d \sum_{j=1}^d \sum_{1 \leq k \leq \ell \leq d} \frac{\nabla_{\lambda_{\mu_i} \lambda_{\mu_j} \lambda_{v_{k\ell}}} g^*}{g^*} \lambda_{\mu_i} \lambda_{\mu_j} \lambda_{v_{k\ell}} + \sum_{i=1}^d \sum_{1 \leq j \leq k \leq d} \sum_{1 \leq \ell \leq m \leq d} \frac{\nabla_{\lambda_{\mu_i} \lambda_{v_{jk}} \lambda_{v_{\ell m}}} g^*}{g^*} \lambda_{\mu_i} \lambda_{v_{jk}} \lambda_{v_{\ell m}} \\ & + \frac{1}{3!} \sum_{1 \leq i \leq j \leq d} \sum_{1 \leq k \leq \ell \leq d} \sum_{1 \leq m \leq n \leq d} \frac{\nabla_{\lambda_{v_{ij}} \lambda_{v_{k\ell}} \lambda_{v_{mn}}} g^*}{g^*} \lambda_{v_{ij}} \lambda_{v_{k\ell}} \lambda_{v_{mn}}. \end{aligned} \quad (42)$$

The first term in (42) can be written as

$$\sum_{i=1}^d \lambda_{\mu_i} \sum_{1 \leq j \leq k \leq \ell \leq d} \frac{\nabla_{\lambda_{\mu_i} \lambda_{\mu_j} \lambda_{v_{k\ell}}} g^*}{g^*} \sum_{(t_1, t_2, t_3) \in p_{12}(j, k, \ell)} \lambda_{\mu_{t_1}} \lambda_{v_{t_2 t_3}} = \frac{\nabla_{\lambda^{\otimes 3}} g^*}{g^*} O(|\lambda| |\lambda_{\mu v}|).$$

From a similar argument, the second term in (42) is also written as $(\nabla_{\lambda^{\otimes 3}} g^*/g^*) O(|\lambda| |\lambda_{\mu v}|)$. In order to bound the third term in (42), observe that, for any sequence a_{ijklmn} ,

$$\begin{aligned} & \sum_{1 \leq i \leq j \leq d} \sum_{1 \leq k \leq \ell \leq d} \sum_{1 \leq m \leq n \leq d} a_{ijklmn} \lambda_{v_{ij}} \lambda_{v_{k\ell}} \lambda_{v_{mn}} \\ & = \sum_{1 \leq m \leq n \leq d} \lambda_{v_{mn}} \sum_{1 \leq i \leq j \leq k \leq \ell \leq d} a_{ijklmn} \left(12 \sum_{(t_1, t_2, t_3, t_4) \in p_{22}(i, j, k, \ell)} \lambda_{v_{t_1 t_2}} \lambda_{v_{t_3 t_4}} \right. \\ & \quad \left. + b(\alpha) \sum_{(t_1, t_2, t_3, t_4) \in p(i, j, k, \ell)} \lambda_{\mu_{t_1}} \lambda_{\mu_{t_2}} \lambda_{\mu_{t_3}} \lambda_{\mu_{t_4}} \right) \\ & \quad - b(\alpha) \sum_{i=1}^d \sum_{j=1}^d \sum_{k=1}^d \lambda_{\mu_i} \lambda_{\mu_j} \lambda_{\mu_k} \sum_{1 \leq \ell \leq m \leq n \leq d} a_{ijklmn} \sum_{(t_1, t_2, t_3) \in p_{12}(\ell, m, n)} \lambda_{\mu_{t_1}} \lambda_{v_{t_2 t_3}}. \end{aligned}$$

Therefore, (42) can be written as $(\nabla_{\psi^{\otimes 3}} g^*/g^*) [O(|\lambda| |\lambda_{\mu^4}|) + O(|\lambda| |\lambda_{\mu v}|)]$. We proceed to bound (40). The terms in (40) with $p \geq 1$ are written as $(\nabla_{\psi^{\otimes 4}} g^*/g^*) [O(|\lambda| |\lambda_{\mu v}|) + O(|\lambda| |\dot{\eta}|)]$ because they contain either $\sum_{i=1}^d \sum_{1 \leq j \leq k \leq d} \sum_{1 \leq \ell \leq m \leq d} \lambda_{\mu_i} \lambda_{v_{jk}} \lambda_{v_{\ell m}}$ or $\sum_{i=1}^d \lambda_{\mu_i} \dot{\eta}$. The term with $p = 0$ is written as $(\nabla_{\psi^{\otimes 4}} g^*/g^*) [O(|\lambda| |\lambda_{\mu^4}|) + O(|\lambda| |\lambda_{\mu v}|)]$ from a similar argument to bound (42).

It remains to bound (41). For the first term in (41), observe that, for any sequence a_{ijklm} ,

$$\begin{aligned}
& \sum_{i=1}^d \sum_{j=1}^d \sum_{k=1}^d \sum_{\ell=1}^d \sum_{m=1}^d a_{ijklm} \lambda_{\mu_i} \lambda_{\mu_j} \lambda_{\mu_k} \lambda_{\mu_\ell} \lambda_{\mu_M} \\
&= \frac{1}{b(\alpha)} \sum_{m=1}^d \lambda_{\mu_M} \sum_{1 \leq i \leq j \leq k \leq \ell \leq d} a_{ijklm} \left(b(\alpha) \sum_{(t_1, t_2, t_3, t_4) \in p(i, j, k, \ell)} \lambda_{\mu_{t_1}} \lambda_{\mu_{t_2}} \lambda_{\mu_{t_3}} \lambda_{\mu_{t_4}} \right. \\
&\quad \left. + 12 \sum_{(t_1, t_2, t_3, t_4) \in p_{22}(i, j, k, \ell)} \lambda_{v_{t_1 t_2}} \lambda_{v_{t_3 t_4}} \right) \\
&\quad - \frac{b(\alpha)}{12} \sum_{1 \leq i \leq j \leq d} \lambda_{v_{ij}} \sum_{1 \leq k \leq \ell \leq m \leq d} a_{ijklm} \sum_{(t_1, t_2, t_3) \in p_{12}(k, \ell, m)} \lambda_{\mu_{t_1}} \lambda_{v_{t_2 t_3}}.
\end{aligned}$$

Therefore, the first term in (41) can be written as $(\nabla_{\psi \otimes \bar{g}} / g^*) [O(|\lambda| |\lambda_{\mu^4}|) + O(|\lambda| |\lambda_{\mu v}|)]$. The second term in (41) is written as $(\nabla_{\psi \otimes \bar{g}} / g^*) [O(|\lambda| |\lambda_{\mu v}|) + O(|\lambda| |\dot{\eta}|) + O(|\lambda| |\lambda_{\mu^4}|)]$ from the same argument as (40), and the stated result follows. \square

8.3 Proof of Proposition 5

The proof is similar to that of Proposition 3 of Kasahara and Shimotsu (2015). Let $\mathbf{t}_\eta := \boldsymbol{\eta} - \boldsymbol{\eta}^*$, so that $\mathbf{t}(\boldsymbol{\psi}, \alpha)$ in (15) is written as $(\mathbf{t}_\eta^\top, \mathbf{t}(\boldsymbol{\lambda}, \alpha)^\top)^\top$. Let

$$\mathbf{G}_n := \nu_n(\mathbf{s}(\mathbf{x}, \mathbf{z})) = \begin{bmatrix} \mathbf{G}_{\eta n} \\ \mathbf{G}_{\lambda n} \end{bmatrix}, \quad \mathbf{G}_{\lambda, \eta n} := \mathbf{G}_{\lambda n} - \mathcal{I}_{\lambda \eta} \mathcal{I}_\eta^{-1} \mathbf{G}_{\eta n}, \quad \mathbf{Z}_{\lambda, \eta n} := \mathcal{I}_{\lambda, \eta}^{-1} \mathbf{G}_{\lambda, \eta n}, \\
\mathbf{t}_{\eta, \lambda} := \mathbf{t}_\eta + \mathcal{I}_\eta^{-1} \mathcal{I}_{\eta \lambda} \mathbf{t}(\boldsymbol{\lambda}, \alpha).$$

Then, we can split the quadratic form in Proposition 4(b) and write it as

$$\sup_{\boldsymbol{\vartheta} \in A_{n\alpha}(\delta)} \left| 2[L_n(\boldsymbol{\psi}_\alpha, \alpha) - L_n(\boldsymbol{\psi}_\alpha^*, \alpha)] - B_n(\sqrt{n} \mathbf{t}_{\eta, \lambda}) - C_n(\sqrt{n} \mathbf{t}(\boldsymbol{\lambda}, \alpha)) \right| = o_p(1), \quad (43)$$

where

$$\begin{aligned}
B_n(\mathbf{t}_{\eta, \lambda}) &= 2\mathbf{t}_{\eta, \lambda}^\top \mathbf{G}_{\eta n} - \mathbf{t}_{\eta, \lambda}^\top \mathcal{I}_{\eta} \mathbf{t}_{\eta, \lambda}, \\
C_n(\mathbf{t}(\boldsymbol{\lambda}, \alpha)) &= 2\mathbf{t}(\boldsymbol{\lambda}, \alpha)^\top \mathbf{G}_{\lambda, \eta n} - \mathbf{t}(\boldsymbol{\lambda}, \alpha)^\top \mathcal{I}_{\lambda, \eta} \mathbf{t}(\boldsymbol{\lambda}, \alpha) \\
&= \mathbf{Z}_{\lambda n}^\top \mathcal{I}_{\lambda, \eta} \mathbf{Z}_{\lambda n} - (\mathbf{t}(\boldsymbol{\lambda}, \alpha) - \mathbf{Z}_{\lambda n})^\top \mathcal{I}_{\lambda, \eta} (\mathbf{t}(\boldsymbol{\lambda}, \alpha) - \mathbf{Z}_{\lambda n}).
\end{aligned} \quad (44)$$

Observe that $2[L_{0,n}(\widehat{\boldsymbol{\gamma}}_0, \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\Sigma}}_0) - L_{0,n}(\boldsymbol{\gamma}^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)] = \max_{\mathbf{t}_\eta} [2\sqrt{n} \mathbf{t}_\eta^\top \mathbf{G}_{\eta n} - n \mathbf{t}_\eta^\top \mathcal{I}_\eta \mathbf{t}_\eta] + o_p(1) = \max_{\mathbf{t}_{\eta, \lambda}} B_n(\sqrt{n} \mathbf{t}_{\eta, \lambda}) + o_p(1)$ from applying Proposition 2 to $L_{0,n}(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and noting that the set of possible values of both $\sqrt{n} \mathbf{t}_\eta$ and $\sqrt{n} \mathbf{t}_{\eta, \lambda}$ approaches \mathbb{R}^{d_η} . In conjunction with (43), we obtain

$$2[L_n(\widehat{\boldsymbol{\psi}}_\alpha, \alpha) - L_{0,n}(\widehat{\boldsymbol{\gamma}}_0, \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\Sigma}}_0)] = C_n(\sqrt{n} \mathbf{t}(\widehat{\boldsymbol{\lambda}}, \alpha)) + o_p(1). \quad (45)$$

For $\mathbf{e} = (e_1, \dots, e_d)^\top \in \{0, 1\}^d$, define

$$\Theta_\lambda^e := \{\boldsymbol{\lambda} \in \Theta_\lambda : |\lambda_{\mu_i}| \geq n^{-1/8}(\log n)^{-1} \text{ if } e_i = 1; |\lambda_{\mu_i}| \leq n^{-1/8}(\log n)^{-1} \text{ if } e_i = 0\}, \quad (46)$$

so that $\Theta_\lambda = \cup_{\mathbf{e} \in \{0,1\}^d} \Theta_\lambda^e$. Define $\tilde{\boldsymbol{\lambda}}^e$ by $C_n(\sqrt{n}\mathbf{t}(\tilde{\boldsymbol{\lambda}}^e, \alpha)) = \max_{\boldsymbol{\lambda} \in \Theta_\lambda^e} C_n(\sqrt{n}\mathbf{t}(\boldsymbol{\lambda}, \alpha))$. Then, we have

$$\mathbf{t}(\tilde{\boldsymbol{\lambda}}^e, \alpha) = O_p(n^{-1/2}), \quad (47)$$

$$2[L_n(\hat{\boldsymbol{\psi}}_\alpha, \alpha) - L_{0,n}(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)] = \max_{\mathbf{e} \in \{0,1\}^d} C_n(\sqrt{n}\mathbf{t}(\tilde{\boldsymbol{\lambda}}^e, \alpha)) + o_p(1), \quad (48)$$

where (47) follows from noting that $C_n(\sqrt{n}\mathbf{t}(\tilde{\boldsymbol{\lambda}}^e, \alpha)) \geq o_p(1)$ and using the argument following (34) in the proof of Proposition 2, and (48) holds because (i) $\max_{\mathbf{e} \in \{0,1\}^d} C_n(\sqrt{n}\mathbf{t}(\tilde{\boldsymbol{\lambda}}^e, \alpha)) \geq 2[L_n(\hat{\boldsymbol{\psi}}_\alpha, \alpha) - L_{0,n}(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)] + o_p(1)$ from the definition of $\mathbf{t}(\tilde{\boldsymbol{\lambda}}^e, \alpha)$ and (45), and (ii) $2[L_n(\hat{\boldsymbol{\psi}}_\alpha, \alpha) - L_{0,n}(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)] \geq \max_{\mathbf{e} \in \{0,1\}^d} C_n(\sqrt{n}\mathbf{t}(\tilde{\boldsymbol{\lambda}}^e, \alpha)) + o_p(1)$ from the definition of $\hat{\boldsymbol{\psi}}_\alpha$ and (43).

We proceed to construct a parameter space $\tilde{\Lambda}_\lambda^e$ that is locally equal to the cone Λ_λ^e defined in (18). Observe that (47) and Lemma 4 imply that, with $c(\alpha) := \alpha(1 - \alpha)$,

$$\begin{aligned} (\tilde{\boldsymbol{\lambda}}_{\boldsymbol{\mu}^v}^e)_{ijk} &= c(\alpha) \sum_{(t_1, t_2, t_3) \in p_{12}(i, j, k)} \tilde{\lambda}_{\mu_{t_1}} \tilde{\lambda}_{v_{t_2 t_3}} \quad \text{for all } \mathbf{e}, \\ (\tilde{\boldsymbol{\lambda}}_{\boldsymbol{\mu}^4}^e)_{ijkl} &= \begin{cases} 12c(\alpha) \sum_{(t_1, t_2, t_3, t_4) \in p_{22}(i, j, k, \ell)} \tilde{\lambda}_{v_{t_1 t_2}} \tilde{\lambda}_{v_{t_3 t_4}} + o_p(n^{-1/2}) & \text{if } \mathbf{e} = \mathbf{0}, \\ c(\alpha)b(\alpha) \sum_{(t_1, t_2, t_3, t_4) \in p(i, j, k, \ell)} \tilde{\lambda}_{\mu_{t_1}} \tilde{\lambda}_{\mu_{t_2}} \tilde{\lambda}_{\mu_{t_3}} \tilde{\lambda}_{\mu_{t_4}} + o_p(n^{-1/2}) & \text{if } \mathbf{e} \neq \mathbf{0} \text{ and } e_i = e_j = e_k = e_\ell = 1, \\ o_p(n^{-1/2}) & \text{otherwise.} \end{cases} \end{aligned} \quad (49)$$

Define

$$\tilde{\Lambda}_\lambda^e := \left(\{(\mathbf{t}_{\boldsymbol{\mu}^v}^e)_{ijk}\}_{1 \leq i \leq j \leq k \leq d}, \{(\mathbf{t}_{\boldsymbol{\mu}^4}^e)_{ijkl}\}_{1 \leq i \leq j \leq k \leq \ell \leq d} \right)^\top \in \mathbb{R}^{d_{\boldsymbol{\mu}^v} + d_{\boldsymbol{\mu}^4}}, \quad (50)$$

where $\mathbf{t}_{\boldsymbol{\mu}^v}^e$ and $\mathbf{t}_{\boldsymbol{\mu}^4}^e$ satisfy, for some $\boldsymbol{\lambda} \in \Theta_\lambda$ and

$$\begin{aligned} (\mathbf{t}_{\boldsymbol{\mu}^v}^e)_{ijk} &= c(\alpha) \sum_{(t_1, t_2, t_3) \in p_{12}(i, j, k)} \lambda_{\mu_{t_1}} \lambda_{v_{t_2 t_3}} \quad \text{for all } \mathbf{e}, \\ (\mathbf{t}_{\boldsymbol{\mu}^4}^e)_{ijkl} &= \begin{cases} 12c(\alpha) \sum_{(t_1, t_2, t_3, t_4) \in p_{22}(i, j, k, \ell)} \lambda_{v_{t_1 t_2}} \lambda_{v_{t_3 t_4}} & \text{if } \mathbf{e} = \mathbf{0}, \\ c(\alpha)b(\alpha) \sum_{(t_1, t_2, t_3, t_4) \in p(i, j, k, \ell)} \lambda_{\mu_{t_1}} \lambda_{\mu_{t_2}} \lambda_{\mu_{t_3}} \lambda_{\mu_{t_4}} & \text{if } \mathbf{e} \neq \mathbf{0} \text{ and } e_i = e_j = e_k = e_\ell = 1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (51)$$

Define $\tilde{\mathbf{t}}_\lambda^e$ by $C_n(\sqrt{n}\tilde{\mathbf{t}}_\lambda^e) = \max_{\mathbf{t}_\lambda \in \tilde{\Lambda}_\lambda^e} C_n(\sqrt{n}\mathbf{t}_\lambda)$. Then, it follows from (49) and (51) that

$\max_{\mathbf{e} \in \{0,1\}^d} C_n(\sqrt{n}\tilde{\mathbf{t}}_{\boldsymbol{\lambda}}^{\mathbf{e}}) = \max_{\mathbf{e} \in \{0,1\}^d} C_n(\sqrt{n}\mathbf{t}(\ddot{\boldsymbol{\lambda}}^{\mathbf{e}}, \alpha)) + o_p(1)$. Therefore,

$$2[L_n(\widehat{\boldsymbol{\psi}}_{\alpha}, \alpha) - L_{0,n}(\widehat{\boldsymbol{\gamma}}_0, \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\Sigma}}_0)] = \max_{\mathbf{e} \in \{0,1\}^d} C_n(\sqrt{n}\tilde{\mathbf{t}}_{\boldsymbol{\lambda}}^{\mathbf{e}}) + o_p(1).$$

The asymptotic distribution of the LRTS follows from applying Theorem 3(c) of Andrews (1999) to $\{C_n(\sqrt{n}\tilde{\mathbf{t}}_{\boldsymbol{\lambda}}^{\mathbf{e}})\}_{\mathbf{e} \in \{0,1\}^d}$. First, Assumption 2 of Andrews (1999) holds trivially for $C_n(\sqrt{n}\tilde{\mathbf{t}}_{\boldsymbol{\lambda}}^{\mathbf{e}})$. Second, Assumption 3 of Andrews (1999) holds with $B_T = n^{1/2}$ because $\mathbf{G}_{\boldsymbol{\lambda}, \boldsymbol{\eta}n} \rightarrow_d \mathbf{G}_{\boldsymbol{\lambda}, \boldsymbol{\eta}} \sim N(0, \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}, \boldsymbol{\eta}})$ and $\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}, \boldsymbol{\eta}}$ is nonsingular. Assumption 4 of Andrews (1999) holds from the same argument as (47). Assumption 5 of Andrews (1999) follows from Assumption 5* of Andrews (1999) because $\widetilde{\Lambda}_{\boldsymbol{\lambda}}^{\mathbf{e}}$ is locally equal to the cone $\Lambda_{\boldsymbol{\lambda}}^{\mathbf{e}}$. Therefore, it follows from Theorem 3(c) of Andrews (1999) that

$$\{C_n(\sqrt{n}\tilde{\mathbf{t}}_{\boldsymbol{\lambda}}^{\mathbf{e}})\}_{\mathbf{e} \in \{0,1\}^d} \rightarrow_d \{(\tilde{\mathbf{t}}_{\boldsymbol{\lambda}}^{\mathbf{e}})^{\top} \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}, \boldsymbol{\eta}} \tilde{\mathbf{t}}_{\boldsymbol{\lambda}}^{\mathbf{e}}\}_{\mathbf{e} \in \{0,1\}^d}, \quad (52)$$

and the stated result follows. \square

8.4 Proof of Proposition 6

For $m = 1, \dots, M_0$, let $\mathcal{N}_m^* \subset \Theta_{\boldsymbol{\vartheta}_{M_0+1}}$ be a sufficiently small closed neighborhood of Υ_{1m}^* , such that $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) < \dots < (\boldsymbol{\mu}_{m-1}, \boldsymbol{\Sigma}_{m-1}) < (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), (\boldsymbol{\mu}_{m+1}, \boldsymbol{\Sigma}_{m+1}) < (\boldsymbol{\mu}_{m+2}, \boldsymbol{\Sigma}_{m+2}) < \dots < (\boldsymbol{\mu}_{M_0+1}, \boldsymbol{\Sigma}_{M_0+1})$ and $\alpha_m, \alpha_{m+1} > 0$ hold and $\Upsilon_{1k}^* \notin \mathcal{N}_m^*$ if $k \neq m$. For $\boldsymbol{\vartheta}_{M_0+1} \in \mathcal{N}_m^*$, we introduce the following one-to-one reparameterization, which is similar to (9):

$$\begin{aligned} \beta_m &:= \alpha_m + \alpha_{m+1}, \quad \tau := \alpha_m / (\alpha_m + \alpha_{m+1}), \\ (\beta_1, \dots, \beta_{m-1}, \beta_{m+1}, \dots, \beta_{M_0-1})^{\top} &:= (\alpha_1, \dots, \alpha_{m-1}, \alpha_{m+2}, \dots, \alpha_{M_0})^{\top}, \\ \begin{pmatrix} \boldsymbol{\mu}_m \\ \boldsymbol{\mu}_{m+1} \\ \boldsymbol{v}_m \\ \boldsymbol{v}_{m+1} \end{pmatrix} &= \begin{pmatrix} \boldsymbol{\nu}_{\boldsymbol{\mu}} + (1 - \tau)\boldsymbol{\lambda}_{\boldsymbol{\mu}} \\ \boldsymbol{\nu}_{\boldsymbol{\mu}} - \tau\boldsymbol{\lambda}_{\boldsymbol{\mu}} \\ \boldsymbol{\nu}_{\boldsymbol{v}} + (1 - \tau)(2\boldsymbol{\lambda}_{\boldsymbol{v}} + C_1\boldsymbol{w}(\boldsymbol{\lambda}_{\boldsymbol{\mu}}\boldsymbol{\lambda}_{\boldsymbol{\mu}}^{\top})) \\ \boldsymbol{\nu}_{\boldsymbol{v}} - \tau(2\boldsymbol{\lambda}_{\boldsymbol{v}} + C_2\boldsymbol{w}(\boldsymbol{\lambda}_{\boldsymbol{\mu}}\boldsymbol{\lambda}_{\boldsymbol{\mu}}^{\top})) \end{pmatrix}, \end{aligned} \quad (53)$$

where $\beta_{M_0} = 1 - \sum_{m=1}^{M_0-1} \beta_m$, $C_1 = -(1/3)(1 + \tau)$, and $C_2 = (1/3)(2 - \tau)$, and we suppress the dependence of $(\boldsymbol{\lambda}_{\boldsymbol{\mu}}, \boldsymbol{\nu}_{\boldsymbol{\mu}}, \boldsymbol{\lambda}_{\boldsymbol{v}}, \boldsymbol{\nu}_{\boldsymbol{v}})$ on τ . With this reparameterization, the null restriction $(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = (\boldsymbol{\mu}_{m+1}, \boldsymbol{\Sigma}_{m+1})$ implied by $H_{0,1m}$ holds if and only if $(\boldsymbol{\lambda}_{\boldsymbol{\mu}}, \boldsymbol{\lambda}_{\boldsymbol{v}}) = \mathbf{0}$. Collect the reparameterized parameters except for τ into one vector $\boldsymbol{\psi}_{\tau}^m$, and let $\boldsymbol{\psi}_{\tau}^{m*}$ denote its true value. Define the reparameterized density as

$$\begin{aligned} g^m(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi}_{\tau}^m, \tau) &:= \beta_m \left[\tau f_v \left(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}, \boldsymbol{\nu}_{\boldsymbol{v}} + (1 - \tau)(2\boldsymbol{\lambda}_{\boldsymbol{v}} + C_1\boldsymbol{w}(\boldsymbol{\lambda}_{\boldsymbol{\mu}}\boldsymbol{\lambda}_{\boldsymbol{\mu}}^{\top})) \right) \right. \\ &\quad \left. + (1 - \tau) f_v \left(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}, \boldsymbol{\nu}_{\boldsymbol{\mu}} - \tau\boldsymbol{\lambda}_{\boldsymbol{\mu}}, \boldsymbol{\nu}_{\boldsymbol{v}} - \tau(2\boldsymbol{\lambda}_{\boldsymbol{v}} + C_2\boldsymbol{w}(\boldsymbol{\lambda}_{\boldsymbol{\mu}}\boldsymbol{\lambda}_{\boldsymbol{\mu}}^{\top})) \right) \right] \\ &\quad + \sum_{j=1}^{m-1} \beta_j f_v(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) + \sum_{j=m+1}^{M_0} \beta_j f_v(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}, \boldsymbol{\mu}_{j+1}, \boldsymbol{\Sigma}_{j+1}). \end{aligned}$$

Define the local PMLE of $\boldsymbol{\psi}_\tau^m$ by

$$\widehat{\boldsymbol{\psi}}_\tau^m := \arg \max_{\boldsymbol{\psi}_\tau^m \in \mathcal{N}_m^*} PL_n^m(\boldsymbol{\psi}_\tau^m, \tau), \quad (54)$$

where $PL_n^m(\boldsymbol{\psi}_\tau^m, \tau) := \sum_{i=1}^n \log[g^m(\mathbf{X}_i | \mathbf{Z}_i; \boldsymbol{\psi}_\tau^m, \tau)] + p_n(\boldsymbol{\psi}_\tau^m)$. Because $\boldsymbol{\psi}_\tau^{m*}$ is the only parameter value in \mathcal{N}_m^* that generates true density, $\widehat{\boldsymbol{\psi}}_\tau^m - \boldsymbol{\psi}_\tau^{m*} = o_p(1)$ follows from a straightforward extension of Proposition 3.

Define the penalized LRT statistic for testing $H_{0,1m}$ as $PLR_{n,1m}(\epsilon_\tau) := \max_{\tau \in [\epsilon_\tau, 1-\epsilon_\tau]} 2\{PL_n^m(\widehat{\boldsymbol{\psi}}_\tau^m, \tau) - PL_{0,n}(\widehat{\boldsymbol{\vartheta}}_{M_0})\}$ for some $\epsilon_\tau \in (0, 1/2)$. The stated result holds if

$$(LR_{n,11}(\epsilon_\tau), \dots, LR_{n,1M_0}(\epsilon_\tau))^\top \rightarrow_d (v_1, \dots, v_{M_0})^\top \quad (55)$$

for any $\epsilon_\tau \in (0, 1/2)$, where $v_m := \max_{\mathbf{e} \in \{0,1\}^d} \left((\widehat{\mathbf{t}}_{\boldsymbol{\lambda},m}^{\mathbf{e}})^\top \mathcal{I}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^m \widehat{\mathbf{t}}_{\boldsymbol{\lambda},m}^{\mathbf{e}} \right)$. We proceed to show (55). Observe that as in (13), the first, second, and third derivatives of $\log[g^m(\mathbf{x} | \mathbf{z}; \boldsymbol{\psi}_\tau^m, \tau)]$ w.r.t. $\boldsymbol{\lambda}_\mu$ and its first derivative w.r.t. $\boldsymbol{\lambda}_v$ become zero when evaluated at $\boldsymbol{\psi}_\tau^m = \boldsymbol{\psi}_\tau^{m*}$. Consequently, $PL_n^m(\boldsymbol{\psi}_\tau^m, \tau) - PL_n^m(\boldsymbol{\psi}_\tau^{m*}, \tau)$ admits the same expansion as $L_n(\boldsymbol{\psi}_\alpha, \alpha) - L_n(\boldsymbol{\psi}_\alpha^*, \alpha)$ in Proposition 4 by replacing $(\mathbf{t}(\boldsymbol{\psi}_\alpha, \alpha), \mathbf{s}(\mathbf{x}, \mathbf{z}), \mathcal{I})$ with $(\mathbf{t}_m(\boldsymbol{\psi}_\tau^m, \tau), \mathbf{s}_m(\mathbf{x}, \mathbf{z}), \mathcal{I}^m)$, where $(\mathbf{s}_m(\mathbf{x}, \mathbf{z}), \mathcal{I}^m)$ is defined in the same manner as $(\mathbf{s}(\mathbf{x}, \mathbf{z}), \mathcal{I})$ but using $(\tilde{\mathbf{s}}_\eta, \mathbf{s}_{\mu v}^m, \mathbf{s}_{\mu^4}^m)$ in place of $(\mathbf{s}_\eta, \mathbf{s}_\lambda)$. Then, (55) follows from the repeating the proof of Proposition 5 for each local MLE by replacing \mathbf{G}_n with $\mathbf{G}_{n,m} := \nu_n(\mathbf{s}_m(\mathbf{x}, \mathbf{z}))$ and collecting the results while noting that $(\mathbf{G}_{n,1}^\top, \dots, \mathbf{G}_{n,M_0}^\top)^\top \rightarrow_d (\mathbf{G}_1^\top, \dots, \mathbf{G}_{M_0}^\top)^\top$. \square

8.5 Proof of Proposition 7

Let $\omega_{n,m}^e$ be the sample counterpart of $(\widehat{\mathbf{t}}_{\boldsymbol{\lambda},m}^{\mathbf{e}})^\top \mathcal{I}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^m \widehat{\mathbf{t}}_{\boldsymbol{\lambda},m}^{\mathbf{e}}$ in Proposition 6 such that the local LRT statistic satisfies $2[L_n^h(\widehat{\boldsymbol{\psi}}_\tau^h, \tau) - L_{0,n}(\widehat{\boldsymbol{\vartheta}}_{m_0})] = \max\{\omega_{n,m}^e\} + o_p(1)$, where $\widehat{\boldsymbol{\psi}}_\tau^m$ is the local MLE defined in (54).

For $\tau \in (0, 1)$, define $\boldsymbol{\vartheta}_{M_0+1}^{m*}(\tau) := \{\boldsymbol{\vartheta}_{M_0+1} \in \Upsilon_{1m}^* : \alpha_m/(\alpha_m + \alpha_{m+1}) = \tau\}$, which gives the true density. Observe that from Assumption 2 and $|x| \leq 1 + |x|^3$, we have $p_{nm}(\boldsymbol{\Sigma}_m) - p_n(\boldsymbol{\Sigma}_m^*) = o_p(n^{1/6})|\mathbf{v}_j - \mathbf{v}_j^*| = o_p(1 + n^{1/2}|\mathbf{v}_j - \mathbf{v}_j^*|^3) = o_p(1 + n^{1/2}(|\boldsymbol{\lambda}_v|^3 + |\boldsymbol{\lambda}_\mu|^6))$. Therefore, in view of the bound on the third term in (42) and on the first term in (41) in the proof of Proposition 4, for any $\boldsymbol{\vartheta}_{M_0+1}$ with $\alpha_m/(\alpha_m + \alpha_{m+1}) = \tau \in (0, 1)$ and whose corresponding $\mathbf{t}_m(\boldsymbol{\psi}_\tau^m)$ is $O_p(n^{-1/2})$, we have

$$PL_n(\boldsymbol{\vartheta}_{M_0+1}) - p_n(\boldsymbol{\vartheta}_{M_0+1}) - PL_n(\boldsymbol{\vartheta}_{M_0+1}^{m*}(\tau)) + p_n(\boldsymbol{\vartheta}_{M_0+1}^{m*}(\tau)) = PL_n(\boldsymbol{\vartheta}_{M_0+1}) - PL_n(\boldsymbol{\vartheta}_{M_0+1}^{m*}(\tau)) + o_p(1). \quad (56)$$

First, we show $\text{EM}_n^{h(1)} = \max\{\omega_{n,m}^e\} + o_p(1)$. Because $\boldsymbol{\vartheta}_{M_0+1}^{m*}(\tau_0)$ is the only value of $\boldsymbol{\vartheta}_{M_0+1}$ that yields the true density if $\boldsymbol{\varsigma} \in \Omega_m^*$ and $\alpha_m/(\alpha_m + \alpha_{m+1}) = \tau_0$, $\boldsymbol{\vartheta}_{M_0+1}^{m(1)}(\tau_0)$ equals a reparameterized penalized local MLE in the neighborhood of $\boldsymbol{\vartheta}_{M_0+1}^{m*}(\tau_0)$. Therefore, $\text{EM}_n^{h(1)} = \max\{\omega_{n,m}^1, \omega_{n,m}^2\} + o_p(1)$ follows from the proof of Proposition 6 and (56).

We proceed to show that $\text{EM}_n^{m(K)} = \max\{\omega_{n,m}^e\} + o_p(1)$. Because a generalized EM

step never decreases the likelihood value (Dempster et al., 1977), we have $PL_n(\boldsymbol{\vartheta}_{M_0+1}^{m(K)}(\tau_0)) \geq PL_n(\boldsymbol{\vartheta}_{M_0+1}^{h(1)}(\tau_0))$. Therefore, it follows from induction that $\boldsymbol{\vartheta}_{M_0+1}^{m(K)}(\tau_0) - \boldsymbol{\vartheta}_{M_0+1}^{m^*}(\tau_0) = o_p(1)$ for any finite K . Let $\tilde{\boldsymbol{\vartheta}}_{M_0+1}^m(\tau^{(K)})$ be the maximizer of $PL_n(\boldsymbol{\vartheta}_{M_0+1})$ under the constraint $\alpha_m/(\alpha_m + \alpha_{m+1}) = \tau^{(K)}$ in an arbitrary small closed neighborhood of $\boldsymbol{\vartheta}_{M_0+1}^{m^*}(\tau^{(K)})$; then, we have $PL_n(\tilde{\boldsymbol{\vartheta}}_{M_0+1}^m(\tau^{(K)})) \geq PL_n(\boldsymbol{\vartheta}_{M_0+1}^{m(K)}(\tau_0)) + o_p(1)$ from the consistency of $\boldsymbol{\vartheta}_{M_0+1}^{m(K)}(\tau_0)$. Thus, $2[PL_n(\boldsymbol{\vartheta}_{M_0+1}^{m(K)}(\tau_0)) - L_{0,n}(\tilde{\boldsymbol{\vartheta}}_{M_0})] = \max\{\omega_{n,m}^e\} + o_p(1)$ holds because both $2[PL_n(\tilde{\boldsymbol{\vartheta}}_{M_0+1}^m(\tau^{(K)})) - L_{0,n}(\tilde{\boldsymbol{\vartheta}}_{M_0})]$ and $2[PL_n(\boldsymbol{\vartheta}_{M_0+1}^{m(1)}(\tau_0)) - L_{0,n}(\tilde{\boldsymbol{\vartheta}}_{M_0})]$ can be written as $\max\{\omega_{n,m}^e\} + o_p(1)$. Further, because $PL_n(\boldsymbol{\vartheta}_{M_0+1}^{m(K)}(\tau_0)) \geq PL_n(\boldsymbol{\vartheta}_{M_0+1}^{m(1)}(\tau_0)) \geq PL_n(\boldsymbol{\vartheta}_{M_0+1}^{m^*}(\tau_0)) + o_p(1)$, it follows from applying Proposition 5 to $\boldsymbol{\vartheta}_{M_0+1}^{m(K)}(\tau_0)$ in conjunction with (56) that $EM_n^{m(K)} = \max\{\omega_{n,m}^e\} + o_p(1)$ holds for all m . The stated result follows from the definition of $EM_n^{(K)}$. \square

9 Auxiliary results and their proofs

Lemma 1. Let $f_v(\mathbf{x}; \boldsymbol{\mu}, \mathbf{v}) := (2\pi)^{-d/2} (\det \mathbf{S}(\mathbf{v}))^{-1/2} \exp(-(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S}(\mathbf{v})^{-1} (\mathbf{x} - \boldsymbol{\mu})/2)$ denote the density of a d -variate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)^\top$ and variance $\mathbf{S}(\mathbf{v})$ with $\mathbf{v} = \{v_{ij}\}_{1 \leq i \leq j \leq d}$ as specified in (8). Then, the following holds for any $t_1, t_2, t_3, t_4, t_5, t_6 \in \{1, \dots, d\}$:

$$\begin{aligned} \frac{\partial f_v(\mathbf{x}; \boldsymbol{\mu}, \mathbf{v})}{\partial v_{t_1 t_2}} &= \frac{1}{2} \frac{\partial^2 f_v(\mathbf{x}; \boldsymbol{\mu}, \mathbf{v})}{\partial \mu_{t_1} \partial \mu_{t_2}}, & \frac{\partial^2 f_v(\mathbf{x}; \boldsymbol{\mu}, \mathbf{v})}{\partial v_{t_1 t_2} \partial v_{t_3 t_4}} &= \frac{1}{4} \frac{\partial^4 f_v(\mathbf{x}; \boldsymbol{\mu}, \mathbf{v})}{\partial \mu_{t_1} \partial \mu_{t_2} \partial \mu_{t_3} \partial \mu_{t_4}}, \\ \frac{\partial^3 f_v(\mathbf{x}; \boldsymbol{\mu}, \mathbf{v})}{\partial v_{t_1 t_2} \partial v_{t_3 t_4} \partial v_{t_5 t_6}} &= \frac{1}{8} \frac{\partial^6 f_v(\mathbf{x}; \boldsymbol{\mu}, \mathbf{v})}{\partial \mu_{t_1} \partial \mu_{t_2} \partial \mu_{t_3} \partial \mu_{t_4} \partial \mu_{t_5} \partial \mu_{t_6}}. \end{aligned}$$

Proof. Henceforth, we suppress $(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $(\mathbf{x}; \boldsymbol{\mu}, \mathbf{v})$ and from $f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $f_v(\mathbf{x}; \boldsymbol{\mu}, \mathbf{v})$ unless confusions might arise. In view of the definition of $\mathbf{S}(\mathbf{v})$ in (8), the following holds for any function $g(\boldsymbol{\Sigma})$ of $\boldsymbol{\Sigma}$:

$$\frac{\partial g(\mathbf{S}(\mathbf{v}))}{\partial v_{t_1 t_2}} = \frac{\partial g(\boldsymbol{\Sigma})/\partial \Sigma_{t_1 t_2} + \partial g(\boldsymbol{\Sigma})/\partial \Sigma_{t_2 t_1}}{2} = \frac{\partial g(\boldsymbol{\Sigma})}{\partial \Sigma_{t_1 t_2}}. \quad (57)$$

Let \mathbf{s}_i denote the i th column of $\boldsymbol{\Sigma}^{-1}$, and let s_{ij} denote the (i, j) th element of $\boldsymbol{\Sigma}^{-1}$. A direct calculation gives $\partial^2 f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})/\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}^\top = -\boldsymbol{\Sigma}^{-1} f + \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} f$ and $\partial f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})/\partial \boldsymbol{\Sigma} = -(1/2)\boldsymbol{\Sigma}^{-1} f + (1/2)\boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} f$. Therefore, the first result follows immediately from (57).

To prove the second result, we first derive $\partial^4 f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})/\partial \mu_{t_1} \partial \mu_{t_2} \partial \mu_{t_3} \partial \mu_{t_4}$. Noting that $\partial \mathbf{s}_j^\top (\mathbf{x} - \boldsymbol{\mu})/\partial \mu_i = -s_{ji}$ and $\partial f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})/\partial \mu_i = \mathbf{s}_i^\top (\mathbf{x} - \boldsymbol{\mu}) f$ and differentiating $\partial^2 f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})/\partial \mu_{t_1} \partial \mu_{t_2} = [-s_{t_1 t_2} + \mathbf{s}_{t_1}^\top (\mathbf{x} - \boldsymbol{\mu}) \mathbf{s}_{t_2}^\top (\mathbf{x} - \boldsymbol{\mu})] f$ with respect to μ_{t_3} and μ_{t_4} , we obtain

$$\frac{\partial^4 f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \mu_{t_1} \partial \mu_{t_2} \partial \mu_{t_3} \partial \mu_{t_4}} = \left(\sum_{\{i,j\}, \{k,\ell\}} s_{t_i t_j} s_{t_k t_\ell} - \sum_{\{i,j\}, \{k\}, \{\ell\}} s_{t_i t_j} \mathbf{s}_{t_k}^\top (\mathbf{x} - \boldsymbol{\mu}) \mathbf{s}_{t_\ell}^\top (\mathbf{x} - \boldsymbol{\mu}) + \prod_{i=1}^4 \mathbf{s}_{t_i}^\top (\mathbf{x} - \boldsymbol{\mu}) \right) f, \quad (58)$$

where $\sum_{\{i,j\}, \{k,\ell\}}$ denotes the sum over all 3 possible partitions of $\{1, 2, 3, 4\}$ into $\{\{i, j\}, \{k, \ell\}\}$, and $\sum_{\{i,j\}, \{k\}, \{\ell\}}$ denotes the sum over all 6 possible partitions of $\{1, 2, 3, 4\}$ into three sets

$\{\{i, j\}, \{k\}, \{\ell\}\}$. Recall that

$$\frac{\partial f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \Sigma_{t_1 t_2}} = (1/2)[-s_{t_1 t_2} + \mathbf{s}_{t_1}^\top (\mathbf{x} - \boldsymbol{\mu}) \mathbf{s}_{t_2}^\top (\mathbf{x} - \boldsymbol{\mu})] f. \quad (59)$$

Let $\mathbf{1}_i$ denote a $d \times 1$ vector whose elements are 0 except for the i th element, which is 1. We then have $s_{t_1 t_2} = \mathbf{1}_{t_1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}_{t_2}$ and $\mathbf{s}_{t_1} = \boldsymbol{\Sigma}^{-1} \mathbf{1}_{t_1}$. Using the symmetry of $\boldsymbol{\Sigma}$, we obtain

$$\begin{aligned} \frac{\partial s_{t_1 t_2}}{\partial \Sigma_{t_3 t_4}} &= \frac{\partial (s_{t_1 t_2} + s_{t_2 t_1})/2}{\partial \Sigma_{t_3 t_4}} \\ &= \frac{\partial}{\partial \Sigma_{t_3 t_4}} \left(\mathbf{1}_{t_1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}_{t_2} + \mathbf{1}_{t_2}^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}_{t_1} \right) / 2 \\ &= -(1/2) \left(\boldsymbol{\Sigma}^{-1} \mathbf{1}_{t_1} \mathbf{1}_{t_2}^\top \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1} \mathbf{1}_{t_2} \mathbf{1}_{t_1}^\top \boldsymbol{\Sigma}^{-1} \right)_{t_3 t_4} \\ &= -(1/2) (s_{t_3 t_1} s_{t_2 t_4} + s_{t_3 t_2} s_{t_1 t_4}), \end{aligned}$$

and

$$\begin{aligned} \frac{\partial \mathbf{s}_{t_1}^\top (\mathbf{x} - \boldsymbol{\mu})}{\partial \Sigma_{t_3 t_4}} &= \frac{\partial}{\partial \Sigma_{t_3 t_4}} \left(\mathbf{1}_{t_1}^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}_{t_1} \right) / 2 \\ &= -(1/2) \left(\boldsymbol{\Sigma}^{-1} \mathbf{1}_{t_1} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \mathbf{1}_{t_1}^\top \boldsymbol{\Sigma}^{-1} \right)_{t_3 t_4} \\ &= -(1/2) \left(s_{t_3 t_1} (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{s}_{t_4} + \mathbf{s}_{t_3}^\top (\mathbf{x} - \boldsymbol{\mu}) s_{t_1 t_4} \right). \end{aligned}$$

Therefore, taking the derivative of the right hand side of (59) with respect to $\Sigma_{t_3 t_4}$ gives

$$\begin{aligned} \frac{\partial^2 f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \Sigma_{t_1 t_2} \partial \Sigma_{t_3 t_4}} &= \frac{1}{4} \left[s_{t_3 t_1} s_{t_2 t_4} + s_{t_3 t_2} s_{t_1 t_4} - \left(s_{t_3 t_1} (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{s}_{t_4} + \mathbf{s}_{t_3}^\top (\mathbf{x} - \boldsymbol{\mu}) s_{t_1 t_4} \right) \mathbf{s}_{t_2}^\top (\mathbf{x} - \boldsymbol{\mu}) \right. \\ &\quad \left. - \mathbf{s}_{t_1}^\top (\mathbf{x} - \boldsymbol{\mu}) \left(s_{t_3 t_2} (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{s}_{t_4} + \mathbf{s}_{t_3}^\top (\mathbf{x} - \boldsymbol{\mu}) s_{t_2 t_4} \right) \right] f \\ &\quad + \frac{1}{2} \left(-s_{t_1 t_2} + \mathbf{s}_{t_1}^\top (\mathbf{x} - \boldsymbol{\mu}) \mathbf{s}_{t_2}^\top (\mathbf{x} - \boldsymbol{\mu}) \right) \frac{\partial f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \Sigma_{t_3 t_4}} \\ &= \frac{1}{4} \left(\sum_{\{i, j\}, \{k, \ell\}} s_{t_i t_j} s_{t_k t_\ell} - \sum_{\{i, j\}, \{k\}, \{\ell\}} s_{t_i t_j} \mathbf{s}_{t_k}^\top (\mathbf{x} - \boldsymbol{\mu}) \mathbf{s}_{t_\ell}^\top (\mathbf{x} - \boldsymbol{\mu}) + \prod_{i=1}^4 \mathbf{s}_{t_i}^\top (\mathbf{x} - \boldsymbol{\mu}) \right) f. \end{aligned} \quad (60)$$

Comparing this with (58) and using (57) gives the second result.

For the third result, differentiating (58) with respect to μ_{t_5} and μ_{t_6} gives

$$\begin{aligned} & \frac{\partial^6 f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \mu_{t_1} \partial \mu_{t_2} \partial \mu_{t_3} \partial \mu_{t_4} \partial \mu_{t_5} \partial \mu_{t_6}} \\ &= \left(- \sum_{\{i,j\},\{k,\ell\},\{m,n\}} s_{t_i t_j} s_{t_k t_\ell} s_{t_m t_n} + \sum_{\{i,j\},\{k,\ell\},\{m\},\{n\}} s_{t_i t_j} s_{t_k t_\ell} \mathbf{s}_{t_m}^\top (\mathbf{x} - \boldsymbol{\mu}) \mathbf{s}_{t_n}^\top (\mathbf{x} - \boldsymbol{\mu}) \right. \\ & \quad \left. - \sum_{\{i,j\},\{k,\ell,m,n\}} s_{t_i t_j} \mathbf{s}_{t_k}^\top (\mathbf{x} - \boldsymbol{\mu}) \mathbf{s}_{t_\ell}^\top (\mathbf{x} - \boldsymbol{\mu}) \mathbf{s}_{t_m}^\top (\mathbf{x} - \boldsymbol{\mu}) \mathbf{s}_{t_n}^\top (\mathbf{x} - \boldsymbol{\mu}) + \prod_{i=1}^6 \mathbf{s}_{t_i}^\top (\mathbf{x} - \boldsymbol{\mu}) \right) f, \end{aligned} \quad (61)$$

where $\sum_{\{i,j\},\{k,\ell\},\{m,n\}}$ denotes the sum over all 15 possible partitions of $\{1, 2, 3, 4, 5, 6\}$ into $\{\{i, j\}, \{k, \ell\}, \{m, n\}\}$, $\sum_{\{i,j\},\{k,\ell\},\{m\},\{n\}}$ denotes the sum over all 45 possible partitions of $\{1, 2, 3, 4, 5, 6\}$ into three sets $\{\{i, j\}, \{k, \ell\}, \{m\}, \{n\}\}$, and $\sum_{\{i,j\},\{k,\ell,m,n\}}$ denotes the sum over all 15 possible partitions of $\{1, 2, 3, 4, 5, 6\}$ into $\{\{i, j\}, \{k, \ell, m, n\}\}$. Differentiating (60) with respect to $\Sigma_{t_5 t_6}$ gives (61) divided by 8, and the third result follows. \square

Lemma 2. *Let $f(x; \boldsymbol{\beta})$ be the density function of a random variable X with parameter $\boldsymbol{\beta}$. Then, $E_{\boldsymbol{\beta}^*}[\nabla_{\boldsymbol{\beta}^*}^k f(x; \boldsymbol{\beta}^*)/f(x; \boldsymbol{\beta}^*)] = 0$ if $f(x; \boldsymbol{\beta})$ is k times differentiable in $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}^*$.*

Proof. The stated result follows from differentiating both sides of $\int f(x; \boldsymbol{\beta}) dx = 1$ k times with respect to $\boldsymbol{\beta}$ and evaluating at $\boldsymbol{\beta}^*$. \square

Lemma 3. *Suppose that $g(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi}_\alpha, \alpha)$ is given by (11), where $\boldsymbol{\psi} = (\boldsymbol{\eta}^\top, \boldsymbol{\lambda}_\mu, \boldsymbol{\lambda}_v)^\top$ and $\boldsymbol{\eta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\nu}_\mu, \boldsymbol{\nu}_v)^\top$. Let g^* and ∇g^* denote $g(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi}_\alpha, \alpha)$ and $\nabla g(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi}_\alpha, \alpha)$ evaluated at $(\boldsymbol{\psi}_\alpha^*, \alpha)$, respectively. Let ∇f^* denote $\nabla f(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. Then, with $b(\alpha) := -(2/3)(\alpha^2 - \alpha + 1) < 0$,*

- (a) for $k = 1, 2, 3$ and $\ell = 0, 1, \dots$, $\nabla_{\boldsymbol{\lambda}_\mu^{\otimes k} \otimes \boldsymbol{\eta}^{\otimes \ell}} g^* = \mathbf{0}$;
- (b) $\nabla_{\lambda_{\mu_i} \lambda_{\mu_j} \lambda_{\mu_k} \lambda_{\mu_\ell}} g^* = \alpha(1 - \alpha)b(\alpha) \nabla_{\mu_i \mu_j \mu_k \mu_\ell} f^*$;
- (c) for $\ell = 0, 1, \dots$, $\nabla_{\boldsymbol{\lambda}_v \otimes \boldsymbol{\eta}^{\otimes \ell}} g^* = \mathbf{0}$;
- (d) $\nabla_{\lambda_{\mu_i} \lambda_{v_j k}} g^* = \alpha(1 - \alpha) \nabla_{\mu_i \mu_j \mu_k} f^*$;
- (e) $\nabla_{\lambda_{v_j j} \lambda_{v_k \ell}} g^* = \alpha(1 - \alpha) \nabla_{\mu_i \mu_j \mu_k \mu_\ell} f^*$.

Proof. We prove part (a) for $\ell = 0$ first. Suppress all arguments in $g(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi}_\alpha, \alpha)$ and $f_v(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}, \boldsymbol{\mu}, \mathbf{v})$ except for $\boldsymbol{\lambda}_\mu$, and rewrite as follows:

$$g(\boldsymbol{\lambda}_\mu) = \alpha f_v((1 - \alpha)\boldsymbol{\lambda}_\mu, (1 - \alpha)C_1 \mathbf{w}(\boldsymbol{\lambda}_\mu \boldsymbol{\lambda}_\mu^\top)) + (1 - \alpha) f_v(-\alpha \boldsymbol{\lambda}_\mu, -\alpha C_2 \mathbf{w}(\boldsymbol{\lambda}_\mu \boldsymbol{\lambda}_\mu^\top)). \quad (62)$$

For a composite function $h(\mathbf{a}, \mathbf{r}(\mathbf{a}))$ of a $d \times 1$ vector $\mathbf{a} = (a_1, \dots, a_d)^\top$, the following result holds:

$$\begin{aligned} \nabla_{a_{i_1} \dots a_{i_k}} h(\mathbf{a}, \mathbf{r}(\mathbf{a})) &= \{(\nabla_{a_{i_1}} + \nabla_{u_{i_1}}) \cdots (\nabla_{a_{i_k}} + \nabla_{u_{i_k}})\} h(\mathbf{a}, \mathbf{r}(\mathbf{u}))|_{\mathbf{u}=\mathbf{a}} \\ &= \sum_{j=0}^k \sum_{p(j, \{i_1, \dots, i_k\})} \nabla_{u_{t_1} \dots u_{t_j} a_{t_{j+1}} \dots a_{t_k}} h(\mathbf{a}, \mathbf{r}(\mathbf{u}))|_{\mathbf{u}=\mathbf{a}}, \end{aligned} \quad (63)$$

where $\sum_{p(j, \{i_1, \dots, i_k\})}$ denotes the sum over all the partitions of $\{i_1, \dots, i_k\}$ into two sets $\{t_1, \dots, t_j\}$ and $\{t_{j+1}, \dots, t_k\}$. Because applying (63) to the right hand side of (62) gives the derivatives of $g(\boldsymbol{\lambda}_\mu)$, we derive $\nabla_{u_{t_1} \dots u_{t_j}} f_v((1-\alpha)\boldsymbol{\lambda}_\mu, (1-\alpha)C_1 \mathbf{w}(\mathbf{u}\mathbf{u}^\top))|_{\mathbf{u}=\mathbf{0}}$. Let $\tilde{c} := (1-\alpha)C_1$. For notational convenience, if $i > j$, define $\nabla_{v_{ij}} h(\mathbf{v}) := \nabla_{v_{ji}} h(\mathbf{v})$ for any function $h(\mathbf{v})$. Using the fact $\nabla_{u_k} w_{ij}(\mathbf{u}\mathbf{u}^\top) = 2u_i \mathbb{I}\{j = k\} + 2u_j \mathbb{I}\{i = k\}$, we obtain

$$\begin{aligned} \nabla_{u_{t_1}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top)) &= \sum_{i=1}^d \sum_{j=i}^d \nabla_{v_{ij}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top)) \tilde{c} \nabla_{u_{t_1}} w_{ij}(\mathbf{u}\mathbf{u}^\top) \\ &= 2 \sum_{i=1}^{t_1} \nabla_{v_{it_1}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top)) \tilde{c} u_i + 2 \sum_{j=t_1+1}^d \nabla_{v_{t_1 j}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top)) \tilde{c} u_j \\ &= 2 \sum_{i=1}^d \nabla_{v_{t_1 i}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top)) \tilde{c} u_i. \end{aligned}$$

Differentiating the right hand side with respect to u_{t_2} gives

$$\nabla_{u_{t_1} u_{t_2}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top)) = 4 \sum_{i=1}^d \sum_{j=1}^d \nabla_{v_{t_1 i} v_{t_2 j}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top)) \tilde{c}^2 u_i u_j + 2 \nabla_{v_{t_1 t_2}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top)) \tilde{c}.$$

Differentiating the right hand side with respect to u_{t_3} gives

$$\begin{aligned} \nabla_{u_{t_1} u_{t_2} u_{t_3}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top)) &= 8 \sum_{i=1}^d \sum_{j=1}^d \sum_{k=1}^d \nabla_{v_{t_1 i} v_{t_2 j} v_{t_3 k}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top)) \tilde{c}^3 u_i u_j u_k \\ &\quad + 4 \sum_{i=1}^d \nabla_{v_{t_1 i} v_{t_2 t_3}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top)) \tilde{c}^2 u_i + 4 \sum_{j=1}^d \nabla_{v_{t_1 t_3} v_{t_2 j}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top)) \tilde{c}^2 u_j \\ &\quad + 4 \sum_{k=1}^d \nabla_{v_{t_1 t_2} v_{t_3 k}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top)) \tilde{c}^2 u_k. \end{aligned}$$

Finally, evaluating these derivatives at $\mathbf{u} = \mathbf{0}$ and differentiating $\nabla_{u_{t_1} u_{t_2} u_{t_3}} f_v(\cdot, \tilde{c}\mathbf{w}(\mathbf{u}\mathbf{u}^\top))$ with

respect to u_{t_4} and evaluating at $\mathbf{u} = \mathbf{0}$ gives

$$\begin{aligned}
& \nabla_{u_{t_1}} f_v(\cdot, \tilde{\mathbf{w}}(\mathbf{u}\mathbf{u}^\top))|_{\mathbf{u}=\mathbf{0}} = 0, \\
& \nabla_{u_{t_1}u_{t_2}} f_v(\cdot, \tilde{\mathbf{w}}(\mathbf{u}\mathbf{u}^\top))|_{\mathbf{u}=\mathbf{0}} = 2\tilde{c}\nabla_{v_{t_1}t_2} f_v(\cdot, \tilde{\mathbf{w}}(\mathbf{u}\mathbf{u}^\top)), \\
& \nabla_{u_{t_1}u_{t_2}u_{t_3}} f_v(\cdot, \tilde{\mathbf{w}}(\mathbf{u}\mathbf{u}^\top))|_{\mathbf{u}=\mathbf{0}} = 0, \\
& \nabla_{u_{t_1}u_{t_2}u_{t_3}u_{t_4}} f_v(\cdot, \tilde{\mathbf{w}}(\mathbf{u}\mathbf{u}^\top))|_{\mathbf{u}=\mathbf{0}} = 4\tilde{c}^2\nabla_{v_{t_1}t_4v_{t_2}t_3} f_v(\cdot, \mathbf{0}) + 4\tilde{c}^2\nabla_{v_{t_1}t_3v_{t_2}t_4} f_v(\cdot, \mathbf{0}) \\
& \quad + 4\tilde{c}^2\nabla_{v_{t_1}t_2v_{t_3}t_4} f_v(\cdot, \mathbf{0}),
\end{aligned} \tag{64}$$

and a similar result holds for $\nabla_{u_{t_1}\dots u_{t_j}} \lambda_\mu^{k-j} f_v((1-\alpha)\lambda_\mu, (1-\alpha)C_1\mathbf{w}(\mathbf{u}\mathbf{u}^\top))|_{\mathbf{u}=\mathbf{0}}$ and $\nabla_{u_{t_1}\dots u_{t_j}} \lambda_\mu^{k-j} f(-\alpha\lambda_\mu, -\alpha C_2\mathbf{w}(\mathbf{u}\mathbf{u}^\top))|_{\mathbf{u}=\mathbf{0}}$.

Differentiating (62) with respect to λ_μ and using (63), (64), $C_1 - C_2 = -1$, $3((1-\alpha)C_1 + \alpha C_2) = 2\alpha - 1$, and Lemma 1, we obtain

$$\begin{aligned}
& \nabla_{\lambda_\mu} g(\mathbf{0}) = \mathbf{0}, \\
& \nabla_{\lambda_{\mu_i}\lambda_{\mu_j}} g(\mathbf{0}) = \alpha(1-\alpha)\nabla_{\mu_i\mu_j} f_v(\mathbf{0}, \mathbf{0}) + 2\alpha(1-\alpha)(C_1 - C_2)\nabla_{v_{ij}} f_v(\mathbf{0}, \mathbf{0}) = 0, \\
& \nabla_{\lambda_{\mu_i}\lambda_{\mu_j}\lambda_{\mu_k}} g(\mathbf{0}) = \alpha(1-\alpha)(1-2\alpha)\nabla_{\mu_i\mu_j\mu_k} f_v(\mathbf{0}, \mathbf{0}) \\
& \quad + 3\alpha(1-\alpha)((1-\alpha)C_1 + \alpha C_2)2\nabla_{\mu_i v_{jk}} f_v(\mathbf{0}, \mathbf{0}) = 0,
\end{aligned}$$

and part (a) for $\ell = 0$ follows. Repeating the same argument with $\nabla_{\eta^{\otimes \ell}} g(\lambda_\mu, \eta)$ gives part (a) for $\ell \geq 1$.

For part (b), differentiating (62) and using (63), (64), and Lemma 1 gives

$$\begin{aligned}
& \nabla_{\lambda_{\mu_i}\lambda_{\mu_j}\lambda_{\mu_k}\lambda_{\mu_\ell}} g(\mathbf{0}) \\
& = \alpha(1-\alpha)[(1-\alpha)^3 + \alpha^3]\nabla_{\mu_i\mu_j\mu_k\mu_\ell} f_v(\mathbf{0}, \mathbf{0}) + 6\alpha(1-\alpha)((1-\alpha)^2 C_1 - \alpha^2 C_2)2\nabla_{\mu_i\mu_j v_{k\ell}} f_v(\mathbf{0}, \mathbf{0}) \\
& \quad + 12\alpha(1-\alpha)((1-\alpha)C_1^2 + \alpha C_2^2)\nabla_{v_{ij}v_{k\ell}} f_v(\mathbf{0}, \mathbf{0}) \\
& = \alpha(1-\alpha)[-(2/3)(\alpha^2 - \alpha + 1)]\nabla_{\mu_i\mu_j\mu_k\mu_\ell} f_v(\mathbf{0}, \mathbf{0}),
\end{aligned}$$

and the stated result follows because $\nabla_{\mu_i\mu_j\mu_k\mu_\ell} f_v(\mathbf{0}, \mathbf{0}) = \nabla_{\mu_i\mu_j\mu_k\mu_\ell} f(\mathbf{0}, \mathbf{0})$. Part (c) follows from a direct calculation.

Parts (d) and (e) follow from direct calculation and using (63), (64) and Lemma 1. \square

Lemma 4. *Let $\mathbf{e} \in \{0, 1\}^d$, and suppose $\lambda = (\lambda_\mu^\top, \lambda_v^\top)^\top \in \Theta_\lambda^{\mathbf{e}}$ satisfies $\mathbf{t}(\lambda, \alpha) = O_p(n^{-1/2})$ for some $\alpha \in (0, 1)$, where $\Theta_\lambda^{\mathbf{e}}$ and $\mathbf{t}(\lambda, \alpha)$ are defined in (46) and (15), respectively. Then, the following result holds for $1 \leq i \leq d$:*

- (a) *If $e_i = 1$, then $\lambda_{\mu_i}^{\mathbf{e}} = O_p(n^{-1/8})$;*
 - (b) *If $e_i = 0$, then $\lambda_{\mu_i}^{\mathbf{e}} = O_p(n^{-1/8}(\log n)^{-1})$;*
 - (c) *If $e_i = 1$ for any i , then $\lambda_v^{\mathbf{e}} = O_p(n^{-3/8}(\log n)^3)$.*
- $$\tag{65}$$

Proof. Observe that $\mathbf{t}(\boldsymbol{\lambda}, \alpha) = O_p(n^{-1/2})$ implies

$$\lambda_{\mu v_{ii}}^e = \alpha(1 - \alpha)\lambda_{\mu_i}^e \lambda_{v_{ii}}^e = O_p(n^{-1/2}), \quad (66)$$

$$\lambda_{\mu v_{ij}}^e = \alpha(1 - \alpha)(\lambda_{\mu_i}^e \lambda_{v_{ij}}^e + \lambda_{\mu_j}^e \lambda_{v_{ii}}^e) = O_p(n^{-1/2}), \quad (67)$$

$$\lambda_{\mu v_{ijk}}^e = \alpha(1 - \alpha)(\lambda_{\mu_i}^e \lambda_{v_{ji}}^e + \lambda_{\mu_j}^e \lambda_{v_{ik}}^e + \lambda_{\mu_k}^e \lambda_{v_{ij}}^e) = O_p(n^{-1/2}), \quad (68)$$

$$\lambda_{\mu_{iii}^4}^e = \alpha(1 - \alpha)[12(\lambda_{v_{ii}}^e)^2 + b(\alpha)(\lambda_{\mu_i}^e)^4] = O_p(n^{-1/2}). \quad (69)$$

First, observe that $\lambda_{v_{ii}}^e = O_p(n^{-3/8} \log n)$ if $e_i = 1$ in view of $|\lambda_{\mu_i}^e| \geq n^{-1/8}(\log n)^{-1}$ and (66). Part (a) follows from substituting this to (69). Part (b) follows from the definition of Θ_{λ}^{ξ} . We prove part (c) by dividing part (c) into the following six cases, where i, j, k are all distinct; (c1) $\lambda_{v_{ii}}^e = O_p(n^{-3/8} \log n)$ if $e_i = 1$; (c2) $\lambda_{v_{ij}}^e = O_p(n^{-3/8}(\log n)^2)$ if $(e_i, e_j) = (1, 1)$; (c3) $\lambda_{v_{ij}}^e = O_p(n^{-3/8}(\log n)^2)$ if $(e_i, e_j) = (1, 0)$ or $(0, 1)$; (c4) $\lambda_{v_{ii}}^e = O_p(n^{-3/8}(\log n)^2)$ if $(e_i, e_j) = (0, 1)$; (c5) $\lambda_{v_{ij}}^e = O_p(n^{-3/8}(\log n)^3)$ if $(e_i, e_j, e_k) = (0, 0, 1)$. (c1) is already proven. (c2) holds because we have $|\lambda_{\mu_i}^e| \geq n^{-1/8}(\log n)^{-1}$ and $\lambda_{\mu_i}^e \lambda_{v_{ij}}^e = O_p(n^{-1/2} \log n)$, which follows from (67) and parts (a)(c1). For (c3), observe that, when $(e_i, e_j) = (1, 0)$, we have $\lambda_{\mu_i}^e \lambda_{v_{ij}}^e = O_p(n^{-1/2} \log n)$ from (67) and parts (b)(c1). Therefore, (c3) holds because $|\lambda_{\mu_i}^e| \geq n^{-1/8}(\log n)^{-1}$. When $(e_i, e_j) = (0, 1)$, (c3) is proven similarly by using $\lambda_{\mu v_{ijj}}^e = O_p(n^{-1/2})$ in place of (67). For (c4), observe that $\lambda_{\mu_j}^e \lambda_{v_{ii}}^e = O_p(n^{-1/2} \log n)$ from (67) and parts (b)(c2). Therefore, (c4) holds because $|\lambda_{\mu_j}^e| \geq n^{-1/8}(\log n)^{-1}$. Finally, (c5) holds because $|\lambda_{\mu_k}^e| \geq n^{-1/8}(\log n)^{-1}$ and $\lambda_{\mu_k}^e \lambda_{v_{ij}}^e = O_p(n^{-1/2}(\log n)^2)$, which follows from (68) and parts (b)(c3). \square

Proposition 8. *Suppose that Assumptions 2, 3, and 5 hold. If $\boldsymbol{\vartheta}_{m_0+1}^{m(k)}(\tau_0) - \boldsymbol{\vartheta}_{m_0+1}^{m^*}(\tau_0) = o_p(1)$, then $\alpha_m^{(k+1)}/[\alpha_m^{(k+1)} + \alpha_{m+1}^{(k+1)}] - \tau_0 = o_p(1)$.*

Proof. We suppress (τ_0) from $\boldsymbol{\vartheta}_{M_0+1}^{m(k)}(\tau_0)$ and $\boldsymbol{\vartheta}_{M_0+1}^{m^*}(\tau_0)$. The proof is similar to the proof of Lemma 3 of Li and Chen (2010). Let $f_i(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $f_i(\boldsymbol{\vartheta}_{M_0+1})$ denote $f(\mathbf{X}_i | \mathbf{Z}_i; \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $f_{M_0+1}(\mathbf{X}_i | \mathbf{Z}_i; \boldsymbol{\vartheta}_{M_0+1})$, respectively. Applying a Taylor expansion to $\alpha_m^{(k+1)} = n^{-1} \sum_{i=1}^n w_{ih}^{(k)}$ and using $\boldsymbol{\vartheta}_{M_0+1}^{m(k)} - \boldsymbol{\vartheta}_{M_0+1}^{m^*} = o_p(1)$, we obtain

$$\alpha_m^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \frac{\alpha_m^{(k)} f_i(\boldsymbol{\gamma}^{(k)}, \boldsymbol{\mu}_m^{(k)}, \boldsymbol{\Sigma}_m^{(k)})}{f_i(\boldsymbol{\vartheta}_{M_0+1}^{h(k)})} = \frac{1}{n} \sum_{i=1}^n \frac{\tau_0 \alpha_m^* f_i(\boldsymbol{\gamma}^*, \boldsymbol{\mu}_m^*, \boldsymbol{\Sigma}_m^*)}{f_i(\boldsymbol{\vartheta}_{M_0+1}^{m^*})} + o_p(1) = \tau_0 \alpha_m^* + o_p(1),$$

where the last equality follows from $E[f_i(\boldsymbol{\gamma}^*, \boldsymbol{\mu}_m^*, \boldsymbol{\Sigma}_m^*)/f_i(\boldsymbol{\vartheta}_{M_0+1}^{m^*})] = 1$ and the law of large numbers. A similar argument gives $\alpha_m^{(k+1)} = (1 - \tau_0)\alpha_m^* + o_p(1)$, and the stated result follows. \square

References

Alexandrovich, G. (2014), ‘‘A Note on the Article ‘Inference for Multivariate Normal Mixtures’ by J. Chen and X. Tan,’’ *Journal of Multivariate Analysis*, 129, 245–248.

- 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.
- Azaïs, J.-M., Gassiat, É., and Mercadier, C. (2009), “The Likelihood Ratio Test for General Mixture Models with or without Structural Parameter,” *ESAIM: Probability and Statistics*, 13, 301–327.
- Chen, H. and Chen, J. (2001), “The Likelihood Ratio Test for Homogeneity in Finite Mixture Models,” *Canadian Journal of Statistics*, 29, 201–215.
- (2003), “Tests for Homogeneity in Normal Mixtures in the Presence of a Structural Parameter,” *Statistica Sinica*, 13, 351–365.
- Chen, H., Chen, J., and Kalbfleisch, J. D. (2004), “Testing for a Finite Mixture Model with Two Components,” *Journal of the Royal Statistical Society, Series B*, 66, 95–115.
- Chen, J. (1995), “Optimal Rate of Convergence for Finite Mixture Models,” *Annals of Statistics*, 23, 221–233.
- Chen, J. and Li, P. (2009), “Hypothesis Test for Normal Mixture Models: The EM Approach,” *Annals of Statistics*, 37, 2523–2542.
- Chen, J., Li, P., and Fu, Y. (2012), “Inference on the Order of a Normal Mixture,” *Journal of the American Statistical Association*, 107, 1096–1105.
- Chen, J. and Tan, X. (2009), “Inference for Multivariate Normal Mixtures,” *Journal of Multivariate Analysis*, 100, 1367–1383.
- 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.
- Dacunha-Castelle, D. and Gassiat, E. (1999), “Testing the Order of a Model using Locally Conic Parametrization: Population Mixtures and Stationary ARMA Processes,” *Annals of Statistics*, 27, 1178–1209.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum Likelihood from Incomplete Data via EM Algorithm (with Discussion),” *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- Garel, B. (2001), “Likelihood Ratio Test for Univariate Gaussian Mixture,” *Journal of Statistical Planning and Inference*, 96, 325–350.
- (2005), “Asymptotic Theory of the Likelihood Ratio Test for the Identification of a Mixture,” *Journal of Statistical Planning and Inference*, 131, 271–296.
- Ghosh, J. K. and Sen, P. K. (1985), “On the Asymptotic Performance of the Log-likelihood Ratio Statistic for the Mixture Model and Related Results,” in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, eds. Le Cam, L. and Olshen, R., Belmont, CA: Wadsworth, vol. 2, pp. 789–806.

- Hartigan, J. (1985), “Failure of Log-likelihood Ratio Test,” in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, eds. Le Cam, L. and Olshen, R., Berkeley: University of California Press, vol. 2, pp. 807–810.
- Kasahara, H. and Shimotsu, K. (2015), “Testing the Number of Components in Normal Mixture Regression Models,” *Journal of the American Statistical Association*, 110, 1632–1645.
- Kiefer, J. and Wolfowitz, J. (1956), “Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters,” *Annals of Mathematical Statistics*, 27, 887–906.
- 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.
- Pollard, D. (1990), *Empirical Processes: Theory and Applications*, vol. 2 of *CBMS Conference Series in Probability and Statistics*, Hayward, CA: Institute of Mathematical Statistics.
- R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
- Rotnitzky, A., Cox, D. R., Bottai, M., and Robins, J. (2000), “Likelihood-based Inference with Singular Information Matrix,” *Bernoulli*, 6, 243–284.
- Titterton, 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.

Table 1: Type I errors (%) of the EM test of $H_0 : M = 1$

Level	$n = 200$			$n = 400$		
	$K = 1$	$K = 2$	$K = 3$	$K = 1$	$K = 2$	$K = 3$
Model 1: $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$						
10%	9.2	9.3	9.6	10.4	10.0	10.1
5%	5.2	5.0	4.9	4.3	4.4	4.4
1%	1.2	1.1	1.3	1.1	1.1	1.1
Model 2: $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$						
10%	9.4	9.3	9.4	10.0	9.9	9.7
5%	6.1	5.9	6.0	4.5	4.6	4.5
1%	1.5	1.5	1.4	1.2	1.2	1.3

Table 2: Powers (%) of the EM test of $H_0 : M = 1$

Level	$n = 200$			$n = 400$		
	$K = 1$	$K = 2$	$K = 3$	$K = 1$	$K = 2$	$K = 3$
Model 1: $\boldsymbol{\mu}_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$						
10%	38.9	38.7	38.0	67.9	67.5	67.3
5%	26.5	26.3	26.1	53.9	54.3	54.0
1%	10.6	10.8	10.8	30.9	30.6	30.5
Model 2: $\boldsymbol{\mu}_1 = \begin{pmatrix} -0.5 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$						
10%	94.7	94.8	94.8	100.0	100.0	100.0
5%	92.0	91.9	91.8	99.9	99.9	99.9
1 %	81.2	81.3	81.2	99.7	99.7	99.7

In both models, $\boldsymbol{\alpha}$ is set to $(\alpha_1, \alpha_2) = (0.7, 0.3)$.

Table 3: Type I errors (%) of the EM test of $H_0 : M = 2$

Level	$n = 200$			$n = 400$		
	$K = 1$	$K = 2$	$K = 3$	$K = 1$	$K = 2$	$K = 3$
	$\alpha = \begin{pmatrix} 0.7 \\ 0.3 \end{pmatrix}, \mu_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \mu_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$					
10%	7.5	7.6	7.7	5.7	5.6	5.9
5%	4.0	4.3	4.2	3.3	3.5	3.4
1%	0.9	1.0	1.0	0.3	0.3	0.3

Table 4: Parameter specifications for testing the power of the EM test of $H_0 : M = 2$

	α	μ_1	μ_2	μ_3	Σ_1	Σ_2	Σ_3
Model 1	$\begin{pmatrix} 0.15 \\ 0.35 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} -2 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 2 \end{pmatrix}$	$\begin{pmatrix} 2 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
Model 2	$\begin{pmatrix} 0.15 \\ 0.35 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} -2 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -1 \\ 4 \end{pmatrix}$	$\begin{pmatrix} 2 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

Table 5: Powers (%) of the EM test of $H_0 : M = 2$

Level	$n = 200$			$n = 400$		
	$K = 1$	$K = 2$	$K = 3$	$K = 1$	$K = 2$	$K = 3$
	Model 1					
10%	26.4	26.3	26.5	30.1	30.3	30.5
5%	17.7	16.7	16.8	22.6	22.6	22.7
1%	5.7	5.6	5.6	8.7	8.5	8.5
	Model 2					
10%	44.5	44.6	44.5	70.5	70.6	70.7
5%	36.1	36.1	36.0	65.4	65.1	65.1
1 %	17.0	16.9	17.1	48.0	47.7	47.2