Testing the Number of Components in Finite Mixture Normal Regression Model with Panel Data

Yu Hao*
Faculty of Business and Economics
The University of Hong Kong
haoyu@hku.hk

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

May 15, 2023

Abstract

This paper introduces a likelihood ratio-based test for examining the null hypothesis of an M_0-component model versus an alternative $(M_0 + 1)$-component model within the context of normal mixture panel regression. Contrary to the cross-sectional normal mixture, we demonstrate that the first-order derivative of the density function for the variance parameter in the panel normal mixture is linearly independent from its second-order derivative for the mean parameter. However, similar to the cross-sectional normal mixture, the likelihood ratio test statistic for the panel normal mixture remains unbounded. To manage this unboundedness, we employ a penalized maximum likelihood estimator and derive the asymptotic distribution of penalized likelihood ratio test and Expectation-Maximization test statistics using a fourth-order Taylor expansion of the log-likelihood function for reparameterized parameters. A sequential hypothesis testing approach is developed for consistently estimating the number of components. Simulation experiments reveal good finite sample performance of the proposed tests. We apply these tests to estimate the number of production technology types for the finite mixture Cobb-Douglas production function model.

*Address for correspondence: Yu(Jasmine) Hao, Faculty of Business and Economics, The University of Hong Kong. We are very grateful for the comments from Chun Pang Chow, Vadim Marmer, and Kevin Song. We also thank to the IAAE grant at the 2019 IAAE Conference. This research is support by the Natural Science and Engineering Research Council of Canada.

1 Introduction

Finite mixture models offer a natural representation of heterogeneity across a finite number of classes. Due to their flexibility, they have been employed in empirical applications across various fields since the proposal of a two-component normal mixture model by Pearson (1894). In economics, finite mixtures are frequently used to model unobserved individual-specific effects in labor economics, health economics, and industrial organization, among others.[1] Theoretical properties and examples of applications have been discussed by several authors, such as Lindsay (1995), Titterington et al. (1985), and McLachlan and Peel (2004).

The number of components is a crucial parameter in finite mixture models. In economic applications, the number of components often represents the number of unobservable types or abilities. Choosing an arbitrary number of parameters may lead to overestimation or underestimation of the level of heterogeneity. Using too few components may result in biased estimation due to overlooked unobserved heterogeneity, while employing too many components can be computationally costly and ill-behaved because of potential identification problems. Thus, developing a statistical procedure to determine the number of components is essential.

Testing for the number of components in normal mixture regression models has been a long-standing unsolved problem. The regularity conditions of the likelihood ratio test (LRT) for standard asymptotic analysis fail in finite mixture models due to issues such as non-identifiable parameters, the singularity of the Fisher Information matrix, and the true parameter being on the boundary of the parameter space. Numerous papers have been written on the subject of LRT 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; Chen et al., 2014), and the asymptotic distribution of the LRT statistic for general finite mixture models has been derived as a function of the Gaussian process (Dacunha-Castelle and Gassiat, 1999; Azaïs et al., 2009; Liu and Shao, 2003; Zhu and Zhang, 2004). However, the key assumptions in these works are violated in cross-sectional normal regression models because normal mixtures possess additional undesirable mathematical properties: (i) the Fisher information for testing is not finite, (ii) the log-likelihood function is unbounded, and (iii) the second derivative of the density function for the mean parameter is linearly dependent on its first derivative for the variance parameter. The asymptotic distribution of the LRT statistics of a cross-sectional univariate finite mixture normal regression model is analyzed by Kasahara and Shimotsu (2015), while its multivariate extension

[1] For example, Heckman and Singer (1984) use the finite mixture model to provide an alternative method to account for the unobserved heterogeneity in the analysis of single-spell duration times of unemployed workers. Keane and Wolpin (1997) and Cameron and Heckman (1998) analyze a dynamic model of schooling and occupational choices with unobserved heterogeneous human capital. Likewise, finite mixture models have been applied in health economics. Deb and Trivedi (1997) develop a finite mixture negative binomial count model that accounts for unobserved dispersion of elderly medical care utilization. In industrial organizations, modelling consumer segmentation in marketing such as Kamakura and Russell (1989) and Andrews and Currim (2003) is a venue of application.

is developed by Kasahara and Shimotsu (2019). Amengual et al. (2022) develops a score-type test for a cross-sectional normal mixture model.

This paper develops a likelihood-ratio-based test for determining the number of components in finite mixture normal regression models with panel data, where outcome variables are conditionally independent across periods given latent type within each unit. To the best of our knowledge, it is not known in the existing literature whether the aforementioned problems (i)-(iii) of the cross-sectional normal mixture still arise in the panel normal mixture or not. Furthermore, no likelihood-based test has yet been developed for testing the null hypothesis of an M_0-component model against an alternative (M_0+1)-component model for $M_0 \geq 1$ in the panel normal regression mixture models with conditional independent errors.[2]

We show that problems related to (i) and (ii) arise, but the higher-order degeneracy of problem (iii) disappears in the panel normal mixture models with conditional independence. Following Chen and Li (2009) and Kasahara and Shimotsu (2015), we consider a penalized likelihood ratio test (PLRT) and an Expectation-Maximization (EM) test to deal with the unboundedness and analyze the asymptotic distribution of the PLRT using a reparameterization orthogonal to the direction in which the Fisher information matrix is singular. The likelihood ratio of an $(M_0 + 1)$-component model against the M_0-component model is approximated with local quadratic-form expansion with squares and cross-products of the reparameterized parameters. We demonstrate that the asymptotic null distributions of the penalized likelihood ratio test statistic (PLRTS) and the EM test statistic are characterized by the maximum of M_0 random variables, which we can easily simulate. Building on the PLRT and EM tests, we propose a sequential hypothesis testing approach for consistently estimating the number of components. In simulations, our proposed PLRT and EM tests demonstrate favorable finite sample properties. Moreover, a sequential hypothesis testing approach accurately selects the correct number of components with high frequency, surpassing selection procedures based on the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

This paper makes several contributions. First, it analyzes the likelihood-ratio-based test for the number of components in the panel normal regression mixture models with conditional independence. Kasahara and Shimotsu (2015) and Kasahara and Shimotsu (2019) analyze the likelihood-ratio-based tests for the number of components in the cross-sectional univariate normal mixture regression models and the multivariate normal mixture models, respectively. We demonstrate that the asymptotic distribution of PLRT and EM test for the panel normal regression mixture models with conditionally independent errors differs from those of the univariate/multivariate normal mixture models in the aforementioned two papers because the higher-order dependency does not occur when the repeated measurement of outcome variables is available in panel data. Further-

[2]Kasahara and Shimotsu (2014) develops a procedure to estimate a lower bound on the number of components consistently in finite mixture models in which each component distribution has independent marginals, which includes the panel normal regression mixture models with conditionally independent errors as a special case.

more, we develop a sequential hypothesis testing approach for consistently estimating the number of components.

Second, while it is well known that the log-likelihood function of normal mixture models is unbounded (Hartigan, 1985), it is unknown if the related unboundedness problem arises in the panel data. We show that the likelihood ratio test statistic is unbounded in the panel normal mixture models with conditionally independent errors when the time dimension of panel data is finite. This unboundedness causes over-rejection of the likelihood ratio test. We introduce a penalty function to prevent the likelihood ratio test statistics from being unbounded, where we use computational experiments to determine the data-driven penalty function. We develop an R package NormalRegPanelMixture (Hao, 2017) that contains the EM test module and asymptotic distribution simulation module.

Third, our empirical analysis of the number of production technology types using panel data from Japanese and Chilean manufacturing firms provides strong evidence for substantial heterogeneity in production function coefficients across firms within narrowly defined industries. This is an important contribution to the literature on production function estimation, where most existing empirical applications assume the homogeneity of production function coefficients across firms using the standard production function estimation methods developed by Olley and Pakes (1996), Levinsohn and Petrin (2003), and Ackerberg et al. (2015). Our empirical finding suggests that it is essential to incorporate unobserved heterogeneity in the production function coefficients across firms in applications (Li and Sasaki, 2017; Doraszelski and Jaumandreu, 2018; Balat et al., 2019; Kasahara et al., 2022).

The EM test approach was introduced by Li et al. (2009) and Chen and Li (2009) to test homogeneity in finite mixture models. Li and Chen (2010) developed an EM test for the null hypothesis of M_0 components applicable to general $M_0 \geq 2$, while Kasahara and Shimotsu (2015) proposed an EM test for normal regression mixture models to test the null of $M_0 \geq 2$. The EM approach has also been applied to test homogeneity in multivariate mixtures (Niu et al., 2011) and subgroup analyses (Shen and He, 2015). More recently, Liu et al. (2018) extended the EM test to mixtures of the general location-scale family distribution, and Kasahara and Shimotsu (2019) developed an EM test for multivariate normal mixture models. Building upon the prior literature, this paper develops an EM test for panel normal regression mixture models with conditionally independent errors.

Identifying and estimating latent group structure in panel data have received attention in recent literature (Kasahara and Shimotsu, 2009; Ando and Bai, 2016; Bonhomme and Manresa, 2015; Lin and Ng, 2012; Lu and Su, 2017; Su et al., 2016). Finite mixture modeling provides a practical, model-based approach to determining unobserved group structures. Choosing the number of groups is often a prerequisite for classifying each individual's group membership. We can estimate the number of groups in panel data regression models by applying our proposed sequential

hypothesis testing approach.

The rest of the paper is organized as follows. In Section 2, we define the finite normal mixture panel regression model. In Section 3, we demonstrate the PLRT for testing the homogeneity of normal mixture panel regression against a two-component model as a precursor to obtaining the general M_0 components test. Section 4 generalizes the result to testing M_0 components against $M_0 + 1$ components. Section 5 introduces the EM test for testing M_0 components against $M_0 + 1$ components. Section 6 derives the asymptotic distribution of the PLRT and EM tests under local alternatives, while Section 7 develops a consistent estimator for the number of components based on sequential hypothesis testing. Section 8 presents the simulated results of the tests. Section 9 provides an empirical application. Let $:=$ denote "equals by definition." Boldface letters denote vectors or matrices.

2 Heteroskedastic finite mixture panel normal regression model

We consider finite mixture normal regression models with panel data, where the panel length T is fixed and the number of cross-sectional observations n goes to infinity. Define $\boldsymbol{w} := \{y_t, \boldsymbol{x}_t, \boldsymbol{z}_t\}_{t=1}^T$ with $y_t \in \mathbb{R}, \boldsymbol{x}_t \in \mathbb{R}^q, \boldsymbol{z}_t \in \mathbb{R}^p$. Given $M \geq 2$, denote the density of a $M-$component model that represents the conditional density function of $\{y_t\}_{t=1}^T$ given $\{\boldsymbol{x}_t, \boldsymbol{z}_t\}_{t=1}^T$ as

$$f_M(\boldsymbol{w}; \boldsymbol{\vartheta}_M) = \sum_{j=1}^{M} \alpha_j f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\theta}_j), \tag{1}$$

where $\boldsymbol{\vartheta}_M = (\boldsymbol{\alpha}^\top, \boldsymbol{\theta}_1^\top, ..., \boldsymbol{\theta}_M^\top, \gamma^\top)^\top \in \Theta_{\boldsymbol{\vartheta}_M}$, $\boldsymbol{\alpha}^\top := (\alpha_1, ..., \alpha_{M-1})$, $\alpha_M = 1 - \sum_{j=1}^{M-1} \alpha_j$, and

$$f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\theta}_j) = \prod_{t=1}^{T} \frac{1}{\sigma_j} \phi\left(\frac{y_t - \mu_j - \boldsymbol{x}_t^\top \boldsymbol{\beta}_j - \boldsymbol{z}_t^\top \boldsymbol{\gamma}}{\sigma_j}\right) \tag{2}$$

is the j-th component density function with $\mu_j \in \Theta_\mu \subset \mathbb{R}$, $\sigma_j^2 \in \Theta_\sigma \subset \mathbb{R}_{++}$, $\boldsymbol{\beta}_j \in \Theta_{\boldsymbol{\beta}} \subset \mathbb{R}^q$, $\boldsymbol{\gamma} \in \Theta_{\boldsymbol{\gamma}} \subset \mathbb{R}^p$, and $\phi(t) = (2\pi)^{-1/2} \exp(-\frac{t^2}{2})$ is the standard normal probability density function. We collect the component-specific parameters into $\boldsymbol{\theta}_j := (\mu_j, \sigma_j^2, \boldsymbol{\beta}_j^\top)^\top \in \Theta_{\boldsymbol{\theta}}$ while the regression coefficient $\boldsymbol{\gamma}$ for a vector $\boldsymbol{z}$ is assumed to be common across components.

The number of components, denoted by M_0, is defined as the smallest integer M such that the data density of $\boldsymbol{w}$ admits the representation (1). Consider a random sample of n with panel length of T independent observations $\{\boldsymbol{W}_i\}_{i=1}^n$ where $\boldsymbol{W}_i = \{(Y_{it}, \boldsymbol{X}_{it}^\top, \boldsymbol{Z}_{it}^\top)^\top\}_{t=1}^T$ from a true M_0-component density $f_M(\boldsymbol{w}; \boldsymbol{\vartheta}_{M_0}^*)$ defined in equation (1) with $\boldsymbol{\vartheta}_{M_0}^* = ((\boldsymbol{\alpha}^*)^\top, (\boldsymbol{\theta}_1^*)^\top, ..., (\boldsymbol{\theta}_{M_0}^*)^\top, (\gamma^*)^\top)^\top$. The superscript $*$ signifies the true parameter value. Because component distributions can be identified only up to permutation, we assume that $\mu_1^* < \mu_2^* <$

$\cdots < \mu^*_{M_0}$ for identification.[3]

Our goal is to test

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

3 Likelihood ratio test for $H_0 : M = 1$ against $H_A : M = 2$

We begin by developing the PLRT to test the null hypothesis $H_0 : M = 1$ against the alternative hypothesis $H_1 : M = 2$. Consider a random sample of n with a panel length of T independent observations $\{\boldsymbol{W}_i\}_{i=1}^n$, where $\boldsymbol{W}_i = \{(Y_{it}, \boldsymbol{X}_{it}^\top, \boldsymbol{Z}_{it}^\top)^\top\}_{t=1}^T$, drawn from a true one-component density $f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\theta})$ defined in equation (2). Now consider a two-component mixture density function

$$f_2(\boldsymbol{w}; \boldsymbol{\vartheta}_2) = \alpha f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\theta}_1) + (1 - \alpha) f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\theta}_2),$$

where $\boldsymbol{\vartheta}_2 = (\alpha, \boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \boldsymbol{\gamma}^\top)^\top \in \Theta_{\boldsymbol{\vartheta}_2}$, and α is the mixing probability of the first component. The two-component model can generate the true one-component density in two cases: (1) $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}^*$; (2) $\alpha = 0$ or 1. Consequently, the null hypothesis $H_0 : M = 1$ can be partitioned into two sub-hypotheses: $H_{01} : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ and $H_{02} : \alpha(1 - \alpha) = 0$. The regularity conditions of the LRTS for a standard asymptotic analysis fail in finite mixture models: under H_{01}, α is not identified, and the Fisher information matrix for the other parameters becomes singular; under H_{02}, α is on the boundary of the parameter space, and either $\boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_2$ is not identified.

As discussed in the introduction, analyzing the asymptotic distribution of the LRTS for the cross-sectional normal mixture is challenging due to its undesirable mathematical properties (cf., Chen and Li, 2009): (i) the Fisher information for testing H_{02} is not finite, (ii) the log-likelihood function is unbounded (Hartigan, 1985), and (iii) the first-order derivative of $f_2(\boldsymbol{w}; \boldsymbol{\vartheta}_2)$ with respect to σ_j^2 is linearly dependent on its second-order derivative with respect to μ_j. The presence of problems (i)-(iii) in panel normal mixture models with $T \geq 2$ is not well-understood in the existing literature because, to the best of our knowledge, no existing studies have examined them.

Regarding problem (i), we note that the issue of infinite Fisher information for testing H_{02} also arises in the panel normal mixture model. For brevity, let us consider the case without $(\boldsymbol{X}, \boldsymbol{Z})$. The score for testing $H_{02} : \alpha = 0$ takes the form

$$\left.\frac{\partial f_2(\boldsymbol{W}; \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)}{\partial \alpha}\right|_{\alpha=0, \mu_2=\mu^*, \sigma_2^2=\sigma^{*2}} = \frac{f(\boldsymbol{W}; \mu_1, \sigma_1^2)}{f(\boldsymbol{W}; \mu^*, \sigma^{*2})} - 1,$$

where $f(\boldsymbol{W}; \mu, \sigma^2) = \prod_{t=1}^T \phi((y_t - \mu)/\sigma)/\sigma$, and $\phi(\cdot)$ is the standard normal density function.

[3] More generally, we may consider a lexicographical order: $\boldsymbol{\theta}_1^* < \boldsymbol{\theta}_2^* < \cdots < \boldsymbol{\theta}_{M_0}^*$.

When $\sigma_1^2 > 2\sigma^{*2}$, $\mathbb{E}[\{f(\boldsymbol{W};\mu_1,\sigma_1^2)/f(\boldsymbol{W};\mu^*,\sigma^{*2})-1\}^2] = \infty$. For more details, please refer to Proposition 5. Because the infinite Fisher information causes difficulty in deriving the asymptotic distribution under H_{02}, this paper focuses on testing H_{01}. We define $\Upsilon_1^* := \{(\alpha,\boldsymbol{\gamma},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \in \Theta_{\vartheta_2} : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}^* \text{ and } \boldsymbol{\gamma} = \boldsymbol{\gamma}^*\}$, which is the subspace of Θ_{ϑ_2} that corresponds to H_{01}. Note that because of our focus on H_{01}, our test may not have power against the local alternatives with $\alpha_n \to 0$. We analyze the asymptotic distribution of the PLRTS under the contiguous local alternatives in Section 6.

Related to problem (ii), the LRTS in normal mixture models with panel data becomes unbounded as the sample size n goes to ∞. Define the likelihood ratio statistic with respect to the true parameter under H_0 as:

$$LR_n^*(\boldsymbol{\vartheta}_2) := 2\left\{\sum_{i=1}^n \log f_2(\boldsymbol{W}_i;\boldsymbol{\vartheta}_2) - \sum_{i=1}^n \log f(\boldsymbol{W}_i;\boldsymbol{\gamma}^*,\boldsymbol{\theta}^*)\right\},$$

where f_2 is the density of the two-component finite mixture distribution in (1) with $M = 2$, and $((\boldsymbol{\gamma}^*)^\top,(\boldsymbol{\theta}^*)^\top)^\top$ is the true parameter value under H_0. Let $\tilde{\boldsymbol{\vartheta}}_{2,n}$ be the maximum likelihood estimator for the two-component model, i.e., $\tilde{\boldsymbol{\vartheta}}_{2,n} = \arg\max_{\boldsymbol{\vartheta}_2\in\Theta_{\boldsymbol{\vartheta}_2}} LR_n^*(\boldsymbol{\vartheta}_2)$.

Proposition 1. *Suppose the true model is described by the one-component model with* $(\boldsymbol{\gamma},\boldsymbol{\theta}) = (\boldsymbol{\gamma}^*,\boldsymbol{\theta}^*)$. *Then, for any positive constant* $M > 0$, $\Pr\left(LR_n^*(\tilde{\boldsymbol{\vartheta}}_{2,n}) \le M\right) \to 0$ *as* $n \to \infty$.

To deal with unboundedness, we consider a maximum penalized likelihood estimator (PMLE) as in Chen and Tan (2009) using the following penalty function:

$$\tilde{p}_n(\boldsymbol{\vartheta}_M) := \sum_{j=1}^M p_n(\sigma_j^2) \quad \text{with} \quad p_n(\sigma_j^2) := -a_n\{\sigma_0^2/\sigma_j^2 + \log(\sigma_j^2/\sigma_0^2) - 1\}. \tag{3}$$

This penalty function circumvents the problem of unbounded log likelihood by preventing a variance parameter estimate from nearing zero. The parameter a_n is selected such that the penalty's impact becomes asymptotically negligible for the distribution of the PMLE. Refer to the conditions C1-C3 in the proof of Proposition 6.

Let

$$\hat{\boldsymbol{\vartheta}}_2 = \arg\max_{\boldsymbol{\vartheta}_2\in\Theta_{\boldsymbol{\vartheta}_2}} \sum_{i=1}^n \log f_2(\boldsymbol{W}_i;\boldsymbol{\vartheta}_2) + \tilde{p}_n(\boldsymbol{\vartheta}_2)$$

denote the Penalized Maximum Likelihood Estimator (PMLE, hereafter) under the two-component model. Define a set of parameter values for the two-component density that generates the true one-component density by $\Theta_2^* := \{(\alpha,\boldsymbol{\gamma},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \in \Theta_{\vartheta_2} : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}^* \text{ and } \boldsymbol{\gamma} = \boldsymbol{\gamma}^*; \alpha = 1 \text{ and } \theta_1 = \theta^*; \alpha = 0 \text{ and } \theta_2 = \theta^*\}$. $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are component-specific parameters, while $\boldsymbol{\gamma}$ is a parameter vector common across components. The following proposition establishes the consistency

of the PMLE.

Assumption 1. *(a) $\boldsymbol{X}$ and $\boldsymbol{Z}$ have finite second moments, and $Pr(\boldsymbol{X}^\top\boldsymbol{\beta}+\boldsymbol{Z}^\top\boldsymbol{\gamma}\neq\boldsymbol{X}^\top\boldsymbol{\beta}^*+\boldsymbol{Z}^\top\boldsymbol{\gamma}^*)>0$ for $(\boldsymbol{\beta}^\top,\boldsymbol{\gamma}^\top)^\top\neq((\boldsymbol{\beta}^*)^\top,(\boldsymbol{\gamma}^*)^\top)^\top$. (b) $a_n>0$ and $a_n=o(n^{1/4})$ in the penalty function (3).*

Proposition 2. *Suppose that Assumption 1 holds. Then under the null hypothesis $H_0: M_0=1$, $\inf_{\boldsymbol{\vartheta}_2\in\Theta_2^*}|\hat{\boldsymbol{\vartheta}}_2-\boldsymbol{\vartheta}_2|\to_p 0$.*

It should be noted that $f_2(\boldsymbol{w};\boldsymbol{\vartheta}_2^*)=f(\boldsymbol{w};\boldsymbol{\gamma}^*,\boldsymbol{\theta}^*)$ for any $\boldsymbol{\vartheta}_2^*\in\Theta_2^*$. Consequently, Proposition 2 suggests that the PMLE $\hat{\boldsymbol{\vartheta}}_2$ converges in probability to a set of parameters for which the true density function $f(\boldsymbol{w};\boldsymbol{\gamma}^*,\boldsymbol{\theta}^*)$ emerges within the space of two-component density functions.

For problem (iii), we show that in normal mixture models with panel data, the first-order derivative of $f_2(\boldsymbol{w};\boldsymbol{\vartheta}_2)$ with respect to σ_j^2 is *not* linearly dependent with its second-order derivative with respect to μ_j (See Proposition 3(c)). Consequently, the panel mixture model (1) with the component density function (2) is strongly identifiable, and the best rate of convergence for estimating the mixing distribution is $n^{-1/4}$ when the number of components is unknown (c.f., Chen, 1995). See Proposition 4(a). In contrast, the strong identifiability does not hold for the cross-sectional normal mixture, and its convergence rate becomes as slow as $n^{-1/8}$ when the number of components is over-specified (c.f., Kasahara and Shimotsu, 2015).

As in any finite mixture models, however, the standard asymptotic analysis breaks down in testing $H_{01}: \boldsymbol{\theta}_1=\boldsymbol{\theta}_2=\boldsymbol{\theta}^*$ because α is not identified under H_{01}; in addition, the first-order derivative at the true value $\boldsymbol{\vartheta}_2^*=(\alpha,(\boldsymbol{\theta}^*)^\top,(\boldsymbol{\theta}^*)^\top,(\boldsymbol{\gamma}^*)^\top)^\top$ are linear dependent as

$$\nabla_{\boldsymbol{\theta}_1}\log f_2(\boldsymbol{w};\boldsymbol{\vartheta}_2^*)=\frac{\alpha}{1-\alpha}\nabla_{\boldsymbol{\theta}_2}\log f_2(\boldsymbol{w};\boldsymbol{\vartheta}_2^*). \tag{4}$$

To deal with this linear dependency, we analyze the asymptotic distribution of LRTS by developing a higher-order approximation for the log-likelihood function.

To extract the direction of Fisher Information matrix singularity, we adapt the reparameterization approach by Kasahara and Shimotsu (2012) and consider the following one-to-one reparameterization of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ given α:

$$\begin{pmatrix}\boldsymbol{\lambda}\\ \boldsymbol{\nu}\end{pmatrix}:=\begin{pmatrix}\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\\ \alpha\boldsymbol{\theta}_1+(1-\alpha)\boldsymbol{\theta}_2\end{pmatrix}\text{ so that }\begin{pmatrix}\boldsymbol{\theta}_1\\ \boldsymbol{\theta}_2\end{pmatrix}=\begin{pmatrix}\boldsymbol{\nu}+(1-\alpha)\boldsymbol{\lambda}\\ \boldsymbol{\nu}-\alpha\boldsymbol{\lambda}\end{pmatrix}, \tag{5}$$

where $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$ are both $(q+2)\times 1$ reparameterized parameter vectors with $\boldsymbol{\nu}=(\nu_\mu,\nu_\sigma,\boldsymbol{\nu}_{\boldsymbol{\beta}}^\top)^\top$ and $\boldsymbol{\lambda}=(\lambda_\mu,\lambda_\sigma,(\boldsymbol{\lambda}_\beta)^\top)^\top=(\mu_1-\mu_2,\sigma_1^2-\sigma_2^2,(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)^\top)^\top$. We also write $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ as $\boldsymbol{\theta}=(\theta_1,\theta_2,\theta_3,...,\theta_{q+2})^\top:=(\mu,\sigma^2,\beta_1,...\beta_q)^\top$ and $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\lambda_3,...,\lambda_{q+2})^\top:=(\lambda_\mu,\lambda_\sigma,\lambda_{\beta_1},...,\lambda_{\beta_q})^\top$.

This reparameterization is essential for analyzing the asymptotic distribution of the PLRTS in light of the linear dependency in (4). The reparameterized parameter $\boldsymbol{\lambda}$ captures a deviation from the one component model, where its first-order derivatives of the log density are identically equal

to zero under $H_0 : M = 1$. Consequently, this reparameterization facilitates the derivation of an approximate quadratic-form criterion function, which is based on the fourth-order Taylor series approximation of the log-likelihood function, in order to characterize the asymptotic distribution of the LRTS.

Define the space for reparameterized parameters as

$$\boldsymbol{\psi} := (\boldsymbol{\gamma}^\top, \boldsymbol{\nu}^\top, \boldsymbol{\lambda}^\top)^\top \in \Theta_{\boldsymbol{\psi}},$$

where $\Theta_{\boldsymbol{\psi}} = \{\boldsymbol{\psi} : \boldsymbol{\gamma} \in \Theta_{\boldsymbol{\gamma}}, \boldsymbol{\nu} + (1-\alpha)\boldsymbol{\lambda} \in \Theta_{\boldsymbol{\theta}}, \boldsymbol{\nu} - \alpha\boldsymbol{\lambda} \in \Theta_{\boldsymbol{\theta}}\}$. Under the null hypothesis $H_{01} : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}^*$, we have $\boldsymbol{\lambda} = (0, \ldots, 0)^\top$ and $\boldsymbol{\nu} = \boldsymbol{\theta}^*$. We rewrite the reparameterized parameters under null hypothesis to be $(\boldsymbol{\psi}^*)^\top = ((\boldsymbol{\gamma}^*)^\top, (\boldsymbol{\theta}^*)^\top, 0, \ldots, 0)^\top$. Under the reparameterized parameter space, the density function and its logarithm are expressed as

$$\begin{aligned} &g(\boldsymbol{w}; \boldsymbol{\psi}, \alpha) = \alpha f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu} + (1-\alpha)\boldsymbol{\lambda}) + (1-\alpha) f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu} - \alpha\boldsymbol{\lambda}) \quad \text{and} \\ &l(\boldsymbol{w}; \boldsymbol{\psi}, \alpha) = \log g(\boldsymbol{w}; \boldsymbol{\psi}, \alpha). \end{aligned} \tag{6}$$

Write $\boldsymbol{\psi}$ as $\boldsymbol{\psi} = (\boldsymbol{\eta}^\top, \boldsymbol{\lambda}^\top)^\top$ with $\boldsymbol{\eta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\nu}^\top)^\top$, where $\boldsymbol{\eta}^* = ((\boldsymbol{\gamma}^*)^\top, (\boldsymbol{\nu}^*)^\top)^\top$ and $\boldsymbol{\lambda}^* = \mathbf{0}$. Denote the parameter space of $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ by $\Theta_{\boldsymbol{\eta}} \subset \mathbb{R}^{p+q+2}$ and $\Theta_{\boldsymbol{\lambda}} \subset \mathbb{R}^{q+2}$, respectively.

Under this reparameterization, the first-order derivatives of the reparameterized log density with respect to the reparameterized parameters $\boldsymbol{\eta}$ is identical to those under the one-component model, and the first-order derivative with respect to $\boldsymbol{\lambda}$ is a zero vector:

$$\nabla_{\boldsymbol{\eta}^\top} l(\boldsymbol{w}; \boldsymbol{\psi}^*, \alpha) = \frac{\nabla_{(\boldsymbol{\gamma}^\top, \boldsymbol{\theta}^\top)^\top} f(\boldsymbol{w}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*)}{f(\boldsymbol{w}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*)} \quad \text{and} \quad \nabla_{\boldsymbol{\lambda}} l(\boldsymbol{w}; \boldsymbol{\psi}^*, \alpha) = \mathbf{0}. \tag{7}$$

With $\nabla_{\boldsymbol{\lambda}} l(\boldsymbol{w}; \boldsymbol{\psi}^*, \alpha) = 0$, the Fisher information matrix is singular, and the standard quadratic approximation fails. Consequently, the information on $\boldsymbol{\lambda}$ is provided by the second-order derivative of $l(\boldsymbol{w}; \boldsymbol{\psi}, \alpha)$ with respect to $\boldsymbol{\lambda}$. We use second-order derivative with respect to $\boldsymbol{\lambda}$ to identify $\boldsymbol{\lambda}$:

$$\nabla_{\boldsymbol{\lambda}\boldsymbol{\lambda}^\top} l(\boldsymbol{w}; \boldsymbol{\psi}^*, \alpha) = \alpha(1-\alpha) \frac{\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^\top} f(\boldsymbol{w}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*)}{f(\boldsymbol{w}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*)}. \tag{8}$$

When α is bounded away from 0 and 1, the elements of $\nabla_{\boldsymbol{\lambda}\boldsymbol{\lambda}^\top} l(\boldsymbol{W}; \boldsymbol{\psi}^*, \alpha)$ are mean-zero random variables.

Note that, unlike the cross-sectional models analyzed by Kasahara and Shimotsu (2015), there exists no collinearity between these first and second-order derivatives for the panel models. This distinction is indeed important as it highlights the differences in the asymptotic distribution of the LRTS for the panel models compared to the cross-sectional models. The absence of collinearity between the first and second-order derivatives in panel models leads to different convergence

rates and asymptotic properties.

Let f^* and ∇f^* denote $f(\boldsymbol{W};\boldsymbol{\gamma}^*,\boldsymbol{\theta}^*)$ and $\nabla f(\boldsymbol{W};\boldsymbol{\gamma}^*,\boldsymbol{\theta}^*)$. Define the vector $\boldsymbol{s}(\boldsymbol{W})$ as

$$\boldsymbol{s}(\boldsymbol{W}) = \begin{pmatrix} \boldsymbol{s}_{\boldsymbol{\eta}}(\boldsymbol{W}) \\ \boldsymbol{s}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W}) \end{pmatrix}, \quad \text{where } \underset{(p+q+2)\times 1}{\boldsymbol{s}_{\boldsymbol{\eta}}(\boldsymbol{W})} := \frac{\nabla_{(\boldsymbol{\gamma}^\top,\boldsymbol{\theta}^\top)^\top} f^*}{f^*} \quad \text{and} \quad \underset{((q+2)(q+1)/2)\times 1}{\boldsymbol{s}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W})} := \frac{\widetilde{\nabla}_{\boldsymbol{\theta\theta}^\top} f^*}{f^*}. \tag{9}$$

The term $\widetilde{\nabla}_{\boldsymbol{\theta\theta}^\top} f^*$ denotes the second-order derivatives of the density function f^* with respect to the parameters $\boldsymbol{\theta}$. Coefficients c_{jk} are employed to adjust the scaling of these second-order derivatives. The function $\boldsymbol{s}(\boldsymbol{w})$ comprises the second-order derivatives of the log-likelihood function with respect to the reparameterized parameter $\boldsymbol{\lambda}$. This function, $\boldsymbol{s}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{w})$, serves as a score function for identifying $\boldsymbol{\lambda}$. Consequently, $\boldsymbol{s}(\boldsymbol{w})$ is referred to as a score function. An explicit expression for the score function $\boldsymbol{s}(\boldsymbol{w})$ can be derived using Hermite polynomials, as elaborated in Appendix B.2.

Collect the relevant normalized reparameterized parameters and define $\boldsymbol{t}(\boldsymbol{\psi},\alpha)$ as

$$\boldsymbol{t}(\boldsymbol{\psi},\alpha) = \begin{pmatrix} \boldsymbol{t}_{\boldsymbol{\eta}} \\ \boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda},\alpha) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\eta}-\boldsymbol{\eta}^* \\ \alpha(1-\alpha)\boldsymbol{v}(\boldsymbol{\lambda}) \end{pmatrix}, \tag{10}$$

where $v(\boldsymbol{\lambda})$ is a vector of unique elements of $\boldsymbol{\lambda\lambda}^\top$ given by

$$v(\boldsymbol{\lambda}) = (\lambda_1\lambda_1, ..., \lambda_{q+2}\lambda_{q+2}, \lambda_1\lambda_2, ..., \lambda_{q+1}\lambda_{q+2})^\top \tag{11}$$

of which length is $q_\lambda := (q+2)(q+3)/2$.

Let $L_n(\boldsymbol{\psi},\alpha) := \sum_{i=1}^n l(\boldsymbol{W}_i;\boldsymbol{\psi}^*,\alpha)$ be the reparameterized log-likelihood function and define the normalized score vector

$$\boldsymbol{S}_n := n^{-1/2}\sum_{i=1}^n \boldsymbol{s}(\boldsymbol{W}_i).$$

Then, taking the fourth order Taylor expansion of $L_n(\boldsymbol{\psi},\alpha)$ around $(\boldsymbol{\psi}^*,\alpha)$, we may write $2\{L_n(\boldsymbol{\psi},\alpha) - L_n(\boldsymbol{\psi}^*,\alpha)\}$ as a quadratic function of $\sqrt{n}\boldsymbol{t}(\boldsymbol{\psi},\alpha)$ as

$$2\{L_n(\boldsymbol{\psi},\alpha) - L_n(\boldsymbol{\psi}^*,\alpha)\} = 2(\sqrt{n}\boldsymbol{t}(\boldsymbol{\psi},\alpha))^\top \boldsymbol{S}_n - (\sqrt{n}\boldsymbol{t}(\boldsymbol{\psi},\alpha))^\top \boldsymbol{\mathcal{I}}_n(\sqrt{n}\boldsymbol{t}(\boldsymbol{\psi},\alpha)) + R_n(\boldsymbol{\psi},\alpha) \tag{12}$$

$$= \boldsymbol{G}_n^\top \boldsymbol{\mathcal{I}}_n \boldsymbol{G}_n - \left[\sqrt{n}\boldsymbol{t}(\boldsymbol{\psi},\alpha) - \boldsymbol{G}_n\right]^\top \boldsymbol{\mathcal{I}}_n \left[\sqrt{n}\boldsymbol{t}(\boldsymbol{\psi},\alpha) - \boldsymbol{G}_n\right] + R_n(\boldsymbol{\psi},\alpha), \tag{13}$$

where $\boldsymbol{\mathcal{I}}_n$ is the negative of the sample Hessian defined in the proof of Proposition 3 while $\boldsymbol{G}_n := \boldsymbol{\mathcal{I}}_n^{-1}\boldsymbol{S}_n$. Let $\boldsymbol{\mathcal{I}} = \mathbb{E}[\boldsymbol{s}(\boldsymbol{W})\boldsymbol{s}(\boldsymbol{W})^\top]$.

Assumption 2. *(a) $\boldsymbol{X}$ and $\boldsymbol{Z}$ have finite 8-th moments. (b) $\mathbb{E}[\boldsymbol{U}\boldsymbol{U}^\top]$ is non-singular, where $\boldsymbol{U} = [1, \boldsymbol{X}^\top, \boldsymbol{Z}^\top]^\top$.*

Proposition 3. *Suppose that assumption 1 and 2 hold. Then, under $H_0 : M = 1$, for $\alpha \in (0,1)$, (a) for any $\delta > 0$, $\limsup_{n\to\infty} \Pr(\sup_{\boldsymbol{\psi}\in\Theta_{\boldsymbol{\psi}}:||\boldsymbol{\psi}-\boldsymbol{\psi}^*||\leq\kappa} |R_n(\boldsymbol{\psi},\alpha)| > \delta(1+||n\boldsymbol{t}(\boldsymbol{\psi},\alpha)||^2)) \to 0$ as $\kappa \to 0$, (b) $\boldsymbol{S}_n \xrightarrow{d} \boldsymbol{S} \sim N(0,\boldsymbol{\mathcal{I}})$, and (c) $\boldsymbol{\mathcal{I}}_n \xrightarrow{p} \boldsymbol{\mathcal{I}}$, where $\boldsymbol{\mathcal{I}}$ is finite and non-singular.*

The non-singularity of $\boldsymbol{\mathcal{I}}$ in Proposition 3(c) highlights the difference between the cross-sectional normal mixture and the panel data normal mixture models. In particular, as shown in equation (75) in Appendix B.2, the first-order derivative of $f_2(\boldsymbol{w};\boldsymbol{\vartheta}_2)$ with respect to σ_j^2 is linearly independent of its second-order derivative with respect to μ_j when $T \geq 2$, ensuring that the high-order degeneracy of problem (iii) does not arise. Intuitively, the availability of repeated observations within each individual unit provides better identification, even for over-parameterized models, and reduces the degree of higher-order degeneracy.

The set of feasible values of $\sqrt{n}\boldsymbol{t}(\boldsymbol{\psi},\alpha)$ is given by the shifted and re-scaled parameter space for $(\boldsymbol{\eta}, v(\boldsymbol{\lambda}))$ defined as $\Lambda_n := \sqrt{n}(\Theta_{\boldsymbol{\eta}} - \eta^*) \times \sqrt{n}\alpha(1-\alpha)v(\Theta_{\boldsymbol{\lambda}})$, where $v(A) := \{t \in \mathbb{R}^{q_\lambda} : t = v(\lambda) \text{ for some } \lambda \in A \subset \mathbb{R}^{q+2}\}$. Because $\Lambda_n/\sqrt{n}$ is locally approximated by a cone $\Lambda := \mathbb{R}^{p+q+2} \times v(\mathbb{R}^{q+2})$, we may apply Lemma 2 of Andrews (1999) to approximate the distribution of the supremum of the right-hand side of (13) as

$$\max_{\boldsymbol{\psi}\in\Theta_{\boldsymbol{\psi}}} 2\{L_n(\boldsymbol{\psi},\alpha) - L_n(\boldsymbol{\psi}^*,\alpha)\} \xrightarrow{d} \boldsymbol{G}^\top \boldsymbol{\mathcal{I}} \boldsymbol{G} - \inf_{\boldsymbol{t}\in\Lambda} (\boldsymbol{t}-\boldsymbol{G})' \boldsymbol{\mathcal{I}} (\boldsymbol{t}-\boldsymbol{G}),$$

where $\boldsymbol{G} = \boldsymbol{\mathcal{I}}^{-1}\boldsymbol{S} \sim N(0, \boldsymbol{\mathcal{I}}^{-1})$. This allows us to characterize the asymptotic distribution of the LRTS.

For each $\alpha \in (0,1)$, define the reparameterized PMLE by

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}\in\Theta_{\boldsymbol{\psi}}} L_n(\boldsymbol{\psi},\alpha) + \sum_{j=1}^{2} p_n(\sigma_j^2(\boldsymbol{\psi},\alpha)) \tag{14}$$

with $\hat{\boldsymbol{\psi}} := (\hat{\boldsymbol{\gamma}}^\top, \hat{\boldsymbol{\nu}}^\top, \hat{\boldsymbol{\lambda}}^\top)^\top$, where $\Theta_{\boldsymbol{\psi}}$ is defined as the space of $\boldsymbol{\psi}$ so that the $\boldsymbol{\vartheta}_2$ implied is in $\Theta_{\boldsymbol{\vartheta}}$ and $\sigma_j^2(\boldsymbol{\psi},\alpha)$ is the value of σ_j implied by the value of $\boldsymbol{\psi}$ and α (e.g., $\sigma_1^2(\boldsymbol{\psi},\alpha) = \nu_\sigma + (1-\alpha)\lambda_\sigma$).

Let $(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\theta}}_0)$ be the one-component MLE that maximizes the one-component likelihood function $L_{0,n}(\boldsymbol{\gamma},\boldsymbol{\theta}) := \sum_{i=1}^n \log f(\boldsymbol{W}_i;\boldsymbol{\gamma},\boldsymbol{\theta})$. Define the LRTS and the PLRTS of testing H_{01}, respectively, with a small positivity constant ϵ on α as

$$LR_n := \max_{\alpha\in[\epsilon,1-\epsilon]} 2\{L_n(\hat{\boldsymbol{\psi}},\alpha) - L_{0,n}(\hat{\boldsymbol{\gamma}}_0,\hat{\boldsymbol{\theta}}_0)\} \text{ and } PLR_n := LR_n + \sum_{j=1}^{2} p_n(\sigma_j^2(\hat{\boldsymbol{\psi}},\alpha)). \tag{15}$$

The hard bound is imposed on the values of α in order to avoid an issue of the infinite Fisher information for testing H_{02}. However, the LRTS may have a reduced power if the true value of α does not satisfy the constraint $[\epsilon, 1-\epsilon]$ given an ad hoc constant $\epsilon > 0$. For this reason, we also

develop the EM-test in Section 5 which does not impose a direct constraint on the value of α.

With $\boldsymbol{s}(\boldsymbol{W})$ in (9), partition $\boldsymbol{\mathcal{I}} = \mathbb{E}[\boldsymbol{s}(\boldsymbol{W})\boldsymbol{s}(\boldsymbol{W})^\top]$ and define

$$\boldsymbol{\mathcal{I}} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}} & \boldsymbol{\mathcal{I}}_{\boldsymbol{\eta\lambda}} \\ \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\eta}} & \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\lambda}} \end{pmatrix}, \quad \boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}} = \mathbb{E}[\boldsymbol{s}_{\boldsymbol{\eta}}(\boldsymbol{W})\boldsymbol{s}_{\boldsymbol{\eta}}(\boldsymbol{W})^\top], \quad \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\eta}} = \mathbb{E}[\boldsymbol{s}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W})\boldsymbol{s}_{\boldsymbol{\eta}}(\boldsymbol{W})^\top], \quad \boldsymbol{\mathcal{I}}_{\boldsymbol{\eta\lambda}} = \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\eta}}^\top,$$
$$\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\lambda}} = \mathbb{E}[\boldsymbol{s}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W})\boldsymbol{s}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W})^\top], \quad \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} = \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\lambda}} - \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\eta}}\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}}^{-1}\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta\lambda}}, \quad \text{and} \quad \boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}} := (\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}})^{-1}\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}},$$

where $\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}} \sim N(0, \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}})$. Define a set that characterizes the feasible values of $\sqrt{n}\boldsymbol{t}_\lambda(\boldsymbol{\lambda}, \alpha)$ when $n \to \infty$ by the cone

$$\Lambda_{\boldsymbol{\lambda}} = \Big\{\sqrt{n}\alpha(1-\alpha)v(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \Theta_{\boldsymbol{\lambda}}\Big\}.$$

Define $\hat{\boldsymbol{t}}_\lambda$ by

$$r_\lambda(\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}) = \inf_{\boldsymbol{t}_{\boldsymbol{\lambda}} \in \Lambda_{\boldsymbol{\lambda}}} r_\lambda(\boldsymbol{t}_{\boldsymbol{\lambda}}), \quad r_\lambda(\boldsymbol{t}_{\boldsymbol{\lambda}}) := (\boldsymbol{t}_{\boldsymbol{\lambda}} - \boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}(\boldsymbol{t}_{\boldsymbol{\lambda}} - \boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}}), \tag{16}$$

where $\hat{\boldsymbol{t}}_\lambda$ is a projection of a random Gaussian random variable $\boldsymbol{G}_{\boldsymbol{\lambda}}$ on a cone $\Lambda_{\boldsymbol{\lambda}}$.

The following proposition establishes the asymptotic distribution of LRTS or PLRTS under the null hypothesis $H_0 : M = 1$.

Proposition 4. *Suppose that assumptions 1 and 2 hold. Under the null hypothesis $H_0 : M_0 = 1$, (a) $\boldsymbol{t}(\hat{\boldsymbol{\psi}}, \alpha) = O_p(n^{-1/2})$ for any $\alpha \in (0,1)$, (b) $LR_n \xrightarrow{d} (\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}$ and $PLR_n \xrightarrow{d} (\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}} +$ plim$_{n\to\infty} \sum_{j=1}^2 p_n(\sigma_j^2(\hat{\boldsymbol{\psi}}, \alpha))$.*

Proposition 4(a) implies that $\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}^* = O_p(n^{-1/4})$ for $j = 1, 2$. The $n^{1/4}$ convergence rate is a consequence of the linear dependency in (4), where the identification of the parameter $\boldsymbol{\theta}$ relies on the fourth-order Taylor approximation of the log-likelihood function. This rate is also the best convergence rate for an over-parameterized mixture under the strong identifiability condition (Chen, 1995). When we choose the penalty function so that $\sum_{j=1}^2 p_n(\sigma_j^2(\hat{\boldsymbol{\psi}}, \alpha)) = o_p(1)$ under the null hypothesis of $M = 1$, PLR_n has the same asymptotic null distribution as that of LR_n.

4 Likelihood ratio test for $H_0 : M = M_0$ against $H_A : M = M_0 + 1$

In this section, we build upon the analysis from the previous section and derive the asymptotic distribution of the PLRTS for testing the null hypothesis of M_0 components against an alternative of $(M_0 + 1)$ components, where $M_0 \geq 2$.

Consider a random sample of n with a panel length of T independent observations $\{\boldsymbol{W}_i\}_{i=1}^n$, where $\boldsymbol{W}_i = \{(Y_{it}, \boldsymbol{X}_{it}^\top, \boldsymbol{Z}_{it}^\top)^\top\}_{t=1}^T$ from an M_0-component density $f_{M_0}(\boldsymbol{w}; \boldsymbol{\vartheta}_{M_0})$ defined in equa-

tion (17):

$$f_{M_0}(\boldsymbol{w};\boldsymbol{\vartheta}^*_{M_0}) = \sum_{j=1}^{M_0} \alpha^*_j f(\boldsymbol{w};\boldsymbol{\gamma}^*,\boldsymbol{\theta}^*_j), \tag{17}$$

where $\boldsymbol{\vartheta}^*_{M_0} = (\boldsymbol{\theta}^*_1, \boldsymbol{\theta}^*_2, \ldots, \boldsymbol{\theta}^*_{M_0}, \alpha^*_1, \ldots, \alpha^*_{M_0-1}, \boldsymbol{\gamma}^*) \in \Theta_{\boldsymbol{\vartheta}_{M_0}}$ and $\alpha^*_{M_0} = 1 - \sum_{j=1}^{M-1} \alpha^*_j$.

Let the density of the (M_0+1)-component model be defined by:

$$f_{M_0+1}(\boldsymbol{w};\boldsymbol{\vartheta}_{M_0+1}) = \sum_{j=1}^{M_0+1} \alpha_j f(\boldsymbol{w};\boldsymbol{\gamma},\boldsymbol{\theta}_j), \tag{18}$$

where $\boldsymbol{\vartheta}_{M_0+1} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{M_0+1}, \alpha_1, \ldots, \alpha_{M_0}, \boldsymbol{\gamma}) \in \Theta_{\vartheta_{M_0+1}}$ as defined in (17). We assume $\mu^*_1 < \mu^*_2, \ldots, < \mu^*_{M_0}$ in the true parameters for identification.

The (M_0+1)-component model (18) gives rise to the true density (17) in two different cases: (i) two components have the same mixing parameter, and (ii) one component has zero mixing proportion. Accordingly, we partition the null hypothesis of $H_0 : M = M_0$ into two as $H_0 = H_{01} \cup H_{02}$, with $H_{01} : \boldsymbol{\theta}_h = \boldsymbol{\theta}_{h+1} = \boldsymbol{\theta}^*_h$ for some $h = 1 \ldots, M_0$, and $H_{02} : \alpha_h = 0$ for some $h = 1, \ldots, M_0 + 1$.

We first analyze the infinite Fisher information problem for testing H_{02}. Partition H_{02} as $H_{02} = \cup_{h=1}^{M_0} H_{0,2h}$, where $H_{0,2h} : \alpha_h = 0$. Define the subset of $\Theta_{\vartheta_{M_0+1}}$ corresponding to $H_{0,2h}$ as

$$\begin{aligned}\Upsilon^*_{2h} = \{\boldsymbol{\vartheta}_{M_0+1} \in \Theta_{\boldsymbol{\vartheta}_{M_0+1}} : \alpha_h = 0; (\alpha_j, \mu_j, \sigma_j) = (\alpha^*_j, \mu^*_j, \sigma^*_j) \text{ for } j < h; \\ (\alpha_j, \mu_j, \sigma_j) = (\alpha^*_{j-1}, \mu^*_{j-1}, \sigma^*_{j-1}) \text{ for } j > h\}.\end{aligned}$$

The score for testing $H_{0,2h} : \alpha_h = 0$ takes the form $\nabla_{\alpha_h} \log f_{M_0+1}(\boldsymbol{W}_i, \boldsymbol{\vartheta}_{M_0+1}) = [f(\boldsymbol{W}_i; \mu_h, \sigma^2_h) - f(\boldsymbol{W}_i; \mu^*_{M_0}, \sigma^{2*}_{M_0})]/f_{M_0}(\boldsymbol{W}_i, \boldsymbol{\vartheta}^*_{M_0})$. Because (μ_h, σ^2_h) is not identified when $\alpha_h = 0$, the Fisher information matrix of the LRTS for testing $H_{0,2h} : \alpha_h = 0$ depends on the supremum of the variance of $\nabla_{\alpha_h} \log f_{M_0+1}(\boldsymbol{W}_i; \boldsymbol{\vartheta}_{M_0+1})$ over $\boldsymbol{\vartheta}_{M_0+1} \in \Upsilon^*_{2h}$. The Fisher information is infinite unless there is an a priori restriction on the values of σ^2_j.

Proposition 5. $\sup_{\vartheta_{M_0+1} \in \Upsilon^*_{2h}} \mathbb{E}[\{\nabla_{\alpha_h} \log f_{M_0+1}(\boldsymbol{W}_i, \vartheta_{M_0+1})\}^2] < \infty$ *if and only if* $\max\{\sigma^2 : \sigma \in \Theta_\sigma\} < 2\max\{\sigma^{2*}_1, \ldots, \sigma^{2*}_{M_0}\}$.

Since the restriction on the values of σ^2_j in Proposition 5 is difficult to justify and not easy to enforce in practice, we focus on testing H_{01}.

Partition H_{01} as $H_{01} = \cup_{h=1}^{M_0} H_{0,1h}$, where $H_{0,1h} : \boldsymbol{\theta}_h = \boldsymbol{\theta}_{h+1}$ with $\mu_1 < \cdots < \mu_h = \mu_{h+1} < \cdots < \mu_{M_0+1}$. We impose these inequality constraints on μ_j for component identification. There are M_0 ways to describe the M_0 component null model in the space of (M_0+1) component models, each way corresponding to the null hypothesis of $H_{0,1h} : \boldsymbol{\theta}_h = \boldsymbol{\theta}_{h+1}$ for $h = 1, 2, ..., M_0$. Testing $H_{0,1h} : \boldsymbol{\theta}_h = \boldsymbol{\theta}_{h+1}$ in the M_0-component null models is similar to testing $H_{01} : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ in the one

component null model in section 3.

Define the subset of $\Theta_{\vartheta_{M_0+1}}$ corresponding to $H_{0,1h}$ as:

$$\begin{aligned}\Upsilon_{1h}^* := \Big\{ &\boldsymbol{\vartheta}_{M_0+1} \in \Theta_{\vartheta_{M_0+1}} : \alpha_h + \alpha_{h+1} = \alpha_h^* \text{ and } \boldsymbol{\theta}_h = \boldsymbol{\theta}_{h+1} = \boldsymbol{\theta}_h^*; \boldsymbol{\gamma} = \boldsymbol{\gamma}^*; \alpha_j = \alpha_j^* \\ &\text{and } \boldsymbol{\theta}_j = \boldsymbol{\theta}_j^* \text{ for } 1 \le j < h; \alpha_j = \alpha_{j-1}^* \text{ and } \boldsymbol{\theta}_j = \boldsymbol{\theta}_{j-1}^* \text{ for } h+1 \le j \le M_0+1 \Big\}\end{aligned} \tag{19}$$

for $h = 1, \ldots, M_0$. The set $\Upsilon_1^* := \cup_{h=1}^{M_0} \Upsilon_{1h}^*$ corresponds to $H_{01} = \cup_{h=1}^{M_0} H_{0,1h}$.

Suppose the null hypothesis of $M = M_0$ holds with the true density (17). Because any parameter in $\Upsilon_1^* = \cup_{h=1}^{M_0} \Upsilon_{1h}^*$ can generate the true density $f_{M_0}(\boldsymbol{w}; \boldsymbol{\vartheta}_{M_0}^*) = \sum_{j=1}^{M_0} \alpha_0^{j*} f(\boldsymbol{w}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}_j^*)$, we need to restrict the estimators under the (M_0+1)-component model to be in a neighborhood of Υ_{1h}^* in order to test $H_{0,1h}$.

Recall that $\mu_1^* < \mu_2^* \ldots < \mu_{M_0}^*$. Let $\underline{\Theta}_\mu$ and $\overline{\Theta}_\mu$ denote the lower bound and upper bounds of Θ_μ. Define $D_1^* = [\underline{\Theta}_\mu, \frac{\mu_1^*+\mu_2^*}{2}] \times \Theta_\beta \times \Theta_{\sigma^2}$, $D_h^* = [\frac{\mu_{h-1}^*+\mu_h^*}{2}, \frac{\mu_h^*+\mu_{h+1}^*}{2}] \times \Theta_\beta \times \Theta_{\sigma^2}$ for $h = 2, \ldots, M_0-1$, $D_{M_0}^* = [\frac{\mu_{M_0-1}^*+\mu_{M_0}^*}{2}, \overline{\Theta}_\mu] \times \Theta_\beta \times \Theta_{\sigma^2}$. Then $D_h^* \subset \Theta_\theta$ is a neighborhood containing θ_h^* but not θ_j^* for $j \neq h$. For $h = 1, \ldots M_0$, give a small positive constant $\epsilon > 0$, define a restricted parameter space $\boldsymbol{\Psi}_h^* \subset \Theta_{\vartheta_{M_0+1}}(\epsilon)$ as

$$\boldsymbol{\Psi}_h^* = \left\{ \begin{array}{r} \alpha_1, \ldots, \alpha_{M_0+1} \in [\epsilon, 1-\epsilon]; \sum_{j=1}^{M_0+1} \alpha_j = 1; \gamma \in \Theta_\gamma; \theta \in \Theta_\theta : \boldsymbol{\theta}_j \in D_j^* \text{ for } j = 1, \ldots, h-1; \\ \boldsymbol{\theta}_h, \boldsymbol{\theta}_{h+1} \in D_h^*; \boldsymbol{\theta}_j \in D_{j-1}^* \text{ for } j = h+2, \ldots, M_0+1. \end{array} \right\} \tag{20}$$

Note that $\boldsymbol{\Psi}_h^* \cap \Upsilon_{1h}^* \neq \emptyset$ and $\boldsymbol{\Psi}_h^* \cap \Upsilon_{1l}^* = \emptyset$ if $h \neq l$ while $\cup_{h=1}^{M_0} \boldsymbol{\Psi}_h^* = \Theta_{\vartheta_{M_0+1}}(\epsilon)$.

Let $\hat{\boldsymbol{\Psi}}_h^*$ and $\hat{D}_h^*$ be consistent estimators of $\boldsymbol{\Psi}_h^*$ and D_h^*, which can be constructed from a consistent estimator of $\boldsymbol{\vartheta}_{M_0}^*$ in the M_0-component model. We test $H_{0,1h} : \boldsymbol{\theta}_h = \boldsymbol{\theta}_{h+1}$ by estimating the (M_0+1)-component model under the restriction that $\vartheta^{M_0+1} \in \hat{\boldsymbol{\Psi}}_h^*$.

For $h = 1, 2, ..., M_0$, define the "local" PMLE that maximizes the log-likelihood function of the (M_0+1)-component model under the constrain that $\boldsymbol{\vartheta}_{M_0+1} \in \hat{\boldsymbol{\Psi}}_h^*$ in (20) by

$$\hat{\boldsymbol{\vartheta}}_{M_0+1}^h = \underset{\boldsymbol{\vartheta}_{M_0+1} \in \hat{\boldsymbol{\Psi}}_h^*}{\arg\max} \; L_{M_0+1,n}(\boldsymbol{\vartheta}_{M_0+1}) + \tilde{p}_n(\boldsymbol{\vartheta}_{M_0+1}),$$

where

$$L_{M,n}(\boldsymbol{\vartheta}_M) := \sum_{i=1}^n \log f_M(\boldsymbol{W}_i; \boldsymbol{\vartheta}_M) \quad \text{and} \quad \tilde{p}_n(\boldsymbol{\vartheta}_M) := \sum_{j=1}^M p_n(\sigma_j^2; \hat{\sigma}_{0,j}^2).$$

with

$$p_n(\sigma_j^2; \hat{\sigma}_{0,j}^2) := -a_n \{ \hat{\sigma}_{0,j}^2 / \sigma_j^2 + \log(\sigma_j^2 / \hat{\sigma}_{0,j}^2) - 1 \}, \tag{21}$$

where $\hat{\sigma}^2_{0,j}$ is a root-n consistent estimator of $\sigma^2_{0,j}$ from M_0-component model under the null hypothesis. Because $\hat{\sigma}^2_j - \sigma^2_{0,j} = O_p(n^{-1/4})$ under the null hypothesis (c.f., Proposition 4(a)), $p_n(\hat{\sigma}^2_j; \hat{\sigma}^2_{0,j}) = o_p(1)$ when a_n is chosen to be $o(n^{1/4})$.

Under $H_0 : M = M_0$, $\mathbf{\Psi}^*_h$ contains a set of parameters Υ^*_{1h} defined in (19) such that $f_{M_0+1}(\boldsymbol{w}; \boldsymbol{\vartheta}_{M_0+1})$ is equal to $f_{M_0}(\boldsymbol{w}; \boldsymbol{\vartheta}^*_{M_0})$ for any $\boldsymbol{\vartheta}_{M_0+1} \in \Upsilon^*_{1h}$ and is therefore the density function from which the data is generated. These penalized likelihood estimators are consistent.

Proposition 6. *Suppose that Assumption 1 holds. Then, under the null hypothesis* $H_0 : M = M_0$, $\inf_{\boldsymbol{\vartheta}_{M_0+1} \in \mathbf{\Psi}^*_h} |\hat{\boldsymbol{\vartheta}}^h_{M_0+1} - \boldsymbol{\vartheta}_{M_0+1}| \xrightarrow{p} 0$ *for* $h = 1, 2, ..., M_0$.

Consider the local PLRTS for testing $H_{0,1h} : \boldsymbol{\vartheta}_h = \boldsymbol{\vartheta}_{h+1}$ defined by

$$PLR^{M_0,h}_n := 2\{L_{M_0+1,n}(\hat{\boldsymbol{\vartheta}}^h_{M_0+1}) + \tilde{p}_n(\hat{\boldsymbol{\vartheta}}^h_{M_0+1}) - L_{M_0,n}(\hat{\boldsymbol{\vartheta}}_{M_0})\} \quad \text{for } h = 1, 2, ..., M_0.$$

The test utilizing the local PLRTS, denoted as $PLR^{M_0,h}_n$, possesses power solely against local alternatives within the restricted parameter space of $\mathbf{\Psi}^*_h$. To guarantee power against local alternatives over a wide range of directions, we consider the PLRTS characterized by the maximum of the local PLRTS for $h = 1, ..., M_0$, as defined by

$$PLR_n(M_0) := \max\{PLR^{M_0,1}_n, PLR^{M_0,2}_n, ..., PLR^{M_0,M_0}_n\}. \tag{22}$$

Because $\Theta_{\vartheta_{M_0+1}}(\epsilon) = \cup^{M_0}_{h=1} \hat{\mathbf{\Psi}}^*_h$, $PLR_n(M_0)$ is identical to $\max_{\boldsymbol{\vartheta}_{M_0+1} \in \Theta_{\boldsymbol{\vartheta}_{M_0+1}}(\epsilon)}\{L_{M_0+1,n}(\boldsymbol{\vartheta}_{M_0+1}) + \tilde{p}_n(\boldsymbol{\vartheta}_{M_0+1})\} - L_{M_0,n}(\hat{\boldsymbol{\vartheta}}_{M_0})$.

To derive the asymptotic null distribution of $PLR_n(M_0)$, collect the score vector for testing $H_{0,1h}$ for $h = 1, \ldots, M_0$ into one vector as

$$\tilde{\boldsymbol{s}}(\boldsymbol{W}) = \begin{pmatrix} \tilde{\boldsymbol{s}}_{\boldsymbol{\eta}}(\boldsymbol{W}) \\ \tilde{\boldsymbol{s}}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W}) \end{pmatrix}, \text{ where } \underset{(M_0+p+q+1)\times 1}{\tilde{\boldsymbol{s}}_{\boldsymbol{\eta}}(\boldsymbol{W})} = \begin{pmatrix} \boldsymbol{s}_{\boldsymbol{\alpha}}(\boldsymbol{W}) \\ \boldsymbol{s}_{(\boldsymbol{\gamma},\boldsymbol{\nu})}(\boldsymbol{W}) \end{pmatrix} \text{ and } \tilde{\boldsymbol{s}}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W}) = \begin{pmatrix} \boldsymbol{s}^1_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W}) \\ \vdots \\ \boldsymbol{s}^{M_0}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W}) \end{pmatrix}, \tag{23}$$

where

$$\begin{aligned} \boldsymbol{s}_{\boldsymbol{\alpha}}(\boldsymbol{W}) &= \begin{pmatrix} f(\boldsymbol{W}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*_1) - f(\boldsymbol{W}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*_{M_0}) \\ \vdots \\ f(\boldsymbol{W}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*_{M_0-1}) - f(\boldsymbol{W}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*_{M_0}) \end{pmatrix} \Big/ f_{M_0}(\boldsymbol{W}; \boldsymbol{\vartheta}^*_{M_0}), \\ \boldsymbol{s}_{(\boldsymbol{\gamma},\boldsymbol{\nu})}(\boldsymbol{W}) &= \sum^{M_0}_{j=1} \alpha^*_j \nabla_{(\boldsymbol{\gamma},\boldsymbol{\nu})} f(\boldsymbol{W}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*_j) / f_{M_0}(\boldsymbol{W}; \boldsymbol{\vartheta}^*_{M_0}), \\ \boldsymbol{s}^h_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W}) &= \widetilde{\nabla}_{\boldsymbol{\theta}_h \boldsymbol{\theta}^\top_h} f(\boldsymbol{W}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*_h) / f_{M_0}(\boldsymbol{W}; \boldsymbol{\vartheta}^*_{M_0}) \quad \text{for } h = 1, 2, ..., M_0, \end{aligned} \tag{24}$$

with $\widetilde{\nabla}_{\boldsymbol{\theta}_h\boldsymbol{\theta}_h^\top} f(\boldsymbol{W};\boldsymbol{\gamma}^*,\boldsymbol{\theta}_h^*) := (c_{11}\nabla_{\theta_{h1}\theta_{h1}} f^*, ..., c_{(q+2)(q+2)}\nabla_{\theta_{h,q+2}\theta_{h,q+2}} f^*, c_{12}\nabla_{\theta_{h1}\theta_{h2}} f^*, ..., c_{(q+1)(q+2)}\nabla_{\theta_{h,q+1}\theta_{h,q+2}} f^*)^\top$ for $\boldsymbol{\theta}_h := (\theta_{h1}, \theta_{h2}, \theta_{h3}, ..., \theta_{h,q+2})^\top := (\mu_h, \sigma_h^2, \beta_{h1}, ..., \beta_{hq})^\top$ and $c_{jk} = 1/2$ for $j \neq k$ and $c_{jk} = 1$ for $j = k$. Define

$$\begin{aligned} \tilde{\boldsymbol{\mathcal{I}}} &:= \mathbb{E}[\tilde{\boldsymbol{s}}(\boldsymbol{W})\tilde{\boldsymbol{s}}(\boldsymbol{W})^\top], \quad \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\eta}} := \mathbb{E}[\tilde{\boldsymbol{s}}_{\boldsymbol{\eta}}(\boldsymbol{W})\tilde{\boldsymbol{s}}_{\boldsymbol{\eta}}(\boldsymbol{W})^\top], \quad \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda\eta}} := \mathbb{E}[\tilde{\boldsymbol{s}}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W})\tilde{\boldsymbol{s}}_{\boldsymbol{\eta}}(\boldsymbol{W})^\top], \\ \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\eta\lambda}} &:= \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda\eta}}^\top, \quad \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda\lambda}} := \mathbb{E}[\tilde{\boldsymbol{s}}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W})\tilde{\boldsymbol{s}}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W})^\top], \quad \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} := \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda\lambda}} - \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda\eta}}\tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\eta}}^{-1}\tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\eta\lambda}}. \end{aligned} \tag{25}$$

Then, the asymptotic distribution of the normalized score function is given by

$$\tilde{\boldsymbol{S}}_n := \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\boldsymbol{s}}(\boldsymbol{W}_i) \xrightarrow{d} \tilde{\boldsymbol{S}} \sim N(\boldsymbol{0}, \tilde{\boldsymbol{\mathcal{I}}}),$$

where, in view of (23), $\tilde{\boldsymbol{S}}$ may be partitioned as $\tilde{\boldsymbol{S}} = (\tilde{\boldsymbol{S}}_{\boldsymbol{\eta}}^\top, \tilde{\boldsymbol{S}}_{\boldsymbol{\lambda\lambda}}^\top)^\top$ with $n^{-1/2}\sum_{i=1}^n \tilde{\boldsymbol{s}}_{\boldsymbol{\eta}}(\boldsymbol{W}_i) \xrightarrow{d} \tilde{\boldsymbol{S}}_{\boldsymbol{\eta}}$ and $n^{-1/2}\sum_{i=1}^n \tilde{\boldsymbol{s}}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W}_i) \xrightarrow{d} \tilde{\boldsymbol{S}}_{\boldsymbol{\lambda\lambda}}$.

Let $\tilde{\boldsymbol{S}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} := (\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^1, \ldots, \boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^{M_0})^\top := \tilde{\boldsymbol{S}}_{\boldsymbol{\lambda\lambda}} - \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda\eta}}\tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\eta}}^{-1}\tilde{\boldsymbol{S}}_{\boldsymbol{\eta}} \sim N(0, \tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda},\boldsymbol{\eta}})$ be a $\mathbb{R}^{M_0(q+2)(q+1)/2}$-valued random vector. For $h = 1, 2, ..., M_0$, define $\tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h := \mathbb{E}[\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h(\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h)^\top]$ and $\boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h := (\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h)^{-1}\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h$.

Define $\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^h$ analogously to $\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}$ as:

$$r_{\boldsymbol{\lambda}}^h(\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^h) = \inf_{\boldsymbol{t}_{\boldsymbol{\lambda}}^h \in \Lambda_{\boldsymbol{\lambda}}} r^h(\boldsymbol{t}_{\boldsymbol{\lambda}}^h); \quad r_{\boldsymbol{\lambda}}^h(\boldsymbol{t}_{\boldsymbol{\lambda}}^h) := (\boldsymbol{t}_{\boldsymbol{\lambda}}^h - \boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h)^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h (\boldsymbol{t}_{\boldsymbol{\lambda}}^h - \boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h) \quad \text{for } h = 1, 2, ..., M_0. \tag{26}$$

The local quadratic-form approximation of the log-likelihood function $LR_n^{M_0,h}$ around $\Upsilon_{1h}^* \subset \Theta_{\vartheta_{M_0+1}}$ shares an identical structure to the approximation we derived in Section 3 in testing H_{01} in the test of homogeneity. Consequently, we can show that $PLR_n^{M_0,h} \xrightarrow{d} (\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^h)^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^h$. Then, given (22), the asymptotic null distribution of the PLRTS for testing H_{01} is given by the maximum over $(\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^h)^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^h$'s for $h = 1, 2, ..., M_0$.

Assumption 3. *(a) $\alpha_j^* \in (\epsilon, 1-\epsilon)$ for $j = 1, \ldots, M_0$. (b) $\tilde{\boldsymbol{\mathcal{I}}}$ is non-singular. (c) a_n in (21) satisfies $a_n = O(1)$.*

Proposition 7. *Suppose that Assumptions 1-3 are satisfied. Then under the null hypothesis $H_0 : M = M_0$, $PLR_n(M_0) \xrightarrow{d} \max\{(\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^1)^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^1 \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^1, \ldots, (\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^{M_0})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^{M_0} \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^{M_0}\}$.*

The asymptotic null distribution of $PLR_n(M_0)$ is non-standard but it is straightforward to simulate the random variable from the asymptotic null distribution using the estimates. Specifically, we simulate a draw of $\tilde{\boldsymbol{S}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} = (\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^1, \ldots, \boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^{M_0})^\top$ from $N(0, \hat{\tilde{\boldsymbol{\mathcal{I}}}}_{\boldsymbol{\lambda},\boldsymbol{\eta}})$, where $\hat{\tilde{\boldsymbol{\mathcal{I}}}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ is a sample analogue estimator of $\tilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$. Then, compute $\boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h = (\hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h)^{-1}\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h$ and obtain $\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^h$ analogously to (26) using an estimator of $\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h$ for $h = 1, ..., M_0$, and a simulated random draw is computed as $\max\{(\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^1)^\top \hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^1 \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^1, \ldots, (\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^{M_0})^\top \hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^{M_0} \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^{M_0}\}$. Appendix B.2-B.3 present an expression for score functions using Hermit polynomials.

5 EM test for $H_0 : M = M_0$ against $H_A : M = M_0 + 1$

This section develops an EM test used for testing the hypothesis $H_0 : M = M_0$ against the alternative hypothesis $H_A : M = M_0 + 1$. A key limitation of the PLRT, as discussed in the previous section, is that the computation of mixing probabilities, denoted as α_j, is subject to a hard constraint, which is dictated by an arbitrary choice of bounds. The EM test, on the other hand, circumvents the need for imposing an explicit constraint on the α_j values. It achieves this by performing a limited number of EM steps, starting from a predetermined set of α_j values. The EM test approach offers certain advantages, including computational simplicity and less stringent assumptions.

Let $\mathcal{T}$ be a finite set of numbers in $(0, 0.5]$ with $0.5 \in \mathcal{T}$, and let $p(\tau) \leq 0$ be a penalty term that is continuous in τ, $p(0.5) = 0$, and $p(\tau) \to -\infty$ as τ goes to 0. Specifically, we choose

$$p(\tau) := \log(2 \min\{\tau, 1 - \tau\}).$$

For each $\tau_0 \in \mathcal{T}$, let $\tau^{(1)}(\tau_0) = \tau_0$ and define the restricted penalized MLE by

$$\boldsymbol{\vartheta}^{h(1)}_{M_0+1}(\tau_0) = \underset{\boldsymbol{\vartheta}_{M_0+1} \in \Theta^h_{\boldsymbol{\vartheta}_{M_0+1}}(\tau)}{\arg\max} PL_n(\boldsymbol{\vartheta}_{M_0+1}, \tau_0)$$

where $\Theta^h_{\boldsymbol{\vartheta}_{M_0+1}}(\tau_0) := \{\boldsymbol{\theta} \in \hat{\boldsymbol{\Psi}}_h : \alpha_h/(\alpha_h + \alpha_{h+1}) = \tau_0\}$ and

$$PL_n(\boldsymbol{\vartheta}_{M_0+1}, \tau) := L_{M_0+1,n}(\boldsymbol{\vartheta}_{M_0+1}) + \tilde{p}_n(\boldsymbol{\vartheta}_{M_0+1}) + p(\tau).$$

Starting from $(\boldsymbol{\vartheta}^{h(1)}_{M_0+1}(\tau_0), \tau^{h(1)}(\tau_0))$ with $\tau^{h(1)}(\tau_0) = \tau_0$, update $\boldsymbol{\vartheta}^{h(k)}_{M_0+1}(\tau_0)$ and $\tau^{h(k)}(\tau_0)$ by the following generalized EM algorithm. Denote the estimators after the k-th round of EM algorithm iteration by $\vartheta^{h(k)}_{M_0+1}$ and $\tau^{h(k)}$. In the E-step, for $i = 1, \ldots, N$ and $j = 1, \ldots, M_0 + 1$, compute the weight for observation i and type j as:

$$\begin{aligned}
w^{(k)}_{ij} &= \begin{cases} \alpha^{(k)}_j f(\boldsymbol{W}_i; \boldsymbol{\gamma}^{(k)}, \boldsymbol{\theta}^{(k)}_j)/f_{M_0+1}(\boldsymbol{W}_i; \boldsymbol{\vartheta}^{h(k)}_{M_0+1}(\tau_0)), j = 1, \ldots, h-1, \\ \alpha^{(k)}_{j-1} f(\boldsymbol{W}_i; \boldsymbol{\gamma}^{(k)}, \boldsymbol{\theta}^{(k)}_j)/f_{M_0+1}(\boldsymbol{W}_i; \boldsymbol{\vartheta}^{h(k)}_{M_0+1}(\tau_0)), j = h+2, \ldots, M_0+1, \end{cases} \\
w^{(k)}_{ih} &= \tau^{h(k)} \alpha^{(k)}_h f(\boldsymbol{W}_i; \boldsymbol{\gamma}^{(k)}, \boldsymbol{\theta}^{(k)}_h)/f_{M_0+1}(\boldsymbol{W}_i; \boldsymbol{\vartheta}^{h(k)}_{M_0+1}(\tau_0)), \\
w^{(k)}_{i,h+1} &= (1 - \tau^{h(k)}) \alpha^{(k)}_h f(\boldsymbol{W}_i; \boldsymbol{\gamma}^{(k)}, \boldsymbol{\theta}^{(k)}_{h+1})/f_{M_0+1}(\boldsymbol{W}_i; \boldsymbol{\vartheta}^{h(k)}_{M_0+1}(\tau_0)),
\end{aligned} \tag{27}$$

where, for brevity, we drop the superscript h and its dependency on τ_0 from the notations such as $w^{h(k)}_{ij}(\tau_0)$.

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

$$\alpha_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} w_{ij}^{(k)} \quad \text{for } j = 1, \ldots, M_0+1 \quad \text{and}$$
$$\tau^{h(k+1)} = \arg\min_{\tau} \left\{ \sum_{i=1}^{n} w_{ih}^{(k)} \log(\tau) + \sum_{i=1}^{n} w_{i,h+1}^{(k)} \log(1-\tau) + p(\tau) \right\}.$$

We also update $\boldsymbol{\theta}_j$ and $\boldsymbol{\gamma}$ as

$$(\sigma_j^{(k+1)})^2 = \arg\min_{\sigma_j^2} \left\{ \sum_{i=1}^{n} w_i^{j(k)} \sum_{t=1}^{T} (y_{it} - \mu_j^{(k+1)} - \boldsymbol{z}_{it}^\top \boldsymbol{\gamma}^{(k+1)} - \boldsymbol{x}_{it}^\top \boldsymbol{\beta}_j^{(k+1)})^2 + p_n(\sigma_j^2) \right\},$$
$$\boldsymbol{\gamma}^{(k+1)} = \left(\sum_{i=1}^{n}\sum_{t=1}^{T} \boldsymbol{z}_{it}\boldsymbol{z}_{it}^\top \right)^{-1} \left(\sum_{i=1}^{n}\sum_{t=1}^{T} \boldsymbol{z}_{it} \left(y_{it} - \sum_{j=1}^{M_0+1} w_{ij}^{(k)} \tilde{\boldsymbol{x}}_{it}^\top \begin{pmatrix} \mu_j^{(k)} \\ \boldsymbol{\beta}_j^{(k)} \end{pmatrix} \right) \right), \quad \text{and}$$
$$\begin{pmatrix} \mu_j^{(k+1)} \\ \boldsymbol{\beta}_j^{(k+1)} \end{pmatrix} = \left(\sum_{i=1}^{n} w_{ij}^{(k)} \sum_{t=1}^{T} \tilde{\boldsymbol{x}}_{it}\tilde{\boldsymbol{x}}_{it}^\top \right)^{-1} \left(\sum_{i=1}^{n} w_{ij}^{(k)} \sum_{t=1}^{T} \tilde{\boldsymbol{x}}_{it}(y_{it} - \boldsymbol{z}_{it}^\top \boldsymbol{\gamma}^{(k+1)}) \right),$$

where $\tilde{\boldsymbol{x}}_{it} = (1, \boldsymbol{x}_{it}^\top)^\top$. In the updating procedure, $\boldsymbol{\vartheta}_{M_0+1}^{h(k+1)}(\tau_0)$ is not restricted to be in $\hat{\boldsymbol{\Psi}}_h^*$.

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

$$M_n^{h(k)}(\tau_0) := 2\left\{ PL_n(\boldsymbol{\vartheta}_{M_0+1}^{h(k)}(\tau_0), \tau^{h(k)}(\tau_0)) - L_{M_0,n}(\hat{\boldsymbol{\vartheta}}_{M_0}) \right\}. \tag{28}$$

With a pre-determined finite number K, define the *local* EM test statistic by taking maximum of $M_n^{h(k)}(\tau_0)$ across different τ_0's as

$$EM_n^h := \max\{M_n^{h(K)}(\tau_0) : \tau_0 \in \mathcal{T}\}. \tag{29}$$

The test statistic EM_n^h tests $H_{0,1h} : \boldsymbol{\theta}_h = \boldsymbol{\theta}_{h+1}$ and has a power against the local alternative that splits the h-th component of the null M_0-component model into two different components. To achieve power against a wide range of local alternatives, we consider the EM test statistic that takes the maximum of M_0 local EM test statistics:

$$EM_n(M_0) := \max\{EM_n^{1(K)}, \ldots, EM_n^{M_0(K)}\}. \tag{30}$$

Proposition 8. *Suppose that Assumptions 1–3 hold and* $\{0.5\} \in \mathcal{T}$. *Then, under the null hypothesis* $H_0 : M = M_0$, *for any finite* K, $EM_n(M_0) \xrightarrow{d} \max\{(\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^1)^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^1 \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^1, \ldots, (\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^{M_0})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^{M_0} \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^{M_0}\}$.

Therefore, the asymptotic null distribution of EM test statistic $EM_n(M_0)$ is the same as that of the PLRTS.

6 Asymptotic Distribution under Local Alternatives

We derive the asymptotic distribution of PLRTS and EM test static under local alternatives. For brevity, we focus on testing $H_0 : M = 1$ against $H_A : M = 2$. Consider the following local alternative to the homogeneous model $f(\boldsymbol{w}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*)$ with $\boldsymbol{\theta}^* = (\mu^*, \sigma^{*2}, (\boldsymbol{\beta}^*)^\top)^\top$. For brevity, we omit the common parameter $\boldsymbol{\gamma}$ in this section. In a reparameterized parameter, $\boldsymbol{\psi}^* = ((\boldsymbol{\nu}^*)^\top, (\boldsymbol{\lambda}^*)^\top)^\top$. For $\alpha^* \in (0,1)$ and a local parameter $\boldsymbol{h} = (\boldsymbol{h}_{\boldsymbol{\nu}}^\top, \boldsymbol{h}_{\boldsymbol{\lambda}}^\top)^\top$ with $\boldsymbol{h}_{\boldsymbol{\lambda}} \in v(\boldsymbol{\Theta}_{\boldsymbol{\lambda}})$, we consider a sequence of contiguous local alternatives $(\alpha_n, \boldsymbol{\psi}_n^\top)^\top = (\alpha_n, \boldsymbol{\nu}_n^\top, \boldsymbol{\lambda}_n^\top) \in \boldsymbol{\Theta}_\alpha \times \boldsymbol{\Theta}_{\boldsymbol{\nu}} \times \boldsymbol{\Theta}_{\boldsymbol{\lambda}}$ such that, with $\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}, \alpha)$ given by (10),

$$\boldsymbol{h}_{\boldsymbol{\nu}} = \sqrt{n}(\boldsymbol{\nu}_n - \boldsymbol{\nu}^*), \quad \boldsymbol{h}_{\boldsymbol{\lambda}} = \sqrt{n}\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_n, \alpha_n), \quad \text{and} \quad \alpha_n = \alpha^* + o(1). \tag{31}$$

Equivalently, the non-reparameterized contiguous local alternatives are given by

$$\boldsymbol{\theta}_{1,n} = \boldsymbol{\nu}_n + (1 - \alpha_n)\boldsymbol{\lambda}_n \quad \text{and} \quad \boldsymbol{\theta}_{2,n} = \boldsymbol{\nu}_n - \alpha_n \boldsymbol{\lambda}_n \tag{32}$$

for $\boldsymbol{\nu}_n = \boldsymbol{\nu}^* + n^{-1/2}\boldsymbol{h}_\nu$ and $\boldsymbol{\lambda}_n = (\lambda_{1,n}, \lambda_{2,n}, ..., \lambda_{q+2,n})^\top$ with

$$\lambda_{j,n} = n^{-1/4}(\alpha_n(1 - \alpha_n))^{-1/2} h_{\lambda,j} \quad \text{for } j = 1, ..., q+2,$$

where $\boldsymbol{h}_{\boldsymbol{\lambda}} = (h_{\lambda,1}^2, ..., h_{\lambda,q+2}^2, h_{\lambda,1}h_{\lambda,2}, ..., h_{\lambda,q+1}h_{\lambda,q+2})^\top$. The local alternatives are of order $n^{1/4}$ rather than $n^{1/2}$. See the discussion after Proposition 4.

The following proposition provides the asymptotic distribution of the PLRT and EM test statistics under contiguous local alternatives.

Proposition 9. *Suppose that the assumptions in Proposition 8 hold for $M_0 = 1$. Consider a sequence of contiguous local alternatives $\boldsymbol{\vartheta}_{2,n} = (\alpha_n, \boldsymbol{\theta}_{1,n}^\top, \boldsymbol{\theta}_{2,n}^\top)^\top$ given in (32), where α_n and $\boldsymbol{\lambda}_n$ satisfy (31). Then, under $H_{1,n} : \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_{2,n}$, we have $PLR_n(1), EM_n(1) \xrightarrow{d} (\tilde{\boldsymbol{t}}_{\boldsymbol{\lambda}})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} \tilde{\boldsymbol{t}}_{\boldsymbol{\lambda}}$, where $\tilde{\boldsymbol{t}}_{\boldsymbol{\lambda}}$ has the same distribution as $\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}$ in Proposition 4 but replacing $\boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ with $(\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}})^{-1}\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}} + \boldsymbol{h}_{\boldsymbol{\lambda}}$.*

Importantly, a set of contiguous local alternatives considered in (32) excludes a sequence such that $\alpha_n \to 0$ or 1.

7 Sequential Hypothesis Testing

To estimate the number of components, we sequentially test $H_0 : M = r$ against $H_1 : M = r + 1$ starting from $r = 1$, and then $r = 2, \ldots, \bar{M}$, where $\bar{M}$ is the upper bound for the number of components, which is assumed to be larger than M_0. The first value for r that leads to a nonrejection

of H_0 gives our estimate for M_0. Robin and Smith (2000) develops a similar sequential hypothesis test for estimating the rank of a matrix.

For $M = 1, \ldots, \bar{M}$, let $c^M_{1-q_n}$ denote the $100(1-q_n)$ percentile of the cumulative distribution function of a random variable $\max\{(\hat{\boldsymbol{t}}^1_{\boldsymbol{\lambda}})^\top \boldsymbol{\mathcal{I}}^1_{\boldsymbol{\lambda},\boldsymbol{\eta}} \hat{\boldsymbol{t}}^1_{\boldsymbol{\lambda}}, \ldots, (\hat{\boldsymbol{t}}^M_{\boldsymbol{\lambda}})^\top \boldsymbol{\mathcal{I}}^M_{\boldsymbol{\lambda},\boldsymbol{\eta}} \hat{\boldsymbol{t}}^M_{\boldsymbol{\lambda}}\}$ for $M = M_0$ in Propositions 7 and 8. Let $\hat{c}^M_{1-q_n}$ be a consistent estimator of $c^M_{1-q_n}$. Then, our estimator based on sequential hypothesis testing (SHT, hereafter) is defined as

$$\begin{aligned}
\hat{M}_{\text{PLR}} &= \min_{M \in \{0,\ldots,\bar{M}\}} \{M : PLR_n(r) \geq \hat{c}^r_{1-q_n}, r = 0, \ldots, M-1, PLR_n(M) \ < \hat{c}^M_{1-q_n}\}, \\
\hat{M}_{\text{EM}} &= \min_{M \in \{0,\ldots,\bar{M}\}} \{M : EM_n(r) \geq \hat{c}^r_{1-q_n}, r = 0, \ldots, M-1, EM_n(M) \ < \hat{c}^M_{1-q_n}\}. \qquad (33)
\end{aligned}$$

The estimators $\hat{M}_{\text{PLR}}$ and $\hat{M}_{\text{EM}}$ depend on the choice of the significance level q_n. The following proposition states that $\hat{M}_{\text{PLR}}$ and $\hat{M}_{\text{EM}}$ converge to M_0 in probability as $n \to \infty$ if we choose q_n such that q_n such that $-n^{-1} \ln q_n = o(1)$ and $q_n = o(1)$.

Let $Q^M_n(\boldsymbol{\vartheta}_M) := n^{-1} \sum_{i=1}^n \ln f_M(\boldsymbol{w}_i; \boldsymbol{\vartheta}_M)$ and $Q^M(\boldsymbol{\vartheta}_M) := \mathbb{E}[\ln f_M(\boldsymbol{w}_i; \boldsymbol{\vartheta}_M)]$, where $f_M(\boldsymbol{w}_i; \boldsymbol{\vartheta}_M)$ is defined in (1) for $M = 1, ..., \bar{M}$.

Assumption 4. *For $M = 1, ..., M_0 - 1$, (a) $Q^M(\boldsymbol{\vartheta}_M)$ has a unique maximum at $\boldsymbol{\vartheta}^*_M$ in $\Theta_{\boldsymbol{\vartheta}_M}$; (b) $\Theta_{\boldsymbol{\vartheta}_M}$ is compact; (c) $\boldsymbol{\vartheta}^*_M$ is interior to $\Theta_{\boldsymbol{\vartheta}_M}$; (d) $B^M(\boldsymbol{\vartheta}^*_M) := \mathbb{E}\left\{\nabla_{\boldsymbol{\vartheta}_M} \ln f_M(\boldsymbol{w}_i; \boldsymbol{\vartheta}_M) \nabla_{\boldsymbol{\vartheta}^\top_M} \ln f_M(\boldsymbol{w}_i; \boldsymbol{\vartheta}_M)\right\}$ is nonsingular; (e) $A^M(\boldsymbol{\vartheta}^*_M) := \mathbb{E}\left\{\nabla_{\boldsymbol{\vartheta}_M \boldsymbol{\vartheta}^\top_M} \ln f_M(\boldsymbol{w}_i; \boldsymbol{\vartheta}_M)\right\}$ has constant rank in some open neighborhood of $\boldsymbol{\vartheta}^*_M$; (f) $Q^{M+1}(\boldsymbol{\vartheta}^*_{M+1}) - Q^M(\boldsymbol{\vartheta}^*_M) > 0$.*

Proposition 10. *Suppose that $M_0 < \bar{M}$ and Assumptions 1-4 hold. If we choose q_n such that $-n^{-1} \ln q_n = o(1)$ and $q_n = o(1)$, then $\hat{M}_{PLR} - M_0 = o_p(1)$ and $\hat{M}_{EM} - M_0 = o_p(1)$.*

Assumption 4 (a)-(e) ensure the consistency and asymptotic normality of $\hat{\boldsymbol{\vartheta}}_M$, where (c)-(e) correspond to Assumption A6 of White (1982). Per Assumption 4(f), the Kullback-Leibler Information Criterion of the model relative to the true M_0 components model strictly decreases as the number of components M increases, for $M < M_0$.

8 Simulation

In this section, we examine the finite sample performance of the EM test and PLRT by simulation. We test $H_0 : M = M_0$ against $H_1 : M = M_0 + 1$ for the model with $M_0 = 2$ and 3.

8.1 Choice of penalty function

We have developed a data-dependent empirical formula for a_n by selecting a formula that ensures empirical rejection probabilities match the nominal size (5%) across various null models and sam-

ple sizes, as reported in Table 11 in Appendix D. Specifically, for the model without conditioning variables, we have derived the following data-dependent empirical formula for testing the null hypotheses of $M_0 = 1, 2, 3, 4$:

$$a_n = \begin{cases} \left(1 + \exp\left\{\frac{\hat{\rho}_1^{M_0}}{\hat{\rho}_4^{M_0}} + \frac{\hat{\rho}_2^{M_0}}{\hat{\rho}_4^{M_0}}\frac{1}{T} + \frac{\hat{\rho}_3^{M_0}}{\hat{\rho}_4^{M_0}}\frac{1}{n}\right\}\right)^{-1}, & M_0 = 1 \\ \left(1 + \exp\left\{\frac{\hat{\rho}_1^{M_0}}{\hat{\rho}_4^{M_0}} + \frac{\hat{\rho}_2^{M_0}}{\hat{\rho}_4^{M_0}}\frac{1}{T} + \frac{\hat{\rho}_3^{M_0}}{\hat{\rho}_4^{M_0}}\frac{1}{n} + \frac{\hat{\rho}_5^{M_0}}{\hat{\rho}_4^{M_0}}\log\left(\frac{\omega(\boldsymbol{\vartheta}_{M_0}; M_0)}{1-\omega(\boldsymbol{\vartheta}_{M_0}; M_0)}\right)\right\}\right)^{-1}, & M_0 = 2, 3, 4, \end{cases} \tag{34}$$

where $\omega(\boldsymbol{\vartheta}_{M_0}; M_0)$ is the misclassification probability as defined in Melnykov and Maitra (2010) for each of the null models. The parameters $\hat{\rho}_1^{M_0}$, $\hat{\rho}_2^{M_0}$, $\hat{\rho}_3^{M_0}$, $\hat{\rho}_4^{M_0}$, and $\hat{\rho}_5^{M_0}$ are chosen as follows. Across different null models, sample sizes, and various candidate values of a_n, we estimate the empirical rejection probabilities at the 5% significance level by simulations and denote them by $\hat{s}$. For example, when testing $H_0 : M_0 = 2$, we repeatedly simulate the 500 datasets under each of the 48 null model parameters and sample sizes $(N, T, \alpha, \mu, \sigma) \in \{100, 500\} \times \{2, 5, 10\} \times \{(0.5, 0.5), (0.2, 0.8)\} \times \{(-1, 1), (-0.5, 0.5), (-0.5, 0.8)\} \times \{(1, 1), (1.5, 0.75), (0.8, 1.2)\}$ and test the null hypothesis of $H_0 : M_0 = 2$ by the EM test using one of the six values of $a_n \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4\}$. For each of $108 \times 6 = 648$ combinations of the parameter values, sample sizes, and a_n values, let $\hat{s}$ denote the fraction of simulated datasets that led to the rejection of the null hypothesis at 5% significance level. Using these 648 "observations" of $\{\hat{s}, N, T, \omega(\boldsymbol{\vartheta}_2; 2), a_n\}$, we run the following regression:

$$\begin{aligned} &\log\left(\frac{\hat{s}}{1-\hat{s}}\right) - \log\left(\frac{0.05}{1-0.05}\right) \\ &= \begin{cases} \rho_1^{M_0} + \rho_2^{M_0}\frac{1}{T} + \rho_3^{M_0}\frac{1}{n} + \rho_4^{M_0}\log\left(\frac{a_n}{1-a_n}\right), & M_0 = 1 \\ \rho_1^{M_0} + \rho_2^{M_0}\frac{1}{T} + \rho_3^{M_0}\frac{1}{n} + \rho_4^{M_0}\log\left(\frac{a_n}{1-a_n}\right) + \rho_5^{M_0}\log\left(\frac{\omega(\boldsymbol{\vartheta}_{M_0}; M_0)}{1-\omega(\boldsymbol{\vartheta}_{M_0}; M_0)}\right), & M_0 = 2, 3, 4, \end{cases} \end{aligned}$$

where $\hat{\rho}_1^{M_0}$, $\hat{\rho}_2^{M_0}$, $\hat{\rho}_3^{M_0}$, $\hat{\rho}_4^{M_0}$, and $\hat{\rho}_5^{M_0}$ in (34) denotes the corresponding estimates. Table 12 in the Appendix reports the estimates. Note that the data-dependent formula (34) is obtained by setting $\hat{s} = 0.05$ and solving for a_n in the above equation.

For the model with conditioning variables, we find that the value of a_n that gives accurate Type I errors is sensitive to the dimension of covariates, and developing a data-dependent empirical formula for a_n is difficult. Consequently, we choose a constant value of a_n that depends only on the number of components $M_0 = 1, 2, 3$, and 4 as follows: $a_n = 0.1617$ if $M_0 = 1$; $a_n = 0.0025$ if $M_0 = 2$; $a_n = 0.0567$ if $M_0 = 3$; $a_n = 0.4858$ if $M_0 = 4$; $a_n = 0.5$ if $M_0 \geq 5$. These penalty terms for the regression with covariates are chosen by averaging the prediction of the penalty function for the null parameters used in the simulations. For example, the penalty term for $M_0 = 2$ is chosen by generating a_n using the formula for all the combinations of $(N, T, \alpha, \mu, \sigma)$ in Table 11 for $M_0 = 2$

and taking the average across the predicted $\hat{a}_n$'s. For $M_0 \geq 5$, we use the parametric bootstrap method to obtain the critical values for our empirical application, where we set $a_n = 0.5$.

8.2 Simulation results

Table 1 displays the simulated Type I error rates for the EM test when examining the null hypothesis $H_0 : M = 2$ against the alternative hypothesis $H_1 : M = 3$. A total of 2000 repetitions were employed for the asymptotic distribution, while 1000 repetitions were used for the bootstrap distribution. Moreover, the PLRT with simulated critical values was considered.

The table presents results for four distinct null models, as explained in the table's footnote. Utilizing the asymptotic distribution, the EM test sizes generally approximate the nominal 5% level. Nonetheless, the test may be undersized in instances where $T \geq 5$. Furthermore, the test size is larger when the mixing proportions are equal ($\boldsymbol{\alpha} = (0.5, 0.5)$) in comparison to when they are unequal ($\boldsymbol{\alpha} = (0.2, 0.8)$). The bootstrapped EM test demonstrates satisfactory performance.

For the PLRT, 2000 repetitions were conducted, and results were reported for cases where a constraint was applied to $\alpha_j \in [\epsilon, 1 - \epsilon]$ with $\epsilon = 0.1$. The value of a_n for the PLRT was chosen to be ten times larger than its value for the EM test. The findings suggest that the PLRT is slightly oversized.

Table 2 reports the rejection frequency of testing $H_0 : M_0 = 2$ under 12 alternative three-component mixture models, as elaborated in the table's footnote. For both EM test and PLRT, the test power is greater when distances between μ_j's are larger and equal, such as $(\mu_1, \mu_2, \mu_3) = (-1, 0, 1)$ or $(-1.5, 0, 1.5)$, as opposed to unbalanced distances like $(-1, 0, 2)$ or $(-0.5, 0, 1.5)$. The power is also improved when the mixture probabilities are equal ($\boldsymbol{\alpha} = (1/3, 1/3, 1/3)$) rather than unequal ($\boldsymbol{\alpha} = (1/4, 1/2, 1/4)$). The power increases with both the time-dimension T and cross-sectional sample size N. Reflecting a larger actual rejection frequency of the PLRT under $H_0 : M_0 = 2$ in Table 1, the power of the PLRT is often higher than that of the EM test, although the EM test sometimes has higher power, especially when the mixing probabilities are unequal.

Table 3 displays the simulated Type I error rates of the EM test using the asymptotic distribution for testing $H_0 : M_0 = 3$ against $H_1 : M_0 = 4$. Six null models are considered with varying $(\alpha_1, \alpha_2, \alpha_3)$ and (μ_1, μ_2, μ_3) values. The EM test generally yields accurate Type I errors.

The Type I error rates of the EM test with conditioning variables under the null $M_0 = 2$ are examined using 500 repetitions. Results presented in Table 4 indicate a slightly oversized test for small samples with $(N, T) = (200, 2)$, but overall, the finite sample properties are satisfactory.

In our empirical application examining production function heterogeneity in Japan and Chile, we find evidence that the number of components is frequently greater than five when we sequentially apply our EM test to estimate the number of components. Consequently, we also investigate the performance of the sequential hypothesis testing (SHT) using the EM test in comparison to the

AIC and the BIC when the data is generated from a five-component model in a realistic setting. Specifically, we simulate 100 datasets from the estimated five-component model of the Chilean textile industry in our empirical application and apply these three methods to select the number of components in each of the 100 datasets. Here, we apply the EM test at the 5 percent significance level to sequentially test the null hypothesis $H_0 : M = M_0$ for $M_0 = 1, 2, ..., 7$, and we determine the number of components to be M_0 when we fail to reject $H_0 : M = M_0$ as in (33).

Table 5 presents the frequency at which the three methods select the number of components in this simulation. The table demonstrates that the proposed sequential hypothesis test selects the correct number of components 72 % of the time, while it underestimates the true number of components 25 % of the time. Conversely, the AIC overestimates the number of components 86 % of the time, and the BIC underestimates the number of components by selecting a four-component model 41 % of the time, accurately estimating the number of components 58% of the time. Overall, in this simulation, our proposed sequential hypothesis testing approach outperforms both the AIC and the BIC.

Table 1: Sizes (in %) of EM test and PLRT of $H_0 : M_0 = 2$ against $H_A : M_0 = 3$ at the 5% level

	EM Test						**EM Test**						**PLRT**					
	Asymptotic						Parametric Bootstrap						Asymptotic					
T	3		5		8		3		5		8		3		5		8	
N	200	400	200	400	200	400	200	400	200	400	200	400	200	400	200	400	200	400
(A, C)	5.3	4.8	4	3.8	4	2.95	4.6	6.2	6.2	4.6	4.8	5.2	7.7	6.6	6.75	6.7	6.5	5.5
(A, D)	5.9	4.9	5	5	4.45	4	5.4	4.8	5.2	5.6	5	5.8	5.4	5.15	5.1	6.1	5.55	6.2
(B, C)	3.8	2.5	3.45	3.05	3.6	3.25	3.6	5.6	4.2	5.2	4	5.4	6.25	5.45	6.45	6.05	5.05	6
(B, D)	4.8	4.6	3.5	3.15	3.55	3.95	3.6	3.6	5.8	4	6.2	4.6	2.35	4.4	3.9	4.85	4.95	5.2

[1] A and B refer to respectively $(\alpha_1, \alpha_2) = (0.5, 0.5)$ and $(0.2, 0.8)$, while C and D refer to $(\mu_1, \mu_2) = (-1, 1)$ and $(-0.5, 0.5)$, respectively.

[2] The variance is set to $(\sigma_1, \sigma_2) = (0.8, 1.2)$. The asymptotic simulations are based on 2000 repetitions and the bootstrap simulation is based on 1000 repetitions.

Table 2: Powers (in %) of EM test and PLRT of $H_0 : M_0 = 2$ against $H_A : M_0 = 3$ at the 5% level

	A				B			
N	100		500		100		500	
T	2	5	2	5	2	5	2	5
	EM test							
(C, G)	20.9	81.6	57.6	100.0	20.5	82.7	62.6	100.0
(C, H)	49.2	99.9	99.9	100.0	38.4	98.7	98.8	100.0
(C, I)	12.1	20.4	18.0	62.6	10.6	20.4	16.8	65.8
(D, G)	77.9	100.0	100.0	100.0	86.5	100.0	100.0	100.0
(D, H)	57.4	100.0	100.0	100.0	42.8	100.0	100.0	100.0
(D, I)	16.0	59.5	31.8	99.9	13.8	70.8	40.3	100.0
(E, G)	93.0	100.0	100.0	100.0	94.0	100.0	100.0	100.0
(E, H)	83.8	100.0	100.0	100.0	70.7	100.0	100.0	100.0
(E, I)	25.7	97.0	80.2	100.0	30.7	96.8	83.1	100.0
(F, G)	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0
(F, H)	93.5	100.0	100.0	100.0	85.3	100.0	100.0	100.0
(F, I)	40.8	99.9	98.2	100.0	52.1	100.0	99.5	100.0
	PLRT							
(C, G)	22.7	85.1	56.5	100.0	23.2	82.6	58.7	100.0
(C, H)	57.1	100.0	99.8	100.0	43.2	99.7	99.1	100.0
(C, I)	12.0	21.0	12.4	66.1	11.3	22.1	12.4	69.3
(D, G)	79.9	100.0	100.0	100.0	87.6	100.0	100.0	100.0
(D, H)	65.3	100.0	100.0	100.0	49.1	100.0	100.0	100.0
(D, I)	14.8	63.6	28.6	100.0	13.6	75.2	36.7	100.0
(E, G)	91.5	100.0	100.0	100.0	93.7	100.0	100.0	100.0
(E, H)	86.8	100.0	100.0	100.0	75.9	100.0	100.0	100.0
(E, I)	28.7	97.2	77.7	100.0	33.0	97.9	85.5	100.0
(F, G)	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0
(F, H)	96.4	100.0	100.0	100.0	89.5	100.0	100.0	100.0
(F, I)	45.7	100.0	98.4	100.0	57.2	100.0	99.8	100.0

Notes: A and B refer to $(\alpha_1, \alpha_2, \alpha_3) = (1/3, 1/3, 1/3)$ and $(1/4, 1/2, 1/4)$, respectively; C, D, E,and F refer to $(\mu_1, \mu_2, \mu_3) = (-0.5, 0, 1.5), (-1, 0, 1), (-1, 0, 2), (-1.5, 0, 1.5)$, respectively; G, H, I refer to $(\sigma_1, \sigma_2, \sigma_3) = (0.6, 0.6, 1.2), (0.6, 1.2, 0.6), (1, 1, 1)$.

Table 3: Sizes (in %) of **EM test** of $H_0 : M_0 = 3$ against $H_A : M_0 = 4$ at 5% level

	(A,C)	(A,D)	(A,E)	(B,C)	(B,D)	(B,E)
100,2	5.95	5.15	5.05	5.05	5.85	4.40
500,2	5.60	5.55	5.25	5.10	5.65	4.05
100,5	4.30	6.00	4.20	5.15	5.10	5.70
500,5	4.20	4.55	3.95	4.50	4.15	4.15

Notes: A and B refer to $(\alpha_1, \alpha_2, \alpha_3) = (1/3, 1/3, 1/3)$ and $(0.25, 0.5, 0.25)$, respectively, while C, D, E refer to $(\mu_1, \mu_2, \mu_3) = (-4, 0, 4)$, $(-4, 0, 6)$ and $(-6, 0, 6)$, respectively. The variance is set to $(\sigma_1, \sigma_2, \sigma_3) = (0.75, 1.5, 0.75)$. The asymptotic simulations are based on 2000 repetitions and the bootstrap simulation is based on 1000 repetitions.

Table 4: Sizes of **EM test** of $H_0 : M_0 = 2$ against $H_A : M_0 = 3$ with conditioning variables

(N, T)	(A, C, E)	(A, C, F)	(A, D, E)	(A, D, F)	(B, C, E)	(B, C, F)	(B, D, E)	(B, D, F)
$(200, 2)$	8.4	8.2	7.4	8.8	8.6	8.2	7.4	3.6
$(500, 2)$	4.6	3.2	3.2	2.2	4.8	4.8	3.6	3.6
$(200, 5)$	4.0	1.8	3.0	2.6	2.2	2.0	2.2	3.2
$(500, 5)$	2.2	1.2	1.6	1.4	3.0	2.0	1.8	2.0

Notes: A and B refer to $(\mu_1, \mu_2) = (-1, 1)$ and $(-0.5, 0.5)$, respectively, while C and D refer to $(\beta_1, \beta_2) = (1, 1)$ and $(-1, 1)$, respectively. E and F refer to $(\sigma_1, \sigma_2) = (0.3, 0.1)$ and $(0.1, 0.1)$. The mixing proportion is set to $(\alpha_1, \alpha_2) = (0.2, 0.8)$. The asymptotic simulations are based on 500 repetitions.

Table 5: Frequency of Number of Components with the Simulated Data

M	1	2	3	4	5	6	7
SHT with EM test	0	0	0	0.26	0.72	0.02	0
AIC	0	0	0	0.01	0.13	0.31	0.55
BIC	0	0	0	0.41	0.58	0.01	0

[1] The data are generated using the estimated parameters based on the Chilean textile industry with five-components and panel length $T = 3$, where $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.16076522, 0.32454077, 0.09025875, 0.35478905, 0.06964622)$, $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5) = (-1.241241, -0.33803875, 0.4480291, 0.52379553, 1.4139465)$, $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (0.451833, -0.05988709, -0.2453261, -0.03106076, 0.2053708)$, $(\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) = (0.9933480, 0.4585760, 0.9954302, 0.4116855, 0.1863346)$. We use the panel length and sample size that are equal to those in the dataset, i.e., $n = 196$ and $T = 3$.

[2] The results are based on 100 repetitions.

[3] Each cell indicates the proportion of times that the model selection indicates a M-component model.

9 Empirical Application

In this section, we conduct an empirical application of our proposed test for the number of components in a finite mixture production function model, the identification of which is analyzed in Kasahara et al. (2022). Specifically, we estimate the number of types of input elasticities in production functions using panel data from Japanese publicly traded firms in the machinery industry, as well as data from Chilean manufacturing firms.

9.1 Production Function and First Order Condition

Consider the input and output panel data of n firms over T years, $\{\{Y_{it}, V_{it}, L_{it}, K_{it}\}_{t=1}^{T}\}_{i=1}^{N}$, where Y_{it}, V_{it} L_{it}, K_{it} represent the output, intermediate input, labor, and capital of firm i in year t, respectively. We denote the logarithm of corresponding variables by lowercase letters as $(y_{it}, v_{it}, l_{it}, k_{it})$, with, for example, $y_{it} = \log(Y_{it})$.

We employ a finite mixture specification to capture unobserved heterogeneity in a firm's input elasticities. We are interested in testing the number of production technology types. Assume there are M discrete types of production technologies and define the latent random variable $D_i \in \{1, 2, \ldots, M\}$ to represent the production technology type of firm i. If $D_i = j$, then firm i is of type j. The population proportion of type j is denoted by $\alpha_j = \Pr(D = j)$. The production function for type j is Cobb-Douglas and the output is related to inputs as

$$Y_{it} = \exp(\epsilon_{it}) F_t^j(V_{it}, L_{it}, K_{it}, \omega_{it}) \tag{35}$$

with

$$F_t^j(V_{it}, L_{it}, K_{it}, \omega_{it}) := \exp(\gamma_t^j + \omega_{it}) V_{it}^{\delta_{v,j}} L_{it}^{\delta_{\ell,j}} K_{it}^{\delta_{k,j}},$$

where γ_t^j represents the aggregate productivity shock of type j in year t; ω_{it} is the serially correlated productivity shock; and ϵ_{it} is the idiosyncratic productivity shock.

We assume that an intermediate input V_{it} is flexibly chosen by firm i after observing aggregate shock γ_t^j and serially correlated productivity shock ω_{it}. The variable ϵ_{it} represents a mean-zero i.i.d. random variable, the realization of which is unknown when the intermediate input V is selected. Denote the information available to a firm for making decisions on V_{it} by $\mathcal{I}_{it}$.Denote the information available to a firm for making decisions on V_{it} by $\mathcal{I}_{it}$.

In order to identify the intermediate input elasticity of the production function, we introduce the following assumptions (c.f., Kasahara et al. (2022)).

Assumption 5. *(a) Each firm belongs to one of M types, and the probability of being type j is given by $\alpha_j = P(D_i = j)$ with $\sum_{j=1}^M \alpha_j = 1$. (b) For the j^{th} type of production technology at time t, the output is expressed in terms of input as in (35), where $\epsilon_{it} \sim N(0, \sigma_j^2)$ are i.i.d across i's and t's. ω_{it} follows an exogenous first-order stationary Markov process given by $\omega_{it} = h^j(\omega_{it-1}) + \eta_{it}$ where, conditional on $\mathcal{I}_{it-1}$, η_{it} is a mean-zero i.i.d. random variable. (c) $(\gamma_t^j, \omega_{it}) \in \mathcal{I}_{it}$ and $\epsilon_{it} \notin \mathcal{I}_{it}$.*

Assumption 6. *(a) Firms are price-takers in both output and input markets, where $P_{Y,t}$ and $P_{V,t}$ are the prices of output and intermediate input in year t. (b) $(P_{Y,t}, P_{V,t})$ are observed by firms at the beginning of the period before V_{it} is chosen.*

Assumption 7. *V_{it}'s are chosen at time t by maximizing the expected profit conditional on information $\mathcal{I}_{it}$ at time t and conditional on the value of (K_{it}, L_{it}). The profit maximization problem for firms with type j technology is given by*

$$V_{it} = \arg\max_V P_{Y,t} \mathbb{E}[\exp(\epsilon_{it}) | D_i = j] F_t^j(V, K_{it}, L_{it}, \omega_{it}) - P_{V,t} V. \tag{36}$$

In Assumption 5(a), each firm's production function belongs to one of the M types. Assumption 5(b) assumes that the idiosyncratic productivity shock follows a normal distribution. Assumption 5(c) assumes that both the aggregate shock γ_t^j and the serially correlated productivity shock ω_{it} are observed when intermediate inputs are chosen, but idiosyncratic productivity shocks are unknown. Assumption 6 states that firms observe input and output prices when deciding on V_{it}. Assumption 7 assumes that V_{it} is chosen to maximize the current expected period profit conditional on the value of (K_{it}, L_{it}).[4]

[4]We are agnostic about the timing of choosing K_{it} and L_{it} as long as they are either determined before V_{it} or simultaneously chosen with V_{it}. It is reasonable to assume that capital input K_{it} is determined before the value of V_{it} is

Given the above assumptions 5, 6, and 7, we derive an empirical specification based on the first-order condition of the profit maximization problem (36), following the idea developed by Gandhi et al. (2020) and extending to a finite mixture production function modeled by Kasahara et al. (2022). Note that $E[\exp(\epsilon_{it})|D_i = j] = \exp(\sigma_j^2/2)$ for $\epsilon_{it} \sim N(0, \sigma_j^2)$. Then, because $\delta_{v,j} = \frac{\partial F_i^j(V_{it},K_{it},L_{it})/\partial V_{it}}{F_i^j(V_{it},K_{it},L_{it})/V_{it}}$ for the Cobb-Douglas production function, the first-order condition with respect to V_{it} in (36) together with the production function (35) implies that:

$$s_{it} = \log \delta_{v,j} + \frac{1}{2}\sigma_j^2 - \epsilon_{it} \quad \text{for } D_i = j, \tag{37}$$

where

$$s_{it} := \log\left(\frac{P_{V,t}V_{it}}{P_{Y,t}Y_{it}}\right)$$

is the logarithm of the ratio of intermediate input cost to revenue.

Collect the observed data as $\boldsymbol{W}_i = \{s_{it}, \log K_{it}\}_{t=1}^T$. Let $\mu_j = \log \delta_{v,j} + \frac{1}{2}\sigma_j^2$ and define a type-specific parameter to be $\boldsymbol{\theta}_j = (\mu_j, \sigma_j)$, where $\delta_{v,j}$ can be identified from $\boldsymbol{\theta}_j$ as $\delta_{v,j} = \exp(\mu_j - \sigma_j^2/2)$. Collect the parameters of each type and the mixing probability as $\boldsymbol{\vartheta}_M = (\alpha_1, \ldots, \alpha_{M-1}, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_M^\top)^\top$. Recall that $\epsilon_{it} \overset{iid}{\sim} N(0, \sigma_j^2)$ over i and t conditional on the technology type $D_i = j$. Then, from (37), we can write the density function of $s_{i1}, ..., s_{iT}$ as a mixture of type-specific likelihood density similar to the density function in equation (1):

$$f_M(\boldsymbol{W}_i; \boldsymbol{\vartheta}_M) = \sum_{j=1}^M \alpha_j \prod_{t=1}^T \frac{1}{\sigma_j}\phi\left(\frac{s_{it} - \mu_j}{\sigma_j}\right). \tag{38}$$

The penalized maximum likelihood estimator is defined as

$$\hat{\boldsymbol{\vartheta}}_M = \arg\max_{\boldsymbol{\vartheta}_M} \sum_{i=1}^n \log f_M(\boldsymbol{W}_i; \boldsymbol{\vartheta}_M) + \tilde{p}_n(\boldsymbol{\vartheta}_M).$$

As an alternative specification, we allow the elasticity of output for intermediate input to be a function of $\log K_{it}$ as $\log \delta_{v,j} = \beta_{0,j} + \beta_{k,j} \log K_{it}$. This results in the logarithm of the ratio of intermediate input cost to revenue being linearly related to $\log K_{it}$ as $s_{it} = \mu_j + \beta_{k,j} \log K_{it} - \epsilon_{it}$ for $D_i = j$ with $\mu_j = \beta_{0,j} + \frac{1}{2}\sigma_j^2$. In this case, the conditional density function of $\{s_{it}\}_{t=1}^T$ given $\{\log K_{it}\}_{t=1}^T$ is

$$f_M(\boldsymbol{W}_i; \boldsymbol{\vartheta}_M) = \sum_{j=1}^M \alpha_j \prod_{t=1}^T \frac{1}{\sigma_j}\phi\left(\frac{s_{it} - \mu_j - \beta_{k,j}\log K_{it}}{\sigma_j}\right). \tag{39}$$

chosen. On the other hand, labor input L_{it} may be flexibly chosen simultaneously with V_{it} after γ_t^j and ω_{it} are observed. Even when labor input is simultaneously chosen with intermediate input, equation (36) and the corresponding first-order condition characterize the intermediate input choice once we interpret L_{it} in (36) as the optimal value chosen by firm i as discussed in Ackerberg et al. (2015).

In addition, we consider a specification in which we include not only $\log K_{it}$ but also $\log L_{it}$ as a regressor:

$$f_M(\boldsymbol{W}_i; \boldsymbol{\vartheta}_M) = \sum_{j=1}^{M} \alpha_j \prod_{t=1}^{T} \frac{1}{\sigma_j} \phi \left(\frac{s_{it} - \mu_j - \beta_{k,j} \log K_{it} - \beta_{\ell,j} \log L_{it}}{\sigma_j} \right). \tag{40}$$

9.2 Empirical result

We apply the EM test to two producer-level data sets to determine the number of production technology types. We used the production data from the Japanese publicly traded firms from 2003 to 2007 and the Chilean manufacturing plants from 1992 to 1996.[5] We cleaned the data and used the firms/plants with continuous data entry for five years to ensure that we had balanced panel data. We focus on the three largest industries in terms of the number of firms and plants for each country (chemical, machine, and electronics for Japan and food products, fabricated metal products, and textiles for Chile). Table 6 presents the summary statistics for the revenue share of intermediate materials and the log of gross output in these industries. The within-industry standard deviations of the revenue share of intermediate materials are substantial across all industries, suggesting that intermediate input elasticities differ across firms within the narrowly defined industries.

[5] Please refer to Kasahara et al. (2021) and Kasahara and Rodrigue (2008) for the details of the datasets of the Japanese publicly traded firms and the Chilean manufacturing plants, respectively.

Table 6: Descriptive statistics for the revenue share of intermediate material and the log of gross output for the Japanese firms and the Chilean plants

Panel A: : Japanese publicly traded firms						
			$\frac{P_{V,t}V_{it}}{P_{Y,t}Y_{it}}$		$\log(Y_{it})$	
Industry	NObs	n	$mean$	sd	$mean$	sd
Chemical	805	161	0.34	0.15	17.52	1.24
Machine	790	158	0.50	0.16	17.31	1.35
Electronics	775	155	0.45	0.18	17.54	1.27
Panel B: Chilean plants						
			$\frac{P_{V,t}V_{it}}{P_{Y,t}Y_{it}}$		$\log(Y_{it})$	
Industry	NObs	n	$mean$	sd	$mean$	sd
Food products	4645	929	0.65	0.15	10.62	1.66
Fabricated metal products	1260	252	0.53	0.18	11.00	1.37
Textiles	1130	226	0.58	0.19	11.01	1.32

[1] The summary statistics are based on the Japanese firm-level data from 2003 to 2007 and the Chilean plant-level data from 1992 - 1996. All observations with $\log(V_{it}/Y_{it}) \leq -3$ and $\log(V_{it}/Y_{it}) > \log(2)$ are removed. The data set is a balanced panel, i.e., we kept firms/plants that are continuously observed for these five years.

[2] The variable $\frac{P_{V,t}V_{it}}{P_{Y,t}Y_{it}}$ is defined as the revenue share of the intermediate input, where $P_{V,t}$ is the average price of the intermediate input at time t, $P_{Y,t}$ is the average price of the output, V_{it} is the quantity of the intermediate input and Y_{it} is the quantity of the output.

Table 7: The EM test for Japanese producer without conditioning variables

		M=1	M=2	M=3	M=4	M=5
				$T = 3$		
Chemical	*EM*	436.37^{***}	239.83^{***}	130.1^{***}	126.4^{***}	63.24^{***}
	BIC	805.55	383.43	157.5	41.62	-70.46
Electronics	*EM*	563.94^{***}	186.67^{***}	115.82^{***}	81.06^{***}	47.76^{***}
	BIC	814.01	264.27	91.67	-10.39	-77.2
Machine	*EM*	434.91^{***}	194.48^{***}	72.83^{***}	56.94^{***}	54.77^{***}
	BIC	458.72	37.85	-142.28	-200.74	-242.71
				$T = 4$		
Chemical	*EM*	629.22^{***}	308.6^{***}	181.39^{***}	177.38^{***}	96.35^{***}
	BIC	1071.45	456.54	162.15	-4.99	-168.01
Electronics	*EM*	803.15^{***}	282.32^{***}	167.83^{***}	106.43^{***}	89.93^{***}
	BIC	1081.48	292.68	24.54	-484.46	
Machine	*EM*	620.95^{***}	292.52^{***}	118.37^{***}	102.57^{***}	75.32^{***}
	BIC	609.1	2.14	-276.04	-380.16	-467.96
				$T = 5$		
Chemical	*EM*	818.38^{***}	386.08^{***}	219.13^{***}	209.42^{***}	118.25^{***}
	BIC	1331.53	527.48	155.86	-48.53	-243.73
Electronics	*EM*	1024.86^{***}	375.29^{***}	226.01^{***}	134.53^{***}	126.36^{***}
	BIC	1343.12	332.61	-28.32	-239.31	-359.17
Machine	*EM*	819.98^{***}	389.69^{***}	156.44^{***}	149.98^{***}	96.32^{***}
	BIC	775.75	-30.17	-406.59	-548.81	-683.96

[1] The estimation is based on the revenue share of intermediate material. [2] *, **, *** indicate the result is significant at 10%, 5% and 1% levels respectively.

Table 8: The EM test for Chilean producer without conditioning variables

		M=1	M=2	M=3	M=4	M=5
				$T = 3$		
Food products	*EM*	805.51^{***}	637.77^{***}	204.92^{***}	80.54^{***}	72.41^{***}
	BIC	422.55	-371.13	-991.96	-1176.61	-1236.82
Fabricated metal products	*EM*	238.84^{***}	68.91^{***}	26.24^{***}	24.42^{***}	21.82^{***}
	BIC	719.74	496.49	444.02	433.01	425
Textiles	*EM*	229.87^{***}	146.17^{***}	64.76^{***}	27.06^{***}	29.98^{**}
	BIC	635.37	418.28	288.34	236.9	223.34
				$T = 4$		
Food products	*EM*	1165.08^{***}	874.27^{***}	257.49^{***}	130.61^{***}	139.59^{***}
	BIC	419.47	-730.83	-1586.11	-1825.87	-1938.03
Fabricated metal products	*EM*	362.1^{***}	120.7^{***}	41.6^{***}	43.68^{***}	20.95^{***}
	BIC	905.9	559.3	453.41	427.34	399.82
Textiles	*EM*	325.17^{***}	222.28^{***}	74.19^{***}	47.58^{***}	51.65^{***}
	BIC	821.73	510.98	303.8	243.51	210.77
				$T = 5$		
Food products	*EM*	1553.9^{***}	1010.31^{***}	290.02^{***}	172.46^{***}	155.25^{***}
	BIC	471.66	-1066.71	-2057.71	-2329.38	-2484.82
Fabricated metal products	*EM*	478.94^{***}	176.5^{***}	58.96^{***}	59.37^{***}	33.19^{***}
	BIC	1101.11	637.21	477.1	433.62	389.54
Textiles	*EM*	428.29^{***}	280.46^{***}	103.41^{***}	56.63^{***}	53.57^{***}
	BIC	968.16	556.01	289.55	201.41	160

[1] The estimation is based on the revenue share of intermediate material. [2] *, **, *** indicate the result is significant at 10%, 5% and 1% levels respectively.

To determine the number of components, we test the null hypothesis $H_0 : M = M_0$ against $H_1 : M = M_0 + 1$ by applying the EM test at the 5 percent significance level sequentially for $M_0 = 1, \ldots, 5$. If we fail to reject the null hypothesis at a certain $M_0 = M$, then we conclude that there are M types of intermediate input elasticities. We consider both the models without conditioning variable (38) and the models with conditioning variable (39)-(40).

Table 7 and 8 report the result of the EM test for the model without conditioning variable (38) from the Japanese and the Chilean industries, respectively, with the panel length of $T = 3, 4, 5$ and the null model of $M = 1, ..., 5$. For all industries in both countries and all panel lengths, we reject the null hypothesis of $H_0 : M = M_0$ for all $M_0 = 1, 2, 3, 4$, and 5 at five percent signifi-

cance level, indicating that the number of types for intermediate input elasticities is at least five types. This result reflects a considerable and persistent heterogeneity in the revenue share of intermediate materials across firms or plants, providing strong evidence for substantial heterogeneity in intermediate input elasticities across firms' production functions in Japanese and Chilean producers. Our findings serve as a caution against the conventional empirical practice of estimating the Cobb-Douglas production function, which assumes that the elasticity parameters are common across firms. Given the strong evidence of heterogeneity in production function coefficients, incorporating heterogeneity in production function coefficients in empirical applications is warranted and should be encouraged.

On the other hand, one possible reason for the estimated number of technology types being greater than five is that the assumption of the Cobb-Douglas production function may be too restrictive. When the production function is not Cobb-Douglas, the revenue share of intermediate materials generally depends on the value of production inputs (Gandhi et al., 2020). For this reason, we test the number of technology types when the revenue share of intermediate materials depends on the value of capital input as well as labor input by estimating the models (39)-(40).

Table 9 presents the results of the sequential hypothesis test and the BIC when estimating the mixture regression model with $\log K_{it}$ in (39) using data with a panel length of $T = 3$. For the Japanese Chemical, Electronics, and Machinery industries, the sequential hypothesis test suggests that the data is generated from seven to nine-component models; concurrently, the BIC selects models with at least ten components. For the Chilean Food industry, the sequential test indicates a ten-component model, while the BIC chooses an eight-component model. In contrast, the sequential hypothesis test and the BIC, respectively, select models with seven and six components for the Chilean Fabricated Metal Products industry and the Chilean Textile industry.

Table 10 reports the results for the model that includes both $\log K_{it}$ and $\log L_{it}$ as regressors. Across six industries, both the results of the sequential hypothesis test and the BIC in Table 10 select models with at least five components, providing evidence for substantial heterogeneity in production technology across firms and plants. Comparing the results of Table 10 with those of Table 9, the selected number of components for the model with $\log K_{it}$ and $\log L_{it}$ is smaller than that for the model with only $\log K$. This suggests that the number of components may be overestimated if we do not consider a sufficiently flexible production function specification by excluding some regressors.

Table 9: The EM test and the BIC (**Dependent Variable**: $\log \frac{P_{V,t}V_{it}}{P_{Y,t}Y_{it}}$, **Regressor**: $\log K_{it}$)

M_0	1	2	3	4	5	6	7	8	9	10
Japanese Chemical										
EM	459.4^{***}	236.36^{***}	125.42^{***}	118.36^{***}	87.63^{***}	53.72^{***}	38.69^{***}	34.07^{**}	36.47	-
BIC	1384.76	943.61	726.53	620.32	518.86	449.92	413.49	394.46	381.48	366.09
Japanese Electronics										
EM	560.06^{***}	213.82^{***}	116.29^{***}	78.81^{***}	47.05^{***}	40.77^{***}	27.4^{**}	29.02	-	-
BIC	1332.14	788.19	593.44	495.74	434.15	406.77	385.45	372.63	367.31	351.17
Japanese Machine										
EM	433.19^{***}	202.92^{***}	80.42^{***}	76.82^{***}	53.83^{***}	34.62^{**}	55.65	-	-	-
BIC	1355.6	940.49	757	696.06	638.48	617.4	588.94	568.71	555.15	544.51
Chilean Food Products										
EM	816.06^{***}	489.37^{***}	169.14^{***}	80.88^{***}	80.63^{***}	52.67^{***}	31.29^{***}	17.16^{**}	20.55^{***}	-60.46
BIC	6759.39	5962.74	5499.3	5356.47	5301.31	5241.91	5210.27	5200.71	5210.77	5222.29
Chilean Fabricated Metal Products										
EM	199.35^{***}	63.25^{***}	49.24^{***}	30.27^{***}	15.73^{**}	18.25^{**}	10.88	-	-	-
BIC	1923.64	1744.72	1699.97	1670.93	1661.03	1659.54	1665.08	1669.02	1680.96	1695.54
Chilean Textile										
EM	201.86^{***}	95.17^{***}	61.43^{***}	31.17^{***}	14.12^{*}	17.45^{**}	7.94	-	-	-
BIC	1681.91	1499.99	1424.93	1380.93	1368.65	1364.94	1365.72	1370.72	1382.83	1392.24

[1] The estimation is based on the revenue share of intermediate material using the panel data of length $T = 3$.

[2] *, **, *** indicate the result is significant at 10%, 5% and 1% levels respectively.

Table 10: The EM test and the BIC (**Dependent Variable**: $\log \frac{P_{V,t}V_{it}}{P_{Y,t}Y_{it}}$, **Regressors**: $\log K_{it}$ and $\log L_{it}$)

M_0	1	2	3	4	5	6	7	8	9	10
Japanese Chemical										
EM	412.35^{***}	224.09^{***}	141.59^{***}	132.24^{**}	121.56	-	-	-	-	-
BIC	1294.05	905.44	705.72	587.3	490.07	479.74	389.05	390.29	382.28	372.69
Japanese Electronics										
EM	573.11^{***}	218.38^{***}	116.07^{***}	94.76^{**}	47.73	-	-	-	-	-
BIC	1336.69	784.95	590.73	498.55	426.23	389.91	372.05	371.64	359.15	368.25
Japanese Machine										
EM	468.06^{***}	204.01^{***}	93.35^{***}	81.62^{***}	62.00^{***}	37.04^{***}	14.21	-	-	-
BIC	1360.56	915.69	736.2	676.26	625.66	596.45	564.34	548.7	536.78	539.64
Chilean Food Products										
EM	805.09^{***}	478.64^{***}	177.08^{***}	84.13^{***}	80.96^{***}	51.97^{***}	32.3^{**}	19.50	-	-
BIC	6732.11	5952.7	5506.55	5362.27	5309.37	5257.78	5233.37	5229.9	5242.9	5258.41
Chilean Fabricated Metal Products										
EM	204.45^{***}	63.57^{***}	49.42^{***}	28.61^{***}	18.32	-	-	-	-	-
BIC	1926.06	1747.29	1709.44	1685.39	1678.71	1680.54	1685.02	1696.19	1703.21	1723.56
Chilean Textile										
EM	203.69^{***}	90.69^{***}	58.4^{***}	32.55^{***}	16.19	-	-	-	-	-
BIC	1673.99	1495.55	1431.18	1394.54	1382.59	1373.03	1368.8	1382.42	1394.3	1394.09

[1] The estimation is based on the revenue share of intermediate material using the panel data of length $T = 3$.

[2] *, **, *** indicate the result is significant at 10%, 5% and 1% levels respectively.

10 Conclusion

The selection of the number of components in a finite normal mixture panel regression model is a crucial practical issue that must be addressed with care. Arbitrarily choosing the number of components can result in biased estimates, invalid inference, and reduced credibility of the final outcomes. To tackle this issue, this study proposes the PLRT and an EM test and derives their asymptotic distribution for the null hypothesis of a model with M_0 components against the alternative hypothesis with $(M_0 + 1)$ components. We also develop a procedure to consistently select the number of components by sequentially applying the PLRT and EM tests. Through a simulation exercise, we demonstrate that the proposed sequential hypothesis testing procedure exhibits good performance in finite samples.

As an empirical application, we estimate the number of production technology types using producer-level panel data from Japan and Chile. We find that most industries in our dataset exhibit a level of heterogeneity that requires a five or more-component mixture model when using the Cobb-Douglas production specification or a specification in which the elasticity of inputs depends on capital and labor input linearly. This suggests strong evidence for the presence of unobserved heterogeneity in technology types. One important caveat of our empirical exercise is that the class of production functions we investigate may be restrictive. Investigating production function heterogeneity with more flexible function forms is an important future research topic.

References

[1] Ackerberg, D. A., Caves, K., and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451.

[2] 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.

[3] Amengual, D., Bei, X., Carrasco, M., and Sentana, E. (2022). Score-type tests for normal mixtures. Working papers, CEMFI.

[4] Ando, T. and Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1):163–191.

[5] Andrews, D. (1999). Estimation When a Parameter is on a Boundary. *Econometrica*, 67(6):1341–1383.

[6] Andrews, R. L. and Currim, I. S. (2003). Retention of latent segments in regression-based marketing models. *International Journal of Research in Marketing*, 20(4):315–321.

[7] Azaïs, J.-M., Gassiat, E., and Mercadier, C. (2009). The likelihood ratio test for general mixture models with or without structural parameter. *ESAIM: Probability and Statistics*, 13:301—327.

[8] Balat, J., Brambilla, I., and Sasaki, Y. (2019). Heterogeneous firms: Skilled- labor productivity and the destination of exports.

[9] Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

[10] Cameron, S. V. and Heckman, J. J. (1998). Life cycle schooling and dynamic selection bias: models and evidence for five cohorts of American males. *Journal of Political Economy*, 106(2):262–333.

[11] Chen, H. and Chen, J. (2001). The likelihood ratio test for homogeneity in finite mixture models. *Canadian Journal of Statistics*, 29:201–215.

[12] Chen, H. and Chen, J. (2003). Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statistica Sinica*, 13:351–365.

[13] 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.

[14] Chen, J. (1995). Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233.

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

[16] Chen, J. and Tan, X. (2009). Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, 100(7):1367–1383.

[17] Chen, X., Ponomareva, M., and Tamer, E. (2014). Likelihood inference in some finite mixture models. *Journal of Econometrics*, 182(1):87–99.

[18] 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.

[19] 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.

[20] Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics*, 12(3):313–336.

[21] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

[22] Doraszelski, U. and Jaumandreu, J. (2018). Measuring the bias of technological change. *Journal of Political Economy*, 126(3):1027–1084.

[23] Foutz, R. V. and Srivastava, R. C. (1977). The performance of the likelihood ratio test when the model is incorrect. *The Annals of Statistics*, 5(6):1183–1194.

[24] Gandhi, A., Navarro, S., and Rivers, D. A. (2020). On the Identification of Gross Output Production Functions. *Journal of Political Economy*, 128(8):2973–3016.

[25] Garel, B. (2001). Likelihood ratio test for univariate Gaussian mixture. *Journal of Statistical Planning and Inference*, 96:325–350.

[26] Garel, B. (2005). Asymptotic theory of the likelihood ratio test for the identification of a mixture. *Journal of Statistical Planning and Inference*, 131:271–296.

[27] 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 Le Cam, L. and Olshen, R., editors, *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 789–806. Belmont, CA: Wadsworth.

[28] Hao, J. (2017). *NormalRegPanelMixture: Finite Mixture Model with Normal Panel Data*. R package version 1.0.

[29] Hartigan, J. (1985). Failure of log-likelihood ratio test. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 807–810. University of California Press 2. Berkeley.

[30] Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52(2):271–320.

[31] Kamakura, W. and Russell, G. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, 26(4):379–390.

[32] Kasahara, H. and Rodrigue, J. (2008). Does the use of imported intermediates increase productivity? plant-level evidence. *Journal of Development Economics*, 87(1):106–118.

[33] Kasahara, H., Schrimpf, P., and Suzuki, M. (2022). Identification and estimation of production function with unobserved heterogeneity. Technical report, ESRI Discussion Paper Series No.368.

[34] Kasahara, H. and Shimotsu, K. (2009). Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices. *Econometrica*, 77(1):135–175.

[35] Kasahara, H. and Shimotsu, K. (2012). Testing the number of components in finite mixture models.

[36] Kasahara, H. and Shimotsu, K. (2014). Non-parametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):97–111.

[37] Kasahara, H. and Shimotsu, K. (2015). Testing the number of components in normal mixture regression models. *Journal of the American Statistical Association*, 110(512):1632–1645.

[38] Kasahara, H. and Shimotsu, K. (2019). Testing the Order of Multivariate Normal Mixture Models.

[39] Kasahara, H., Suzuki, M., and Sawada, Y. (2021). The effect of bank recapitalization policy on credit allocation and corporate investment: Evidence from a banking crisis in japan: Economic and social research institute. Technical report, ESRI Discussion Paper Series No.365.

[40] Keane, M. P. and Wolpin, K. I. (1997). The career decisions of young men. *Journal of Political Economy*, 105(3):473–522.

[41] Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, third edition edition.

[42] Lemdani, M. and Pons, O. (1997). Likelihood ratio tests for genetic linkage. *Statistics and Probability Letters*, 33:15–22.

[43] Levinsohn, J. and Petrin, A. (2003). Estimating Production Functions Using Inputs to Control for Unobservables. *Review of Economic Studies*, pages 317–341.

[44] Li, P. and Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association*, 105:1084–1092.

[45] Li, P., Chen, J., and Marriott, P. (2009). Non-finite Fisher information and homogeneity: An EM approach. *Biometrika*, 96:411–426.

[46] Li, T. and Sasaki, Y. (2017). Constructive identification of heterogeneous elasticities in the cobb-douglas production function.

[47] Lin, C.-C. and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1:42–55.

[48] Lindsay, B. G. (1995). Mixture Models: Theory, Geometry and Applications NSF-CBMS Regional Conference Series in Probability and Statistics. *Source: NSF-CBMS Regional Conference Series in Probability and Statistics*, 5:1–163.

[49] Liu, G., Fu, Y., Li, P., and Pu, X. (2018). Using differential variability to increase the power of the homogeneity test in a two-sample problem. *Statistica Sinica*, 28(1):27–41.

[50] Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics*, 31:807–832.

[51] Lu, X. and Su, L. (2017). Determining the number of groups in latent panel structures with an application to income and democracy. *Quantitative Economics*, 8(3):729–760.

[52] McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*.

[53] Melnykov, V. and Maitra, R. (2010). Finite mixture models and model-based clustering. *Statistics Survey*, 4:80–116.

[54] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier.

[55] Niu, X., Li, P., and Zhang, P. (2011). Testing homogeneity in a multivariate mixture model. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 39(2):218–238.

[56] Olley, G. S. and Pakes, A. (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica*, pages 1263–1297.

[57] Robin, J.-M. and Smith, R. J. (2000). Tests of rank. *Econometric Theory*, 16(2):151–175.

[58] Shen, J. and He, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association*, 110(509):303–312.

[59] Su, L., Zhentao, S., and Phillips, P. (2016). Identifying latent structures in panel data. *Econometrica*, 84:2215–2264.

[60] Titterington, D. M., Smith, A. F., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley.

[61] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

[62] Zhu, H.-T. and Zhang, H. (2004). Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society, Series B*, 66:3–16.

A Proofs of propositions

Proof of Proposition 1. We first consider a model with intercept parameter and variance parameter but without covariates with $\mathbf{W}_i = \{y_{it}\}_{t=1}^T$.

Define

$$s_i^2 = \frac{1}{T-1}\sum_{t=1}^{T}(Y_{it} - \bar{Y}_i)^2 \quad \text{with} \quad \bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it},$$

where s_i^2 follows the chi-square distribution with the $T-1$ degrees of freedom. Let $i^* = \arg\min_{i=1,\ldots,n}\{s_i^2\}$ so that $s_{i^*}^2 = \min\{s_1^2, \ldots, s_n^2\}$ be the minimum of s_i^2 across all i's. We consider a sequence of parameters $\boldsymbol{\vartheta}_{2,n} = (\alpha_n, \boldsymbol{\theta}_{1,n}^\top, \boldsymbol{\theta}_{2,n}^\top)^\top$ with $\alpha_n = 1/n$, $\boldsymbol{\theta}_{1,n} = (\mu_{1,n}, \sigma_{1,n}^2)^\top = (\bar{Y}_{i^*}, s_{i^*}^2)^\top$, and $\boldsymbol{\theta}_{2,n} = \boldsymbol{\theta}^* = (\mu^*, \sigma^*)^\top$ for all n. Because $LR_n^*(\boldsymbol{\vartheta}_{2,n}) \leq LR_n^*(\tilde{\boldsymbol{\vartheta}}_{2,n})$, it suffices to show that $LR_n^*(\boldsymbol{\vartheta}_{2,n})$ is unbounded in probability.

Define

$$\ell(\boldsymbol{W}_i; \boldsymbol{\theta}) := \log f(\boldsymbol{W}_i; \boldsymbol{\theta}) = -\frac{T}{2}\log\sigma^2 - \frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\left(\frac{Y_{it}-\mu}{\sigma}\right)^2.$$

Then, the likelihood ratio test statistic for a two-component mixture is written as:

$$\begin{aligned} LR_n^*(\boldsymbol{\vartheta}_{2,n}) &= 2\left\{\sum_{i=1}^{n}\log\left(\alpha_n\prod_{t=1}^{T}\frac{1}{\sigma_{1,n}}\phi\left(\frac{Y_{it}-\mu_{1,n}}{\sigma_{1,n}}\right) + (1-\alpha_n)\prod_{t=1}^{T}\frac{1}{\sigma^*}\phi\left(\frac{Y_{it}-\mu^*}{\sigma^*}\right)\right) - \sum_{i=1}^{n}\ell(\boldsymbol{W}_i;\boldsymbol{\theta}^*)\right\} \\ &= 2\sum_{i\neq i^*}\left\{\log\left(\exp(\log\alpha_n + \ell(\boldsymbol{W}_i;\boldsymbol{\theta}_{1,n})) + \exp(\log(1-\alpha_n) + \ell(\boldsymbol{W}_i;\boldsymbol{\theta}^*))\right) - \ell(\boldsymbol{W}_i;\boldsymbol{\theta}^*)\right\} \\ &\quad + 2\left\{\log\left(\exp(\log\alpha_n + \ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}_{1,n})) + \exp(\log(1-\alpha_n) + \ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}^*))\right) - \ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}^*)\right\}. \end{aligned} \tag{41}$$

The first term on the right hand side of (41) can be re-written as:

$$= 2(n-1)\log\left(\frac{n-1}{n}\right) + 2\sum_{i\neq i^*}\log\left(1 + \frac{1}{n-1}\exp(\ell(\boldsymbol{W}_i;\boldsymbol{\theta}_{1,n}) - \ell(\boldsymbol{W}_i;\boldsymbol{\theta}^*))\right),$$

which is bounded from below by -1 as $n \to \infty$ because $\lim_{n\to\infty} 2(n-1)\log\left(\frac{n-1}{n}\right) = -1$ and $\log\left(1 + \frac{1}{n-1}\exp(\ell(\boldsymbol{W}_i;\boldsymbol{\theta}_{1,n}) - \ell(\boldsymbol{W}_i;\boldsymbol{\theta}^*))\right) \geq 0$ for all n.

The second term on the right-hand side of (41) is written as

$$2\{-\log n + \ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}_{1,n})\} + 2\log\left(1 + (n-1)\exp(\ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}^*) - \ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}_{1,n}))\right) - 2\ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}^*), \tag{42}$$

where $2\{-\log n + \ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}_{1,n})\}$ diverges to infinity as $n \to \infty$ by Lemma 1, while the second

term in (42) is bounded below from zero and the third term, is bounded in probability because $\ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}^*) = O_p(1)$. Therefore, for any $M < \infty$, we have $\Pr\left(LR_n^*(\boldsymbol{\vartheta}_{2,n}) \leq M\right) \to 0$ as $n \to \infty$. The stated result follows from $LR_n^*(\boldsymbol{\vartheta}_{2,n}) \leq LR_n^*(\tilde{\boldsymbol{\vartheta}}_{2,n})$ for all n.

For a model with covariates, we may consider a sequence of parameters $\boldsymbol{\vartheta}_{2,n} = (\alpha_n, \boldsymbol{\theta}_{1,n}^\top, \boldsymbol{\theta}_{2,n}^\top, \boldsymbol{\gamma}_n^\top)^\top$ with $\alpha_n = 1/n$, $\boldsymbol{\theta}_{1,n} = (\mu_{1,n}, \sigma_{1,n}^2, \boldsymbol{\beta}_{1,n}^\top)^\top = (\bar{Y}_{i^*} - \bar{\boldsymbol{Z}}_{i^*}^\top \boldsymbol{\gamma}^*, s_{i^*}^2, \mathbf{0}^\top)^\top$ with $\bar{\boldsymbol{Z}}_{i^*} = (1/T)\sum_{t=1}^T \boldsymbol{Z}_{it}$, $\boldsymbol{\theta}_{2,n} = \boldsymbol{\theta}^* = (\mu^*, \sigma^*, (\boldsymbol{\beta}^*)^\top)^\top$, and $\boldsymbol{\gamma}_n = \boldsymbol{\gamma}^*$. Then, repeating the above argument, the state result follows. □

Proof of Proposition 2. The stated result follows from repeating the proof of Proposition 6. □

Proof of Proposition 3. The proof follows that of Proposition 2 in Kasahara and Shimotsu (2012). For a vector $\boldsymbol{x}$ and a function $f(\boldsymbol{x})$, let $\nabla_{\boldsymbol{x}^k} f(\boldsymbol{x})$ denote its k-th derivative with respect to $\boldsymbol{x}$, which can be a multidimensional array. Observe that, for any finite k and for a neighborhood $\mathcal{N}$ of $\boldsymbol{\psi}^*$, we obtain

$$\begin{aligned} &E||\nabla_{\boldsymbol{\psi}^k} g(\boldsymbol{W}_i;\boldsymbol{\psi}^*,\alpha)/g(\boldsymbol{W}_i;\boldsymbol{\psi}^*,\alpha)||^2 < \infty, \\ &E||\sup_{\boldsymbol{\psi}\in\Theta_{\boldsymbol{\psi}}\cap\mathcal{N}} \nabla_{\boldsymbol{\psi}^k} \log g(\boldsymbol{W}_i;\boldsymbol{\psi},\alpha)||^2 < \infty, \end{aligned} \tag{43}$$

because each element of $\nabla_{\boldsymbol{\psi}^k} \log g(y|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\psi},\alpha)$ is written as a sum of products of Hermite polynomials. Note also that the following holds:

$$\nabla_{\eta\lambda_j} L_n(\boldsymbol{\psi}^*,\alpha) = 0, \quad \nabla_{\lambda_i\lambda_j\lambda_k} L_n(\boldsymbol{\psi}^*,\alpha) = O_p(n^{1/2}), \tag{44}$$

$$\nabla_{\eta\eta\lambda_i} L_n(\boldsymbol{\psi}^*,\alpha) = O_p(n), \quad \nabla_{\eta\eta\eta} L_n(\boldsymbol{\psi}^*,\alpha) = O_p(n), \tag{45}$$

where equation (44) follows from Proposition 3(a)(c) and (43) while equation (45) is a simple consequence of (43). Furthermore, for a neighborhood $\mathcal{N}$ of $\boldsymbol{\psi}^*$,

$$\sup_{\boldsymbol{\psi}\in\Theta_{\boldsymbol{\psi}}\cap\mathcal{N}} \left| n^{-1}\nabla^{(4)} L_n(\boldsymbol{\psi},\alpha) - E\nabla^{(4)} \log g(\boldsymbol{W}_i;\boldsymbol{\psi},\alpha) \right| = o_p(1), \tag{46}$$

$$E\nabla^{(4)} g(\boldsymbol{W}_i;\boldsymbol{\psi},\alpha) \text{ is continuous in } \psi\in\Theta_{\boldsymbol{\psi}}\cap\mathcal{N}. \tag{47}$$

Equations (46) and (47) follow from Lemma 2.4 of Newey and McFadden (1994) and the fact that $\nabla_{\boldsymbol{\psi}^k} \log g(\boldsymbol{w};\boldsymbol{\psi},\alpha)$ is written as a sum of products of Hermite polynomials.

Taking a fourth-order Taylor expansion of $L_n(\boldsymbol{\psi},\alpha)$ around $\boldsymbol{\psi}^*$ and using (43) and (44), we can

write $L_n(\boldsymbol{\psi},\alpha) - L_n(\boldsymbol{\psi}^*,\alpha)$ as the sum of relevant terms and the remainder term as follows:

$$
\begin{aligned}
&L_n(\boldsymbol{\psi},\alpha) - L_n(\boldsymbol{\psi}^*,\alpha) = \\
&\quad \nabla_{\boldsymbol{\eta}} L_n^*(\boldsymbol{\eta}-\boldsymbol{\eta}^*) + \frac{1}{2!}(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^\top \nabla_{\boldsymbol{\eta}\boldsymbol{\eta}^\top} L_n^*(\boldsymbol{\eta}-\boldsymbol{\eta}^*) + \frac{1}{2!}\sum_{i=1}^{q+2}\sum_{j=1}^{q+2} \nabla_{\lambda_i\lambda_j} L_n^* \lambda_i\lambda_j \qquad (48)\\
&\quad + \frac{3}{3!}\sum_{i=1}^{q+2}\sum_{j=1}^{q+2} (\boldsymbol{\eta}-\boldsymbol{\eta}^*)^\top \nabla_{\boldsymbol{\eta}\lambda_i\lambda_j} L_n^* \lambda_i\lambda_j \qquad (49)\\
&\quad + \frac{1}{4!}\sum_{i=1}^{q+2}\sum_{j=1}^{q+2}\sum_{k=1}^{q+2}\sum_{\ell=1}^{q+2} \nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell} L_n^* \lambda_i\lambda_j\lambda_k\lambda_\ell + R_n(\boldsymbol{\psi},\alpha), \qquad (50)
\end{aligned}
$$

where ∇L_n^* denotes the derivative of $L_n(\boldsymbol{\psi},\alpha)$ evaluated at $(\boldsymbol{\psi}^*,\alpha)$. In view of (44)-(45), the remainder term is written as

$$
\begin{aligned}
R_n(\boldsymbol{\psi},\alpha) &= O_p(n^{1/2})\sum_{i=1}^{q+2}\sum_{j=1}^{q+2}\sum_{k=1}^{q+2}\lambda_i\lambda_j\lambda_k + O_p(n)\left(\sum_{i=1}^{q+2}||\boldsymbol{\eta}-\boldsymbol{\eta}^*||^2\lambda_i + ||\boldsymbol{\eta}-\boldsymbol{\eta}^*||^3\right) \qquad (51)\\
&+ O_p(n)\sum_{i=1}^{q+2}\sum_{j=1}^{q+2}\sum_{k=1}^{q+2}\left(||\boldsymbol{\eta}-\boldsymbol{\eta}^*||^4 + ||\boldsymbol{\eta}-\boldsymbol{\eta}^*||^3|\lambda_i| + ||\boldsymbol{\eta}-\boldsymbol{\eta}^*||^2|\lambda_i\lambda_j| + ||\boldsymbol{\eta}-\boldsymbol{\eta}^*||\lambda_i\lambda_j\lambda_k|\right) \qquad (52)\\
&+ \frac{1}{4!}\sum_{i=1}^{q+2}\sum_{j=1}^{q+2}\sum_{k=1}^{q+2}\sum_{\ell=1}^{q+2}\{\nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell} L_n(\boldsymbol{\psi}^\dagger,\alpha) - \nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell} L_n(\boldsymbol{\psi}^*,\alpha)\}\lambda_i\lambda_j\lambda_k\lambda_\ell \qquad (53)
\end{aligned}
$$

with $\boldsymbol{\psi}^\dagger$ being between $\boldsymbol{\psi}$ and $\boldsymbol{\psi}^*$. Because $||\sqrt{n}t(\boldsymbol{\psi},\alpha)||^2 = n||\boldsymbol{\eta}-\boldsymbol{\eta}^*||^2 + n\sum_{i=1}^{q+2}\sum_{j=1}^{i}\alpha^2(1-\alpha)^2|\lambda_i\lambda_j|^2$, the right hand side of (51) and the terms in (52) are bounded by $O_p(1)(||\sqrt{n}t(\boldsymbol{\psi},\alpha)|| + ||\sqrt{n}t(\boldsymbol{\psi},\alpha)||^2)(||\boldsymbol{\eta}-\boldsymbol{\eta}^*|| + ||\lambda||)$. In view of (46) and (47), (53) is bounded by $||\sqrt{n}t(\boldsymbol{\psi},\alpha)||^2[d(\boldsymbol{\psi}^\dagger) + o_p(1)]$ with $d(\boldsymbol{\psi}^\dagger)\to 0$ as $\boldsymbol{\psi}^\dagger\to\boldsymbol{\psi}^*$, where a function $d(\boldsymbol{\psi}^\dagger)$ corresponds to $n^{-1}\mathbb{E}[\nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell} L_n(\boldsymbol{\psi}^\dagger,\alpha) - \nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell} L_n(\boldsymbol{\psi}^*,\alpha)]$. Therefore, $R_n(\boldsymbol{\psi},\alpha) = (1+||\sqrt{n}t(\boldsymbol{\psi},\alpha)||)^2[d(\boldsymbol{\psi}^\dagger) + o_p(1) + O_p(||\boldsymbol{\psi}-\boldsymbol{\psi}^*||)]$, and part (a) follows.

Part (b) follows from Lemman 3(c)(d), the Lindeberg-Levy central limit theorem, and the finiteness of $\boldsymbol{\mathcal{I}}$ in part (c).

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

$$
\boldsymbol{\mathcal{I}}_n = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}n} & \boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}\boldsymbol{\lambda}n} \\ \boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}\boldsymbol{\lambda}n}^\top & \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}n} \end{pmatrix}, \quad \boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}n}: (p+q+2)\times(p+q+2), \quad \boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}\boldsymbol{\lambda}n}: (p+q+2)\times q_\lambda, \quad \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}n}: q_\lambda\times q_\lambda,
$$

where q_λ represents the number of unique terms in $\sum_{i=1}^{q+2}\sum_{j=1}^{q+2}\sum_{k=1}^{q+2}\sum_{\ell=1}^{q+2}\lambda_i\lambda_j\lambda_k\lambda_\ell$. $\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}n}$ is given by $\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}n} = -n^{-1}\nabla_{\boldsymbol{\eta}\boldsymbol{\eta}^\top} L_n(\psi^*,\alpha)$. For $\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}\boldsymbol{\lambda}n}$, let $A_{ij} = n^{-1}\nabla_{\boldsymbol{\eta}\lambda_i\lambda_j} L_n(\psi^*,\alpha)$

and write the term in (49) as $(n/2)\sum_{i=1}^{q+2}\sum_{j=1}^{q+2}(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^\top A_{ij}\lambda_i\lambda_j = n\sum_{i=1}^{q+2}\sum_{j=1}^{i} c_{ij}(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^\top A_{ij}\lambda_i\lambda_j$, where the c_{ij}'s are defined when we introduce $\widetilde{\nabla}_{\boldsymbol{\theta}\boldsymbol{\theta}^\top} f^*$ after (9). Then, by defining $\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}\boldsymbol{\lambda}n} = -(c_{11}A_{11},\ldots,c_{qq}A_{qq},c_{12}A_{12},\ldots,c_{q-1,q}A_{q-1,q})/\alpha(1-\alpha)$, the term in (49) equals $-n(\boldsymbol{\eta}-\boldsymbol{\eta}^*)^\top\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}\boldsymbol{\lambda}n}[\alpha(1-\alpha)v(\boldsymbol{\lambda})]$. For $\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}n}$, define $B_{ijk\ell} = n^{-1}(8/4!)\nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell}L_n(\boldsymbol{\psi}^*,\alpha)$ so that the first term in (50) is written as $(n/8)\sum_{i=1}^{q+2}\sum_{j=1}^{q+2}\sum_{k=1}^{q+2}\sum_{\ell=1}^{q+2}B_{ijk\ell}\lambda_i\lambda_j\lambda_k\lambda_\ell = (n/2)\sum_{i=1}^{q+2}\sum_{j=1}^{i}\sum_{k=1}^{q+2}\sum_{\ell=1}^{k}c_{ij}c_{k\ell}B_{ijk\ell}\lambda_i\lambda_j\lambda_k\lambda_\ell$. Define $\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}n}$ such that the $(ij,k\ell)$'s element of $\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}n}$ is $-c_{ij}c_{k\ell}B_{ijk\ell}/\alpha^2(1-\alpha)^2$, where the ij's run over $\{(1,1),\ldots,(q,q),(1,2),\ldots,(q-1,q)\}$. Then, the first term in (50) equals $-(n/2)[\alpha(1-\alpha)v(\boldsymbol{\lambda})]'\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}n}[\alpha(1-\alpha)v(\boldsymbol{\lambda})]$. With this definition of $\boldsymbol{\mathcal{I}}_n$, the expansion (48)-(50) is written as (12) in terms of $\sqrt{n}t(\psi,\alpha)$.

We now show that $\boldsymbol{\mathcal{I}}_n \to_p \boldsymbol{\mathcal{I}}$. $\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}n} \to_p \boldsymbol{\mathcal{I}}_\eta$ holds trivially. For $\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}\boldsymbol{\lambda}n}$, it follows from Lemma 3(c) and the law of large numbers that $A_{ij} \to_p -\mathbb{E}[\nabla_{\boldsymbol{\eta}} l(\boldsymbol{W};\psi^*,\alpha)\nabla_{\lambda_i\lambda_j} l(\boldsymbol{W};\psi^*,\alpha)]$, giving $\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}\boldsymbol{\lambda}n} \to_p E\left[\boldsymbol{s}_{\boldsymbol{\eta}}\boldsymbol{s}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^\top/\alpha(1-\alpha)\right] = \boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}\boldsymbol{\lambda}}$. For $\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}n}$, Lemma 3(d) and the law of large numbers imply that $\sum_{i=1}^{q+2}\sum_{j=1}^{q+2}\sum_{k=1}^{q+2}\sum_{\ell=1}^{q+2}B_{ijk\ell}\lambda_i\lambda_j\lambda_k\lambda_\ell \to_p$
$-\sum_{i=1}^{q+2}\sum_{j=1}^{q+2}\sum_{k=1}^{q+2}\sum_{\ell=1}^{q+2}E[\nabla_{\lambda_i\lambda_j} l(\boldsymbol{W};\psi^*,\alpha)\nabla_{\lambda_k\lambda_\ell} l(\boldsymbol{W};\psi^*,\alpha)]\lambda_i\lambda_j\lambda_k\lambda_\ell$, where the factor $(8/4!) = 1/3$ in $B_{ijk\ell}$ and the three derivatives on the right hand side of Lemma 3(d) cancel each other. Therefore, we have $\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}n} \to_p E\left[\boldsymbol{s}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}\boldsymbol{s}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^\top/\alpha^2(1-\alpha)^2\right] = \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}}$, and $\boldsymbol{\mathcal{I}}_n \to_p \boldsymbol{\mathcal{I}}$ follows.

We complete the proof of part (c) by showing that $\boldsymbol{\mathcal{I}} = E[\boldsymbol{s}(\boldsymbol{W})\boldsymbol{s}(\boldsymbol{W})^\top]$ is finite and non-singular. Note that $\boldsymbol{s}(\boldsymbol{W})$ can be expressed in Hermite polynomials as in (74). Then, the finiteness of $\boldsymbol{\mathcal{I}}$ follows from Assumption 2(a) and the definition of Hermite polynomials.

To show that $\boldsymbol{\mathcal{I}}$ is positive definite, it suffices to show that there exists no multi-collinearity in $\boldsymbol{s}(\boldsymbol{w})$. Suppose, on the contrary, that $\boldsymbol{s}(\boldsymbol{w})$ is multi-collinear and that there exists a non-zero vector $\boldsymbol{a}$ that solves the equation $\boldsymbol{a}^\top\boldsymbol{s}(\boldsymbol{w}) = 0$ for all values of $\boldsymbol{w}$. Partition $s(\boldsymbol{w})$ as $\boldsymbol{s}(\boldsymbol{w}) = (\boldsymbol{s}_{(\mu)}^\top, \boldsymbol{s}_{(\beta)}^\top)^\top$ with $\boldsymbol{s}_{(\mu)} = (s_\mu, s_\sigma, s_{\lambda_{\mu\mu}}, s_{\lambda_{\mu\sigma}}, s_{\lambda_{\sigma\sigma}})^\top$ and $\boldsymbol{s}_{(\beta)} = (\boldsymbol{s}_{\boldsymbol{\beta}}^\top, \boldsymbol{s}_{\boldsymbol{\gamma}}^\top, \boldsymbol{s}_{\lambda_{\mu\boldsymbol{\beta}}}^\top, \boldsymbol{s}_{\lambda_{\sigma\boldsymbol{\beta}}}^\top, \boldsymbol{s}_{\lambda_{\boldsymbol{\beta}\boldsymbol{\beta}}}^\top)^\top$, where $\boldsymbol{s}(\boldsymbol{w})$ is defined in (9) and (74). Similarly, partition $\boldsymbol{a}$ as $\boldsymbol{a} = (\boldsymbol{a}_{(\mu)}^\top, \boldsymbol{a}_{(\beta)}^\top)^\top$ so that

$$\boldsymbol{a}^\top\boldsymbol{s}(\boldsymbol{w}) = \boldsymbol{a}_{(\mu)}^\top\boldsymbol{s}_{(\mu)} + \boldsymbol{a}_{(\beta)}^\top\boldsymbol{s}_{(\beta)}. \tag{54}$$

By Assumption 2(b) and the property of Hermite polynomials, if $\boldsymbol{a}^\top\boldsymbol{s}(\boldsymbol{w}) = \boldsymbol{0}$ for all $\boldsymbol{w}$, then $\boldsymbol{a}_{(\beta)} = \boldsymbol{0}$.

Then, in view of (54), the stated result follows if we can show that $\boldsymbol{a}_{(\mu)}^\top\boldsymbol{s}_{(\mu)} = \boldsymbol{0}$ for all $\boldsymbol{w}$ implies $\boldsymbol{a}_{(\mu)} = 0$. Suppose that

$$\begin{aligned}\boldsymbol{a}_{(\mu)}^\top\boldsymbol{s}_{(\mu)} = {} & a_\mu\sum_{t=1}^{T}H_{i,t}^{1*} + (a_\sigma + a_{\lambda_{\mu\mu}})\sum_{t=1}^{T}H_{i,t}^{2*} + \frac{a_{\lambda_{\mu\mu}}}{2}\sum_{t=1}^{T}\sum_{s\neq t}H_{i,t}^{1*}H_{i,s}^{1*} \\ & + a_{\lambda_{\mu\sigma}}\sum_{t=1}^{T}H_{i,t}^{3*} + a_{\lambda_{\mu\sigma}}\sum_{t=1}^{T}\sum_{s\neq t}H_{i,t}^{1*}H_{i,s}^{2*} + 3a_{\lambda_{\sigma\sigma}}\sum_{t=1}^{T}H_{i,t}^{4*} + \frac{a_{\lambda_{\sigma\sigma}}}{2}\sum_{t=1}^{T}\sum_{s\neq t}H_{i,t}^{2*}H_{i,s}^{2*} = 0\end{aligned}$$

for all $\boldsymbol{w}$, where $H_{i,t}^{j*}$ for $j = 1, 2, 3$ is defined in (73) in Appendix B.2.

Because the above equation hold for all values of $\boldsymbol{w}$, with the property of the Hermite polynomials, we have $a_\mu = 0, (a_\sigma + a_{\lambda_{\mu\mu}}) = 0, a_{\lambda_{\mu\mu}} = 0, a_{\lambda_{\mu\sigma}} = 0, a_{\lambda_{\sigma\sigma}} = 0$. This implies that $\boldsymbol{a}_{(\mu)} = 0$. Therefore, there exists no multi-collinearity in $\boldsymbol{s}(\boldsymbol{w})$, and $\boldsymbol{\mathcal{I}}$ is non-singular, proving part (c). □

Proof of Proposition 4. The proof is similar to that of Proposition 3 in Kasahara and Shimotsu (2015).

The proof of part (a) closely follows the proof of Theorem 1 of Andrews (1999). Let $\boldsymbol{T}_n := \boldsymbol{\mathcal{I}}_n^{1/2}\sqrt{n}\boldsymbol{t}(\hat{\boldsymbol{\psi}}_\alpha, \alpha)$. Then, in view of (12), we have

$$\begin{aligned} o_p(1) &\leq L_n(\hat{\boldsymbol{\psi}}_\alpha, \alpha) - L_n(\boldsymbol{\psi}^*, \alpha) \\ &= \boldsymbol{T}_n'\boldsymbol{\mathcal{I}}_n^{-1/2}\boldsymbol{S}_n - \frac{1}{2}||\boldsymbol{T}_n||^2 + R_n(\hat{\boldsymbol{\psi}}_\alpha, \alpha) \\ &= O_p(||\boldsymbol{T}_n||) - \frac{1}{2}||\boldsymbol{T}_n||^2 + (1 + ||\boldsymbol{\mathcal{I}}_n^{-1/2}\boldsymbol{T}_n||)^2 o_p(1) \\ &= ||\boldsymbol{T}_n||O_p(1) - \frac{1}{2}||\boldsymbol{T}_n||^2 + o_p(||\boldsymbol{T}_n||) + o_p(||\boldsymbol{T}_n||^2) + o_p(1), \end{aligned}$$

where the third equality holds because $\boldsymbol{\mathcal{I}}_n^{-1/2}\boldsymbol{S}_n = O_p(1)$ and $R_n(\hat{\boldsymbol{\psi}}_\alpha, \alpha) = o_p((1 + ||\boldsymbol{\mathcal{I}}_n^{-1/2}\boldsymbol{T}_n||)^2)$ from Propositions 2 and 3. Rearranging this equation yields $||\boldsymbol{T}_n||^2 \leq 2||\boldsymbol{T}_n||O_p(1) + o_p(1)$. Denote the $O_p(1)$ term by ς_n. Then, $(||\boldsymbol{T}_n|| - \varsigma_n)^2 \leq \varsigma_n^2 + o_p(1) = O_p(1)$; taking its square root gives $||\boldsymbol{T}_n|| \leq O_p(1)$. In conjunction with $\boldsymbol{\mathcal{I}}_n \to_p \boldsymbol{\mathcal{I}}$, we obtain $\sqrt{n}\boldsymbol{t}(\hat{\boldsymbol{\psi}}_\alpha, \alpha) = O_p(1)$, and part (a) follows.

For part (b), noting that $L_n(\boldsymbol{\psi}^*, \alpha) = L_{0,n}(\boldsymbol{\gamma}_0^*, \boldsymbol{\theta}_0^*)$, write

$$LR_n = \max_{\alpha\in[\epsilon,1-\epsilon]} 2\{L_n(\hat{\boldsymbol{\psi}}_\alpha, \alpha) - L_n(\boldsymbol{\psi}^*, \alpha)\} - 2\{L_{0,n}(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\theta}}_0) - L_{0,n}(\boldsymbol{\gamma}_0^*, \boldsymbol{\theta}_0^*)\}. \tag{55}$$

Define

$$\boldsymbol{S}_n = \begin{pmatrix} \boldsymbol{S}_{\boldsymbol{\eta}n} \\ \boldsymbol{S}_{\boldsymbol{\lambda}n} \end{pmatrix} := \begin{pmatrix} n^{-1/2}\sum_{i=1}^n \boldsymbol{s}_{\boldsymbol{\eta}}(\boldsymbol{W}_i) \\ n^{-1/2}\sum_{i=1}^n \boldsymbol{s}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{W}_i) \end{pmatrix}, \quad \begin{aligned} \boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}n} &:= \boldsymbol{S}_{\boldsymbol{\lambda}n} - \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}\boldsymbol{\eta}}\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}}^{-1}\boldsymbol{S}_{\boldsymbol{\eta}n}, \quad \boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}n} := \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^{-1}\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}n}, \\ \boldsymbol{t}_{\boldsymbol{\eta},\boldsymbol{\lambda}} &:= \boldsymbol{t}_{\boldsymbol{\eta}} - \boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}}\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}\boldsymbol{\lambda}}^{-1}\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}, \alpha), \end{aligned}$$

and split the quadratic form in (12) to obtain

$$2\{L_n(\boldsymbol{\psi}, \alpha) - L_n(\boldsymbol{\psi}^*, \alpha)\} = B_n(\sqrt{n}\boldsymbol{t}_{\boldsymbol{\eta},\boldsymbol{\lambda}}) + C_n(\sqrt{n}\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}, \alpha)) + R_n(\boldsymbol{\psi}, \alpha), \tag{56}$$

where

$$\begin{aligned} B_n(\boldsymbol{t}_{\boldsymbol{\eta},\boldsymbol{\lambda}}) &= 2\boldsymbol{t}_{\boldsymbol{\eta},\boldsymbol{\lambda}}^\top\boldsymbol{S}_{\boldsymbol{\eta}n} - \boldsymbol{t}_{\boldsymbol{\eta},\boldsymbol{\lambda}}^\top\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}}\boldsymbol{t}_{\boldsymbol{\eta},\boldsymbol{\lambda}}, \\ C_n(\boldsymbol{t}_{\boldsymbol{\lambda}}) &= 2\boldsymbol{t}_{\boldsymbol{\lambda}}^\top\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}n} - \boldsymbol{t}_{\boldsymbol{\lambda}}^\top\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}\boldsymbol{t}_{\boldsymbol{\lambda}} \\ &= \boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}n}^\top\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}\boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}n} - (\boldsymbol{t}_{\boldsymbol{\lambda}} - \boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}n})^\top\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}(\boldsymbol{t}_{\boldsymbol{\lambda}} - \boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}n}), \end{aligned} \tag{57}$$

with $\boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}^n} \xrightarrow{d} \boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\eta}} = (\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}})^{-1}\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ and $\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}^n} \xrightarrow{d} \boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}} \sim N(\boldsymbol{0}, \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}})$. Also, $R_n(\hat{\boldsymbol{\psi}}_\alpha, \alpha) = o_p(1)$ holds from Proposition 3(a) and $\sqrt{n}\boldsymbol{t}(\hat{\boldsymbol{\psi}}_\alpha, \alpha) = O_p(1)$.

Because $\Delta_{(\gamma,\theta)} f(x; \hat{\boldsymbol{\gamma}}_0^*, \hat{\boldsymbol{\theta}}_0^*)$ is identical to $\Delta_{\boldsymbol{\eta}} f(x; \boldsymbol{\psi}^*, \alpha)$, a standard analysis gives $2[L_{0,n}(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\theta}}_0) - L_{0,n}(\boldsymbol{\gamma}_0^*, \boldsymbol{\theta}_0^*)] = \max_{\boldsymbol{t}_{\boldsymbol{\eta}}} B_n(\sqrt{n}\boldsymbol{t}_{\boldsymbol{\eta}}) + o_p(1)$. Note that the possible values of both $\sqrt{n}\boldsymbol{t}_{\boldsymbol{\eta}}$ and $\sqrt{n}\boldsymbol{t}_{\boldsymbol{\eta},\boldsymbol{\lambda}}$ approaches $\mathbb{R}^{q+2}$. Therefore, in view of (56)-(57), we can write equation (55) as

$$LR_n = C_n(\sqrt{n}\boldsymbol{t}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\lambda}}, \alpha)) + o_p(1), \tag{58}$$

where $\hat{\boldsymbol{\lambda}}$ is defined in (14).

The asymptotic distribution of LR_n follows from applying Theorem 3(c) of (Andrews, 1999, p.1362) to (56) and (58). First, Assumption 2 of Andrews (1999) holds because Assumption 2* of Andrews (1999) hold because of Proposition 2(a). Second, Assumption 3 of Andrews (1999) holds with $B_T = n^{1/2}$ and $T = n$ because $\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}^n} \xrightarrow{d} \boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}} \sim N(\boldsymbol{0}, \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}})$ and $\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ is non-singular. Assumption 4 of Andrews (1999) holds from part (a). Assumption 5 of Andrews (1999) follows from Assumption 5* and Lemma 3 of Andrews (1999) with $b_T = n^{1/2}$ because $\alpha(1-\alpha)v(\Theta_{\boldsymbol{\lambda}})$ is locally equal to $\Lambda_{\boldsymbol{\lambda}}$. Therefore, it follows from Theorem 3(c) of Andrews (1999) that $C_n(\sqrt{n}\boldsymbol{t}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\lambda}}, \alpha)) \xrightarrow{d} (\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}})^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}$, where $\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}$ is defined by (16).

□

Proof of Proposition 5. Under $H_{2,0}$, we obtain the $\vartheta_{M_0+1} \in \Upsilon_{2h}^*$,

$$\begin{aligned}
&\mathbb{E}[\{\nabla_{\alpha_h} \log f_{M_0+1}(\boldsymbol{W}_i, \vartheta_{M_0+1})\}^2] \\
&= \int \frac{\{f(\boldsymbol{w}; \boldsymbol{\mu}_h, \boldsymbol{\sigma}_h) - f(\boldsymbol{w}; \boldsymbol{\mu}_{M_0}^*, \boldsymbol{\sigma}_{M_0}^*)\}^2}{\sum_{j=1}^{M_0} \alpha_j^* f(\boldsymbol{w}; \boldsymbol{\mu}_j^*, \boldsymbol{\sigma}_j^*)} d\boldsymbol{w} \\
&= \int \frac{\{f(\boldsymbol{w}; \boldsymbol{\mu}_h, \boldsymbol{\sigma}_h)\}^2}{\sum_{j=1}^{M_0} \alpha_j^* f(\boldsymbol{w}; \boldsymbol{\mu}_j^*, \boldsymbol{\sigma}_j^*)} d\boldsymbol{w} + \int \frac{\{f(\boldsymbol{w}; \boldsymbol{\mu}_{M_0}^*, \boldsymbol{\sigma}_{M_0}^*)\}^2}{\sum_{j=1}^{M_0} \alpha_j^* f(\boldsymbol{w}; \boldsymbol{\mu}_j^*, \boldsymbol{\sigma}_j^*)} d\boldsymbol{w} - 2\int \frac{f(\boldsymbol{w}; \boldsymbol{\mu}_h, \boldsymbol{\sigma}_h) f(\boldsymbol{w}; \boldsymbol{\mu}_{M_0}^*, \boldsymbol{\sigma}_{M_0}^*)}{\sum_{j=1}^{M_0} \alpha_j^* f(\boldsymbol{w}; \boldsymbol{\mu}_j^*, \boldsymbol{\sigma}_j^*)} d\boldsymbol{w}.
\end{aligned} \tag{59}$$

The latter two terms on the right-hand side of (59) are bounded because $f(\boldsymbol{w}; \boldsymbol{\mu}_{M_0}^*, \boldsymbol{\sigma}_{M_0}^*) / \sum_{j=1}^{M_0} \alpha_j^* f(\boldsymbol{w}; \boldsymbol{\mu}_j^*, \boldsymbol{\sigma}_j^*) \leq (1/\alpha_{M_0}^*)$ for any $\boldsymbol{w}$ and $f(\boldsymbol{w}; \mu, \sigma)$ integrates to one. Therefore, the left-hand side of (59) goes to infinity if and only if the first term on the right-hand side of (59) goes to infinity.

Because $\max_j \alpha_j \leq \sum_j^{M_0} \alpha_j \leq M_0 \max_j \alpha_j$, we obtain

$$\frac{1}{M_0} \frac{\{f(\boldsymbol{w}; \boldsymbol{\mu}_h, \boldsymbol{\sigma}_h)\}^2}{\max_j \{\alpha_j^* f(\boldsymbol{w}; \boldsymbol{\mu}_j^*, \boldsymbol{\sigma}_j^*)\}} \leq \frac{\{f(\boldsymbol{w}; \boldsymbol{\mu}_h, \boldsymbol{\sigma}_h)\}^2}{\sum_{j=1}^{M_0} \alpha_j^* f(\boldsymbol{w}; \boldsymbol{\mu}_j^*, \boldsymbol{\sigma}_j^*)} \leq \frac{\{f(\boldsymbol{w}; \boldsymbol{\mu}_h, \boldsymbol{\sigma}_h)\}^2}{\max_j \{\alpha_j^* f(\boldsymbol{w}; \boldsymbol{\mu}_j^*, \boldsymbol{\sigma}_j^*)\}}.$$

Without loss of generality, we assume that $\sigma_{M_0}^* = \max\{\sigma_1^*, \ldots, \sigma_{M_0}^2\}$ and the maximum is unique. Then there exists $M \in (0, \infty)$, such that $\max_j \{\alpha_j^* f(\boldsymbol{w}, \boldsymbol{\mu}_j^*, \boldsymbol{\sigma}_j^2)\} = \alpha_{M_0}^* f(\boldsymbol{w}, \boldsymbol{\mu}_{M_0}^*, \boldsymbol{\sigma}_{M_0}^2)$

when $|y_t| > M \; \forall t = 1, \ldots, T$. Note that

$$\begin{aligned}\frac{\{f(\boldsymbol{w}; \boldsymbol{\mu}_h, \boldsymbol{\sigma}_h)\}^2}{f(\boldsymbol{w}; \boldsymbol{\mu}^*_{M_0}, \boldsymbol{\sigma}^*_{M_0})\}} &= \prod_{t=1}^{T} \frac{(\sigma^*_{M_0})^2}{(2\pi)^{1/2}\sigma_h^2} \exp\left\{-\frac{1}{\sigma_h^2}(y_t - \mu_h)^2 + \frac{1}{2(\sigma^*_{M_0})^2}(y_t - \mu^*_{M_0})^2\right\} \\ &= \frac{(\sigma^*_{M_0})^{2T}}{(2\pi)^{1/2}\sigma_h^{2T}} \exp\left\{-\frac{1}{\sigma_h^2}\sum_{t=1}^{T}(y_t - \mu_h)^2 + \frac{1}{2(\sigma^*_{M_0})^2}\sum_{t=1}^{T}(y_t - \mu^*_{M_0})^2\right\}.\end{aligned} \tag{60}$$

The stated result follows because the integral of this over $|y| \geq M$ is finite if $\sigma_h^2/\sigma^{2*}_{M_0} < 2$ and infinite if $\sigma_h^2/\sigma^{2*}_{M_0} > 2$. When $\sigma_h^2/\sigma^{2*}_{M_0} = 2$, it is finite if $\mu_h = \mu^*_{M_0}$ and infinite if $\mu_h \neq \mu^*_{M_0}$. □

Proof of Proposition 6. Our panel data model can be viewed as a special case of the T-dimensional multivariate normal mixture models, where the variance-covariance matrix for each component is given by a $T \times T$ diagonal matrix, $\Sigma_j := \text{diag}(\sigma_j^2, \ldots, \sigma_2^2)$. Chen and Tan (2009) provides the consistency proof for the penalized maximum likelihood estimator for a multivariate normal mixture under their conditions C1-C3 for the penalty function but Alexandrovich (2014) identifies a soft spot in the proof of Chen and Tan (2009) and provides an alternative consistency proof by strengthening the condition C3 of Chen and Tan (2009). Their $p_n(G)$ and $\tilde{p}_n(G)$ correspond to our $\tilde{p}_n(\boldsymbol{\vartheta}_2)$ and $p_n(\boldsymbol{\vartheta}_2)$, respectively; consequently, the condition C1 and C2 in Chen and Tan (2009) and a version of condition C3 strengthned by Alexandrovich (2014) can be stated in our notation as follows:

C1. The penalty function is written as $\tilde{p}_n(\boldsymbol{\vartheta}_M) = \sum_{j=1}^{M} p_n(\sigma_j^2)$

C2. For any fixed $\boldsymbol{\vartheta}_M$ with $\sigma_j^2 > 0$ for $j = 1, 2, ..., M$, we have $\tilde{p}_n(\boldsymbol{\vartheta}_M) = o(n)$ and $\sup_{\boldsymbol{\vartheta}_M \in \boldsymbol{\Theta}_M} \max\{0, \tilde{p}_n(\boldsymbol{\vartheta}_M)\} = o(n)$. In addition, $\tilde{p}_n(\boldsymbol{\vartheta}_M)$ is differentiable with respect to $\boldsymbol{\vartheta}_M$ and as $n \to \infty$, $\nabla_{\boldsymbol{\vartheta}_M} \tilde{p}_n(\boldsymbol{\vartheta}_M) = o(n^{1/2})$ at any fixed $\boldsymbol{\vartheta}$ such that $\sigma_j^2 > 0$ for $j = 1, 2, ..., M$.

A version of C3 by Alexandrovich (2014). For large enough n, $p_n(\sigma_j^2) \leq \left(\frac{3}{4}\sqrt{n \log\log n}\right) \log(\sigma_j^2)$, when $\sigma_j^2 < cn^{-2}$ for some $c > 0$.

The consistency of the PMLE, $\hat{\boldsymbol{\vartheta}}_{M_0}$ and $\hat{\boldsymbol{\vartheta}}_{M_0+1}$, follows from Theorems 1 and 3 of Chen and Tan (2009) and Corollary 3 of Alexandrovich (2014) if we can show that the above three conditions hold for our penalty function (3). Given (3), C1 trivially holds. Under Assumption 1(b), C2 also holds because $a_n = O(n^{1/4-\zeta})$ with $\zeta > 0$ implies $a_n = o(n)$ or $o(n^{1/2})$, and $\Delta_{\sigma_j^2} \tilde{p}_n(\boldsymbol{\vartheta}) = -a_n(-\sigma_0^2/(\sigma_j^2) + 1/\sigma_j^2) = o(n^{1/4})$ if $\sigma_j^2 > 0$. For C3, suppose that $\sigma_j^2 < n^{-2}$. Then, because $a_n = o(n^{1/4})$ and $a_n > 0$, $p_n(\sigma_j) = -a_n\left(\sigma_j^{-2}\sigma_0^2 + 2\log(\sigma_j/\sigma_0) - 1\right) < -c_n\left(n^{9/4}\sigma_0^2 - 2n^{5/4}\log(n\sigma_0) - n^{1/4}\right) < \left(\frac{3}{4}\sqrt{n \log\log n}\right) 2\log(n)$ when n is large enough, where c_n is a sequence of positive numbers that are bounded. Therefore, $\tilde{p}_n(\boldsymbol{\vartheta}_M)$ satisfies the above three conditions, and the stated result follows follows from Theorems 1 and 3 of Chen and Tan (2009) and Corollary 3 of Alexandrovich (2014). □

Proof of Proposition 7. For $h = 1, \ldots, M_0$, let $\mathcal{N}_h^* \subset \Theta_{\boldsymbol{\vartheta}_{M_0+1}}(\epsilon)$ be a sufficiently small closed neighbourhood of Υ_{1h}^* such that $\alpha_h, \alpha_{h+1} > 0$ hold and $\Upsilon_{1k}^* \notin \mathcal{N}_h^*$ if $k \neq h$. Consider the following one-to-one reparameterization from the (M_0+1)-component model parameter $\boldsymbol{\vartheta}_{M_0+1} = (\alpha_1, \ldots, \alpha_{M_0}, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_h^\top, \boldsymbol{\theta}_{h+1}^\top, \ldots, \boldsymbol{\theta}_{M_0+1}^\top, \boldsymbol{\gamma}^\top)^\top$. Similarly to (5), the one-to-one reparamterization for testing the null hypothesis $H_{0,1h}$ is given by

$$\begin{pmatrix} \boldsymbol{\lambda}_h \\ \boldsymbol{\nu}_h \end{pmatrix} := \begin{pmatrix} \boldsymbol{\theta}_h - \boldsymbol{\theta}_{h+1} \\ \tau\boldsymbol{\theta}_h + (1-\tau)\boldsymbol{\theta}_{h+1} \end{pmatrix} \text{ so that } \begin{pmatrix} \boldsymbol{\theta}_h \\ \boldsymbol{\theta}_{h+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\nu} + (1-\tau)\boldsymbol{\lambda} \\ \boldsymbol{\nu} - \tau\boldsymbol{\lambda} \end{pmatrix}$$

and reparameterize α_j for $j = 1, 2, ..., M_0$ as

$$(\pi_1, \ldots, \pi_{h-1}, \pi_h, \pi_{h+1}, \ldots, \pi_{M_0-1}) = (\alpha_1, \ldots, \alpha_{h-1}, (\alpha_h + \alpha_{h+1}), \alpha_{h+2}, \ldots, \alpha_{M_0})$$
$$\tau = \alpha_h/(\alpha_h + \alpha_{h+1})$$

so that $\pi_h = \alpha_h + \alpha_{h+1}$ and $\pi_{M_0} = 1 - \sum_{j=1}^{M_0-1} \pi_j$.

Collect the reparameterized parameters except τ as

$$\boldsymbol{\psi}_{h,\tau} = (\boldsymbol{\eta}^\top, \boldsymbol{\lambda}_h^\top)^\top \quad \text{with} \quad \boldsymbol{\eta} = (\pi_1, \ldots, \pi_{M_0-1}, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_{h-1}^\top, \boldsymbol{\nu}_h^\top, \boldsymbol{\theta}_{h+2}^\top, \ldots, \boldsymbol{\theta}_{M_0+1}^\top, \boldsymbol{\gamma})^\top.$$

In the reparameterized model, the null restriction $\boldsymbol{\theta}_h = \boldsymbol{\theta}_{h+1}$ implied by $H_{0,1h}$ holds if and only if $\boldsymbol{\lambda}_h = 0$. Under $H_{0,1h}$, we have $\lambda_h^* = 0$ and $\eta^* = (\alpha_1^*, ..., \alpha_{M_0-1}^*, (\boldsymbol{\theta}_1^*)^\top, \ldots, (\boldsymbol{\theta}_{M_0}^*)^\top, (\boldsymbol{\gamma}^*)^\top)^\top$. Define the log-likelihood under the reparameterized parameters as

$$f_{M_0+1}^h(\boldsymbol{w}; \boldsymbol{\psi}_{h,\tau}, \tau) = \pi_h g^h(\boldsymbol{w}, \boldsymbol{\psi}_{h,\tau}, \tau) + \sum_{j=1}^{h-1} \pi_j f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\theta}_j) + \sum_{j=h}^{M_0} \pi_{j+1} f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\theta}_{j+1}),$$

where $g^h(\boldsymbol{w}, \boldsymbol{\psi}_{h,\tau}, \tau)$ is defined similarly to (6) as

$$g^h(\boldsymbol{w}, \boldsymbol{\psi}_{h,\tau}, \tau) = \tau f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu}_h + (1-\tau)\boldsymbol{\lambda}_h) + (1-\tau) f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu}_h - \tau\boldsymbol{\lambda}_h). \tag{61}$$

Define the local penalized MLE of $\boldsymbol{\psi}_{h,\tau}$ by

$$\hat{\boldsymbol{\psi}}_{h,\tau} := \underset{\boldsymbol{\psi}_{h,\tau} \in \mathcal{N}_h^*}{\arg\max} L_n^h(\boldsymbol{\psi}_{h,\tau}, \tau) + \sum_{j=1}^{M_0} p_n(\sigma_j^2(\boldsymbol{\psi}_{h,\tau}, \tau)), \tag{62}$$

where $L_n^h(\boldsymbol{\psi}_{h,\tau}, \tau) := \sum_{i=1}^N \log g^h(\boldsymbol{W}_i; \boldsymbol{\psi}_{h,\tau}, \tau)$ and $\sigma_j^2(\boldsymbol{\psi}_{h,\tau}, \tau)$ is the value of σ_j^2 implied by the value of $\boldsymbol{\psi}_{h,\tau}$ and τ. Because $\boldsymbol{\psi}_{h,\tau}^*$ is the only parameter value in $\mathcal{N}_h^*$ that generates the true density, $\hat{\boldsymbol{\psi}}_{h,\tau} - \boldsymbol{\psi}_{h,\tau}^* = o_p(1)$ follows Proposition 4.

For $\epsilon \in (0, 1/2)$, define the LRTS for testing $H_{0,1h}$ as $LR_n^{M_0,h} := \max_{\tau \in [\epsilon, 1-\epsilon]} 2(L_n^h(\hat{\boldsymbol{\psi}}_{h,\tau}, \tau) -$

$L_{0,n}(\hat{\boldsymbol{\vartheta}}_{M_0}))$ Because $\hat{\sigma}_j^2 - \sigma_{0,j}^2 = O_p(n^{-1/4})$ under the null hypothesis (c.f., Proposition 4(a)), we have $\tilde{p}_n(\boldsymbol{\vartheta}_{M_0+1}) = o_p(1)$ by Assumption 3(c), and $PLR_n^{M_0,h} - LR_n^{M_0,h} = o_p(1)$ follows for $h = 1, ..., M_0$.

Then, in view of (22), the stated result holds if

$$(LR_n^{M_0,1}, \ldots, LR_n^{M_0,M_0})^\top \xrightarrow{d} (\hat{\boldsymbol{t}}_\lambda^1)^\top \boldsymbol{\mathcal{I}}_{\eta,\lambda}^1 (\hat{\boldsymbol{t}}_\lambda^1), \ldots, (\hat{\boldsymbol{t}}_\lambda^{M_0})^\top \boldsymbol{\mathcal{I}}_{\eta,\lambda}^{M_0} (\hat{\boldsymbol{t}}_\lambda^{M_0}))^\top. \tag{63}$$

Observe that $L_n^h(\boldsymbol{\psi}_{h,\tau}, \tau) - L_n^h(\boldsymbol{\psi}_{h,\tau}^*, \tau)$ admits the same expansion as $L_n(\hat{\boldsymbol{\psi}}, \alpha) - L_n(\boldsymbol{\psi}^*, \alpha)$ in (13) and (56) by replacing $(\alpha, \boldsymbol{t}(\boldsymbol{\psi}, \alpha), \boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}, \alpha), \boldsymbol{S}_n, \boldsymbol{G}_n, \boldsymbol{\mathcal{I}}_n, R_n(\boldsymbol{\psi}, \alpha))$ with $(\tau, \boldsymbol{t}^h(\boldsymbol{\psi}^h, \tau), \boldsymbol{t}_{\boldsymbol{\lambda}}^h(\boldsymbol{\lambda}^h, \tau), \boldsymbol{S}_n^h, \boldsymbol{G}_n^h, \boldsymbol{\mathcal{I}}_n^h, R_n^h(\boldsymbol{\psi}^h, \tau))$, where $(\boldsymbol{S}_n^h, \boldsymbol{\mathcal{I}}_n^h)$ is defined similarly to $(\boldsymbol{S}_n, \boldsymbol{\mathcal{I}}_n)$ but replacing $(\boldsymbol{s}_{\boldsymbol{\eta}}, \boldsymbol{s}_{\boldsymbol{\lambda\lambda}})$ with $(\tilde{\boldsymbol{s}}_{\boldsymbol{\eta}}, \boldsymbol{s}_{\boldsymbol{\lambda\lambda}}^h)$ while $\boldsymbol{G}_n^h := (\boldsymbol{\mathcal{I}}_n^h)^{-1} \boldsymbol{S}_n^h$. Applying the proof of Proposition 3, we have $\boldsymbol{S}_n^h \xrightarrow{d} \boldsymbol{S}^h \sim N(\boldsymbol{0}, \boldsymbol{\mathcal{I}}^h)$ and $\boldsymbol{\mathcal{I}}_n^h \xrightarrow{p} \boldsymbol{\mathcal{I}}^h$. Then, (62) follows from the proof of Propositions 3 and 4 for each local penalized MLE by replacing $(\boldsymbol{G}_n, \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}, \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}})$ with $(\boldsymbol{G}_n^h, \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^h, \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h)$, and collecting the results while noting that $(\boldsymbol{S}_n^1, ..., \boldsymbol{S}_n^{M_0}) \xrightarrow{d} (\boldsymbol{S}^1, ..., \boldsymbol{S}^{M_0})$. □

Proof of Proposition 8. The proof is similar to that of Proposition 7 in Kasahara and Shimotsu (2015). Let ω_n^h denote the sample counterpart of $(\hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^h)^\top \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h \hat{\boldsymbol{t}}_{\boldsymbol{\lambda}}^h$ in Proposition 7 such that the LRTS satisfies $2\{L_n^h(\hat{\boldsymbol{\psi}}_{h,\tau}, \tau) - L_{0,n}(\hat{\boldsymbol{\vartheta}}_{M_0})\} = \omega_n^h + o_p(1)$, where $\hat{\boldsymbol{\psi}}_{h,\tau}$ is the local penalized MLE as defined (62) and ω_n^h is defined similarly to $C_n(\sqrt{n}\boldsymbol{t}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\lambda}}, \alpha))$ in (58) but replacing $(\boldsymbol{t}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\lambda}}, \alpha), \boldsymbol{S}_n, \boldsymbol{G}_n, \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}})$ with $(\boldsymbol{t}_{\boldsymbol{\lambda}}^h(\hat{\boldsymbol{\lambda}}^h, \tau), \boldsymbol{S}_n, \boldsymbol{G}_n^h, \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}^h)$ defined in the proof of Proposition 7.

First, we show that $M_n^{h(1)}(\tau_0) = 2\{PL_n^h(\boldsymbol{\vartheta}_{M_0+1}^{h(1)}(\tau_0), \tau_0) - L_{0,n}(\hat{\boldsymbol{\vartheta}}_{M_0})\} = \omega_n^h + p(\tau_0) + o_p(1)$. Define $\boldsymbol{\vartheta}_{M_0+1}^{h*}(\tau_0)$ by the value of $\boldsymbol{\vartheta}_{M_0+1}$ in $\Theta_{\boldsymbol{\vartheta}_{M_0+1}}^h(\tau_0) := \{\boldsymbol{\vartheta} \in \boldsymbol{\Psi}_h^* : \alpha_h/(\alpha_h + \alpha_{h+1}) = \tau_0\}$. Because $\boldsymbol{\vartheta}_{M_0+1}^{h*}(\tau_0)$ is the only value of $\boldsymbol{\vartheta}_{M_0+1}$ that yields the true density if $\boldsymbol{\vartheta}_{M_0+1} \in \boldsymbol{\Psi}_h^*$ in (20) and $\alpha_h/(\alpha_h + \alpha_{h+1}) = \tau_0$, $\boldsymbol{\vartheta}_{M_0+1}^{h(1)}(\tau_0)$ equals a reparameterized penalized local MLE in the neighborhood of $\boldsymbol{\vartheta}_{M_0+1}^{h*}(\tau_0)$, and $\boldsymbol{\vartheta}_{M_0+1}^{h(1)}(\tau_0) - \boldsymbol{\vartheta}_{M_0+1}^{h*}(\tau_0) = o_p(1)$ holds in view of Proposition 6. Furthermore, by the consistency of $\sigma_j^{h(1)}$ and $a_n = O(1)$, we have $\tilde{p}(\boldsymbol{\vartheta}_{M_0+1}^{h(1)}(\tau_0)) \xrightarrow{p} 0$. Therefore, $M_n^{h(1)}(\tau_0) = \omega_n^h + p(\tau_0) + o_p(1)$ follows from repeating the proof of Proposition 7. Finally, $EM_n^{(1)} \xrightarrow{d} \max_{h=1}^{M_0}\{\omega_n^h\}$ holds because $\{0.5\} \in \mathcal{T}$ and $p(0.5) = 0$.

We proceed to show that $M_n^{h(K)}(\tau_0) = \omega_n^h + p(\tau_0) + o_p(1)$ for any finite K. Because a generalized EM step never decreases likelihood (Dempster et al., 1977), we have

$$PL_n(\boldsymbol{\vartheta}_{M_0+1}^{h(K)}(\tau_0), \tau^{h(K)}(\tau_0)) > PL_n(\boldsymbol{\vartheta}_{M_0+1}^{h(1)}(\tau_0), \tau^{h(1)}(\tau_0)). \tag{64}$$

Therefore, it follows from Theorem 1 of Chen and Tan (2009), Lemma 4 in Appendix B, and induction that $\boldsymbol{\vartheta}_{M_0+1}^{h(K)}(\tau_0) - \boldsymbol{\vartheta}_{M_0+1}^{h*} = o_p(1)$ for any finite K. Let $\tilde{\boldsymbol{\vartheta}}_{M_0+1}^h$ be the maximizer of $PL_{M_0+1}(\boldsymbol{\vartheta}_{M_0+1}, \tau^{h(K)}(\tau_0))$ under the constraint of $\alpha_h/(\alpha_h + \alpha_{h+1}) = \tau^{h(K)}(\tau_0)$ in an arbitrary small neighbourhood of $\boldsymbol{\vartheta}_{M_0+1}^{h*}(\tau^{(K)})$. Then, $2\{PL_n^h(\tilde{\boldsymbol{\vartheta}}_{M_0+1}^h, \tau^{h(K)}(\tau_0)) - L_{0,n}(\hat{\boldsymbol{\vartheta}}_{M_0})\} = \omega_n^h + p(\tau_0) + o_p(1)$ holds from the definition of $\tilde{\boldsymbol{\vartheta}}_{M_0+1}^h$ and $\tilde{p}(\tilde{\boldsymbol{\vartheta}}_{M_0+1}^h) \xrightarrow{p} 0$ by repeating the proof

of Proposition 7. It also follows from the consistency of $\boldsymbol{\vartheta}_{M_0+1}^{h(K)}(\tau_0)$ that $PL_n(\tilde{\boldsymbol{\vartheta}}_n^h, \tau^{h(K)}(\tau_0)) \geq PL_n(\boldsymbol{\vartheta}_{M_0+1}^{h(K)}(\tau_0), \tau^{h(K)}(\tau_0)) + o_p(1)$. Therefore, in view of (64), we have

$$PL_n(\tilde{\boldsymbol{\vartheta}}_n^h, \tau^{h(K)}(\tau_0)) \geq PL_n(\boldsymbol{\vartheta}_{M_0+1}^{h(K)}(\tau_0), \tau^{h(K)}(\tau_0)) + o_p(1) \geq PL_n(\boldsymbol{\vartheta}_{M_0+1}^{h(1)}(\tau_0), \tau^{h(1)}(\tau_0)). \quad (65)$$

Finally, because $2\{PL_n(\tilde{\boldsymbol{\vartheta}}_{M_0+1}^h, \tau^{h(K)}(\tau_0)) - L_{0,n}(\hat{\boldsymbol{\vartheta}}_{M_0})\} = \omega_n^h + p(\tau_0) + o_p(1)$ and $2\{PL_n(\boldsymbol{\vartheta}_{M_0+1}^{h(1)}(\tau_0), \tau^{h(1)}(\tau_0)) - L_{0,n}(\hat{\boldsymbol{\vartheta}}_{M_0})\} = \omega_n^h + p(\tau_0) + o_p(1)$, it follows from (65) that $M_n^{h(K)}(\tau_0) = 2\{PL_n(\boldsymbol{\vartheta}_{M_0+1}^{h(K)}(\tau_0)) - L_{0,n}(\hat{\boldsymbol{\vartheta}}_{M_0})\} = \omega_n^h + p(\tau_0) + o_p(1)$ holds for all h. The stated result then follows from the definition of $EM_n^{(K)}$ and $\{0.5\} \in \mathcal{T}$. □

Proof of Proposition 9. Let $\boldsymbol{\psi}_n = ((\boldsymbol{\nu}^*)^\top, \boldsymbol{\lambda}_n^\top)^\top$ be the value of $\boldsymbol{\psi}$ under $H_{1,n} : \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_{2,n}$ and let $\boldsymbol{h} = (\mathbf{0}^\top, \boldsymbol{h}_{\boldsymbol{\lambda}}^\top)^\top$, where $\boldsymbol{h}_{\boldsymbol{\lambda}}$ is defined by (31). Let $\mathbb{P}_{\boldsymbol{\vartheta}}$ be the probability measure on $\{\boldsymbol{W}_i\}_{i=1}^n$ under $\boldsymbol{\vartheta}$. Denote the log-likelihood ratio of $\mathbb{P}_{\boldsymbol{\vartheta}_{2,n}}$ to $\mathbb{P}_{\boldsymbol{\vartheta}_2^*}$ by $\log\left(\frac{d\mathbb{P}_{\boldsymbol{\vartheta}_n}}{d\mathbb{P}_{\boldsymbol{\vartheta}^*}}\right) = L_n(\boldsymbol{\psi}_n, \alpha^*) - L_n(\boldsymbol{\psi}^*, \alpha^*)$. Then, it follows from (12) and Proposition 3 that

$$\log \frac{d\mathbb{P}_{\boldsymbol{\vartheta}_{2,n}}}{d\mathbb{P}_{\boldsymbol{\vartheta}_2^*}} = \boldsymbol{h}\boldsymbol{S}_n - \boldsymbol{h}^\top \boldsymbol{\mathcal{I}} \boldsymbol{h}/2 + o_p(1) \quad \text{under } \mathbb{P}_{\boldsymbol{\vartheta}_2^*}. \quad (66)$$

Furthermore, because $\boldsymbol{S}_n \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\mathcal{I}})$ under $\mathbb{P}_{\boldsymbol{\vartheta}_2^*}$, $\frac{d\mathbb{P}_{\boldsymbol{\vartheta}_n}}{d\mathbb{P}_{\boldsymbol{\vartheta}^*}}$ converges in distribution under $\mathbb{P}_{\boldsymbol{\vartheta}_2^*}$ to $\exp(N(\mu, \sigma^2))$ with $\mu = -(1/2)\boldsymbol{h}^\top \boldsymbol{\mathcal{I}} \boldsymbol{h}$ and $\sigma^2 = \boldsymbol{h}^\top \boldsymbol{\mathcal{I}} \boldsymbol{h}$ so that $E(\exp(N(\mu, \sigma^2)) = 1$. Consequently, $\mathbb{P}_{\boldsymbol{\vartheta}_{2,n}}$ is mutually contiguous with respect to $\mathbb{P}_{\boldsymbol{\vartheta}_2^*}$ from Le Cam's First Lemma (see, e.g., Corollary 12.3.1 of Lehmann and Romano, 2005) and, in view of (66), we have

$$\begin{pmatrix} \boldsymbol{S}_n \\ \log \frac{d\mathbb{P}_{\boldsymbol{\vartheta}_{2,n}}}{d\mathbb{P}_{\boldsymbol{\vartheta}_2^*}} \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} \mathbf{0} \\ -(1/2)\boldsymbol{h}^\top \boldsymbol{\mathcal{I}} \boldsymbol{h} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\mathcal{I}} & \boldsymbol{\mathcal{I}}\boldsymbol{h} \\ \boldsymbol{h}^\top \boldsymbol{\mathcal{I}} & \boldsymbol{h}^\top \boldsymbol{\mathcal{I}} \boldsymbol{h} \end{pmatrix} \right) \quad \text{under } \mathbb{P}_{\boldsymbol{\vartheta}_2^*}.$$

and

$$\boldsymbol{S}_n \xrightarrow{d} N(\boldsymbol{\mathcal{I}}\boldsymbol{h}, \boldsymbol{\mathcal{I}}) \quad \text{under } \mathbb{P}_{\boldsymbol{\vartheta}_{2,n}}$$

from Le Cam's Third Lemma (see, e.g., 12.3.2 of Lehmann and Romano, 2005). Therefore, the proof of Proposition 4 goes through under $\mathbb{P}_{\boldsymbol{\vartheta}_{2,n}}$ if we replace $\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}n} \xrightarrow{d} \boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ with $\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}n} \xrightarrow{d} \boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}} + (\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}} - \boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}\boldsymbol{\lambda}}\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}}^{-1}\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}\boldsymbol{\lambda}})\boldsymbol{h}_{\boldsymbol{\lambda}} = \boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\eta}} + \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}\boldsymbol{h}_{\boldsymbol{\lambda}}$, and the stated result follows. □

Proof of Proposition 10. We provide a proof for $\hat{M}_{\text{PLR}}$. The consistency proof for $\hat{M}_{\text{EM}}$ is similar. We first prove that, when $M < M_0$, $\Pr(PLR_n(M) > \hat{c}_{1-q_n}^M) \to 1$ as $n \to \infty$. Let $\tilde{Q}_n^M(\boldsymbol{\vartheta}_M) := Q_n^M(\boldsymbol{\vartheta}_M) + n^{-1}\tilde{p}_n(\boldsymbol{\vartheta}_M)$ for $M \geq 2$. By Assumptions 3(c) and 4(b), $n^{-1}\tilde{p}_n(\boldsymbol{\vartheta}_M) \xrightarrow{p} 0$ uniformly over $\Theta_{\boldsymbol{\vartheta}_M}$. Then, it follows from Lemma 2.4 of Newey and McFadden (1994) that

$$\sup_{\boldsymbol{\vartheta}_M \in \Theta_{\boldsymbol{\vartheta}_M}} |\tilde{Q}_n^M(\boldsymbol{\vartheta}_M) - Q^M(\boldsymbol{\vartheta}_M)| = o_p(1), \quad (67)$$

and Assumption 4(a)(b) and the standard consistency proof (e.g., Theorem 2.1 of Newey and McFadden, 1994) give $\hat{\boldsymbol{\vartheta}}_M \xrightarrow{p} \boldsymbol{\vartheta}^*_M$ for $M < M_0$. Furthermore, by Assumption 3(c), $n^{-1/4}\nabla \tilde{p}_n(\boldsymbol{\vartheta}_M) = o_p(1)$ uniformly over $\Theta_{\boldsymbol{\vartheta}_M}$, and it follows from the argument in Theorem 3.2 of White (1982) that

$$\sqrt{n}(\hat{\boldsymbol{\vartheta}}_M - \boldsymbol{\vartheta}^*_M) \xrightarrow{p} N(0, A^M(\boldsymbol{\vartheta}^*_M)^{-1} B^M(\boldsymbol{\vartheta}^*_M) A^M(\boldsymbol{\vartheta}^*_M)^{-1}). \tag{68}$$

Then, from (67), (68), and the mean value expansion, we have $\tilde{Q}^M_n(\hat{\boldsymbol{\vartheta}}_M) - Q^M(\boldsymbol{\vartheta}^*_M) = O_p(n^{-1/2})$, and

$$\frac{PLR_n(M)}{n} := \tilde{Q}^{M+1}_n(\hat{\boldsymbol{\vartheta}}_{M+1}) - \tilde{Q}^M_n(\hat{\boldsymbol{\vartheta}}_M) = Q^{M+1}(\boldsymbol{\vartheta}^*_{M+1}) - Q^M(\boldsymbol{\vartheta}^*_M) + o_p(1).$$

Because $Q^{M+1}(\boldsymbol{\vartheta}^*_{M+1}) - Q^M(\boldsymbol{\vartheta}^*_M) > 0$ by Assumption 4(f), $PLR_n(M)/n \to \infty$ as $n \to \infty$. By Lemma 2, $-n^{-1}\ln q_n = o(1)$ and $\hat{c}^M_{1-q_n} - c^M_{1-q_n} = o_p(1)$ implies that $n^{-1}\hat{c}^M_{1-q_n} = o_p(1)$. Therefore, when $M < M_0$, we have $\Pr(PLR_n(M) > \hat{c}^M_{1-q_n}) = \Pr(PLR_n(M)/n > \hat{c}^M_{1-q_n}/n) \to 1$ as $n \to \infty$.

When $M = M_0$, because $PLR_n(M_0) = O_p(1)$ by Proposition 7 and $\hat{c}^M_{1-q_n} \to \infty$ by $q_n = o(1)$, $\Pr(PLR_n(M_0) > \hat{c}^{M_0}_{1-q_n}) \to 0$ as $n \to \infty$. □

B Auxiliary results and their proofs

B.1 Lemmas

Lemma 1. *For any $M < \infty$, $\Pr\left(-\log n + \ell(\boldsymbol{W}_{i^*}; \bar{Y}_{i^*}, s^2_{i^*}) < M\right) \to 0$ as $n \to \infty$.*

Proof of Lemma 1. Because $\sum_{t=1}^T \frac{(Y_{it}-\mu)^2}{s^2_{i^*}} = T - 1$ when $i = i^*$, we have

$$\begin{aligned} -\log n + \ell(\boldsymbol{W}_{i^*}; \bar{Y}_{i^*}, s^2_{i^*}) &= -\log n - \frac{T}{2}\log s^2_{i^*} - \frac{T}{2}\log(2\pi) - \frac{T-1}{2} \\ &= -\log\left(Cn(s^2_{i^*})^{T/2}\right), \end{aligned} \tag{69}$$

for some positive constant C.

Therefore, to prove the stated result, it suffices to show that, for any $\epsilon > 0$, $\Pr(n(s^2_{i^*})^{T/2} > \epsilon) \to 0$ as $n \to \infty$. Given the property of the first-order statistic, the distribution of $s^2_{i^*}$ is given by $1 - [1 - F_{T-1}(s)]^n$, where $F_{T-1}(s)$ is the cumulative distribution function for chi-squared variables of degree $T - 1$. It follows that

$$\Pr(n(s^2_{i^*})^{T/2} > \epsilon) = [1 - F_{T-1}((\epsilon/n)^{2/T})]^n.$$

When $T = 3$, $1 - F_{T-1}(s) = e^{-s/2}$ and $\Pr(n(s^2_{i^*})^{T/2} > \epsilon) = e^{-Cn^{1/3}}$ for some positive constant C and, therefore, $\Pr(n(s^2_{i^*})^{T/2} > \epsilon) \to 0$ as $n \to \infty$, and the stated result follows.

For general $T \geq 2$, write

$$\left[1 - F_{T-1}\big((\epsilon/n)^{2/T}\big)\right]^n = \left\{\left[1 - F_{T-1}\big((\epsilon/n)^{2/T}\big)\right]^{\frac{1}{F_{T-1}((\epsilon/n)^{2/T})}}\right\}^{nF_{T-1}((\epsilon/n)^{2/T})}. \tag{70}$$

Then, because $(1-F)^{\frac{1}{F}} \to \frac{1}{e}$ when $F \to 0$, the stated result follows from (70) if we can show

$$\frac{F_{T-1}((\epsilon x)^{2/T})}{x} \to \infty \text{ as } x \to 0$$

for $x = 1/n$. By applying L'Hôpital's rule, we have

$$\lim_{x\to 0} \frac{F_{T-1}((\epsilon x)^{2T})}{x} = \lim_{x\to 0} f_{T-1}((\epsilon x)^{2/T})\epsilon^{2T} x^{2/T-1},$$

where f_k is the PDF of χ-square distribution with k degrees of freedom. Note that $f_{T-1}((\epsilon x)^{2/T}) = \frac{1}{2^{(T-1)/2}\Gamma((T-1)/2)}((\epsilon x)^{2/T})^{(T-1)/2-1} e^{-((\epsilon x)^{2/T})/2}$, then

$$\lim_{x\to 0} f_{T-1}((\epsilon x)^{2/T}) x^{2/T-1} = \lim_{x\to 0} C_{T,\epsilon} e^{-((\epsilon x)^{2/T})/2} x^{-\frac{1}{T}} = \infty,$$

where $C_{T,\epsilon} = \frac{\epsilon^{(T-1)/T}}{2^{(T-1)/2}\Gamma((T-1)/2)}$, because $e^{-((\epsilon x)^{2/T})/2} \to 1$ and $x^{-\frac{1}{T}} \to \infty$ as $x \to 0$ for any finite $T \geq 2$. Therefore, $\lim_{x\to 0} \frac{F_{T-1}((\epsilon x)^{2T})}{x} = \infty$ and the stated result for $T \geq 2$ follows from (70).

□

Lemma 2. *Suppose that assumptions in Proposition 10 hold. If $-n^{-1}\ln q_n = o(1)$, then $n^{-1}c^M_{1-q_n} = o(1)$.*

Proof. For brevity of notation, write $c_n = c^M_{1-q_n}$. By Theorem 2.1 of Foutz and Srivastava (1977), $PLR_n(M) \xrightarrow{d} \sum_{j=1}^K b_j \chi^2_j$ for $0 < b_j < \infty$ and K is finite, where χ^2_1, ..., χ^2_K are independent chi-square random variables with one degree of freedom. Then, we have

$$q_n = \Pr\left(\sum_{j=1}^K b_j \chi^2_j \geq c_n\right) \leq \sum_{j=1}^K \Pr\left(\chi^2_j \geq \frac{c_n}{b_j}\right) \leq \frac{K}{\sqrt{1-2t}} \exp\left(-t\frac{c_n}{b^*}\right) \quad \text{for } 0 < t < \frac{1}{2}$$

with $b^* = \arg\max\{b_1, ..., b_K\}$, where the last inequality follows from a Chernoff bound: $\Pr\left(\chi^2_j \geq \frac{c_n}{b^*}\right) \leq \frac{\mathbb{E}[\exp(t(\chi^2_j - 1))]}{\exp(t(\frac{c_n}{b^*} - 1))} = \frac{1}{\sqrt{1-2t}} \exp\left(-t\frac{c_n}{b^*}\right)$ for $0 < t < \frac{1}{2}$. Therefore, $-\frac{\ln q_n}{n} \geq -\frac{1}{n}\ln\left(\frac{K}{\sqrt{1-2t}}\right) + \frac{1}{2b^*_j}\frac{c_n}{n}$, and the stated result follows.

□

Lemma 3. *Suppose that $g(\boldsymbol{w}; \boldsymbol{\psi}, \alpha)$ is defined as (6), where $\boldsymbol{\psi} = (\boldsymbol{\eta}^\top, \boldsymbol{\lambda}^\top)^\top$. Let g^*, ∇g^*, $\nabla \log g^*$ denote $g(\boldsymbol{W}; \boldsymbol{\psi}, \alpha)$, $\nabla g(\boldsymbol{W}; \boldsymbol{\psi}, \alpha)$, and $\nabla \log g(\boldsymbol{W}; \boldsymbol{\psi}, \alpha)$ evaluated at $(\boldsymbol{\psi}, \alpha)$, respectively. Let ∇f^* denote $\nabla f(\boldsymbol{W}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*)$. The following statements hold.*

(a) For $l = 0, 1, \ldots, \nabla_{(\boldsymbol{\lambda}\otimes\boldsymbol{\eta}^{\otimes l})^\top} g^* = 0$*;*

(b) $\nabla_{(\boldsymbol{\lambda}^{\otimes 2})^\top} g^* = \alpha(1-\alpha)\nabla_{(\boldsymbol{\theta}^{\otimes 2})^\top} f^*$*;*

(c) $\mathbb{E}[\nabla_{\lambda_i\lambda_j} \log g^*] = 0$*,* $\mathbb{E}[\nabla_{\lambda_i\lambda_j\lambda_k} \log g^*] = 0$*, and* $\mathbb{E}[\nabla_{\eta\lambda_i\lambda_j} \log g^*] = -\mathbb{E}[\nabla_\eta \log g^* \nabla_{\lambda_i\lambda_j} \log g^*]$*;*

(d) $\mathbb{E}[\nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell} \log g^*] = -\mathbb{E}[\nabla_{\lambda_i\lambda_j} \log g^* \nabla_{\lambda_k\lambda_\ell} \log g^* + \nabla_{\lambda_i\lambda_k} \log g^* \nabla_{\lambda_j\lambda_\ell} \log g^* + \nabla_{\lambda_i\lambda_\ell} \log g^* \nabla_{\lambda_j\lambda_k} \log g^*]$.

Proof of Lemma 3. Recall that

$$g(\boldsymbol{w}; \boldsymbol{\psi}, \alpha) = \alpha f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu} + (1-\alpha)\boldsymbol{\lambda}) + (1-\alpha) f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu} - \alpha\boldsymbol{\lambda}).$$

First we show that for $l = 0$ holds for (a), $\nabla_{\boldsymbol{\lambda}} g^* = \alpha(1-\alpha)\nabla_{\boldsymbol{\theta}} f^* - \alpha(1-\alpha)\nabla_{\boldsymbol{\theta}} f^* = 0$. For $l > 0$, by Fubini's Theorem, we have

$$\begin{aligned}
\nabla_{(\boldsymbol{\lambda}^{\otimes 2})^\top} g &= \nabla_{\boldsymbol{\lambda}}\Big(\alpha\nabla_{(\boldsymbol{\gamma},\boldsymbol{\theta})^{\otimes l}} f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu} + (1-\alpha)\boldsymbol{\lambda}) + (1-\alpha)\nabla_{(\boldsymbol{\gamma},\boldsymbol{\theta})^{\otimes l}} f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu} - \alpha\boldsymbol{\lambda})\Big|_{\boldsymbol{\nu}=\boldsymbol{\theta}^*, \boldsymbol{\lambda}=\mathbf{0}}\Big) \\
&= \Big(\alpha(1-\alpha)\nabla_{(\boldsymbol{\gamma}^{\otimes l}, \boldsymbol{\theta}^{\otimes l+1})} f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu} + (1-\alpha)\boldsymbol{\lambda}) - \alpha(1-\alpha)\nabla_{(\boldsymbol{\gamma}^{\otimes l}, \boldsymbol{\theta}^{\otimes l+1})} f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu} - \alpha\boldsymbol{\lambda})\Big|_{\boldsymbol{\nu}=\boldsymbol{\theta}^*, \boldsymbol{\lambda}=\mathbf{0}}\Big) \\
&= 0.
\end{aligned}$$

To show part (b), note that

$$\begin{aligned}
\nabla_{(\boldsymbol{\lambda}^{\otimes 2})^\top} g &= \nabla_{\boldsymbol{\lambda}}\Big(\alpha(1-\alpha)\nabla_{\boldsymbol{\lambda}^\top} f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu} - \alpha(1-\alpha)\boldsymbol{\lambda}) + (1-\alpha)\nabla_{\boldsymbol{\lambda}^\top} f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu} - \alpha\boldsymbol{\lambda})\Big) \\
&= \alpha(1-\alpha)^2\nabla_{(\boldsymbol{\lambda}^{\otimes 2})^\top} f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu} + \alpha^2(1-\alpha)\boldsymbol{\lambda}) + (1-\alpha)\nabla_{(\boldsymbol{\lambda}^{\otimes 2})^\top} f(\boldsymbol{w}; \boldsymbol{\gamma}, \boldsymbol{\nu} - \alpha\boldsymbol{\lambda})\big|_{\boldsymbol{\nu}=\boldsymbol{\theta}^*, \boldsymbol{\lambda}=\mathbf{0}} \\
&= \nabla_{(\boldsymbol{\lambda}^{\otimes 2})^\top} f^*.
\end{aligned}$$

For parts (c) and (d), observe that $\int \nabla_{\lambda_i} \log g(\boldsymbol{w}; \psi, \alpha) g(\boldsymbol{w}; \psi, \alpha) dx = 0$ holds for any ψ in the interior of Θ_ψ, and differentiating this equation w.r.t. λ_j gives

$$\int \{\nabla_{\lambda_i\lambda_j} \log g(\boldsymbol{w}; \psi, \alpha) + \nabla_{\lambda_i} \log g(\boldsymbol{w}; \psi, \alpha) \nabla_{\lambda_j} \log g(\boldsymbol{w}; \psi, \alpha)\} g(\boldsymbol{w}; \psi, \alpha) dx = 0. \tag{71}$$

Evaluating (71) at $\psi = \psi^*$ in conjunction with part (a) gives the first equation in part (c). Differentiating (71) w.r.t. λ_k or η and evaluating at $\psi = \psi^*$ give the latter two equations in part (c). Part (d) follows from differentiating (71) w.r.t. λ_k and λ_ℓ and evaluating at $\psi = \psi^*$ in conjunction with parts (a)(c).

□

Lemma 4. *Suppose that the assumptions of Proposition 8 hold. If* $\boldsymbol{\vartheta}^{h(K)}_{M_0+1}(\tau_0) - \boldsymbol{\vartheta}^{h*}_{M_0+1}(\tau_0) = o_p(1)$ *and* $\tau^{(K)} - \tau_0 = o_p(1)$*, then (a)* $\alpha_m^{(K+1)}/[\alpha_h^{(K+1)} + \alpha_{h+1}^{(K+1)}] - \tau_0 = o_p(1)$ *and (b)* $\tau^{(K+1)} - \tau_0 = o_p(1)$.

Proof. The proof is similar to the proof of Lemma 3 of Chen and Li (2009) and Lemma 10 in Appendix D of Kasahara and Shimotsu (2019). We suppress (τ_0) from $\boldsymbol{\vartheta}^{h(K)}_{M_0+1}(\tau_0)$ and $\boldsymbol{\vartheta}^{h*}_{M_0+1}(\tau_0)$. We suppress $\boldsymbol{Z}$ for brevity. Let $f_i(\boldsymbol{\gamma}, \boldsymbol{\theta}_j)$ and $f_i(\boldsymbol{\vartheta}_{M_0+1})$ denote $f(\boldsymbol{W}_i; \boldsymbol{\gamma}, \boldsymbol{\theta}_j)$ in (2) and $f_{M_0+1}(\boldsymbol{W}_i; \boldsymbol{\vartheta}_{M_0+1})$ in (18), respectively. Applying a Taylor expansion to $\alpha_h^{(K+1)} = n^{-1}\sum_{i=1}^n w_{ih}^{(K)}$ and using $\boldsymbol{\vartheta}^{h(K)}_{M_0+1} - \boldsymbol{\vartheta}^{h*}_{M_0+1} = o_p(1)$, we obtain

$$\begin{aligned}\alpha_m^{(K+1)} &= \frac{1}{n}\sum_{i=1}^n \frac{\tau^{(K)}(\alpha_h^{(K)} + \alpha_{h+1}^{(K)}) f_i(\boldsymbol{\gamma}^{(K)}, \boldsymbol{\theta}_h^{(K)})}{f_i(\boldsymbol{\vartheta}^{h(K)}_{M_0+1})} \\ &= \frac{1}{n}\sum_{i=1}^n \frac{\tau_0\alpha_h^* f_i(\boldsymbol{\gamma}^*, \boldsymbol{\theta}_h^*)}{f_i(\boldsymbol{\vartheta}^{h*}_{M_0+1})} + o_p(1) = \tau_0\alpha_h^* + o_p(1),\end{aligned}$$

where the last equality follows from $\mathbb{E}[f_i(\boldsymbol{\gamma}^*, \boldsymbol{\theta}_h^*)/f_i(\boldsymbol{\vartheta}^{h*}_{M_0+1})] = 1$ and the law of large numbers. A similar argument gives $\alpha_{h+1}^{(K+1)} = (1-\tau_0)\alpha_h^* + o_p(1)$, and part (a) follows.

For part (b), define $H(\tau) := \sum_{i=1}^n w_{ih}^{(K)}\log(\tau) + \sum_{i=1}^n w_{i,h+1}^{(K)}\log(1-\tau) = n\alpha_h^{(K+1)}\log(\tau) + n\alpha_{h+1}^{(K+1)}$, then $\tau^{(K+1)}$ maximizes $H(\tau) + p(\tau)$. $H(\tau)$ is maximized at $\tilde{\tau} := \alpha_h^{(K+1)}/(\alpha_h^{(K+1)} + \alpha_{h+1}^{(K+1)}) = (\tau_0\alpha_h^* + o_p(1))/(\tau_0\alpha_h^* + (1-\tau_0)\alpha_h^* + o_p(1)) = \tau_0 + o_p(1)$. Observe that, with $\bar{\tau}$ between $\tau^{(K+1)}$ and $\tilde{\tau}$,

$$p(\tilde{\tau}) \le p(\tilde{\tau}) - p(\tau^{(K+1)}) \le H(\tau^{(K+1)}) - H(\tilde{\tau}) = H''(\bar{\tau})(\tau^{(K+1)} - \tilde{\tau})^2, \tag{72}$$

where the first inequality follows from $p(\tau) \le 0$, the second inequality holds because $\tau^{(K+1)}$ maximizes $H(\tau) + p(\tau)$, and the last equality follows from expanding $H(\tau^{(K+1)})$ twice around $\tilde{\tau}$ and noting that $H'(\tilde{\tau}) = 0$ because $\tilde{\tau}$ maximizes $H(\tau)$. Note that $H''(\tau) = -n \times \left\{ \frac{\alpha_h^{(K+1)}}{\tau^2} + \frac{\alpha_{h+1}^{(K+1)}}{(1-\tau)^2} \right\} < 0$ and $\inf_\tau H''(\tau) \ge -n(\alpha_h^{(K+1)} + \alpha_{h+1}^{(K+1)})$. Therefore, in view of $\tilde{\tau} - \tau_0 = o_p(1)$ and (72), we have $(\tau^{(K+1)} - \tilde{\tau})^2 \le p(\tilde{\tau})/H''(\bar{\tau}) = O_p(n^{-1})$, and part (b) holds.

□

B.2 Score function for testing $H_0 : m = 1$ against $H_A : m = 2$

$H^j(\cdot)$ is defined as the j-th order Hermite polynomial. $H^1(t) = t$, $H^2(t) = t^2 - 1$, $H^3(t) = t^3 - 3t$, and $H^4(t) = t^4 - 6t^2 + 3$. As shown in Kasahara and Shimotsu (2015) supplement material, the derivative of $\{\frac{1}{\sigma}\phi(\frac{t}{\sigma})\}$ is

$$\frac{\nabla_{\mu^m}\nabla_{(\sigma^2)^\ell}\{\frac{1}{\sigma}\phi(\frac{t}{\sigma})\}}{\{\frac{1}{\sigma}\phi(\frac{t}{\sigma})\}} = \left(\frac{1}{2}\right)^\ell \left(\frac{1}{\sigma}\right)^{m+2\ell} H^{m+2\ell}\left(\frac{t}{\sigma}\right).$$

Let

$$f^* = f(\boldsymbol{w};\gamma^*,\theta^*), \nabla f^* = \nabla f(\boldsymbol{w};\gamma^*,\theta^*), H_{i,t}^{j*} = \frac{1}{\sigma^* j!} H^j\left(\frac{y_{it} - \boldsymbol{x}_{it}^\top \boldsymbol{\beta}^* - \boldsymbol{z}_{it}^\top \boldsymbol{\gamma}^* - \mu^*}{\sigma^*}\right), \tag{73}$$

then the first-order derivatives of the density functions are

$$\nabla_\mu f^* = f^* \sum_{t=1}^T \frac{1}{\sigma} H_{i,t}^{1*}; \nabla_{\sigma^2} f^* = f^* \sum_{t=1}^T \frac{1}{2}\frac{1}{\sigma^2} H_{it}^{2*};$$
$$\nabla_{\boldsymbol{\beta}} f^* = f^* \sum_{t=1}^T \frac{1}{\sigma} H_{it}^{1*} \boldsymbol{x}_{it}; \nabla_{\boldsymbol{\gamma}} f^* = f^* \sum_{t=1}^T \frac{1}{\sigma} H_{it}^{1*} \boldsymbol{z}_{it}.$$

The score function defined in (9) is then written in terms of the Hermite polynomials:

$$\boldsymbol{s}_{\boldsymbol{\eta}} = \begin{pmatrix} s_\mu \\ s_\sigma \\ \boldsymbol{s}_{\boldsymbol{\beta}} \\ \boldsymbol{s}_{\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^T H_{i,t}^{1*} \\ \sum_{t=1}^T H_{i,t}^{2*} \\ \sum_{t=1}^T H_{i,t}^{1*} \boldsymbol{x}_{it} \\ \sum_{t=1}^T H_{i,t}^{1*} \boldsymbol{z}_{it} \end{pmatrix}, \qquad \boldsymbol{s}_{\boldsymbol{\lambda\lambda}} = \begin{pmatrix} s_{\lambda_\mu\lambda_\mu} \\ s_{\lambda_\mu\lambda_\sigma} \\ s_{\lambda_\sigma\lambda_\sigma} \\ \boldsymbol{s}_{\lambda_\mu\lambda_{\boldsymbol{\beta}}} \\ \boldsymbol{s}_{\lambda_\sigma\lambda_{\boldsymbol{\beta}}} \\ \boldsymbol{s}_{\lambda_{\boldsymbol{\beta}}\lambda_{\boldsymbol{\beta}}} \end{pmatrix}, \tag{74}$$

where

$$\begin{pmatrix} s_{\lambda_{\mu\mu}} \\ s_{\lambda_{\mu\sigma}} \\ s_{\lambda_{\sigma\sigma}} \\ \boldsymbol{s}_{\lambda_{\mu\boldsymbol{\beta}}} \\ \boldsymbol{s}_{\lambda_{\sigma\boldsymbol{\beta}}} \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^T H_{i,t}^{2*} + \frac{1}{2}\sum_{t=1}^T \sum_{s\neq t} H_{1,i,t}^{1*} H_{i,s}^{1*} \\ 3\sum_{t=1}^T H_{i,t}^{3*} + \sum_{t=1}^T \sum_{s\neq t} H_{i,t}^{1*} H_{i,s}^{2*} \\ 3\sum_{t=1}^T H_{i,t}^{4*} + \frac{1}{2}\sum_{t=1}^T \sum_{s\neq t} H_{i,t}^{2*} H_{i,t}^{2*} \\ 2\sum_{t=1}^T H_{i,t}^{2*} \boldsymbol{x}_{it} + \sum_{t=1}^T \sum_{s\neq t} H_{i,t}^{1*} H_{i,s}^{1*} \boldsymbol{x}_{it} \\ 3\sum_{t=1}^T H_{i,t}^{3*} \boldsymbol{x}_{it} + 2\sum_{t=1}^T \sum_{s\neq t} H_{i,t}^{1*} H_{i,s}^{2*} \boldsymbol{x}_{it} \end{pmatrix}, \quad \text{and}$$

$$\boldsymbol{s}_{\lambda_{\boldsymbol{\beta}\boldsymbol{\beta}}} = \begin{pmatrix} \sum_{t=1}^T H_{i,t}^{2*} x_{it,1}^2 + \frac{1}{2}\sum_{t=1}^T \sum_{s\neq t} H_{i,t}^{1*} x_{it,1} H_{i,s}^{1*} x_{is,1} \\ \vdots \\ \sum_{t=1}^T H_{i,t}^{2*} x_{it,q}^2 + \frac{1}{2}\sum_{t=1}^T \sum_{s\neq t} H_{i,t}^{1*} x_{it,q} H_{i,s}^{1*} x_{is,q} \\ 2\sum_{t=1}^T H_{i,t}^{2*} x_{it,1} x_{it,2} + \sum_{t=1}^T \sum_{s\neq t} H_{i,t}^{1*} x_{it,1} H_{i,s}^{1*} x_{is,2} \\ \vdots \\ 2\sum_{t=1}^T H_{i,t}^{2*} x_{it,1} x_{it,q} + \sum_{t=1}^T \sum_{s\neq t} H_{i,t}^{1*} x_{it,1} H_{i,s}^{1*} x_{is,q} \\ 2\sum_{t=1}^T H_{i,t}^{2*} x_{it,2} x_{it,3} + \sum_{t=1}^T \sum_{s\neq t} H_{i,t}^{1*} x_{it,2} H_{i,s}^{1*} x_{is,3} \\ \vdots \\ 2\sum_{t=1}^T H_{i,t}^{2*} x_{it,q-1} x_{it,q} + \sum_{t=1}^T \sum_{s\neq t} H_{i,t}^{1*} x_{it,q-1} H_{i,s}^{1*} x_{is,q} \end{pmatrix}. \tag{75}$$

When $T = 1$, the score functions are as follow:

$$\boldsymbol{s_\eta} = \begin{pmatrix} s_\mu \\ s_\sigma \\ \boldsymbol{s_\beta} \\ \boldsymbol{s_\gamma} \end{pmatrix} = \begin{pmatrix} H_i^{1*} \\ H_i^{2*} \\ H_i^{1*}\boldsymbol{x}_i \\ H_i^{1*}\boldsymbol{z}_i \end{pmatrix}, \quad \begin{pmatrix} s_{\lambda_{\mu\mu}} \\ s_{\lambda_{\mu\sigma}} \\ s_{\lambda_{\sigma\sigma}} \\ \boldsymbol{s}_{\lambda_{\mu\boldsymbol{\beta}}} \\ \boldsymbol{s}_{\lambda_{\sigma\boldsymbol{\beta}}} \end{pmatrix} = \begin{pmatrix} H_i^{2*} \\ 3H_i^{3*} \\ 3H_i^{4*} \\ 2H_i^{2*}\boldsymbol{x}_i \\ 3H_i^{3*}\boldsymbol{x}_i \end{pmatrix}, \quad \text{and } \boldsymbol{s}_{\lambda_{\boldsymbol{\beta\beta}}} = \begin{pmatrix} H_i^{2*}x_{i,1}^2 \\ \vdots \\ H_i^{2*}x_{i,q}^2 \\ 2H_i^{2*}x_{i,1}x_{i,2} \\ \vdots \\ 2H_i^{2*}x_{i,1}x_{i,q} \\ 2H_i^{2*}x_{i,2}x_{i,3} \\ \vdots \\ 2H_i^{2*}x_{i,q-1}x_{i,q} \end{pmatrix}. \tag{76}$$

Notice that s_σ and $s_{\lambda_{\mu\mu}}$ are perfect collinear and, therefore, the Fisher information matrix associated with the proposed score function is singular under this reparameterization for data with $T = 1$.

B.3 Score function for testing $H_0 : m = M_0$ against $H_A : m = M_0 + 1$

The derivative of the reparameterized density w.r.t λ at ψ_τ^{h*} is zero similar to testing homogeneity case. With the constraint $\pi^{M_0} = 1 - \sum_{j=1}^{M_0-1} \pi^j$. The score functions $s_{\eta i}$'s contain the first-order derivatives w.r.t π's γ and ν at ψ_τ^{h*}:

$$\begin{aligned} \nabla_{\pi^j} l^h(\boldsymbol{w}; \psi_\tau^{h*}, \tau) &= \frac{f(\boldsymbol{w}; \gamma^*, \theta_0^{j*}) - f(\boldsymbol{w}; \gamma^*, \theta_0^{M_0*})}{\sum_{j=1}^{M_0} \alpha_0^{j*} f(\boldsymbol{w}; \gamma^*, \theta_0^{j*})}; \\ \nabla_\gamma l^h(\boldsymbol{w}; \psi_\tau^{h*}, \tau) &= \frac{\sum_{j=1}^{M_0} \alpha_0^{j*} \nabla_\gamma f(\boldsymbol{w}; \gamma^*, \theta_0^{j*})}{\sum_{j=1}^{M_0} \alpha_0^{j*} f(\boldsymbol{w}; \gamma^*, \theta_0^{j*})}; \\ \nabla_\nu l^h(\boldsymbol{w}; \psi_\tau^{h*}, \tau) &= \frac{\nabla_\theta f(\boldsymbol{w}; \gamma^*, \theta_0^{h*})}{\sum_{j=1}^{M_0} \alpha_0^{j*} f(\boldsymbol{w}; \gamma^*, \theta_0^{j*})}. \end{aligned} \tag{77}$$

Define $H_{j,i,t}^{b*}$ as an abridged expression for $\frac{1}{b!}\frac{1}{\sigma_0^*} H^b(\frac{y_{it} - \mu_0^{j*} - x_{it}'\beta_0^{j*} - z_{it}'\gamma^*}{\sigma_0^{j*}})$. Define the weight w_i^{j*} as

$$w_i^{j*} = \frac{\alpha_0^{j*} f(\{\boldsymbol{W}_{it}\}_{t=1}^T; \gamma^*, \theta_0^{j*})}{f_{M_0}(\{\boldsymbol{W}_{it}\}_{t=1}^T; \vartheta_{M_0}^*)}, j = 1, \ldots, M_0,$$

where $f_{M_0}(\{\boldsymbol{W}_{it}\}_{t=1}^T; \vartheta_{M_0}^*)$ is defined by equation (17).

As shown in section B.3, the score functions are:

$$\boldsymbol{s}_{\boldsymbol{\alpha}}(\boldsymbol{w}_i) = \begin{pmatrix} \frac{f(\boldsymbol{w}|\theta_0^{1*}) - f(\boldsymbol{w}|\theta_0^{M_0*})}{\sum_l \alpha_0^{l*} f(\boldsymbol{w}|\theta_0^{l*})} \\ \vdots \\ \frac{f(\boldsymbol{w}|\theta_0^{M_0-1*}) - f(\boldsymbol{w}|\theta_0^{M_0*})}{\sum_l \alpha_0^{l*} f(\{\boldsymbol{W}_{it}^*\}_{t=1}^T|\theta_0^{l*})} \end{pmatrix}, \boldsymbol{s}_{\mu}(\boldsymbol{w}_i) = \begin{pmatrix} w_i^{1*} \sum_{t=1}^T H_{1,i,t}^{1*} \\ \vdots \\ w_i^{M_0*} \sum_{t=1}^T H_{M_0,i,t}^{1*} \end{pmatrix}, \boldsymbol{s}_{\boldsymbol{\beta}}(\boldsymbol{w}_i) = \begin{pmatrix} w_i^{1*} \sum_{t=1}^T H_{1,i,t}^{1*} x_{it} \\ \vdots \\ w_i^{M_0*} \sum_{t=1}^T H_{M_0,i,t}^{1*} x_{it} \end{pmatrix},$$

$$\boldsymbol{s}_{\sigma}(\boldsymbol{w}_i) = \begin{pmatrix} w_i^{1*} \sum_{t=1}^T H_{1,i,t}^{2*} \\ \vdots \\ w_i^{M_0*} \sum_{t=1}^T H_{M_0,i,t}^{2*} \end{pmatrix}, \boldsymbol{s}_{\boldsymbol{\gamma}}(\boldsymbol{w}_i) = \begin{pmatrix} w_i^{1*} \sum_{t=1}^T H_{1,i,t}^{1*} z_{it} \\ \vdots \\ w_i^{M_0*} \sum_{t=1}^T H_{M_0,i,t}^{1*} z_{it} \end{pmatrix}.$$

The score function for $\boldsymbol{s}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^h$ is obtained analogously to $\boldsymbol{s}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}$ by replacing $H_{i,t}^{b*}$ with $H_{h,i,t}^{b*}$ for $b = 1, .., 4$ so that

$$\boldsymbol{s}_{\boldsymbol{\lambda}_{\mu\sigma}}^h(\boldsymbol{w}_i) = w_i^{h*} \begin{pmatrix} \sum_{t=1}^T H_{h,i,t}^{2*} + \frac{1}{2} \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t}^{1*} H_{h,i,s}^{1*} \\ 3 \sum_{t=1}^T H_{h,i,t}^{4*} + \frac{1}{2} \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t}^{2*} H_{h,i,t}^{2*} \\ 3 \sum_{t=1}^T H_{h,i,t}^{3*} + \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t}^{1*} H_{h,i,s}^{2*} \\ 2 \sum_{t=1}^T H_{h,i,t}^{2*} x_{it} + \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t}^{1*} x_{it} H_{h,i,s}^{1*} \\ 3 \sum_{t=1}^T H_{h,i,t}^{3*} x_{it} + 2 \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t}^{1*} x_{it} H_{h,i,s}^{2*} \end{pmatrix},$$

$$\boldsymbol{s}_{\boldsymbol{\lambda}_{\beta}}^h(\boldsymbol{w}_i) = w_i^{h*} \begin{pmatrix} \sum_{t=1}^T H_{h,i,t}^{2*} x_{it,1}^2 + \frac{1}{2} \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t}^{1*} x_{it,1} H_{h,i,s}^{1*} x_{is,1} \\ \vdots \\ \sum_{t=1}^T H_{h,i,t}^{2*} x_{it,q}^2 + \frac{1}{2} \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t}^{1*} x_{it,q} H_{h,i,s}^{1*} x_{is,q} \\ 2 \sum_{t=1}^T H_{h,i,t}^{2*} x_{it,1} x_{it,2} + \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t}^{1*} x_{it,1} H_{h,i,s}^{1*} x_{is,2} \\ \vdots \\ 2 \sum_{t=1}^T H_{h,i,t}^{2*} x_{it,1} x_{it,q} + \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t}^{1*} x_{it,1} H_{h,i,s}^{1*} x_{is,q} \\ 2 \sum_{t=1}^T H_{h,i,t}^{2*} x_{it,2} x_{it,3} + \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t}^{1*} x_{it,2} H_{h,i,s}^{1*} x_{is,3} \\ \vdots \\ 2 \sum_{t=1}^T H_{h,i,t}^{2*} x_{it,q-1} x_{it,q} + \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t}^{1*} x_{it,q-1} H_{h,i,s}^{1*} x_{is,q} \end{pmatrix}.$$

B.4 How to simulate the asymptotic distribution

C Other tables

Table 11: Parameter specification for null models with $M_0 = 1, 2, 3, 4$

	$M_0 = 1$
N	$\{100, 500\}$
T	$\{2, 5, 10\}$
a_n	$(0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4)$
	$M_0 = 2$
N	$\{100, 500\}$
T	$\{2, 5, 10\}$
α	$\{(0.5, 0.5); (0.2, 0.8)\}$
μ	$\{(-1, 1), (-0.5, 0.5), (-0.8, 0.8)\}$
σ	$\{(1, 1), (1.5, 0.75), (0.8, 1.2)\}$
a_n	$(0.01, 0.05, 0.1, 0.2, 0.3, 0.4)$
	$M_0 = 3$
N	$\{100, 500\}$
T	$\{2, 10\}$
α	$\{(1/3, 1/3, 1/3); (0.25, 0.5, 0.25)\}$
μ	$\{(-4, 0, 4); (-4, 0, 5); (-5, 0, 5); (-4, 0, 6); (-5, 0, 6); (-6, 0, 6)\}$
σ	$\{(1, 1, 1); (0.75, 1.5, 0.75)\}$
a_n	$(0.01, 0.05, 0.1, 0.2, 0.3, 0.4)$
	$M_0 = 4$
N	$\{100, 500\}$
T	$\{2, 10\}$
α	$\{(0.25, 0.25, 0.25, 0.25)\}$
μ	$\{(-4, -1, 1, 4); (-5, -1, 1, 5); (-6, -2, 2, 6); (-6, -1, 2, 5); (-5, 0, 2, 4); (-6, 0, 2, 4)\}$
σ	$\{(1, 1, 1, 1); (1, 0.75, 0.5, 0.25)\}$
a_n	$(0.01, 0.05, 0.1, 0.2, 0.3, 0.4)$

Table 12: The estimated a_n-function based on the simulated nominal size

	Dependent variable			
	$\log\left(\frac{\hat{s}}{1-\hat{s}}\right) - \log\left(\frac{0.05}{1-0.05}\right)$			
	(1)	(2)	(3)	(4)
$1/T$	0.776^{***}	-0.288^{***}	0.611^{***}	0.258^{***}
	(0.238)	(0.074)	(0.050)	(0.087)
$1/N$	28.143^{***}	4.637	21.156^{***}	8.585^{**}
	(10.127)	(3.124)	(2.524)	(4.334)
$\log\left(\frac{a_n}{1-a_n}\right)$	-0.016	-0.101^{***}	-0.111^{***}	-0.128^{***}
	(0.019)	(0.009)	(0.007)	(0.030)
$\log\left(\frac{\omega(\boldsymbol{\vartheta}_{M_0};M_0)}{1-\omega(\boldsymbol{\vartheta}_{M_0};M_0)}\right)$		-0.197^{***}	0.002	-0.013^{***}
		(0.029)	(0.006)	(0.003)
Constant	-0.616^{***}	-0.811^{***}	-0.680^{***}	-0.735^{***}
	(0.113)	(0.047)	(0.060)	(0.068)
Observations	48	648	576	288

Note: *p<0.1; **p<0.05; ***p<0.01