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

Yu Hao∗  Hiroyuki Kasahara
Faculty of Business and Economics  Vancouver School of Economics
The University of Hong Kong  The University of British Columbia
haoyu@hku.hk  hkasahar@mail.ubc.ca

June 6, 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 in the context of normal mixture panel regression. We demonstrate that the first-order derivative of the density function for the variance parameter in the normal mixture panel regression model is linearly independent from its second-order derivative for the mean parameter, unlike the cross-sectional normal mixture. 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 use a penalized maximum likelihood estimator and derive the asymptotic distribution of the 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 that the proposed tests have good finite sample performance. We apply these tests to estimate the number of production technology types for a finite mixture Cobb-Douglas production function model. Our findings suggest heterogeneity in output elasticities for intermediate goods, indicating firm-specific variation in production functions beyond Hicks-neutral productivity terms.

∗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 are also grateful for the IAAE grant at the 2019 IAAE Conference. This research is supported 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. Because of their flexibility, they have been used in empirical applications in various fields since the proposal of a two-component normal mixture model by [1]. In economics, finite mixtures are frequently used to model unobserved individual-specific effects in labor economics, health economics, and industrial organization, as well as in other fields [1]. The theoretical properties of finite mixture models and examples of their applications have been discussed by numerous authors, such as Titterington et al. (1985), Lindsay (1995), 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 of individuals. 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, while using too many components can be computationally costly and the model becomes singular because of 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 long been an unsolved problem. The regularity conditions of the likelihood ratio test (LRT) for standard asymptotic analysis fail in finite mixture models because of 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 the LRT for the number of components (see, e.g., Ghosh and Sen, 1985; Chernoff and Lander, 1993; Lemdani and Pons, 1997; Chen and Chen, 2001, 2003; Chen et al., 2004; Garel, 2001, 2003; 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; Liu and Shao, 2003; Zhu and Zhang, 2004; Azaïs et al., 2009). 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.

Kasahara and Shimotsu (2015) analyze the asymptotic distribution of the LRT statistics of a cross-sectional univariate finite mixture normal regression model, and Kasahara and Shimotsu (2019)[1] for example, Heckman and Singer (1984) use the finite mixture model to provide an alternative method of accounting for unobserved heterogeneity in the analysis of the 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 the unobserved dispersion of medical care utilization by the elderly. Kamakura and Russell (1989) and Andrews and Currin (2003) model consumer segmentation in marketing in industrial organizations.

---

[1] For example, Heckman and Singer (1984) use the finite mixture model to provide an alternative method of accounting for unobserved heterogeneity in the analysis of the 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 the unobserved dispersion of medical care utilization by the elderly. Kamakura and Russell (1989) and Andrews and Currin (2003) model consumer segmentation in marketing in industrial organizations.
develop a multivariate extension. Develop 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 the latent type within each unit. To the best of our knowledge, it is not known in the literature whether the aforementioned problems (i)–(iii) of the cross-sectional normal mixture still arise in the panel normal mixture. 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.

We show that problems related to (i) and (ii) arise, but the higher-order degeneracy of problem (iii) disappears in 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 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 to the literature. First, it analyzes the likelihood ratio-based test for the number of components in panel normal regression mixture models with conditional independence. Kasahara and Shimotsu (2015) and Kasahara and Shimotsu (2019) analyze likelihood ratio-based tests for the number of components in cross-sectional univariate normal mixture regression models and multivariate normal mixture models, respectively. We demonstrate that the asymptotic distribution of the PLRT and EM tests for panel normal regression mixture models with conditionally independent errors differs from that of the univariate/multivariate normal mixture models in the aforementioned two papers because higher-order dependency does not occur when the repeated measurement of the outcome variables is available in the panel data.

Kasahara and Shimotsu (2014) develop 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 panel normal regression mixture models with conditionally independent errors as a special case.
Furthermore, 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 whether the related unboundedness problem arises in panel data. We show that the likelihood ratio test statistic is unbounded in panel normal mixture models with conditionally independent errors when the time dimension of panel data is finite. This unboundedness causes over-rejection of the LRT. We introduce a penalty function to prevent the likelihood ratio test statistics from being unbounded, using computational experiments to determine the data-driven penalty function. We also develop an R package NormalRegPanelMixture (Hao, 2017) that contains the EM test module and asymptotic distribution simulation module.

Third, we conduct an empirical analysis of the number of production technology types using panel data from Japanese and Chilean manufacturing firms and provide strong evidence of substantial heterogeneity in production function coefficients across firms within narrowly defined industries. This is an important contribution to the literature on production function estimation, as most 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 into 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) develop an EM test for the null hypothesis of \( M_0 \) components applicable to general \( M_0 \geq 2 \), and Kasahara and Shimotsu (2015) propose an EM test for normal regression mixture models to test the null of \( M_0 \geq 2 \). The EM approach is also applied to test homogeneity in multivariate mixtures (Niu et al., 2011) and subgroup analyses (Shen and He, 2015). More recently, Liu et al. (2018) extend the EM test to mixtures of the general location-scale family distribution, and Kasahara and Shimotsu (2019) develop an EM test for multivariate normal mixture models. Building upon the literature, this paper develops an EM test for panel normal regression mixture models with conditionally independent errors.

The identification and estimation of latent group structures in panel data has received attention in recent studies (Kasahara and Shimotsu, 2009; Lin and Ng, 2012; Bonhomme and Manresa, 2015; Ando and Bai, 2016; Su et al., 2016; Lu and Su, 2017). 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 this paper is organized as follows. In Section 2, we define the finite normal mixture
where \( y \) with \( \theta \) represents the conditional density function of \( n \) fixed and the number of cross-sectional observations. We collect the component-specific parameters into \( \theta \). We consider finite mixture normal regression models with panel data, where the panel length is the \( j \)-th component density function with \( \mu_j \in \Theta_\mu \subset \mathbb{R}, \sigma_j^2 \in \Theta_{\sigma} \subset \mathbb{R}_{++}, \beta_j \in \Theta_\beta \subset \mathbb{R}^q, \gamma \in \Theta_\gamma \subset \mathbb{R}^p \), and \( \phi(t) = (2\pi)^{-1/2} \exp(-t^2/2) \) is the standard normal probability density function.

We collect the component-specific parameters into \( \theta \), and the regression coefficient \( \gamma \) for a vector \( z \) 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 \( w \) admits the representation (1). Consider a random sample of \( n \) with a panel length of \( T \) independent observations \( \{W_i\}_{i=1}^n \), where \( W_i = \{(Y_{it}, X_{it}^T, Z_{it}^T)\}_{t=1}^T \) from a true \( M_0 \)-component density \( f_M(w; \theta^*_M) \) defined in equation (1) with \( \theta^*_M = ((\alpha^*)^T, (\beta^*)^T) \). The superscript * signifies the true parameter value. Because component distributions can be identified only up to permutation, we assume that \( \mu^*_1 < \mu^*_2 < \cdots < \mu^*_M \) for identification.

3More generally, we may consider a lexicographical order: \( \theta^*_1 < \theta^*_2 < \cdots < \theta^*_M \).
Our goal is to test

\[ H_0 : M = M_0 \] 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 \( \{W_i\}_{i=1}^n \), where \( W_i = \{(Y_{it}, X_{it}^\top, Z_{it}^\top)^\top\}_{t=1}^T \), drawn from a true one-component density \( f(w; \gamma, \theta) \) defined in equation (2). Now consider a two-component mixture density function

\[
f_2(w; \theta_2) = \alpha f(w; \gamma, \theta_1) + (1 - \alpha) f(w; \gamma, \theta_2),
\]

where \( \theta_2 = (\alpha, \theta_1^\top, \theta_2^\top, \gamma^\top)^\top \in \Theta_{\theta_2} \) and \( \alpha \) is the mixing probability of the first component. The two-component model can generate the true one-component density in two cases: (1) \( \theta_1 = \theta_2 = \theta^* \) and (2) \( \alpha = 0 \) or 1. Consequently, the null hypothesis \( H_0 : M = 1 \) can be partitioned into two subhypotheses: \( H_{01} : \theta_1 = \theta_2 \) and \( H_{02} : \alpha(1 - \alpha) = 0 \). The regularity conditions of the LRTS for a standard asymptotic analysis fail in finite mixture models: under \( H_{01} \), \( \alpha \) is not identified, and the Fisher information matrix for the other parameters becomes singular; under \( H_{02} \), \( \alpha \) on the boundary of the parameter space, and either \( \theta_1 \) or \( \theta_2 \) is not identified.

As discussed in the introduction, analyzing the asymptotic distribution of the LRTS for the cross-sectional normal mixture is challenging because of 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 \( f_2(w; \theta_2) \) with respect to \( \sigma^2 \) is linearly dependent on its second-order derivative with respect to \( \mu_j \). The presence of problems (i)–(iii) in panel normal mixture models with \( T \geq 2 \) is not well understood in the literature because, to the best of our knowledge, no studies have examined them so far.

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

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

where \( f(W; \mu, \sigma^2) = \prod_{t=1}^T \phi((y_{it} - \mu)/\sigma)/\sigma \) and \( \phi(\cdot) \) is the standard normal density function. When \( \sigma_1^2 > 2\sigma^2 \), \( \mathbb{E}[(f(W; \mu_1, \sigma_1^2)/f(W; \mu^*, \sigma^2) - 1)^2] = \infty \). For more details, please refer to Proposition 3. Because the infinite Fisher information causes difficulty in deriving the asymptotic distribution under \( H_{02} \), this paper focuses on testing \( H_{01} \). We define \( \gamma_{t}^* := \{(\alpha, \gamma, \theta_1, \theta_2) \in \Theta_{\theta_2} : \theta_1 = \theta_2 = \theta^* \text{ and } \gamma = \gamma^* \} \), which is the subspace of \( \Theta_{\theta_2} \) that corresponds to \( H_{01} \). Note that because we focus
on $H_0$, our test may not have power against the local alternatives with $a_n \to 0$. 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 $\infty$. Define the likelihood ratio statistic with respect to the true parameter under $H_0$ as

$$LR_n^*(\theta_2) := 2 \left\{ \frac{1}{n} \sum_{i=1}^{n} \log f_2(W_i; \theta_2) - \frac{1}{n} \sum_{i=1}^{n} \log f(W_i; \gamma^*, \theta^*) \right\},$$

where $f_2$ is the density of the two-component finite mixture distribution in (1) with $M = 2$ and $(\gamma^*, (\theta^*)^\top)^\top$ is the true parameter value under $H_0$. Let $\hat{\theta}_{2,n}$ be the maximum likelihood estimator for the two-component model, i.e., $\hat{\theta}_{2,n} = \arg \max_{\theta_2 \in \Theta_{\theta_2}} LR_n^*(\theta_2)$.

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

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

$$\tilde{p}_n(\Theta_M) := \sum_{j=1}^{M} p_n(\sigma_j^2) \quad \text{with} \quad p_n(\sigma_j^2) := -a_n\{\sigma_j^2/\sigma^2_0 + \log(\sigma_j^2/\sigma^2_0) - 1\}. \quad (3)$$

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 conditions C1–C3 in the proof of Proposition [6].

Let

$$\hat{\theta}_2 = \arg \max_{\theta_2 \in \Theta_{\theta_2}} \frac{1}{n} \sum_{i=1}^{n} \log f_2(W_i; \theta_2) + \tilde{p}_n(\theta_2)$$

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

**Assumption 1.** (a) $X$ and $Z$ have finite second moments, and $\Pr(X^\top \beta + Z^\top \gamma \neq X^\top \beta^* + Z^\top \gamma^*) > 0$ for $(\beta^*, \gamma^*)^\top \neq ((\beta^*)^\top, (\gamma^*)^\top)^\top$. (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 $H_0 : M_0 = 1$, $\inf_{\theta_2 \in \Theta_{\theta_2}} |\hat{\theta}_2 - \theta_2| \to_p 0$. 

6
It should be noted that \( f_2(w; \theta_2^*) = f(w; \gamma^*, \theta^*) \) for any \( \theta_2^* \in \Theta_2^* \). Consequently, Proposition 2 suggests that the PMLE \( \hat{\theta}_2 \) converges in probability to a set of parameters for which the true density function \( f(w; \gamma^*, \theta^*) \) 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 \( f_2(w; \theta_2) \) with respect to \( \sigma_j^2 \) is not linearly dependent with its second-order derivative with respect to \( \mu_j \) (See Proposition 3(c)). Consequently, the panel mixture model with the component density function \( \Theta_2 \) 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 (cf. 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 (cf. Kasahara and Shimotsu, 2015).

As in any finite mixture models, however, the standard asymptotic analysis breaks down in testing \( H_{01} : \theta_1 = \theta_2 = \theta^* \) because \( \alpha \) is not identified under \( H_{01} \); in addition, the first-order derivative at the true value \( \hat{\theta}_2^* = (\alpha, (\theta^*)^T, (\theta^*)^T, (\gamma^*)^T)^T \) is linear dependent as

\[
\nabla_{\theta}, \log f_2(w; \theta_2^*) = \frac{\alpha}{1-\alpha} \nabla_{\theta} \log f_2(w; \theta_2^*). \tag{4}
\]

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

To extract the direction of the Fisher information matrix singularity, we adapt the reparameterization approach by Kasahara and Shimotsu (2012) and consider the following one-to-one reparameterization of \( \theta_1 \) and \( \theta_2 \) given \( \alpha \):

\[
\begin{pmatrix}
\lambda \\
\nu
\end{pmatrix} := \begin{pmatrix}
\theta_1 - \theta_2 \\
\alpha \theta_1 + (1-\alpha) \theta_2
\end{pmatrix} \quad \text{so that} \quad \begin{pmatrix}
\theta_1 \\
\theta_2
\end{pmatrix} = \begin{pmatrix}
\nu + (1-\alpha) \lambda \\
\nu - \alpha \lambda
\end{pmatrix}, \tag{5}
\]

where \( \nu \) and \( \lambda \) are both \( (q+2) \times 1 \) reparameterized parameter vectors with \( \nu = (\nu_\mu, \nu_\sigma, \nu_\beta^T)^T \) and \( \lambda = (\lambda_\mu, \lambda_\sigma, (\lambda_\beta)^T)^T = (\mu_1 - \mu_2, \sigma_1^2 - \sigma_2^2, (\beta_1 - \beta_2)^T)^T \). We also write \( \theta \) and \( \lambda \) as \( \theta = (\theta_1, \theta_2, \theta_3, \ldots, \theta_{q+2})^T := (\mu, \sigma^2, \beta_1, \ldots, \beta_q)^T \) and \( \lambda = (\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_{q+2})^T := (\lambda_\mu, \lambda_\sigma, \lambda_\beta_1, \ldots, \lambda_\beta_q)^T \).

This reparameterization is essential for analyzing the asymptotic distribution of the PLRTS in light of the linear dependency in (4). The reparameterized parameter \( \lambda \) 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, to characterize the asymptotic distribution of the LRTS.
Define the space for reparameterized parameters as

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

where $$\Theta_{\psi} = \{ \psi : \gamma \in \Theta_\gamma, \nu + (1 - \alpha)\lambda \in \Theta_\theta, \nu - \alpha \lambda \in \Theta_\theta \}.$$ Under the null hypothesis $$H_{01} : \theta_1 = \theta_2 = \theta^*$$, we have $$\lambda = (0, \ldots, 0)^T$$ and $$\nu = \theta^*$$. We rewrite the reparameterized parameters under the null hypothesis as $$(\psi^*)^T = ((\gamma^*)^T, (\theta^*)^T, 0, \ldots, 0)^T$$. Under the reparameterized parameter space, the density function and its logarithm are expressed as

$$
g(w; \psi, \alpha) = \alpha f(w; \gamma, \nu + (1 - \alpha)\lambda) + (1 - \alpha) f(w; \gamma, \nu - \alpha \lambda) \quad \text{and} \quad (6)$$

$$
l(w; \psi, \alpha) = \log g(w; \psi, \alpha).
$$

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

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

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

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

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

When $$\alpha$$ is bounded away from 0 and 1, the elements of $$\nabla_{\lambda\lambda^T} l(W; \psi^*, \alpha)$$ 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 with the cross-sectional models. The absence of collinearity between the first- and second-order derivatives in the panel models leads to different convergence rates and asymptotic properties.

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

\[ s(W) = \begin{pmatrix} s_\eta(W) \\ s_{\lambda\lambda}(W) \end{pmatrix}, \quad \text{where} \quad s_\eta(W) := \frac{\nabla(\eta^T \theta^T) f^*}{f^*}, \quad \text{and} \quad s_{\lambda\lambda}(W) := \frac{\nabla_{\lambda\lambda} f^*}{f^*}. \] (9)

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

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

\[ t(\psi, \alpha) = \begin{pmatrix} t_\psi \\ t_\lambda(\alpha) \end{pmatrix} = \begin{pmatrix} \eta - \eta^* \\ \alpha(1 - \alpha)v(\lambda) \end{pmatrix}, \] (10)

where \( v(\lambda) \) is a vector of unique elements of \( \lambda\lambda^T \) given by

\[ v(\lambda) = (\lambda_1\lambda_1, ..., \lambda_q\lambda_{q+2}, \lambda_1\lambda_2, ..., \lambda_{q+1}\lambda_{q+2})^T, \] (11)

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

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

\[ S_n := n^{-1/2} \sum_{i=1}^n s(W_i). \]

Then, taking the fourth-order Taylor expansion of \( L_n(\psi, \alpha) \) around \( (\psi^*, \alpha) \), we may write

\[ 2\{L_n(\psi, \alpha) - L_n(\psi^*, \alpha)\} \] as a quadratic function of \( \sqrt{n}t(\psi, \alpha) \) as

\[ 2\{L_n(\psi, \alpha) - L_n(\psi^*, \alpha)\} = 2(\sqrt{n}t(\psi, \alpha))^T S_n - (\sqrt{n}t(\psi, \alpha))^T \mathcal{I}_n(\sqrt{n}t(\psi, \alpha)) + R_n(\psi, \alpha) \] (12)

\[ = G_n^T \mathcal{I}_n G_n - \left[ \sqrt{n}t(\psi, \alpha) - G_n \right]^T \mathcal{I}_n \left[ \sqrt{n}t(\psi, \alpha) - G_n \right] + R_n(\psi, \alpha), \] (13)

where \( \mathcal{I}_n \) is the negative of the sample Hessian defined in the proof of Proposition 3 and \( G_n := \mathcal{I}_n^{-1} S_n \). Let \( \mathcal{I} = E[s(W)s(W)^T] \).

**Assumption 2.** (a) \( X \) and \( Z \) have finite 8-th moments. (b) \( E[UU^T] \) is non-singular, where \( U = [1, X^T, Z^T]^T \).

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

The non-singularity of \( \mathcal{I} \) 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 (7.2) in Appendix B.2 the first-order derivative of \( f_2(\omega; \theta_2) \) with respect to \( \sigma_j^2 \) is linearly independent of its second-order derivative with respect to \( \mu_j \) when \( T \geq 2 \), which ensures that the higher-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 \( \sqrt{n} t(\psi, \alpha) \) is given by the shifted and rescaled parameter space for \( (\eta, v(\lambda)) \) defined as \( \Lambda_n := \sqrt{n}(\Theta_\eta - \eta^*) \times \sqrt{n}(1 - \alpha)v(\Theta_\lambda) \), where \( v(A) := \{ t \in \mathbb{R}^q : t = v(\lambda) \} \) 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 can apply Lemma 2 of Andrews (1999) to approximate the distribution of the supremum of the right-hand side of (13) as

\[
\max_{\psi \in \Theta_\psi} 2\{ L_n(\psi, \alpha) - L_n(\psi^*, \alpha) \} \overset{d}{\to} G^T IG - \inf_{t \in \Lambda} (t - G)^T I(t - G),
\]

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

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

\[
\hat{\psi} = \arg \max_{\psi \in \Theta_\psi} L_n(\psi, \alpha) + \sum_{j=1}^2 p_n(\sigma_j^2(\psi, \alpha))
\]

with \( \hat{\psi} := (\hat{\gamma}^T, \hat{\nu}^T, \hat{\lambda})^T \), where \( \Theta_\psi \) is defined as the space of \( \psi \) such that the \( \theta_2 \) implied is in \( \Theta_\theta \) and \( \sigma_j^2(\psi, \alpha) \) is the value of \( \sigma_j \) implied by the value of \( \psi \) and \( \alpha \) (e.g., \( \sigma_1^2(\psi, \alpha) = \nu_\sigma + (1 - \alpha)\lambda_\sigma \)).

Let \( (\hat{\gamma}_0, \hat{\theta}_0) \) be the one-component MLE that maximizes the one-component likelihood function

\[ L_{0,n}(\gamma, \theta) := \sum_{i=1}^n \log f(W_i; \gamma, \theta). \]

Define the LRTS and the PLRTS of testing \( H_{01} \) with a small positivity constant \( \epsilon \) on \( \alpha \) as, respectively,

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

A hard bound is imposed on the values of \( \alpha \) to avoid an issue of the infinite Fisher information for testing \( H_{02} \). However, the LRTS may have reduced power if the true value of \( \alpha \) does not satisfy the constraint \( [\epsilon, 1-\epsilon] \) given an ad hoc constant \( \epsilon > 0 \). For this reason, in Section 5 we also develop the EM test, which does not impose a direct constraint on the value of \( \alpha \).
With \( s(W) \) in (9), partition \( \mathcal{I} = \mathbb{E}[s(W)s(W)^\top] \) and define
\[
\mathcal{I} = \begin{pmatrix}
\mathcal{I}_\eta & \mathcal{I}_{\eta\lambda} \\
\mathcal{I}_{\lambda\eta} & \mathcal{I}_{\lambda\lambda}
\end{pmatrix}, \quad \mathcal{I}_\eta = \mathbb{E}[s_\eta(W)s_\eta(W)^\top], \quad \mathcal{I}_{\eta\lambda} = \mathbb{E}[s_{\lambda\eta}(W)s_\eta(W)^\top], \quad \mathcal{I}_{\lambda\eta} = \mathcal{I}_{\eta\lambda}^\top, \quad \mathcal{I}_{\lambda\lambda} = \mathbb{E}[s_{\lambda\lambda}(W)s_{\lambda\lambda}(W)^\top],
\]
where \( S_{\lambda\eta} \sim N(0, \mathcal{I}_{\lambda\eta}) \). Define a set that characterizes the feasible values of \( \sqrt{n}t_\lambda(\lambda, \alpha) \) when \( n \to \infty \) by the cone
\[
\Lambda_\lambda = \left\{ \sqrt{n}(1 - \alpha)v(\lambda) : \lambda \in \Theta_\lambda \right\}.
\]

Define \( \hat{t}_\lambda \) as
\[
r_\lambda(\hat{t}_\lambda) = \inf_{t_\lambda \in \Lambda_\lambda} r_\lambda(t_\lambda), \quad r_\lambda(t_\lambda) := (t_\lambda - G_{\lambda, \eta})(\lambda - G_{\lambda, \eta}),
\]
where \( \hat{t}_\lambda \) is a projection of a random Gaussian random variable \( G_\lambda \) on a cone \( \Lambda_\lambda \).

The following proposition establishes the asymptotic distribution of the 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) \( t(\hat{\psi}, \alpha) = O_p(n^{-1/2}) \) for any \( \alpha \in (0, 1) \), (b) \( LR_n \overset{d}{\to} (\hat{t}_\lambda)^\top \mathcal{I}_{\lambda, \eta}\hat{t}_\lambda \) and \( PLR_n \overset{d}{\to} (\hat{t}_\lambda)^\top \mathcal{I}_{\lambda, \eta}\hat{t}_\lambda + \text{plim}_{n \to \infty} \sum_{j=1}^2 \rho_n(\sigma_j^2(\hat{\psi}, \alpha)) \).

Proposition 4(a) implies that \( \hat{\theta}_j - \theta^* = O_p(n^{-1/4}) \) for \( j = 1, 2 \). The \( n^{1/4} \) convergence rate is a consequence of the linear dependency in (1), where the identification of the parameter \( \theta \) 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 \( \sum_{j=1}^2 \rho_n(\sigma_j^2(\hat{\psi}, \alpha)) = o_p(1) \) under the null hypothesis of \( M = 1 \), \( PLR_n \) has the same asymptotic null distribution as \( LR_n \).

### 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 \( \{W_i\}_{i=1}^n \), where \( W_i = \{(Y_{it}, X_{it}^\top, Z_{it}^\top)^\top \}_{t=1}^T \) from an \( M_0 \)-component density \( f_{M_0}(w; \theta_{M_0}) \) defined in equation (17):
\[
f_{M_0}(w; \theta_{M_0}) = \sum_{j=1}^{M_0} \alpha_j^* f(w; \gamma_j^*, \theta_j^*), \quad (17)
\]
where \( \theta^*_M = (\theta^*_1, \theta^*_2, \ldots, \theta^*_M, \alpha^*_1, \ldots, \alpha^*_M, \gamma^*) \in \Theta_{\theta_M} \) and \( \alpha^*_M = 1 - \sum_{j=1}^{M-1} \alpha^*_j \).

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

\[
f_{M_0+1}(w; \theta_{M_0+1}) = \sum_{j=1}^{M_0+1} \alpha_j f(w; \gamma, \theta_j),
\]

where \( \theta_{M_0+1} = (\theta_1, \theta_2, \ldots, \theta_{M_0+1}, \alpha_1, \ldots, \alpha_{M_0}, \gamma) \in \Theta_{\theta_{M_0+1}} \) as defined in (17). We assume \( \mu^*_1 < \mu^*_2, \ldots, < \mu^*_M \) in the true parameters for identification.

The \((M_0 + 1)\)-component model (18) gives rise to the true density (17) in two 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 \( H_{01} : \theta_h = \theta_{h+1} = \theta^*_h \) 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 \( H_{02} = \cup_{h=1}^{M_0} H_{02h} \), where \( H_{02h} : \alpha_h = 0 \). Define the subset of \( \Theta_{\theta_{M_0+1}} \) corresponding to \( H_{02h} \) as

\[
\Upsilon_{2h} = \{ \theta_{M_0+1} \in \Theta_{\theta_{M_0+1}} : \alpha_h = 0; (\alpha_j, \mu_j, \sigma_j) = (\alpha^*_j, \mu^*_j, \sigma^*_j) \text{ for } j < h; (\alpha_j, \mu_j, \sigma_j) = (\alpha^*_{j-1}, \mu^*_{j-1}, \sigma^*_{j-1}) \text{ for } j > h \}.
\]

The score for testing \( H_{02h} : \alpha_h = 0 \) takes the form

\[
\nabla_{\alpha_h} \log f_{M_0+1}(W_i; \theta_{M_0+1}) = \frac{f(W_i; \mu_h, \sigma^*_h)}{f(M_0(W_i; \theta^*_M))}. \]

Because \( \mu^*_h \) is not identified when \( \alpha_h = 0 \), the Fisher information matrix of the LRTS for testing \( H_{02h} : \alpha_h = 0 \) depends on the supremum of the variance of \( \nabla_{\alpha_h} \log f_{M_0+1}(W_i; \theta_{M_0+1}) \) over \( \theta_{M_0+1} \in \Upsilon_{2h} \). The Fisher information is infinite unless there is an a priori restriction on the values of \( \sigma^*_j \).

**Proposition 5.** \( \sup_{\theta_{M_0+1} \in \Upsilon_{2h}} \mathbb{E}[\{ \nabla_{\alpha_h} \log f_{M_0+1}(W_i; \theta_{M_0+1}) \}^2] < \infty \) if and only if \( \max\{\sigma^2 : \sigma \in \Theta_\sigma\} \leq 2 \max\{\sigma^2_1, \ldots, \sigma^2_{M_0}\} \).

Because the restriction on the values of \( \sigma^*_j \) in Proposition 5 is difficult to justify and not easy to enforce in practice, we focus on testing \( H_{01} \).

Partition \( H_{01} \) as \( H_{01} = \cup_{h=1}^{M_0} H_{01h} \), where \( H_{01h} : \theta_h = \theta_{h+1} \) with \( \mu_1 < \cdots < \mu_h = \mu_{h+1} < \cdots < \mu_{M_0+1} \). We impose these inequality constraints on \( \mu_j \) 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, and each way corresponds to the null hypothesis of \( H_{01h} : \theta_h = \theta_{h+1} \) for \( h = 1, 2, \ldots, M_0 \). Testing \( H_{01h} : \theta_h = \theta_{h+1} \) in the \( M_0 \)-component null models is similar to testing \( H_{01} : \theta_1 = \theta_2 \) in the one-component null model in Section 3.
Define the subset of $\Theta_{\partial M_0+1}$ corresponding to $H_{0,1h}$ as
\[
Y_{1h}^*: = \left\{ \vartheta_{M_0+1} \in \Theta_{\partial M_0+1} : \alpha_h + \alpha_{h+1} = \alpha^*_h \text{ and } \theta_h = \theta_{h+1} = \vartheta^*_h; \gamma = \gamma^*; \gamma_j = \alpha^*_j \right\}
\] for $h = 1, \ldots, M_0$.

Let $\check{\Psi}^*_h$ denote the subset of $\check{\Phi}^*_h$ where $h \leq M_0$ and $\check{\Psi}^*_h \subseteq (\check{\Phi}^*_h)_{\partial M_0+1}$,

\[
\check{\Psi}^*_h = \left\{ \alpha_1, \ldots, \alpha_{M_0+1} \in [\epsilon, 1 - \epsilon] ; \sum_{j=1}^{M_0+1} \alpha_j = 1; \gamma \in \Theta_{\gamma}; \theta \in \Theta_{\theta}; \vartheta^*_j \in D^*_j \right\}
\] for $h = 1, \ldots, M_0$. Let $\check{\Psi}^*_h \cap Y_{1h}^* = \emptyset$ if $h \neq l$, and $\check{\Psi}^*_h \cap Y_{1h}^* = \emptyset$ if $h = l$.

For $h = 1, 2, \ldots, M_0$, define the local PMLE that maximizes the log-likelihood function of the $(M_0 + 1)$-component model under the constraint that $\check{\vartheta}^*_h \in \check{\Psi}^*_h$ by estimating the $(M_0 + 1)$-component model under the restriction that $\check{\vartheta}^*_h \in \check{\Psi}^*_h$.

\[
\check{\vartheta}^*_h = \arg \max_{\vartheta \in \check{\Psi}^*_h} L_{M_0+1,n}(\vartheta_{M_0+1}) + \check{p}_n(\vartheta_{M_0+1}),
\]

where
\[
L_{M,n}(\vartheta) := \sum_{i=1}^{n} \log f_M(W_i; \vartheta) \quad \text{and} \quad \check{p}_n(\vartheta) := \sum_{j=1}^{M} \check{p}_n(\sigma^2_j; \check{\sigma}^2_{0,j}).
\]

with
\[
\check{p}_n(\sigma^2_j; \check{\sigma}^2_{0,j}) := -a_n \{ \check{\sigma}^2_{0,j} \sigma^2_j + \log(\sigma^2_j / \check{\sigma}^2_{0,j}) - 1 \},
\] for $h = 1, 2, \ldots, M_0$.
alternatives within the restricted parameter space of the local PLRTS for
natives over a wide range of directions, we consider the PLRTS characterized by the maximum of
Suppose that Assumption 1 holds. Then, under the null hypoth 
esis from which the data are generated. These penalized like lihood estimators are consistent.

Under \( H_0 : M = M_0 \), \( \Psi_h^M \) contains a set of parameters \( \Gamma_1^h \) defined in \( \Delta \) such that
\( f_{M_0+1}(w; \theta_{M_0+1}) \) is equal to \( f_{M_0}(w; \theta_{M_0}^* \) for any \( \theta_{M_0+1} \in \Gamma_1^h \) and is therefore the density function from which the data are generated. These penalized likelihood estimators are consistent.

**Proposition 6.** Suppose that Assumption \( \Box \) holds. Then, under the null hypoth 
esis \( H_0 : M = M_0 \), \( \inf_{\theta_{M_0+1} \in \Psi_h^M} |\theta_{M_0+1} - \hat{\theta}_{M_0+1}| \rightarrow 0 \) for \( h = 1, 2, ..., M_0 \).

Consider the local PLRTS for testing \( H_0 : \theta_h = \theta_{h+1} \) defined by
\[
PLR_{n,h}^{M_0,\theta} := 2\{L_{M_0+1,n}(\hat{\theta}_{M_0}^h) + \hat{p}_n(\hat{\theta}_{M_0}^h) - L_{M_0,n}^\theta(\hat{\theta}_{M_0})\} \quad \text{for } h = 1, 2, ..., M_0.
\]

The test utilizing the local PLRTS, denoted by \( PLR_{n,h}^{M_0,\theta} \), possesses power solely against local alternatives within the restricted parameter space of \( \Psi_h^M \). 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
\[
PLR_{n}^{M_0}(M_0) := \max\{PLR_{n,1}^{M_0,1}, PLR_{n,2}^{M_0,2}, ..., PLR_{n,M_0}^{M_0,M_0}\}.
\]

Because \( \Theta_{\theta_{M_0+1}}(\epsilon) = \bigcup_{h=1}^{M_0} \hat{\Psi}_h^M \), \( PLR_{n}^{M_0}(M_0) \) is identical to \( \max_{\theta_{M_0+1} \in \Theta_{\theta_{M_0+1}}(\epsilon)} \{L_{M_0+1,n}(\theta_{M_0+1}) + \hat{p}_n(\hat{\theta}_{M_0+1}) - L_{M_0,n}^\theta(\hat{\theta}_{M_0})\} \).

To derive the asymptotic null distribution of \( PLR_{n}^{M_0}(M_0) \), collect the score vector for testing \( H_{0,1h} \) for \( h = 1, ..., M_0 \) into one vector as
\[
\tilde{s}(W) = \begin{pmatrix} \tilde{s}_\eta(W) \\ \tilde{s}_{\lambda\lambda}(W) \end{pmatrix}
\]
where \( \tilde{s}_\eta(W) = \begin{pmatrix} s_\alpha(W) \\ s_{(\gamma,\nu)}(W) \end{pmatrix} \) and \( \tilde{s}_{\lambda\lambda}(W) = \begin{pmatrix} s_{L\lambda}^1(W) \\ s_{L\lambda}^M(W) \end{pmatrix} \),
\[
(23)
\]
where
\[
s_\alpha(W) = \begin{pmatrix} f(W; \gamma^*, \theta_{M_0}^*) - f(W; \gamma^*, \theta_{M_0}^*) \\ \vdots \\ f(W; \gamma^*, \theta_{M_0-1}^*) - f(W; \gamma^*, \theta_{M_0}^*) \end{pmatrix} / f_{M_0}(W; \theta_{M_0}^*),
\]
\[
s_{(\gamma,\nu)}(W) = \sum_{j=1}^{M_0} \alpha_j \nabla_{(\gamma,\nu)} f(W; \gamma^*, \theta_{M_0}^*) / f_{M_0}(W; \theta_{M_0}^*),
\]
\[
\tilde{s}_{L\lambda}(W) = \tilde{\nabla}_{\theta_h \theta_h^*} f(W; \gamma^*, \theta_{h}^*) / f_{M_0}(W; \theta_{M_0}^*) \quad \text{for } h = 1, 2, ..., M_0,
\]
with \( \tilde{\nabla}_{\theta_h \theta_h^*} f(W; \gamma^*, \theta_{h}^*) := (c_1 \nabla_{\theta_{h1} \theta_{h1}} f^*, ..., c_{(q+2)(q+2)} \nabla_{\theta_{h,q+2} \theta_{h,q+2}} f^*, c_{(q+1)(q+2)} \nabla_{\theta_{h+1} \theta_{h+1}} f^*, ..., c_{(q+2)(q+2)} \nabla_{\theta_{h+1} \theta_{h+1}} f^*)^T \)}
for \( \theta \in (\theta_{h1}, \theta_{h2}, \theta_{h3}, ..., \theta_{h,q+2})^T := (\mu_h, \sigma_h^2, \beta_{h1}, ..., \beta_{hq})^T \) and \( c_{jk} = 1/2 \) for \( j \neq k \) and \( c_{jk} = 1 \) for \( j = k \). Define

\[
\tilde{T} := \mathbb{E}[\tilde{s}(W)^\top], \quad \tilde{T}_\eta := \mathbb{E}[\tilde{s}_\eta(W)^\top], \quad \tilde{T}_{\lambda \eta} := \mathbb{E}[\tilde{s}_{\lambda \eta}(W)^\top], \quad \tilde{T}_{\lambda \lambda} := \tilde{T}_{\lambda \eta} - \tilde{T}_{\eta \lambda} \tilde{T}_\eta^{-1} \tilde{T}_{\eta \lambda}.
\]

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

\[
\tilde{S}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{s}(W_i) \xrightarrow{d} \tilde{S} \sim N(0, \tilde{T}),
\]

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

Let \( \tilde{S}_{\lambda \eta} := (\tilde{S}_{\lambda \eta}^1, ..., \tilde{S}_{\lambda \eta}^M)^\top \) be a \( \mathbb{R}^{M_0(q+2)(q+1)/2} \)-valued random vector. For \( h = 1, 2, ..., M_0 \), define \( \hat{\lambda}_{\lambda \eta}^h := \mathbb{E}[\tilde{s}_{\lambda \eta}(\tilde{S}_{\lambda \eta}^h)^\top] \) and \( G_{\lambda \eta}^h := (\hat{\lambda}_{\lambda \eta}^h)^{-1} \tilde{S}_{\lambda \eta}^h \).

Define \( \hat{\lambda}_h \) analogously to \( \hat{\lambda}_{} \) as

\[
\varphi_h^h(\hat{\lambda}_h) = \inf_{\tilde{t}_h \in \mathcal{A}} \varphi_h^h(\tilde{t}_h): \quad \varphi_h^h(\tilde{t}_h) := (\tilde{t}_h - G_{\lambda \eta}^h)^\top \tilde{S}_{\lambda \eta}^h (\tilde{t}_h - G_{\lambda \eta}^h) \quad \text{for } h = 1, 2, ..., M_0.
\]

The local quadratic-form approximation of the log-likelihood function \( LR_n^{M_0,h} \) around \( \hat{\lambda}_{} \) has an identical structure to the approximation that we derive in Section 3 in testing \( H_0 \) in the test of homogeneity. Consequently, we can show that \( PLR_n^{M_0,h} \xrightarrow{d} (\hat{\lambda}_{}^h)^\top \tilde{T}_{\lambda \eta}^h \bar{t}_h \). Then, given (22), the asymptotic null distribution of the PLRTS for testing \( H_0 \) is given by the maximum over \( (\hat{\lambda}_{}^h)^\top \tilde{T}_{\lambda \eta}^h \bar{t}_h \) for \( h = 1, 2, ..., M_0 \).

**Assumption 3.** (a) \( \alpha^*_h \in (\epsilon, 1 - \epsilon) \) for \( j = 1, ..., M_0 \). (b) \( \tilde{T} \) is non-singular. (c) \( a_n \) in (27) satisfies \( a_n = O(1) \).

**Proposition 7.** Suppose that Assumptions 1-3 are satisfied. Then, under the null hypothesis \( H_0 : M = M_0 \), \( PLR_n(M_0) \xrightarrow{d} \max(\{ (\hat{\lambda}_{}^1)^\top \tilde{T}_{\lambda \eta}^1 \bar{t}_1, ..., (\hat{\lambda}_{}^{M_0})^\top \tilde{T}_{\lambda \eta}^{M_0} \bar{t}_1 \} \).

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 \( \tilde{S}_{\lambda \eta} = (\tilde{S}_{\lambda \eta}^1, ..., \tilde{S}_{\lambda \eta}^{M_0})^\top \) from \( N(0, \tilde{T}_{\lambda \eta}) \), where \( \tilde{T}_{\lambda \eta} \) is a sample analogue estimator of \( \tilde{T}_{\lambda \eta} \). Then, compute \( G_{\lambda \eta}^h = (\tilde{T}_{\lambda \eta}^h)^{-1} \tilde{S}_{\lambda \eta}^h \) and obtain \( \tilde{t}_h \) analogously to (26) using an estimator of \( \tilde{T}_{\lambda \eta}^h \) for \( h = 1, ..., M_0 \) and a simulated random draw is computed as \( \max(\{ (\hat{\lambda}_{}^1)^\top \tilde{T}_{\lambda \eta}^1 \bar{t}_1, ..., (\hat{\lambda}_{}^{M_0})^\top \tilde{T}_{\lambda \eta}^{M_0} \bar{t}_1 \} \). Appendixes B.2 and B.3 present an expression for the 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 $\alpha_j$, is subject to a hard constraint, which is dictated by an arbitrary choice of bounds. The EM test, in contrast, circumvents the need to impose an explicit constraint on the $\alpha_j$ values. It achieves this by performing a limited number of EM steps, starting from a predetermined set of $\alpha_j$ values. The EM test approach offers certain advantages, including computational simplicity and less stringent assumptions.

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

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

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

$$\hat{\theta}^{h(1)}_{M_0 + 1}(\tau_0) = \arg \max_{\theta \in \Theta_{M_0 + 1}^n} PL_n(\theta, \tau_0),$$

where $\Theta_{M_0 + 1}^n(\tau_0) := \{\theta \in \hat{\Psi}_h : (\alpha_h + \alpha_{h+1}) = \tau_0\}$ and

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

Starting from $(\hat{\theta}^{h(1)}_{M_0 + 1}(\tau_0), \tau^{h(1)}(\tau_0))$ with $\tau^{h(1)}(\tau_0) = \tau_0$, update $\hat{\theta}^{h(k)}_{M_0 + 1}(\tau_0)$ 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 $\hat{\theta}^{h(k)}_{M_0 + 1}$ 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

$$w_{ij}^{(k)} = \begin{cases} 
\alpha_j^{(k)} f(W_i; \gamma_j^{(k)}, \theta_j^{(k)}) / f_{M_0 + 1}(W_i; \hat{\theta}^{h(k)}_{M_0 + 1}(\tau_0)), & j = 1, \ldots, h - 1, \\
\alpha_{j-1}^{(k)} f(W_i; \gamma_j^{(k)}, \theta_j^{(k)}) / f_{M_0 + 1}(W_i; \hat{\theta}^{h(k)}_{M_0 + 1}(\tau_0)), & j = h + 2, \ldots, M_0 + 1,
\end{cases}$$

$$w_{ih}^{(k)} = \tau^{h(k)} \alpha_h^{(k)} f(W_i; \gamma_h^{(k)}, \theta_h^{(k)}) / f_{M_0 + 1}(W_i; \hat{\theta}^{h(k)}_{M_0 + 1}(\tau_0)),$$

$$w_{i,h+1}^{(k)} = (1 - \tau^{h(k)}) \alpha_h^{(k)} f(W_i; \gamma_h^{(k)}, \theta_h^{(k)}) / f_{M_0 + 1}(W_i; \hat{\theta}^{h(k)}_{M_0 + 1}(\tau_0)),$$

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

16
We also update $\theta_j$ and $\gamma$ as

\[
\begin{align*}
\left(\sigma_j^{(k+1)}\right)^2 &= \arg\min_{\sigma^2} \left\{ \sum_{i=1}^{n} w_i^{(k)} \sum_{t=1}^{T} \left( y_{it} - \mu_j^{(k+1)} - z_{it} \gamma^{(k+1)} - x_{it} \beta_j^{(k+1)} \right)^2 + p_n(\sigma_j^2) \right\}, \\
\gamma^{(k+1)} &= \left( \sum_{i=1}^{n} \sum_{t=1}^{T} z_{it} x_{it}^T \right)^{-1} \left( \sum_{i=1}^{n} \sum_{t=1}^{T} z_{it} \left( y_{it} - \sum_{j=1}^{M_0+1} w_{ij}^{(k)} x_{it} \beta_j^{(k)} \right) \right), \quad \text{and} \quad \\
\left( \mu_j^{(k+1)} \beta_j^{(k+1)} \right) &= \left( \sum_{i=1}^{n} w_{ij}^{(k)} \sum_{t=1}^{T} x_{it} x_{it}^T \right)^{-1} \left( \sum_{i=1}^{n} w_{ij}^{(k)} \sum_{t=1}^{T} x_{it} \left( y_{it} - z_{it} \gamma^{(k+1)} \right) \right),
\end{align*}
\]

where $x_{it} = (1, x_{it}^T)$. In the updating procedure, $\hat{\theta}_j^{(k+1)}(\tau_0)$ is not restricted to be in $\hat{\Psi}_h$.

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

\[
M_n^{h(k)}(\tau_0) := 2 \left\{ P L_n(\theta_j^{(k+1)}(\tau_0), \tau^{(k+1)}(\tau_0)) - L_{M_0,n}(\hat{\theta}_M) \right\}.
\]

With a predetermined finite number $K$, define the local EM test statistic by taking the maximum of $M_n^{h(k)}(\tau_0)$ across different values of $\tau_0$ as

\[
EM_n^{h(K)} := \max \{ M_n^{h(k)}(\tau_0) : \tau_0 \in \mathcal{T} \}.
\]

The test statistic $EM_n^{h(k)}$ tests $H_{0,1h} : \theta_h = \theta_{h+1}$ 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:

\[
EM_n(M_0) := \max \{ EM_n^{1(K)}, \ldots, EM_n^{M_0(K)} \}.
\]

Proposition 8. Suppose that Assumptions \[\text{[H]}\] hold and \{(0.5) \in \mathcal{T}. Then, under the null hypothesis $H_0 : M = M_0$, for any finite $K$, $EM_n(M_0) \xrightarrow{d} \max \{ (\hat{t}_A)^T I_{\lambda,\eta}^{1}, \ldots, (\hat{t}_A)^T I_{\lambda,\eta}^{M_0} \}$. Therefore, the asymptotic null distribution of the 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 the PLRTS and EM test statistic under local alternatives. For brevity, we focus on testing \( H_0 : M = 1 \) against \( H_1 : M = 2 \). Consider the following local alternative to the homogeneous model \( f(w; \gamma, \theta) \) with \( \theta^* = (\mu^*, \sigma^*, (\beta^*)^\top) \). For brevity, we omit the common parameter \( \gamma \) in this section. In a reparameterized parameter, \( \psi^* = ((\nu^*)^\top, (\lambda^*)^\top)^\top \).

For \( \alpha^* \in (0, 1) \) and a local parameter \( h = (h_\nu^\top, h_\lambda^\top) \) with \( h_\lambda \in \nu(\Theta_\lambda) \), we consider a sequence of contiguous local alternatives \((\alpha_n, \psi_n^\top) = (\alpha_n, \nu_n^\top, \lambda_n^\top) \in \Theta_n \times \Theta_\nu \times \Theta_\lambda \) such that, with \( t_\lambda(\lambda, \alpha) \) given by \((10)\),

\[
\begin{align*}
    h_\nu &= \sqrt{n}(\nu_n - \nu^*), \quad h_\lambda = \sqrt{n}t_\lambda(\lambda_n, \alpha_n), \quad \text{and} \quad \alpha_n = \alpha^* + o(1). \quad (31)
\end{align*}
\]

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

\[
    \theta_{1,n} = \nu_n + (1 - \alpha_n)\lambda_n \quad \text{and} \quad \theta_{2,n} = \nu_n - \alpha_n\lambda_n \quad (32)
\]

for \( \nu_n = \nu^* + n^{-1/2}h_\nu \) and \( \lambda_n = (\lambda_{1,n}, \lambda_{2,n}, \ldots, \lambda_{q+2,n})^\top \) with

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

where \( h_\lambda = (h_{\lambda,1}^2, \ldots, h_{\lambda,q+2}^2, h_{\lambda,1}h_{\lambda,2}, \ldots, h_{\lambda,q+1}h_{\lambda,q+2})^\top \). The local alternatives are of order \( n^{1/4} \) rather than \( n^{1/2} \). See the discussion following 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\([3]\) hold for \( M_0 = 1 \). Consider a sequence of contiguous local alternatives \( \theta_{2,n} = (\alpha_n, \theta_{1,n}^\top, \theta_{2,n}^\top) \) given in \((32)\), where \( \alpha_n \) and \( \lambda_n \) satisfy \((31)\). Then, under \( H_{1,n} : \theta = \theta_{2,n} \), we have \( \text{PLRT}_n(1), \text{EM}_n(1) \overset{d}{\rightarrow} (\tilde{t}_\lambda)^\top I_{\lambda,\eta}^{-1}t_\lambda \), where \( \tilde{t}_\lambda \) has the same distribution as \( t_\lambda \) in Proposition\([4]\) but \( G_{\lambda,\eta} \) is replaced with \( (I_{\lambda,\eta})^{-1}S_{\lambda,\eta} + h_\lambda \).

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 \( r = 2, \ldots, M \), where \( M \) 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 $M_0$. Develop a similar sequential hypothesis test for estimating the rank of a matrix.

For $M = 1, \ldots, M$, let $c_{1-q_n}^M$ denote the $100(1 - q_n)$ percentile of the cumulative distribution function of a random variable $\max\{Z^1, Z^2, \ldots, Z^M\}$ for $M = M_0$ in Propositions 7 and 8. Let $c_{1-q_n}^M$ be a consistent estimator of $c_{1-q_n}^M$. Then, our estimator based on sequential hypothesis testing (SHT, hereafter) is defined as

\[
\hat{M}_{\text{PLR}} = \min_{M \in \{1, \ldots, \hat{M}\}} \{M : \text{PLR}_n(r) \geq c_{1-q_n}^M, r = 0, \ldots, M - 1, \text{PLR}_n(M) < c_{1-q_n}^M\},
\]

\[
\hat{M}_{\text{EM}} = \min_{M \in \{1, \ldots, \hat{M}\}} \{M : \text{EM}_n(r) \geq c_{1-q_n}^M, r = 0, \ldots, M - 1, \text{EM}_n(M) < c_{1-q_n}^M\}. \tag{33}
\]

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 $n \to \infty$ when $-n^{-1} \ln q_n = o(1)$ and $q_n = o(1)$.

Let $Q_n^M(\theta_M) := n^{-1} \sum_{i=1}^n \ln f_M(w_i; \theta_M)$ and $Q^M(\theta_M) := \mathbb{E}[\ln f_M(w_i; \theta_M)]$, where $f_M(w_i; \theta_M)$ is defined in (1) for $M = 1, \ldots, \hat{M}$.

**Assumption 4.** For $M = 1, \ldots, M_0 - 1$, (a) $Q^M(\theta_M)$ has a unique maximum at $\theta^*_M$ in $\Theta_{\theta_M}$; (b) $\Theta_{\theta_M}$ is compact; (c) $\theta^*_M$ is interior to $\Theta_{\theta_M}$; (d) $B^M(\theta^*_M) := \mathbb{E}\left\{\nabla_{\theta_M} \ln f_M(w_i; \theta_M) \nabla_{\theta_M}^\top \ln f_M(w_i; \theta_M)\right\}$ is non-singular; and (e) $A^M(\theta^*_M) := \mathbb{E}\left\{\nabla_{\theta_M} \theta^*_M \ln f_M(w_i; \theta_M)\right\}$ has a constant rank in some open neighborhood of $\theta^*_M$; (f) $Q^{M+1}(\theta^*_M) - Q^M(\theta^*_M) > 0$.

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

Assumptions 4(a)–(e) ensure the consistency and asymptotic normality of $\hat{\theta}_M$, 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 $M_0$-component 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 the 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 develop a data-dependent empirical formula for $a_n$ by selecting a formula that ensures that the empirical rejection probabilities match the nominal size (5%) across various null models and
sample sizes, as reported in Table 11 in Appendix D. Specifically, for the model without conditioning variables, we derive the following data-dependent empirical formula for testing the null hypotheses of $M_0 = 1, 2, 3, 4$:

$$a_n = \begin{cases} 
(1 + \exp \left( \frac{\hat{\rho}_4^{M_0} + \hat{\rho}_2^{M_0}}{\hat{\rho}_4^{M_0} + \hat{\rho}_2^{M_0}} \frac{1}{n} \right) \right)^{-1}, & M_0 = 1 \\
(1 + \exp \left( \frac{\hat{\rho}_4^{M_0} + \hat{\rho}_2^{M_0}}{\hat{\rho}_4^{M_0} + \hat{\rho}_2^{M_0}} \frac{1}{n} + \hat{\rho}_4^{M_0} \log \left( \frac{\omega(\hat{\vartheta}_M; M_0)}{1 - \omega(\hat{\vartheta}_M; M_0)} \right) \right)^{-1}, & M_0 = 2, 3, 4,
\end{cases}$$  \hspace{1cm} (34)

where $\omega(\hat{\vartheta}_M; M_0)$ 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 the $108 \times 6 = 648$ combinations of the parameter values, sample sizes, and $a_n$ values, let $\hat{s}$ denote the fraction of simulated datasets that lead to the rejection of the null hypothesis at the 5% significance level. Using these 648 observations of $\{\hat{s}, N, T, \omega(\hat{\vartheta}_2; 2), a_n\}$, we run the following regression:

$$\log \left( \frac{\hat{s}}{1 - \hat{s}} \right) - \log \left( \frac{0.05}{1 - 0.05} \right) = \begin{cases} 
\hat{\rho}_1^{M_0} + \hat{\rho}_2^{M_0} \frac{1}{n} + \hat{\rho}_3^{M_0} \frac{1}{n} + \hat{\rho}_4^{M_0} \log \left( \frac{a_n}{1 - a_n} \right), & M_0 = 1 \\
\hat{\rho}_1^{M_0} + \hat{\rho}_2^{M_0} \frac{1}{n} + \hat{\rho}_3^{M_0} \frac{1}{n} + \hat{\rho}_4^{M_0} \log \left( \frac{a_n}{1 - a_n} \right) + \hat{\rho}_5^{M_0} \log \left( \frac{\omega(\hat{\vartheta}_M; M_0)}{1 - \omega(\hat{\vartheta}_M; M_0)} \right), & M_0 = 2, 3, 4,
\end{cases}$$

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) denote 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 $M_0 = 1, 2, 3, 4$ 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 $M_0 = 4$; and $a_n = 0.5$ if $M_0 \geq 5$. These penalty terms for the regression with covariates are chosen by averaging the predictions 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 of the combinations of $(N, T, \alpha, \mu, \sigma)$ in Table

20
for $M_0 = 2$ and taking the average across the predicted values of $\hat{a}_n$. 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 we examine the null hypothesis $H_0 : M = 2$ against the alternative hypothesis $H_1 : M = 3$. A total of 2,000 repetitions are used for the asymptotic distribution, and 1,000 repetitions are used for the bootstrap distribution. Moreover, the PLRT with simulated critical values is considered.

The table presents the 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 ($\alpha = (0.5, 0.5)$) than when they are unequal ($\alpha = (0.2, 0.8)$). The bootstrapped EM test demonstrates satisfactory performance.

For the PLRT, 2,000 repetitions are conducted, and the results are reported for cases where a constraint is applied to $\alpha_j \in [\epsilon, 1 - \epsilon]$ with $\epsilon = 0.1$. The value of $a_n$ for the PLRT is chosen to be 10 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 the EM test and the PLRT, the test power is greater when the distances between $\mu_j$s are larger and equal, such as $(-1, 0, 1)$ or $(-1.5, 0, 1.5)$, as opposed to unbalanced distances such as $(-1, 0, 2)$ or $(-0.5, 0, 1.5)$. The power is also improved when the mixture probabilities are equal ($\alpha = (1/3, 1/3, 1/3)$) relative to when they are unequal ($\alpha = (1/4, 1/2, 1/4)$). The power increases with both the time dimension $T$ and the 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 $(\mu_1, \mu_2, \mu_3)$ 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. The 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 5 when we sequentially apply our EM test to estimate the number of components. We also investigate the performance of
the SHT using the EM test in comparison with the AIC and the BIC when the data are 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% 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 frequencies at which the three methods select the number of components in this simulation. The table demonstrates that the proposed SHT 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 and accurately estimates the number of components 58% of the time. Overall, in this simulation, our proposed SHT approach outperforms both the AIC and the BIC.
Table 1: Sizes (in %) of the EM Test and the 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 |
| $T$   | $N$   | 3      | 5     | 8     | 3     | 5     | 8     | 3     | 5     | 8     | 3     | 5     | 8     |
| 200   | 400    | 200    | 400    | 200    | 400    | 200    | 400    | 200    | 400    | 200    | 400    | 200    | 400    | 200    | 400    | 200    | 400    | 200    | 400    |
| (A, C) | 5.3    | 4.8    | 4.0   | 3.8   | 4.0   | 2.95  | 4.6    | 6.2    | 6.2    | 4.8    | 4.6    | 4.8    | 5.2    | 7.7    | 6.6    | 6.75   | 6.7    | 6.5    | 5.5    | 5.5    |
| (A, D) | 5.9    | 4.9    | 5.0   | 5.0   | 4.45  | 4.0   | 5.4    | 4.8    | 5.2    | 5.6    | 5.0   | 5.0   | 5.8    | 5.4    | 5.15   | 5.1    | 6.1    | 5.55   | 6.2    | 6.2    |
| (B, C) | 3.8    | 2.5    | 3.05  | 3.05  | 3.45  | 3.6   | 3.25   | 3.6    | 4.2    | 5.2    | 4.0   | 4.0   | 5.4    | 6.25   | 5.45   | 6.45   | 6.05   | 5.05   | 5.05   | 6.05   |
| (B, D) | 4.8    | 4.6    | 3.5   | 3.5   | 3.15  | 3.55  | 3.95   | 3.6    | 4.0    | 5.8    | 4.0   | 4.0   | 4.6    | 2.35   | 4.4    | 3.9    | 4.85   | 4.95   | 5.2    | 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 2,000 repetitions and the bootstrap simulation is based on 1,000 repetitions.
Table 2: Powers (in %) of the EM Test and the PLRT of $H_0 : M_0 = 2$ Against $H_A : M_0 = 3$ at the 5% Level

| N  | A         |         | B         |         |
|----|-----------|---------|-----------|---------|
|    | 100       | 500     | 100       | 500     |
| T  | 2         | 5       | 2         | 5       |

| EM test          |         |         |         |         |
| (C, G)           | 20.9    | 81.6    | 57.6    | 100.0   |
| (C, H)           | 49.2    | 99.9    | 99.9    | 100.0   |
| (C, I)           | 12.1    | 20.4    | 18.0    | 62.6    |
| (D, G)           | 77.9    | 100.0   | 100.0   | 100.0   |
| (D, H)           | 57.4    | 100.0   | 100.0   | 100.0   |
| (D, I)           | 16.0    | 59.5    | 31.8    | 99.9    |
| (E, G)           | 93.0    | 100.0   | 100.0   | 100.0   |
| (E, H)           | 83.8    | 100.0   | 100.0   | 100.0   |
| (E, I)           | 25.7    | 97.0    | 80.2    | 100.0   |
| (F, G)           | 99.9    | 100.0   | 100.0   | 100.0   |
| (F, H)           | 93.5    | 100.0   | 100.0   | 100.0   |
| (F, I)           | 40.8    | 99.9    | 98.2    | 100.0   |

| PLRT              |         |         |         |         |
| (C, G)           | 22.7    | 85.1    | 56.5    | 100.0   |
| (C, H)           | 57.1    | 100.0   | 99.8    | 100.0   |
| (C, I)           | 12.0    | 21.0    | 12.4    | 66.1    |
| (D, G)           | 79.9    | 100.0   | 100.0   | 100.0   |
| (D, H)           | 65.3    | 100.0   | 100.0   | 100.0   |
| (D, I)           | 14.8    | 63.6    | 28.6    | 100.0   |
| (E, G)           | 91.5    | 100.0   | 100.0   | 100.0   |
| (E, H)           | 86.8    | 100.0   | 100.0   | 100.0   |
| (E, I)           | 28.7    | 97.2    | 77.7    | 100.0   |
| (F, G)           | 99.9    | 100.0   | 100.0   | 100.0   |
| (F, H)           | 96.4    | 100.0   | 100.0   | 100.0   |
| (F, I)           | 45.7    | 100.0   | 98.4    | 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 $(\mu_1, \mu_2, \mu_3) = (-0.5, 0, 1.5), (-1, 0, 1), (-1, 0, 2)$, and $(-1.5, 0, 1.5)$, respectively; and $G, H$, and $I$ refer to $(\sigma_1, \sigma_2, \sigma_3) = (0.6, 0.6, 1.2), (0.6, 1.2, 0.6)$, and $(1, 1, 1)$, respectively.
Table 3: Sizes (in %) of the EM Test of $H_0: M_0 = 3$ Against $H_A: M_0 = 4$ at the 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, and $C, D,$ and $E$ 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 2,000 repetitions and the bootstrap simulation is based on 1,000 repetitions.

Table 4: Sizes of the EM Test of $H_0: M_0 = 2$ Against $H_A: M_0 = 3$ with Conditioning Variables

|     | (A, C, E) | (A, C, F) | (A, D, E) | (A, D, F) | (B, C, E) | (B, C, F) | (B, D, E) | (B, D, F) |
|-----|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| (N, T) |           |           |           |           |           |           |           |           |
| (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, and $C$ 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 $(0.1, 0.1)$, respectively. 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 the 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 $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.16076522, 0.32454077, 0.09025875, 0.35478905, 0.06964622)$, $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5) = (-1.241241, -0.33803875, 0.4480291, 0.52379553, 1.4139465)$, $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (0.451833, -0.05988709, -0.2453261, -0.03106076, 0.2053708)$, and $(\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) = (0.9933480, 0.4585760, 0.9954302, 0.4116855, 0.1863346)$. 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 an $M$-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 and 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, $\{Y_{it}, V_{it}, L_{it}, K_{it}\}_{i=1}^N_{t=1}$, where $Y_{it}$, $V_{it}$, $L_{it}$, and $K_{it}$ represent the output, intermediate input, labor, and capital of firm $i$ in year $t$, respectively. We denote the logarithms of the corresponding variables by lowercase letters as $(y_{it}, v_{it}, l_{it}, k_{it})$, with, for example, $y_{it} = \log(Y_{it})$.

We use a finite mixture specification to capture the unobserved heterogeneity in a firm’s input elasticities. We are interested in testing the number of production technology types. Assume that 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

$$Y_{it} = \exp(\epsilon_{it})F^j_t(V_{it}, L_{it}, K_{it}, \omega_{it}),$$

(35)
with
\[ F_t^j(V_{it}, L_{it}, K_{it}, \omega_{it}) := \exp(\gamma^j_t + \omega_{it})V_{it}^j L_{it}^{S_{it,j}} K_{it}^{S_k,j}, \]
where \( \gamma^j_t \) represents the aggregate productivity shock of type \( j \) in year \( t \), \( \omega_{it} \) is the serially correlated productivity shock, and \( \epsilon_{it} \) is the idiosyncratic productivity shock.

We assume that an intermediate input \( V_{it} \) is flexibly chosen by firm \( i \) after observing the aggregate shock \( \gamma^j_t \) and the serially correlated productivity shock \( \omega_{it} \). The variable \( \epsilon_{it} \) 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 \( I_{it} \).

To identify the intermediate input elasticity of the production function, we introduce the following assumptions (cf. 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 values of \( i \) and \( t \). \( \omega_{it} \) follows an exogenous first-order stationary Markov process given by 
\[ \omega_{it} = h^j(\omega_{it-1}) + \eta_{it} \] where, conditional on \( I_{it-1} \), \( \eta_{it} \) is a mean-zero i.i.d. random variable. (c) \( (\gamma^j_t, \omega_{it}) \) \( \in I_{it} \) and \( \epsilon_{it} \not\in I_{it} \).

**Assumption 6.** (a) Firms are price-takers in both output and input markets, where \( P_{Y,t} \) and \( P_{V,t} \) are, respectively, 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.** Values of \( V_{it} \) are chosen at time \( t \) by maximizing the expected profit conditional on information \( I_{it} \) 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
\[ V_{it} = \arg \max_{V} P_{it} \mathbb{E}[\exp(\epsilon_{it})|D_i = j]F_t^j(V, K_{it}, L_{it}, \omega_{it}) - P_{V,t}V. \] (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 \( \gamma^j_t \) and the serially correlated productivity shock \( \omega_{it} \) 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 \( (K_{it}, L_{it}) \).

\[ ^4 \text{We are agnostic about the timing of choosing } K_{it} \text{ and } L_{it} \text{ as long as they are either determined before } V_{it} \text{ or simultaneously chosen with } V_{it}. \] It is reasonable to assume that capital input \( K_{it} \) is determined before the value of \( V_{it} \) is chosen.
Given the above Assumptions [5, 6, and 7] we derive an empirical specification based on the first-order condition of the profit maximization problem (35), following the idea developed by Gandhi et al. (2020) and extending it 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 
\[
\delta_{v,j} = \frac{\partial P_i(V_{it}, K_{it}, L_{it})/\partial V_{it}}{P_i(V_{it}, K_{it}, L_{it})/V_{it}}
\]
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
\[
s_{it} = \log \delta_{v,j} + \frac{1}{2}\sigma_j^2 - \epsilon_{it} \quad \text{for } D_i = j,
\]
where
\[
s_{it} := \log \left( \frac{P_i V_{it}}{F_{it}} \right)
\]
is the logarithm of the ratio of the intermediate input cost to revenue.

Collect the observed data as \( W_i = \{s_{it}, \log K_{it}\}_{t=1}^T \). Let \( \mu_j = \log \delta_{v,j} + \frac{1}{2}\sigma_j^2 \) and define a type-specific parameter to be \( \theta_j = (\mu_j, \sigma_j) \), where \( \delta_{v,j} \) can be identified from \( \theta_j \) as \( \delta_{v,j} = \exp(\mu_j - \sigma_j^2/2) \). Collect the parameters of each type and the mixing probability as \( \theta_M = (\alpha_1, \ldots, \alpha_{M-1}, \theta_1^T, \ldots, \theta_M^T)^T \). Recall that \( \epsilon_{it} \sim N(0, \sigma_j^2) \) over \( i \) and \( t \) conditional on the technology type \( D_i = j \). Then, from (37), we can write the density function of \( s_{i1}, \ldots, s_{iT} \) as a mixture of type-specific likelihood density similar to the density function in equation (1):
\[
f_M(W_i; \theta_M) = \sum_{j=1}^M \alpha_j \prod_{t=1}^T \frac{1}{\sigma_j} \phi \left( \frac{s_{it} - \mu_j}{\sigma_j} \right). \tag{38}
\]
The PMLE is defined as
\[
\hat{\theta}_M = \arg \max_{\theta_M} \sum_{i=1}^n \log f_M(W_i; \theta_M) + \tilde{p}_n(\theta_M).
\]

As an alternative specification, we allow the elasticity of output for intermediate input to be a function of \( \log K_{it} \) as \( \delta_{v,j} = \beta_{0,j} + \beta_{k,j} \log K_{it} \). 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
\[
f_M(W_i; \theta_M) = \sum_{j=1}^M \alpha_j \prod_{t=1}^T \frac{1}{\sigma_j} \phi \left( \frac{s_{it} - \mu_j - \beta_{k,j} \log K_{it}}{\sigma_j} \right). \tag{39}
\]
However, labor input \( L_{it} \) may be flexibly chosen simultaneously with \( V_{it} \) after \( \gamma_i^j \) and \( \omega_{it} \) are observed. Even when labor input is simultaneously chosen with intermediate input, equation (38) 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:

$$f_M(W_i; \theta_M) = \sum_{j=1}^{M} \alpha_j \prod_{t=1}^{T} \frac{1}{\sigma_j} \phi \left( \frac{s_{it} - \mu_j - \beta_{k,j} \log K_{it} - \beta_{l,j} \log L_{it}}{\sigma_j} \right).$$

(40)

### 9.2 Empirical results

We apply the EM test to two producer-level datasets to determine the number of production technology types. We use the production data from Japanese publicly traded firms from 2003 to 2007 and Chilean manufacturing plants from 1992 to 1996.\footnote{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.} We clean the data and use the firms/plants with continuous data entry for five years to ensure that we have 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 the intermediate input elasticities differ across firms within the narrowly defined industries.
Table 6: Descriptive Statistics for the Revenue Share of Intermediate Materials and the Log of Gross Output for the Japanese Firms and Chilean Plants

Panel A: Japanese publicly traded firms

| Industry    | NObs | n  | \(\frac{P_{V,t}V_{it}}{P_{Y,t}Y_{it}}\) mean | sd | \(\log(Y_{it})\) 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

| Industry               | NObs | n  | \(\frac{P_{V,t}V_{it}}{P_{Y,t}Y_{it}}\) mean | sd | \(\log(Y_{it})\) 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 Japanese firm-level data from 2003 to 2007 and Chilean plant-level data from 1992 to 1996. All observations with \(\log(V_{it}/Y_{it}) \leq -3\) and \(\log(V_{it}/Y_{it}) > \log(2)\) are removed. The dataset is a balanced panel; i.e., we keep 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 Producers 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 materials.
2 *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.
Table 8: The EM Test for Chilean Producers 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     | -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     | -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     | -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 materials.
2 *, **, and *** indicate significance at the 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% 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 variables (38) and the models with conditioning variables (39)–(40).

Tables 7 and 8 report the results of the EM test for the model without conditioning variables (38) for the Japanese and the Chilean industries with a panel length of $T = 3, 4, 5$ and a null model of $M = 1, \ldots, 5$. 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 the 5% significance level, which indicates
that there are at least five types of intermediate input elasticities. This result reflects the 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 among 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 elasticity parameters are common across firms. Given the strong evidence of heterogeneity in the production function coefficients, incorporating heterogeneity in production function coefficients in empirical applications is warranted and should be encouraged.

One possible reason for the estimated number of technology types being greater than 5 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 values of capital input and labor input by estimating models (39)–(40).

Table 9 presents the results of the SHT and the BIC when we estimate 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 SHT suggests that the data are generated from seven- to nine-component models; concurrently, the BIC selects models with at least 10 components. For the Chilean food industry, the SHT indicates a 10-component model, while the BIC chooses an eight-component model. In contrast, the SHT 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, the SHT and the BIC in Table 10 both 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: \( \log \frac{P_{V,t}}{P_{V,t}} \), 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 materials using panel data of length \( T = 3 \).

2 *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.
Table 10: The EM Test and the BIC (Dependent Variable: $\frac{P_{V,t}}{P_{Y,t}}$, Regressors: $\log K_{it}$ and $\log L_{it}$)

| Mo  | 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 materials using panel data of length $T = 3$.
2 *, **, and *** indicate significance at the 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. Arbitrary choice of the number of components can result in biased estimates and invalid inferences, and can reduce the 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 SHT 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 provides strong evidence of the presence of unobserved heterogeneity in technology types. One important caveat of our empirical exercise is that the class of production functions that 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] 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.

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

[5] 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.

[6] 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.
[7] Balat, J., Brambilla, I., and Sasaki, Y. (2019). Heterogeneous firms: Skilled-labor productivity and the destination of exports.

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

[9] 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.

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

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

[12] 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.

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

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

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

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

[17] 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.

[18] 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.

[19] 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.
[20] 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.

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

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.
[33] Kasahara, H. and Shimotsu, K. (2009). Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices. *Econometrica*, 77(1):135–175.

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

[35] 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.

[36] 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.

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

[38] 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.

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

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

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

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

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

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

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

[46] 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.
[47] 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.

[48] 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.

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

[50] 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.

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

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

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

[54] 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.

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

[56] 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.

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

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

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

[60] 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 an intercept parameter and a variance parameter but without covariates with $W_i = \{y_{it}\}_{t=1}^T$.

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 a chi-square distribution with $T - 1$ degrees of freedom. Let $i^* = \arg \min_{i=1,\ldots,n} \{s_i^2\}$ so that $s_{i^*}^2 = \min\{s_1^2, \ldots, s_n^2\}$ is the minimum of $s_i^2$ across all values of $i$. We consider a sequence of parameters $\theta_{2,n} = (\alpha_n, \theta_{1,n}, \theta_{2,n})^T$ with $\alpha_n = 1/n, \theta_{1,n} = (\mu_{1,n}, \sigma_{1,n}^2)^T = (\tilde{Y}_{i^*}, s_{i^*}^2)^T$, and $\theta_{2,n} = \theta^* = (\mu^*, \sigma^*)^T$ for all $n$. Because $LR_n^*(\theta_{2,n}) \leq LR_n^*(\tilde{\theta}_{2,n})$, it suffices to show that $LR_n^*(\theta_{2,n})$ is unbounded in probability.

Define

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

Then, the LRT statistic for a two-component mixture is written as

$$LR_n^*(\theta_{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) \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(W_i; \theta^*) \}

= 2 \sum_{i \neq i^*} \{ \log (\exp(\log \alpha_n + \ell(W_i; \theta_{1,n})) + \exp(1 - \alpha_n + \ell(W_i; \theta^*)) - \ell(W_i; \theta^*)) \}

+ 2 \{ \log (\exp(\log \alpha_n + \ell(W_{i^*}; \theta_{1,n})) + \exp(1 - \alpha_n + \ell(W_{i^*}; \theta^*)) - \ell(W_{i^*}; \theta^*)) \}. \quad (41)

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

$$= 2(n - 1) \log \left( \frac{n - 1}{n} \right) + 2 \sum_{i \neq i^*} \log \left( 1 + \frac{1}{n - 1} \exp(\ell(W_i; \theta_{1,n}) - \ell(W_{i^*}; \theta^*)) \right),$$

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 $\log \left( 1 + \frac{1}{n - 1} \exp(\ell(W_i; \theta_{1,n}) - \ell(W_{i^*}; \theta^*)) \right) \geq 0$ for all $n$.

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

$$2 \{ -\log n + \ell(W_{i^*}; \theta_{1,n}) \} + 2 \log (1 + (n - 1) \exp(\ell(W_{i^*}; \theta^*) - \ell(W_{i^*}; \theta_{1,n}))) - 2\ell(W_{i^*}; \theta^*), \quad (42)$$

where $2 \{ -\log n + \ell(W_{i^*}; \theta_{1,n}) \}$ diverges to infinity as $n \to \infty$ by Lemma[1] the second term in (42)
is bounded below from zero, and the third term is bounded in probability because $\ell(W_i; \theta^*) = O_p(1)$. Therefore, for any $M < \infty$, we have $\Pr \left( LR_n^\ast(\theta_2, n) \leq M \right) \to 0$ as $n \to \infty$. The stated result then follows from $LR_n^\ast(\theta_2, n) \leq LR_n^\ast(\hat{\theta}_2, n)$ for all $n$.

For a model with covariates, we can consider a sequence of parameters $\theta_{2,n} = (\alpha_n, \theta_1^T, \theta_2^T, \gamma_n^T)^T$ with $\alpha_n = 1/n$, $\theta_1, \theta_2, \gamma_n = (\mu^*, \sigma^*, (\beta^*)^T)^T$, and $\gamma_n = \gamma^*$. Then, repeating the above argument, the stated 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 Shimosu (2012). For a vector $x$ and a function $f(x)$, let $\nabla_x f(x)$ denote its $k$-th derivative with respect to $x$, which can be a multidimensional array. Observe that for any finite $k$ and for a neighborhood $N$ of $\psi^*$, we obtain

$$E||\nabla_{\psi^k} g(W_i; \psi^*, \alpha) / g(W_i; \psi^*, \alpha)||^2 < \infty,$$

$$E\left|\sup_{\psi \in \Theta \cap N} \nabla_{\psi^k} \log g(W_i; \psi, \alpha)\right|^2 < \infty$$

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

$$\nabla_{\eta^k} L_n(\psi^*, \alpha) = 0, \quad \nabla_{\lambda, \lambda, \lambda} L_n(\psi^*, \alpha) = O_p(n^{1/2}),$$

$$\nabla_{\eta \eta^k} L_n(\psi^*, \alpha) = O_p(n), \quad \nabla_{\eta \eta^\eta} L_n(\psi^*, \alpha) = O_p(n),$$

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

$$\sup_{\psi \in \Theta \cap N} \left| n^{-1} \nabla^{(4)} L_n(\psi, \alpha) - E\nabla^{(4)} \log g(W_i; \psi, \alpha) \right| = o_p(1),$$

$$E\nabla^{(4)} g(W_i; \psi, \alpha)$$

is continuous in $\psi \in \Theta \cap N$.

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

Taking a fourth-order Taylor expansion of $L_n(\psi, \alpha)$ around $\psi^*$ and using (43) and (44), we can
write $L_n(\psi, \alpha) - L_n(\psi^*, \alpha)$ as the sum of the relevant terms and the remainder term as follows:

$$L_n(\psi, \alpha) - L_n(\psi^*, \alpha) =$$

$$\nabla_{\eta} L^*_{\eta}(\eta - \eta^*) + \frac{1}{2!} (\eta - \eta^*)^T \nabla_{\eta\eta} L^*_{\eta}(\eta - \eta^*) + \frac{1}{2!} \sum_{i=1}^{q+2} \sum_{j=1}^{q+2} \nabla_{\lambda_i\lambda_j} L^*_{\eta}(\eta - \eta^*)$$

(48)

$$+ \frac{3}{3!} \sum_{i=1}^{q+2} \sum_{j=1}^{q+2} (\eta - \eta^*)^T \nabla_{\eta\lambda_i\lambda_j} L^*_{\eta}(\eta - \eta^*)$$

(49)

$$+ \frac{1}{4!} \sum_{i=1}^{q+2} \sum_{j=1}^{q+2} \sum_{k=1}^{q+2} \nabla_{\lambda_i\lambda_j\lambda_k\lambda_k} L^*_{\eta}(\eta - \eta^*)$$

(50)

where $\nabla L^*_{\eta}$ denotes the derivative of $L_n(\psi, \alpha)$ evaluated at $(\psi^*, \alpha)$. In view of (44) and (45), the remainder term is written as

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

(51)

$$+ O_p(n) \sum_{i=1}^{q+2} \sum_{j=1}^{q+2} (||\eta - \eta^*||^4 + ||\eta - \eta^*||^3 |\lambda_i| + ||\eta - \eta^*||^2 |\lambda_i \lambda_j| + ||\eta - \eta^*|| |\lambda_i \lambda_j \lambda_k|)$$

(52)

$$+ \frac{1}{4!} \sum_{i=1}^{q+2} \sum_{j=1}^{q+2} \sum_{k=1}^{q+2} \sum_{\ell=1}^{q+2} \{ \nabla_{\lambda_i\lambda_j\lambda_k\lambda_k} L_n(\psi^*, \alpha) - \nabla_{\lambda_i\lambda_j\lambda_k\lambda_k} L_n(\psi^*, \alpha) \} \lambda_i \lambda_j \lambda_k \lambda_k \lambda_k$$

(53)

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

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

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

$$\mathcal{I}_n = \left( \begin{array}{cc} \mathcal{I}_{\eta\eta} & \mathcal{I}_{\eta\lambda} \\ \mathcal{I}_{\eta\lambda} & \mathcal{I}_{\lambda\lambda} \end{array} \right), \quad \mathcal{I}_{\eta\eta} : (p + q + 2) \times (p + q + 2), \quad \mathcal{I}_{\eta\lambda} : (p + q + 2) \times q_{\lambda}, \quad \mathcal{I}_{\lambda\lambda} : q_{\lambda} \times q_{\lambda},$$

where $q_{\lambda}$ represents the number of unique terms in $\sum_{i=1}^{q+2} \sum_{j=1}^{q+2} \sum_{k=1}^{q+2} \sum_{\ell=1}^{q+2} \lambda_i \lambda_j \lambda_k \lambda_k$. $\mathcal{I}_{\eta\eta}$ is given by $\mathcal{I}_{\eta\eta} = -n^{-1} \nabla_{\eta\eta} L_n(\psi^*, \alpha)$. For $\mathcal{I}_{\eta\lambda}$, let $A_{ij} = n^{-1} \nabla_{\eta\lambda_i \lambda_j} L_n(\psi^*, \alpha)$ and write
the term in (49) as \((n/2) \sum_{i=1}^{q+2} \sum_{j=1}^{q+2} (\eta - \eta^*)^T A_{ij} \lambda_i \lambda_j = n \sum_{i=1}^{q+2} \sum_{j=1}^{q+2} c_{ij} (\eta - \eta^*)^T A_{ij} \lambda_i \lambda_j\), where the values of \(c_{ij}\) are defined when we introduce \(\nabla_\theta f^*\) after (9). Then, by defining \(\mathcal{I}_{\eta \lambda n} = -(c_{11} A_{11} + \ldots + c_{qq} A_{qq} + c_{12} A_{12} + \ldots + c_{q-1,q} A_{q-1,q})/(1 - \alpha)\), the term in (49) equals 
\[-n(\eta - \eta^*)^T \mathcal{I}_{\eta \lambda n} \alpha(1-\alpha)v(\lambda)\]. For \(\mathcal{I}_{\lambda n}\), define \(B_{ijkl} = n^{-1}(8/\alpha! \nabla_{\lambda_i \lambda_j} \lambda_k \lambda_\ell \mathcal{L}_n(\psi^*, \alpha)\) so that the first term in (50) is written as \((n/8) \sum_{i=1}^{q+2} \sum_{j=1}^{q+2} \sum_{k=1}^{q+2} \sum_{\ell=1}^{q+2} B_{ijkl} \lambda_i \lambda_j \lambda_k \lambda_\ell = (n/2) \sum_{i=1}^{q+2} \sum_{j=1}^{q+2} \sum_{k=1}^{q+2} \sum_{\ell=1}^{q+2} c_{ijkl} B_{ijkl} \lambda_i \lambda_j \lambda_k \lambda_\ell\). Define \(\mathcal{I}_{\lambda n}\) such that the \((ij, k\ell)\) element of \(\mathcal{I}_{\lambda n}\) is 
\[-c_{ijkl} B_{ijkl}/\alpha^2(1 - \alpha)^2\], where the values of \(ij\) 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(\lambda)\] \(\mathcal{I}_{\lambda n} \alpha(1-\alpha)v(\lambda)\). With this definition of \(\mathcal{I}_{\lambda n}\), the expansion (48-50) is written as (12) in terms of \(\sqrt{m}(\psi, \alpha)\).

We now show that \(\mathcal{I}_n \rightarrow_p \mathcal{I}\). \(\mathcal{I}_{\eta n} \rightarrow_p \mathcal{I}_0\) holds trivially. For \(\mathcal{I}_{\lambda n}\), it follows from Lemma(3c) and the law of large numbers that \(A_{ij} \rightarrow_p -E[\nabla_\eta l(W; \psi^*, \alpha) \nabla_{\lambda_i \lambda_j} l(W; \psi^*, \alpha)]\), giving \(\mathcal{I}_{\eta \lambda n} \rightarrow_p E[\mathcal{I}_{\lambda n} \alpha(1-\alpha)] = \mathcal{I}_{\lambda n}\). For \(\mathcal{I}_{\lambda n}\), Lemma(3d) and the law of large numbers imply that \(\sum_{i=1}^{q+2} \sum_{j=1}^{q+2} \sum_{k=1}^{q+2} \sum_{\ell=1}^{q+2} B_{ijkl} \lambda_i \lambda_j \lambda_k \lambda_\ell \rightarrow_p -\sum_{i=1}^{q+2} \sum_{j=1}^{q+2} \sum_{k=1}^{q+2} \sum_{\ell=1}^{q+2} E[\nabla_{\lambda_i \lambda_j} l(W; \psi^*, \alpha) \nabla_{\lambda_k \lambda_\ell} l(W; \psi^*, \alpha)] \lambda_i \lambda_j \lambda_k \lambda_\ell\), where the factor \((8/4!) = 1/3\) in \(B_{ijkl}\) and the three derivatives on the right-hand side of Lemma(3d) cancel each other out. Therefore, we have \(\mathcal{I}_{\lambda n} \rightarrow_p E[\mathcal{I}_{\lambda n} \alpha(1-\alpha)]^2 = \mathcal{I}_{\lambda n}\) and \(\mathcal{I}_n \rightarrow_p \mathcal{I}\) follows.

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

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

\[
a^\top s(w) = a_{(\mu)}^\top s_{(\mu)} + a_{(\beta)}^\top s_{(\beta)}.
\]

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

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

\[
a_{(\mu)}^\top s_{(\mu)} = a_{(\mu)}^\top \sum_{t=1}^{T} H_{1t}^1 + (a_\alpha + a_{\lambda_{\mu\mu}}) \sum_{t=1}^{T} H_{1t}^2 + \frac{a_{\lambda_{\mu\sigma}}}{2} \sum_{t=1}^{T} \sum_{s \neq t} H_{1t}^1 H_{1s}^1 + a_{\lambda_{\mu\sigma}} \sum_{t=1}^{T} H_{1t}^3 + \frac{a_{\lambda_{\sigma\sigma}}}{2} \sum_{t=1}^{T} \sum_{s \neq t} H_{1t}^2 H_{1s}^2 = 0
\]
for all \( w \), where \( H_i^j \) for \( j = 1, 2, 3 \) is defined in (73) in Appendix B.2.

Because the above equation holds for all values of \( w \), with the property of the Hermite polynomials, we have \( a_\mu = 0, (a_\sigma + a_{\lambda_\mu}) = 0, a_{\lambda_\mu} = 0, a_{\lambda_\sigma} = 0 \). This implies that \( a_\omega = 0 \). Therefore, no multicollinearity exists in \( s(w) \) and \( \mathcal{T} \) 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 \( T_n := \mathcal{T}_n^{1/2} \sqrt{n}t(\hat{\psi}_\alpha, \alpha) \). Then, in view of (12), we have

\[
0_p(1) \leq L_n(\hat{\psi}_\alpha, \alpha) - L_n(\psi^*, \alpha) = T_n^{-1/2}S_n - \frac{1}{2}||T_n||^2 + R_n(\hat{\psi}_\alpha, \alpha)
\]

\[
= O_p(||T_n||) - \frac{1}{2}||T_n||^2 + (1 + ||\mathcal{T}_n^{-1/2}T_n||)2o_p(1)
\]

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

For part (b), noting that \( L_n(\psi^*, \alpha) = L_{0,\alpha}(\gamma_0^*, \theta_0^*) \), write

\[
LR_n = \max_{\alpha \in \mathcal{I}} \{ L_n(\hat{\psi}_\alpha, \alpha) - L_n(\psi^*, \alpha) \} - 2\{ L_{0,\alpha}(\gamma_0^*, \theta_0^*) \}.
\] 

(55)

Define

\[
S_n = \begin{pmatrix} S_{\eta_\mu} \\ S_{\lambda_\mu} \end{pmatrix} := \begin{pmatrix} n^{-1/2} \sum_{i=1}^n s_\eta(W_i) \\ n^{-1/2} \sum_{i=1}^n s_\lambda(W_i) \end{pmatrix}, \quad S_{\lambda, \eta} := \mathcal{T}_{\lambda, \eta} \mathcal{T}_n^{-1} S_{\eta_\mu}, \quad G_{\lambda, \eta} := \mathcal{T}_{\lambda, \eta}^{-1} S_{\lambda, \eta}, \quad t_{\eta, \lambda} := t_\eta - \mathcal{T}_n \mathcal{T}_n^{-1} t_\lambda(\lambda, \alpha),
\]

and split the quadratic form in (12) to obtain

\[
2\{ L_n(\psi, \alpha) - L_n(\psi^*, \alpha) \} = B_n(\sqrt{n}t_{\eta, \lambda}) + C_n(\sqrt{n}t_\lambda(\lambda, \alpha)) + R_n(\psi, \alpha),
\] 

(56)

where

\[
B_n(t_{\eta, \lambda}) = 2t_{\eta, \lambda}^T S_{\eta_\mu} - t_{\eta, \lambda}^T \mathcal{T}_n t_{\eta, \lambda}, \quad C_n(t_\lambda) = 2t_\lambda^T S_{\lambda, \eta} - t_\lambda^T \mathcal{T}_n t_\lambda
\]

(57)

\[
G_{\lambda, \eta} = \mathcal{T}_{\lambda, \eta} G_{\lambda, \eta} \mathcal{T}_n - (t_\lambda - G_{\lambda, \eta})^T \mathcal{T}_{\lambda, \eta} (t_\lambda - G_{\lambda, \eta}),
\]
with $G_{\lambda, \eta} \overset{d}{=} G_{\lambda, \eta} = (I_{\lambda, \eta})^{-1} S_{\lambda, \eta}$ and $S_{\lambda, \eta} \overset{d}{=} S_{\lambda, \eta} \sim N(0, I_{\lambda, \eta})$. In addition, $R_n(\hat{\psi}_\alpha, \alpha) = o_p(1)$ holds from Proposition 3(a) and $\sqrt{n} t(\hat{\omega}_\alpha, \alpha) = O_p(1)$.

Because $\Delta(\gamma, \alpha) f(x; \gamma^*, \alpha)$ is identical to $\Delta f(x; \psi^*, \alpha)$, a standard analysis gives $2[L_0.n(\gamma_{0}, \theta_0) - L_{0.n}(\gamma_{0}, \theta_0)] = \max_{t_n} B_n(\sqrt{n} t_{\gamma, \alpha}) + o_p(1)$. Note that the possible values of both $\sqrt{n} t_{\gamma, \alpha}$ and $\sqrt{n} \omega_{\gamma, \alpha}$ approach $\mathbb{R}$.

Therefore, in view of (56) and (57), we can write equation (55) as

$$LR_n = C_n(\sqrt{n} \omega_{\gamma, \alpha}) + o_p(1),$$

(58)

where $\hat{\lambda}$ is as 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) holds because of Proposition 2(a). Second, Assumption 3 of Andrews (1999) holds with $B_T = n^{1/2}$ and $T = n$ because $S_{\lambda, \eta} \overset{d}{=} S_{\lambda, \eta} \sim N(0, I_{\lambda, \eta})$ and $I_{\lambda, \eta}$ 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 $\alpha(1 - \alpha) v(\Theta_{\lambda})$ is locally equal to $\Lambda_{\lambda}$. Therefore, it follows from Theorem 3(c) of Andrews (1999) that $C_n(\sqrt{n} \omega_{\gamma, \alpha}) \overset{d}{\to} (i_{\lambda})^1 I_{\lambda, \eta} i_{\lambda}$, where $i_{\lambda}$ is defined by (16).

Proof of Proposition 3

Under $H_{2,0}$, we obtain $\theta_{M_0 + 1} \in \mathbb{Y}_{2h}$.

$$E[\{\nabla_{\alpha_k} \log f_{M_0 + 1}(W_i, \theta_{M_0 + 1})\}^2] = \int \left\{ f(w; \mu_\alpha, \sigma_\alpha) - f(w; \mu_\alpha^{*}, \sigma_\alpha^{*}) \right\}^2 dw$$

$$= \int \sum_{j=1}^{M_0} \alpha_j^* f(w; \mu_j^*, \sigma_j^*) \sum_{j=1}^{M_0} \alpha_j^* f(w; \mu_j^*, \sigma_j^*) \frac{f(w; \mu_\alpha, \sigma_\alpha)}{\sum_{j=1}^{M_0} \alpha_j^* f(w; \mu_j^*, \sigma_j^*)} \frac{f(w; \mu_\alpha, \sigma_\alpha)}{\sum_{j=1}^{M_0} \alpha_j^* f(w; \mu_j^*, \sigma_j^*)} dw$$

(59)

The latter two terms on the right-hand side of (59) are bounded because $f(w; \mu_\alpha, \sigma_\alpha) / \sum_{j=1}^{M_0} \alpha_j^* f(w; \mu_j^*, \sigma_j^*) \leq (1/\alpha_j^*)$ for any $w$ and $f(w; \mu, \sigma)$ 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 $\max_j \alpha_j \leq \sum_{j=1}^{M_0} \alpha_j \leq M_0 \max_j \alpha_j$, we obtain

$$\frac{1}{M_0 \max_j \alpha_j} \left\{ f(w; \mu_\alpha, \sigma_\alpha) \right\}^2 \leq \frac{\sum_{j=1}^{M_0} \alpha_j^* f(w; \mu_j^*, \sigma_j^*) \sum_{j=1}^{M_0} \alpha_j^* f(w; \mu_j^*, \sigma_j^*)}{\sum_{j=1}^{M_0} \alpha_j^* f(w; \mu_j^*, \sigma_j^*)} \leq \frac{\sum_{j=1}^{M_0} \alpha_j^* f(w; \mu_j^*, \sigma_j^*)}{\max_j \alpha_j f(w; \mu_j^*, \sigma_j^*)}.$$

Without loss of generality, we assume that $\sigma_{M_0} = \max\{\sigma_1^*, \ldots, \sigma_{M_0}^*\}$ and that the maximum is unique. Then, there exists $M \in (0, \infty)$, such that $\max_j \alpha_j f(w; \mu_j^*, \sigma_j^*) = \alpha_{M_0}^* f(w; \mu_{M_0}^*, \sigma_{M_0}^*)$.
when $|y_t| > M \forall t = 1, \ldots, T$. Note that

$$
\left\{ f(w; \mu_h, \sigma^2_h) \right\}^2 = \prod_{t=1}^{T} \frac{(\sigma^2_{M_0})^2}{(2\pi)^{T/2} \sigma^2_h} \exp \left\{ -\frac{1}{\sigma^2_h} (y_t - \mu_h)^2 + \frac{1}{2(\sigma^2_{M_0})^2} (y_t - \mu^*_{M_0})^2 \right\}
$$

and

$$
= \frac{(\sigma^2_{M_0})^{2T}}{(2\pi)^{T/2} \sigma^2_h} \exp \left\{ -\frac{1}{\sigma^2_h} \sum_{t=1}^{T} (y_t - \mu_h)^2 + \frac{1}{2(\sigma^2_{M_0})^2} \sum_{t=1}^{T} (y_t - \mu^*_{M_0})^2 \right\}. \tag{60}
$$

The stated result follows because the integral of the right-hand side of (60) over $|y| \geq M$ is finite if $\sigma^2_h/\sigma^2_{M_0} < 2$ and infinite if $\sigma^2_h/\sigma^2_{M_0} > 2$. When $\sigma^2_h/\sigma^2_{M_0} = 2$, it is finite if $\mu_h = \mu^*_{M_0}$ and infinite if $\mu_h \neq \mu^*_{M_0}$.

**Proof of Proposition 3.** 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. Chen and Tan (2009) provide the consistency proof for the PMLE for a multivariate normal mixture under their conditions C1–C3 for the penalty function. However, Alexandrovich (2014) identifies a weakness in the proof of Chen and Tan (2009) and provides an alternative consistency proof by strengthening condition C3 of Chen and Tan (2009). Their $p_n(G)$ and $\hat{p}_n(G)$ correspond to our $\hat{p}_n(\theta_2)$ and $p_n(\theta_2)$, respectively; consequently, the conditions C1 and C2 in Chen and Tan (2009) and a version of condition C3 strengthened by Alexandrovich (2014) can be stated in our notation as follows:

**C1.** The penalty function is written as $\hat{p}_n(\theta_M) = \sum_{j=1}^{M} p_n(\sigma^2_j)$.

**C2.** For any fixed $\theta_M$ with $\sigma^2_j > 0$ for $j = 1, 2, \ldots, M$, we have $\hat{p}_n(\theta_M) = o(n)$ and sup $\theta_M \in \Theta_M \max \{0, \hat{p}_n(\theta_M)\} = o(n)$. In addition, $\hat{p}_n(\theta_M)$ is differentiable with respect to $\theta_M$ and as $n \to \infty$, $\nabla_{\theta_M} \hat{p}_n(\theta_M) = o(n^{1/2})$ at any fixed $\theta$ such that $\sigma^2_j > 0$ for $j = 1, 2, \ldots, M$.

**A version of C3 by Alexandrovich (2014).** For a sufficiently large $n$, $p_n(\sigma^2_j) \leq (\frac{\sqrt{n}}{\sqrt{\log \log n}}) \log(\sigma^2_j)$ when $\sigma^2_j < cn^{-2}$ for some $c > 0$.

The consistency of the PMLE, $\hat{\theta}_{M_0}$ and $\hat{\theta}_{M_0+1}$, 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 $\zeta > 0$ implying $a_n = o(n)$ or $o(n^{1/2})$, and $\Delta_2 \hat{p}_n(\theta) = -a_n(-\sigma^2_0/(\sigma^2_j) + 1/\sigma^2_j) = o(n^{1/4})$ if $\sigma^2_j > 0$. For C3, suppose that $\sigma^2_j < n^{-2}$. Then, because $a_n = o(n^{1/4})$ and $a_n > 0$, $p_n(\sigma_j) = -a_n(\sigma^2_j + 2 \log(\sigma_j/\sigma_0) - 1) < -c_n(n^{9/4} - 2n^{5/4} \log(n\sigma_0) - n^{1/4}) < (\frac{\sqrt{n}}{\sqrt{\log \log n}}) \log(\sigma^2_j)$ when $n$ is sufficiently large, where $c_n$ is a sequence of positive numbers that are bounded. Therefore, $\hat{p}_n(\theta_M)$ satisfies the above three conditions, and the stated result 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 $\mathcal{N}^*_h \subset \Theta_{M_0+1}(\epsilon)$ be a sufficiently small closed neighborhood of $\Upsilon^*_{1h}$ such that $\alpha_h, \alpha_{h+1} > 0$ holds and $\Upsilon^*_{1k} \not\subset \mathcal{N}^*_h$ if $k \neq h$. Consider the following one-to-one reparameterization from the $(M_0 + 1)$-component model parameter $\theta_{M_0+1} = (\alpha_1, \ldots, \alpha_{M_0}, \theta^1, \ldots, \theta^h, \theta^h_{h+1}, \ldots, \theta^M_{M_0+1}, \gamma^T)^T$. Similar to [5], the one-to-one reparameterization for testing the null hypothesis $H_{0,1h}$ is given by

$$
\begin{pmatrix}
\lambda_h \\
\nu_h
\end{pmatrix} := \begin{pmatrix}
\theta_h - \theta_{h+1} \\
\tau \theta_h + (1 - \tau) \theta_{h+1}
\end{pmatrix}
$$

so that

$$
\begin{pmatrix}
\theta_h \\
\theta_{h+1}
\end{pmatrix} = \begin{pmatrix}
\nu + (1 - \tau) \lambda \\
\nu - \tau \lambda
\end{pmatrix}.
$$

and $\alpha_j$ is reparameterized for $j = 1, 2, \ldots, M_0$ 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 $\tau$ as

$$
\psi_{h,\tau} = (\eta^T, \chi^T)\text{ with } \eta = (\pi_1, \ldots, \pi_{M_0-1}, \theta^1, \ldots, \theta^{h-1}, \nu_h, \theta^h_{h+1}, \ldots, \theta^M_{M_0+1}, \gamma)^T.
$$

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

$$
f_{M_0+1}^h(w; \psi_{h,\tau}, \tau) = \pi_h g^h(w, \psi_{h,\tau}, \tau) + \sum_{j=1}^{h-1} \pi_j f(w; \gamma, \theta_j) + \sum_{j=h}^{M_0} \pi_{j+1} f(w; \gamma, \theta_{j+1}),
$$

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

$$
g^h(w, \psi_{h,\tau}, \tau) = \tau f(w; \gamma, \nu_h + (1 - \tau) \lambda_h) + (1 - \tau) f(w; \gamma, \nu_h - \tau \lambda_h).
$$

(61)

Define the local PMLE of $\psi_{h,\tau}$ by

$$
\hat{\psi}_{h,\tau} := \arg \max_{\psi_{h,\tau} \in \mathcal{N}^*_h} L^h_n(\psi_{h,\tau}, \tau) + \sum_{j=1}^{M_0} p_n(\sigma^2_j(\psi_{h,\tau}, \tau)),
$$

(62)

where $L^h_n(\psi_{h,\tau}, \tau) := \sum_{i=1}^{N} \log g^h(W_i; \psi_{h,\tau}, \tau)$ and $\sigma^2_j(\psi_{h,\tau}, \tau)$ is the value of $\sigma^2_j$ implied by the values of $\psi_{h,\tau}$ and $\tau$. Because $\psi^*_h, \tau$ is the only parameter value in $\mathcal{N}^*_h$ that generates the true density, $\hat{\psi}_{h,\tau} - \psi^*_h, \tau = o_p(1)$ follows Proposition[4]

For $\epsilon \in (0, 1/2)$, define the LRTS for testing $H_{0,1h}$ as $LR^M_n := \max_{\tau \in \epsilon, 1-\epsilon} \{2(L^h_n(\hat{\psi}_{h,\tau}, \tau) - 48$
Because \( \sigma_j^2 = \sigma_j^2 \) under the null hypothesis (cf. Proposition 3(a)), we have \( \rho_n(\theta_{M_0+1}) = o_p(1) \) by Assumption 3(c), and \( \text{PL}_{n}^{M_0,h} - \text{LR}_{n}^{M_0,h} = o_p(1) \) follows for \( h = 1, \ldots, M_0 \).

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

\[
(\text{LR}_{n}^{M_0,1}, \ldots, \text{LR}_{n}^{M_0,M_0})^\top \xrightarrow{d} (\tilde{t}_{\lambda}^{1}, \ldots, \tilde{t}_{\lambda}^{M_0})^\top \times (\tilde{\lambda}_{0,\lambda}^{1}, \ldots, \tilde{\lambda}_{0,\lambda}^{M_0})^\top. \tag{63}
\]

Observe that \( L_n^h(\psi, \tau, \lambda) = L_n^h(\psi^*, \tau, \lambda) \) admits the same expansion as \( L_n(\psi, \alpha) - L_n(\psi^*, \alpha) \) in (13) and (56) when \((\alpha, t(\psi, \alpha), t_\lambda(\lambda, \alpha), S_n, G_n, \mathcal{I}_n, R_n(\psi, \alpha)) \) is replaced with \((\tau, t^h(\psi, \tau), t_\lambda^h(\lambda, \tau), S_n^h, G_n^h, \mathcal{I}_n^h, R_n^h(\psi, \tau)) \) where \((S_n^h, \mathcal{I}_n^h) \) is defined similarly to \((S_n, \mathcal{I}_n) \) but \((\tilde{s}_n, \tilde{s}_{\lambda,\lambda}) \) and \( G_n^h := (\mathcal{I}_n^h)^{-1} S_n^h \). Applying the proof of Proposition 3, we have \( S_n^h \xrightarrow{d} S^h \sim N(0, \mathcal{I}^h) \) and \( \mathcal{I}_n^h \xrightarrow{p} \mathcal{I}^h \). Then, (62) follows from the proofs of Propositions 3 and 5 for each local PMLE when \((G_n, \tilde{t}_\lambda, \mathcal{I}_{\lambda,\eta}) \) is replaced with \((G_n^h, \tilde{t}_\lambda^h, \mathcal{I}_{\lambda,\eta}^h) \) and the results are collected; note that \((S_n^h, \ldots, S_n^{M_0}) \xrightarrow{d} (S^1, \ldots, S^{M_0}) \).

**Proof of Proposition 8** The proof is similar to that of Proposition 7 in Kasahara and Shimotsu (2015). Let \( \omega^h \) denote the sample counterpart of \((\tilde{\lambda}_{0,\lambda}^h)_{0}^\top \mathcal{I}_{\lambda,\eta}^h \tilde{\lambda}_{0,\lambda}^h \) in Proposition 7 such that the LRTS satisfies

\[
2\{L_n(\hat{\psi}_{h,\tau,\lambda}) - L_n(\hat{\theta}_{M_0})\} = \omega^h + o_p(1),
\]

where \( \hat{\psi}_{h,\tau} \) is the local PMLE as defined in (62) and \( \omega^h \) is defined similarly to \( C_n(\sqrt{\tilde{t}_\lambda(\hat{\lambda}_{0,\alpha}))}) \) in (58) but with \((t_\lambda(\lambda, \alpha), S_n, G_n, \mathcal{I}_n) \) replaced with \((t_\lambda^h(\lambda, \tau), S_n^h, G_n^h, \mathcal{I}_{\lambda,\eta}^h) \) in the proof of Proposition 7.

First, we show that \( M_n^{h(1)}(\tau_0) = 2\{PL_n^{h}(\hat{\theta}_{M_0+1}^{h(1)}(\tau_0), \tau_0) - L_{0,n}(\hat{\theta}_{M_0})\} = \omega^h + p(\tau_0) + o_p(1) \).

Define \( \hat{\theta}_{M_0+1}^{h(1)}(\tau_0) = \{ \theta \in \Psi_h : \alpha_h/(\alpha_h + \alpha_{h+1}) = \tau_0 \} \).

Because \( \hat{\theta}_{M_0+1}^{h(1)}(\tau_0) \) is the only value of \( \hat{\theta}_{M_0+1} \) that yields the true density if \( \hat{\theta}_{M_0+1} \in \Psi_h \) in (20) and \( \alpha_h/(\alpha_h + \alpha_{h+1}) = \tau_0 \), \( \hat{\theta}_{M_0+1}(\tau_0) \) equals a reparameterized local PMLE in the neighborhood of \( \hat{\theta}_{M_0+1}(\tau_0) \), and \( \hat{\theta}_{M_0+1}(\tau_0) = o_p(1) \) holds in view of Proposition 6. Furthermore, by the consistency of \( \hat{\tau}^{h(1)} \) and \( \alpha_0 = O(1) \), we have \( \hat{\theta}_{M_0+1}^{h(1)}(\tau_0) \xrightarrow{p} \tau_0 \). Therefore, \( M_n^{h(1)}(\tau_0) = \omega^h + p(\tau_0) + o_p(1) \) holds from repeating the proof of Proposition 2. Finally, \( EM_n^{h(1)} \xrightarrow{d} \max_{h=1}^{M_0} \omega^h \) holds because \( \{0.5\} \times T = 0 \) and \( p(0.5) = 0 \).

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

\[
\text{PL}_n(\hat{\theta}_{M_0+1}^{h(K)}(\tau_0), \tau^{h(K)}(\tau_0)) > \text{PL}_n(\hat{\theta}_{M_0+1}^{h(1)}(\tau_0), \tau^{h(1)}(\tau_0)). \tag{64}
\]

Therefore, it follows from Theorem 1 of Chen and Tan (2009), Lemma 4 in Appendix 8, and induction that \( \hat{\theta}_{M_0+1}^{h(K)}(\tau_0) = \hat{\theta}_{M_0+1}^{h(1)}(\tau_0) + o_p(1) \) for any finite \( K \). Let \( \hat{\theta}_{M_0+1} \) be the maximizer of \( \text{PL}_{M_0+1}(\theta_{M_0+1}, \tau^{h(K)}(\tau_0)) \) under the constraint of \( \alpha_h/(\alpha_h + \alpha_{h+1}) = \tau^{h(K)}(\tau_0) \) in an arbitrary small neighborhood of \( \hat{\theta}_{M_0+1}(\tau^{h(K)}) \). Then, \( 2\{PL_n(\hat{\theta}_{M_0+1}^{h(K)}(\tau_0), \tau^{h(K)}(\tau_0)) - L_{0,n}(\hat{\theta}_{M_0})\} = \omega^h + p(\tau_0) + o_p(1) \) holds from the definition of \( \hat{\theta}_{M_0+1} \) and \( \hat{\theta}(\hat{\theta}_{M_0+1}) \xrightarrow{p} 0 \) by repeating the proof.
of Proposition [7] It also follows from the consistency of \( \hat{\theta}^{h(K)}_{M_0+1}(\tau_0) \) that \( PL_n(\hat{\theta}^h_{M_0+1}(\tau_0), \tau^h(K)(\tau_0)) \geq PL_n(\hat{\theta}^h_{M_0+1}(\tau_0), \tau^h(K)(\tau_0)) + o_p(1) \). Therefore, in view of (64), we have

\[
PL_n(\hat{\theta}^h_{n}, \tau^h(K)(\tau_0)) \geq PL_n(\hat{\theta}^h_{M_0+1}(\tau_0), \tau^h(K)(\tau_0)) + o_p(1) \geq PL_n(\hat{\theta}^{h(1)}_{M_0+1}(\tau_0), \tau^{h(1)}(\tau_0)). \tag{65}
\]

Finally, because \( 2\{PL_n(\hat{\theta}^h_{M_0+1}, \tau^h(K)(\tau_0)) - L_{0,n}(\hat{\theta}_{M_0})\} = \omega^h_n + p(\tau_0) + o_p(1) \) and \( 2\{PL_n(\hat{\theta}^{h(1)}_{M_0+1}(\tau_0), \tau^{h(1)}(\tau_0)) - L_{0,n}(\hat{\theta}_{M_0})\} = \omega^{h(1)}_n + p(\tau_0) + o_p(1) \), it follows from (65) that \( M^h(K)(\tau_0) = 2\{PL_n(\hat{\theta}^h_{M_0+1}(\tau_0)) - L_{0,n}(\hat{\theta}_{M_0})\} \) for all \( h \). The stated result then follows from the definition of \( E M^h(K) \) and \( \{0.5\} \in \mathcal{T} \).

**Proof of Proposition 9** Let \( \psi_n = (\nu^*, \lambda_n^\top) \) be the value of \( \psi \) under \( H_{1,n} : \vartheta = \vartheta_{2,n} \), and let \( h = (0^\top, \lambda_n^\top) \), where \( \lambda_n \) is defined by (31). Let \( P_{\vartheta} \) be the probability measure on \( \{W_i\}_{i=1}^n \) under \( \vartheta \). Denote the log-likelihood ratio of \( P_{\vartheta_{2,n}} \) to \( P_{\vartheta_2} \) by \( \log \left( \frac{dP_{\vartheta_{2,n}}}{dP_{\vartheta_2}} \right) = L_n(\psi_n, \alpha^*) - L_n(\psi^*, \alpha^*) \). Then, it follows from (12) and Proposition 3 that

\[
\log \frac{dP_{\vartheta_{2,n}}}{dP_{\vartheta_2}} = hS_n - h^\top I h/2 + o_p(1) \quad \text{under} \quad P_{\vartheta_2}. \tag{66}
\]

Furthermore, because \( S_n \xrightarrow{d} N(0, I) \) under \( P_{\vartheta_2} \), \( \frac{dP_{\vartheta_{2,n}}}{dP_{\vartheta_2}} \) converges in distribution under \( P_{\vartheta_2} \) to \( \exp(N(\mu, \sigma^2)) \) with \( \mu = -(1/2)h^\top I h \) and \( \sigma^2 = h^\top I h \) so that \( E(\exp(N(\mu, \sigma^2))) = 1 \). Consequently, \( P_{\vartheta_{2,n}} \) is mutually contiguous with respect to \( P_{\vartheta_2} \) 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

\[
\left( \begin{array}{c}
S_n \\
\log \frac{dP_{\vartheta_{2,n}}}{dP_{\vartheta_2}}
\end{array} \right) \xrightarrow{d} \mathcal{N}\left( \begin{array}{c}
0 \\
-\frac{1}{2}h^\top I h
\end{array} \right), \quad \left( I, I h \right) \quad \text{under} \quad P_{\vartheta_2}
\]

and

\[
S_n \xrightarrow{d} N(I h, I) \quad \text{under} \quad P_{\vartheta_{2,n}}
\]

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 \( P_{\vartheta_{2,n}} \) if we replace \( S_{\lambda, \eta} \xrightarrow{d} S_{\lambda, \eta} \) with \( S_{\lambda, \eta} \xrightarrow{d} S_{\lambda, \eta} + (I \lambda - I \eta \lambda I^{-1} I \lambda) h \), and the stated result follows.

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

\[
\sup_{\vartheta_M \in \Theta_{\vartheta_M}} |\hat{Q}_{n}^M(\vartheta_M) - Q^M(\vartheta_M)| = o_p(1), \tag{67}
\]
and Assumption 4(a) and (b) and the standard consistency proof (e.g., Theorem 2.1 of Newey and McFadden, 1994) give \( \hat{\theta}_M \) to \( \theta_M^* \) for \( M < M_0 \). Furthermore, by Assumption 5(c), 
\( n^{-1/4} \sqrt{\eta_n}(\hat{\theta}_M) = o_p(1) \) uniformly over \( \Theta_{\theta_M^*} \), and it follows from the argument in Theorem 3.2 of White (1982) that 
\[
\sqrt{n}(\hat{\theta}_M - \theta_M^*) \xrightarrow{p} N(0, A_M^{M^*^{-1}}B_M^{M^*^{-1}}A_M^{M^*^{-1}}). \tag{68}
\]

Then, from (67), (68), and the mean value expansion, we have 
\( \hat{Q}_n^M(\hat{\theta}_M) - Q^M(\theta_M^*) = O_p(n^{-1/2}) \), and 
\[
\frac{PLR_n(M)}{n} := \hat{Q}_{n+1}(\hat{\theta}_{M+1}) - \hat{Q}_n^M(\hat{\theta}_M) = Q_{n+1}^M(\theta_{M+1}^*) - Q^M(\theta_M^*) + o_p(1).
\]
Because \( Q_{n+1}^M(\theta_{M+1}^*) - Q^M(\theta_M^*) > 0 \) by Assumption 4(f), \( PLR_n(M)/n \to \infty \) as \( n \to \infty \). By Lemma 2, \( -n^{-1} \log q_n = o(1) \) and \( c_{1-q_n}^M - c_{1-q_n}^M = o_p(1) \) implies that \( n^{-1/2}c_{1-q_n}^M = o_p(1) \). Therefore, when \( M < M_0 \), we have 
\[
Pr(PLR_n(M) > c_{1-q_n}^M) = Pr(PLR_n(M)/n > c_{1-q_n}^M/n) \to 1 \quad \text{as} \quad n \to \infty.
\]

When \( M = M_0 \), because \( PLR_n(M_0) = O_p(1) \) by Proposition 2 and \( c_{1-q_n}^M \to \infty \) by \( q_n = o(1) \), 
\[
Pr(PLR_n(M_0) > c_{M_0}^M) \to 0 \quad \text{as} \quad n \to \infty. \]

B Auxiliary results and their proofs

B.1 Lemmas

Lemma 1. For any \( M < \infty \), 
\[
Pr(-\log n + \ell(W_{i*}; \hat{Y}_{i*}, s_{i*}^2) < M) \to 0 \quad \text{as} \quad n \to \infty.
\]

Proof of Lemma 2. Because \( \sum_{t=1}^T \frac{(Y_t - \mu)^2}{s_t^2} = T - 1 \) when \( i = i^* \), we have
\[
-\log n + \ell(W_{i*}; \hat{Y}_{i*}, s_{i*}^2) = -\log n - \frac{T}{2} \log s_{i*}^2 - \frac{T}{2} \log(2\pi) - \frac{T - 1}{2}
\]
\[
= -\log \left(Cn(s_{i*}^2)^{T/2}\right), \tag{69}
\]
for some positive constant \( C \).

Therefore, to prove the stated result, it suffices to show that for any \( \epsilon > 0 \), 
\[
Pr(n(s_{i*}^2)^{T/2} > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.
\]
Given the property of the first-order statistic, the distribution of \( s_{i*}^2 \) 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
\[
Pr(n(s_{i*}^2)^{T/2} > \epsilon) = [1 - F_{T-1}((\epsilon/n)^{2/T})]^n.
\]
When \( T = 3 \), \( 1 - F_{T-1}(s) = e^{-s/2} \) and 
\[
Pr(n(s_{i*}^2)^{T/2} > \epsilon) = e^{-Cn^{1/3}} \text{ for some positive constant } C,
\]

51
and therefore, \( \Pr(n(s_n^2)^{T/2} > \epsilon) \to 0 \) as \( n \to \infty \), and the stated result follows.

For general \( T \geq 2 \), write

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

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

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

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

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

where \( f_k \) is the PDF of the \( \chi \)-square distribution with \( k \) degrees of freedom. Note that

\[
f_{T-1}((\epsilon x)^{2/T}) = \frac{1}{2^{(T-1)/2}(\pi(T-1)/2)!} (\epsilon x)^2/T(T-1)/2 e^{-(\epsilon x)^2/2};
\]

then,

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

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

\[\Box\]

**Lemma 2.** Suppose that the assumptions in Proposition 10 hold. If \( -n^{-1} \ln q_n = o(1) \), then \( n^{-1} c_1^{M_{q_n}} = o(1) \).

**Proof.** For brevity of notation, write \( c_n = c_n^{M_{q_n}} \). By Theorem 2.1 of Foutz and Srivastava (1977), \( PLR_n(M) \xrightarrow{d} \sum_{j=1}^{K} b_j \chi_j^2 \) for \( 0 < b_j < \infty \) and \( K \) finite, where \( \chi_1^2, \ldots, \chi_K^2 \) are independent chi-square random variables with one degree of freedom. Then, we have

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

with \( b^* = \arg \max \{ b_1, \ldots, b_K \} \), where the last inequality follows from a Chernoff bound:

\[
\Pr \left( \chi_j^2 \geq \frac{c_n}{b^*} \right) \leq \frac{E[\exp(t(\chi_j^2-1))]}{\exp(t(\chi_j^2-1))} = \frac{1}{\sqrt{1-2t}} \exp \left( -t \frac{c_n}{b^*} \right) \quad \text{for} \quad 0 < t < \frac{1}{2}.
\]

Therefore, \( -\frac{1}{n} \ln q_n \geq -\frac{1}{n} \ln \left( \frac{K}{\sqrt{1-2t}} \right) + \frac{1}{20} \frac{c_n}{b^*} \), and the stated result follows.

\[\Box\]

52
Lemma 3. Suppose that \( g(w; \psi, \alpha) \) is defined as [3], where \( \psi = (\eta^\top, \lambda^\top)^\top \). Let \( g^*, \nabla g^*, \) and \( \nabla \log g^* \) denote \( g(\mathbf{W}; \psi, \alpha), \nabla g(\mathbf{W}; \psi, \alpha), \) and \( \nabla \log g(\mathbf{W}; \psi, \alpha) \) evaluated at \( (\psi, \alpha) \), respectively. Let \( \nabla f^* \) denote \( \nabla f(\mathbf{W}; \gamma^*, \theta^*) \). The following statements hold.

(a) For \( l = 0, 1, \ldots, \nabla (\lambda \otimes \eta)^\top g^* = 0; \)

(b) \( \nabla (\lambda^\otimes_2)^\top g^* = (1 - \alpha) \nabla (g^\otimes_2)^\top f^*; \)

(c) \( \mathbb{E}[\nabla_{\lambda_1, \lambda_1} \log g^*] = 0, \mathbb{E}[\nabla_{\lambda_1, \lambda_2} \log g^*] = 0, \) and \( \mathbb{E}[\nabla_{\eta \lambda, \lambda_1} \log g^*] = -\mathbb{E}[\nabla_{\eta} \log g^* \nabla_{\lambda_1} \lambda_1 \log g^*]; \)

(d) \( \mathbb{E}[\nabla_{\lambda_1, \lambda_2} \lambda_1 \log g^*] = -\mathbb{E}[\nabla_{\lambda_1, \lambda_2} \lambda_1 \log g^* + \nabla_{\lambda_1, \lambda_2} \log g^* \nabla_{\lambda_1} \lambda_2 \log g^*]; \)

Proof of Lemma 3. Recall that

\[
g(w; \psi, \alpha) = \alpha f(w; \gamma, \nu + (1 - \alpha) \lambda) + (1 - \alpha) f(w; \gamma, \nu - \alpha \lambda).\]

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

\[
\nabla (\lambda^\otimes_2)^\top g = \nabla_{\lambda} \left( \alpha \nabla_{(\gamma, \theta)}^\otimes f(w; \gamma, \nu + (1 - \alpha) \lambda) + (1 - \alpha) \nabla_{(\gamma, \theta)}^\otimes f(w; \gamma, \nu - \alpha \lambda) \right)_{\nu=\theta^*, \lambda=0}
\]

\[
= \alpha (1 - \alpha) \nabla_{(\gamma^\otimes_1, \theta^\otimes_1)} f(w; \gamma, \nu + (1 - \alpha) \lambda) - \alpha (1 - \alpha) \nabla_{(\gamma^\otimes_1, \theta^\otimes_1)} f(w; \gamma, \nu - \alpha \lambda)_{\nu=\theta^*, \lambda=0}
\]

\[
= 0.
\]

To show part (b), note that

\[
\nabla (\lambda^\otimes_2)^\top g = \nabla_{\lambda} \left( \alpha (1 - \alpha) \nabla_{\lambda} f(w; \gamma, \nu - \alpha (1 - \alpha) \lambda) + (1 - \alpha) \nabla_{\lambda} f(w; \gamma, \nu - \alpha \lambda) \right)
\]

\[
= \alpha (1 - \alpha)^2 \nabla (\lambda^\otimes_2)^\top f(w; \gamma, \nu + (1 - \alpha) \lambda) + (1 - \alpha) \nabla (\lambda^\otimes_2)^\top f(w; \gamma, \nu - \alpha \lambda)_{\nu=\theta^*, \lambda=0}
\]

\[
= \nabla (\lambda^\otimes_2)^\top f^*.
\]

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

\[
\int \{ \nabla_{\lambda_i} \log g(w; \psi, \alpha) + \nabla_{\lambda_i} \log g(w; \psi, \alpha) \nabla_{\lambda_j} \log g(w; \psi, \alpha) \} g(w; \psi, \alpha) dx = 0. \quad (71)
\]

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

\[ \square \]
Lemma 4. Suppose that the assumptions of Proposition 8 hold. If \( \vartheta_{M_0+1}(\tau_0) - \vartheta_{M_0+1}(\tau_0) = o_p(1) \) and \( \tau^{(K)} - \tau_0 = o_p(1) \) and \( \alpha_{M_0+1}(\tau_0) = o_p(1) \), then \( \alpha_{M_0+1}(\tau_0) - \alpha_{M_0+1}(\tau_0) = o_p(1) \).

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 \( \tau_0 \) from \( \vartheta_{M_0+1}(\tau_0) \) and \( \vartheta_{M_0+1}(\tau_0) \). We suppress \( Z \) for brevity. Let \( f_i(\gamma, \theta_j) \) and \( f_i(\theta_{M_0+1}) \) denote \( f(W_i; \gamma, \theta_j) \) in (2) and \( f_{M_0+1}(W_i; \theta_{M_0+1}) \) in (18), respectively. Applying a Taylor expansion to \( \alpha_{M_0+1}(\tau_0) = n^{-1} \sum_{i=1}^n w_{ih} \) and using \( \vartheta_{M_0+1} - \vartheta_{M_0+1} = o_p(1) \), we obtain

\[
\alpha_{M_0+1}(\tau_0) = \frac{1}{n} \sum_{i=1}^n \tau^{(K)}(\alpha_{M_0+1} + \alpha_{M_0+1}) f_i(\theta_{M_0+1}) f_i(\theta_{M_0+1}) = \frac{1}{n} \sum_{i=1}^n \tau_0 \alpha_{M_0+1} f_i(\theta_{M_0+1}) + o_p(1) = \tau_0 \alpha_{M_0+1} + o_p(1),
\]

where the last equality follows from \( \mathbb{E}[f_i(\gamma^*, \theta_j^*)/f_i(\theta_{M_0+1})] = 1 \) and the law of large numbers. A similar argument gives \( \alpha_{M_0+1}(\tau_0) = (1 - \tau_0) \alpha_{M_0+1} + o_p(1) \), and part (a) follows.

For part (b), define \( H(\tau) := \sum_{i=1}^n w_{ih}^{(K)} \log(\tau) + \sum_{i=1}^n w_{ih}^{(K)} \log(1 - \tau) = n \alpha_{M_0+1}^{(K)} \log(\tau) + n \alpha_{M_0+1}^{(K)} \log(1 - \tau) \), then \( \tau^{(K)} \) maximizes \( H(\tau) + p(\tau) \). \( H(\tau) \) is maximized at \( \tilde{\tau} := \alpha_{M_0+1}^{(K)} + \alpha_{M_0+1}^{(K)} = (\tau_0 \alpha_{M_0+1} + o_p(1))/\tau_0 \alpha_{M_0+1} + o_p(1) = \tau_0 + o_p(1) \). Observe that with \( \tilde{\tau} = \frac{\tau^{(K)} - \tau}{2} \),

\[
p(\tilde{\tau}) \leq p(\tilde{\tau}) - p(\tau^{(K)}) \leq H(\tau^{(K)}) - H(\tilde{\tau}) = H''(\tilde{\tau})(\tau^{(K)} - \tilde{\tau})^2,
\]

where the first inequality follows from \( p(\tau) \leq 0 \), the second inequality holds because \( \tau^{(K)} \) maximizes \( H(\tau) + p(\tau) \), and the last equality follows from expanding \( H(\tau^{(K)}) \) twice around \( \tilde{\tau} \) and noting that \( H''(\tilde{\tau}) = 0 \) because \( \tilde{\tau} \) maximizes \( H(\tau) \). Note that \( H''(\tau) = -n \times \left\{ \alpha_{M_0+1}^{(K)} + \alpha_{M_0+1}^{(K)} \right\} < 0 \)

and \( \inf_{\tau} H''(\tau) \geq -n(\alpha_{M_0+1}^{(K)} + \alpha_{M_0+1}^{(K)}) \). Therefore, in view of \( \tilde{\tau} - \tau_0 = o_p(1) \) and (72), we have \( (\tau^{(K)} - \tilde{\tau})^2 \leq p(\tilde{\tau})/H''(\tilde{\tau}) = O_p(n^{-1}) \), and part (b) holds.

\[\square\]

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 the supplementary material of Kasahara and Shimotsu (2015), the derivative of \( \{ \frac{1}{n} \phi(\tau) \} \) is

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

54
Let
\[ f^* = f(w; \gamma^*, \theta^*), \nabla f^* = \nabla f(w; \gamma^*, \theta^*), H_{i,t}^j = \frac{1}{\sigma^j} H_i^j \left( \frac{y_{it} - x_{it}^T \beta - z_{it}^T \gamma^* - \mu^*}{\sigma^*} \right); \quad (73) \]
then, the first-order derivatives of the density functions are
\[ \nabla_{\mu} f^* = f^* \sum_{t=1}^{T} \frac{1}{\sigma} H_{i,t}^{1*}; \nabla_{\gamma} f^* = f^* \sum_{t=1}^{T} \frac{1}{2 \sigma^2} H_{i,t}^{2*}; \]
\[ \nabla_{\beta} f^* = f^* \sum_{t=1}^{T} \frac{1}{\sigma} H_{i,t}^{1*} x_{it}; \nabla_{\gamma} f^* = f^* \sum_{t=1}^{T} \frac{1}{\sigma} H_{i,t}^{1*} z_{it}. \]
The score function defined in (9) is then written in terms of the Hermite polynomials:
\[ s_{\eta} = \begin{pmatrix} s_{\mu} \\ s_{\sigma} \\ s_{\beta} \\ s_{\gamma} \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{T} H_{i,t}^{1*} \\ \sum_{t=1}^{T} H_{i,t}^{2*} \\ \sum_{t=1}^{T} H_{i,t}^{1*} x_{it} \\ \sum_{t=1}^{T} H_{i,t}^{1*} z_{it} \end{pmatrix}, \quad s_{\lambda\lambda} = \begin{pmatrix} s_{\lambda_{\mu}\lambda_{\mu}} \\ s_{\lambda_{\mu}\lambda_{\sigma}} \\ s_{\lambda_{\mu}\lambda_{\beta}} \\ s_{\lambda_{\mu}\lambda_{\gamma}} \\ s_{\lambda_{\sigma}\lambda_{\sigma}} \\ s_{\lambda_{\mu}\lambda_{\beta}} \\ s_{\lambda_{\gamma}\lambda_{\gamma}} \end{pmatrix}, \quad (74) \]
where
\[ \begin{pmatrix} s_{\lambda_{\mu}\lambda_{\mu}} \\ s_{\lambda_{\mu}\lambda_{\sigma}} \\ s_{\lambda_{\mu}\lambda_{\beta}} \\ s_{\lambda_{\mu}\lambda_{\gamma}} \\ s_{\lambda_{\sigma}\lambda_{\sigma}} \\ s_{\lambda_{\mu}\lambda_{\beta}} \\ s_{\lambda_{\gamma}\lambda_{\gamma}} \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{T} H_{i,t}^{2*} + \frac{1}{2} \sum_{t=1}^{T} \sum_{s \neq t} H_{i,t}^{1*} H_{i,s}^{1*} \\ 3 \sum_{t=1}^{T} H_{i,t}^{3*} + \sum_{t=1}^{T} \sum_{s \neq t} H_{i,t}^{1*} H_{i,s}^{2*} \\ 3 \sum_{t=1}^{T} H_{i,t}^{2*} + \frac{1}{2} \sum_{t=1}^{T} \sum_{s \neq t} H_{i,t}^{2*} H_{i,s}^{2*} \\ 2 \sum_{t=1}^{T} H_{i,t}^{1*} x_{it} + \sum_{t=1}^{T} \sum_{s \neq t} H_{i,t}^{1*} H_{i,s}^{1*} x_{it} \\ 3 \sum_{t=1}^{T} H_{i,t}^{1*} x_{it} + 2 \sum_{t=1}^{T} \sum_{s \neq t} H_{i,t}^{1*} H_{i,s}^{1*} x_{it} \end{pmatrix}, \quad \text{and} \]
\[ s_{\lambda_{\beta\beta}} = \begin{pmatrix} \sum_{t=1}^{T} H_{i,t}^{2*} x_{it,1}^{2} + \frac{1}{2} \sum_{t=1}^{T} \sum_{s \neq t} H_{i,t}^{1*} x_{it,1} H_{i,s}^{1*} x_{is,1} \\ \vdots \\ \sum_{t=1}^{T} H_{i,t}^{2*} x_{it,q}^{2} + \frac{1}{2} \sum_{t=1}^{T} \sum_{s \neq t} H_{i,t}^{1*} x_{it,q} H_{i,s}^{1*} x_{is,q} \\ 2 \sum_{t=1}^{T} H_{i,t}^{1*} x_{it,1} x_{it,2} + \sum_{t=1}^{T} \sum_{s \neq t} H_{i,t}^{1*} x_{it,1} H_{i,s}^{1*} x_{is,2} \\ \vdots \\ 2 \sum_{t=1}^{T} H_{i,t}^{1*} x_{it,1} x_{it,q} + \sum_{t=1}^{T} \sum_{s \neq t} H_{i,t}^{1*} x_{it,1} H_{i,s}^{1*} x_{is,q} \\ 2 \sum_{t=1}^{T} H_{i,t}^{1*} x_{it,2} x_{it,3} + \sum_{t=1}^{T} \sum_{s \neq t} H_{i,t}^{1*} x_{it,2} H_{i,s}^{1*} x_{is,3} \\ \vdots \\ 2 \sum_{t=1}^{T} H_{i,t}^{1*} x_{it,q-1} x_{it,q} + \sum_{t=1}^{T} \sum_{s \neq t} H_{i,t}^{1*} x_{it,q-1} H_{i,s}^{1*} x_{is,q} \end{pmatrix}. \quad (75) \]
When $T = 1$, the score functions are as follows:

$$s_\eta = \begin{pmatrix} s_\mu \\ s_\sigma \\ s_\beta \\ s_\gamma \end{pmatrix} = \begin{pmatrix} H_1^s \\ H_2^s \\ H_1^s x_i \\ H_1^s z_i \end{pmatrix}, \quad \begin{pmatrix} s_{\lambda,\mu} \\ s_{\lambda,\sigma} \\ s_{\lambda,\beta} \\ s_{\lambda,\gamma} \end{pmatrix} = \begin{pmatrix} H_2^s \\ 3H_3^s \\ 2H_2^s x_i \\ 3H_3^s x_i \end{pmatrix}, \quad \text{and } s_{\lambda,\beta} = \begin{pmatrix} H_2^s x_i^2 \\ \vdots \\ 2H_2^s x_i x_i,1 x_i,2 \\ 2H_2^s x_i,1 x_i,1 x_i,2 x_i,3 \\ \vdots \\ 2H_2^s x_i,1 x_i,1 x_i,q \\ 2H_2^s x_i,2 x_i,3 \\ \vdots \end{pmatrix} \quad (76)$$

Note that $s_\sigma$ and $s_{\lambda,\mu}$ are perfectly 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. $\lambda$ at $\psi_H^{*}$ is identically zero similarly to the test of the homogeneity case. The values of the score function $s_{\theta_i}$ contain the first-order derivatives w.r.t. the parameters $\gamma$ and $\nu$ at $\psi_H^{*}$:

$$\nabla_{\psi} l^h(w; \psi_H^{*}, \tau) = \frac{f(w; \gamma^*, \theta_0^{M_0}*) - f(w; \gamma^*, \theta_0^{M_0}*)}{\sum_{i=1}^{M_0} \alpha_0^* f(w; \gamma^*, \theta_0^{M_0}*)};$$

$$\nabla_{\gamma} l^h(w; \psi_H^{*}, \tau) = \frac{\sum_{j=1}^{M_0} \alpha_0^* \nabla_{\gamma} f(w; \gamma^*, \theta_0^{M_0}*)}{\sum_{j=1}^{M_0} \alpha_0^* f(w; \gamma^*, \theta_0^{M_0}*)};$$

$$\nabla_{\nu} l^h(w; \psi_H^{*}, \tau) = \frac{\nabla_{\nu} f(w; \gamma^*, \theta_0^{M_0}*)}{\sum_{j=1}^{M_0} \alpha_0^* f(w; \gamma^*, \theta_0^{M_0}*)}.$$  \quad (77)

Define $H_{j,i,t}^{h,*}$ as an abridged expression for $\frac{1}{\sigma_0^*} \frac{1}{M_0} H^h_{j,i,t} \frac{u_{i,t}^j - \mu_{i,t}^j - x_{i,t}^j \beta_{M_0}^j}{\sigma_0^j}$. Define the weight $w_{i,t}^{j,*}$ as

$$w_{i,t}^{j,*} = \frac{\alpha_0^j f(\{W_{i,t}\}_{t=1}^{T}; \gamma^*, \theta_0^{M_0}*)}{f_{M_0}(\{W_{i,t}\}_{t=1}^{T}; \theta_0^{M_0})}, j = 1, \ldots, M_0,$$

where $f_{M_0}(\{W_{i,t}\}_{t=1}^{T}; \theta_0^{M_0})$ is defined by equation (17).
As shown in section B.3, the score functions are

\[ s_\alpha(w_i) = \left( \frac{f(w_i|\theta_1^*) - f(w_i|\theta_0^*)}{\sum_i \alpha_t f(w_i|\theta_0^*)} \right) \]

\[ s_\beta(w_i) = \left( \begin{array}{c} w_i^1 \sum_{t=1}^T H_{1,i,t}^1 \\ \vdots \\ w_i^{M_0} \sum_{t=1}^T H_{M_0,i,t}^1 \end{array} \right) \]

\[ s_\gamma(w_i) = \left( \begin{array}{c} w_i^1 \sum_{t=1}^T H_{1,i,t}^1 x_{it} \\ \vdots \\ w_i^{M_0} \sum_{t=1}^T H_{M_0,i,t}^1 x_{it} \end{array} \right) \]

The score function for \( s_{\lambda\lambda}^h \) is obtained analogously to \( s_{\lambda\lambda} \) by replacing \( H_{1,i,t}^{h,s} \) with \( H_{b,i,t}^{h,s} \) for \( b = 1, \ldots, 4 \) so that

\[ s_{\lambda\lambda}^h(w_i) = w_i^{h,s} \]

\[ s_{\lambda\lambda}^h(w_i) = \left( \begin{array}{c} \sum_{t=1}^T H_{h,i,t}^2 + \frac{1}{2} \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t}^1 H_{h,i,s}^1 \\ 3 \sum_{t=1}^T H_{h,i,t}^2 + \frac{1}{2} \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t}^1 H_{h,i,s}^1 \\ 2 \sum_{t=1}^T H_{h,i,t} x_{it} + \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t} x_{it} H_{h,i,s}^1 \\ 3 \sum_{t=1}^T H_{h,i,t} x_{it} + 2 \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t} x_{it} H_{h,i,s}^1 \end{array} \right) \]

\[ s_{\lambda\lambda}^h(w_i) = \left( \begin{array}{c} \sum_{t=1}^T H_{h,i,t}^2 x_{it,1}^2 + \frac{1}{2} \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t} x_{it,1} H_{h,i,s} x_{is,1} \\ \vdots \\ 2 \sum_{t=1}^T H_{h,i,t} x_{it,1} x_{it,2} + \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t} x_{it,1} x_{it,q} H_{h,i,s} x_{is,q} \\ 2 \sum_{t=1}^T H_{h,i,t} x_{it,2} x_{it,1} + \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t} x_{it,2} x_{it,q} H_{h,i,s} x_{is,q} \\ \vdots \\ 2 \sum_{t=1}^T H_{h,i,t} x_{it,q-1} x_{it,q} + \sum_{t=1}^T \sum_{s \neq t} H_{h,i,t} x_{it,q-1} x_{is,q} H_{h,i,s} x_{is,q} \end{array} \right) \]
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\} |
| $\alpha_n$ | (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\} |
| $\alpha$ | \{(0.5, 0.5): (0.2, 0.8)\} |
| $\mu$ | \{(-1, 1), (-0.5, 0.5), (-0.8, 0.8)\} |
| $\sigma$ | \{(1, 1), (1.5, 0.75), (0.8, 1.2)\} |
| $\alpha_n$ | (0.01, 0.05, 0.1, 0.2, 0.3, 0.4) |

| $M_0 = 3$ |
|---|
| $N$ | \{100, 500\} |
| $T$ | \{2, 10\} |
| $\alpha$ | \{(1/3, 1/3, 1/3): (0.25, 0.5, 0.25)\} |
| $\mu$ | \{(-4, 0, 4), (-4, 0, 5), (-5, 0, 5), (-4, 0, 6), (-5, 0, 6), (-6, 0, 6)\} |
| $\sigma$ | \{(1, 1, 1): (0.75, 1.5, 0.75)\} |
| $\alpha_n$ | (0.01, 0.05, 0.1, 0.2, 0.3, 0.4) |

| $M_0 = 4$ |
|---|
| $N$ | \{100, 500\} |
| $T$ | \{2, 10\} |
| $\alpha$ | \{(0.25, 0.25, 0.25, 0.25)\} |
| $\mu$ | \{(-4, -1, 1, 4), (-5, -1, 1, 5), (-6, -2, 2, 6), (-6, -1, 2, 5), (-5, 0, 2, 4), (-6, 0, 2, 4)\} |
| $\sigma$ | \{(1, 1, 1, 1): (1, 0.75, 0.5, 0.25)\} |
| $\alpha_n$ | (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 | (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)     |
| $\log \left( \frac{\hat{s}_{a_n}}{\hat{s}_{a_n}} \right)$ | -0.016      | -0.101***   | -0.111***   | -0.128***   |
|                   | (0.019)     | (0.009)     | (0.007)     | (0.030)     |
| $\log \left( \frac{\omega(\theta_{M0};M_0)}{\omega(\theta_{M_n};M_0)} \right)$ | -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 .