Online Appendix to "Estimating the Number of Components in Panel Data Finite Mixture Regression Models with an Application to Production Function Heterogeneity"

Yu Hao Hiroyuki Kasahara

July 2, 2026

This Online Appendix is for online publication only and accompanies the main text. Cross-references of the form § n, Table n, and equation (n) refer to the main text; appendix material is numbered A–B below.

A Appendix

A.1 Likelihood ratio test for testing $H_{02} : \alpha(1-\alpha) = 0$

We derive the asymptotic distribution of the LRT statistic for testing the hypothesis $H_{02} : \alpha(1-\alpha) = 0$ for testing $H_0 : M_0 = 1$ against $H_1 : M_0 = 2$. Our analysis focuses on the null hypothesis $\alpha = 0$, noting that the case $\alpha = 1$ follows by symmetry.

Consider the following restricted parameter space in which $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are bounded away from each other while α can take the value of 0 and 1:

$$\widetilde{\Theta}_{\boldsymbol{\vartheta}_2}(\boldsymbol{c}) := \{\boldsymbol{\vartheta}_2 \in \Theta_{\boldsymbol{\vartheta}_2} : ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|| \geq c_1,\ \min_j \sigma_j^2 \geq c_2 \sum_{k=1}^{2} \alpha_k \sigma_k^2,\ \sum_{k=1}^{2} \alpha_k \sigma_k^2 \geq c_3\}.$$

Define the MLE $\tilde{\boldsymbol{\vartheta}}_2$ in this restricted parameter space by

$$\ell_n^2(\tilde{\boldsymbol{\vartheta}}_2) = \max_{\boldsymbol{\vartheta}_2 \in \widetilde{\Theta}_{\boldsymbol{\vartheta}_2}(c)} \ell_n^2(\boldsymbol{\vartheta}_2). \tag{1}$$

Following the proof of Proposition (3), $\tilde{\boldsymbol{\vartheta}}_2$ can be shown to be consistent under Assumptions 1 and 3.

Throughout this subsection, $h(\boldsymbol{w};\boldsymbol{\theta})$ denotes the component-specific density, corresponding to $f(\boldsymbol{w};\boldsymbol{\theta}_j)$ in the notation of the main text. To simplify the asymptotic representation and regularity conditions, we use the parameter $\boldsymbol{\lambda} := \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$ and reparameterize $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ to $(\boldsymbol{\lambda}, \boldsymbol{\theta}_2)$ so that the model parameter is $\boldsymbol{\Psi} = (\boldsymbol{\theta}_2, \boldsymbol{\lambda}, \alpha) \in \Theta_{\boldsymbol{\Psi}}$. Under $H_0 : M_0 = 1$, the two-components model replicates the true one-component model when $\boldsymbol{\theta}_2 = \boldsymbol{\theta}^*$ and $\alpha = 0$. Given that $\boldsymbol{\lambda}$ is unidentified when $\alpha = 0$, we follow Andrews (2001) by deriving the limit of the LRT statistic for each $\boldsymbol{\lambda} := \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$ in $\Theta_{\boldsymbol{\lambda}}(c_1) := \{\boldsymbol{\lambda} \in \Theta_{\boldsymbol{\lambda}} : ||\boldsymbol{\lambda}|| \geq c_1\}$ for $c_1 > 0$.

Define the reparameterized log-density as $\log g_2(\boldsymbol{w};\boldsymbol{\theta}_2,\boldsymbol{\lambda},\alpha) = \log(\alpha h(\boldsymbol{w};\boldsymbol{\theta}_2+\boldsymbol{\lambda}) + (1-\alpha)h(\boldsymbol{w};\boldsymbol{\theta}_2))$. Collect the partial derivative of $\log g_2(\boldsymbol{w};\boldsymbol{\theta}_2,\boldsymbol{\lambda},\alpha)$ with respect to $\boldsymbol{\theta}_2$ and its right partial derivative with respect to α evaluated at $(\boldsymbol{\theta}_2,\boldsymbol{\lambda},\alpha) = (\boldsymbol{\theta}^*,\boldsymbol{\lambda},0)$ as

$$\boldsymbol{s}(\boldsymbol{w};\boldsymbol{\lambda}) := \begin{pmatrix} \boldsymbol{s}_{\boldsymbol{\theta}_2}(\boldsymbol{w}) \\ s_\alpha(\boldsymbol{w};\boldsymbol{\lambda}) \end{pmatrix} := \begin{pmatrix} \nabla_{\boldsymbol{\theta}_2} \log g_2(\boldsymbol{w};\boldsymbol{\theta}^*,\boldsymbol{\lambda},0) \\ \nabla_\alpha \log g_2(\boldsymbol{w};\boldsymbol{\theta}^*,\boldsymbol{\lambda},0) \end{pmatrix} = \begin{pmatrix} \frac{\nabla_{\boldsymbol{\theta}} h(\boldsymbol{w};\boldsymbol{\theta}^*)}{h(\boldsymbol{w};\boldsymbol{\theta}^*)} \\ \frac{h(\boldsymbol{w};\boldsymbol{\theta}^*+\boldsymbol{\lambda})-h(\boldsymbol{w};\boldsymbol{\theta}^*)}{h^*(\boldsymbol{w};\boldsymbol{\theta}^*)} \end{pmatrix}. \tag{2}$$

Define $\boldsymbol{\mathcal{I}}(\boldsymbol{\lambda}) := E[\boldsymbol{s}(\boldsymbol{W};\boldsymbol{\lambda})\boldsymbol{s}(\boldsymbol{W};\boldsymbol{\lambda})^\top]$. Analogously to (22), define

$$\begin{aligned}
&\boldsymbol{\mathcal{I}}(\boldsymbol{\lambda}) = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}_2} & \boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}_2\alpha}(\boldsymbol{\lambda}) \\ \boldsymbol{\mathcal{I}}_{\alpha\boldsymbol{\theta}_2}(\boldsymbol{\lambda}) & \boldsymbol{\mathcal{I}}_\alpha(\boldsymbol{\lambda}) \end{pmatrix}, \quad \boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}_2} = \mathbb{E}[\boldsymbol{s}_{\boldsymbol{\theta}_2}(\boldsymbol{W})\boldsymbol{s}_{\boldsymbol{\theta}_2}(\boldsymbol{W})^\top], \quad \boldsymbol{\mathcal{I}}_{\alpha\boldsymbol{\theta}_2}(\boldsymbol{\lambda}) = \mathbb{E}[s_\alpha(\boldsymbol{W};\boldsymbol{\lambda})\boldsymbol{s}_{\boldsymbol{\theta}_2}(\boldsymbol{W})^\top], \\
&\boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}_2\alpha}(\boldsymbol{\lambda}) = \boldsymbol{\mathcal{I}}_{\alpha\boldsymbol{\theta}_2}(\boldsymbol{\lambda})^\top, \quad \boldsymbol{\mathcal{I}}_\alpha(\boldsymbol{\lambda}) = \mathbb{E}[s_\alpha(\boldsymbol{W})^2], \text{ and } \boldsymbol{\mathcal{I}}_{\alpha,\boldsymbol{\theta}_2}(\boldsymbol{\lambda}) = \boldsymbol{\mathcal{I}}_\alpha(\boldsymbol{\lambda}) - \boldsymbol{\mathcal{I}}_{\alpha\boldsymbol{\theta}_2}(\boldsymbol{\lambda})\boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}_2}^{-1}\boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}_2\alpha}(\boldsymbol{\lambda}).
\end{aligned} \tag{3}$$

Let $\{\boldsymbol{S}(\boldsymbol{\lambda}) = (\boldsymbol{S}_{\boldsymbol{\theta}_2}, S_\alpha(\boldsymbol{\lambda})) : \boldsymbol{\lambda} \in \widetilde{\Theta}_{\boldsymbol{\lambda}}(c_1)\}$ be a mean zero vector-valued Gaussian process such that $\mathbb{E}[\boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{S}(\boldsymbol{\lambda})^\top] = \boldsymbol{\mathcal{I}}(\boldsymbol{\lambda})$, where $\boldsymbol{S}_{\boldsymbol{\theta}_2}$ is independent of $\boldsymbol{\lambda}$, $S_\alpha(\boldsymbol{\lambda})$ is 1×1, and $\boldsymbol{\mathcal{I}}(\boldsymbol{\lambda})$ is defined in (3). Let $S_{\alpha,\boldsymbol{\theta}_2}(\boldsymbol{\lambda}) := S_\alpha(\boldsymbol{\lambda}) - \boldsymbol{\mathcal{I}}_{\alpha\boldsymbol{\theta}_2}(\boldsymbol{\lambda})\boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}_2}^{-1}\boldsymbol{S}_{\boldsymbol{\theta}_2}$.

Define the LRT statistics for testing $H_{02} : \alpha(1-\alpha) = 0$ by

$$\widetilde{LR}_n^2(\boldsymbol{c}) := 2\left\{\ell_n^2(\widetilde{\boldsymbol{\vartheta}}_2) - \ell_n^1(\hat{\boldsymbol{\theta}}_0)\right\} = 2\left\{\max_{(\boldsymbol{\theta}_2,\boldsymbol{\lambda},\alpha)\in\Theta_{\boldsymbol{\theta}}\times\widetilde{\Theta}_{\boldsymbol{\lambda}}(c_1)\times[0,1/2]} \ell_n^2(\boldsymbol{\theta}_2+\boldsymbol{\lambda},\boldsymbol{\theta}_2,\alpha) - \ell_n^1(\hat{\boldsymbol{\theta}}_0)\right\}. \tag{4}$$

Assumption A.1. *(a) $\boldsymbol{\theta}^*$ is in the interior of $\Theta_{\boldsymbol{\theta}}$. (b) $h(\boldsymbol{w};\boldsymbol{\theta})$ is twice continuously differentiable on $\Theta_{\boldsymbol{\theta}}$. (c) $\boldsymbol{\mathcal{I}}(\boldsymbol{\lambda})$ defined in (3) is finite and positive definite uniformly in $\Theta_{\boldsymbol{\lambda}}(c_1)$. (d) Assumption 3 holds when we replace $\bar{\Theta}_{\boldsymbol{\vartheta}_2}(\boldsymbol{c})$ with $\tilde{\Theta}_{\boldsymbol{\vartheta}_2}(\boldsymbol{c})$.*

Proposition A.1. *Suppose Assumptions 1 and A.1 hold. Then, (a) $\inf_{\boldsymbol{\vartheta}_2\in\Theta^*_{\boldsymbol{\vartheta}_2}} ||\widetilde{\boldsymbol{\vartheta}}_2 - \boldsymbol{\vartheta}_2|| \to 0$ almost surely, and (b) $\widetilde{LR}_n^2(\boldsymbol{c}) \xrightarrow{d} \sup_{\boldsymbol{\lambda}\in\Theta_{\boldsymbol{\lambda}}(c_1)} \left(\max\{0, \boldsymbol{\mathcal{I}}_{\alpha,\boldsymbol{\theta}_2}(\boldsymbol{\lambda})^{-1/2} S_{\alpha,\boldsymbol{\theta}_2}(\boldsymbol{\lambda})\}\right)^2$.*

Assumption A.1(b) can be verified for a class of mixture models we consider in this paper.

A necessary condition for Assumption A.1(c) is $\sup_{\boldsymbol{\lambda}\in\widetilde{\Theta}_{\boldsymbol{\lambda}}(c_1)} \mathbb{E}[\nabla_\alpha \log g_2(\boldsymbol{w};\boldsymbol{\theta}^*,\boldsymbol{\lambda},0)^2] < \infty$, which is violated for the finite mixture normal regression models when $\sigma_j^2 > 2\sigma^{*2}$.

One approach is to impose a restriction that $\sigma_1^2 \le 2\sigma^{*2}$. A difficulty arises in imposing this constraint when estimating a two-component model, given that σ^{*2} is unknown. A possible solution is to use the estimated variance from the one-component model, setting:

$$\sigma_j^2 \in \left[0,\, 2\hat{\sigma}_0^2 - c\right] \text{ for } j = 1,2,$$

where $\hat{\sigma}_0^2$ is the estimated variance from the one-component model and c is a small positive constant. This ensures the asymptotic validity of the condition $\sigma_j^2 < 2\sigma^{*2}$. However, in finite samples, there remains a positive probability that values of σ_j^2 within this random interval may violate the constraint.

A.2 Asymptotic analysis of the penalised maximum likelihood estimator in Section 5.3

In this appendix, we extend the previous asymptotic analysis for testing $H_0 : M = M_0$ against $H_1 : M = M_0 + 1$ to the more general setting of testing $H_0 : M = M_0$ against $H_1 : M = M_1$, where $M_1 > M_0 + 1$ to derive conditions for $\ell_n^M(\hat{\boldsymbol{\vartheta}}_M) - \ell_n^{M_0}(\boldsymbol{\vartheta}^*_{M_0}) = O_p(1)$.

There are two primary challenges in extending our earlier analysis. First, while there are exactly M_0 ways for the $(M_0 + 1)$-component model to replicate the M_0-component model, the number of ways for an M_1-component model to replicate an M_0-component model can be substantially larger when $M_1 > M_0 + 1$. Second, the degree of singularity in the Fisher Information Matrix may become more severe, necessitating the higher-order expansions beyond the second-order for an asymptotic analysis.

To deal with the first challenge, given any integer M that is greater than M_0, we extend the definition of the restricted parameter space (16) by partitioning the M component-specific parameters into M_0 ordered, non-empty subsets. Explicitly, we consider a partition $\mathcal{T} = (T_1, T_2, \dots, T_{M_0})$ of the index set $\{1, 2, \dots, M\}$, subject to the following conditions: (i) $T_h \cap T_{h'} = \emptyset$ for $h \neq h'$, and $|T_h| \geq 1$, (ii) $\bigcup_{h=1}^{M_0} T_h = \{1, 2, \dots, M\}$, (iii) If $j \in T_h$ and $k \in T_{h'}$ with $h < h'$, then $j < k$. The total number of such ordered, non-empty, consecutive partitions is $J_M := \binom{M-1}{M_0-1}$. Let $\{\mathcal{T}^j\}_{j=1}^{J_M}$ denote the set of all these partitions, where each $\mathcal{T}^j = (T_1^j, T_2^j, \dots, T_{M_0}^j)$.

Additionally, consider a partition $\{\Theta^*_{\boldsymbol{\theta},h}\}_{h=1}^{M_0}$ of the parameter space $\Theta_{\boldsymbol{\theta}}$, defined by (15), such that each $\Theta^*_{\boldsymbol{\theta},h}$ forms a neighborhood containing $\boldsymbol{\theta}^*_h$ but excluding $\boldsymbol{\theta}^*_j$ for all $j \neq h$. Then, for each partition $\mathcal{T}^j$, $j = 1, 2, \dots, J_M$, the *restricted parameter set* (not to be confused with the reparameterized parameter vector $\boldsymbol{\Psi}_M^j$ defined below) is:

$$\boldsymbol{\Psi}^*_{\mathcal{T}^j} = \left\{ \boldsymbol{\vartheta}_M \in \bar{\Theta}_{\boldsymbol{\vartheta}_M}(\boldsymbol{c}) : \sum_{j=1}^{M} \alpha_j = 1, \quad \boldsymbol{\theta}_j \in \Theta^*_{\boldsymbol{\theta},h} \text{ for all } j \in T_h^j, \quad h = 1, \dots, M_0 \right\}.$$

Each $\boldsymbol{\Psi}^*_{\mathcal{T}^j}$ uniquely represents a distinct configuration in which the components from an M-component model are distributed across M_0 components; the union of $\boldsymbol{\Psi}^*_{\mathcal{T}^j}$ over $j = 1, 2, ..., J_M$ covers the parameter space $\bar{\Theta}_{\boldsymbol{\vartheta}_M}(\boldsymbol{c})$.

Define the *local* MLE that maximises the log-likelihood function of the M-component model under the constraint that $\boldsymbol{\vartheta}_M \in \boldsymbol{\Psi}^*_{\mathcal{T}^j}$ by

$$\ell_n^M(\widehat{\boldsymbol{\vartheta}}_M^j) := \arg \sup_{\boldsymbol{\vartheta}_M \in \boldsymbol{\Psi}^*_{\mathcal{T}^j}} \ell_n^M(\boldsymbol{\vartheta}_M).$$

Because $\cup_{j=1}^{J_M} \boldsymbol{\Psi}^*_{\mathcal{T}^j} = \bar{\Theta}_{\boldsymbol{\vartheta}_M}(\boldsymbol{c})$,

$$\ell_n^M(\hat{\boldsymbol{\vartheta}}_M) - \ell_n^{M_0}(\boldsymbol{\vartheta}^*_{M_0}) = \max_{j=1,2,\dots,J_M} \{\ell_n^M(\hat{\boldsymbol{\vartheta}}_M^j) - \ell_n^{M_0}(\boldsymbol{\vartheta}^*_{M_0})\}. \tag{5}$$

In view of (5), $\ell_n^M(\hat{\boldsymbol{\vartheta}}_M) - \ell_n^{M_0}(\boldsymbol{\vartheta}^*_{M_0}) = O_p(1)$ follows if we are able to show that $\ell_n^M(\hat{\boldsymbol{\vartheta}}_M^j) - \ell_n^{M_0}(\boldsymbol{\vartheta}^*_{M_0}) = O_p(1)$ for any $j = 1, ..., J_M$.

To deal with the singularity of the Fisher Information Matrix for the M-component density function $g_M(\boldsymbol{w};\boldsymbol{\vartheta}_M)$, we consider the following one-to-one reparameterization of $\boldsymbol{\vartheta}_M$. Specifically, for $j=1,2,\ldots,J_M$, we reparameterize $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_{M-1})$ by $\beta_h^j=\sum_{k\in T_h^j}\alpha_k$, $\boldsymbol{\tau}_h^j=(\tau_{h,1}^j,...,\tau_{h,|T_h^j|-1}^j)^\top$, and $\tau_{h,k}^j=\alpha_k/\beta_h^j$ for $k\in T_h^j$ and $h=1,\ldots,M_0$. Given the index set $\mathcal{T}^j=(T_1^j,\ldots,T_{M_0}^j)$ with $T_h^j=(k_h,k_h+1,\ldots,k_h+|T_h^j|-1)$, and applying an additional reparameterization of $\{\boldsymbol{\theta}_k\}_{k=1}^M$ defined by (7), the reparameterized density function of the M-component model is expressed as:

$$g_M(\boldsymbol{w};\boldsymbol{\Psi}_M^j,\boldsymbol{\tau}^j)=\sum_{h=1}^{M_0}\beta_h^j\left(\sum_{k\in T_h^j}\tau_k^j h(\boldsymbol{w};\tilde{\boldsymbol{\theta}}_k^j)\right),\tag{6}$$

where

$$\begin{pmatrix}\tilde{\boldsymbol{\theta}}_{k_h}^j\\ \tilde{\boldsymbol{\theta}}_{k_h+1}^j\\ \vdots\\ \tilde{\boldsymbol{\theta}}_{k_h+|T_h^j|-1}^j\end{pmatrix}=\begin{pmatrix}\boldsymbol{\nu}_h^j-\sum_{\ell=2}^{|T_h^j|}\tau_{h,\ell}^j\boldsymbol{\lambda}_{h,\ell}^j\\ \boldsymbol{\nu}_h^j+\tau_{h,1}^j\boldsymbol{\lambda}_{h,2}^j\\ \vdots\\ \boldsymbol{\nu}_h^j+\tau_{h,1}^j\boldsymbol{\lambda}_{h,|T_h^j|-1}^j\end{pmatrix},\quad h=1,\ldots,M_0.\tag{7}$$

Collecting the reparameterized parameters, we set $\boldsymbol{\Psi}_M^j:=(\boldsymbol{\eta}^j,\boldsymbol{\lambda}^j)^\top$ (in subscripts and function arguments, we write $\boldsymbol{\psi}_M^j$ for typographic clarity), $\boldsymbol{\tau}^j:=(\boldsymbol{\tau}_1^j,\ldots,\boldsymbol{\tau}_{M_0}^j)^\top$, $\boldsymbol{\eta}^j:=(\boldsymbol{\beta}^j,\boldsymbol{\nu}^j)$, $\boldsymbol{\beta}^j:=(\beta_1^j,\ldots,\beta_{M_0-1}^j)^\top$, $\boldsymbol{\nu}^j:=(\boldsymbol{\nu}_1^j,\ldots,\boldsymbol{\nu}_{M_0}^j)^\top$, and $\boldsymbol{\lambda}^j:=(\boldsymbol{\lambda}_1^j,\ldots,\boldsymbol{\lambda}_{M_0}^j)^\top$, where $\boldsymbol{\nu}_h^j:=(\boldsymbol{\nu}_{h,2}^j,\ldots,\boldsymbol{\nu}_{h,|T_h^j|}^j)^\top$, $\boldsymbol{\lambda}_h^j:=(\boldsymbol{\lambda}_{h,2}^j,\ldots,\boldsymbol{\lambda}_{h,|T_h^j|}^j)^\top$, and $\boldsymbol{\tau}_h^j:=(\tau_{h,1}^j,\ldots,\tau_{h,|T_h^j|-1}^j)$ for $h=1,\ldots,M_0$. When the data is generated from the M_0-components model density $g_{M_0}(\boldsymbol{w};\boldsymbol{\vartheta}_{M_0}^*):=\sum_{h=1}^{M_0}\alpha_h^*h(\boldsymbol{w};\boldsymbol{\theta}_h^*)$, we have $\tilde{\boldsymbol{\theta}}_k^{j*}=\boldsymbol{\theta}_h^*$ for all $k\in T_h^j$, hence $\boldsymbol{\nu}_h^{j*}=\boldsymbol{\theta}_h^*$ and $\boldsymbol{\lambda}_h^{j*}=\mathbf{0}$ while $\boldsymbol{\tau}^j$ is not identified. In this scenario, since $\beta_h^{j*}=\alpha_h^*$ and $\sum_{k\in T_h^j}\tau_k^j h(\boldsymbol{w};\tilde{\boldsymbol{\theta}}_k^{j*})=h(\boldsymbol{w};\boldsymbol{\theta}_h^*)$ in (6), the density function $g_M(\boldsymbol{w};\boldsymbol{\Psi}_M^{j*},\boldsymbol{\tau}^j)$ coincides with the true M_0-component model density $g_{M_0}(\boldsymbol{w};\boldsymbol{\vartheta}_{M_0}^*)$.

With this reparameterization, the first-order derivatives satisfy:

$$\nabla_{\boldsymbol{\nu}_h^j}\log g_M(\boldsymbol{w};\boldsymbol{\psi}_M^{j*},\boldsymbol{\tau}^j)=\frac{\nabla_{\boldsymbol{\theta}_h}h(\boldsymbol{w};\boldsymbol{\theta}_h^*)}{g_{M_0}(\boldsymbol{w};\boldsymbol{\vartheta}_{M_0}^*)},\quad \nabla_{\boldsymbol{\lambda}_h^j}\log g_M(\boldsymbol{w};\boldsymbol{\psi}_M^{j*},\boldsymbol{\tau}^j)=\mathbf{0},\quad h=1,\ldots,M_0.\tag{8}$$

Because $\nabla_{\boldsymbol{\lambda}_h^j}\log g_M(\boldsymbol{w};\boldsymbol{\psi}_M^{j*},\boldsymbol{\tau}^j)=\mathbf{0}$, the Fisher information matrix is singular, and the standard quadratic approximation fails. As analyzed in Section 5.1, when $M=M_0+1$, the unique elements of $\nabla_{\boldsymbol{\lambda}_h^j\boldsymbol{\lambda}_h^j}\log g_M(\boldsymbol{w};\boldsymbol{\psi}_M^{j*},\boldsymbol{\tau}^j)$ plays the role of score function in (2) to identify $\boldsymbol{\lambda}_h^j$. When M is much larger than M_0, we need higher order derivatives beyond the second order derivatives to identify $\boldsymbol{\lambda}_h^j$.

Denote the density ratio by

$$l_{\boldsymbol{\Psi}_M^j\boldsymbol{\tau}^j,i}:=\frac{g_M(\boldsymbol{W}_i;\boldsymbol{\Psi}_M^j,\boldsymbol{\tau}^j)}{g_{M_0}(\boldsymbol{W}_i;\boldsymbol{\vartheta}_{M_0}^*)}$$

so that $\ell_n(\boldsymbol{\Psi}_M^j,\boldsymbol{\tau}^j)-\ell_n(\boldsymbol{\Psi}_M^{j*},\boldsymbol{\tau}^j)=\sum_{i=1}^n\log l_{\boldsymbol{\Psi}_M^j\boldsymbol{\tau}^j,i}$.

Let $\boldsymbol{\lambda}^j_{h,\ell_1} \otimes \boldsymbol{\lambda}^j_{h,\ell_2} \otimes \cdots \otimes \boldsymbol{\lambda}^j_{h,\ell_p}$ denote the tensor containing all interactions among the elements of the vectors $\boldsymbol{\lambda}^j_{h,\ell_1}, \boldsymbol{\lambda}^j_{h,\ell_2}, \cdots, \boldsymbol{\lambda}^j_{h,\ell_p}$, where $\otimes$ denotes the Kronecker product. We use the higher-order derivatives of $l_{\boldsymbol{\Psi}^j_M \tau^j, i}$ with respect to $\boldsymbol{\lambda}^j_h$ evaluated at $\boldsymbol{\Psi}_M = \boldsymbol{\Psi}^*_M$ to identify $\boldsymbol{\lambda}^j_h$: for $h = 1, 2..., M_0$ and for $p \geq 2$,

$$\nabla_{\boldsymbol{\lambda}^j_{h,\ell_1} \otimes \boldsymbol{\lambda}^j_{h,\ell_2} \otimes \cdots \otimes \boldsymbol{\lambda}^j_{h,\ell_p}} l_{\boldsymbol{\Psi}^j_M \tau^j, i} = \gamma^j_{h,\ell_1\ell_2\cdots\ell_p}(\boldsymbol{\tau}^j_h) \frac{\alpha^*_h \nabla_{\boldsymbol{\theta}^{\otimes p}} h^*_{h,i}}{g^*_{M_0,i}} \quad \text{for } (\ell_1, ..., \ell_p) \in \{2, ..., |T^j_h|\}^p,$$

where $\boldsymbol{\theta}^{\otimes p}$ represents $\boldsymbol{\theta} \otimes \boldsymbol{\theta} \cdots \otimes \boldsymbol{\theta}$ (p times), $h^*_{h,i} := h(W_i; \boldsymbol{\theta}^*_h)$, and $g^*_{M_0,i} := g_{M_0}(\boldsymbol{W}_i; \boldsymbol{\vartheta}^*_{M_0})$. For $p = 2, 3$, we have $\gamma^j_{h,\ell_1\ell_2}(\boldsymbol{\tau}^j_h) = \tau^j_{h,1}\tau^j_{h,\ell_1}(\tau^j_{h,\ell_2} + \delta_{\ell_1\ell_2})$ and $\gamma^j_{h,\ell_1\ell_2\ell_3}(\boldsymbol{\tau}^j_h) = \tau^j_{h,1}\tau^j_{h,\ell_1}(\delta_{\ell_1\ell_2}\delta_{\ell_2\ell_3}(\tau^j_{h,1})^2 - \tau^j_{h,\ell_2}\tau^j_{h,\ell_3})$ with δ_{ij} being Kronecker delta. The elements of $\nabla_{\boldsymbol{\lambda}^j_{h,\ell_1} \otimes \boldsymbol{\lambda}^j_{h,\ell_2} \otimes \cdots \otimes \boldsymbol{\lambda}^j_{h,\ell_p}} l_{\boldsymbol{\Psi}^j_M \tau^j, i}$ are ave-zero random variables.

Let $q := \dim(\boldsymbol{\theta})$. Let $\widetilde{\text{vech}}_p(\nabla_{\boldsymbol{\theta}^{\otimes p}} h^*/g^*)$ extract all *unique* elements of q-dimensional symmetric array $\nabla_{\boldsymbol{\theta}^{\otimes p}} h^*/g^*$, multiply by its frequency, and stacks them into a vector. For example, for $p = 3$ and $q \geq 3$, $\widetilde{\text{vech}}_3\left(\nabla_{\boldsymbol{\theta}\otimes\boldsymbol{\theta}\otimes\boldsymbol{\theta}} h^*/g^*\right)$ with $\boldsymbol{\theta} = (\theta_1, ..., \theta_q)^\top$ will contain (i) q elements of the form $\nabla_{\theta_i^3} h^*/g^*$, (ii) $q(q-1)$ elements of the form $3\nabla_{\theta_i^2\theta_j} h^*/g^*$ for $i \neq j$, and (iii) $\binom{q}{3}$ of the form $6\nabla_{\theta_i\theta_j\theta_k} h^*/g^*$ for $i \neq j \neq k$.

Choose p^j_h so that the dimension of the unique elements of $\frac{\nabla_{\boldsymbol{\theta}\otimes\boldsymbol{\theta}} h^*_{h,i}}{g^*_{M_0,i}}, \frac{\nabla_{\boldsymbol{\theta}\otimes\boldsymbol{\theta}\otimes\boldsymbol{\theta}} h^*_{h,i}}{g^*_{M_0,i}}, ..., \frac{\nabla_{\boldsymbol{\theta}^{\otimes p_M}} h^*_{h,i}}{g^*_{M_0,i}}$ is at least as large as the number of elements in $\boldsymbol{\lambda}^j_h$ but no larger than necessary:

$$\begin{aligned} &\dim\left(\widetilde{\text{vech}}_2\left(\frac{\alpha^*_h \nabla_{\boldsymbol{\theta}\otimes\boldsymbol{\theta}} h^*_{h,i}}{g^*_{M_0,i}}\right)\right) + \cdots + \dim\left(\widetilde{\text{vech}}_{p^j_h}\left(\frac{\alpha^*_h \nabla_{\boldsymbol{\theta}^{\otimes p^j_h}} h^*_{h,i}}{g^*_{M_0,i}}\right)\right) \geq \dim(\boldsymbol{\lambda}^j_h) \\ &> \dim\left(\widetilde{\text{vech}}_2\left(\frac{\alpha^*_h \nabla_{\boldsymbol{\theta}\otimes\boldsymbol{\theta}} h^*_{h,i}}{g^*_{M_0,i}}\right)\right) + \cdots + \dim\left(\widetilde{\text{vech}}_{p^j_h - 1}\left(\frac{\alpha^*_h \nabla_{\boldsymbol{\theta}^{\otimes (p^j_h - 1)}} h^*_{h,i}}{g^*_{M_0,i}}\right)\right). \end{aligned} \tag{9}$$

The value of p^j_h indicates a necessary order of local expansions for identifying $\boldsymbol{\lambda}^j_h$. For example, consider the case of testing $H_0 : M_0 = 1$ against $H_1 : M_0 = M_1$ for any $M_1 > 1$. Then, $g^*_{M_0,i} = h^*_i$ and $J_{M_1} = 1$. In this case, for $p = 2$, because $\dim\left(\widetilde{\text{vech}}_2\left(\frac{\nabla_{\boldsymbol{\theta}\otimes\boldsymbol{\theta}} h^*_i}{h^*_i}\right)\right) = \frac{q(q+1)}{2}$ and $\dim(\boldsymbol{\lambda}) = q(M_1 - 1)$, (9) is satisfied if $\frac{(q+1)}{2} \geq M_1 - 1$. For $p = 3$ and $q \geq 3$, the condition is $\frac{q(q+1)}{2} + \left(q + q(q-1) + \frac{q(q-1)(q-2)}{6}\right) \geq q(M_1 - 1) > \frac{q(q+1)}{2}$, which simplifies to $\frac{(q+5)(q+1)}{6} + 1 \geq M_1 > \frac{q+3}{2}$. For $p = 4$ and $q \geq 4$, the condition is given by $\frac{(q+5)(q+1)}{6} + \frac{q^3 - 5q^2 + 10q - 4}{2} + 1 \geq M_1 > \frac{(q+5)(q+1)}{6} + 1$, where $\boldsymbol{\lambda}$ is identified for up to $M_1 = 18$ if $q = 4$.

For $h = 1, ..., M_0$, let

$$\boldsymbol{v}_{\boldsymbol{\lambda}^j_h \boldsymbol{\tau}^j_h} := \begin{pmatrix} \text{vech}_2\left(\sum_{\ell_1=2}^{|T^j_h|}\sum_{\ell_2=2}^{|T^j_h|} \gamma^j_{h,\ell_1\ell_2}(\boldsymbol{\tau}^j_h)\boldsymbol{\lambda}^j_{h,\ell_1} \otimes \boldsymbol{\lambda}^j_{h,\ell_2}\right) \\ \vdots \\ \text{vech}_{p^j_h}\left(\sum_{\ell_1=2}^{|T^j_h|}\cdots\sum_{\ell_{p_M}=2}^{|T^j_h|} \gamma^j_{h,\ell_1\ell_2\cdots\ell_{p_M}}(\boldsymbol{\tau}^j_h)\boldsymbol{\lambda}^j_{h,\ell_1} \otimes \boldsymbol{\lambda}^j_{h,\ell_2} \otimes \cdots \otimes \boldsymbol{\lambda}^j_{h,\ell_{p^j_h}}\right) \end{pmatrix},$$

where $\mathrm{vech}_p(\cdot)$ is defined similarly to $\widetilde{\mathrm{vech}}_p(\cdot)$ as an operator that extracts all unique elements of q-dimensional symmetric array, but without multiplying each element by its frequency, so that the elements of $\mathrm{vech}_p(\boldsymbol{\lambda}_{h,\ell_1} \otimes \cdots \otimes \boldsymbol{\lambda}_{h,\ell_p})$ are conformable to those of $\widetilde{\mathrm{vech}}_p(\nabla_{\boldsymbol{\theta}^{\otimes p}} h_h^*/g_{M_0}^*)$.

Collect the relevant normalized reparameterized parameters and define $\boldsymbol{t}_{\psi_M^j \tau^j}$ as

$$\boldsymbol{t}_{\psi_M^j \tau^j} := \begin{pmatrix} \boldsymbol{\eta} - \boldsymbol{\eta}^* \\ \boldsymbol{v}_{\lambda_1^j \tau_1^j} \\ \vdots \\ \boldsymbol{v}_{\lambda_{M_0}^j \tau_{M_0}^j} \end{pmatrix}. \tag{10}$$

For $j = 1, \dots, J_M$, define the vector $\boldsymbol{s}^j(\boldsymbol{W})$ as

$$\boldsymbol{s}^j(\boldsymbol{W}) = \begin{pmatrix} \boldsymbol{s}_{\boldsymbol{\eta}}^j(\boldsymbol{W}) \\ \boldsymbol{s}_{\boldsymbol{\lambda}}^j(\boldsymbol{W}) \end{pmatrix} \quad \text{with} \quad \boldsymbol{s}_{\boldsymbol{\eta}}^j(\boldsymbol{W}) := \begin{pmatrix} \boldsymbol{s}_{\boldsymbol{\alpha}}^j(\boldsymbol{W}) \\ \boldsymbol{s}_{\boldsymbol{\nu}}^j(\boldsymbol{W}) \end{pmatrix} \quad \text{and} \quad \boldsymbol{s}_{\boldsymbol{\lambda}}^j(\boldsymbol{W}) := \begin{pmatrix} \boldsymbol{s}_{\lambda_1^j}^j(\boldsymbol{W}) \\ \vdots \\ \boldsymbol{s}_{\lambda_{M_0}^j}^j(\boldsymbol{W}) \end{pmatrix}, \tag{11}$$

where

$$\boldsymbol{s}_{\boldsymbol{\alpha}}^j(\boldsymbol{W}) = \begin{pmatrix} \frac{h_1(\boldsymbol{W};\boldsymbol{\theta}_1^*) - h_{M_0}(\boldsymbol{W};\boldsymbol{\theta}_{M_0}^*)}{g_{M_0}(\boldsymbol{W};\boldsymbol{\vartheta}_{M_0}^*)} \\ \vdots \\ \frac{h_{M_0-1}(\boldsymbol{W};\boldsymbol{\theta}_{M_0-1}^*) - h_{M_0}(\boldsymbol{W};\boldsymbol{\theta}_{M_0}^*)}{g_{M_0}(\boldsymbol{W};\boldsymbol{\vartheta}_{M_0}^*)} \end{pmatrix}, \quad \boldsymbol{s}_{\boldsymbol{\nu}}^j(\boldsymbol{W}) = \begin{pmatrix} \frac{\alpha_1^* \nabla_{\boldsymbol{\theta}} h(\boldsymbol{W};\boldsymbol{\theta}_1^*)}{g_{M_0}(\boldsymbol{W};\boldsymbol{\vartheta}_{M_0}^*)} \\ \vdots \\ \frac{\alpha_{M_0}^* \nabla_{\boldsymbol{\theta}} h(\boldsymbol{W};\boldsymbol{\theta}_{M_0}^*)}{g_{M_0}(\boldsymbol{W};\boldsymbol{\vartheta}_{M_0}^*)} \end{pmatrix}, \quad \text{and} \tag{12}$$

$$\boldsymbol{s}_{\lambda_h^j}^j(\boldsymbol{W}) = \begin{pmatrix} \widetilde{\mathrm{vech}}_2\left(\frac{\alpha_h^* \nabla_{\boldsymbol{\theta}\otimes\boldsymbol{\theta}} h(\boldsymbol{W};\boldsymbol{\theta}_h^*)}{g_{M_0}(\boldsymbol{W};\boldsymbol{\vartheta}_{M_0}^*)}\right) \\ \vdots \\ \widetilde{\mathrm{vech}}_{p_h^j}\left(\frac{\alpha_h^* \nabla_{\boldsymbol{\theta}^{\otimes p_h^j}} h(\boldsymbol{W};\boldsymbol{\theta}_h^*)}{g_{M_0}(\boldsymbol{W};\boldsymbol{\vartheta}_{M_0}^*)}\right) \end{pmatrix} \quad \text{for } h = 1, \dots, M_0. \tag{13}$$

We denote $\boldsymbol{s}_i^j := \boldsymbol{s}^j(\boldsymbol{W}_i)$. We assume that $l_{\boldsymbol{\Psi}_M^j \tau^j, i}$ can be expanded around $l_{\boldsymbol{\Psi}_M^j \tau^j, i} = 1$ as follows. Let $P_n(\boldsymbol{s}_i^j(\boldsymbol{s}_i^j)^\top) := n^{-1} \sum_{i=1}^n \boldsymbol{s}_i^j(\boldsymbol{s}_i^j)^\top$ and $\nu_n(\boldsymbol{s}_i^j) := n^{-1/2} \sum_{i=1}^n [\boldsymbol{s}_i^j - \mathbb{E}_{\boldsymbol{\theta}^*}[\boldsymbol{s}_i^j]]$.

For $\epsilon > 0$, define the neighborhood of the set $\boldsymbol{\Psi}_M^j$ corresponding to the null hypothesis $H_0 : M_0 = M$ as follows:

$$\mathcal{N}_\epsilon = \{\boldsymbol{\Psi}_M^j \in \Theta_{\boldsymbol{\Psi}_M^j} : \sup_{\boldsymbol{\tau}^j \in \Theta_{\boldsymbol{\tau}^j, c_1}} \|\boldsymbol{t}_{\psi_M^j \tau^j}\| < \epsilon\},$$

where $\Theta_{\boldsymbol{\tau}^j, c_1}$ denotes the parameter space of $\boldsymbol{\tau}^j$, as implied by the parameter space $\Theta_{\boldsymbol{\alpha}, c_1}$ of $\boldsymbol{\alpha}$, which restricts the values of τ_h^j to be bounded away from 0 and 1.

Assumption A.2. *For $M = M_0 + 1, \dots, \bar{M}$, the following assumption holds when the data is generated* *from the M_0-component model. For all $j = 1, \ldots, J_M$ and $i = 1, \ldots, n$, $l_{\mathbf{\Psi}_M^j \boldsymbol{\tau}^j, i} - 1$ admits an expansion*

$$l_{\mathbf{\Psi}_M^j \boldsymbol{\tau}^j, i} - 1 = \left(\boldsymbol{t}_{\psi_M^j \boldsymbol{\tau}^j}\right)^\top \boldsymbol{s}_i^j + r_{\mathbf{\Psi}_M^j \boldsymbol{\tau}^j, i}, \tag{14}$$

where $\boldsymbol{t}_{\psi_M^j \boldsymbol{\tau}^j}$ satisfies $\mathbf{\Psi}_M^j \to \mathbf{\Psi}_M^{j*}$ if $\sup_{\boldsymbol{\tau}^j \in \Theta_{\tau^j, c_1}} ||\boldsymbol{t}_{\psi_M^j \boldsymbol{\tau}^j}|| \to 0$, and $(\boldsymbol{s}_i^j, r_{\mathbf{\Psi}_M^j \boldsymbol{\tau}^j, i})$ satisfy, for some $C \in (0, \infty)$, $\delta > 0$, $\epsilon > 0$, and $\rho \in (0, 1)$, (a) $\mathbb{E}_{\boldsymbol{\vartheta}_{M_0}^*} ||\boldsymbol{s}_i^j||^{2+\delta} < C$, (b) $||P_n(\boldsymbol{s}_i^j (\boldsymbol{s}_i^j)^\top) - \boldsymbol{\mathcal{I}}^j|| = o_p(1)$, where $\boldsymbol{\mathcal{I}}^j := \mathbb{E}[\boldsymbol{s}_i^j (\boldsymbol{s}_i^j)^\top]$ is finite and non-singular, (c) $\mathbb{E}_{\boldsymbol{\theta}^*}[\sup_{\mathbf{\Psi}_M^j \in \mathcal{N}_\varepsilon} |r_{\mathbf{\Psi}_M^j \boldsymbol{\tau}^j, i} / (||\boldsymbol{t}_{\psi_M^j \boldsymbol{\tau}^j}|| ||\mathbf{\Psi}_M^j - \mathbf{\Psi}_M^{j*}||)|^2] < \infty$, (d) $\sup_{\mathbf{\Psi}_M^j \in \mathcal{N}_\varepsilon} [\nu_n(r_{\mathbf{\Psi}_M^j \boldsymbol{\tau}^j, i}) / (||\boldsymbol{t}_{\psi_M^j \boldsymbol{\tau}^j}|| ||\mathbf{\Psi}_M^j - \mathbf{\Psi}_M^{j*}||)] = O_p(1)$, and (e) $\sup_{\mathbf{\Psi}_M^j \in \mathcal{N}_\varepsilon} ||\nu_n(\boldsymbol{s}_i^j)|| = O_p(1)$.

Assumption A.2(b) is a key regularity condition that requires the unique elements of $\nabla_{\boldsymbol{\theta}} h_h^* / g_{M_0}^*$, $\nabla_{\boldsymbol{\theta} \otimes \boldsymbol{\theta}} h_h^* / g_{M_0}^*$, ..., $\nabla_{\boldsymbol{\theta}^{\otimes p_h}} h_h^* / g_{M_0}^*$ are linearly independent and their expectation is finite.

Proposition A.2. *Suppose that Assumption A.2 holds. Then, for any $c > 0$,*

$$\sup_{\mathbf{\Psi}_M^j \in \mathcal{N}_{c/\sqrt{n}}} \left| \ell_n(\mathbf{\Psi}_M^j, \boldsymbol{\tau}^j) - \ell_n(\mathbf{\Psi}_M^{j*}, \boldsymbol{\tau}^j) - \sqrt{n} \boldsymbol{t}_{\mathbf{\Psi}_M^j \boldsymbol{\tau}^j}^\top \nu_n(\boldsymbol{s}_i^j) + n \boldsymbol{t}_{\mathbf{\Psi}_M^j \boldsymbol{\tau}^j}^\top \boldsymbol{\mathcal{I}}^j \boldsymbol{t}_{\mathbf{\Psi}_M^j \boldsymbol{\tau}^j} / 2 \right| = o_p(1).$$

The following proposition expands $\ell_n(\mathbf{\Psi}_M^j, \boldsymbol{\tau}^j)$ in $A_{\varepsilon_n}(\eta) := \{\mathbf{\Psi}_M \in \mathcal{N}_{\varepsilon_n} : \ell_n(\mathbf{\Psi}_M^j, \boldsymbol{\tau}^j) - \ell_n(\mathbf{\Psi}_M^{j*}, \boldsymbol{\tau}^j) \geq -\eta\}$ for some $\eta \in [0, \infty)$ and for $\varepsilon_n \to 0$ slowly to ensure $\Pr(\hat{\mathbf{\Psi}}_M \in \mathcal{N}_{\varepsilon_n}) = 1$.

Proposition A.3. *Suppose that Assumption A.2 holds. Then, for all $j = 1, \ldots, J_M$, for any $\eta > 0$, and for any $\{\varepsilon_n : n = 1, 2, \ldots\}$ such that $\varepsilon_n \to 0$ but $\Pr(\hat{\mathbf{\Psi}}_M \in \mathcal{N}_{\varepsilon_n}) = 1$,*
(a) $\sup_{\mathbf{\Psi}_M \in A_{\varepsilon_n}(\eta)} \left| \boldsymbol{t}_{\psi_M^j \boldsymbol{\tau}^j} \right| = O_p(n^{-1/2})$, (b) for any $c > 0$,

$$\sup_{\mathbf{\Psi}_M \in A_{\varepsilon_n}(\eta) \cup \mathcal{N}_{c/\sqrt{n}}} \left| \ell_n(\mathbf{\Psi}_M^j, \boldsymbol{\tau}^j) - \ell_n(\mathbf{\Psi}_M^{j*}, \boldsymbol{\tau}^j) - \sqrt{n} \boldsymbol{t}_{\psi_M^j \boldsymbol{\tau}^j} \nu_n(\boldsymbol{s}_i^j) + n \boldsymbol{t}_{\psi_M^j \boldsymbol{\tau}^j} \boldsymbol{\mathcal{I}}^j \boldsymbol{t}_{\psi_M^j \boldsymbol{\tau}^j} / 2 \right| = o_p(1),$$

and (c) $\sup_{\mathbf{\Psi}_M \in A_{\varepsilon_n}(\eta)} \left| \ell_n(\mathbf{\Psi}_M^j, \boldsymbol{\tau}^j) - \ell_n(\mathbf{\Psi}_M^{j*}, \boldsymbol{\tau}^j) \right| = O_p(1)$.

Because a consistent MLE is in $A_{\varepsilon_n}(\eta)$ by definition, Proposition A.3 implies that $\ell_n(\hat{\mathbf{\Psi}}_M^j, \boldsymbol{\tau}^j) - \ell_n(\mathbf{\Psi}_M^{j*}, \boldsymbol{\tau}^j) = O_p(1)$ for a consistent MLE. Together, Propositions A.2 and A.3 verify Assumption A.4, completing the justification for BIC consistency in Proposition A.6.

A.3 Technical Details for LRT

A.3.1 Parameter Space Partition and Local Restrictions

Recall that $\mu_1^* < \mu_2^* < \cdots < \mu_{M_0}^*$. Let $\underline{\Theta}_\mu$ and $\overline{\Theta}_\mu$ denote the lower and upper bounds of Θ_μ. Define

$$\Theta_{\boldsymbol{\theta}, h}^* = \begin{cases} [\underline{\Theta}_\mu, \frac{\mu_1^* + \mu_2^*}{2}] \times \Theta_\beta \times \Theta_{\sigma^2}, & h = 1 \\ [\frac{\mu_{h-1}^* + \mu_h^*}{2}, \frac{\mu_h^* + \mu_{h+1}^*}{2}] \times \Theta_\beta \times \Theta_{\sigma^2}, & 2 \leq h \leq M_0 - 1 \\ [\frac{\mu_{M_0-1}^* + \mu_{M_0}^*}{2}, \overline{\Theta}_\mu] \times \Theta_\beta \times \Theta_{\sigma^2}, & h = M_0. \end{cases} \tag{15}$$

Then $\{\Theta^*_{\boldsymbol{\theta},h}\}_{h=1}^{M_0}$ partitions $\Theta_{\boldsymbol{\theta}}$ such that $\Theta^*_{\boldsymbol{\theta},h}$ contains $\boldsymbol{\theta}^*_h$ but not $\boldsymbol{\theta}^*_j$ for $j \neq h$. Define the restricted parameter space $\boldsymbol{\Psi}^*_h \subset \bar{\Theta}_{\boldsymbol{\vartheta}_{M_0+1}}(\boldsymbol{c})$ as

$$\boldsymbol{\Psi}^*_h = \left\{ \begin{array}{l} \boldsymbol{\vartheta}_{M_0+1} \in \bar{\Theta}_{\boldsymbol{\vartheta}_{M_0+1}}(\boldsymbol{c}) : \sum_{j=1}^{M_0+1} \alpha_j = 1; \\ \boldsymbol{\theta}_j \in \Theta^*_{\boldsymbol{\theta},j} \text{ for } j \leq h-1; \quad \boldsymbol{\theta}_h, \boldsymbol{\theta}_{h+1} \in \Theta^*_{\boldsymbol{\theta},h}; \\ \boldsymbol{\theta}_j \in \Theta^*_{\boldsymbol{\theta},j-1} \text{ for } j \geq h+2 \end{array} \right\}. \tag{16}$$

By construction, $\boldsymbol{\Psi}^*_h \cap \Theta^*_{\boldsymbol{\vartheta}_{M_0+1},1h} \neq \emptyset$, $\boldsymbol{\Psi}^*_h \cap \Theta^*_{\boldsymbol{\vartheta}_{M_0+1},1l} = \emptyset$ for $h \neq l$, and $\cup_{h=1}^{M_0} \boldsymbol{\Psi}^*_h = \bar{\Theta}_{\boldsymbol{\vartheta}_{M_0+1}}(\boldsymbol{c})$.

The subset $\Theta^*_{\boldsymbol{\vartheta}_{M_0+1},1h}$ corresponding to $H_{0,1h}$ is defined as

$$\Theta^*_{\boldsymbol{\vartheta}_{M_0+1},1h} := \left\{ \begin{array}{l} \boldsymbol{\vartheta}_{M_0+1} \in \Theta_{\vartheta_{M_0+1}} : \alpha_h, \alpha_{h+1} > 0; \quad \alpha_h + \alpha_{h+1} = \alpha^*_h; \\ \boldsymbol{\theta}_h = \boldsymbol{\theta}_{h+1} = \boldsymbol{\theta}^*_h; \quad \alpha_j = \alpha^*_j, \boldsymbol{\theta}_j = \boldsymbol{\theta}^*_j \text{ for } j < h; \\ \alpha_j = \alpha^*_{j-1}, \boldsymbol{\theta}_j = \boldsymbol{\theta}^*_{j-1} \text{ for } j > h+1 \end{array} \right\} \tag{17}$$

for $h = 1, \ldots, M_0$.

A.3.2 Explicit Score Function via Hermite Polynomials

$H^j(\cdot)$ 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 $\{\frac{1}{\sigma}\phi(\frac{t}{\sigma})\}$ is

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

Let $f^* = f(\boldsymbol{W}; \theta^*) = \prod_{t=1}^T f^*_t(\boldsymbol{W}; \theta^*)$, where f^*_t denotes the component density for observation t, which may be either a normal or a mixture density. Let $\nabla f^* = \nabla f(\boldsymbol{W}; \theta^*)$ denote its gradient.

Define the Hermite polynomials evaluated at $Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta}^* - \mu^*$ as follows:

$$H^{j*}_t = \frac{1}{(\sigma^*)^j} H^j\left(\frac{Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta}^* - \mu^*}{\sigma^*}\right). \tag{18}$$

A.3.3 Score function for the model (1) with normal density ((2))-((3))

Let $f^*_t = \frac{1}{\sigma^*}\phi\left(\frac{Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta}^* - \mu^*}{\sigma^*}\right)$. Then, the first-order derivatives of the normal density function ((2))-((3)) with respect to the parameters are given by:

$$\nabla_\mu f^* = f^* \sum_{t=1}^T H^{1*}_t, \qquad \nabla_{\sigma^2} f^* = f^* \sum_{t=1}^T \frac{1}{2} H^{2*}_t,$$

$$\nabla_{\boldsymbol{\beta}} f^* = f^* \sum_{t=1}^T H^{1*}_t \boldsymbol{X}_t.$$

The score function defined in (2) is then written in terms of the Hermite polynomials:

$$\boldsymbol{s}_{\boldsymbol{\eta}}(\boldsymbol{W}) = \begin{pmatrix} s_\mu \\ s_\sigma \\ \boldsymbol{s}_{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^T H_t^{1*} \\ \sum_{t=1}^T H_t^{2*} \\ \sum_{t=1}^T H_t^{1*} \boldsymbol{x}_t \end{pmatrix} \quad \text{and} \quad \boldsymbol{s}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W}) = \begin{pmatrix} s_{\lambda_\mu \lambda_\mu} \\ s_{\lambda_\mu \lambda_\sigma} \\ s_{\lambda_\sigma \lambda_\sigma} \\ \boldsymbol{s}_{\lambda_\mu \lambda_{\boldsymbol{\beta}}} \\ \boldsymbol{s}_{\lambda_\sigma \lambda_{\boldsymbol{\beta}}} \\ \boldsymbol{s}_{\lambda_{\boldsymbol{\beta}} \lambda_{\boldsymbol{\beta}}} \end{pmatrix}, \tag{19}$$

where

$$\begin{pmatrix} s_{\lambda_\mu \lambda_\mu} \\ s_{\lambda_\mu \lambda_\sigma} \\ s_{\lambda_\sigma \lambda_\sigma} \\ \boldsymbol{s}_{\lambda_\mu \lambda_{\boldsymbol{\beta}}} \\ \boldsymbol{s}_{\lambda_\sigma \lambda_{\boldsymbol{\beta}}} \\ \boldsymbol{s}_{\lambda_{\boldsymbol{\beta}} \lambda_{\boldsymbol{\beta}}} \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^T \left[H_t^{2*} + \sum_{s \neq t} H_t^{1*} H_s^{1*} \right] \\ \frac{1}{2} \sum_{t=1}^T \left[H_t^{3*} + \sum_{s \neq t} H_t^{1*} H_s^{2*} \right] \\ \frac{1}{4} \sum_{t=1}^T \left[H_t^{4*} + \sum_{s \neq t} H_t^{2*} H_s^{2*} \right] \\ \sum_{t=1}^T \left[H_t^{2*} + \sum_{s \neq t} H_t^{1*} H_s^{1*} \right] \boldsymbol{X}_t \\ \frac{1}{2} \sum_{t=1}^T \left[H_t^{3*} + \sum_{s \neq t} H_t^{1*} H_s^{2*} \right] \boldsymbol{X}_t \\ \sum_{t=1}^T \left[H_t^{2*} \widetilde{\text{vech}}(\boldsymbol{X}_t \boldsymbol{X}_t^\top) + \sum_{s \neq t} H_t^{1*} H_s^{1*} \widetilde{\text{vech}}(\boldsymbol{X}_t \boldsymbol{X}_s^\top) \right] \end{pmatrix}. \tag{20}$$

A.3.4 Score function for the model (1) with normal mixture density (2)–(3) when $K_\epsilon = 2$

For brevity, we present the score function for the model (1) with a two-component normal mixture (2)–(3) when there are no covariates. The score function for the model with covariates can be analogously derived. Let

$$f^* = \prod_{t=1}^T f_t^* = \prod_{t=1}^T \left(\tau^* f_{1t}^* + (1 - \tau^*) f_{2t}^* \right),$$

where $f_{kt}^* = \frac{1}{\sigma^*} \phi \left(\frac{Y_t - \mu_k^*}{\sigma^*} \right)$ for $k = 1, 2$. In this model, we omit the covariates $\boldsymbol{X}_t$ and the parameter $\boldsymbol{\beta}^*$ for simplicity. The score function with covariates can be derived similarly to the previous section. Define the b-th order normalized Hermite polynomial for k-th component evaluated at $Y_t - \mu_k^*$ as follows:

$$H_{kt}^{b*} = \frac{1}{(\sigma^*)^b} H^b \left(\frac{Y_t - \mu_k^*}{\sigma^*} \right). \tag{21}$$

Let

$$\gamma_{1t} := \frac{\tau^* f_{1t}^*}{\tau^* f_{1t}^* + (1 - \tau^*) f_{2t}^*}, \quad \gamma_{2t} := 1 - \gamma_{1t}.$$

The score function defined in (2) is then written as:

$$
\boldsymbol{s}_{\boldsymbol{\eta}}(\boldsymbol{W})=\begin{pmatrix} s_\tau \\ s_{\mu_1} \\ s_{\mu_2} \\ s_\sigma \end{pmatrix}=\begin{pmatrix} \sum_{t=1}^T\left(\frac{\gamma_{1t}}{\tau^*}-\frac{\gamma_{2t}}{1-\tau^*}\right) \\ \sum_{t=1}^T \gamma_{1t}H_{1t}^{1*} \\ \sum_{t=1}^T \gamma_{2t}H_{2t}^{1*} \\ \sum_{t=1}^T \gamma_{1t}H_{1t}^{2*}+\gamma_{2t}H_{2t}^{2*} \end{pmatrix} \quad \text{and} \quad \boldsymbol{s}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W})=\begin{pmatrix} s_{\lambda_\tau\lambda_\tau} \\ s_{\lambda_{\mu_1}\lambda_{\mu_1}} \\ s_{\lambda_{\mu_2}\lambda_{\mu_2}} \\ s_{\lambda_{\mu_1}\lambda_{\mu_2}} \\ s_{\lambda_\tau\lambda_{\mu_1}} \\ s_{\lambda_\tau\lambda_{\mu_2}} \\ s_{\lambda_{\mu_1}\lambda_\sigma} \\ s_{\lambda_{\mu_2}\lambda_\sigma} \\ s_{\lambda_\sigma\lambda_\sigma} \\ s_{\lambda_\tau\lambda_\sigma} \end{pmatrix}, \tag{22}
$$

where

$$
\begin{pmatrix} s_{\lambda_\tau\lambda_\tau} \\ s_{\lambda_{\mu_1}\lambda_{\mu_1}} \\ s_{\lambda_{\mu_2}\lambda_{\mu_2}} \\ s_{\lambda_{\mu_1}\lambda_{\mu_2}} \\ s_{\lambda_\tau\lambda_{\mu_1}} \\ s_{\lambda_\tau\lambda_{\mu_2}} \\ s_{\lambda_{\mu_1}\lambda_\sigma} \\ s_{\lambda_{\mu_2}\lambda_\sigma} \\ s_{\lambda_\sigma\lambda_\sigma} \\ s_{\lambda_\tau\lambda_\sigma} \end{pmatrix}=\begin{pmatrix} \sum_{t=1}^T\sum_{s\neq t}\left(\frac{\gamma_{1t}}{\tau^*}-\frac{\gamma_{2t}}{1-\tau^*}\right)\left(\frac{\gamma_{1s}}{\tau^*}-\frac{\gamma_{2s}}{1-\tau^*}\right) \\ \sum_{t=1}^T\left(\gamma_{1t}H_{1t}^{2*}+\sum_{s\neq t}\gamma_{1t}\gamma_{1s}H_{1t}^{1*}H_{1s}^{1*}\right) \\ \sum_{t=1}^T\left(\gamma_{2t}H_{2t}^{2*}+\sum_{s\neq t}\gamma_{2t}\gamma_{2s}H_{2t}^{1*}H_{2s}^{1*}\right) \\ \sum_{t=1}^T\left(\sum_{s\neq t}\gamma_{1t}\gamma_{2s}H_{1t}^{1*}H_{2s}^{1*}\right) \\ \sum_{t=1}^T\left(\frac{\gamma_{1t}}{\tau^*}H_{1t}^{1*}+\sum_{s\neq t}\gamma_{1t}H_{1t}^{1*}\left(\frac{\gamma_{1s}}{\tau^*}-\frac{\gamma_{2s}}{1-\tau^*}\right)\right) \\ \sum_{t=1}^T\left(-\frac{\gamma_{2t}}{1-\tau^*}H_{2t}^{1*}+\sum_{s\neq t}\gamma_{2t}H_{2t}^{1*}\left(\frac{\gamma_{1s}}{\tau^*}-\frac{\gamma_{2s}}{1-\tau^*}\right)\right) \\ \frac{1}{2}\sum_{t=1}^T\left(\gamma_{1t}H_{1t}^{3*}+\sum_{s\neq t}\gamma_{1t}H_{1t}^{1*}\left(\gamma_{1s}H_{1s}^{2*}+\gamma_{2s}H_{2s}^{2*}\right)\right) \\ \frac{1}{2}\sum_{t=1}^T\left(\gamma_{2t}H_{2t}^{3*}+\sum_{s\neq t}\gamma_{2t}H_{2t}^{1*}\left(\gamma_{1s}H_{1s}^{2*}+\gamma_{2s}H_{2s}^{2*}\right)\right) \\ \frac{1}{4}\sum_{t=1}^T\left(\left(\gamma_{1t}H_{1t}^{4*}+\gamma_{2t}H_{2t}^{4*}\right)+\sum_{s\neq t}\left(\gamma_{1t}H_{1t}^{2*}+\gamma_{2t}H_{2t}^{2*}\right)\left(\gamma_{1s}H_{1s}^{2*}+\gamma_{2s}H_{2s}^{2*}\right)\right) \\ \frac{1}{2}\sum_{t=1}^T\left(\frac{\gamma_{1t}}{\tau^*}H_{1t}^{2*}-\frac{\gamma_{2t}}{1-\tau^*}H_{2t}^{2*}+\sum_{s\neq t}\left(\gamma_{1t}H_{1t}^{2*}+\gamma_{2t}H_{2t}^{2*}\right)\left(\frac{\gamma_{1s}}{\tau^*}-\frac{\gamma_{2s}}{1-\tau^*}\right)\right) \end{pmatrix}. \tag{23}
$$

A.3.5 Score function for testing $H_0 : M = M_0$ against $H_A : M = M_0 + 1$

We only present score function for the model (1) with normal density ((2))-((3)).

Let $g^* = g(\boldsymbol{W};\vartheta^*_{M_0}) = \sum_{j=1}^{M_0}\alpha_j^* f_j^*$ denote the true M_0-component model as in equation (11), where $f_j^* = f(\boldsymbol{W};\theta_j^*) = \prod_{t=1}^T \frac{1}{\sigma_j^*}\phi\left(\frac{Y_t-\mu_j^*}{\sigma_j^*}\right)$ is the j-th component density. In this section, we omit the covariate $\boldsymbol{x}_{it}$ and the parameters β_j^* for simplicity; the derivation with covariates can be done analogously to the previous section. Define $H_{j,it}^{b*}$ as a shorthand for the b-th order normalized Hermite polynomial evaluated at $\frac{Y_{it}-\mu_j^*}{\sigma_j^*}$, i.e., $H_{j,t}^{b*} = \frac{1}{(\sigma_j^*)^b}H^b\left(\frac{Y_t-\mu_j^*}{\sigma_j^*}\right)$. Define the weight w_i^{j*} as

$w_j^* = \frac{\alpha_j^* f_j^*}{\sum_l \alpha_l^* f_l^*}, \quad j = 1, \ldots, M_0$. The score functions are

$$\boldsymbol{s}_{\boldsymbol{\alpha}}(\boldsymbol{W}) = \begin{pmatrix} \frac{w_1^*}{\alpha_1^*} - \frac{w_{M_0}^*}{\alpha_{M_0}^*} \\ \vdots \\ \frac{w_{M_0-1}^*}{\alpha_{M_0-1}^*} - \frac{w_{M_0}^*}{\alpha_{M_0}^*} \end{pmatrix}, \qquad \boldsymbol{s}_{\mu}(\boldsymbol{W}) = \begin{pmatrix} w_1^* \sum_{t=1}^T H_{1,t}^{1*} \\ \vdots \\ w_{M_0}^* \sum_{t=1}^T H_{M_0,t}^{1*} \end{pmatrix},$$

$$\boldsymbol{s}_{\sigma}(\boldsymbol{W}) = \begin{pmatrix} w_1^* \sum_{t=1}^T H_{1,t}^{2*} \\ \vdots \\ w_{M_0}^* \sum_{t=1}^T H_{M_0,t}^{2*} \end{pmatrix}, \qquad \boldsymbol{s}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{W}) = \begin{pmatrix} \boldsymbol{s}_{\lambda_\mu\lambda_\mu} \\ \boldsymbol{s}_{\lambda_\mu\lambda_\sigma} \\ \boldsymbol{s}_{\lambda_\sigma\lambda_\sigma} \end{pmatrix},$$

where

$$\boldsymbol{s}_{\lambda_\mu\lambda_\mu} = \begin{pmatrix} w_1^* \sum_{t=1}^T \left[H_{1,t}^{2*} + \sum_{s\neq t} H_{1,t}^{1*} H_{1,s}^{1*} \right] \\ \vdots \\ w_{M_0}^* \sum_{t=1}^T \left[H_{M_0,t}^{2*} + \sum_{s\neq t} H_{M_0,t}^{1*} H_{M_0,s}^{1*} \right] \end{pmatrix}, \boldsymbol{s}_{\lambda_\mu\lambda_\sigma} = \begin{pmatrix} w_1^* \frac{1}{2} \sum_{t=1}^T \left[H_{1,t}^{3*} + \sum_{s\neq t} H_{1,t}^{1*} H_{1,s}^{2*} \right] \\ \vdots \\ w_{M_0}^* \frac{1}{2} \sum_{t=1}^T \left[H_{M_0,t}^{3*} + \sum_{s\neq t} H_{M_0,t}^{1*} H_{M_0,s}^{2*} \right] \end{pmatrix},$$

$$\boldsymbol{s}_{\lambda_\sigma\lambda_\sigma} = \begin{pmatrix} w_1^* \frac{1}{4} \sum_{t=1}^T \left[H_{1,t}^{4*} + \sum_{s\neq t} H_{1,t}^{2*} H_{1,s}^{2*} \right] \\ \vdots \\ w_{M_0}^* \frac{1}{4} \sum_{t=1}^T \left[H_{M_0,t}^{4*} + \sum_{s\neq t} H_{M_0,t}^{2*} H_{M_0,s}^{2*} \right] \end{pmatrix}.$$

A.4 Asymptotic distribution under local alternatives

We derive the asymptotic distribution of the LRTS under local alternatives. For brevity, we focus on testing $H_0 : M_0 = 1$ against $H_A : M_0 = 2$. Consider the following local alternative to the homogeneous model $f(\boldsymbol{w}; \boldsymbol{\gamma}^*, \boldsymbol{\theta}^*)$ with $\boldsymbol{\theta}^* = (\mu^*, \sigma^{*2}, (\boldsymbol{\beta}^*)^\top)^\top$. For brevity, we omit the common parameter $\boldsymbol{\gamma}$ in this section. In a reparameterized parameter, $\boldsymbol{\Psi}^* = ((\boldsymbol{\nu}^*)^\top, (\boldsymbol{\lambda}^*)^\top)^\top$. For $\alpha^* \in (0,1)$ and a local parameter $\boldsymbol{h} = (\boldsymbol{h}_{\boldsymbol{\nu}}^\top, \boldsymbol{h}_{\boldsymbol{\lambda}}^\top)^\top$ with $\boldsymbol{h}_{\boldsymbol{\lambda}} \in v(\boldsymbol{\theta}_{\boldsymbol{\lambda}})$, we consider a sequence of contiguous local alternatives $(\alpha_n, \boldsymbol{\psi}_n^\top)^\top = (\alpha_n, \boldsymbol{\nu}_n^\top, \boldsymbol{\lambda}_n^\top) \in \boldsymbol{\Theta}_\alpha \times \boldsymbol{\Theta}_{\boldsymbol{\nu}} \times \boldsymbol{\Theta}_{\boldsymbol{\lambda}}$ such that, with $\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}, \alpha)$ given by (35),

$$\boldsymbol{h}_{\boldsymbol{\nu}} = \sqrt{n}(\boldsymbol{\nu}_n - \boldsymbol{\nu}^*), \quad \boldsymbol{h}_{\boldsymbol{\lambda}} = \sqrt{n}\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_n, \alpha_n), \quad \text{and} \quad \alpha_n = \alpha^* + o(1). \tag{24}$$

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

$$\boldsymbol{\theta}_{1,n} = \boldsymbol{\nu}_n + (1 - \alpha_n)\boldsymbol{\lambda}_n \quad \text{and} \quad \boldsymbol{\theta}_{2,n} = \boldsymbol{\nu}_n - \alpha_n \boldsymbol{\lambda}_n \tag{25}$$

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

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

where $\boldsymbol{h}_{\boldsymbol{\lambda}} = (h_{\lambda,1}^2, 2h_{\lambda,1}h_{\lambda,2}, h_{\lambda,2}^2, ..., 2h_{\lambda,q-1}h_{\lambda,q}, h_{\lambda,q}^2)^\top$. 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 LRT test statistics under contiguous local alternatives.

Proposition A.4. *Suppose that the assumptions in Proposition 6 hold for $M_0 = 1$. Consider a sequence of contiguous local alternatives $\vartheta_{2,n} = (\alpha_n, \theta_{1,n}^\top, \theta_{2,n}^\top)^\top$ given in (25), where α_n and λ_n satisfy (24). Then, under $H_{1,n} : \vartheta = \vartheta_{2,n}$, we have $LR_n \xrightarrow{d} (\tilde{t}_\lambda)^\top \mathcal{I}_{\lambda,\nu} \tilde{t}_\lambda$, where $\tilde{t}_\lambda$ has the same distribution as $\widehat{t}_\lambda$ in Proposition 4 but $G_{\lambda,\nu}$ is replaced with $(\mathcal{I}_{\lambda,\nu})^{-1} S_{\lambda,\nu} + h_\lambda$.*

The result follows from Le Cam's third lemma. To apply it, one needs the joint convergence under $P^n_{\nu^*}$ of the score vector $S_{\lambda,\nu,n}$ and the log-likelihood ratio $\Lambda_n := \log(dP^n_{\vartheta_{2,n}}/dP^n_{\nu^*})$, together with the LAN property $\Lambda_n \xrightarrow{d} N(-\frac{1}{2}h^\top \mathcal{I} h,\ h^\top \mathcal{I} h)$ and cross-covariance $\mathrm{Cov}(S_{\lambda,\nu}, \Lambda) = \mathcal{I}_{\lambda,\nu} h_\lambda$. For regular (non-singular Fisher information) models this is standard; for the present singular mixture model (where the Fisher information in the λ direction is degenerate at $\lambda = \mathbf{0}$) the required LAN structure — joint convergence of $(S_{\lambda,\nu,n}, \Lambda_n)$ and the cross-covariance identity — follows from the local-likelihood-ratio expansion in Kasahara and Shimotsu (2015) adapted to the reparameterised score $\sqrt{n} t_\lambda$. Given this LAN structure, Le Cam's third lemma shifts $G_{\lambda,\nu}$ from $N(\mathbf{0}, \mathcal{I}^{-1}_{\lambda,\nu})$ to $N(h_\lambda, \mathcal{I}^{-1}_{\lambda,\nu})$, yielding the stated distribution.

The set of contiguous local alternatives considered in (25) excludes a sequence such that $\alpha_n \to 0$ or 1.

A.5 Proof of Propositions

Proof of Proposition 1. We first consider a model with an intercept parameter and a variance parameter but without covariates with $\mathbf{W}_i = \{y_{it}\}_{t=1}^T$ for the model (1) with normal density ((2))-((3)).

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

Define

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

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

$$\begin{aligned}
LR_n^*(\vartheta_{2,n}) &= 2\left\{\sum_{i=1}^n \log\left(\alpha_n \prod_{t=1}^T \frac{1}{\sigma_{1,n}}\phi\left(\frac{Y_{it} - \mu_{1,n}}{\sigma_{1,n}}\right) + (1-\alpha_n)\prod_{t=1}^T \frac{1}{\sigma^*}\phi\left(\frac{Y_{it} - \mu^*}{\sigma^*}\right)\right) - \sum_{i=1}^n \ell(\mathbf{W}_i; \theta^*)\right\} \\
&= 2\sum_{i \neq i^*}\left\{\log\left(\exp(\log\alpha_n + \ell(\mathbf{W}_i; \theta_{1,n})) + \exp(\log(1-\alpha_n) + \ell(\mathbf{W}_i; \theta^*))\right) - \ell(\mathbf{W}_i; \theta^*)\right\} \\
&\quad + 2\left\{\log\left(\exp(\log\alpha_n + \ell(\mathbf{W}_{i^*}; \theta_{1,n})) + \exp(\log(1-\alpha_n) + \ell(\mathbf{W}_{i^*}; \theta^*))\right) - \ell(\mathbf{W}_{i^*}; \theta^*)\right\}.
\end{aligned} \tag{26}$$

The first term on the right-hand side of (26) can be rewritten as

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

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

The second term on the right-hand side of (26) is written as

$$2\{-\log n+\ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}_{1,n})\}+2\log\left(1+(n-1)\exp(\ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}^*)-\ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}_{1,n}))\right)-2\ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}^*), \quad (27)$$

where $2\{-\log n+\ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}_{1,n})\}$ diverges to infinity as $n\to\infty$ by Lemma 2, the second term in (27) is bounded below by zero, and the third term is bounded in probability because $\ell(\boldsymbol{W}_{i^*};\boldsymbol{\theta}^*)=O_p(1)$. Therefore, for any $M<\infty$, we have $\Pr\left(LR_n^*(\boldsymbol{\vartheta}_{2,n})\leq M\right)\to 0$ as $n\to\infty$.

For a model with covariates for the model (1) with normal density ((2))-((3)), we can consider a sequence of parameters $\boldsymbol{\vartheta}_{2,n}=(\alpha_n,\boldsymbol{\theta}_{1,n}^\top,\boldsymbol{\theta}_{2,n}^\top)^\top$ with $\alpha_n=1/n$, $\boldsymbol{\theta}_{1,n}=(\mu_{1,n},\sigma_{1,n}^2,\boldsymbol{\beta}_{1,n}^\top)^\top=(\bar{Y}_{i^*},s_{i^*}^2,\boldsymbol{0}^\top)^\top$ with $\boldsymbol{\theta}_{2,n}=\boldsymbol{\theta}^*=(\mu^*,\sigma^{*2},(\boldsymbol{\beta}^*)^\top)^\top$. Then, repeating the above argument, the stated result follows.

For the model (1) with normal mixture density (2)–(3), suppose that $\mu_{j1}<\mu_{j2}$ for $j=1,2$. Then, in view of (44), we can bound the log-likelihood function from below as

$$\begin{aligned}&\sum_{i=1}^n\log\left(\alpha\prod_{t=1}^T\sum_{k=1}^{K_\epsilon}\tau_{1k}\frac{1}{\sigma_1}\phi\left(\frac{Y_{it}-\mu_{1k}}{\sigma_1}\right)+(1-\alpha)\prod_{t=1}^T\sum_{k=1}^{K_\epsilon}\tau_{2k}\frac{1}{\sigma_2}\phi\left(\frac{Y_{it}-\mu_{2k}}{\sigma_2}\right)\right)\\ \geq&\sum_{i=1}^n\log\left(\alpha\prod_{t=1}^T\frac{1}{\sigma_1}\phi\left(\frac{Y_{it}-\mu_{11}}{\sigma_1}\right)+(1-\alpha)\prod_{t=1}^T\frac{1}{\sigma_2}\phi\left(\frac{Y_{it}-\mu_{21}}{\sigma_2}\right)\right).\end{aligned}$$

Therefore, the LRT statistic for a two-component mixture, $LR_n^*(\boldsymbol{\vartheta}_{2,n})$, is bounded below by the right hand side of (26). Then, repeating the argument following (26), the stated result follows from Lemma 2. This proves part (i).

□

Proof of Proposition 2. Note that, for $Z=\frac{Y-\mu}{\sigma}$,

$$\frac{\partial}{(\partial\mu)}\left[\tfrac{1}{\sigma}\,\phi(Z)\right]=\tfrac{Z}{\sigma}\,\frac{1}{\sigma}\,\phi(Z),\quad \frac{\partial^2}{(\partial\mu)^2}\left[\tfrac{1}{\sigma}\,\phi(Z)\right]=\tfrac{Z^2-1}{\sigma^2}\,\frac{1}{\sigma}\,\phi(Z),$$

$$\text{and}\quad\frac{\partial}{\partial\sigma^2}\left[\tfrac{1}{\sigma}\,\phi(Z)\right]=\tfrac{1}{2\sigma^2}\,(Z^2-1)\,\frac{1}{\sigma}\,\phi(Z).$$

Hence, for $j=1$ (the case $j=2$ follows identically with $\bar{\alpha}$ replaced by $1-\bar{\alpha}$), a straightforward calculation gives

$$\left.\frac{\partial^2 g_2}{(\partial\mu_j)^2}\right|_{(\bar{\alpha},\theta^*,\theta^*)}=\bar{\alpha}\,f^*\,\frac{1}{\sigma^{*2}}\left[(Z_1+\cdots+Z_T)^2-T\right]\quad\text{and}\quad\left.\frac{\partial g_2}{\partial\sigma_j^2}\right|_{(\bar{\alpha},\theta^*,\theta^*)}=\bar{\alpha}\,f^*\,\frac{1}{2\,\sigma^{*2}}\sum_{t=1}^T(Z_t^2-1), \quad (28)$$

where

$$Z_t := \frac{Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta}^* - \mu^*}{\sigma^*}, \qquad t = 1, \dots, T,$$

is the standardised residual obtained by conditioning on $\{\boldsymbol{X}_t\}$; all expectations below are taken conditionally on $\{\boldsymbol{X}_t\}$.

Suppose, for contradiction, that real constants a, b exist such that

$$\left.\frac{\partial^2 g_2}{(\partial \mu_j)^2}\right|_{(\bar{\alpha},\theta^*,\theta^*)} \;=\; a \;+\; b\; \left.\frac{\partial g_2}{\partial \sigma_j^2}\right|_{(\bar{\alpha},\theta^*,\theta^*)}$$

with positive probability. Since $\bar{\alpha} f^*/\sigma^{*2} > 0$ almost surely, substituting (28) into the above equation and dividing both sides by $\bar{\alpha} f^*/\sigma^{*2}$ gives

$$(Z_1 + \cdots + Z_T)^2 \;-\; T \;=\; \frac{a\,\sigma^{*2}}{\bar{\alpha} f^*} \;+\; \frac{b}{2}\sum_{t=1}^{T}\big(Z_t^2 - 1\big). \tag{29}$$

We derive a contradiction by considering two cases.

Case 1: $a = 0$. The right-hand side of (29) reduces to $\frac{b}{2}\sum_t (Z_t^2 - 1)$, so the equation becomes

$$\Big(1 - \tfrac{b}{2}\Big)\sum_{t=1}^{T} Z_t^2 \;+\; 2\sum_{1\le s<r\le T} Z_s Z_r \;=\; T \;-\; \tfrac{bT}{2},$$

after expanding $(Z_1 + \cdots + Z_T)^2 = \sum_t Z_t^2 + 2\sum_{s<r} Z_s Z_r$. The left-hand side is a polynomial in $(Z_1, \dots, Z_T)$ whose cross-terms $Z_s Z_r$ have coefficient 2, while the right-hand side is a constant. Because $T \ge 2$ the cross-terms are present, so this is a nonzero polynomial that cannot equal a constant on a set of positive Lebesgue measure. Since $(Z_1, \dots, Z_T)$ has a density, the identity fails almost surely, a contradiction.

Case 2: $a \neq 0$. Recall that $f^* = \prod_{t=1}^{T} \frac{1}{\sigma^*}\,\phi(Z_t) = (\sigma^*)^{-T}(2\pi)^{-T/2}\exp\big(-\frac{1}{2}\sum_t Z_t^2\big)$, so $1/f^* = (\sigma^*)^{T}(2\pi)^{T/2}\exp\big(\frac{1}{2}\sum_t Z_t^2\big)$. Rearranging (29) isolates the non-polynomial term:

$$\frac{a\,\sigma^{*2}}{\bar{\alpha}}\,(\sigma^*)^{T}(2\pi)^{T/2}\exp\Big(\tfrac{1}{2}\sum_{t=1}^{T} Z_t^2\Big) \;=\; P(Z_1, \dots, Z_T),$$

where P collects the remaining polynomial terms. Since $a \neq 0$, the left-hand side is a nonzero multiple of $\exp(\frac{1}{2}\sum_t Z_t^2)$, which grows super-polynomially along any ray $\|Z\| \to \infty$. No polynomial can match this growth, so the equality cannot hold on any set of positive Lebesgue measure. Since $(Z_1, \dots, Z_T)$ has a density, this is again a contradiction.

In both cases we reach a contradiction, so no constants a, b satisfy the supposed identity with positive probability, and

$$\Pr\Big[\left.\tfrac{\partial^2 g_2}{(\partial \mu_j)^2}\right|_{(\bar{\alpha},\theta^*,\theta^*)} \;=\; a \;+\; b\left.\tfrac{\partial g_2}{\partial \sigma_j^2}\right|_{(\bar{\alpha},\theta^*,\theta^*)}\Big] \;=\; 0.$$

□

Proof of Proposition 3. Our proof closely follows the proof of Theorem 3.3 in Hathaway (1985) by verifying Assumptions 1, 2, 3, and 5 of Kiefer and Wolfowitz (1956).

We first consider a model with the component-specific density function (1) with ((2))-((3)). Because our model has additional free parameters $\boldsymbol{\beta}_j$s, as in the proof of Kasahara and Shimotsu (2015), we consider the joint density of $m_q := M(q+1)$ observations instead of $M+1$ observations in Hathaway (1985, p. 798), where $q := \dim(\boldsymbol{\beta})$. The joint density function of m_q observations is itself a mixture of M^{m_q} components, where each component is given by $\prod_{j=1}^{m_q}\prod_{t=1}^{T} f(Y_{jt};\mu_{i_j}+\boldsymbol{X}_{jt}^{\top}\boldsymbol{\beta}_{i_j},\sigma_{i_j})$ for some choices $i_j \in \{1,\dots,M\}$, with the density of $N(\mu,\sigma^2)$ denoted by $f(y;\mu,\sigma) := \frac{1}{\sigma}\phi((y-\mu)/\sigma)$.

Assumptions 1 and 2 of Kiefer and Wolfowitz (1956) are easily verified for the joint density of m_q observations. Assumption 3 (strict KL identification) holds because Gaussian mixture densities with distinct mixing distributions are distinct L^1 functions, so the true parameter is the unique maximiser of the expected log-likelihood; see Kasahara and Shimotsu (2015, Lemma A.2) for the analogous argument. This extends to the regression-mixture case: if two parameter vectors generate the same conditional mixture density $g_M(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\vartheta}_M)$ for almost all $\boldsymbol{X}$, then fixing $\boldsymbol{X}=\boldsymbol{x}$ and applying the standard Gaussian-mixture identifiability argument (which uses only the Y-marginal) shows the component means $\mu_j+\boldsymbol{x}^{\top}\boldsymbol{\beta}_j$ and variances must match. Since this must hold for a.e. $\boldsymbol{x}$, varying $\boldsymbol{x}$ across a set of positive measure forces $\boldsymbol{\beta}_j=\tilde{\boldsymbol{\beta}}_j$ and $\mu_j=\tilde{\mu}_j$ for each j, reducing to the covariate-free case. We verify Assumption 5 of Kiefer and Wolfowitz (1956) for the joint density function of m_q observations by showing that

$$E\left[\log\prod_{j=1}^{m_q}\prod_{t=1}^{T} f(Y_{jt};\mu_{i_j}^{*}+\boldsymbol{X}_{jt}^{\top}\boldsymbol{\beta}_{i_j}^{*},\sigma_{i_j}^{*})\right] > -\infty \tag{30}$$

for $\boldsymbol{\vartheta}_M^{*}\in\Theta_M^{*}$ and that

$$E\sup_{\boldsymbol{\vartheta}_M\in\bar{\Theta}_{\boldsymbol{\vartheta}_M}(\boldsymbol{c})}\left[\log\prod_{j=1}^{m_q}\prod_{t=1}^{T} f(Y_{jt};\mu_{i_j}+\boldsymbol{X}_{jt}^{\top}\boldsymbol{\beta}_{i_j},\sigma_{i_j})\right] < \infty \tag{31}$$

for all component choices $i_j\in\{1,\dots,M\}$, which correspond to equations (3.1) and (3.4) in Hathaway (1985), respectively. (30) follows from the argument in the proof of Theorem 3.3 of Hathaway (1985).

For (31), because $\boldsymbol{\vartheta}_M\in\bar{\Theta}_{\boldsymbol{\vartheta}_M}(\boldsymbol{c})$, there exists $c\in(0,1]$ such that $\min_{j,k}\sigma_j/\sigma_k > c$. Proceeding as in Hathaway (1985, pp. 798–799), we can show that

$$\sup_{\boldsymbol{\vartheta}_M\in\bar{\Theta}_{\boldsymbol{\vartheta}_M}(\boldsymbol{c})}\log\left[\prod_{j=1}^{m_q}\prod_{t=1}^{T} f(Y_{jt};\mu_{i_j}+\boldsymbol{X}_{jt}^{\top}\boldsymbol{\beta}_{i_j},\sigma_{i_j})\right]$$

is no larger than, for some $\ell\in\{1,\dots,M\}$ and $j_1,j_2,\dots,j_{q+1}\in\{1,\dots,m_q\}$,

$$\sup_{\mu_\ell,\boldsymbol{\beta}_\ell,\sigma_\ell}\log\left[\delta(\sigma_\ell)\prod_{r=1}^{q+1}\prod_{t=1}^{T} f(Y_{j_r t};\mu_\ell+\boldsymbol{X}_{j_r t}^{\top}\boldsymbol{\beta}_\ell,\sigma_\ell)\right], \tag{32}$$

where $\delta(\sigma_\ell) = (2\pi)^{-T(M-1)(q+1)/2}(c\sigma_\ell)^{-T(M-1)(q+1)}$, because $f(Y_{jt};\mu_{i_j}+\boldsymbol{X}_{jt}^{\top}\boldsymbol{\beta}_{i_j},\sigma_{i_j}) =$

$(2\pi)^{-1/2}(\sigma_{i_j})^{-1}\exp(-\{Y_{jt} - (\mu_{i_j} + \boldsymbol{X}_{jt}^\top\boldsymbol{\beta}_{i_j})\}^2/2\sigma_{i_j}^2) \leq (2\pi)^{-1/2}(\sigma_{i_j})^{-1} \leq (2\pi)^{-1/2}(c\sigma_\ell)^{-1}$ for $j \notin \{j_1, j_2, ..., j_{q+1}\}$.

Note that $\prod_{q=1}^{q+1}\prod_{t=1}^{T} f(Y_{j_qt}; \mu_\ell + \boldsymbol{x}_{j_qt}^\top\boldsymbol{\beta}_\ell, \sigma_\ell)$ is the likelihood function of a linear Gaussian model. Therefore, the maximised value of (32) equals $C_1 - C_2\log(SSR)$, where C_1 and C_2 are a finite constant that depends only on M, k, and T while SSR is the sum of squared residuals obtained from regressing $\{\{Y_{j_qt}\}_{t=1}^T\}_{q=1}^{q+1}$ on $\{\{1, \boldsymbol{x}_{j_qt}\}_{t=1}^T\}_{q=1}^{q+1}$. Because there are $T(q+1)$ observations and $q+1$ regression parameters, the residual degrees of freedom are $(T-1)(q+1)$, which are positive under $T \geq 2$. Thus $SSR > 0$ a.s., and the finite moment conditions in Assumption 3(a) ensure $\mathbb{E}|\log(SSR)| < \infty$. Therefore, the expected value of (32) is finite, and (31) holds. This verifies Assumption 5 of Kiefer and Wolfowitz (1956), and the stated consistency result follows.

When the component-specific density function is given by (1) with (2)–(3), where the number of sub-components in (3) is K, the joint density function of m_q observations can be written as a mixture of $M^{m_q} \times K^{m_qT}$ components, where each component is given by $\prod_{j=1}^{m_q}\prod_{t=1}^{T} f(Y_{jt}; \mu_{i_ji_k} + \boldsymbol{X}_{jt}^\top\boldsymbol{\beta}_{i_j}, \sigma_{i_j})$ for some choices $i_j \in \{1, \ldots, M\}$ and $i_k \in \{1, ..., K\}$. Then, repeating the same argument as above, we can show that the expected value of $\sup_{\boldsymbol{\vartheta}_M \in \Theta_{\boldsymbol{\vartheta}_M}(\boldsymbol{c})} \log\left[\prod_{j=1}^{m_q}\prod_{t=1}^{T} f(Y_{jt}; \mu_{i_ji_k} + \boldsymbol{X}_{jt}^\top\boldsymbol{\beta}_{i_j}, \sigma_{i_j})\right]$ is bounded, and Assumption 5 of Kiefer and Wolfowitz (1956) holds. Because Assumptions 1, 2, and 3 of Kiefer and Wolfowitz (1956) are also easily verified, the consistency result follows.

□

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

Given the value of $\boldsymbol{c}$ and $\alpha \in [c_1, 1-c_1]$, define the space for reparameterized parameters as $\boldsymbol{\psi} := (\boldsymbol{\nu}^\top, \boldsymbol{\lambda}^\top)^\top \in \Theta_{\boldsymbol{\Psi}_\alpha}$, where $\Theta_{\psi_\alpha} = \{\boldsymbol{\psi} = (\boldsymbol{\nu}^\top, \boldsymbol{\lambda}^\top)^\top : (\alpha, \boldsymbol{\nu} + (1-\alpha)\boldsymbol{\lambda}, \boldsymbol{\nu} - \alpha\boldsymbol{\lambda}) \in \Theta_{\boldsymbol{\vartheta}_2}(\boldsymbol{c})\}$. Under the null hypothesis $H_{01} : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}^*$, we have $\boldsymbol{\lambda} = \boldsymbol{0}$ and $\boldsymbol{\nu} = \boldsymbol{\theta}^*$. We rewrite the reparameterized parameters under the null hypothesis as $\boldsymbol{\psi}^* = ((\boldsymbol{\theta}^*)^\top, \boldsymbol{0}^\top)^\top$. We denote the reparameterized density function and its logarithm as

$$g(\boldsymbol{w}; \boldsymbol{\psi}, \alpha) := \alpha f(\boldsymbol{w}; \boldsymbol{\nu} + (1-\alpha)\boldsymbol{\lambda}) + (1-\alpha) f(\boldsymbol{w}; \boldsymbol{\nu} - \alpha\boldsymbol{\lambda}) \text{ and } l(\boldsymbol{w}; \boldsymbol{\psi}, \alpha) = \log g(\boldsymbol{w}; \boldsymbol{\psi}, \alpha). \quad (33)$$

Let $L_n(\boldsymbol{\psi}, \alpha) := \sum_{i=1}^n l(\boldsymbol{W}_i; \boldsymbol{\psi}, \alpha)$ be the reparameterized log-likelihood function. For each $\alpha \in [c_1, 1-c_1]$, define the reparameterized MLE as

$$\hat{\boldsymbol{\psi}}_\alpha = \arg\max_{\boldsymbol{\psi} \in \Theta_{\boldsymbol{\Psi}_\alpha}} L_n(\boldsymbol{\psi}, \alpha). \quad (34)$$

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

$$\boldsymbol{t}(\boldsymbol{\psi}, \alpha) = \begin{pmatrix} \boldsymbol{t}_{\boldsymbol{\nu}} \\ \boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}, \alpha) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\nu} - \boldsymbol{\nu}^* \\ \alpha(1-\alpha)\boldsymbol{v}(\boldsymbol{\lambda}) \end{pmatrix}, \quad (35)$$

where $v(\boldsymbol{\lambda})$ is given by (23).

As discussed in the proof of Lemma 1, taking the fourth-order Taylor expansion of $L_n(\boldsymbol{\psi}, \alpha)$ around $(\boldsymbol{\Psi}^*, \alpha)$, we may write $2\{L_n(\boldsymbol{\psi}, \alpha) - L_n(\boldsymbol{\psi}^*, \alpha)\}$ as a quadratic function of $\sqrt{n}\boldsymbol{t}(\boldsymbol{\psi}, \alpha)$ as

$$\begin{aligned} 2\{L_n(\boldsymbol{\psi}, \alpha) - L_n(\boldsymbol{\psi}^*, \alpha)\} &= 2(\sqrt{n}\boldsymbol{t}(\boldsymbol{\psi}, \alpha))^\top \boldsymbol{S}_n - (\sqrt{n}\boldsymbol{t}(\boldsymbol{\psi}, \alpha))^\top \boldsymbol{\mathcal{I}}_n(\sqrt{n}\boldsymbol{t}(\boldsymbol{\psi}, \alpha)) + R_n(\boldsymbol{\psi}, \alpha) \quad (36) \\ &= \boldsymbol{G}_n^\top \boldsymbol{\mathcal{I}}_n \boldsymbol{G}_n - \left[\sqrt{n}\boldsymbol{t}(\boldsymbol{\psi}, \alpha) - \boldsymbol{G}_n\right]^\top \boldsymbol{\mathcal{I}}_n \left[\sqrt{n}\boldsymbol{t}(\boldsymbol{\psi}, \alpha) - \boldsymbol{G}_n\right] + R_n(\boldsymbol{\psi}, \alpha), \quad (37) \end{aligned}$$

where $\boldsymbol{S}_n := n^{-1/2}\sum_{i=1}^n \boldsymbol{s}(\boldsymbol{W}_i)$ and $\boldsymbol{G}_n := \boldsymbol{\mathcal{I}}_n^{-1}\boldsymbol{S}_n$, where $\boldsymbol{\mathcal{I}}_n$ is the negative of the sample Hessian defined in the proof of Lemma 1.

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

$$LR_n = \max_{\alpha\in[c_1,1-c_1]} 2\{L_n(\hat{\boldsymbol{\psi}}_\alpha,\alpha) - L_n(\boldsymbol{\psi}^*,\alpha)\} - 2\{L_{0,n}(\hat{\boldsymbol{\gamma}}_0,\hat{\boldsymbol{\theta}}_0) - L_{0,n}(\boldsymbol{\gamma}_0^*,\boldsymbol{\theta}_0^*)\}. \tag{38}$$

Define

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

and split the quadratic form in (36) to obtain

$$2\{L_n(\boldsymbol{\psi},\alpha) - L_n(\boldsymbol{\psi}^*,\alpha)\} = B_n(\sqrt{n}\boldsymbol{t}_{\nu,\lambda}) + C_n(\sqrt{n}\boldsymbol{t}_\lambda(\boldsymbol{\lambda},\alpha)) + R_n(\boldsymbol{\Psi},\alpha), \tag{39}$$

where

$$\begin{aligned} B_n(\boldsymbol{t}_{\nu,\lambda}) &= 2\boldsymbol{t}_{\nu,\lambda}^\top \boldsymbol{S}_{\nu n} - \boldsymbol{t}_{\nu,\lambda}^\top \boldsymbol{\mathcal{I}}_\nu \boldsymbol{t}_{\nu,\lambda},\\ C_n(\boldsymbol{t}_\lambda) &= 2\boldsymbol{t}_\lambda^\top \boldsymbol{S}_{\lambda,\nu n} - \boldsymbol{t}_\lambda^\top \boldsymbol{\mathcal{I}}_{\lambda,\nu}\boldsymbol{t}_\lambda\\ &= \boldsymbol{G}_{\lambda,\nu n}^\top \boldsymbol{\mathcal{I}}_{\lambda,\nu}\boldsymbol{G}_{\lambda,\nu n} - (\boldsymbol{t}_\lambda - \boldsymbol{G}_{\lambda,\nu n})^\top \boldsymbol{\mathcal{I}}_{\lambda,\nu}(\boldsymbol{t}_\lambda - \boldsymbol{G}_{\lambda,\nu n}), \end{aligned} \tag{40}$$

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

Because $\Delta_{(\gamma,\theta)} f(x;\hat{\boldsymbol{\gamma}}_0^*,\hat{\boldsymbol{\theta}}_0^*)$ is identical to $\Delta_{\boldsymbol{\nu}} f(x;\boldsymbol{\Psi}^*,\alpha)$, a standard analysis gives $2[L_{0,n}(\hat{\boldsymbol{\gamma}}_0,\hat{\boldsymbol{\theta}}_0) - L_{0,n}(\boldsymbol{\gamma}_0^*,\boldsymbol{\theta}_0^*)] = \max_{\boldsymbol{t}_\nu} B_n(\sqrt{n}\boldsymbol{t}_\nu) + o_p(1)$. Note that the possible values of both $\sqrt{n}\boldsymbol{t}_\nu$ and $\sqrt{n}\boldsymbol{t}_{\nu,\lambda}$ approach $\mathbb{R}^q$. Therefore, in view of (39) and (40), we can write equation (38) as

$$LR_n = \max_{\alpha\in[c_1,1-c_1]} C_n(\sqrt{n}\boldsymbol{t}_\lambda(\hat{\boldsymbol{\lambda}}_\alpha,\alpha)) + o_p(1), \tag{41}$$

where $\hat{\boldsymbol{\lambda}}_\alpha$ is as defined in (34).

The asymptotic distribution of LR_n follows from applying Theorem 3(c) of (Andrews, 1999, p. 1362) to (39) and (41). First, Assumption 2 of Andrews (1999) holds because Assumption 2* of Andrews (1999) holds by Lemma 1(a) (the remainder bound) together with Proposition 2. Second, Assumption 3 of Andrews (1999) holds with $B_T = n^{1/2}$ and $T = n$ because $\boldsymbol{S}_{\lambda,\nu n} \xrightarrow{d} \boldsymbol{S}_{\lambda,\nu} \sim N(\boldsymbol{0},\boldsymbol{\mathcal{I}}_{\lambda,\nu})$ by Lemma 1(b) and $\boldsymbol{\mathcal{I}}_{\lambda,\nu}$ is non-singular by Lemma 1(c) (for $K_\epsilon = 1$; for $K_\epsilon = 2$, non-singularity follows from Lemma 2(c) instead). 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 Λ_λ. To confirm that Lemma 3 of Andrews (1999) applies, note that Λ_λ is a closed convex cone: Θ_λ is a simplex constraint (finitely many linear inequalities), so its tangent cone at any point is a polyhedral cone (intersection of finitely many halfspaces); scaling by the positive scalar $\alpha(1-\alpha)$ preserves closedness and convexity. The projection onto Λ_λ therefore exists and is unique. Assumption 6 of Andrews (1999) (uniform equicontinuity of the score in a shrinking neighbourhood) holds because, by Assumption 4(a), the log-likelihood function has uniformly bounded second derivatives in a neighbourhood of $(\boldsymbol{\Psi}^*,\alpha)$; a standard mean-value argument then gives the required modulus-of-continuity bound. The second-order tangent cone condition required for Theorem 3(c) holds because Λ_λ is a polyhedral cone (the nonnegative orthant, arising from the simplex constraints on the mixing proportions); for polyhedral cones the second-order tangent cone at any point equals the tangent cone itself, so the condition is automatically satisfied for any $q \geq 1$. Therefore, it follows from Theorem 3(c) of Andrews (1999) that $C_n(\sqrt{n}\boldsymbol{t}_\lambda(\hat{\boldsymbol{\lambda}}, \alpha)) \xrightarrow{d} (\hat{\boldsymbol{t}}_\lambda)^\top \boldsymbol{\mathcal{I}}_{\lambda,\nu}\hat{\boldsymbol{t}}_\lambda$, where $\hat{\boldsymbol{t}}_\lambda$ is defined by (24).

□

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

$$\begin{aligned}
&\mathbb{E}[\{\nabla_{\alpha_h} \log f_{M_0+1}(\boldsymbol{W}_i, \vartheta_{M_0+1})\}^2] \\
&= \int \frac{\{f(\boldsymbol{w};\boldsymbol{\theta}_h) - f(\boldsymbol{w};\boldsymbol{\theta}^*_{M_0})\}^2}{\sum_{j=1}^{M_0} \alpha_j^* f(\boldsymbol{w};\boldsymbol{\theta}_j^*)} d\boldsymbol{w} \\
&= \int \frac{\{f(\boldsymbol{w};\boldsymbol{\theta}_h)\}^2}{\sum_{j=1}^{M_0} \alpha_j^* f(\boldsymbol{w};\boldsymbol{\theta}_j^*)} d\boldsymbol{w} + \int \frac{\{f(\boldsymbol{w};\boldsymbol{\theta}^*_{M_0})\}^2}{\sum_{j=1}^{M_0} \alpha_j^* f(\boldsymbol{w};\boldsymbol{\theta}_j^*)} d\boldsymbol{w} - 2\int \frac{f(\boldsymbol{w};\boldsymbol{\theta}_h) f(\boldsymbol{w};\boldsymbol{\theta}^*_{M_0})}{\sum_{j=1}^{M_0} \alpha_j^* f(\boldsymbol{w};\boldsymbol{\theta}_j^*)} d\boldsymbol{w}.
\end{aligned} \tag{42}$$

The latter two terms on the right-hand side of (42) are bounded because $f(\boldsymbol{w};\boldsymbol{\theta}^*_{M_0})/\sum_{j=1}^{M_0} \alpha_j^* f(\boldsymbol{w};\boldsymbol{\theta}_j^*) \leq (1/\alpha^*_{M_0})$ for any $\boldsymbol{w}$ and $f(\boldsymbol{w};\boldsymbol{\theta})$ integrates to one. Therefore, the left-hand side of (42) goes to infinity if and only if the first term on the right-hand side of (42) goes to infinity.

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

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

Without loss of generality, we assume that $\sigma^*_{M_0} = \max\{\sigma^*_1, \ldots, \sigma^*_{M_0}\}$ and that the maximum is unique. We focus on models without covariates because the law of iterated expectations implies that the stated result also holds for models with covariates if it holds for models without covariates.

We first prove the case for the component-specific density function $f(\boldsymbol{w};\boldsymbol{\theta}) = \prod_{t=1}^T f(y_t;\boldsymbol{\theta})$, where $f(y_t;\boldsymbol{\theta})$ is given by ((2))–((3)) with $\boldsymbol{\beta}_j = \boldsymbol{0}$. Under the normal density function ((2))–((3)) without covariates, there exists a sufficiently large but finite positive constant B, such that $\max_j\{\alpha_j^* f(\boldsymbol{w}, \mu_j^*, \sigma_j^2)\} = \alpha^*_{M_0} f(\boldsymbol{w}, \mu^*_{M_0}, \sigma^2_{M_0})$ when $|y_t| > B$ for all $t = 1, \ldots, T$. Note that

$$\begin{aligned}
\frac{\{f(\boldsymbol{w};\mu_h,\sigma_h)\}^2}{f(\boldsymbol{w};\mu^*_{M_0},\sigma^*_{M_0})} &= \prod_{t=1}^T \frac{\sigma^*_{M_0}}{(2\pi)^{1/2}\sigma_h^2} \exp\left\{-\frac{1}{\sigma_h^2}(y_t-\mu_h)^2 + \frac{1}{2(\sigma^*_{M_0})^2}(y_t-\mu^*_{M_0})^2\right\} \\
&= \left(\frac{\sigma^*_{M_0}}{(2\pi)^{1/2}\sigma_h^2}\right)^T \exp\left\{-\frac{1}{\sigma_h^2}\sum_{t=1}^T(y_t-\mu_h)^2 + \frac{1}{2(\sigma^*_{M_0})^2}\sum_{t=1}^T(y_t-\mu^*_{M_0})^2\right\}.
\end{aligned} \tag{43}$$

Then, the integral of the right-hand side of (43) over $|y_t| \geq B$ for $t = 1, ..., T$ is infinite if $\sigma_h^2/\sigma^{2*}_{M_0} > 2$, and the stated result holds. Conversely, when $\sigma_h^2 < 2\sigma^{*2}_{M_0}$, the leading coefficient of y_t^2 in the exponent is strictly negative, so the integrand is bounded by a product of Gaussian densities and the integral is finite. At the boundary $\sigma_h^2 = 2\sigma^{*2}_{M_0}$, the y_t^2 coefficient in the exponent is identically zero; the integrand then decays at most polynomially in $|y_t|$ (the linear terms $(\mu_h - \mu^*)y_t/\sigma^{*2}$

dominate), so the integral over $|y_t| \geq B$ remains infinite and the stated result holds in this case as well.

When $f(y_t;\boldsymbol{\theta})$ is given by normal mixture density (2)–(3), suppose that $\mu_{j1} < \mu_{j2} < \cdots < \mu_{K_\epsilon}$ for all $j = 1, ..., M_0$. Then, there exists a constant B such that, when $y_t > B$ for all $t = 1, \ldots, T$, we have $\max_j\{\alpha_j^* f(\boldsymbol{w}, \boldsymbol{\theta}_j^*)\} = \alpha_{M_0}^* f(\boldsymbol{w}, \boldsymbol{\theta}_{M_0}^*)$ and

$$\frac{1}{\sigma_j}\exp\left(-\frac{1}{2}\left(\frac{y_t-\mu_{K_\epsilon}}{\sigma_j}\right)^2\right) \geq \sum_{k=1}^{K_\epsilon}\tau_{jk}\frac{1}{\sigma_j}\exp\left(-\frac{1}{2}\left(\frac{y_t-\mu_{jk}}{\sigma_j}\right)^2\right) \geq \frac{1}{\sigma_j}\exp\left(-\frac{1}{2}\left(\frac{y_t-\mu_{j1}}{\sigma_j}\right)^2\right). \tag{44}$$

Then, it follows that

$$\frac{\{f(\boldsymbol{w};\boldsymbol{\theta}_h)\}^2}{f(\boldsymbol{w};\boldsymbol{\theta}_{M_0}^*)} \geq \left(\frac{\sigma_{M_0}^*}{(2\pi)^{1/2}\sigma_h^2}\right)^T \exp\left\{-\frac{1}{\sigma_h^2}\sum_{t=1}^T (y_t-\mu_{1h})^2 + \frac{1}{2(\sigma_{M_0}^*)^2}\sum_{t=1}^T (y_t-\mu_{M_0K}^*)^2\right\},$$

where the integral of the right-hand side over $y_t > B$ for $t = 1, ..., T$ is infinite if $\sigma_h^2/\sigma_{M_0}^{2*} > 2$ as in (43).

□

Proof of Proposition 6. Collect the score vector for testing $H_{0,1h}$ for $h = 1, \ldots, M_0$ into one vector as

$$\tilde{s}(\boldsymbol{W}) = \begin{pmatrix}\tilde{s}_\eta(\boldsymbol{W})\\ \tilde{s}_{\lambda\lambda}(\boldsymbol{W})\end{pmatrix}, \text{ where } \underset{(M_0+p+q+1)\times 1}{\tilde{s}_\eta(\boldsymbol{W})} = \begin{pmatrix}\boldsymbol{s}_{\boldsymbol{\alpha}}(\boldsymbol{W})\\ \boldsymbol{s}_{(\nu)}(\boldsymbol{W})\end{pmatrix} \text{ and } \tilde{s}_{\lambda\lambda}(\boldsymbol{W}) = \begin{pmatrix}\boldsymbol{s}_{\lambda\lambda}^1(\boldsymbol{W})\\ \vdots \\ \boldsymbol{s}_{\lambda\lambda}^{M_0}(\boldsymbol{W})\end{pmatrix}, \tag{45}$$

where

$$\begin{aligned}\boldsymbol{s}_{\boldsymbol{\alpha}}(\boldsymbol{W}) &= \begin{pmatrix} f(\boldsymbol{W};\boldsymbol{\theta}_1^*) - f(\boldsymbol{W};\boldsymbol{\theta}_{M_0}^*)\\ \vdots \\ f(\boldsymbol{W};\boldsymbol{\theta}_{M_0-1}^*) - f(\boldsymbol{W};\boldsymbol{\theta}_{M_0}^*)\end{pmatrix}\Bigg/ f_{M_0}(\boldsymbol{W};\boldsymbol{\vartheta}_{M_0}^*),\\ \boldsymbol{s}_{\boldsymbol{\nu}}(\boldsymbol{W}) &= \sum_{j=1}^{M_0}\alpha_j^*\frac{\nabla_{\boldsymbol{\nu}} f(\boldsymbol{W};\boldsymbol{\theta}_j^*)}{f_{M_0}(\boldsymbol{W};\boldsymbol{\vartheta}_{M_0}^*)}, \text{ and } \boldsymbol{s}_{\lambda\lambda}^h(\boldsymbol{W}) = \frac{\widetilde{\nabla}_{\boldsymbol{\theta}_h\boldsymbol{\theta}_h^\top} f(\boldsymbol{W};\boldsymbol{\theta}_h^*)}{f_{M_0}(\boldsymbol{W};\boldsymbol{\vartheta}_{M_0}^*)} \quad \text{for } h = 1, 2, ..., M_0,\end{aligned} \tag{46}$$

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

$$\begin{aligned}\tilde{\mathcal{I}} &:= \mathbb{E}[\tilde{s}(\boldsymbol{W})\tilde{s}(\boldsymbol{W})^\top], \quad \tilde{\mathcal{I}}_\eta := \mathbb{E}[\tilde{s}_\eta(\boldsymbol{W})\tilde{s}_\eta(\boldsymbol{W})^\top], \quad \tilde{\mathcal{I}}_{\lambda\eta} := \mathbb{E}[\tilde{s}_{\lambda\lambda}(\boldsymbol{W})\tilde{s}_\eta(\boldsymbol{W})^\top],\\ \tilde{\mathcal{I}}_{\eta\lambda} &:= \tilde{\mathcal{I}}_{\lambda\eta}^\top, \quad \tilde{\mathcal{I}}_{\lambda\lambda} := \mathbb{E}[\tilde{s}_{\lambda\lambda}(\boldsymbol{W})\tilde{s}_{\lambda\lambda}(\boldsymbol{W})^\top], \quad \tilde{\mathcal{I}}_{\lambda,\eta} := \tilde{\mathcal{I}}_{\lambda\lambda} - \tilde{\mathcal{I}}_{\lambda\eta}\tilde{\mathcal{I}}_\eta^{-1}\tilde{\mathcal{I}}_{\eta\lambda}.\end{aligned} \tag{47}$$

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

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

where, in view of (45), $\tilde{S}$ may be partitioned as $\tilde{S} = (\tilde{S}_{\eta}^{\top}, \tilde{S}_{\lambda\lambda}^{\top})^{\top}$ 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 $\tilde{S}_{\lambda,\eta} := (S_{\lambda,\eta}^{1}, \ldots, S_{\lambda,\eta}^{M_0})^{\top} := \tilde{S}_{\lambda\lambda} - \tilde{\mathcal{I}}_{\lambda\eta}\tilde{\mathcal{I}}_{\eta}^{-1}\tilde{S}_{\eta} \sim N(0, \tilde{\mathcal{I}}_{\lambda,\eta})$ be a $\mathbb{R}^{M_0(q)(q+1)/2}$-valued random vector. For $h = 1, 2, ..., M_0$, define $\tilde{\mathcal{I}}_{\lambda,\eta}^{h} := \mathbb{E}[S_{\lambda,\eta}^{h}(S_{\lambda,\eta}^{h})^{\top}]$ and $G_{\lambda,\eta}^{h} := (\mathcal{I}_{\lambda,\eta}^{h})^{-1}S_{\lambda,\eta}^{h}$.

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

$$r_{\lambda}^{h}(\hat{t}_{\lambda}^{h}) = \inf_{t_{\lambda}^{h}\in\Lambda_{\lambda}} r^{h}(t_{\lambda}^{h}); \quad r_{\lambda}^{h}(t_{\lambda}^{h}) := (t_{\lambda}^{h} - G_{\lambda,\eta}^{h})^{\top}\mathcal{I}_{\lambda,\eta}^{h}(t_{\lambda}^{h} - G_{\lambda,\eta}^{h}) \quad \text{for } h = 1, 2, ..., M_0. \tag{48}$$

The local quadratic-form approximation of the log-likelihood function $LR_{n}^{M_0,h}$ around $\Theta_{\vartheta_{M_0+1},1h}^{*} \subset \Theta_{\vartheta_{M_0+1}}$ has an identical structure to the approximation that we derive in Section 5.1 in testing H_{01} in the test of homogeneity. Specifically, the bounded ratios $\alpha_j^* f_j^*/g_{M_0}^*$ ensure that Lemma 1 conditions transfer to the M_0-component case, and the Hermite polynomial structure of the score is preserved under the reparameterisation (λ_h, ν_h). Consequently, we can show that $LR_{n}^{M_0,h} \xrightarrow{d} (\hat{t}_{\lambda}^{h})^{\top}\mathcal{I}_{\lambda,\eta}^{h}\hat{t}_{\lambda}^{h}$. Then, given (25), the asymptotic null distribution of the LRTS for testing H_{01} is given by the maximum over $(\hat{t}_{\lambda}^{h})^{\top}\mathcal{I}_{\lambda,\eta}^{h}\hat{t}_{\lambda}^{h}$s for $h = 1, 2, ..., M_0$. We now proceed with the proof.

For $h = 1, \ldots, M_0$, let $\mathcal{N}_h^* \subset \Theta_{\vartheta_{M_0+1}}(c)$ be a sufficiently small closed neighborhood of $\Theta_{\vartheta_{M_0+1},1h}^{*}$ such that $\alpha_h, \alpha_{h+1} > 0$ holds and $\Theta_{\vartheta_{M_0+1},1k}^{*} \not\subset \mathcal{N}_h^*$ if $k \neq h$. Consider the following one-to-one reparameterization from the (M_0+1)-component model parameter $\vartheta_{M_0+1} = (\alpha_1, \ldots, \alpha_{M_0}, \theta_1^{\top}, \ldots, \theta_h^{\top}, \theta_{h+1}^{\top}, \ldots, \theta_{M_0+1}^{\top})^{\top}$. Similar to (21), the one-to-one reparameterization for testing the null hypothesis $H_{0,1h}$ is given by

$$\begin{pmatrix}\lambda_h\\ \nu_h\end{pmatrix} := \begin{pmatrix}\theta_h - \theta_{h+1}\\ \tau\theta_h + (1-\tau)\theta_{h+1}\end{pmatrix} \text{ so that } \begin{pmatrix}\theta_h\\ \theta_{h+1}\end{pmatrix} = \begin{pmatrix}\nu + (1-\tau)\lambda\\ \nu - \tau\lambda\end{pmatrix},$$

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

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

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

Collect the reparameterized parameters except τ as

$$\psi_{h,\tau} = (\eta^{\top}, \lambda_h^{\top})^{\top} \quad \text{with} \quad \eta = (\pi_1, \ldots, \pi_{M_0-1}, \theta_1^{\top}, \ldots, \theta_{h-1}^{\top}, \nu_h^{\top}, \theta_{h+2}^{\top}, \ldots, \theta_{M_0+1}^{\top})^{\top}.$$

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 $\eta^* = (\alpha_1^*, ..., \alpha_{M_0-1}^*, (\theta_1^*)^{\top}, \ldots, (\theta_{M_0}^*)^{\top})^{\top}$. Define the log-likelihood under the reparameterized parameters as

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

where $g^h(\boldsymbol{w}, \boldsymbol{\psi}_{h,\tau}, \tau)$ is defined similarly to (33) as

$$g^h(\boldsymbol{w}, \boldsymbol{\psi}_{h,\tau}, \tau) = \tau f(\boldsymbol{w}; \boldsymbol{\nu}_h + (1-\tau)\boldsymbol{\lambda}_h) + (1-\tau) f(\boldsymbol{w}; \boldsymbol{\nu}_h - \tau\boldsymbol{\lambda}_h). \tag{49}$$

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

$$\hat{\boldsymbol{\psi}}_{h,\tau} := \arg\max_{\boldsymbol{\psi}_{h,\tau} \in \mathcal{N}_h^*} L_n^h(\boldsymbol{\psi}_{h,\tau}, \tau), \tag{50}$$

where $L_n^h(\boldsymbol{\psi}_{h,\tau}, \tau) := \sum_{i=1}^n \log f_{M_0+1}^h(\boldsymbol{W}_i; \boldsymbol{\psi}_{h,\tau}, \tau)$. Because $\boldsymbol{\psi}_{h,\tau}^*$ is the only parameter value in $\mathcal{N}_h^*$ that generates the true density, $\hat{\boldsymbol{\psi}}_{h,\tau} - \boldsymbol{\psi}_{h,\tau}^* = o_p(1)$ follows.

Define the LRTS for testing $H_{0,1h}$ as $LR_n^{M_0,h} := \max_{\tau \in [c_1, 1-c_1]} 2(L_n^h(\hat{\boldsymbol{\psi}}_{h,\tau}, \tau) - L_{0,n}(\hat{\boldsymbol{\vartheta}}_{M_0}))$. Then, in view of (25), the stated result holds if

$$(LR_n^{M_0,1}, \ldots, LR_n^{M_0,M_0})^\top \xrightarrow{d} (\hat{\boldsymbol{t}}_\lambda^1)^\top \boldsymbol{\mathcal{I}}_{\eta,\lambda}^1 (\hat{\boldsymbol{t}}_\lambda^1), \ldots, (\hat{\boldsymbol{t}}_\lambda^{M_0})^\top \boldsymbol{\mathcal{I}}_{\eta,\lambda}^{M_0} (\hat{\boldsymbol{t}}_\lambda^{M_0}))^\top. \tag{51}$$

Observe that $L_n^h(\boldsymbol{\psi}_{h,\tau}, \tau) - L_n^h(\boldsymbol{\psi}_{h,\tau}^*, \tau)$ admits the same expansion as $L_n(\hat{\boldsymbol{\psi}}, \alpha) - L_n(\boldsymbol{\psi}^*, \alpha)$ in (37) and (39) when $(\alpha, \boldsymbol{t}(\boldsymbol{\psi}, \alpha), \boldsymbol{t}_\lambda(\boldsymbol{\lambda}, \alpha), \boldsymbol{S}_n, \boldsymbol{G}_n, \boldsymbol{\mathcal{I}}_n, R_n(\boldsymbol{\Psi}, \alpha))$ is replaced with $(\tau, \boldsymbol{t}^h(\boldsymbol{\psi}^h, \tau), \boldsymbol{t}_\lambda^h(\boldsymbol{\lambda}^h, \tau), \boldsymbol{S}_n^h, \boldsymbol{G}_n^h, \boldsymbol{\mathcal{I}}_n^h, R_n^h(\boldsymbol{\Psi}^h, \tau))$, where $(\boldsymbol{S}_n^h, \boldsymbol{\mathcal{I}}_n^h)$ is defined similarly to $(\boldsymbol{S}_n, \boldsymbol{\mathcal{I}}_n)$ but $(\boldsymbol{s}_\eta, \boldsymbol{s}_{\lambda\lambda})$ is replaced with $(\tilde{\boldsymbol{s}}_\eta, \boldsymbol{s}_{\lambda\lambda}^h)$ and $\boldsymbol{G}_n^h := (\boldsymbol{\mathcal{I}}_n^h)^{-1}\boldsymbol{S}_n^h$. Applying the proof of Lemma 1, we have $\boldsymbol{S}_n^h \xrightarrow{d} \boldsymbol{S}^h \sim N(\boldsymbol{0}, \boldsymbol{\mathcal{I}}^h)$ and $\boldsymbol{\mathcal{I}}_n^h \xrightarrow{p} \boldsymbol{\mathcal{I}}^h$. Then, (51) follows from the proof of Lemma 1 and Proposition 4 for each local MLE when $(\boldsymbol{G}_n, \hat{\boldsymbol{t}}_\lambda, \boldsymbol{\mathcal{I}}_{\lambda,\eta})$ is replaced with $(\boldsymbol{G}_n^h, \hat{\boldsymbol{t}}_\lambda^h, \boldsymbol{\mathcal{I}}_{\lambda,\eta}^h)$ and the results are collected. Joint convergence $(\boldsymbol{S}_n^1, \ldots, \boldsymbol{S}_n^{M_0}) \xrightarrow{d} (\boldsymbol{S}^1, \ldots, \boldsymbol{S}^{M_0})$ holds by the multivariate CLT applied to the stacked score vector $(\boldsymbol{s}^1(\boldsymbol{W}_i)^\top, \ldots, \boldsymbol{s}^{M_0}(\boldsymbol{W}_i)^\top)^\top$, which has finite second moments under Assumption 5(a). The full covariance matrix of the limit, including off-diagonal blocks $\mathrm{Cov}(\boldsymbol{S}^h, \boldsymbol{S}^{h'}) = E[\boldsymbol{s}^h(\boldsymbol{W})\boldsymbol{s}^{h'}(\boldsymbol{W})^\top]$ for $h \neq h'$, need not be zero; the stated limit in Proposition 6 is the maximum of M_0 jointly Gaussian quadratic forms whose dependence structure is fully determined by this covariance matrix. Applying the continuous mapping theorem to the jointly Gaussian limit $(\boldsymbol{S}^1, \ldots, \boldsymbol{S}^{M_0})$, together with the continuous mapping $(\boldsymbol{S}_n^h, \boldsymbol{\mathcal{I}}_n^h) \mapsto (\hat{\boldsymbol{t}}_\lambda^h)^\top \boldsymbol{\mathcal{I}}_{\lambda,\nu}^h \hat{\boldsymbol{t}}_\lambda^h$ (continuous by the local MLE projection argument in the proof of Proposition 4), yields joint convergence of $(LR_n^{M_0,1}, \ldots, LR_n^{M_0,M_0})$. A second application of the continuous mapping theorem with the continuous max functional then gives (51). The bootstrap in Proposition 7 preserves this joint distribution consistently, so the critical-value approximation remains valid regardless of the cross-covariance structure. $\square$

Proof of Proposition 8. We first prove that when $M < M_0$, $\Pr(LR_n^M > \hat{c}_{1-q_n}^M) \to 1$ as $n \to \infty$. Under Assumption 6, by the standard extremum estimator consistency argument (e.g., Theorem 2.1 of Newey and McFadden, 1994) gives $\hat{\boldsymbol{\vartheta}}_M \xrightarrow{p} \boldsymbol{\vartheta}_M^*$ for $M \leq M_0$, where $\boldsymbol{\vartheta}_M^*$ is the pseudo-true parameter that minimizes the Kullback–Leibler divergence within the M-component model. Because $\boldsymbol{\vartheta}_M^*$ lies in the interior of $\bar{\Theta}_{\boldsymbol{\vartheta}_M}(\boldsymbol{c})$ (the mixing proportions are bounded away from 0 and 1, and the variance ratios are bounded below), it follows from Theorem 3.2 of White (1982) that, for $M \leq M_0$,

$$\sqrt{n}(\hat{\boldsymbol{\vartheta}}_M - \boldsymbol{\vartheta}_M^*) \xrightarrow{d} N(0, A^M(\boldsymbol{\vartheta}_M^*)^{-1} B^M(\boldsymbol{\vartheta}_M^*) A^M(\boldsymbol{\vartheta}_M^*)^{-1}). \tag{52}$$

Then, from (52) and the mean value expansion — noting that $M+1 \le M_0$ for $M = 1, \dots, M_0 - 1$, so Assumption 6 and the asymptotic-normality result (52) apply to both the M-component and the $(M+1)$-component estimators — we have $Q_n^M(\hat{\vartheta}_M) - Q^M(\vartheta_M^*) = O_p(n^{-1/2})$ and

$$\frac{LR_n^M}{n} := 2\left(Q_n^{M+1}(\hat{\vartheta}_{M+1}) - Q_n^M(\hat{\vartheta}_M)\right) = 2\left(Q^{M+1}(\vartheta_{M+1}^*) - Q^M(\vartheta_M^*)\right) + o_p(1)$$

for $M = 1, 2, ..., M_0 - 1$.

Because $Q^{M+1}(\vartheta_{M+1}^*) - Q^M(\vartheta_M^*) > 0$ by Assumption 6(f), LR_n^M/n converges to the positive constant $2(Q^{M+1}(\vartheta_{M+1}^*) - Q^M(\vartheta_M^*))$ as $n \to \infty$. It remains to show $n^{-1}\hat{c}_{1-q_n}^M = o_p(1)$. Since the bootstrap LRT statistic $LR_n^{*,M}$ converges in distribution to the same chi-bar-squared limit F_M as the theoretical statistic (Proposition 7(a)–(b)), applying the Chernoff bound in the proof of Lemma 5 to the bootstrap distribution gives $n^{-1}\hat{c}_{1-q_n}^M \xrightarrow{p} 0$ directly, without appealing to quantile consistency. Therefore, when $M < M_0$, we have $\Pr(LR_n^M > \hat{c}_{1-q_n}^M) = \Pr(LR_n^M/n > \hat{c}_{1-q_n}^M/n) \to 1$ as $n \to \infty$.

When $M = M_0$, because $LR_n^{M_0} = O_p(1)$ by Proposition 6 and $\hat{c}_{1-q_n}^M \to \infty$ by $q_n = o(1)$, $\Pr(LR_n^{M_0} > \hat{c}_{1-q_n}^{M_0}) \to 0$ as $n \to \infty$. □

Proof of Proposition A.5. We first prove that $\lim_{n\to\infty} \Pr(\widehat{M}_{PL} < M_0) = 0$. To show $\Pr(\hat{M}_{PL} < M_0) \to 0$, it suffices to prove that for each $M < M_0$,

$$\Pr\big(p\ell_n^M > p\ell_n^{M_0}\big) \;\to\; 0.$$

By definition,

$$\frac{p\ell_n^M - p\ell_n^{M_0}}{n} = (Q_n^M(\hat{\vartheta}_M) - Q_n^{M_0}(\hat{\vartheta}_{M_0})) - \frac{p_{n,k_M} - p_{n,k_{M_0}}}{n}. \tag{53}$$

From $p_{n,k} = o(n)$ in Assumption A.3(c),

$$\frac{p_{n,k_M} - p_{n,k_{M_0}}}{n} \to 0. \tag{54}$$

By the union bound and the triangle inequality, for any $\epsilon > 0$,

$$\Pr\Big(\Big|Q_n^M(\hat{\vartheta}_M) - Q_n^{M_0}(\hat{\vartheta}_{M_0}) - \Big(Q^M(\vartheta_M^*) - Q^{M_0}(\vartheta_{M_0}^*)\Big)\Big| > 4\epsilon\Big)$$

is at most

$$\Pr\Big(\big|Q^M(\hat{\vartheta}_M) - Q^M(\vartheta_M^*)\big| > \epsilon\Big) + \Pr\Big(\Big|Q^{M_0}(\hat{\vartheta}_{M_0}) - Q^{M_0}(\vartheta_{M_0}^*)\Big| > \epsilon\Big) \tag{55}$$

$$+ \Pr\Big(\big|Q_n^M(\hat{\vartheta}_M) - Q^M(\hat{\vartheta}_M)\big| > \epsilon\Big) + \Pr\Big(\Big|Q_n^{M_0}(\hat{\vartheta}_{M_0}) - Q^{M_0}(\hat{\vartheta}_{M_0})\Big| > \epsilon\Big). \tag{56}$$

The consistency of $\hat{\vartheta}_M$ and $\hat{\vartheta}_{M_0}$, together with the Continuous Mapping Theorem, implies that the two terms in (55) converge to 0 as $n \to \infty$.

For small $\delta > 0$, let $B(\delta, \vartheta_M^*)$ and $B(\delta, \vartheta_{M_0}^*)$ be open balls in the parameter spaces Θ_{ϑ_M} and $\Theta_{\vartheta_{M_0}}$, respectively. Since $\hat{\vartheta}_M$ and $\hat{\vartheta}_{M_0}$ are consistent, with probability approaching one, the two terms in (56) are bounded by

$$\Pr\Big(\sup_{\boldsymbol{\vartheta}_M\in B(\delta,\boldsymbol{\vartheta}^*_M)}\big|Q_n^M(\boldsymbol{\vartheta}_M)-Q^M(\boldsymbol{\vartheta}_M)\big|>\epsilon\Big)\quad\text{and}\quad\Pr\Big(\sup_{\boldsymbol{\vartheta}_{M_0}\in B(\delta,\boldsymbol{\vartheta}^*_{M_0})}\Big|Q_n^{M_0}(\boldsymbol{\vartheta}_{M_0})-Q^{M_0}(\boldsymbol{\vartheta}_{M_0})\Big|>\epsilon\Big).$$

By applying Lemma 2.4 of Newey and McFadden (1994) under Assumption 6(b), and noting that $\log g_M(\boldsymbol{w};\boldsymbol{\vartheta}_M)$ is continuous at each $\boldsymbol{\vartheta}_M$ and the envelope function $\sup_{\boldsymbol{\vartheta}_M\in B(\delta,\boldsymbol{\vartheta}^*_M)}|\log g_M(\boldsymbol{w};\boldsymbol{\vartheta}_M)|$ has finite expectation for $M\le M_0$, these two probabilities converge to 0 by the Uniform Law of Large Numbers. Thus, we obtain

$$Q_n^M(\hat{\boldsymbol{\vartheta}}_M)-Q_n^{M_0}(\hat{\boldsymbol{\vartheta}}_{M_0})\xrightarrow{p}Q^M(\theta^*_M)-Q^{M_0}(\theta^*_{M_0}).\tag{57}$$

Therefore, from (53), (54), and (57), and in view of Assumptions 6(f) and A.3(c), we have

$$\frac{p\ell_n^M-p\ell_n^{M_0}}{n}\xrightarrow{p}Q^M(\theta^*_M)-Q^{M_0}(\theta^*_{M_0})<0.$$

Hence, $\Pr(p\ell_n^M>p\ell_n^{M_0})\to 0$ for $M<M_0$, which proves that

$$\lim_{n\to\infty}\Pr(\hat{M}_{PL}<M_0)=0.\tag{58}$$

We proceed to prove that $\lim_{n\to\infty}\Pr(\widehat{M}_{PL}>M_0)=0$ by showing that, for $M=M_0+1,\ldots,\bar{M}$,

$$\Pr(p\ell_n^M>p\ell_n^{M_0})\to 0\quad\text{as } n\to\infty.$$

By definition of $p\ell_n^M$, we have

$$\Pr(p\ell_n^M>p\ell_n^{M_0})=\Pr\left(\frac{\ell_n^M(\hat{\boldsymbol{\vartheta}}_M)-\ell_n^{M_0}(\hat{\boldsymbol{\vartheta}}_{M_0})}{p_{n,k_{M_0}}}+1-\frac{p_{n,k_M}}{p_{n,k_{M_0}}}>0\right).$$

For any $\epsilon>0$, $\lim_{n\to\infty}\Pr\left(\frac{\ell_n^M(\hat{\boldsymbol{\vartheta}}_M)-\ell_n^{M_0}(\hat{\boldsymbol{\vartheta}}_{M_0})}{p_{n,k_{M_0}}}>\epsilon\right)=0$ because $p_{n,k_{M_0}}\to\infty$ and $\ell_n^M(\hat{\boldsymbol{\vartheta}}_M)-\ell_n^{M_0}(\hat{\boldsymbol{\vartheta}}_{M_0})=O_p(1)$ by Assumptions A.3(b) and A.4, respectively. Therefore,

$$\frac{\ell_n^M(\hat{\boldsymbol{\vartheta}}_M)-\ell_n^{M_0}(\hat{\boldsymbol{\vartheta}}_{M_0})}{p_{n,k_{M_0}}}+1-\frac{p_{n,k_M}}{p_{n,k_{M_0}}}\xrightarrow{p}1-\lim_{n\to\infty}\frac{p_{n,k_M}}{p_{n,k_{M_0}}},$$

which is strictly negative by Assumption A.3(d). Thus, $\Pr(p\ell_n^M>p\ell_n^{M_0})\to 0$ as $n\to\infty$, and $\lim_{n\to\infty}\Pr(\widehat{M}_{PL}>M_0)=0$ follows.

□

Proof of Proposition A.6. Because $\ell_n^{M_0}(\hat{\boldsymbol{\vartheta}}_{M_0})-\ell_n^{M_0}(\boldsymbol{\vartheta}^*_{M_0})=O_p(1)$ under standard regularity conditions in Assumption 6, Assumption A.4 holds when $\ell_n^M(\hat{\boldsymbol{\vartheta}}_M)-\ell_n^{M_0}(\boldsymbol{\vartheta}^*_{M_0})=O_p(1)$ in view of $\ell_n^M(\hat{\boldsymbol{\vartheta}}_M)-\ell_n^{M_0}(\hat{\boldsymbol{\vartheta}}_{M_0})=(\ell_n^M(\hat{\boldsymbol{\vartheta}}_M)-\ell_n^{M_0}(\boldsymbol{\vartheta}^*_{M_0}))-(\ell_n^{M_0}(\hat{\boldsymbol{\vartheta}}_{M_0})-\ell_n^{M_0}(\boldsymbol{\vartheta}^*_{M_0}))$. Because a consistent MLE is in $A_{\varepsilon_n}(\eta)$ defined in Appendix A.2, Propositions 3 and A.3 imply that $\ell_n(\hat{\boldsymbol{\Psi}}^j_M,\boldsymbol{\tau}^j)-\ell_n(\boldsymbol{\Psi}^{j*}_M,\boldsymbol{\tau}^j)=O_p(1)$, and the stated result holds. □

Proof of Corollary 1. It is straightforward to verify that the penalty function $p_{n,k} = \frac{k}{2}\log(n)$ satisfies Assumption A.3. Thus, result (a) directly follows from Proposition A.6. For (b), Assumption A.3(c) holds for AIC penalty. Then, repeating the argument in the proof of Proposition A.5 up to (58) proves that $p\ell_n^{M_0}(\hat{\boldsymbol{\vartheta}}_{M_0}) > \max\left\{p\ell_n^1(\widehat{\boldsymbol{\vartheta}}_1), ..., p\ell_n^{M_0-1}(\widehat{\boldsymbol{\vartheta}}_{M_0-1})\right\}$ holds with probability approaching one as $n \to \infty$. This implies that

$$\begin{aligned}\lim_{n\to\infty}\Pr\left(\widehat{M}_{PL} > M_0\right) &\geq \lim_{n\to\infty}\Pr\left(p\ell_n^{M_0+1}(\hat{\boldsymbol{\vartheta}}_{M_0+1}) > \max\left\{p\ell_n^1(\widehat{\boldsymbol{\vartheta}}_1), ..., p\ell_n^{M_0}(\widehat{\boldsymbol{\vartheta}}_{M_0})\right\}\right)\\ &= \lim_{n\to\infty}\Pr\left(p\ell_n^{M_0+1}(\hat{\boldsymbol{\vartheta}}_{M_0+1}) > p\ell_n^{M_0}(\widehat{\boldsymbol{\vartheta}}_{M_0})\right)\\ &= \lim_{n\to\infty}\Pr\left(2\left[\ell_n^{M_0+1}(\hat{\boldsymbol{\vartheta}}_{M_0+1}) - \ell_n^{M_0}(\widehat{\boldsymbol{\vartheta}}_{M_0})\right] > 2(k_{M_0+1} - k_{M_0})\right), \end{aligned} \tag{59}$$

where k_{M_0} and k_{M_0+1} are the number of parameters for M_0 and (M_0+1) components model. Since the term $2\left[\ell_n^{M_0+1}(\hat{\boldsymbol{\vartheta}}_{M_0+1}) - \ell_n^{M_0}(\widehat{\boldsymbol{\vartheta}}_{M_0})\right]$ is $O_p(1)$ and converges in distribution as specified in Proposition 6, and because k_{M_0} and k_{M_0+1} are finite, it follows from (59) that $\lim_{n\to\infty}\Pr\left(\widehat{M}_{PL} > M_0\right) > 0$, and part (b) follows. □

Proof of Proposition A.1. Part (a) can be proven by the same argument as Proposition 3. The only additional requirement is that Assumptions 1, 2, 3, and 5 of Kiefer and Wolfowitz (1956) hold on the restricted parameter space $\tilde{\Theta}_{\boldsymbol{\vartheta}_2}(\boldsymbol{c})$ (which imposes $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \geq c_1$). This is straightforward: $\tilde{\Theta}_{\boldsymbol{\vartheta}_2}(\boldsymbol{c})$ is closed and bounded (compactness is preserved by a lower-bound restriction on a compact set); the envelope condition (Assumption 5/H3.2) depends only on the density form and not on the separation constraint, so it carries over unchanged from Proposition 3. The remaining detail is thus omitted.

The proof for part (b) closely follows Theorem 2(b) in Andrews (2001). For each $\boldsymbol{\lambda} \in \Theta_{\lambda}(c_1)$, we approximate the log-likelihood function $\ell_n^2(\boldsymbol{\theta}_2, \boldsymbol{\lambda}, \alpha)$ around the true parameter values $(\boldsymbol{\theta}^*, 0)$ using the partial derivative with respect to $\boldsymbol{\theta}_2$ and the right partial derivative with respect to α to obtain

$$\begin{aligned}2\{\ell_n^2(\boldsymbol{\theta}_2, \boldsymbol{\lambda}, \alpha) - \ell_n^2(\boldsymbol{\theta}^*, \boldsymbol{\lambda}, 0)\} &= \boldsymbol{G}_n(\boldsymbol{\lambda})^\top \boldsymbol{\mathcal{I}}_n(\boldsymbol{\lambda})\boldsymbol{G}_n(\boldsymbol{\lambda})\\ &\quad - \left[\sqrt{n}\boldsymbol{t}(\boldsymbol{\theta}_2, \alpha) - \boldsymbol{G}_n(\boldsymbol{\lambda})\right]^\top \boldsymbol{\mathcal{I}}_n\left[\sqrt{n}\boldsymbol{t}(\boldsymbol{\theta}_2, \alpha) - \boldsymbol{G}_n(\boldsymbol{\lambda})\right] + R_n(\boldsymbol{\theta}_2, \boldsymbol{\lambda}, \alpha), \end{aligned} \tag{60}$$

where $R_n(\boldsymbol{\theta}_2, \boldsymbol{\lambda}, \alpha)$ is a remainder term, and $\boldsymbol{\mathcal{I}}_n(\boldsymbol{\lambda})$, $\boldsymbol{G}_n(\boldsymbol{\lambda})$, and $\boldsymbol{t}(\boldsymbol{\theta}_2, \alpha)$ are defined as

$$\boldsymbol{\mathcal{I}}_n(\boldsymbol{\lambda}) := n^{-1}\sum_{i=1}^n \boldsymbol{s}(\boldsymbol{W}_i; \boldsymbol{\lambda})\boldsymbol{s}(\boldsymbol{W}_i; \boldsymbol{\lambda})^\top, \ \boldsymbol{G}_n(\boldsymbol{\lambda}) := \boldsymbol{\mathcal{I}}_n(\boldsymbol{\lambda})^{-1}\boldsymbol{S}_n(\boldsymbol{\lambda}), \text{ and } \boldsymbol{t}(\boldsymbol{\theta}_2, \alpha) := \begin{pmatrix}\boldsymbol{\theta}_2 - \boldsymbol{\theta}^* \\ \alpha\end{pmatrix}$$

with $\boldsymbol{S}_n(\boldsymbol{\lambda}) = n^{-1/2}\sum_{i=1}^n \boldsymbol{s}(\boldsymbol{W}_i; \boldsymbol{\lambda})$ and $\boldsymbol{s}(\boldsymbol{W}_i; \boldsymbol{\lambda})$ defined in (2). (θ, π) and $(B_T, D\ell_T(\theta_0, \pi), \mathcal{J}_{T\pi}, Z_{T\pi})$ in Andrews (2001) correspond to our $((\boldsymbol{\theta}_2^\top, \alpha)^\top, \boldsymbol{\lambda})$ and $(n^{1/2}, n^{1/2}\boldsymbol{S}_n(\boldsymbol{\lambda}), \boldsymbol{\mathcal{I}}_n(\boldsymbol{\lambda}), \boldsymbol{G}_n(\boldsymbol{\lambda}))$.

We prove the stated result by applying Theorem 2(b) of Andrews (2001) to (60). (β, δ, π) and $(B_T, G_\pi, \mathcal{J}_\pi, Z_\pi, Z_{\beta\pi})$ in Andrews (2001, pp. 697-699) correspond to our $(\alpha, \boldsymbol{\theta}_2, \boldsymbol{\lambda})$ and $(n^{1/2}, \boldsymbol{S}(\boldsymbol{\lambda}), \boldsymbol{\mathcal{I}}(\boldsymbol{\lambda}), \boldsymbol{G}(\boldsymbol{\lambda}), \boldsymbol{G}_{\alpha,\boldsymbol{\theta}_2}(\boldsymbol{\lambda}))$, where $\boldsymbol{G}(\boldsymbol{\lambda}) := \boldsymbol{\mathcal{I}}(\boldsymbol{\lambda})^{-1}\boldsymbol{S}(\boldsymbol{\lambda})$, $\boldsymbol{G}_{\alpha,\boldsymbol{\theta}_2}(\boldsymbol{\lambda}) := \boldsymbol{\mathcal{I}}_{\alpha,\boldsymbol{\theta}_2}(\boldsymbol{\lambda})^{-1}\boldsymbol{S}_{\alpha,\boldsymbol{\theta}_2}(\boldsymbol{\lambda})$, and ψ in Andrews (2001, pp. 697-699) does not exist in our setting. The stated result then follows because $\boldsymbol{s}_{\boldsymbol{\theta}_2}(\boldsymbol{w})$ is identical to the score of the one-component model and $\hat{\lambda}_{\beta\pi}'(H\mathcal{J}_{*\pi}^{-1}H')^{-1}\hat{\lambda}_{\beta\pi}$ in Theorem 2(b) of Andrews (2001) is distributed as $(\max\{0, \boldsymbol{\mathcal{I}}_{\alpha,\boldsymbol{\theta}_2}(\boldsymbol{\lambda})^{-1/2}\boldsymbol{S}_{\alpha,\boldsymbol{\theta}_2}(\boldsymbol{\lambda})\})^2$. We proceed to verify the assumptions of Theorem 2(b) of Andrews (2001) (hereafter, A-Assumptions 2^{2^*}, 3-5, 7, and 8). A-Assumption 2^{2^*}(a) and (b) follow from our Assumption A.1(a)(b). A-Assumption 2^{2^*}(c) holds because our Assumption A.1(c) and (d) and the uniform law of large numbers imply that $\sup_{\lambda\in\Theta_\lambda(c_1)}||\mathcal{I}_n(\lambda)-\mathcal{I}(\lambda)||\rightarrow_p 0$ and $\mathcal{I}(\lambda)$ is continuous.

A-Assumption 3 follows from Theorem 10.2 of Pollard (1990) if (i) $\widetilde{\Theta}_\lambda$ is totally bounded, (ii) the finite dimensional distributions of $\boldsymbol{G}_n(\cdot)$ converge to those of $\boldsymbol{G}(\cdot)$, and (iii) $\{\boldsymbol{G}_n(\cdot):n\geq 1\}$ is stochastically equicontinuous. Condition (i) follows from the compactness of $\widetilde{\Theta}_\lambda$ while conditions (ii) follow from Assumption A.1(b) and (c) and the multivariate CLT. Condition (iii) can be verified by our Assumption A.1(b) and (c) and Theorem 2 of Andrews (1994) because $\nabla_{(\boldsymbol{\theta}_2^\top,\alpha)^\top}\log g_2(\cdot;\boldsymbol{\theta}_2,\boldsymbol{\lambda},\alpha)$ are Lipschitz functions indexed by a finite dimensional parameter $\boldsymbol{\lambda}$ by Assumption A.1(b).

A-Assumption 4 follows from Lemma 1 of Andrews (2001) because, for each $\boldsymbol{\lambda}\in\widetilde{\Theta}_\lambda(c_1)$, $(\tilde{\boldsymbol{\theta}}_2(\boldsymbol{\lambda}),\tilde{\alpha}(\boldsymbol{\lambda}))=\arg\max_{(\boldsymbol{\theta}_2,\alpha)\in\Theta_{\boldsymbol{\theta}_2}\times[0,1/2]}\ell_n^2(\boldsymbol{\theta}_2,\boldsymbol{\lambda},\alpha)$ converges to $(\boldsymbol{\theta}_2^*,0)$ in probability from the standard consistency proof. A-Assumption 5 holds because (i) the set $[0,1/2]$ equals a nonnegative half-line locally around 0, and (ii) $\Theta_{\boldsymbol{\theta}_2}-\boldsymbol{\theta}_2^*$ is locally equal to $\mathbb{R}^{\dim(\boldsymbol{\theta}_2)}$. A-Assumption 7(a) is not relevant for our problem. A-Assumptions 7(b) and 8 follow from our proof of A-Assumption 5.

□

The following two proofs rely on Lemma 1 stated in the Lemmas subsection below.

Proof of Proposition A.2. For notational brevity, we drop the superscript j from $\boldsymbol{\Psi}_M^j$, $\boldsymbol{\tau}^j$, $\boldsymbol{s}_i^j$ and $\mathcal{I}^j$. By using the Taylor expansion of $2\log(1+x)=2x-x^2(1+o(1))$ for small x, we have uniformly $\boldsymbol{\Psi}_M\in\mathcal{N}_{c/\sqrt{n}}$,

$$\ell_n(\boldsymbol{\Psi}_M,\boldsymbol{\tau})-\ell_n(\boldsymbol{\Psi}_M^*,\boldsymbol{\tau})=2\sum_{k=1}^n\log(1+h_{\boldsymbol{\Psi}_M\tau,i})=nP_n(2h_{\boldsymbol{\Psi}_M\tau,i}-[1+o_p(1)]h^2_{\boldsymbol{\Psi}_M\tau,i}),\tag{61}$$

where $h_{\boldsymbol{\Psi}_M\tau,i}:=\sqrt{l_{\boldsymbol{\Psi}_M\tau,i}}-1$. The stated result holds if we show that

$$\sup_{\boldsymbol{\Psi}_M\in\mathcal{N}_{c/\sqrt{n}}}\left|nP_n(h^2_{\boldsymbol{\Psi}_M\tau,i})-n\boldsymbol{t}^\top_{\boldsymbol{\Psi}_M\tau}\mathcal{I}\boldsymbol{t}_{\boldsymbol{\Psi}_M\tau}/4\right|=o_p(1)\quad\text{and}\tag{62}$$

$$\sup_{\boldsymbol{\Psi}_M\in\mathcal{N}_{c/\sqrt{n}}}|nP_n(h_{\boldsymbol{\Psi}_M\tau,i})-\sqrt{n}\boldsymbol{t}^\top_{\boldsymbol{\Psi}_M\tau}\nu_n(\boldsymbol{s}_i)/2+n\boldsymbol{t}_{\boldsymbol{\Psi}_M\tau}\mathcal{I}\boldsymbol{t}^\top_{\boldsymbol{\Psi}_M\tau}/8|=o_p(1),\tag{63}$$

because then the right-hand side of (61) is equal to $\sqrt{n}\boldsymbol{t}^\top_{\boldsymbol{\Psi}_M\tau}\nu_n(\boldsymbol{s}_i)-\boldsymbol{t}_{\boldsymbol{\Psi}_M\tau}\mathcal{I}\boldsymbol{t}^\top_{\boldsymbol{\Psi}_M\tau}/2+o_p(1)$ uniformly in $\boldsymbol{\Psi}_M\in\mathcal{N}_{c/\sqrt{n}}$.

We first show (62). Let

$$m_{\boldsymbol{\Psi}_M\tau i}:=l_{\boldsymbol{\Psi}_M\tau,i}-1=\boldsymbol{t}^\top_{\boldsymbol{\Psi}_M\tau}\boldsymbol{s}_i+r_{\boldsymbol{\Psi}_M\tau,i}.$$

Observe that

$$\max_{1\leq i\leq n}\sup_{\boldsymbol{\Psi}_M\in\mathcal{N}_{c/\sqrt{n}}}|m_{\boldsymbol{\Psi}_M\tau i}|=\max_{1\leq i\leq n}\sup_{\boldsymbol{\Psi}_M\in\mathcal{N}_{c/\sqrt{n}}}|\boldsymbol{t}^\top_{\boldsymbol{\Psi}_M\tau}\boldsymbol{s}_i+r_{\boldsymbol{\Psi}_M\tau,i}|=o_p(1),\tag{64}$$

from Assumptions A.2(a) and (c) and Lemma 7. Write $4P_n(h^2_{\Psi_M\tau,i})$ as

$$4P_n(h^2_{\Psi_M\tau,i}) = P_n\left(\frac{4(l_{\Psi_M\tau,i}-1)^2}{(\sqrt{l_{\Psi_M\tau,i}}+1)^2}\right) = P_n(l_{\Psi_M\tau,i}-1)^2 - P_n\left((l_{\Psi_M\tau,i}-1)^3\frac{(\sqrt{l_{\Psi_M\tau,i}}+3)}{(\sqrt{l_{\Psi_M\tau,i}}+1)^3}\right). \tag{65}$$

From (14),

$$P_n(l_{\Psi_M\tau,i}-1)^2 = \boldsymbol{t}^\top_{\Psi_M\tau}P_n(\boldsymbol{s}_i\boldsymbol{s}_i')\boldsymbol{t}_{\Psi_M\tau} + \zeta_{\Psi_M\tau n}, \tag{66}$$

where $\zeta_{\Psi_M\tau n} := 2\boldsymbol{t}^\top_{\Psi_M\tau}P_n[\boldsymbol{s}_i r_{\Psi_M\tau,i}] + P_n(r_{\Psi_M\tau,i})^2$. It follows from Assumptions A.2(a)(b)(c) and $(E|XY|)^2 \le E|X|^2E|Y|^2$ that, uniformly in $\Psi_M \in \mathcal{N}_\varepsilon$,

$$|\zeta_{\Psi_M\tau n}| = O_p(||\boldsymbol{t}_{\Psi_M\tau}||^2||\Psi_M-\Psi^*_M||) + O_p(n^{-1}||\boldsymbol{t}_{\Psi_M\tau}||||\Psi_M-\Psi^*_M||) + O_p(n^{-1}||\Psi_M-\Psi^*_M||^2). \tag{67}$$

Then, (62) holds because $||P_n(\boldsymbol{s}_i\boldsymbol{s}_i') - \boldsymbol{\mathcal{I}}|| = o_p(1)$ and the second term on the right-hand side of (65) is bounded by

$$C\sup_{\Psi_M\in\mathcal{N}_{c/\sqrt{n}}} P_n\left[|m_{\Psi_M\tau i}|^3\right] \le o_p(1)\sup_{\Psi_M\in\mathcal{N}_{c/\sqrt{n}}} P_n\left[m^2_{\Psi_M\tau i}\right] = o_p(n^{-1}),$$

from $\sup_{\Psi_M\in\mathcal{N}_{c/\sqrt{n}}}\frac{(\sqrt{l_{\Psi_M\tau,i}}+3)}{(\sqrt{l_{\Psi_M\tau,i}}+1)^3} \le C$, (64), and $P_n(m^2_{\Psi_M\tau i}) = \boldsymbol{t}^\top_{\Psi_M\tau}\boldsymbol{\mathcal{I}}\boldsymbol{t}_{\Psi_M\tau} + o_p(||\boldsymbol{t}_{\Psi_M\tau}||^2)$.

We proceed to show (63). Consider the following expansion of $h_{\Psi_M\tau,i}$:

$$h_{\Psi_M\tau,i} = (l_{\Psi_M\tau,i}-1)/2 - h^2_{\Psi_M\tau,i}/2 = (\boldsymbol{t}^\top_{\Psi_M\tau}\boldsymbol{s}_i + r_{\Psi_M\tau,i})/2 - h^2_{\Psi_M\tau,i}/2. \tag{68}$$

Then, (63) follows from (62), (68), and Assumptions A.2(d), and the stated result follows. □

Proof of Proposition A.3. For brevity, we drop the superscript M from $\boldsymbol{s}_i^M$ and $\boldsymbol{\mathcal{I}}^M$. Define $h_{\Psi_M\alpha,i} := \sqrt{l_{\Psi_M\alpha,i}}-1$. For part (a), it follows from $\log(1+x)\le x$ and $h_{\Psi_M\alpha,i} = (l_{\Psi_M\alpha,i}-1)/2 - h^2_{\Psi_M\alpha,i}/2$ (see (68)) that

$$\ell_n(\Psi_M,\boldsymbol{\alpha}) - \ell_n(\Psi^*_M,\boldsymbol{\alpha}) = 2\sum_{k=1}^n\log(1+h_{\Psi_M\alpha,i}) \le 2nP_n(h_{\Psi_M\alpha,i}) = \sqrt{n}\nu_n(l_{\Psi_M\alpha,i}-1) - nP_n(h^2_{\Psi_M\alpha,i}). \tag{69}$$

Observe that $h^2_{\Psi_M\alpha,i} = (l_{\Psi_M\alpha,i}-1)^2/(\sqrt{l_{\Psi_M\alpha,i}}+1)^2 \ge \mathbb{I}\{l_{\Psi_M\alpha,i}\le\kappa\}(l_{\Psi_M\alpha,i}-1)^2/(\sqrt{\kappa}+1)^2$ for any $\kappa>0$. Therefore,

$$P_n(h^2_{\Psi_M\alpha,i}) \ge (\sqrt{\kappa}+1)^{-2}P_n\left(\mathbb{I}\{l_{\Psi_M\alpha,i}\le\kappa\}(l_{\Psi_M\alpha,i}-1)^2\right). \tag{70}$$

In view of (66), (70) is rewritten as

$$P_n(h^2_{\Psi_M\alpha,i}) \ge (\sqrt{\kappa}+1)^{-2}\left\{\boldsymbol{t}^\top_{\Psi_M\alpha}\left[P_n(\boldsymbol{s}_i\boldsymbol{s}_i') - P_n(\mathbb{I}\{l_{\Psi_M\alpha,i}>\kappa\}\boldsymbol{s}_i\boldsymbol{s}_i')\right]\boldsymbol{t}_{\Psi_M\alpha} + \tilde{\zeta}_{\Psi_M\alpha n}\right\}, \tag{71}$$

where $\tilde{\zeta}_{\Psi_M\alpha n}$ satisfies the bound similar to (67). From Hölder's inequality, we have $P_n(\mathbb{I}\{l_{\Psi_M\alpha,i}>\kappa\}||\boldsymbol{s}_i||^2) \le [P_n(\mathbb{I}\{l_{\Psi_M\alpha,i}>\kappa\})]^{\delta/(2+\delta)}[P_n(||\boldsymbol{s}_i||^{2+\delta})]^{2/(2+\delta)}$, where the right-hand side is no larger than $\kappa^{-\delta/(2+\delta)}O_p(1)$ uniformly in $\Psi_M\in\mathcal{N}_{\varepsilon_n}$ because (i) it follows from $\kappa\mathbb{I}\{l_{\Psi_M\alpha,i}>\kappa\}\le l_{\Psi_M\alpha,i}$ that $P_n(\mathbb{I}\{l_{\Psi_M\alpha,i}>\kappa\}) \le \kappa^{-1}P_n(l_{\Psi_M\alpha,i})$ and $\sup_{\Psi_M\in\mathcal{N}_{\varepsilon_n}}|P_n(l_{\Psi_M\alpha,i})-1| = o_p(1)$ from Assumptions A.2(d)(e), and (ii) $P_n(||\boldsymbol{s}_i||^{2+\delta}) = O_p(1)$ from Assumption A.2(a). Consequently, $\mathbb{P}(\sup_{\boldsymbol{\Psi}_M\in\mathcal{N}_{\varepsilon_n}} P_n(\mathbb{I}\{l_{\boldsymbol{\Psi}_M\boldsymbol{\alpha},i} > \kappa\}||\boldsymbol{s}_i||^2) \geq \lambda_{\min}/4) \to 0$ as $\kappa \to \infty$, where $\lambda_{\min}$ is the smallest eigenvalue of $\boldsymbol{\mathcal{I}}$. Hence, we can write (71) as $P_n(h^2_{\boldsymbol{\Psi}_M\boldsymbol{\alpha},i}) \geq \frac{\tau}{2}(1+o_p(1))\boldsymbol{t}^\top_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}}\boldsymbol{\mathcal{I}}\boldsymbol{t}_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}} + \frac{\tau}{2}\tilde{\zeta}_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}n}$ for $\tau := 2(\sqrt{\kappa}+1)^{-2} > 0$ by taking κ sufficiently large. Because $\sqrt{n}\nu_n(l_{\boldsymbol{\Psi}_M\boldsymbol{\alpha},i}-1) = \sqrt{n}\boldsymbol{t}^\top_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}}\nu_n(\boldsymbol{s}_i) + O_p(\sqrt{n}||\boldsymbol{t}_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}}||||\boldsymbol{\Psi}_M - \boldsymbol{\Psi}^*_M||)$ from Assumptions A.2(d) and because $\tilde{\zeta}_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}n}$ satisfies the bound analogous to (67), it follows from (69) that, uniformly in $\boldsymbol{\Psi}_M \in \mathcal{N}_{\varepsilon_n}$,

$$
\begin{aligned}
-\eta &\leq \ell_n(\boldsymbol{\Psi}_M,\boldsymbol{\alpha}) - \ell_n(\boldsymbol{\Psi}^*_M,\boldsymbol{\alpha}) \\
&\leq \sqrt{n}\boldsymbol{t}^\top_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}}\nu_n(\boldsymbol{s}_i) - \frac{\tau}{2}(1+o_p(1))n\boldsymbol{t}^\top_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}}\boldsymbol{\mathcal{I}}\boldsymbol{t}_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}} + O_p(\sqrt{n}||\boldsymbol{t}_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}}||||\boldsymbol{\Psi}_M - \boldsymbol{\Psi}^*_M||) \\
&\quad + O_p(n||\boldsymbol{t}_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}}||^2||\boldsymbol{\Psi}_M - \boldsymbol{\Psi}^*_M||) + O_p(||\boldsymbol{t}_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}}||||\boldsymbol{\Psi}_M - \boldsymbol{\Psi}^*_M||) + O_p(||\boldsymbol{\Psi}_M - \boldsymbol{\Psi}^*_M||^2). \qquad (72)
\end{aligned}
$$

Let $\boldsymbol{T}_n := \boldsymbol{\mathcal{I}}^{1/2}\sqrt{n}\boldsymbol{t}_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}}$. From (72), Assumptions A.2(b) and (e), and the fact $||\boldsymbol{\Psi}_M - \boldsymbol{\Psi}^*_M|| \to 0$ if $\boldsymbol{t}_{\boldsymbol{\Psi}_M\boldsymbol{\alpha}} \to 0$, we obtain

$$
-\eta \leq \ell_n(\boldsymbol{\Psi}_M,\boldsymbol{\alpha}) - \ell_n(\boldsymbol{\Psi}^*_M,\boldsymbol{\alpha}) \leq ||\boldsymbol{T}_n||O_p(1) - \frac{\tau}{2}||T_n||^2 + o_p(||T_n||^2) + o_p(||T_n||) + o_p(1),
$$

uniformly in $\boldsymbol{\Psi}_M \in \mathcal{N}_{\varepsilon_n}$. By rearranging the inequality and choosing a sufficiently large constant C, we have $||\boldsymbol{T}_n||C - (\tau/2)||\boldsymbol{T}_n||^2 + C \geq 0$ with an arbitrarily high probability. Namely, for any $\delta > 0$, there exist $C, n_0 < \infty$ such that

$$
\mathbb{P}\left(\inf_{\boldsymbol{\Psi}_M\in\mathcal{N}_{\varepsilon_n}} ||\boldsymbol{T}_n||C - (\tau/2)||\boldsymbol{T}_n||^2 + C \geq 0\right) \geq 1-\delta, \quad \text{for all } n > n_0. \qquad (73)
$$

Rearranging the terms inside $\mathbb{P}(\cdot)$ gives $\sup_{\boldsymbol{\Psi}_M\in\mathcal{N}_{\varepsilon_n}}(||\boldsymbol{T}_n|| - (C/\tau))^2 \leq 2C/\tau + (C/\tau)^2$. Taking its square root gives $\mathbb{P}(\sup_{\boldsymbol{\Psi}_M\in\mathcal{N}_{\varepsilon_n}} ||\boldsymbol{T}_n|| \leq C_1) \geq 1-\delta$ for a constant C_1, and part (a) follows. Part (b) follows from part (a) and Proposition A.2. □

A.6 Proof of Bootstrap Validity

The key technical ingredient is the following lemma, which shows that the null distribution of the LRT is invariant under local perturbations of the null parameter. This is the panel analogue of Lemma 12 in the supplementary appendix of Kasahara and Shimotsu (2019).

Lemma 1 (Stability of the LRT under local null perturbations)**.** *Let $C_{\boldsymbol{\nu}}$ denote the set of sequences $\{\boldsymbol{\nu}_n\}_{n\geq 1}$ satisfying $\sqrt{n}(\boldsymbol{\nu}_n - \boldsymbol{\nu}^*) \to \boldsymbol{h}_{\boldsymbol{\nu}}$ for some finite $\boldsymbol{h}_{\boldsymbol{\nu}} \in \mathbb{R}^{d_{\boldsymbol{\nu}}}$. For each such sequence, let $P^n_{\boldsymbol{\nu}_n}$ denote the product probability measure under which $\boldsymbol{W}_1,\ldots,\boldsymbol{W}_n$ are i.i.d. from the M_0-component density $g_{M_0}(\cdot\,;\boldsymbol{\nu}_n)$.*

(a) *Suppose Assumptions 1 and 4 hold and $M_0 = 1$. For every $\{\boldsymbol{\nu}_n\} \in C_{\boldsymbol{\nu}}$, the LRT statistic LR_n under $P^n_{\boldsymbol{\nu}_n}$ converges in distribution to the same limit as under $P^n_{\boldsymbol{\nu}^*}$, namely $\widehat{\boldsymbol{t}}^\top_{\boldsymbol{\lambda}}\boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\nu}}\widehat{\boldsymbol{t}}_{\boldsymbol{\lambda}}$ (Proposition 4).*

(b) *Suppose Assumptions 1, 3, and 5 hold and $M_0 = M \geq 2$. For every $\{\boldsymbol{\nu}_n\} \in C_{\boldsymbol{\nu}}$, the LRT statistic $LR^{M_0}_n$ under $P^n_{\boldsymbol{\nu}_n}$ converges in distribution to the same limit as under $P^n_{\boldsymbol{\nu}^*}$, namely $\max_{h=1,\ldots,M_0}\{(\widehat{\boldsymbol{t}}^h_{\boldsymbol{\lambda}})^\top\boldsymbol{\mathcal{I}}^h_{\boldsymbol{\lambda},\boldsymbol{\nu}}\widehat{\boldsymbol{t}}^h_{\boldsymbol{\lambda}}\}$ (Proposition 6).*

Proof of Lemma 1. We prove part (a); part (b) follows by the same argument applied to each local LRT statistic $LR^{M_0,h}_n$, $h = 1,\ldots,M_0$, in the decomposition (25).

Set $\boldsymbol{h} = (\boldsymbol{h}_{\boldsymbol{\nu}}^{\top}, \mathbf{0}^{\top})^{\top}$. The proof proceeds in four steps.

Step 1 (LAN and contiguity of the null model). Under $P_{\boldsymbol{\nu}^*}^n$, the correctly specified M_0-component model is a regular parametric family: Assumption 6(c) places $\boldsymbol{\vartheta}_{M_0}^*$ in the interior of $\Theta_{\boldsymbol{\vartheta}_{M_0}}$, (d) ensures non-singularity of the information matrix $\boldsymbol{\mathcal{I}}_{\boldsymbol{\nu}} = B^{M_0}(\boldsymbol{\vartheta}_{M_0}^*)$, and (e) provides smooth local identification. These conditions, together with the finite moment requirements in Assumption 4(a), yield the standard LAN expansion

$$\log \frac{dP_{\boldsymbol{\nu}_n}^n}{dP_{\boldsymbol{\nu}^*}^n} = \boldsymbol{h}_{\boldsymbol{\nu}}^{\top} \boldsymbol{S}_{\boldsymbol{\nu}n} - \tfrac{1}{2}\, \boldsymbol{h}_{\boldsymbol{\nu}}^{\top} \boldsymbol{\mathcal{I}}_{\boldsymbol{\nu}}\, \boldsymbol{h}_{\boldsymbol{\nu}} + o_{P_{\boldsymbol{\nu}^*}}(1), \tag{74}$$

where $\boldsymbol{S}_{\boldsymbol{\nu}n} = n^{-1/2} \sum_{i=1}^n \nabla_{\boldsymbol{\nu}} \log g_{M_0}(\boldsymbol{W}_i; \boldsymbol{\nu}^*)$ is the null-model score. Since $\boldsymbol{S}_{\boldsymbol{\nu}n} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\mathcal{I}}_{\boldsymbol{\nu}})$ under $P_{\boldsymbol{\nu}^*}^n$, mutual contiguity of $P_{\boldsymbol{\nu}_n}^n$ and $P_{\boldsymbol{\nu}^*}^n$ follows from Le Cam's first lemma (Lehmann and Romano, 2005, Corollary 12.3.1).

Step 2 (Score shift under contiguity). At the null point $(\boldsymbol{\nu}^*, \boldsymbol{\lambda} = \mathbf{0})$, the two-component density reduces to the one-component density: $g_2(\boldsymbol{W}; \boldsymbol{\nu}^*, \mathbf{0}) = g_1(\boldsymbol{W}; \boldsymbol{\nu}^*)$. Hence the nuisance component of the two-component score coincides with the null-model score: $\boldsymbol{s}_{\boldsymbol{\nu}}(\boldsymbol{W}_i) = \nabla_{\boldsymbol{\nu}} \log g_2(\boldsymbol{W}_i; \boldsymbol{\nu}^*, \mathbf{0}) = \nabla_{\boldsymbol{\nu}} \log g_1(\boldsymbol{W}_i; \boldsymbol{\nu}^*)$. The full score vector $\boldsymbol{S}_n = (\boldsymbol{S}_{\boldsymbol{\nu}n}^{\top}, \boldsymbol{S}_{\boldsymbol{\lambda}n}^{\top})^{\top}$ of the two-component model satisfies $\boldsymbol{S}_n \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\mathcal{I}})$ under $P_{\boldsymbol{\nu}^*}^n$ by Lemma 1(b)–(c). Joint convergence of $(\boldsymbol{S}_n, \log dP_{\boldsymbol{\nu}_n}^n / dP_{\boldsymbol{\nu}^*}^n)$ follows from the multivariate CLT applied to $(\boldsymbol{s}(\boldsymbol{W}_i), \nabla_{\boldsymbol{\nu}} \log g_1(\boldsymbol{W}_i; \boldsymbol{\nu}^*))$ together with the LAN expansion (74). Because $\boldsymbol{s}_{\boldsymbol{\nu}}(\boldsymbol{W}_i)$ appears in both vectors, the cross-covariance is

$$\operatorname{Cov}\big(\boldsymbol{S}_n,\ \boldsymbol{h}_{\boldsymbol{\nu}}^{\top} \boldsymbol{S}_{\boldsymbol{\nu}n}\big) = \boldsymbol{\mathcal{I}}\, \boldsymbol{h},$$

where the last equality uses $\boldsymbol{h} = (\boldsymbol{h}_{\boldsymbol{\nu}}^{\top}, \mathbf{0}^{\top})^{\top}$ and the block structure $\boldsymbol{\mathcal{I}} = \left(\begin{smallmatrix} \boldsymbol{\mathcal{I}}_{\boldsymbol{\nu}} & \boldsymbol{\mathcal{I}}_{\boldsymbol{\nu\lambda}} \\ \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\nu}} & \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\lambda}} \end{smallmatrix}\right)$. By Le Cam's third lemma (Lehmann and Romano, 2005, Corollary 12.3.2) (the corollary is stated for scalar statistics; the vector conclusion follows by applying it to every linear combination $\boldsymbol{a}^{\top} \boldsymbol{S}_n$ and invoking the Cramér–Wold device) under $P_{\boldsymbol{\nu}_n}^n$:

$$\boldsymbol{S}_n \ \xrightarrow{d} N(\boldsymbol{\mathcal{I}}\, \boldsymbol{h},\ \boldsymbol{\mathcal{I}}). \tag{75}$$

Step 3 (Cancellation in the projected score). From (75), under $P_{\boldsymbol{\nu}_n}^n$ the component-wise means are $E[\boldsymbol{S}_{\boldsymbol{\nu}n}] = \boldsymbol{\mathcal{I}}_{\boldsymbol{\nu}}\, \boldsymbol{h}_{\boldsymbol{\nu}}$ and $E[\boldsymbol{S}_{\boldsymbol{\lambda}n}] = \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\nu}}\, \boldsymbol{h}_{\boldsymbol{\nu}}$. The projected score (cf. the proof of Proposition 4) is

$$\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\nu}n} \ := \ \boldsymbol{S}_{\boldsymbol{\lambda}n} - \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\nu}}\, \boldsymbol{\mathcal{I}}_{\boldsymbol{\nu}}^{-1}\, \boldsymbol{S}_{\boldsymbol{\nu}n}.$$

Under $P_{\boldsymbol{\nu}_n}^n$, its mean is

$$E[\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\nu}n}] = \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\nu}}\, \boldsymbol{h}_{\boldsymbol{\nu}} - \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\nu}}\, \boldsymbol{\mathcal{I}}_{\boldsymbol{\nu}}^{-1}\, \boldsymbol{\mathcal{I}}_{\boldsymbol{\nu}}\, \boldsymbol{h}_{\boldsymbol{\nu}} = \mathbf{0},$$

and its variance remains the Schur complement $\operatorname{Var}[\boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\nu}n}] = \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\lambda}} - \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda\nu}}\, \boldsymbol{\mathcal{I}}_{\boldsymbol{\nu}}^{-1}\, \boldsymbol{\mathcal{I}}_{\boldsymbol{\nu\lambda}} =: \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\nu}}$ (using the notation of equation (22)), identical to the distribution under $P_{\boldsymbol{\nu}^*}^n$.

Step 4 (LRT distribution is unchanged). Contiguity of $P_{\boldsymbol{\nu}_n}^n$ and $P_{\boldsymbol{\nu}^*}^n$ preserves convergence in probability (Le Cam's first lemma; Corollary 12.3.1 of Lehmann and Romano 2005). Therefore, $\boldsymbol{\mathcal{I}}_n \xrightarrow{p} \boldsymbol{\mathcal{I}}$ and $R_n(\hat{\boldsymbol{\Psi}}_{\alpha}, \alpha) = o_p(1)$ (Lemma 1(a)(c)) continue to hold under $P_{\boldsymbol{\nu}_n}^n$. Combined with Step 3, the proof of Proposition 4 goes through identically under $P_{\boldsymbol{\nu}_n}^n$. In particular, the LRT expansion (41) gives

$$LR_n = \max_{\alpha \in [c_1, 1-c_1]} C_n\left(\sqrt{n}\, \boldsymbol{t}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\lambda}}_\alpha, \alpha)\right) + o_p(1),$$

where C_n depends on the data only through $\boldsymbol{G}_{\boldsymbol{\lambda},\boldsymbol{\nu}n} := \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\nu}}^{-1} \boldsymbol{S}_{\boldsymbol{\lambda},\boldsymbol{\nu}n} \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda},\boldsymbol{\nu}}^{-1})$ under $P^n_{\boldsymbol{\nu}_n}$ (by Step 3). Hence LR_n has the same limiting distribution under $P^n_{\boldsymbol{\nu}_n}$ as under $P^n_{\boldsymbol{\nu}^*}$. □

Proof of Proposition 7. The proof follows the argument in the proof of Theorem 15.4.2 of Lehmann and Romano (2005); see also the supplementary appendix of Kasahara and Shimotsu (2019), which establishes an analogous bootstrap validity result for multivariate normal mixture models using a similar proof strategy.

Part (a). Under $H_0\colon M_0 = 1$, the MLE $\hat{\boldsymbol{\vartheta}}_1$ of the one-component model satisfies $\sqrt{n}(\hat{\boldsymbol{\vartheta}}_1 - \boldsymbol{\vartheta}_1^*) \xrightarrow{d} H$ for a $P_{\boldsymbol{\nu}^*}$-a.s. finite random variable H, by Assumption 6(a)–(d) applied with $M = 1$ (standard regular-MLE asymptotics under the correctly specified one-component model).

By the Almost Sure Representation Theorem (Lehmann and Romano, 2005, Theorem 11.2.19), there exist random variables $\tilde{\boldsymbol{\vartheta}}_1$ and $\tilde{H}$ defined on a common probability space such that

(i) $\hat{\boldsymbol{\vartheta}}_1$ and $\tilde{\boldsymbol{\vartheta}}_1$ have the same distribution, and

(ii) $\sqrt{n}(\tilde{\boldsymbol{\vartheta}}_1 - \boldsymbol{\vartheta}_1^*) \to \tilde{H}$ almost surely.

Identifying $\boldsymbol{\nu}_n = \tilde{\boldsymbol{\vartheta}}_1$ in the reparameterised two-component model, property (ii) gives $\{\boldsymbol{\nu}_n\} \in C_{\boldsymbol{\nu}}$ with probability one. The bootstrap generates $\boldsymbol{W}_1^*, \dots, \boldsymbol{W}_n^*$ i.i.d. from $g_1(\cdot\,; \tilde{\boldsymbol{\vartheta}}_1)$, and the bootstrap LRT LR_n^* is computed on this sample. By Lemma 1(a), conditionally on $\tilde{\boldsymbol{\vartheta}}_1$,

$$LR_n^* \xrightarrow{d} F_1 \qquad \text{under } P^n_{\tilde{\boldsymbol{\vartheta}}_1} \text{ for a.e. realisation of } \tilde{\boldsymbol{\vartheta}}_1.$$

Since $\hat{\boldsymbol{\vartheta}}_1$ and $\tilde{\boldsymbol{\vartheta}}_1$ have the same distribution (property (i)), it follows that $P^*(LR_n^* \le x \mid \text{data}) \xrightarrow{p} F_1(x)$ at every continuity point x of F_1.

Part (b). Under $H_0\colon M_0 = M$ with $M \ge 2$, the restricted MLE $\hat{\boldsymbol{\vartheta}}_M$ satisfies $\sqrt{n}(\hat{\boldsymbol{\vartheta}}_M - \boldsymbol{\vartheta}_M^*) \xrightarrow{d} H$ by Assumption 6(a)–(d). Before invoking Lemma 1(b), we verify that the $(M+1)$-component bootstrap MLE $\hat{\boldsymbol{\vartheta}}^*_{M+1}$ is interior to $\bar{\Theta}_{\boldsymbol{\vartheta}_{M+1}}(\boldsymbol{c})$ with bootstrap probability approaching 1 in outer probability. Under the bootstrap law, data are drawn from $g_M(\cdot\,; \hat{\boldsymbol{\vartheta}}_M)$ with $\hat{\boldsymbol{\vartheta}}_M \xrightarrow{p} \boldsymbol{\vartheta}_M^*$. By Assumption 5(c), the $(M+1)$-component pseudo-true parameter $\boldsymbol{\vartheta}^*_{M+1}$ is interior to $\bar{\Theta}_{\boldsymbol{\vartheta}_{M+1}}(\boldsymbol{c})$. Standard continuity-of-extremum-estimators arguments (e.g., Newey and McFadden 1994, Theorem 2.1) applied to the bootstrap log-likelihood, which converges uniformly to Q^{M+1} under the bootstrap law, then imply $\hat{\boldsymbol{\vartheta}}^*_{M+1} \xrightarrow{p^*} \boldsymbol{\vartheta}^*_{M+1}$ in outer probability, ensuring interiority with bootstrap probability approaching 1. Repeating the Almost Sure Representation argument with Lemma 1(b) in place of (a) yields the stated result.

Part (c). We verify the two conditions needed for sequential consistency.

Under-fitted models ($M < M_0$): By Assumption 6(f), $n^{-1} LR_n^M \xrightarrow{p} Q^{M+1}(\boldsymbol{\vartheta}^*_{M+1}) - Q^M(\boldsymbol{\vartheta}^*_M) > 0$. The bootstrap critical value $\hat{c}^{*,M}_{1-q_n}$ is the $(1-q_n)$-quantile of the bootstrap distribution, which converges to F_M by parts (a)–(b). Applying the Chernoff bound in the proof of Lemma 5 to the bootstrap null distribution gives $n^{-1}\hat{c}^{*,M}_{1-q_n} \xrightarrow{p} 0$. Hence $H_0\colon M_0 = M$ is rejected with probability approaching one.

Correctly specified model ($M = M_0$): We show $\hat{c}^{*,M_0}_{1-q_n} \xrightarrow{p} \infty$. Parts (a)–(b) establish that $LR^{*,M_0}_n \mid \text{data} \xrightarrow{d} F_{M_0}$ in probability, i.e., $P^*(LR^{*,M_0}_n \leq x \mid \text{data}) \xrightarrow{p} F_{M_0}(x)$ for each continuity point x of F_{M_0}. Since F_{M_0} is continuous, the Polya theorem upgrades this to uniform convergence: $\sup_x |P^*(LR^{*,M_0}_n \leq x \mid \text{data}) - F_{M_0}(x)| \xrightarrow{p} 0$. For any $K > 0$, this gives $P^*(LR^{*,M_0}_n > K \mid \text{data}) \xrightarrow{p} 1 - F_{M_0}(K) > 0$. Since $q_n \to 0$, for all large enough n we have $q_n < 1 - F_{M_0}(K)$, which forces $\hat{c}^{*,M_0}_{1-q_n} > K$ with probability approaching one. Because K is arbitrary, $\hat{c}^{*,M_0}_{1-q_n} \xrightarrow{p} \infty$. Since $LR^{M_0}_n = O_p(1)$ under H_0 (Proposition 6),

$$P\left(LR^{M_0}_n > \hat{c}^{*,M_0}_{1-q_n}\right) \leq P\left(LR^{M_0}_n > K\right) + P\left(\hat{c}^{*,M_0}_{1-q_n} \leq K\right) \;\to\; 1 - F_{M_0}(K) + 0$$

for every $K > 0$. Taking $K \to \infty$ yields $P(LR^{M_0}_n > \hat{c}^{*,M_0}_{1-q_n}) \to 0$, so H_0 is not rejected with probability approaching one. Therefore $\widehat{M}_{\mathrm{LRT}} = M_0 + o_p(1)$. □

Remark 1 (Penalised bootstrap)**.** *In the implementation, both estimation and bootstrap re-estimation maximise the penalised log-likelihood $\ell^{\text{pen}}_n(\boldsymbol{\vartheta}) := \ell_n(\boldsymbol{\vartheta}) + p(\boldsymbol{\vartheta})$, where the logarithmic boundary-repulsion penalty is*

$$p(\boldsymbol{\vartheta}) = (a_\alpha - 1)\sum_{j=1}^{M} \log \alpha_j + \sum_{j=1}^{M}(a_\tau - 1)\sum_{k=1}^{K_\epsilon} \log \tau_{jk} \tag{76}$$

for the static model. The penalty $p(\boldsymbol{\vartheta})$ is C^2 on the interior of $\Theta_{\boldsymbol{\vartheta}_M}$ (as a sum of log *functions of parameters bounded away from zero). Variances are regularised only through the hard floor $\sigma^2_j \geq \delta \hat{s}^2$, which does not correspond to a fixed additive term in $p(\boldsymbol{\vartheta})$. We verify that the penalised bootstrap is first-order asymptotically equivalent to the restricted bootstrap.*

Null-model estimator equivalence. *Under $H_0\colon M_0 = M$, the restricted MLE $\hat{\boldsymbol{\vartheta}}_M$ is interior to $\Theta_{\boldsymbol{\vartheta}_M}$ by Assumption 6(c). At any interior point, the penalty gradient satisfies $\nabla p(\boldsymbol{\vartheta}) = O(1)$ because the logarithmic-penalty derivatives are bounded away from zero at interior points, and the variance floor bounds variances away from zero. The penalised score equation is $\sum_{i=1}^n \mathbf{s}(\mathbf{W}_i; \boldsymbol{\vartheta}) + \nabla p(\boldsymbol{\vartheta}) = \mathbf{0}$. Expanding around $\hat{\boldsymbol{\vartheta}}_M$:*

$$\hat{\boldsymbol{\vartheta}}^{\text{pen}}_M - \hat{\boldsymbol{\vartheta}}_M = -[\nabla^2 \ell_n(\hat{\boldsymbol{\vartheta}}_M)]^{-1} \nabla p(\hat{\boldsymbol{\vartheta}}_M) + o_p(n^{-1}) = O_p(n^{-1}) = o_p(n^{-1/2}), \tag{77}$$

since $\nabla^2 \ell_n = O_p(n)$ (Assumption 6(d)–(e)) and $\nabla p(\hat{\boldsymbol{\vartheta}}_M) = O_p(1)$ by interiority. Consequently, $\sqrt{n}(\hat{\boldsymbol{\vartheta}}^{\text{pen}}_M - \boldsymbol{\vartheta}^*_M) = \sqrt{n}(\hat{\boldsymbol{\vartheta}}_M - \boldsymbol{\vartheta}^*_M) + o_p(1)$, so the Almost Sure Representation argument in Proposition 7 goes through with the penalised MLE in place of the restricted MLE.

LRT equivalence. *The LRT is computed using the sample log-likelihood ℓ_n evaluated at the penalised MLEs. For the null model: $|\ell^M_n(\hat{\boldsymbol{\vartheta}}^{\text{pen}}_M) - \ell^M_n(\hat{\boldsymbol{\vartheta}}_M)| = O_p(\sqrt{n}) \cdot O_p(n^{-1}) = o_p(1)$, by a Taylor expansion around $\hat{\boldsymbol{\vartheta}}_M$ and the score being $O_p(\sqrt{n})$ at the true value. For the alternative $(M+1)$-component model under H_0: the penalised MLE $\hat{\boldsymbol{\vartheta}}^{\text{pen}}_{M+1}$ is also interior (the logarithmic penalties bound the mixing proportions and sub-component weights away from zero, and the variance floor bounds variances away from zero), and the same Taylor argument gives $|\ell^{M+1}_n(\hat{\boldsymbol{\vartheta}}^{\text{pen}}_{M+1}) - \ell^{M+1}_n(\hat{\boldsymbol{\vartheta}}_{M+1})| = o_p(1)$. Combining both sides: $|LR^{\text{pen}}_n - LR_n| = o_p(1)$.*

Bootstrap law. *The same bounds hold conditionally on the data for the bootstrap penalised LRT, because the bootstrap data is generated from the penalised null-model MLE (which is interior) and the bootstrap re-estimation uses the same penalty. Therefore, the penalised bootstrap is first-order asymptotically equivalent* *to the restricted bootstrap, and Proposition 7 applies. The logarithmic-penalty terms in* (76) *diverge to* $-\infty$ *as* $\alpha_j \to 0$ *or* $\tau_{jk} \to 0$*, providing smooth boundary repulsion that complements the hard constraints (including the variance floor) in* $\bar{\Theta}_{\vartheta_M}(\boldsymbol{c})$*.*

A.7 Lemmas

Lemma 2. *For any* $\kappa < \infty$, $\Pr\left(-\log n + \ell(\boldsymbol{W}_{i^*}; \bar{Y}_{i^*}, s_{i^*}^2) < \kappa\right) = \exp(-C_T n^{1/T})$ *as* $n \to \infty$*, where* C_T *for* $T = 2, 3, ..$ *are some positive constant that depends on* κ *and* T*.*

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

$$-\log n + \ell(\boldsymbol{W}_{i^*}; \bar{Y}_{i^*}, s_{i^*}^2) = -\log n - \frac{T}{2}\log s_{i^*}^2 - \frac{T}{2}\log(2\pi) - \frac{T-1}{2}$$
$$= -\log\left(Bn\left(\chi^2_{i^*,T-1}\right)^{T/2}\right) \tag{78}$$

with $B := \left(\frac{2\pi\sigma^{*2}}{T-1}\right)^{T/2}\exp(\frac{T-1}{2}) > 0$ and

$$\chi^2_{i^*,T-1} := \frac{(T-1)s_{i^*}^2}{\sigma^{*2}}.$$

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

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

For any finite $T \geq 2$, write

$$\left[1 - F_{T-1}\left((\epsilon/n)^{2/T}\right)\right]^n = \left\{\left[1 - F_{T-1}\left((\epsilon/n)^{2/T}\right)\right]^{\frac{1}{F_{T-1}((\epsilon/n)^{2/T})}}\right\}^{nF_{T-1}((\epsilon/n)^{2/T})}. \tag{79}$$

Then, because $(1-F)^{\frac{1}{F}} \to \frac{1}{e}$ when $F \to 0$, the stated result follows from (79) if we can show

$$\frac{F_{T-1}((\epsilon x)^{2/T})}{x} = O(x^{-1/T}) \text{ as } x \to 0 \tag{80}$$

for $x = 1/n$, where (79)-(80) imply that $\log\left[1 - F_{T-1}\left((\epsilon/n)^{2/T}\right)\right]^n \leq -C_T n^{1/T}$ for some positive constant C_T when n is sufficiently large.

For small t, the CDF $F_k(t)$ can be expanded as $F_k(t) = \frac{t^{k/2}}{2^{k/2}\Gamma(k/2+1)} + o(t^{k/2})$ as $t \to 0$. Then, letting $k = T-1$ and $t = (\epsilon x)^{2/T} = (\epsilon x)^{2/(k+1)}$, and simplifying the exponent $(x^{2/(k+1)})^{k/2} = x^{k/(k+1)}$, we have

$$F_{T-1}((\epsilon x)^{2/T}) = \frac{(\epsilon x)^{k/(k+1)}}{2^{k/2}\Gamma(k/2+1)} + o((\epsilon x)^{k/(k+1)}) \text{ as } x \to 0.$$

Therefore, given finite $\epsilon > 0$, noting that $x^{k/(k+1)-1} = x^{-1/(k+1)}$, we have

$$\frac{F_{T-1}((\epsilon x)^{2/T})}{x} = \frac{\epsilon^{k/(k+1)}x^{-1/(k+1)}}{2^{k/2}\Gamma(k/2+1)} + o(x^{-1/(k+1)}) = O(x^{-1/(k+1)}) \text{ as } x \to 0,$$

and the equation (80) follows with $k+1 = T$.

□

Lemma 3. *For any $M < \infty$, $\Pr\left(-4\log n + \ell(\boldsymbol{W}_{i^*}; \bar{Y}_{i^*}, s_{i^*}^2) < M\right) \to 1$ as $n \to \infty$.*

Proof. Following the proof of Lemma 2, because $-4\log n + \ell(\boldsymbol{W}_{i^*}; \bar{Y}_{i^*}, s_{i^*}^2) = -\log\left(Bn^4\left(\chi^2_{i^*,T-1}\right)^{T/2}\right)$, it suffices to show that, for any $\epsilon > 0$, $\Pr\left(\chi^2_{i^*,T-1} > \left(\epsilon/n^4\right)^{2/T}\right) = \left[1 - F_{T-1}\left((\epsilon/n^4)^{2/T}\right)\right]^n \to 1$ as $n \to \infty$. For any finite $T \geq 2$, write

$$\left[1 - F_{T-1}\left((\epsilon/n^4)^{2/T}\right)\right]^n = \left\{\left[1 - F_{T-1}\left((\epsilon/n^4)^{2/T}\right)\right]^{\frac{1}{F_{T-1}((\epsilon/n^4)^{2/T})}}\right\}^{nF_{T-1}((\epsilon/n^4)^{2/T})}. \tag{81}$$

Then, because $(1-F)^{\frac{1}{F}} \to \frac{1}{e}$ when $F \to 0$, the stated result follows from (81) if we can show $\frac{F_{T-1}((\epsilon x^4)^{2/T})}{x} \to 0$ as $x \to 0$ for $x = 1/n$. By applying L'Hôpital's rule, we have

$$\lim_{x\to 0}\frac{F_{T-1}((\epsilon x^4)^{2/T})}{x} = \lim_{x\to 0} f_{T-1}((\epsilon x^4)^{2/T})\epsilon^{2/T}(8/T)x^{8/T-1} = \lim_{x\to 0}\tilde{C}_{T,\epsilon}e^{-((\epsilon x^4)^{2/T})/2}x^{3-\frac{4}{T}} = 0,$$

where $\tilde{C}_{T,\epsilon} = \frac{\epsilon^{(T-1)/T}(8/T)}{2^{(T-1)/2}\Gamma((T-1)/2)}$ because $e^{-((\epsilon x^4)^{2/T})/2} \to 1$ and $x^{3-\frac{4}{T}} \to 0$ as $x \to 0$ for any finite $T \geq 2$. □

Proof of Lemma 1. The proof follows that of Proposition 2 in Kasahara and Shimotsu (2012). For a vector $\boldsymbol{x}$ and a function $f(\boldsymbol{x})$, let $\nabla_{\boldsymbol{x}^k} f(\boldsymbol{x})$ denote its k-th derivative with respect to $\boldsymbol{x}$, which can be a multidimensional array. Observe that for any finite k and for a neighborhood $\mathcal{N}$ of $\boldsymbol{\psi}^*$, we obtain

$$\begin{aligned} &E||\nabla_{\boldsymbol{\psi}^k} g(\boldsymbol{W}_i; \boldsymbol{\psi}^*, \alpha)/g(\boldsymbol{W}_i; \boldsymbol{\psi}^*, \alpha)||^2 < \infty, \\ &E||\sup_{\boldsymbol{\psi}\in\Theta_{\boldsymbol{\psi}}\cap\mathcal{N}} \nabla_{\boldsymbol{\psi}^k} \log g(\boldsymbol{W}_i; \boldsymbol{\psi}, \alpha)||^2 < \infty \end{aligned} \tag{82}$$

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

$$\nabla_{\eta\lambda_j} L_n(\boldsymbol{\Psi}^*, \alpha) = 0, \quad \nabla_{\lambda_i\lambda_j\lambda_k} L_n(\boldsymbol{\Psi}^*, \alpha) = O_p(n^{1/2}), \tag{83}$$

$$\nabla_{\eta\eta\lambda_i} L_n(\boldsymbol{\Psi}^*, \alpha) = O_p(n), \quad \nabla_{\eta\eta\eta} L_n(\boldsymbol{\Psi}^*, \alpha) = O_p(n), \tag{84}$$

where equation (83) follows from Lemma 6(a), (c), and (d) and (82) and equation (84) is a simple consequence of (82). For a neighborhood $\mathcal{N}$ of $\boldsymbol{\Psi}^*$,

$$\sup_{\boldsymbol{\Psi}\in\Theta_{\boldsymbol{\Psi}}\cap\mathcal{N}} \left|n^{-1}\nabla^{(4)} L_n(\boldsymbol{\Psi}, \alpha) - E\nabla^{(4)} \log g(\boldsymbol{W}_i; \boldsymbol{\Psi}, \alpha)\right| = o_p(1), \tag{85}$$

$$E\nabla^{(4)} \log g(\boldsymbol{W}_i; \boldsymbol{\Psi}, \alpha) \text{ is continuous in } \psi \in \Theta_{\boldsymbol{\Psi}} \cap \mathcal{N}. \tag{86}$$

Equations (85) and (86) follow from Lemma 2.4 of Newey and McFadden (1994) and the fact that $\nabla_{\psi^k} \log g(\boldsymbol{w}; \boldsymbol{\psi}, \alpha)$ is written as a sum of products of Hermite polynomials.

Taking a fourth-order Taylor expansion of $L_n(\boldsymbol{\psi}, \alpha)$ around $\boldsymbol{\psi}^*$ and using (82) and (83), we can write $L_n(\boldsymbol{\psi}, \alpha) - L_n(\boldsymbol{\psi}^*, \alpha)$ as the sum of the relevant terms and the remainder term as follows:

$$
\begin{aligned}
& L_n(\boldsymbol{\psi}, \alpha) - L_n(\boldsymbol{\psi}^*, \alpha) = \\
& \nabla_{\boldsymbol{\eta}} L_n^*(\boldsymbol{\eta} - \boldsymbol{\eta}^*) + \frac{1}{2!}(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^\top \nabla_{\boldsymbol{\eta}\boldsymbol{\eta}^\top} L_n^*(\boldsymbol{\eta} - \boldsymbol{\eta}^*) + \frac{1}{2!}\sum_{i=1}^{q}\sum_{j=1}^{q} \nabla_{\lambda_i\lambda_j} L_n^* \lambda_i\lambda_j & (87) \\
& + \frac{3}{3!}\sum_{i=1}^{q}\sum_{j=1}^{q} (\boldsymbol{\eta} - \boldsymbol{\eta}^*)^\top \nabla_{\boldsymbol{\eta}\lambda_i\lambda_j} L_n^* \lambda_i\lambda_j & (88) \\
& + \frac{1}{4!}\sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{\ell=1}^{q} \nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell} L_n^* \lambda_i\lambda_j\lambda_k\lambda_\ell + R_n(\boldsymbol{\psi}, \alpha), & (89)
\end{aligned}
$$

where ∇L_n^* denotes the derivative of $L_n(\boldsymbol{\psi}, \alpha)$ evaluated at $(\boldsymbol{\psi}^*, \alpha)$. In view of (83) and (84), the remainder term is written as

$$
\begin{aligned}
R_n(\boldsymbol{\Psi}, \alpha) &= O_p(n^{1/2}) \sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{q} \lambda_i\lambda_j\lambda_k + O_p(n)\left(\sum_{i=1}^{q} ||\boldsymbol{\eta} - \boldsymbol{\eta}^*||^2 \lambda_i + ||\boldsymbol{\eta} - \boldsymbol{\eta}^*||^3 \right) & (90) \\
&+ O_p(n) \sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{q} \left(||\boldsymbol{\eta} - \boldsymbol{\eta}^*||^4 + ||\boldsymbol{\eta} - \boldsymbol{\eta}^*||^3|\lambda_i| + ||\boldsymbol{\eta} - \boldsymbol{\eta}^*||^2|\lambda_i\lambda_j| + ||\boldsymbol{\eta} - \boldsymbol{\eta}^*|| |\lambda_i\lambda_j\lambda_k| \right) & (91) \\
&+ \frac{1}{4!}\sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{\ell=1}^{q} \{\nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell} L_n(\boldsymbol{\psi}^\dagger, \alpha) - \nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell} L_n(\boldsymbol{\psi}^*, \alpha)\} \lambda_i\lambda_j\lambda_k\lambda_\ell & (92)
\end{aligned}
$$

with $\boldsymbol{\psi}^\dagger$ being between $\boldsymbol{\psi}$ and $\boldsymbol{\psi}^*$. Because $||\sqrt{n}t(\boldsymbol{\psi}, \alpha)||^2 = n||\boldsymbol{\eta} - \boldsymbol{\eta}^*||^2 + n\sum_{i=1}^{q}\sum_{j=1}^{i} \alpha^2(1-\alpha)^2|\lambda_i\lambda_j|^2$, the right-hand side of (90) and the terms in (91) are bounded by $O_p(1)(||\sqrt{n}t(\boldsymbol{\psi}, \alpha)|| + ||\sqrt{n}t(\boldsymbol{\psi}, \alpha)||^2)(||\boldsymbol{\eta} - \boldsymbol{\eta}^*|| + ||\lambda||)$. In view of (85) and (86), (92) is bounded by $||\sqrt{n}t(\boldsymbol{\psi}, \alpha)||^2[d(\boldsymbol{\psi}^\dagger) + o_p(1)]$ with $d(\boldsymbol{\psi}^\dagger) \to 0$ as $\boldsymbol{\psi}^\dagger \to \boldsymbol{\psi}^*$, where a function $d(\boldsymbol{\psi}^\dagger)$ corresponds to $n^{-1}\mathbb{E}[\nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell} L_n(\boldsymbol{\psi}^\dagger, \alpha) - \nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell} L_n(\boldsymbol{\psi}^*, \alpha)]$. Therefore, $R_n(\boldsymbol{\psi}, \alpha) = (1 + ||\sqrt{n}t(\boldsymbol{\psi}, \alpha)||)^2[d(\boldsymbol{\psi}^\dagger) + o_p(1) + O_p(||\boldsymbol{\psi} - \boldsymbol{\psi}^*||)]$, and part (a) follows.

Part (b) follows from Lemma 6(c) and (d), the Lindeberg–Levy central limit theorem, and the finiteness of $\boldsymbol{\mathcal{I}}$ in part (c).

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

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

where $q_\lambda = q(q+1)/2$. $\boldsymbol{\mathcal{I}}_{\eta n}$ is given by $\boldsymbol{\mathcal{I}}_{\eta n} = -n^{-1}\nabla_{\eta\eta^\top} L_n(\psi^*, \alpha)$. For $\boldsymbol{\mathcal{I}}_{\eta\lambda n}$, let $A_{ij} = n^{-1}\nabla_{\eta\lambda_i\lambda_j} L_n(\psi^*, \alpha)$ and write the term in (88) as $(n/2)\sum_{i=1}^{q}\sum_{j=1}^{q}(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^\top A_{ij}\lambda_i\lambda_j = n\sum_{i=1}^{q}\sum_{j=1}^{i}(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^\top A_{ij}c_{ij}\lambda_i\lambda_j$, where $c_{ij} = 1/2$ if $i = j$, and $c_{ij} = 1$ if $i \neq j$. Then, by defining $\boldsymbol{\mathcal{I}}_{\eta\lambda n} = -(A_{11}, \ldots, A_{qq}, A_{12}, \ldots, A_{q-1,q})/\alpha(1-\alpha)$, the term in (88) equals $-n(\boldsymbol{\eta} - \boldsymbol{\eta}^*)^\top \boldsymbol{\mathcal{I}}_{\eta\lambda n}[\alpha(1-\alpha)v(\boldsymbol{\lambda})]$. For $\boldsymbol{\mathcal{I}}_{\lambda n}$, define $B_{ijk\ell} = n^{-1}(8/4!)\nabla_{\lambda_i\lambda_j\lambda_k\lambda_\ell} L_n(\boldsymbol{\Psi}^*, \alpha)$ so that the first term in (89) is written as

$(n/8)\sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{\ell=1}^{q} B_{ijk\ell}\lambda_i\lambda_j\lambda_k\lambda_\ell = (n/2)\sum_{i=1}^{q}\sum_{j=1}^{i}\sum_{k=1}^{q}\sum_{\ell=1}^{k} c_{ij}c_{k\ell}B_{ijk\ell}\lambda_i\lambda_j\lambda_k\lambda_\ell$. Define $\boldsymbol{\mathcal{I}}_{\lambda n}$ such that the $(ij,k\ell)$ element of $\boldsymbol{\mathcal{I}}_{\lambda n}$ is $-c_{ij}c_{k\ell}B_{ijk\ell}/\alpha^2(1-\alpha)^2$, where the values of ij run over $\{(1,1),\dots,(q,q),(1,2),\dots,(q-1,q)\}$. Then, the first term in (89) equals $-(n/2)[\alpha(1-\alpha)v(\boldsymbol{\lambda})]'\boldsymbol{\mathcal{I}}_{\lambda n}[\alpha(1-\alpha)v(\boldsymbol{\lambda})]$. With this definition of $\boldsymbol{\mathcal{I}}_n$, the expansion (87)-(89) is written as (36) in terms of $\sqrt{n}t(\psi,\alpha)$.

We now show that $\boldsymbol{\mathcal{I}}_n \to_p \boldsymbol{\mathcal{I}}$. $\boldsymbol{\mathcal{I}}_{\eta n} \to_p \boldsymbol{\mathcal{I}}_\eta$ holds trivially. For $\boldsymbol{\mathcal{I}}_{\eta\lambda n}$, it follows from Lemma 6(c) and the law of large numbers that $A_{ij} \to_p -\mathbb{E}[\nabla_{\boldsymbol{\eta}} l(\boldsymbol{W};\psi^*,\alpha)\nabla_{\lambda_i\lambda_j} l(\boldsymbol{W};\psi^*,\alpha)]$, giving $\boldsymbol{\mathcal{I}}_{\eta\lambda n} \to_p E\left[\boldsymbol{s}_{\boldsymbol{\eta}}\boldsymbol{s}_{\boldsymbol{\lambda\lambda}}^\top/\alpha(1-\alpha)\right] = \boldsymbol{\mathcal{I}}_{\eta\lambda}$. For $\boldsymbol{\mathcal{I}}_{\lambda n}$, Lemma 6(d) and (e) and the law of large numbers imply that $\sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{\ell=1}^{q} B_{ijk\ell}\lambda_i\lambda_j\lambda_k\lambda_\ell \to_p$
$-\sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{q}\sum_{\ell=1}^{q} E[\nabla_{\lambda_i\lambda_j} l(\boldsymbol{W};\psi^*,\alpha)\nabla_{\lambda_k\lambda_\ell} l(\boldsymbol{W};\psi^*,\alpha)]\lambda_i\lambda_j\lambda_k\lambda_\ell$, where the factor $(8/4!) = 1/3$ in $B_{ijk\ell}$ and the three derivatives on the right-hand side of Lemma 6(e) cancel each other out. Therefore, we have $\boldsymbol{\mathcal{I}}_{\lambda n} \to_p E\left[\boldsymbol{s}_{\boldsymbol{\lambda\lambda}}\boldsymbol{s}_{\boldsymbol{\lambda\lambda}}^\top/\alpha^2(1-\alpha)^2\right] = \boldsymbol{\mathcal{I}}_{\boldsymbol{\lambda}}$, and $\boldsymbol{\mathcal{I}}_n \to_p \boldsymbol{\mathcal{I}}$ follows.

We complete the proof of part (c) by showing that $\boldsymbol{\mathcal{I}} = E[\boldsymbol{s}(\boldsymbol{W})\boldsymbol{s}(\boldsymbol{W})^\top]$ is finite and non-singular. Note that $\boldsymbol{s}(\boldsymbol{W})$ can be expressed in Hermite polynomials as in (19). Then, the finiteness of $\boldsymbol{\mathcal{I}}$ follows from Assumption 4(a) and the definition of Hermite polynomials.

To show that $\boldsymbol{\mathcal{I}}$ is positive definite, it suffices to show that there exists no multicollinearity in $\boldsymbol{s}(\boldsymbol{w})$. Suppose, to the contrary, that $\boldsymbol{s}(\boldsymbol{w})$ is multicollinear and that there exists a non-zero vector $\boldsymbol{a}$ that solves the equation $\boldsymbol{a}^\top\boldsymbol{s}(\boldsymbol{w}) = 0$ for all values of $\boldsymbol{w}$. Partition $s(\boldsymbol{w})$ as $\boldsymbol{s}(\boldsymbol{w}) = (\boldsymbol{s}_{(\mu)}^\top, \boldsymbol{s}_{(\beta)}^\top)^\top$ with $\boldsymbol{s}_{(\mu)} = (s_\mu, s_\sigma, s_{\lambda_{\mu\mu}}, s_{\lambda_{\mu\sigma}}, s_{\lambda_{\sigma\sigma}})^\top$ and $\boldsymbol{s}_{(\beta)} = (\boldsymbol{s}_{\boldsymbol{\beta}}^\top, \boldsymbol{s}_{\lambda_{\mu\beta}}^\top, \boldsymbol{s}_{\lambda_{\sigma\beta}}^\top, \boldsymbol{s}_{\lambda_{\beta\beta}}^\top)^\top$, where $\boldsymbol{s}(\boldsymbol{w})$ is defined in (2) and (19). Similarly, partition $\boldsymbol{a}$ as $\boldsymbol{a} = (\boldsymbol{a}_{(\mu)}^\top, \boldsymbol{a}_{(\beta)}^\top)^\top$ so that

$$\boldsymbol{a}^\top\boldsymbol{s}(\boldsymbol{w}) = \boldsymbol{a}_{(\mu)}^\top\boldsymbol{s}_{(\mu)} + \boldsymbol{a}_{(\beta)}^\top\boldsymbol{s}_{(\beta)}. \tag{93}$$

By Assumption 4(b) and the property of Hermite polynomials, if $\boldsymbol{a}^\top\boldsymbol{s}(\boldsymbol{w}) = \boldsymbol{0}$ for all $\boldsymbol{w}$, then $\boldsymbol{a}_{(\beta)} = \boldsymbol{0}$.

Then, in view of (93), the stated result follows if we can show that $\boldsymbol{a}_{(\mu)}^\top\boldsymbol{s}_{(\mu)} = \boldsymbol{0}$ for all $\boldsymbol{w}$ implies $\boldsymbol{a}_{(\mu)} = 0$. Suppose that

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

for all $\boldsymbol{w}$, where H_t^{j*} for $j = 1,2,3$ is defined in (18) in Appendix A.3.2.

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

The proof of part (d) closely follows the proof of Theorem 1 of Andrews (1999). Let $\boldsymbol{T}_n :=$

$\mathcal{I}_n^{1/2}\sqrt{n}\boldsymbol{t}(\hat{\boldsymbol{\Psi}}_\alpha, \alpha)$. Then, in view of (36), we have

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

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

□

Lemma 4 (Linear independence of posterior-weighted Hermite functions)**.** *Let $\tau^* \in (0,1)$, $\sigma^* > 0$, and $\mu_1^*, \mu_2^* \in \mathbb{R}$ with $\mu_1^* \neq \mu_2^*$. Define the densities $\phi_k(y) := (\sigma^*)^{-1}\phi((y-\mu_k^*)/\sigma^*)$ for $k = 1, 2$, the mixture density $f^*(y) := \tau^*\phi_1(y) + (1-\tau^*)\phi_2(y)$, the posterior weights $\gamma_k(y) := \tau_k^*\phi_k(y)/f^*(y)$ with $\tau_1^* := \tau^*$ and $\tau_2^* := 1 - \tau^*$, and the Hermite functions $H_k^{a*}(y) := (\sigma^*)^{-a}H^a((y-\mu_k^*)/\sigma^*)$, where H^a is the probabilists' Hermite polynomial of degree a as defined in* (18)*. Set*

$$\begin{aligned}
V_1(y) &:= \frac{\phi_1(y) - \phi_2(y)}{f^*(y)} = \frac{\gamma_1(y)}{\tau^*} - \frac{\gamma_2(y)}{1-\tau^*}, \\
V_2(y) &:= \gamma_1(y)H_1^{1*}(y), \\
V_3(y) &:= \gamma_2(y)H_2^{1*}(y), \\
V_4(y) &:= \gamma_1(y)H_1^{2*}(y) + \gamma_2(y)H_2^{2*}(y).
\end{aligned} \tag{94}$$

The five functions $\{1, V_1, V_2, V_3, V_4\}$ are linearly independent on $\mathbb{R}$.

Proof. Suppose $c_0 + \sum_{j=1}^4 c_jV_j(y) = 0$ for all $y \in \mathbb{R}$. Multiplying by $f^*(y) > 0$ and using the Rodrigues identity for the probabilists' Hermite polynomials,

$$H_k^{a*}(y)\,\phi_k(y) = (-1)^a\partial_y^a\phi_k(y), \tag{95}$$

yields the identity

$$P_1(y)\,\phi_1(y) + P_2(y)\,\phi_2(y) = 0 \quad \forall y \in \mathbb{R},$$

where P_1, P_2 are polynomials of degree at most 2 in y depending linearly on $(c_0, \ldots, c_4)$. Both sides are real-analytic, so the identity holds for all $y \in \mathbb{R}$.

Without loss of generality assume $\mu_1^* < \mu_2^*$. Dividing by $\phi_2(y) > 0$,

$$P_1(y)\frac{\phi_1(y)}{\phi_2(y)} + P_2(y) = 0, \qquad \frac{\phi_1(y)}{\phi_2(y)} = \exp\left(\frac{(\mu_1^* - \mu_2^*)y}{\sigma^{*2}} + \frac{\mu_2^{*2} - \mu_1^{*2}}{2\sigma^{*2}}\right). \tag{96}$$

Since $\mu_1^* < \mu_2^*$, the exponential factor decays to 0 as $y \to +\infty$ super-polynomially, so $P_1(y)\phi_1(y)/\phi_2(y) \to 0$. Taking $y \to +\infty$ in (96) forces $P_2(y) \to 0$; but P_2 is a polynomial in

y, so $P_2 \equiv 0$. Substituting back, $P_1 \equiv 0$. Reading off the coefficients of the zero polynomials P_1, P_2 in terms of $(c_0, \ldots, c_4)$ yields $c_0 = c_1 = c_2 = c_3 = c_4 = 0$.

The above linear-independence statement for posterior-weighted Hermite functions at two distinct centres is the classical identifiability result for finite normal mixtures with equal variance and distinct means, due to Teicher (1963) and Yakowitz and Spragins (1968); we have given a self-contained derivation specialised to the auxiliary basis required below. □

Remark 2. *The proof of Lemma 4 extends mechanically to any $K_\epsilon \geq 2$ provided the within-type means $\mu_1^*, \ldots, \mu_{K_\epsilon}^*$ are pairwise distinct (Assumption 1). One obtains an identity $\sum_{k=1}^{K_\epsilon} P_k(y)\phi_k(y) = 0$ with P_k polynomials of degree at most $2(K_\epsilon - 1)$; ordering $\mu_1^* < \cdots < \mu_{K_\epsilon}^*$ and taking $y \to +\infty$ peels off $P_{K_\epsilon} \equiv 0$, then $P_{K_\epsilon - 1} \equiv 0$,* etc.*, by iteration.*

Assumption MOD-1 (stated in the main text) is the natural cross-period strengthening of Assumption 4(b). It is automatically implied by the conjunction of Assumption 4(b) and Assumption 3's "identifying covariate variation" clause whenever $(\boldsymbol{X}_t, \boldsymbol{X}_s)$ admits a joint density positive on an open set of $\mathbb{R}^{2q}$. This condition holds in the empirical specifications of Section 7, where $\boldsymbol{X}$ is a continuous panel process with non-degenerate joint distribution across t. It rules out degenerate panel-covariate processes such as perfect persistence ($\boldsymbol{X}_1 = \boldsymbol{X}_2$) or supports concentrated on polynomial varieties (e.g. $\{X_1 X_2 = 0\}$ a.s.) for which Assumption 4(b) by itself does not control the bilinear cross-period design rank.[1]

Proof of Lemma 2. Parts (a), (b), and (d) extend from the proof of Lemma 1 verbatim: each element of $\nabla_{\psi^k} \log g(\boldsymbol{w}; \boldsymbol{\psi}, \alpha)$ for $k \leq 4$ is a finite sum of products of posterior probabilities $\gamma_{kt} \in [0, 1]$, Hermite polynomials of degree at most 4, and (when covariates are present) monomials in $\boldsymbol{X}_t$ of degree at most 2. The derivative bounds (82)–(86) and the Bartlett-style identities of Lemma 6 continue to apply, and the Lindeberg–Lévy CLT, the uniform LLN of Newey and McFadden (1994), and the boundary-MLE argument of Andrews (1999) proceed exactly as in the proof of Lemma 1.

It remains to prove part (c). *Finiteness* of $\boldsymbol{\mathcal{I}} = \mathbb{E}[\boldsymbol{s}(\boldsymbol{W})\boldsymbol{s}(\boldsymbol{W})^\top]$ follows because each entry is a polynomial of degree at most 4 in the standardised residuals $Z_{kt} = (Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta}^* - \mu_k^*)/\sigma^*$ and degree at most 2 in $\boldsymbol{X}_t$; the standardised residuals have all polynomial moments, and Assumption 4 (a) supplies the required 9th moment of $\boldsymbol{X}$ (more than enough to cover degree 4 after squaring).

Non-singularity: suppose, toward a contradiction, that a non-zero vector $\boldsymbol{a}$ satisfies $\boldsymbol{a}^\top \boldsymbol{s}(\boldsymbol{W}) = 0$ a.s. We will deduce $\boldsymbol{a} = \boldsymbol{0}$. Partition $\boldsymbol{a}$ into a first-order block $\boldsymbol{a}_1 := (a_\tau, a_{\mu_1}, a_{\mu_2}, a_\sigma, \boldsymbol{a}_{\boldsymbol{\beta}}^\top)^\top$ and a $(4+q) \times (4+q)$ symmetric matrix $\boldsymbol{A}_{\lambda\lambda}$ collecting the ten $\boldsymbol{\lambda}\boldsymbol{\lambda}^\top$ coefficients $(a_{\lambda_{\tau\tau}}, a_{\lambda_{\mu_1\mu_1}}, a_{\lambda_{\mu_2\mu_2}}, a_{\lambda_{\mu_1\mu_2}}, a_{\lambda_{\tau\mu_1}}, a_{\lambda_{\tau\mu_2}}, a_{\lambda_{\mu_1\sigma}}, a_{\lambda_{\mu_2\sigma}}, a_{\lambda_{\sigma\sigma}}, a_{\lambda_{\tau\sigma}})$ together with the $\boldsymbol{\lambda}_{\boldsymbol{\beta}}$-blocks.

Write the standardised residual $\widetilde{Y}_t := Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta}^*$. By (22)–(23) together with Lemma 6(c), the score admits the U-statistic representation

$$\boldsymbol{a}^\top \boldsymbol{s}(\boldsymbol{W}) = \sum_{t=1}^{T} u_{\boldsymbol{a}_1}(\widetilde{Y}_t, \boldsymbol{X}_t) + \sum_{t \neq s} \boldsymbol{V}(\widetilde{Y}_t)^\top \boldsymbol{M}(\boldsymbol{X}_t, \boldsymbol{X}_s)\, \boldsymbol{V}(\widetilde{Y}_s), \tag{97}$$

where $\boldsymbol{V}(\widetilde{Y}_t) := (V_1(\widetilde{Y}_t), V_2(\widetilde{Y}_t), V_3(\widetilde{Y}_t), V_4(\widetilde{Y}_t))^\top$ is the auxiliary basis of Lemma 4, $u_{\boldsymbol{a}_1}$ collects the single-period first-order and same-t second-order contributions, and $\boldsymbol{M}(\boldsymbol{X}_t, \boldsymbol{X}_s) = \boldsymbol{J}(\boldsymbol{X}_t)^\top \boldsymbol{A}_{\lambda\lambda} \boldsymbol{J}(\boldsymbol{X}_s)$

[1] An adversarial counterexample illustrates the necessity. Take $\tau^* = 1/2$, $\mu_1^* = 0$, $\mu_2^* = 1$, $\sigma^* = 1$, $T = 2$, $q = 1$, and $P_{\boldsymbol{X}}\{(X_1, X_2) \in \{(0,0), (1,0), (0,1)\}\} = 1/3$ each. Then $\mathbb{E}[\boldsymbol{U}\boldsymbol{U}^\top]$ is non-singular (LRT1(b) holds), yet $X_1 X_2 = 0$ a.s., so the matrix $B = \mathrm{diag}(0, 1) \neq \boldsymbol{0}$ satisfies $\boldsymbol{U}_1^\top B \boldsymbol{U}_2 = X_1 X_2 = 0$ a.s. The implication "$\boldsymbol{U}_1^\top B \boldsymbol{U}_2 = 0$ a.s. $\Rightarrow B = \boldsymbol{0}$" therefore requires Assumption MOD-1, not merely Assumption 4(b).

for a deterministic block matrix $\boldsymbol{J}(\boldsymbol{x}) \in \mathbb{R}^{(4+q)\times 4}$ whose first four rows are $\boldsymbol{I}_4$ (no $\boldsymbol{x}$-factor) and whose lower q-block inserts $\boldsymbol{x}$ in the V_2- and V_3-slots (the $\boldsymbol{\beta}$-block columns). The factorisation $\boldsymbol{J}(\boldsymbol{x})\boldsymbol{v} = \boldsymbol{L}(\boldsymbol{v})\boldsymbol{U}$ with $\boldsymbol{U} = (1, \boldsymbol{x}^\top)^\top$ and $\boldsymbol{L}(\boldsymbol{v})$ a deterministic linear map of $\boldsymbol{v} \in \mathbb{R}^4$ records the explicit dependence on $\boldsymbol{U}$.

Step 1 (Mixed-difference identity). Since $f^*(\boldsymbol{y} \mid \boldsymbol{x}) > 0$ everywhere, $\boldsymbol{a}^\top \boldsymbol{s}(\boldsymbol{W}) = 0$ a.s. implies the identity holds for all $(\boldsymbol{Y}, \boldsymbol{X})$ in the joint support. Because $T \geq 2$ (Assumption 3), we may apply the bivariate mixed difference operator $\Delta := \Delta_{\widetilde{y}_1, z_1}\Delta_{\widetilde{y}_2, z_2}$ that holds $\widetilde{Y}_3, \ldots, \widetilde{Y}_T$ and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ fixed and takes

$$\Delta h(\widetilde{Y}_1, \widetilde{Y}_2) := h(\widetilde{y}_1, \widetilde{y}_2) - h(z_1, \widetilde{y}_2) - h(\widetilde{y}_1, z_2) + h(z_1, z_2),$$

for arbitrary $\widetilde{y}_1, \widetilde{y}_2, z_1, z_2 \in \mathbb{R}$. All single-period terms $u_{\boldsymbol{a}_1}(\widetilde{Y}_t, \boldsymbol{X}_t)$ in (97) are additively separable in $(\widetilde{y}_1, \widetilde{y}_2)$ (they depend on at most one of them at a time) and so are annihilated by Δ. Using symmetry $\boldsymbol{M}(\boldsymbol{X}_2, \boldsymbol{X}_1)^\top = \boldsymbol{M}(\boldsymbol{X}_1, \boldsymbol{X}_2)$, the cross-period sum reduces to a single $(t, s) = (1, 2)$ pair after applying Δ:

$$2\left(\boldsymbol{V}(\widetilde{y}_1) - \boldsymbol{V}(z_1)\right)^\top \boldsymbol{M}(\boldsymbol{X}_1, \boldsymbol{X}_2)\left(\boldsymbol{V}(\widetilde{y}_2) - \boldsymbol{V}(z_2)\right) = 0 \tag{98}$$

for almost every $(\widetilde{y}_1, z_1, \widetilde{y}_2, z_2) \in \mathbb{R}^4$ and for $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ in the joint support.

Step 2 ($\boldsymbol{A}_{\lambda\lambda} = \boldsymbol{0}$). By Lemma 4, the five functions $\{1, V_1, V_2, V_3, V_4\}$ are linearly independent on $\mathbb{R}$, so the difference set $\{\boldsymbol{V}(\widetilde{y}) - \boldsymbol{V}(z) : \widetilde{y}, z \in \mathbb{R}\}$ spans $\mathbb{R}^4$. Therefore (98) forces

$$\boldsymbol{M}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{J}(\boldsymbol{X}_1)^\top \boldsymbol{A}_{\lambda\lambda}\boldsymbol{J}(\boldsymbol{X}_2) = \boldsymbol{0} \quad \text{a.s.}$$

For any fixed $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^4$, write $\boldsymbol{J}(\boldsymbol{X}_i)\boldsymbol{v}_i = \boldsymbol{L}(\boldsymbol{v}_i)\boldsymbol{U}_i$. Then $\boldsymbol{U}_1^\top \boldsymbol{L}(\boldsymbol{v}_1)^\top \boldsymbol{A}_{\lambda\lambda}\boldsymbol{L}(\boldsymbol{v}_2)\boldsymbol{U}_2 = 0$ a.s. over the joint distribution of $(\boldsymbol{U}_1, \boldsymbol{U}_2)$. Vectorising, $(\boldsymbol{U}_1 \otimes \boldsymbol{U}_2)^\top \operatorname{vec}\!\big(\boldsymbol{L}(\boldsymbol{v}_1)^\top \boldsymbol{A}_{\lambda\lambda}\boldsymbol{L}(\boldsymbol{v}_2)\big) = 0$ a.s. over $(\boldsymbol{U}_1 \otimes \boldsymbol{U}_2)$. By Assumption MOD-1, $\mathbb{E}[(\boldsymbol{U}_1 \otimes \boldsymbol{U}_2)(\boldsymbol{U}_1 \otimes \boldsymbol{U}_2)^\top]$ is non-singular, hence $\boldsymbol{L}(\boldsymbol{v}_1)^\top \boldsymbol{A}_{\lambda\lambda}\boldsymbol{L}(\boldsymbol{v}_2) = \boldsymbol{0}$ for every fixed $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^4$. Standard basis variations of $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ then force each block of $\boldsymbol{A}_{\lambda\lambda}$ to zero, so $\boldsymbol{A}_{\lambda\lambda} = \boldsymbol{0}$ identically.

Step 3 ($\boldsymbol{a}_1 = \boldsymbol{0}$). With $\boldsymbol{A}_{\lambda\lambda} = \boldsymbol{0}$, the identity (97) reduces to

$$\sum_{t=1}^{T}\Big[a_\tau V_1(\widetilde{Y}_t) + (a_{\mu_1} + \boldsymbol{X}_t^\top \boldsymbol{a}_{\boldsymbol{\beta},1})V_2(\widetilde{Y}_t) + (a_{\mu_2} + \boldsymbol{X}_t^\top \boldsymbol{a}_{\boldsymbol{\beta},2})V_3(\widetilde{Y}_t) + a_\sigma V_4(\widetilde{Y}_t)\Big] = 0 \quad \text{a.s.},$$

where $\boldsymbol{a}_{\boldsymbol{\beta},1}$ and $\boldsymbol{a}_{\boldsymbol{\beta},2}$ are the $\boldsymbol{\beta}$-coefficients in the score components that multiply V_2 and V_3 respectively (they coincide if the $\boldsymbol{\beta}_j$-block is not type-specific; the proof goes through with separate coefficients in either case). Varying $\widetilde{Y}_1$ alone forces the bracketed function of $\widetilde{Y}_1$ to equal a constant. By Lemma 4, the basis $\{1, V_1, V_2, V_3, V_4\}$ is linearly independent, so the coefficient on each V_j must equal zero, while the constant absorbs the only constant-function direction. Hence

$$a_\tau = 0, \quad a_\sigma = 0, \quad a_{\mu_1} + \boldsymbol{X}_1^\top \boldsymbol{a}_{\boldsymbol{\beta},1} = 0, \quad a_{\mu_2} + \boldsymbol{X}_1^\top \boldsymbol{a}_{\boldsymbol{\beta},2} = 0 \quad \text{a.s. } \boldsymbol{X}_1.$$

Writing the last two equations as $(\,a_{\mu_j}, \boldsymbol{a}_{\boldsymbol{\beta},j}^\top\,)\boldsymbol{U}_1 = 0$ a.s. for $j = 1, 2$ and invoking Assumption 4(b) (non-singularity of $\mathbb{E}[\boldsymbol{U}\boldsymbol{U}^\top]$) yields $a_{\mu_j} = 0$ and $\boldsymbol{a}_{\boldsymbol{\beta},j} = \boldsymbol{0}$. Hence $\boldsymbol{a}_1 = \boldsymbol{0}$.

Step 4 (Conclusion). Combining Steps 2 and 3, $\boldsymbol{a} = \boldsymbol{0}$, contradicting the hypothesis. Therefore $\boldsymbol{\mathcal{I}}$ is positive definite.

Quantitative rank certificate. Specialised to the no-covariate sub-system in the (μ, τ, σ)-block (i.e. $\boldsymbol{a}_{\boldsymbol{\beta}} = \boldsymbol{0}$ and $q = 0$), the proof produces a 14×14 linear system on the fourteen non-$\boldsymbol{\beta}$ coordinates of

$\boldsymbol{a}$. A choice of fourteen non-redundant linear functionals from Steps 2 and 3 yields a matrix whose determinant is

$$\det = \frac{1}{16\,\tau^*(1-\tau^*)},$$

strictly positive on $\tau^* \in (0, 1)$. A symbolic verification using SymPy is provided in *memo_keps2_nonsingularity_det.sympy.txt* (run *memo_keps2_nonsingularity_det.sympy.py*); evaluations at $\tau^* \in \{1/10, 1/4, 1/2, 3/4, 9/10\}$ give determinants $\{25/36, 1/3, 1/4, 1/3, 25/36\}$. □

Remark 3. *Lemma 2 converts the non-singularity of $\boldsymbol{I}$ for $K_\epsilon \geq 2$, currently maintained as Assumption 4(c), into an analytical consequence of Assumptions 1 (with $K_\epsilon = 2$), 3, 4 (a)(b)(d), and MOD-1. Propositions 4, 6, and 7 currently rely on the maintained hypothesis for $K_\epsilon = 2$; they therefore hold unconditionally for $K_\epsilon = 2$ under this strengthening. The strengthening is automatically satisfied in the empirical specifications of Section 7, where the input vector $\boldsymbol{x}_{it} = (k_{it}, \ell_{it}, v_{it})^\top$ admits a joint density positive on an open set across consecutive years. The proof extends to $K_\epsilon \geq 3$* mutatis mutandis *via the iterated phase-peeling argument of Remark 2.*

Lemma 5. *Suppose that the assumptions in Proposition 8 hold. If $-n^{-1}\log q_n = o(1)$, then $n^{-1}c^M_{1-q_n} = o(1)$.*

Proof. For brevity of notation, write $c_n = c^M_{1-q_n}$. By Theorem 2.1 of Foutz and Srivastava (1977), $PLR_n(M) \xrightarrow{d} \sum_{j=1}^K b_j\chi_j^2$ for $0 < b_j < \infty$ and K is finite, where χ_1^2, ..., χ_K^2 are independent chi-square random variables with one degree of freedom. Then, we have

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

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

□

Lemma 6. *Suppose that $g(\boldsymbol{w};\boldsymbol{\psi},\alpha)$ is defined as (33), where $\boldsymbol{\psi} = (\boldsymbol{\eta}^\top, \boldsymbol{\lambda}^\top)^\top$. Let g^*, ∇g^*, and $\nabla \log g^*$ denote $g(\boldsymbol{W};\boldsymbol{\psi},\alpha)$, $\nabla g(\boldsymbol{W};\boldsymbol{\psi},\alpha)$, and $\nabla \log g(\boldsymbol{W};\boldsymbol{\psi},\alpha)$ evaluated at $(\boldsymbol{\psi}^*, \alpha)$, respectively. Let ∇f^* denote $\nabla f(\boldsymbol{W};\boldsymbol{\theta}^*)$. The following statements hold.*

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

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

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

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

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

Proof of Lemma 6. Recall that

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

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

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

To show part (b), note that

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

For part (c), $\nabla_{\boldsymbol{\lambda}\boldsymbol{\lambda}^\top} \log g^* = \nabla_{\boldsymbol{\lambda}\boldsymbol{\lambda}^\top} g^*/g^* - (\nabla_{\boldsymbol{\lambda}} \log g^*)(\nabla_{\boldsymbol{\lambda}} \log g^*) = \alpha(1-\alpha)\nabla_{\boldsymbol{\lambda}\boldsymbol{\lambda}^\top} f^*/f^*$ because $\nabla_{\boldsymbol{\lambda}\boldsymbol{\lambda}^\top} g^*/g^* = \alpha(1-\alpha)\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^\top} f^*/f^*$ from part (a) and $\nabla_{\boldsymbol{\lambda}} \log g^* = \nabla_{\boldsymbol{\lambda}} g^*/g^* = 0$ from part (b).

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

$$\int \{\nabla_{\lambda_i\lambda_j} \log g(\boldsymbol{w};\psi,\alpha) + \nabla_{\lambda_i} \log g(\boldsymbol{w};\psi,\alpha)\nabla_{\lambda_j} \log g(\boldsymbol{w};\psi,\alpha)\} g(\boldsymbol{w};\psi,\alpha)dx = 0. \tag{99}$$

Evaluating (99) at $\psi = \psi^*$ in conjunction with part (a) gives the first equation in part (d). Differentiating (99) w.r.t. λ_k or η and evaluating at $\psi = \psi^*$ gives the latter two equations in part (d). Part (e) follows from differentiating (99) w.r.t. λ_k and λ_ℓ and evaluating at $\psi = \psi^*$ in conjunction with parts (a) and (d).

□

Lemma 7. *[Lemma 2.1 of Liu and Shao (2003)] Suppose $X_1, \ldots, X_n$ are i.i.d. random variables with $\max_{1\le i\le n} \mathbb{E}|X_i|^{q+\delta} < C$ for some $\delta > 0$, $q > 0$, and $C \in (0,\infty)$. Then, $\max_{1\le i\le n} |X_i| = o_p(n^{1/q})$.*

Proof. For any $\varepsilon > 0$, we have

$$P\left(\max_{1\le i\le n} |X_i| > \varepsilon n^{1/q}\right) \le \sum_{1\le i\le n} P(|X_i| > \varepsilon n^{1/q}) \le \varepsilon^{-q}\mathbb{E}(|X_1|^q \mathbb{I}\{|X_1| > \varepsilon n^{1/q}\})$$

by a version of the Markov inequality and the i.i.d. assumption. As $n \to \infty$, the right-hand side tends to 0 by the dominated convergence theorem. □

A.8 EM Algorithm

A.8.1 EM Algorithm for the model (1) with normal mixture density (2)–(3)

We introduce latent variables $D_i \in \{1, 2, \ldots, M\}$ and $C_{it} \in \{1, 2, \ldots, K\}$, where the top-level latent variable D_i indicates which of the M top-level components was chosen for observation i, and C_{it} indicates which of the K inner components was chosen for the t-th observation of unit i under the selected top-level component. Then, the complete-data log-likelihood is given as

$$\log L_c = \sum_{i=1}^{n} \sum_{j=1}^{M} \mathbb{I}(D_i = j) \left[\log \alpha_j + \sum_{t=1}^{T} \sum_{k=1}^{K_\epsilon} \mathbb{I}(C_{it} = k) \big(\log \tau_{jk} + \log \phi(Z_{jk,it}) - \log \sigma_j \big) \right],$$

where $Z_{jk,it} := \frac{Y_{it} - \mu_{jk} - \boldsymbol{x}_{it}^\top \boldsymbol{\beta}_j}{\sigma_j}$.

To obtain the penalised maximum a posteriori (MAP) estimator $\hat{\boldsymbol{\vartheta}} = \arg\max_{\boldsymbol{\vartheta} \in \bar{\Theta}(\boldsymbol{c})} \ell_n^{\text{pen}}(\boldsymbol{\vartheta})$, where $\ell_n^{\text{pen}} := \ell_n + p$ with penalty (76), we implement the following EM algorithm by alternating between the E-step and the M-step. The algorithm is initialized with multiple random starting values and iterated until the relative change in log-likelihood falls below 10^{-6} or 5,000 iterations are reached, whichever comes first.

In the E-step, given current estimates $(\alpha_j^{(m)}, \boldsymbol{\theta}_j^{(m)})$, compute the posterior probabilities of the latent variables as

$$\pi_{i,j}^{(m)} = P(D_i = j | \mathbf{W}_i, \{\alpha_j^{(m)}, \boldsymbol{\theta}_j^{(m)}\}) = \frac{\alpha_j^{(m)} f(\mathbf{W}_i; \theta_j^{(m)})}{\sum_{r=1}^{M} \alpha_r^{(m)} f(\mathbf{W}_i; \theta_r^{(m)})} \quad \text{and}$$

$$\gamma_{jk,it}^{(m)} = P(C_{it} = k | \mathbf{W}_i, D_i = j, \boldsymbol{\theta}_j^{(m)}) = \frac{\tau_{jk}^{(m)} \phi(z_{jk,it}^{(m)})}{\sum_{\ell=1}^{K_\epsilon} \tau_{j\ell}^{(m)} \phi(z_{j\ell,it}^{(m)})}$$

for each $j = 1, \ldots, M$ and $k = 1, \ldots, K$.

In the M-step, we maximise the expected complete-data penalised log-likelihood $E[\log L_c | \mathbf{W}, \alpha_j^{(m)}, \boldsymbol{\theta}_j^{(m)}] + p(\boldsymbol{\vartheta})$ as follows: for $j = 1, ..., M$,

1. Update α_j as
$$\alpha_j^{(m+1)} = \frac{\sum_{i=1}^{n} \pi_{i,j}^{(m)}}{n}, \quad \text{with } \sum_{j=1}^{M} \alpha_j^{(m+1)} = 1.$$

2. Update τ_{jk} as
$$\tau_{jk}^{(m+1)} = \frac{\sum_{i=1}^{n} \pi_{i,j}^{(m)} \sum_{t=1}^{T} \gamma_{jk,it}^{(m)}}{\sum_{i=1}^{n} \pi_{i,j}^{(m)} T} \quad \text{for } k = 1, ..., K_\epsilon.$$

3. Update $\boldsymbol{\mu}_j$ as
$$\mu_{jk}^{(m+1)} = \frac{\sum_{i,t} \pi_{i,j}^{(m)} \gamma_{jk,it}^{(m)} (Y_{it} - \mathbf{X}_{it}^\top \beta_j^{(m)})}{\sum_{i,t} \pi_{i,j}^{(m)} \gamma_{jk,it}^{(m)}} \quad \text{for } k = 1, ..., K.$$

4. Given $\boldsymbol{\mu}_j^{(m+1)}$, update $\boldsymbol{\beta}_j$ as

$$\boldsymbol{\beta}_j^{(m+1)} = \arg\min_{\boldsymbol{\beta}_j} \sum_{k=1}^{K_\epsilon} \sum_{i=1}^{n} \sum_{t=1}^{T} \pi_{i,j}^{(m)} \gamma_{jk,it}^{(m)} (Y_{it} - \mu_{jk}^{(m+1)} - \boldsymbol{x}_{it}^\top \boldsymbol{\beta}_j)^2.$$

5. Update σ_j^2

$$\sigma_j^{(m+1)2} = \frac{\sum_{i=1}^{n} \pi_{i,j}^{(m)} \sum_{t=1}^{T} \sum_{k=1}^{K_\epsilon} \gamma_{jk,it}^{(m)} (Y_{it} - \mu_{jk}^{(m+1)} - \boldsymbol{x}_{it}^\top \boldsymbol{\beta}_j)^2}{\sum_{i=1}^{n} \pi_{i,j}^{(m)} T}.$$

We iterate between the E-step and the M-step until convergence.

MAP modifications. Under the priors in Section 5.2, the M-step updates above are replaced by their MAP counterparts: (i) $\alpha_j^{(m+1)} \propto \sum_i \pi_{i,j}^{(m)} + a_\alpha - 1$, normalised to sum to one; (ii) $\tau_{jk}^{(m+1)} \propto \sum_{i,t} \pi_{i,j}^{(m)} \gamma_{jk,it}^{(m)} + a_\tau - 1$, normalised per component; (iii) $\sigma_j^{(m+1)2} = \mathrm{SS}_j / N_j^{\mathrm{eff}}$, subject to $\sigma_j^2 \geq \delta\, \hat{s}^2$. The standard MLE formulas above correspond to $a_\alpha = a_\tau = 1$.

A.9 Information Criteria: AIC and BIC

The consistency of BIC and inconsistency of AIC are stated in Section 5.3 of the main text; we provide formal proofs here under Assumptions A.3–A.4.

We may also estimate the number of components by the penalised maximum likelihood estimator

$$\widehat{M}_{PL} = \arg \max_{M \in \{1,2,\dots,\overline{M}\}} p\ell_n^M(\hat{\boldsymbol{\vartheta}}_M),$$

where

$$p\ell_n^M(\hat{\boldsymbol{\vartheta}}_M) := \ell_n^M(\hat{\boldsymbol{\vartheta}}_M) - p_{n,k_M},$$

and $\hat{\boldsymbol{\vartheta}}_M$ is the MLE defined by (19) while p_{n,k_M} is a model-selection penalization term with $k_M := \dim(\boldsymbol{\vartheta}_M)$ representing the number of estimated parameters for mixture models (1). (The model-selection penalty p_{n,k_M} is distinct from the log-prior penalty $p(\boldsymbol{\vartheta})$ in (76) used for EM regularisation; see Section 5.2.)

The following proposition states that the penalised maximum likelihood estimator of the number of components is consistent under the regularity condition.

Assumption A.3. *For all $n > 1$, the penalty function $p_{n,k}$ satisfies (a) $p_{n,k+1} \geq p_{n,k} > 0$, (b) $\lim_{n\to\infty} p_{n,k} = \infty$, (c) $p_{n,k} = o(n)$, and (d) $\lim_{n\to\infty} \frac{p_{n,k'}}{p_{n,k}} > 1$ for $k < k' \leq \overline{M}$.*

Assumption A.4. *For $M \in \{M_0+1, M_0+2, \dots, \overline{M}\}$, $\ell_n^M(\hat{\boldsymbol{\vartheta}}_M) - \ell_n^{M_0}(\hat{\boldsymbol{\vartheta}}_{M_0}) = O_p(1)$.*

Proposition A.5 (Consistent estimation by the maximum penalised likelihood estimator)**.** *Suppose that Assumptions 6-A.4 hold. Then, $\widehat{M}_{PL} \xrightarrow{p} M_0$.*

Assumption A.3 corresponds to Assumption (C1) of Keribin (2000), specifying conditions for penalty functions. The BIC penalty function, given by $p_{n,k} = \frac{k}{2}\log(n)$, satisfies this assumption, whereas the AIC penalty function, $p_{n,k} = k$, does not satisfy Assumption A.3(b).

Assumption A.4 prevents over-estimation. Appendix A.2 verifies it by extending the Section 5.2 higher-order expansion to $M_1 > M_0 + 1$: the key additional condition is Assumption A.2(b), which requires higher-order linear independence of the score derivatives $\nabla_{\theta_h^{\otimes p}} f/g_{M_0}$ for $p = 1, \ldots, p_h$ and $h = 1, \ldots, M_0$.

The following proposition formalizes the results derived in Appendix A.2.

Proposition A.6. *Suppose that Assumptions 1, 3, 6, A.3, and A.2 hold. Then, Assumption A.4 holds, and* $\widehat{M}_{PL} \xrightarrow{p} M_0$.

Corollary 1 (Consistency of BIC and Inconsistency of AIC)**.** *Suppose Assumptions 1, 3, 6, and A.2 hold. Then, (a) If* $p_{n,k} = \frac{k}{2}\log(n)$ *(BIC penalty), we have* $\widehat{M}_{PL} \xrightarrow{p} M_0$. *(b) If* $p_{n,k} = k$ *(AIC penalty), then* $\lim_{n\to\infty} \Pr(\widehat{M}_{PL} > M_0) > 0$.

Remark 4. *The higher-order linear independence condition in Assumption A.2(b) follows from the panel structure ($T \geq 2$) by the same argument as Proposition 2.*

Leroux (1992) established that, under conditions similar to Assumption A.3, the maximum penalised likelihood method produces an estimator that asymptotically does not underestimate the number of components. Building upon the work of Dacunha-Castelle and Gassiat (1999), Keribin (2000) derived regularity conditions necessary for the consistency of the maximum penalised likelihood estimator. Our Assumption A.2(b) corresponds to condition (P2) in Keribin (2000). Keribin's condition (P2) is not satisfied when M is moderately larger than M_0, as it relies on a second-order expansion of the density ratio for identification. Our Assumption A.2(b) instead accommodates settings where identification requires a higher-order rank condition, thereby extending the applicability of the consistency result.

A.10 Nonparametric estimation of the lower bound of the number of components by the rank test

We also estimate the lower bound of the number of components without imposing the parametric assumption on error distributions by extending a method proposed by Kasahara and Shimotsu (2014) which in turn is based on the rank test of Kleibergen and Paap (2006).

We partition the support of $Y_t \in \mathcal{Y}_t$ into $|\Delta_t|$ mutually exclusive and exhaustive subsets $\Delta_t = \{\delta_1^t, \ldots, \delta_{|\Delta_t|}^t\}$ so that $\mathcal{Y}_t = \cup_{i=1}^{|\Delta_t|}\delta_i$ and $\delta_i \cap \delta_j = \emptyset$ for $i \neq j$.

For each $k \in \mathcal{T} := \{1, 2, ..., T\}$, define $\mathcal{T}_{-k} := \{s \in \mathcal{T} : s \neq t\} = \{1, .., t-1, t+1, ..., T\}$ and let $\mathbf{Y}_{\mathcal{T}_{-k}} = (Y_1, .., Y_{k-1}, Y_{k_1}, ..., Y_T)^\top$ be a vector of Y_ts in the group $\mathcal{T}_{-k}$. Partition the support of $\mathbf{Y}_{\mathcal{T}_{-k}}$ into $|\Delta_{\mathcal{T}_{-k}}|$ mutually exclusive and exhaustive subsets $\Delta_{\mathcal{T}_{-k}} = \{\boldsymbol{\delta}_1^{\mathcal{T}_{-k}}, \ldots, \boldsymbol{\delta}_{|\Delta_{\mathcal{T}_{-k}}|}^{\mathcal{T}_{-k}}\}$. In particular, we construct this partition by unfolding the tensor product $\otimes_{t\neq k}\Delta_t$ using the Cartesian product denoted by $\odot$ as $\Delta_{\mathcal{T}_{-k}} = \odot_{t\neq k}\Delta_t = \Delta_1 \odot \cdots \odot \Delta_{k-1} \odot \Delta_{k+1} \odot \cdots \odot \Delta_T$ so that $|\Delta_{\mathcal{T}_{-k}}| = |\Delta_t|^{T-1}$.

For each $k \in \mathcal{T}$, we construct a $|\Delta_t| \times |\Delta_{\mathcal{T}_{-k}}|$ bivariate probability matrix $\boldsymbol{P}_k$ by arranging $\Pr(Y_k \in \delta_a, \boldsymbol{Y}_{-k} \in \boldsymbol{\delta}_b^{\mathcal{T}_{-k}})$ for partition level $(a, b) = (1, 1), \ldots, (|\Delta_t|, |\Delta_{\mathcal{T}_{-k}}|)$ as

$$\boldsymbol{P}_k = \begin{bmatrix} \Pr(Y_k \in \delta_1, \boldsymbol{Y}_{\mathcal{T}_{-k}} \in \boldsymbol{\delta}_1^{\mathcal{T}_{-k}}) & \cdots & \Pr(Y_k \in \delta_1, \boldsymbol{Y}_{\mathcal{T}_{-k}} \in \boldsymbol{\delta}_{|\Delta_{\mathcal{T}_{-k}}|}^{\mathcal{T}_{-k}}) \\ \vdots & \ddots & \vdots \\ \Pr(Y_k \in \delta_{|\Delta_t|}, \boldsymbol{Y}_{\mathcal{T}_{-k}} \in \boldsymbol{\delta}_1^{\mathcal{T}_{-k}}) & \cdots & \Pr(Y_k \in \delta_{|\Delta_t|}, \boldsymbol{Y}_{-k} \in \boldsymbol{\delta}_{|\Delta_{\mathcal{T}_{-k}}|}^{\mathcal{T}_{-k}}) \end{bmatrix}. \tag{100}$$

Collect the marginal probability distribution of Y_k and $\mathbf{Y}_{-k}$ over Δ_t and $\Delta_{\mathcal{T}_{-k}}$ conditional on being from the j-th component into a vector as

$$\begin{aligned} \boldsymbol{p}_k^j &: = (\Pr(Y_k \in \delta_1 | D = j), \ldots, \Pr(Y_k \in \delta_{|\Delta_t|} | D = j))^\top \quad \text{and} \\ \boldsymbol{q}_k^j &:= (\Pr(\mathbf{Y}_{-k} \in \boldsymbol{\delta}_1^{\mathcal{T}_{-k}} | D = j), \ldots, \Pr(\mathbf{Y}_{-k} \in \boldsymbol{\delta}_{|\Delta_{\mathcal{T}_{-k}}|}^{\mathcal{T}_{-k}} | D = j))^\top, \quad \text{respectively.} \end{aligned}$$

Then, under the conditional independence assumption as in the mixture model (1) but without imposing parametric restrictions, we may represent $\boldsymbol{P}_k$ as

$$\boldsymbol{P}_k = \sum_{j=1}^{M} \alpha^j \boldsymbol{p}_k^j (\boldsymbol{q}_k^j)^\top.$$

Kasahara and Shimotsu (2014) shows the rank of $\boldsymbol{P}_k$ identifies the lower bound of the number of components and develop a sequential hypothesis testing procedure for estimating the rank of $\boldsymbol{P}_k$ when the empirical quantile of the Y_t's are used to construct the partition. Because there are T possible ways to pick different k's out of $\{1, ..., T\}$, we test the maximum of the ranks of $\boldsymbol{P}_k$ across $k = 1, ..., T$.

We first develop a rk-statistic of Kleibergen and Paap (2006) for testing the null hypothesis of $\text{rank}(\boldsymbol{P}_k) = r$. Write the singular value decomposition of $\boldsymbol{P}_k$ as

$$\boldsymbol{P}_k = \mathbf{U}^k \mathbf{S}^k (\mathbf{V}^k)^\top = \begin{bmatrix} \mathbf{U}_{11}^k & \mathbf{U}_{12}^k \\ \mathbf{U}_{21}^k & \mathbf{U}_{22}^k \end{bmatrix} \begin{bmatrix} \mathbf{S}_1^k & 0 \\ 0 & \mathbf{S}_2^k \end{bmatrix} \begin{bmatrix} \mathbf{V}_{11}^k & \mathbf{V}_{12}^k \\ \mathbf{V}_{21}^k & \mathbf{V}_{22}^k \end{bmatrix}^\top,$$

where $\mathbf{U}^k$ is a $|\Delta_t| \times |\Delta_t|$ orthonormal matrix, $\mathbf{V}^k$ is a $|\Delta_{\mathcal{T}_{-k}}| \times |\Delta_{\mathcal{T}_{-k}}|$ orthonormal matrix, and $\mathbf{S}^k$ is a $|\Delta_t| \times |\Delta_{\mathcal{T}_{-k}}|$ diagonal matrix containing the singular values in decreasing order. In the partition of $\mathbf{U}^k$, $\mathbf{V}^k$, and $\mathbf{S}^k$ on the right-hand side, $\mathbf{U}_{11}^k$, $\mathbf{V}_{11}^k$, and $\mathbf{S}_1^k$ are $r \times r$, and the dimensions of the other submatrices are defined conformably. Then, the null hypothesis $\mathcal{H}_0 : \text{rank}(\mathbf{P}_k) = r$ is equivalent to $\mathcal{H}_0 : \mathbf{S}_2^k = 0$. The statistic of Kleibergen and Paap (2006) is based on an orthogonal transformation of $\mathbf{S}_2^k$ given by $\boldsymbol{\Lambda}_r^k = (\mathbf{A}_r^k)^\top \mathbf{P}_k \mathbf{B}_r^k$, where

$$\mathbf{A}_r^k = \begin{bmatrix} \mathbf{U}_{12}^k \\ \mathbf{U}_{22}^k \end{bmatrix} (\mathbf{U}_{22}^k)^{-1} (\mathbf{U}_{22}^k (\mathbf{U}_{22}^k)^\top)^{1/2} \quad \text{and} \quad \mathbf{B}_r^k = \begin{bmatrix} \mathbf{V}_{12}^k \\ \mathbf{V}_{22}^k \end{bmatrix} (\mathbf{V}_{22}^k)^{-1} (\mathbf{V}_{22}^k (\mathbf{V}_{22}^k)^\top)^{1/2}.$$

Let $\widehat{\boldsymbol{P}}_k$ be a sample analogue estimator for $\boldsymbol{P}_k$ for which we have $\sqrt{n}\text{vec}(\widehat{\boldsymbol{P}}_k - \boldsymbol{P}_k) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_k)$. The following proposition follows from Theorem 1 of Kleibergen and Paap (2006).

Proposition A.7. *Suppose that $\sqrt{n}vec(\widehat{\boldsymbol{P}}_k - \boldsymbol{P}_k) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_k)$ and that $\boldsymbol{\Omega}_r^k := ((\mathbf{B}_r^k)^\top \otimes (\mathbf{A}_r^k)^\top) \boldsymbol{\Sigma}_k (\mathbf{B}_r^k \otimes \mathbf{A}_r^k)$ is non-singular. If* $\text{rank}(\mathbf{P}_k) = r$*, then $\sqrt{n}\widehat{\boldsymbol{\lambda}}_r^k \rightarrow_d \mathcal{N}(0, \boldsymbol{\Omega}_r^k)$ as $n \to \infty$, where $\widehat{\boldsymbol{\lambda}}_r^k = vec((\widehat{\mathbf{A}}_r^k)^\top \widehat{\mathbf{P}}_k \widehat{\mathbf{B}}_r^k)$.*

Kleibergen and Paap (2006) proposed the statistic called the rk-statistic:

$$\text{rk}^k(r) = n(\widehat{\boldsymbol{\lambda}}_r^k)^\top (\widehat{\boldsymbol{\Omega}}_r^k)^{-1} \widehat{\boldsymbol{\lambda}}_r^k,$$

where $\widehat{\boldsymbol{\Omega}}_r^k$ is a consistent estimator for $\boldsymbol{\Omega}_r^k$. If the assumptions of proposition A.7 hold, $\text{rk}(r)$ converges in distribution to a $\chi^2((|\Delta_t| - r)(|\Delta_{\mathcal{T}_{-k}}| - r))$ random variable as $n \to \infty$.

When $T \geq 3$, we test the null hypothesis that $\text{rank}(\boldsymbol{P}_k) \leq r$ for each $k = 1, \ldots, T$. By selecting partitions such that $|\Delta_t| = r + 1$, we define the following ave- and max-rk test statistics:

$$\text{ave-rk}(r) = \frac{1}{T}\sum_{k=1}^{T} \text{rk}^k(r) \quad \text{and} \quad \text{max-rk}(r) = \max\{\text{rk}^1(r), ..., \text{rk}^T(r)\}.$$

See Section 3.4 of Kasahara and Shimotsu (2014) for the asymptotic distribution of these statistics.

In practice, we use the Bayesian bootstrap to obtain the bootstrap p-value for testing the rank, rejecting the null hypothesis $\text{rank}(\boldsymbol{P}_k) \leq r$ for all $k = 1, \ldots, T$ at significance level α if the bootstrap p-value is strictly less than α; see Appendix A.11. We estimate the lower bound for the number of components by applying the sequential hypothesis testing procedure described in Section 3.2 of Kasahara and Shimotsu (2014).

A.11 Bootstrap procedure for the ave- and max-rk statistic

We consider the following Bayesian bootstrap by drawing $\{\omega_i^{(b)}\}_{i=1}^n$ for $b = 1, ..., B$, where $\omega_i^{(b)} = \tilde{\omega}_i^{(b)}/\sum_{j=1}^n \tilde{\omega}_j^{(b)}$ with $\tilde{\omega}_i^{(b)} \overset{iid}{\sim} Exp(1)$ for $i = 1, ..., n$, and compute the bootstrap version of $\widehat{\boldsymbol{P}}_k$ as

$$\widehat{\boldsymbol{P}}_k^{(b)} = \begin{bmatrix} \sum_{i=1}^n \omega_i^{(b)} \mathbb{I}(Y_k \in \delta_1, \boldsymbol{Y}_{\mathcal{T}_{-k}} \in \boldsymbol{\delta}_1^{\mathcal{T}_{-k}}) & \cdots & \sum_{i=1}^n \omega_i^{(b)} \mathbb{I}(Y_k \in \delta_1, \boldsymbol{Y}_{\mathcal{T}_{-k}} \in \boldsymbol{\delta}_{|\Delta_{\mathcal{T}_{-k}}|}^{\mathcal{T}_{-k}}) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^n \omega_i^{(b)} \mathbb{I}(Y_k \in \delta_{|\Delta_t|}, \boldsymbol{Y}_{\mathcal{T}_{-k}} \in \boldsymbol{\delta}_1^{\mathcal{T}_{-k}}) & \cdots & \sum_{i=1}^n \omega_i^{(b)} \mathbb{I}(Y_k \in \delta_{|\Delta_t|}, \boldsymbol{Y}_{-k} \in \boldsymbol{\delta}_{|\Delta_{\mathcal{T}_{-k}}|}^{\mathcal{T}_{-k}}) \end{bmatrix}$$

for $k = 1, ..., T$. The same weights are used across different ks to accomodate the dependencies across ks, so that the bootstrapped correlation between $\widehat{\boldsymbol{P}}_k^{(b)}$ and $\widehat{\boldsymbol{P}}_\ell^{(b)}$ for $k \neq \ell$ correctly captures the corresponding sample correlation between $\widehat{\boldsymbol{P}}_k$ and $\widehat{\boldsymbol{P}}_\ell$.

With this bootstrapped $\{\widehat{\boldsymbol{P}}_k^{(b)}\}_{k=1}^T$ for $b = 1, ..., B$, we construct the bootstrapped rk-statistics as

$$\text{rk}^{k(b)}(r) = n(\widehat{\boldsymbol{\lambda}}_r^{k(b)} - \widehat{\boldsymbol{\lambda}}_r^k)^\top (\widehat{\boldsymbol{\Omega}}_r^{k(b)})^{-1} (\widehat{\boldsymbol{\lambda}}_r^{k(b)} - \widehat{\boldsymbol{\lambda}}_r^k) \quad \text{for } k = 1, ..., T, \tag{101}$$

where $\widehat{\boldsymbol{\lambda}}_r^{k(b)}$ and $\widehat{\boldsymbol{\Omega}}_r^{k(b)}$ are computed as described in Proposition A.7 but we compute $\widehat{\mathbf{A}}_r^k$ and $\widehat{\boldsymbol{\Sigma}}_k$ using $\widehat{\boldsymbol{P}}_k^{(b)}$ in place of $\widehat{\boldsymbol{P}}_k$.

Based on $\{\text{rk}^{k(b)}(r)\}_{k=1}^T$ derived as above, we can construct bootstrapped ave- and max-rk-statistics as:

$$\text{ave-rk}^{(b)}(r) = \frac{1}{T}\sum_{k=1}^{T} \text{rk}^{k(b)}(r) \;\text{ and }\; \text{max-rk}^{(b)}(r) = \max\{\text{rk}^{1(b)}(r), ..., \text{rk}^{T(b)}(r)\} \;\text{ for } b = 1, 2, ..., B.$$

We then compute the bootstrap p-value as the empirical proportion of the bootstrapped test statistics $\text{max-rk}^{(b)}(r)$ that equal or exceed the observed statistic $\text{max-rk}(r)$: bootstrap p-value $= \frac{1}{B}\sum_{b=1}^B \mathbb{I}\left(\text{max-rk}^{(b)}(r) \geq \text{max-rk}(r)\right)$. We reject the null hypothesis that $\text{rank}(\boldsymbol{P}_k) \leq r$ for all $k = 1, \ldots, T$ at significance level α if the bootstrap p-value is strictly less than α.

B Additional Tables and Figures

B.1 Controlled-DGP Validation: Size and Power

Table B1 presents size and power for testing $H_0\colon M_0 = 2$ vs $H_1\colon M_0 = 3$ under controlled DGPs with conditionally independent errors, based on 500 simulations with $T = 3$. The size DGP uses heterogeneous variances $(\sigma_1, \sigma_2) = (0.8, 1.2)$ to evaluate size under a realistic null where components differ in both location and scale. The power DGP uses homogeneous variances $(\sigma_1, \sigma_2, \sigma_3) = (1, 1, 1)$ to isolate the effect of mean separation on detection power without confounding from variance differences.

Table B1: Size and Power for testing $H_0 : M_0 = 2$ vs $H_1 : M_0 = 3$ at the 5% level under controlled DGPs

n	Component means	**ave-rk**	**max-rk**	**LR**	**AIC**	**BIC**
Panel A: Size. DGP is two-component, $(\sigma_1, \sigma_2) = (0.8, 1.2)$, $\alpha = 0.5$						
200	$(\mu_1, \mu_2) = (-1, 1)$	3.4	6.0	4.6	9.6	0.0
200	$(\mu_1, \mu_2) = (-0.5, 0.5)$	0.2	0.2	4.6	7.2	0.0
400	$(\mu_1, \mu_2) = (-1, 1)$	5.8	6.4	4.2	7.6	0.0
400	$(\mu_1, \mu_2) = (-0.5, 0.5)$	0.0	0.0	4.6	6.8	0.0
Panel B: Power. DGP is three-component, $\sigma_j = 1$, $\alpha_j = 1/3$						
200	$(\mu_1, \mu_2, \mu_3) = (-0.5, 0, 1.5)$	0.6	0.8	12.0	15.4	0.2
200	$(\mu_1, \mu_2, \mu_3) = (-1.5, 0, 1.5)$	50.8	37.2	99.8	88.6	85.6
400	$(\mu_1, \mu_2, \mu_3) = (-0.5, 0, 1.5)$	3.0	3.2	14.4	23.4	0.0
400	$(\mu_1, \mu_2, \mu_3) = (-1.5, 0, 1.5)$	92.8	81.2	99.8	89.2	100.0

Notes: 500 replications, $T = 3$, conditionally independent normal errors (model 1 with (2)–(3)). Rejection frequencies (%) at 5% level; AIC/BIC report % selecting $M = 3$ over $M = 2$.

Size (Panel A): LRT maintains near-nominal 5% rejection rates (4.2–4.6%) across all scenarios. Rank tests (ave-rk, max-rk) achieve nominal size only with large samples ($n = 400$) and well-separated means, but severely underperform with close means. AIC overestimates (7–10%), while BIC remains conservative (0%).

Power (Panel B): LRT achieves near-100% power with well-separated means ($\mu = (-1.5, 0, 1.5)$) and maintains superior power (12–14%) even with close means, vastly outperforming rank tests (0–3%). BIC matches LRT under clear separation but becomes overly conservative as means converge. AIC consistently overestimates across all scenarios.

B.2 Additional Empirical Results

This appendix reports supplementary results for the empirical analysis in Section 7 of the main paper. Section B.2.1 documents input winsorisation robustness and also serves as the consolidated reference for CRE augmentation and Normal vs. Mixture error comparisons (Table B2). Section B.2.2 reports BIC-selected $\hat{M}$ under a three-component within-type error (K_ϵ=3). Section B.2.3 reports per-type Step 1 coefficients and a cross-panel elasticity comparison for the T=5 panel. Section B.2.4 gives the full per-type intercepts and output elasticities at T=10. Section B.2.6 analyses the misclassification bias arising from the two-step classify-then-GNR procedure. Section B.2.7 explains why hard plant-to-type assignment is not used.

B.2.1 Input Winsorisation Robustness

The baseline in the main paper uses the *raw (no-winsor)* panel. This section documents a robustness check where we winsorise the centred log inputs ($\tilde{k}_{it}, \tilde{\ell}_{it}, \tilde{m}_{it}$) at the 5/95 percentile (across the pooled panel; thresholds fixed pre-bootstrap) and re-run the full pipeline (model selection and Phase 2 GNR) on the winsorised data. A small number of extreme plants exert disproportionate leverage on the per-type GNR Step 2 GMM fit; winsorisation assesses whether these observations drive the baseline $\hat{M}$ selections.[2] Table B2 covers all 16 specification cells (three functional forms $\times$ Normal/Mixture errors $\times$ with/without CRE) and reports both LRT-selected and BIC-selected $\hat{M}$ under the raw data, alongside LRT-selected $\hat{M}$ under the winsorised data.

Effect on $\hat{M}$ selection. Table B2 reports $\hat{M}$ for the no-winsor and winsor 5/95 baselines across all 16 specifications, with BIC shown alongside LRT for the no-winsor case.

[2] A tighter 1/99 threshold was not run: the full 16-specification grid (B=199) takes 18–24 hours per pass, and doubling the cost is not justified given that 5/95 winsorisation already shifts $\hat{M}$ by at most two from the no-winsor baseline in most cells.

Table B2: Sequential bootstrap LRT and BIC selected $\hat{M}$, no-winsor vs. 5/95% winsorised data, with K_ϵ=3 BIC robustness rows. BIC uses $-2\hat{\ell} + k \log N$ (N = plants). Bold rows: preferred Mixture specification at each flexibility level.

Panel A: T = 5

			No winsor						**Winsor 5/95**					
			LRT			*BIC*			*LRT*			*BIC*		
Spec	**Error**	**CRE**	**Me**	**Fo**	**Tx**	**Me**	**Fo**	**Tx**	**Me**	**Fo**	**Tx**	**Me**	**Fo**	**Tx**
CD (q=0)	Normal	–	8	9	9	9	9	9	3	4	4	7	9	10
CD (q=0)	Normal	✓	5	6	5	5	7	6	4	5	4	6	9	8
CD (q=0)	**Mixture**	–	**3**	**4**	**4**	**7**	**7**	**6**	**4**	**5**	**4**	**6**	**9**	**8**
CD (q=0)	**Mixture**	✓	**1**	**4**	**1**	**6**	**7**	**7**	**2**	**5**	**4**	**1**	**8**	**6**
Translog (linear)	Normal	–	7	8	6	9	10	7	2	5	4	4	6	4
Translog (linear)	Normal	✓	7	7	6	7	9	6	3	2	2	3	10	4
Translog (linear)	**Mixture**	–	**2**	**2**	**4**	**6**	**4**	**4**	**3**	**2**	**3**	**6**	**8**	**4**
Translog (linear)	**Mixture**	✓	**2**	**3**	**4**	**4**	**3**	**5**	**2**	**2**	**2**	**3**	**9**	**6**
Translog (exact)	Normal	–	2	8	2	10	10	6	4	6	5	8	10	7
Translog (exact)	Normal	✓	2	5	2	8	10	8	4	4	5	7	10	9
Translog (exact)	**Mixture**	–	**4**	**6**	**4**	**5**	**9**	**6**	**4**	**5**	**4**	**5**	**9**	**4**
Translog (exact)	**Mixture**	✓	**2**	**3**	**3**	**8**	**7**	**5**	**3**	**4**	**2**	**6**	**7**	**5**

Panel B: T = 10

			No winsor						**Winsor 5/95**					
			LRT			*BIC*			*LRT*			*BIC*		
Spec	**Error**	**CRE**	**Me**	**Fo**	**Tx**	**Me**	**Fo**	**Tx**	**Me**	**Fo**	**Tx**	**Me**	**Fo**	**Tx**
CD (q=0)	Normal	–	8	9	8	8	10	8	8	4	5	8	10	10
CD (q=0)	Normal	✓	4	7	5	6	8	6	7	4	3	8	8	9
CD (q=0)	**Mixture**	–	**2**	**6**	**4**	**8**	**10**	**8**	**3**	**3**	**5**	**7**	**9**	**8**
CD (q=0)	**Mixture**	✓	**4**	**4**	**3**	**6**	**8**	**5**	**2**	**3**	**3**	**5**	**8**	**6**
Translog (linear)	Normal	–	7	9	7	10	10	8	5	5	4	6	5	6
Translog (linear)	Normal	✓	5	9	6	8	10	7	3	2	2	5	7	8
Translog (linear)	**Mixture**	–	**2**	**2**	**3**	**7**	**9**	**7**	**5**	**3**	**3**	**7**	**7**	**8**
Translog (linear)	**Mixture**	✓	**2**	**3**	**3**	**4**	**8**	**5**	**2**	**3**	**3**	**5**	**8**	**5**
Translog (exact)	Normal	–	4	6	2	9	10	10	4	4	8	10	10	10
Translog (exact)	Normal	✓	4	4	2	9	10	8	8	4	10	10	10	9
Translog (exact)	**Mixture**	–	**4**	**5**	**2**	**7**	**10**	**8**	**3**	**3**	**4**	**8**	**10**	**6**
Translog (exact)	**Mixture**	✓	**6**	**4**	**2**	**6**	**10**	**9**	**5**	**4**	**3**	**9**	**10**	**10**
CD (q=0)	Mixture (K_ϵ=3)	–	–	–	–	6	9	7	–	–	–	–	–	–
Translog (exact)	Mixture (K_ϵ=3)	–	–	–	–	6	9	7	–	–	–	–	–	–

Notes. LRT: sequential bootstrap LRT at 5% significance (B=199 replications). BIC: argmin over M=1, . . . , 10 of $-2\hat{\ell} + k \log N$, where N is the number of plants and k the number of free parameters; requires no bootstrap simulation (Leroux, 1992). BIC consistently selects larger $\hat{M}$ than LRT (typically by 2–6 units) because the plant-count penalty $\log N$ with $N \approx$ 157–290 is small in finite samples. Both criteria confirm $\hat{M} \geq 2$ across all specifications, establishing the existence of technology heterogeneity. Same EM pipeline throughout (H_0 unconstrained; H_1 Dirichlet penalty $a_\alpha \sum_j \log \alpha_j + a_\tau \sum_{j,k} \log \tau_{jk}$ with a_α=1.5, a_τ=1.1, variance floor $0.01\widehat{\mathrm{Var}}(y)$). Spec labels: "Translog (linear)" = first-order Taylor approximation (q=3); "Translog (exact)" = nonlinear FOC (q=3). K_ϵ=3 rows are no-CRE, no-winsor, BIC-only robustness checks; the sequential LRT and winsorised variants are not run for those rows.

Effect on Pool M=1 elasticity CIs. Winsorisation fixes three borderline zero-coverage cases in the Pool M=1 capital elasticity: (i) Metal T=10 Cobb-Douglas ($[-0.017, +0.108] \rightarrow [+0.016, +0.120]$); (ii) Textiles T=5 Python GNR ($[-0.014, +0.160] \rightarrow [+0.031, +0.163]$); (iii) Metal T=10 Python GNR improves from -0.092 to -0.015 on the lower bound ($[-0.092, +0.278] \rightarrow [-0.015, +0.278]$), though still nominally containing zero. Labour elasticity CIs are negligibly affected, consistent with labour being well-identified regardless of input-tail outliers.

B.2.2 Three-Component Within-Type Error Robustness

Table B2 reports the K_ϵ=3 robustness check in the same number-selection table as the baseline rows. The rerun uses BIC only, no CRE, no winsorisation, and the same Dirichlet penalty $a_\alpha \sum_j \log \alpha_j + a_\tau \sum_{j,k} \log \tau_{jk}$ (a_α=1.5, a_τ=1.1, σ^2 floor $0.01\widehat{\mathrm{Var}}(y)$). For both Cobb-Douglas and Translog (exact), BIC selects Metal/Food/Textiles = 6/9/7. This is larger than the sequential LRT K_ϵ=2 baseline, reflecting both the criterion gap and the richer within-type error distribution, but the qualitative conclusion that all three industries contain multiple latent technology types is unchanged.

B.2.3 Short Panel Robustness: $T = 5$ Results

The main results focus on the T=10 panel (1987–1996), which provides a longer earnings window for the cubic-Markov GMM Step 2. Here we report per-type Step 1 coefficients and a cross-panel elasticity comparison for the T=5 panel (1992–1996). This panel has larger samples (N_{Food}=862, N_{Metal}=242, N_{Tex}=203) but a shorter time dimension. Model selection at T=5 is shown in Table B2 (Panel A, no-winsor columns).

$T = 5$ vs. $T = 10$ model selection. The selected $\hat{M}$ values under the Translog (exact) Mixture specification (no CRE) are Metal/Food/Textiles = 4/6/4 at T=5, compared with 4/5/2 at T=10. The T=5 Food and Textiles selections are higher than their T=10 counterparts, consistent with two mechanisms. First, the larger T=5 samples (862 vs. 669 firms for Food, a 29% increase) give the LRT more power to detect additional types. Second, the T=5 sub-panel (1992–1996) covers a narrower time window and may be more heterogeneous. The T=10 panel (1987–1996) pools a longer growth trajectory and may merge types that are distinct only in the early-to-mid 1990s transition period. Under CRE at T=5, the Translog (exact) Mixture specification selects $\hat{M}$=2 for Metal, $\hat{M}$=3 for Food, and $\hat{M}$=3 for Textiles. The convergence of CRE selections relative to no-CRE suggests that some of the no-CRE $\hat{M}$ is driven by unabsorbed capital-productivity correlation.

B.2.4 Per-Type Translog Coefficients and Elasticities at $T = 10$

Table B3 reports the per-type Translog (exact) production-function coefficients, output elasticities, returns to scale, and Step 2 AR(1) productivity persistence under the headline Translog (exact, K_ϵ=2), no-CRE no-winsor baseline at T=10. The table is organized by industry subpanels, with one row for each latent type. Each type is reported in two rows: the first row gives the point estimates, and the second row gives the corresponding firm-cluster nonparametric-bootstrap confidence intervals. Types are ranked in ascending order of $\varepsilon_{M,j}$. The empirical second-stage post-selection bootstrap line is no longer used; uncertainty is computed by resampling firms, re-estimating the fixed-$\hat{M}$ first-stage mixture, matching bootstrap clusters to the baseline headline types by maximum firm-overlap, and recomputing the per-type GNR estimates within each bootstrap draw.

Table B3: Combined per-type Translog Exact GNR estimates, T=10, no-CRE no-winsor headline baseline. Each type is reported in two rows: point estimates followed by centered 95% firm-cluster bootstrap confidence intervals.

Panel A: Translog Exact production-function coefficients

Type ($\hat{\alpha}$, N)	**Statistic**	$\beta_{K,j}$	$\beta_{L,j}$	$\beta_{M,j}$	$\beta_{KK,j}$	$\beta_{LL,j}$	$\beta_{MM,j}$	$\beta_{KL,j}$	$\beta_{KM,j}$	$\beta_{LM,j}$
Food CIIU 311, T=10**,** N=669**,** $\hat{M}$=5										
T1 ($\hat{\alpha}$=0.078, N=61)	Point estimate	0.118	0.341	0.603	0.014	-0.033	0.044	0.035	-0.018	-0.063
	95% CI	[0.056, 0.189]	[0.198, 0.597]	[0.502, 0.670]	[-0.018, 0.039]	[-0.182, 0.086]	[-0.016, 0.182]	[-0.018, 0.082]	[-0.066, 0.007]	[-0.151, 0.022]
T2 ($\hat{\alpha}$=0.307, N=152)	Point estimate	0.166	0.182	0.668	0.036	0.006	0.048	0.034	-0.047	-0.040
	95% CI	[0.112, 0.303]	[0.044, 0.302]	[0.611, 0.735]	[0.014, 0.087]	[-0.072, 0.104]	[-0.021, 0.122]	[0.002, 0.066]	[-0.128, -0.030]	[-0.097, 0.039]
T3 ($\hat{\alpha}$=0.124, N=166)	Point estimate	0.152	0.236	0.685	0.020	0.068	0.171	0.032	-0.059	-0.092
	95% CI	[0.071, 0.259]	[0.120, 0.444]	[0.617, 0.745]	[-0.015, 0.051]	[-0.031, 0.229]	[0.110, 0.224]	[0.013, 0.062]	[-0.102, -0.010]	[-0.176, -0.031]
T4 ($\hat{\alpha}$=0.225, N=137)	Point estimate	0.182	0.172	0.706	0.075	0.047	0.144	0.066	-0.092	-0.089
	95% CI	[0.126, 0.259]	[0.052, 0.244]	[0.663, 0.755]	[0.036, 0.126]	[-0.015, 0.107]	[0.093, 0.196]	[0.015, 0.092]	[-0.140, -0.051]	[-0.128, -0.004]
T5 ($\hat{\alpha}$=0.266, N=153)	Point estimate	0.140	0.141	0.757	0.026	-0.032	0.043	0.037	-0.037	-0.023
	95% CI	[0.101, 0.243]	[0.070, 0.253]	[0.734, 0.775]	[0.008, 0.064]	[-0.080, 0.044]	[0.028, 0.101]	[0.016, 0.061]	[-0.108, -0.023]	[-0.077, 0.010]
Metal CIIU 381, T=10**,** N=157**,** $\hat{M}$=4										
T1 ($\hat{\alpha}$=0.306, N=48)	Point estimate	0.157	0.342	0.472	0.034	0.141	0.157	-0.026	-0.049	-0.097
	95% CI	[0.076, 0.205]	[0.215, 0.482]	[0.445, 0.503]	[-0.001, 0.056]	[-0.024, 0.244]	[0.138, 0.184]	[-0.063, 0.055]	[-0.070, -0.020]	[-0.197, -0.055]
T2 ($\hat{\alpha}$=0.083, N=12)	Point estimate	0.120	0.601	0.543	-0.019	0.066	0.071	-0.079	0.073	-0.107
	95% CI	[0.029, 10.024]	[0.071, 4.302]	[0.470, 0.679]	[-0.114, 1.760]	[-0.828, 1.178]	[0.026, 0.265]	[-0.187, 0.518]	[0.006, 0.105]	[-0.399, -0.034]
T3 ($\hat{\alpha}$=0.276, N=43)	Point estimate	0.193	0.460	0.566	0.052	0.344	0.223	0.075	-0.107	-0.226
	95% CI	[0.112, 0.295]	[0.313, 0.588]	[0.527, 0.606]	[-0.007, 0.113]	[0.092, 0.549]	[0.173, 0.287]	[0.012, 0.144]	[-0.158, -0.066]	[-0.301, -0.150]
T4 ($\hat{\alpha}$=0.336, N=54)	Point estimate	0.152	0.292	0.569	0.056	0.228	0.166	0.022	-0.060	-0.121
	95% CI	[0.092, 0.200]	[0.215, 0.426]	[0.546, 0.595]	[0.004, 0.085]	[0.091, 0.366]	[0.146, 0.191]	[-0.040, 0.072]	[-0.082, -0.032]	[-0.166, -0.094]
Textiles CIIU 321, T=10**,** N=162**,** $\hat{M}$=2										
T1 ($\hat{\alpha}$=0.555, N=74)	Point estimate	0.222	0.279	0.514	0.024	0.136	0.189	-0.004	-0.075	-0.130
	95% CI	[0.182, 0.281]	[0.218, 0.335]	[0.486, 0.541]	[0.007, 0.050]	[0.056, 0.191]	[0.165, 0.215]	[-0.033, 0.030]	[-0.095, -0.062]	[-0.159, -0.099]
T2 ($\hat{\alpha}$=0.445, N=88)	Point estimate	0.152	0.244	0.653	0.039	0.092	0.159	0.011	-0.068	-0.106
	95% CI	[0.116, 0.187]	[0.193, 0.298]	[0.621, 0.701]	[0.018, 0.057]	[0.029, 0.161]	[0.124, 0.208]	[-0.022, 0.041]	[-0.086, -0.050]	[-0.159, -0.070]

Table B3: Combined per-type Translog Exact GNR estimates, continued.

Panel B: Elasticities, returns to scale, and productivity persistence

Type ($\hat{\alpha}$, N)	**Statistic**	$\varepsilon_{K,j}$	$\varepsilon_{L,j}$	$\varepsilon_{M,j}$	RTS_j	$\hat{\rho}_j$
Food CIIU 311, T=10, N=669, $\hat{M}$=5						
T1 ($\hat{\alpha}$=0.078, N=61)	Point estimate	0.123	0.313	0.599	1.035	0.752
	95% CI	[0.056, 0.193]	[0.183, 0.510]	[0.525, 0.664]	[0.930, 1.177]	[0.577, 0.834]
T2 ($\hat{\alpha}$=0.307, N=152)	Point estimate	0.161	0.182	0.680	1.023	0.758
	95% CI	[0.108, 0.292]	[0.040, 0.307]	[0.633, 0.728]	[0.948, 1.146]	[0.635, 0.850]
T3 ($\hat{\alpha}$=0.124, N=166)	Point estimate	0.151	0.234	0.688	1.073	0.734
	95% CI	[0.066, 0.260]	[0.104, 0.459]	[0.646, 0.734]	[0.933, 1.315]	[0.515, 0.920]
T4 ($\hat{\alpha}$=0.225, N=137)	Point estimate	0.188	0.177	0.699	1.064	0.660
	95% CI	[0.127, 0.267]	[0.045, 0.261]	[0.665, 0.738]	[0.985, 1.199]	[0.532, 0.842]
T5 ($\hat{\alpha}$=0.266, N=153)	Point estimate	0.146	0.150	0.748	1.045	0.811
	95% CI	[0.106, 0.242]	[0.081, 0.281]	[0.714, 0.765]	[0.994, 1.148]	[0.639, 0.898]
Metal CIIU 381, T=10, N=157, $\hat{M}$=4						
T1 ($\hat{\alpha}$=0.306, N=48)	Point estimate	0.146	0.347	0.488	0.981	0.641
	95% CI	[0.075, 0.195]	[0.213, 0.478]	[0.448, 0.536]	[0.890, 1.070]	[0.534, 0.738]
T2 ($\hat{\alpha}$=0.083, N=12)	Point estimate	0.090	0.647	0.516	1.253	0.883
	95% CI	[0.022, 9.143]	[0.126, 3.359]	[0.404, 0.601]	[0.797, 12.846]	[0.632, 0.982]
T3 ($\hat{\alpha}$=0.276, N=43)	Point estimate	0.210	0.482	0.540	1.231	0.516
	95% CI	[0.120, 0.308]	[0.353, 0.601]	[0.496, 0.587]	[1.153, 1.314]	[0.390, 0.618]
T4 ($\hat{\alpha}$=0.336, N=54)	Point estimate	0.155	0.277	0.593	1.025	0.707
	95% CI	[0.091, 0.207]	[0.203, 0.413]	[0.560, 0.630]	[0.971, 1.132]	[0.524, 0.972]
Textiles CIIU 321, T=10, N=162, $\hat{M}$=2						
T1 ($\hat{\alpha}$=0.555, N=74)	Point estimate	0.221	0.292	0.560	1.073	0.709
	95% CI	[0.180, 0.278]	[0.239, 0.342]	[0.517, 0.602]	[1.021, 1.116]	[0.576, 0.825]
T2 ($\hat{\alpha}$=0.445, N=88)	Point estimate	0.170	0.244	0.617	1.031	0.627
	95% CI	[0.134, 0.205]	[0.178, 0.305]	[0.584, 0.652]	[0.986, 1.094]	[0.512, 0.737]

Notes. Types are ranked in ascending order of $\varepsilon_{M,j}$. The reported coefficients use the conventional centered Translog representation; the squared-input coefficients follow the one-half convention in the production function. Inputs are centered at the industry grand means. $\beta_{M,j}$, $\beta_{KM,j}$, $\beta_{LM,j}$, and $\beta_{MM,j}$ come from the GNR Step 1 materials-share equation; the remaining production-function coefficients and $\hat{\rho}_j$ come from GNR Step 2. Confidence intervals are centered 95% firm-cluster nonparametric-bootstrap percentile spreads (B=500), resampling firms, re-estimating the fixed-$\hat{M}$ first-stage mixture, matching bootstrap clusters to the baseline headline types by maximum firm-overlap, and recomputing the per-type GNR estimates.

Type heterogeneity is concentrated in $\beta_{M,j}$: in the clean updated second-stage point estimates, Food spans $\beta_{M,j} \in [0.603, 0.757]$, Metal spans $[0.472, 0.569]$, and Textiles spans $[0.514, 0.653]$. The materials elasticities $\varepsilon_{M,j}$ closely track the materials coefficients, confirming that variation in $\beta_{M,j}$ drives type differentiation. Capital elasticities $\varepsilon_{K,j}$ play a secondary role in type differentiation.

B.2.5 Sensitivity to Regularisation Penalty Strength

Table B4 examines how model-selection accuracy responds to the strength of the Dirichlet soft penalties used during EM estimation. The baseline specification uses $a_\tau = 1.1$ (penalty on τ_{jk}), $a_\alpha = 1.5$ (penalty on α_j), and $\sigma_j^2 \geq 0.01\hat{s}^2$; each row perturbs one penalty parameter while holding the others fixed. The final row imposes a common within-type variance ($\sigma_j^2 = \sigma^2$ for all j) in the Normal specification. The DGP is calibrated to Chilean fabricated metal products (CIIU 381) with $M_0 = 2$ and $T = 5$. BIC achieves 100% correct selection across all configurations. LRT is most sensitive to the τ_{jk} penalty strength, with the baseline $a_\tau = 1.1$ providing good finite-sample performance.

Table B4: Sensitivity of model-selection accuracy to Dirichlet penalty strength ($M_0 = 2$, Metal industry, $K_\epsilon = 2$, $T = 5$; computation in progress)

	Normal DGP				**Mixture DGP**			
Configuration	**BIC(N)**	**LR(N)**	**BIC(Mix)**	**LR(Mix)**	**BIC(N)**	**LR(N)**	**BIC(Mix)**	**LR(Mix)**
Results pending—rerun gen_table_sensitivity.py								

B.2.6 Misclassification Bias and the Two-Step Estimation Procedure

The empirical analysis recovers type-specific production-function parameters by a two-step plug-in: (i) fit the inside-log nonlinear share-equation EM mixture and MAP-assign each plant to $\hat{j}_i \in \{1, \ldots, \hat{M}\}$ at the fitted posterior; (ii) run the GNR Step 1 and Step 2 estimators on each type-conditional sub-sample. Because the Chilean panel is short (T=5 or T=10), MAP classification is imperfect *even in the limit* $n \to \infty$. Misclassified plants contaminate the per-type sub-samples and the per-type GNR estimator converges to a pseudo-true value $\psi_j^*(T) \neq \psi_j^0$.

Exponential bias bound. Let $P_{kj}(T) = \Pr(\hat{j} = j \mid \text{type} = k)$ be the population MAP confusion matrix. Under regularity (Gaussian errors, bounded inputs, identified per-type moment), the bias satisfies

$$\|\psi_j^*(T) - \psi_j^0\| \;\leq\; C_j \cdot e^{-T\rho_j}, \qquad \rho_j = \min_{k \neq j} \bar{C}_{kj},$$

with $\bar{C}_{kj}$ the path-averaged Chernoff information between the share-equation densities of types k and j, and C_j scaling with GNR instrument strength and the type-parameter contrast.

Step-1 vs. Step-2 asymmetry. The classifier observes only the materials-share equation. A misclassified plant therefore resembles the target type on Step 1 (materials elasticity) but carries alien output-side structure (capital, labour, persistence) not seen by the classifier. Monte-Carlo simulations show that at moderate separation (n=277, M_0=2, misclassification rate 18.7% at T=5), the

Step-2 persistence bias $|\Delta\rho|^{\mathrm{MAP}}$=0.042 exceeds the Step-1 materials bias $|\Delta\varepsilon_M|^{\mathrm{MAP}}$=0.0025 by an order of magnitude. The Step-2 capital elasticity bias ($|\Delta\varepsilon_K|^{\mathrm{MAP}}$=0.012 at T=5, falling to 0.002 at T=20) is intermediate. This asymmetry sharpens the "classification is exogenous to the production-function parameters" claim: it holds as $T \to \infty$ but fails at fixed T, with Step-2 parameters carrying substantially more fixed-T bias than Step-1.

Empirical separation. Pairwise path-averaged Chernoff informations estimated from the inside-log nonlinear mixture fits give worst-pair indices $e^{-5\hat{\rho}_j}$ of 17–25% at T=5 across the three industries, matched by the Monte-Carlo actual rate of 18.7% at comparable separation. At T=10 the index falls to 3–6%, so the bias is substantially reduced in the main results.

Remedies. The cleanest fix is joint estimation: a single EM treating type membership as latent throughout, eliminating the fixed-T misclassification bias by construction. The two-step approach used here is retained for comparability with earlier industry-level applications; in the current empirical update we report the corresponding second-stage point estimates.

B.2.7 Hard Plant-to-Type Assignment (Binarization)

An alternative is to assign each plant to a single type by MAP classification ("binarization") and run ordinary GNR on each hard-assigned sub-sample. We do not adopt binarization for three reasons: (i) it discards the signal in plants with non-trivial posterior probability for a second type, conflicting with GNR's identification from the full conditional distribution of productivity (Gandhi et al., 2020); (ii) it amplifies the fixed-T misclassification bias $C_j\, e^{-T\rho_j}$ (Section B.2.6) across every Step-1 and Step-2 moment condition rather than smoothing it via posterior weighting; and (iii) with $\hat{M} \geq 3$ types, discrete membership counts make small-type estimates sensitive to the exact MAP partition. The two-step plug-in used here occupies an intermediate position: it uses MAP assignment but applies GNR to the resulting sub-sample as if it were a clean partition, with the resulting bias bounded as in Section B.2.6. For the T=10 panel, the exponential bias bound yields effective misclassification rates below 6%. This is negligible relative to the magnitude of the cross-type parameter contrasts reported in Table B3.

References

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

[2] Andrews, D. W. K. (1994). Empirical process methods in econometrics. In *Handbook of Econometrics*, volume 4, pages 2247–2794. North-Holland, Amsterdam.

[3] Andrews, D. W. K. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 69:683–734.

[4] 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.

[5] 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.

[6] 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.

[7] Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, 13(2):795–800.

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

[9] 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.

[10] 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.

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

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

[13] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27(4):887–906.

[14] Kleibergen, F. and Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, 133(1):97–126.

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

[16] Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20:1350–1360.

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

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

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

[20] Teicher, H. (1963). Identifiability of finite mixtures. *Annals of Mathematical Statistics*, 34(4):1265–1269.

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

[22] Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214.