Panel Stochastic Frontier Models with Latent Group Structures*

Kazuki Tomioka†
Hiroshima University

Thomas T. Yang‡
Australian National University

Xibin Zhang§
Monash University

April 9, 2026

Abstract

Stochastic frontier models have attracted considerable attention due to the incorporation of an inefficiency term in addition to the conventional error term. In this paper, we propose a general estimation framework for panel stochastic frontier models that accommodates potential heterogeneity through latent group structures. The framework is tailored to the distinctive features of stochastic frontier models and is paired with a practical hybrid estimation procedure that combines individual-level and joint panel estimation. We illustrate the estimation framework using a panel stochastic frontier model that treats the inefficiency term as a random effect, and show that it can be readily extended to a range of fixed effects specifications common in the literature. Simulation studies indicate strong finite-sample performance, and we further demonstrate the practicality of the approach in an empirical application to the cost efficiency of the U.S. commercial banking sector.

Keywords: Classification, Group Structures, Panel Data, Stochastic Frontier

JEL classification: C23, C33, C38, C51

*We thank the editor, Michal Kolesár, an Associate Editor, and two anonymous referees for their helpful comments. We are grateful for Bin Peng, Valentin Zelenyuk, and the participants of the Econometric Society Australasian Meeting 2024 and AE^2 conference 2025 for their helpful feedback.

†Department of Economics, Hiroshima University, Hiroshima, Japan. Email: kazuki.tomioka@anu.edu.au.

‡Corresponding author. Research School of Economics, The Australian National University, Canberra, ACT 2601, Australia. Email: tao.yang@anu.edu.au.

§Department of Econometrics and Business Statistics, Monash University, Caulfield East, Victoria 3145, Australia. Email: xibin.zhang@monash.edu.

1 Introduction

Aigner et al. (1977) and Meeusen and van Den Broeck (1977) introduced the stochastic frontier (SF) model to study the productive (in)efficiency of firms, and such SF models have since attracted considerable attention.[1] One distinct feature of a SF model is the decomposition of the error term, typically expressed as $\varepsilon = v - u$, where v is a random disturbance and $u \geq 0$ is the inefficiency term. Unlike standard regression models, the identification of v and u is crucial in SF modeling because of their economic interpretations.

In this paper, we develop a general estimation framework for time-varying panel SF models that can accommodate potential heterogeneity across firms through latent group structures. To illustrate our methodology, we consider a model similar to Yao et al. (2019)'s that allows for heterogeneous time-varying coefficients:

$$\begin{aligned} y_{it} &= \alpha_i^0 + \alpha_i(\tau_t) + x_{it}'\beta_i(\tau_t) + \varepsilon_{it}, \quad \text{with} \quad \varepsilon_{it} = v_{it} - u_i, \\ &= \alpha_i^0 - u_i + \alpha_i(\tau_t) + x_{it}'\beta_i(\tau_t) + v_{it}, \end{aligned} \tag{1.1}$$

for firm $i = 1, 2, \ldots, N$ and time $t = 1, 2, \ldots, T$, where $\tau_t = t/T \in (0, 1]$. In this specification of the model, α_i^0 and $\alpha_i(\tau_t)$ denote the constant and time-varying intercepts of the frontier, respectively. The term $\beta_i(\tau_t)$ denotes a non-random, time-varying coefficient vector, v_{it} is a zero-mean random error term, and $u_i \geq 0$ represents a firm-specific random inefficiency term. In addition to allowing for heterogeneity in the efficient frontier, we permit the variance of v_{it} to vary across firms, and allow for the possibility that u_i follows a mixture distribution.

Our study is motivated by the role of heterogeneity in the measurement of the efficient frontier and the inefficiency in panel SF models. Heterogeneity can either shift the efficient frontier or distort the location and scale of inefficiency estimates (see, for example, Galan et al., 2014).

[1] We refer readers to Kumbhakar and Lovell (2000) for early developments and Kumbhakar et al. (2022) and Tsionas et al. (2023) for recent advancements and a comprehensive review of the literature.

According to Greene (2005a), the true underlying frontier may include unmeasured firm-specific characteristics that reflect the technology in use. Greene (2005a) was instrumental in expanding SF models to incorporate firm heterogeneity by allowing for either true fixed effects or true random effects. Building on this, heterogeneity was further explored in subsequent work by allowing the inefficiency term to be purely transient or to contain both transient and persistent components (see, for example Colombi et al., 2014; Kumbhakar et al., 2014; Tsionas and Kumbhakar, 2014). These approaches help disentangle unobserved heterogeneity from inefficiency. However, a homogeneous frontier implicitly assumes that all firms operate under the same technology, so any systematic deviation is attributed to inefficiency. This assumption is restrictive and risks conflating inefficiency with unobserved differences in production regimes (Greene, 2005a,b).

The framework we propose addresses this issue by introducing a latent group structure for frontier parameters. Firms are partitioned into a small number of groups, with each group sharing a common frontier that reflects a particular technological regime (e.g., distinct business models or scale-specific production technologies). The finite collection of latent frontiers provides a disciplined approximation to such technological differences, after which inefficiency is measured relative to the appropriate group frontier. We view the latent group structure as providing a statistically disciplined way to capture the unobserved factors that plague inference of inefficiency, offering a middle ground between a fully homogeneous frontier and unrestricted firm-specific frontiers.

Formally, we assume that firms can be classified into one of $K^* \geq 1$ groups. Within each group, firms share a common set of parameters $\{\alpha(\tau_t), \beta(\tau_t), \text{Var}(v_t)\}$. We show that the classification step is consistent, ensuring that firms are benchmarked against the correct regime with probability approaching one. Importantly, we do not impose the same group structure for the distribution of the inefficiency term, u or on the constant term, α^0. We find that defining group membership for $\{\alpha^0, u\}$ in the same way as for the other parameters is inappropriate, the reasons of which we discuss in detail in Section 2.4. Instead, we account for potential heterogeneity in $\alpha^0 - u$ by

modeling it as a mixture of distributions.

The idea of uncovering latent group structures in panel data models has been extensively studied in recent years. Existing methods can be broadly categorized into two main approaches. The first approach relies on a two-step procedure in which individual-level estimation is followed by applying a clustering algorithm such as K-means (Lin and Ng, 2012; Bonhomme and Manresa, 2015; Ando and Bai, 2016), or Hierarchical Agglomerative Clustering (HAC) (Chen, 2019). The second approach, introduced by Su et al. (2016) performs simultaneous estimation and classification by penalizing the joint mean squared errors. Further developments along this line include Su et al. (2019), Huang et al. (2020) and Wang and Su (2021). Conceptually, the latter approach pools all observations across firms for joint estimation and classification, whereas the former relies on individual (non-pooled) firm-specific estimates as the basis of classification.

These existing approaches cannot be applied directly in our setting for the following reason. Estimating the parameters that govern the underlying distribution of u requires pooling observations across firms, typically via maximum likelihood based on the likelihood function in (C.2). However, since the log-likelihood function is complex and highly nonlinear, applying the method of Su et al. (2016) that requires simultaneous estimation and classification for SF models is non-trivial. This leads to a dilemma of whether to pool or not to pool observations for inference. To resolve this, we propose a new hybrid approach that combines pooled and non-pooled estimation methods that is flexible enough to be applied to a broad class of SF models.

Our paper makes several contributions to the literature on SF models and classification methods for panel data models. First, we estimate a robust panel SF model that incorporates heterogeneity across firms. We carefully set out the framework, clarifying what is feasible and what is not, and provide detailed explanations of the underlying rationale.[2] Second, to the best of our knowledge,

[2]Previous work on robust estimation of the SF model has primarily focused on semi-parametric or non-parametric specifications for either the efficient frontier (Park and Simar, 1994; Yao et al., 2019) or the error term distribution

this is one of the few papers in the literature to model and allow the variance of the error term to exhibit latent group structures. A notable study in this area is Loyo and Boot (2024), where they focus on modeling the variance of the error term, primarily for efficiency gains and/or the specific role played by the variance. In contrast, we aim to uncover heterogeneity in both the error and inefficiency terms, for the economic interpretations given in SF models. We emphasize the importance of modeling heterogeneity in both the variance of the error term, v, and the distribution of inefficiency, u, since inefficiency estimates depend on the variance of v in the widely used Jondrow et al. (1982)'s point estimator. Given the importance of the joint modeling of both the variance of v and the distribution of u in SF models, we consider this to be a substantial contribution. Finally, we propose a hybrid estimation procedure that combines firm-level and joint panel estimation to address the potential heterogeneity present in both the frontier and error components. As the description of the SF model in (1.1) alluded, we present the hybrid approach in the main text using a random effects specification of the SF model, and show in Appendix A how it extends to other variants, including the fixed effects SF models studied by Greene (2005a,b); Chen et al. (2014), and Zhou et al. (2020).

The rest of this paper is organized as follows. In Section 2, we introduce the estimation procedure and propose information criteria to determine the number of groups and whether u follows a unique or a mixture of distributions. Section 3 examines the theoretical properties of our procedure, specifying the conditions required for tuning parameters. In Section 4, we evaluate the small-sample performance of our method, providing practical recommendations for tuning parameters that meet the conditions and perform well in simulations. In Section 5, we apply our procedure to a dataset of large U.S. commercial banks, using the recommended tuning parameters from the simulations. Our findings confirm the presence of heterogeneity in the frontiers and a mixture

(Greene, 2005b; Lai and Kumbhakar, 2023). Additionally, there is a growing literature on specification tests for the distribution of inefficiency (Cheng et al., 2024).

distribution for the inefficiency term. The Online Appendix includes several important supplementary results. Appendix A.1 extends our approach to fixed effects SF models under distributional assumptions. Appendix A.2 further generalizes this to fixed effects SF models without any distributional restrictions, allowing for more flexible conditions on the inefficiency term. Appendix A.3 discusses a relaxation of the normalization condition. Appendix A.4 addresses the possibility that the inefficiency term equals zero with positive probability. Appendix B considers the mixture distribution with more than two components. Other parts of the Appendix support the results in the main body of the paper. Specifically, Appendices C and D describe, respectively, the approximate likelihood function of the model and the clustering method used. Proofs of the theorems and propositions from Section 3 are provided in Appendix E, with technical lemmas presented in Appendix H. Additional simulation and application results are included in Appendix F, and Appendix G further discusses the properties of the information matrix in support of Appendix E.

2 Model and Estimation

This section presents the model and the procedure. To facilitate exposition, we assume $\alpha_i^0 = \alpha^0$ and $\text{Var}(u_i) = \sigma_{ui}^2 = \sigma_u^2$ for all i in Sections 2.1 to 2.3. These restrictions are relaxed to the general case subsequently. The technical conditions and main theorems are postponed to the next section.

2.1 The Model

We illustrate our approach using the panel SF model with random effects (RE) and relegate other variants of the model to Appendix A. The particular model we consider modifies Yao et al. (2019)'s

by incorporating heterogeneous, time-varying coefficients:

$$\begin{aligned} y_{it} &= \alpha^0 + \alpha_i(\tau_t) + x_{it}'\beta_i(\tau_t) + \varepsilon_{it} \\ &= \alpha^0 - u_i + \alpha_i(\tau_t) + \sum_{l=1}^{p} x_{itl}\beta_{il}(\tau_t) + v_{it}, \end{aligned} \tag{2.1}$$

noting the temporal restriction $\alpha_i^0 = \alpha^0$, for expositional purposes to be generalized subsequently. The subscript l denotes the l-th element of a vector x_{it}, which is p by 1. In this specification, $\varepsilon_{it} = v_{it} - u_i$, where v_{it} is a mean-zero random error term, and u_i is a non-negative term capturing the inefficiency of firm i.[3] We let $\alpha_i(\cdot)$ and $\beta_i(\cdot)$ evolve smoothly in τ_t to capture the gradual technological drift, regulatory cycles, and business-model adjustments that are pervasive in long panels (for example, banking costs). Smoothness avoids implausible jumps while allowing flexible, low-frequency changes that pooled static frontiers cannot capture.

As in Yao et al. (2019), we assume that

$$v_{it} \sim N(0, \sigma_{vi}^2), \quad u_i \sim |N(0, \sigma_u^2)|, \quad \text{and} \quad v_{it} \perp u_i \perp x_{it}. \tag{2.2}$$

We restrict σ_u^2 to be identical for all i (no group structure) momentarily, again to facilitate exposition. We note that the two parameters where homogeneity is imposed, α^0 and σ_u^2, exhibit distinctive features compared to other parameters. Given the importance of these parameters in SF models, we devote a separate section to discuss these differences in detail in Section 2.4.

The assumption given by (2.2) implies that inefficiency arises from managerial, organizational or behavioral factors that are unrelated to observed input or output variables in the frontier. The resulting panel SF model is RE in the sense of Greene (2005a). In panel SF models, RE is attractive because fixed effects (FE) type likelihoods face an incidental-parameters problem in finite

[3] We describe the model, estimation method, and simulations in terms of the production frontier model, but use the cost frontier model for our application. The only difference is that the inefficiency term, u_i, enters the model negatively (production frontier) or positively (cost frontier). This distinction is minor, and one can let $\varepsilon_{it} \equiv v_{it} + u_i$ for cost frontiers.

T. RE avoids this by treating unit effects probabilistically, which Tsionas and Kumbhakar (2014) emphasize when arguing for a fully likelihood-based/Bayesian route for panel SF model analysis with multiple error components. We adopt this setup primarily to illustrate our proposed approach, although the methodology can be extended to alternative model specifications, such as the four-component panel stochastic frontier model introduced by Tsionas and Kumbhakar (2014) and Lai and Kumbhakar (2023). FE-type models are discussed separately in Appendices A.1 and A.2, as their treatment is relatively straightforward given the procedure developed in the main body of the paper.

We assume that there are $K^* \geq 1$ groups of parameters, and each firm's parameters belong to one of these groups. Mathematically,

$$\left\{\alpha_i\left(\tau_t\right), \beta_i\left(\tau_t\right), \sigma_{vi}\right\} = \sum_{k=1}^{K^*} \left\{\alpha^*_{(k)}\left(\tau_t\right), \beta^*_{(k)}\left(\tau_t\right), \sigma^*_{v(k)}\right\} \mathbf{1}(i \in G_k), \tag{2.3}$$

where $\mathbf{1}(\cdot)$ is the indicator function, equaling 1 if $(\cdot)$ is true and 0 otherwise. Additionally, parameters from different groups are distinct, meaning $\left\{\alpha^*_{(k)}\left(\tau_t\right), \beta^*_{(k)}\left(\tau_t\right), \sigma^*_{v(k)}\right\} \neq \left\{\alpha^*_{(j)}\left(\tau_t\right), \beta^*_{(j)}\left(\tau_t\right), \sigma^*_{v(j)}\right\}$, for $j \neq k$. The group membership sets satisfy $G_j \cap G_k = \emptyset$ and $\bigcup_{k=1}^{K^*} G_k = \{1, 2, \ldots, N\}$.

It is worth noting that we impose the following assumption on each $\alpha^*_{(k)}(s)$:

$$\int_0^1 \alpha^*_{(1)}(s)\,\mathrm{d}s = \int_0^1 \alpha^*_{(2)}(s)\,\mathrm{d}s = \ldots = \int_0^1 \alpha^*_{(K^*)}(s)\,\mathrm{d}s, \tag{2.4}$$

although we generalize it to allow them to differ in Appendix A.3.[4] The normalization adopted here is $\int_0^1 \alpha^*_{(k)}(s)\,\mathrm{d}s = 0$. Clearly, when $K^* = 1$ (the homogeneous case), this normalization is innocuous; however, it is not in the general case due to the restriction in (2.4). When the intercept term does not vary over time, $\alpha_i(s) = 0$. This normalization ensures that $\alpha_i(s)$ captures the time-varying component of the intercept term.

[4] We explain in detail why we impose this condition, how we adjust the procedure without it, and when we recommend relaxing the condition in Appendix A.3.

2.2 Approximation of $\alpha(\cdot)$ and $\beta(\cdot)$

The approximations we adopt are standard in the literature. Let $L^2[0,1] = \{f(s) : \int_0^1 f^2(s)\,\mathrm{d}s < \infty\}$ represent the space of square-integrable functions. The inner product equipped on this space is defined as $\langle f_1, f_2\rangle \equiv \int_0^1 f_1(s) f_2(s)\,\mathrm{d}s$, and the induced norm is $\|f\| = \langle f, f\rangle^{1/2}$. Following Dong and Linton (2018) and Atak et al. (2025), we use cosine functions as basis functions. In particular, $B_0(s) = 1$ and $B_j(s) = \sqrt{2}\cos(j\pi s)$ for $j \geq 1$. The set $\{B_j(s)\}_{j=0}^{\infty}$ then forms an orthonormal basis for the Hilbert space $L^2[0,1]$, such that $\langle B_i, B_j\rangle = \delta_{ij}$, where δ_{ij} is the Kronecker delta.

Suppose $f \in L^2[0,1]$ is κ-th order continuously differentiable. Then, we have

$$\begin{aligned} f(s) &= \sum_{j=0}^{\infty} B_j(s) v_j^0 = \sum_{j=0}^{m-1} B_j(s) v_j^0 + \sum_{j=m}^{\infty} B_j(s) v_j^0 \\ &= \sum_{j=0}^{m-1} B_j(s) v_j^0 + O\left(m^{-\kappa}\right) \equiv \mathbb{B}^m(s)' v^0 + O\left(m^{-\kappa}\right), \end{aligned}$$

where $\mathbb{B}^m(s) \equiv (B_0(s), B_1(s), \ldots, B_{m-1}(s))'$, $v_j^0 = \langle f, B_j\rangle$, and $v^0 = \left(v_0^0, v_1^0, \ldots, v_{m-1}^0\right)'$. Here, $\sum_{j=m}^{\infty} B_j(s) v_j^0$ is the bias term from using only the first $m-1$ terms contained in $\mathbb{B}^m(s)$ to approximate $f(s)$. If $f(s)$ is κ-th order differentiable, the bias term is $\sum_{j=m}^{\infty} B_j(s) v_j^0 = O(m^{-\kappa})$. When $\int_0^1 f(s)\,\mathrm{d}s = 0$ is imposed, we approximate f using $\mathbb{B}_{-0}^m(s) \equiv (B_1(s), \ldots, B_{m-1}(s))'$, since $\int_0^1 B_j(s)\,\mathrm{d}s = 0$ for $j \geq 1$. Similarly,

$$f(s) = \mathbb{B}_{-0}^m(s) v_{-0}^0 + O\left(m^{-\kappa}\right),$$

for some $v_{-0}^0 = \left(v_1^0, v_2^0, \ldots, v_{m-1}^0\right)'$.

We apply this approximation to our case. For each firm i, we have

$$\begin{aligned} y_{it} &= \alpha^0 - u_i + \alpha_i(\tau_t) + \sum_{l=1}^{p} x_{itl}\beta_{il}(\tau_t) + v_{it} \\ &\approx \alpha^0 - u_i + \mathbb{B}_{-0}^m(\tau_t)' \pi_{i0}^0 + \sum_{l=1}^{p} x_{itl}\mathbb{B}^m(\tau_t)' \pi_{il}^0 + v_{it} \\ &\equiv \alpha^0 - u_i + \left[\mathbb{B}_{-0}^m(\tau_t)', (x_{it} \otimes \mathbb{B}^m(\tau_t))'\right] \pi_i^0 + v_{it} \\ &\equiv \alpha^0 - u_i + z_{it}'\pi_i^0 + v_{it} \equiv \tilde{z}_{it}'\tilde{\pi}_i^0 + v_{it}, \end{aligned} \tag{2.5}$$

where the terms $\mathbb{B}_{-0}^{m}(\tau_t)'\pi_{i0}^{0}$ and $\mathbb{B}^{m}(\tau_t)'\pi_{il}^{0}$ represent approximations of $\alpha_i(\tau_t)$ and $\beta_{il}(\tau_t)$, respectively, and

$$
\begin{aligned}
x_{it}\otimes\mathbb{B}^{m}(\tau_t) &\equiv \left(x_{it1}B_0(\tau_t),\ldots,x_{it1}B_{m-1}(\tau_t),\ldots,x_{itp}B_0(\tau_t),\ldots,x_{itp}B_{m-1}(\tau_t)\right)',\\
\pi_i^0 &\equiv \left(\pi_{i0}^{0\prime},\pi_{i1}^{0\prime},\ldots,\pi_{ip}^{0\prime}\right)', \quad \tilde{\pi}_i^0 \equiv \left(\alpha^0-u_i,\pi_{i0}^{0\prime},\pi_{i1}^{0\prime},\ldots,\pi_{ip}^{0\prime}\right)',\\
z_{it} &\equiv \left[\mathbb{B}_{-0}^{m}(\tau_t)',\left(x_{it}\otimes\mathbb{B}^{m}(\tau_t)\right)'\right]', \quad \text{and} \quad \tilde{z}_{it}\equiv(1,z_{it}')'.
\end{aligned}
$$

The last two lines of equation (2.5) represent three equivalent ways of expressing the approximation.

2.3 The Estimation

The variance of inefficiency, σ_u^2 is identified through the skewness in the distribution of α^0-u_i across i. As a result, σ_u^2 cannot be identified or estimated without pooling observations across different i. However, pooling observations for estimation introduces challenges for numerical optimization, since the log-likelihood functions in (C.2) and (C.3) are complex and highly nonlinear. This issue exacerbates as the number of unknown parameters increases and becomes particularly pronounced in pooled estimation with classification methods. This creates a dilemma regarding whether to pool observations for estimation.

To address these challenges, we propose a hybrid procedure that combines estimations with and without pooling. We present the detailed steps of the procedure below.

Step 1: Individual Estimation

Using the approximation from (2.5), we regress y_{it} against $\mathbb{B}^{m}(\tau_t)$ and $x_{it}\otimes\mathbb{B}^{m}(\tau_t)$ for $t=1,2,\ldots,T$ to obtain $\widehat{\tilde{\pi}}_i$. From this we obtain $\hat{\sigma}_{vi}^2$ as the sample variance of the regression residuals.

Specifically, consider the following expression: $Z_{im}\equiv(\tilde{z}_{i1},...,\tilde{z}_{iT})'$, a $T\times m(p+1)$ vector.

Then the OLS estimator for firm i is given by

$$\hat{\tilde{\pi}}_i = \left(Z_{im}' Z_{im}\right)^{-1} Z_{im}' y_i, \tag{2.6}$$

with $y_i = (y_{i1}, \ldots, y_{iT})'$. We then obtain an estimate of σ_{vi}^2 as $\hat{\sigma}_{vi}^2 = \frac{1}{T-1} \sum_{t=1}^{T} \left(y_{it} - \tilde{z}_{it}' \hat{\tilde{\pi}}_i\right)^2$. Excluding the first element in $\hat{\tilde{\pi}}_i$, we let $\hat{\pi}_i$ denote the estimated coefficients associated with z_{it}. The estimates $\hat{\pi}_i$ and $\hat{\sigma}_{vi}$ are collected to form an estimate of ϑ_i: $\hat{\vartheta}_i = \left(\hat{\pi}_i', \hat{\sigma}_{vi}\right)'$, based on which we form groups.

Step 2: Classification

Having obtained $\hat{\vartheta}_1, \hat{\vartheta}_2, \ldots, \hat{\vartheta}_N$ from Step 1, we use the L_2 norm to measure the distance between $\hat{\vartheta}_i$ and $\hat{\vartheta}_j$. Based on this distance measure, we then apply the classical HAC algorithm to the estimates of each firm's functional coefficient to determine group memberships. The HAC is a widely used algorithm for clustering, and several variants of it are employed in heterogeneous panel data models (see, for e.g., Chen (2019)). Details of the HAC method are provided in Appendix D and refer the readers to Everitt et al. (2011) for a comprehensive treatment. Given a value for K, we apply the HAC to obtain an estimate of the group membership, denoted as $\left(\hat{G}_{1|K}, \hat{G}_{2|K}, \ldots, \hat{G}_{K|K}\right)$, which forms a partition of the set $\{1, 2, \ldots, N\}$.

Step 3: Post-Classification Estimation and Determination of K^*

Within each group, we now have significantly more observations available for pooling. Recognizing this, we set the number of sieve terms to $\underline{m}$, which is substantially larger than m. Within each estimated group, $\hat{G}_{k|K}$ for $1 \leq k \leq K$, we conduct post-classification estimation using standard within-panel data estimation methods. Let $\underline{z}_{it} = \left[\mathbb{B}_{-0}^{\underline{m}}(\tau_t)', (x_{it} \otimes \mathbb{B}^{\underline{m}}(\tau_t))'\right]'$ denote the new regressors. At this stage, we do not consider the inefficiency term $\alpha^0 - u_i$. The group specific

coefficient is given by

$$\hat{\pi}_{(k|K)} = \arg\min_{\pi} \sum_{i \in \hat{G}_{k|K}} \sum_{t=1}^{T} \left(\ddot{y}_{it} - \ddot{z}_{it}' \pi \right)^2,$$

where $\ddot{y}_{it} = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it}$, and $\ddot{z}_{it} = z_{it} - \frac{1}{T}\sum_{t=1}^{T} z_{it}$. The estimate of the variance of v_{it} for group $\hat{G}_{k|K}$ is

$$\hat{\sigma}^2_{v(k|K)} = \frac{1}{N_k(T-1)} \sum_{i \in \hat{G}_{k|K}} \sum_{t=1}^{T} \left(\ddot{y}_{it} - \ddot{z}_{it}' \hat{\pi}_{(k|K)} \right)^2,$$

where $N_k = \sharp\{\hat{G}_{k|K}\}$ is the number of elements in $\hat{G}_{k|K}$. For simplicity, we do not explicitly distinguish between $\hat{N}_k = \sharp\{\hat{G}_{k|K^*}\}$ and $N_k = \sharp\{G_{k|K^*}\}$. Similarly, $\hat{\vartheta}_{(k|K)} = \left(\hat{\pi}'_{(k|K)}, \hat{\sigma}_{v(k|K)} \right)'$.

Inspired by the pseudo log-likelihood, we construct an information criterion to determine the optimal number of groups as follows:

$$\begin{aligned} \mathrm{IC}(K, \lambda_{NT}) &= \sum_{k=1}^{K} \left\{ N_k T \log(\hat{\sigma}_{v(k|K)}) + \sum_{i \in \hat{G}_{k|K}} \sum_{t=1}^{T} \frac{\left(\ddot{y}_{it} - \ddot{z}_{it}' \hat{\pi}_{(k|K)} \right)^2}{\hat{\sigma}^2_{v(k|K)}} \right\} + \lambda_{NT} K \\ &= \sum_{k=1}^{K} \left\{ N_k T \log(\hat{\sigma}_{v(k|K)}) + N_k (T-1) \right\} + \lambda_{NT} K, \end{aligned} \tag{2.7}$$

where λ_{NT} is a suitable penalty term. The optimal number of groups is the minimizer of (2.7)

$$\hat{K}(\lambda_{NT}) = \arg\min_{K=1,2,\ldots,\bar{K}} \mathrm{IC}(K, \lambda_{NT}),$$

given a suitable $\bar{K}$. For brevity, we henceforth refer to this as $\hat{K}$. The final group estimates are then given by

$$\hat{\vartheta}_{(k|\hat{K})} = \left(\hat{\pi}_{(k|\hat{K})}, \hat{\sigma}_{v(k|\hat{K})} \right), \quad k = 1, 2, \ldots, \hat{K}.$$

Step 4: Estimation of α^0 and σ_u^2

We estimate α^0 and σ_u^2 pooling all observations via maximum likelihood estimation (MLE). Specifically, the estimate is given by

$$\left(\hat{\alpha}^0, \hat{\sigma}_u^2 \right) = \arg\max_{(s, \delta_u^2)} \sum_{k=1}^{\hat{K}} \sum_{i \in \hat{G}_{k|\hat{K}}} \log f \left(y_i \mid x_i; s, \delta_u^2, \hat{\vartheta}_{(k|\hat{K})} \right),$$

where $y_i = (y_{i1}, \ldots, y_{iT})'$, $x_i = (x_{i1}, \ldots, x_{iT})'$, and $f\left(y_i \mid x_i; s, \delta_u^2, \hat{\vartheta}_{(k|\hat{K})}\right)$ is defined in (C.2), noting that we use the post-classification estimates of ϑ. Since only two parameters, α^0 and σ_u^2, are being estimated at this stage, the numerical optimization is straightforward.

2.4 Inefficiency Term

We now consider the general case where α^0 can be heterogeneous across i, and we will use α_i^0 from this point onward. Modeling the underlying structure of $\alpha_i^0 - u_i$ differs from that of $\{\alpha_i(\tau_t), \beta_i(\tau_t), \sigma_{vi}^2\}$ because u_i is assumed to be random effects, and $\alpha_i^0 - u_i$ naturally varies across i, even when α_i^0 is identical. For this reason, we focus on identifying the distribution of $\alpha_i^0 - u_i$ rather than the actual values. While we can uncover the underlying distribution, consistently estimating group membership remains challenging.

Consider the following example to illustrate this point. Suppose we have two random variables ε_1 and ε_2, with $\varepsilon_1 \sim 0 - |N(0,2)|$ and $\varepsilon_2 \sim 1 - |N(0,1)|$. If we mix i.i.d. realizations of ε_1 and ε_2, such as $\{\varepsilon_{11}, \varepsilon_{12}, \ldots, \varepsilon_{1n}, \varepsilon_{21}, \varepsilon_{22}, \ldots, \varepsilon_{2n}\}$, it is likely that many ε_{1i} and ε_{2j} values lie very close to one another. For example, a small-scale Monte Carlo experiment with $n = 100$ suggests that about 28% of ε_{1i} have at least one ε_{2j} within a radius of 0.01. In such cases, swapping their memberships would likely have a minimal impact on the likelihood function, complicating their distinct identification from the data.

Misclassification of group memberships can have a serious impact on the inefficiency term, unlike parameters at the frontiers, where only similar frontiers can be misclassified together due to low power or minor estimation errors. Continuing the previous example, suppose $\varepsilon_{1i} = 0$ (highly efficient with $u_{1i} = 0$) is misclassified as ε_2, then the inefficiency term for ε_{1i} would be calculated as 1 (indicating inefficiency). Conversely, if $\varepsilon_{2j} = 0$ (originally not efficient with $u_{2j} = 1$) is misclassified as ε_1, then the inefficiency term for ε_{2j} would be calculated as 0 (indicating high efficiency).

Given these challenges and the serious implications of misclassification, we adopt a mixture distribution approach as follows. Suppose there exist an integer $\mathcal{K}^* \geq 1$, such that with probability τ_j^0, it is distributed as $\alpha_{(j)}^0 - \left|N(0, \sigma_{u(j)}^2)\right|$ for $j = 1, 2, ..., \mathcal{K}^* - 1$, and with probability $\tau_{\mathcal{K}^*}^0 = 1 - \tau_1^0 - ... - \tau_{\mathcal{K}^*-1}^0$, as $\alpha_{(\mathcal{K}^*)}^0 - \left|N(0, \sigma_{u(\mathcal{K}^*)}^2)\right|$, where $\left(\alpha_{(j)}^0, \sigma_{u(j)}^2\right)$, $j = 1, 2, ..., \mathcal{K}^*$, are distinct vectors, $0 < \tau_j^0 < 1$, $j = 1, 2, ..., \mathcal{K}^* - 1$, and $1 - \tau_1^0 - ... - \tau_{\mathcal{K}^*-1}^0 > 0$. When $\mathcal{K}^* = 1$, the error distribution is reduced to that of a unique distribution. The mixture distribution on the inefficiency term is similar in spirit to the latent class model in Greene (2005b). However, the latent class model is only a small part of Greene (2005b) and so the treatment is very brief. We examine this issue in depth by proposing an information criterion to determine the number of components, rigorously establish its theoretical properties, and assess its small-sample performance through simulation studies.

The potential presence of a mixture distribution significantly alters the interpretation of the results. With uniquely distributed inefficiency term, we can remove the subscript i from α_i^0 because $\alpha_i^0 - u_i \overset{d}{\sim} \alpha^0 - |N(0, \sigma_u^2)|$. A point estimate of $\alpha^0 - u_i$ is

$$\widehat{\alpha^0 - u_i} = \frac{1}{T} \sum_{t=1}^{T} \left(y_{it} - z_{it}' \hat{\pi}_i\right),$$

where $\hat{\pi}_i$ is a sub-vector of $\hat{\hat{\pi}}_i$ defined in (2.6) in Step 1. Thus, $\alpha^0 - u_i$ can be estimated consistently as $T \to \infty$, allowing us to rank firms according to inefficiency because α^0 is identical across i. However, in the case of a mixture distribution, although we can still consistently estimate $\widehat{\alpha_i^0 - u_i}$ (similar to the above), it is not possible to rank firms as in the former scenario. This limitation arises because memberships, or equivalently, the values of α_i^0, cannot be identified. This observation aligns with the findings for the cross-sectional case discussed in Greene (2005b).

The presence of a mixture distribution in the distribution of $\alpha_i^0 - u_i$ does not impact the estimation of ϑ_i given independence among u_i, v_{it}, and x_{it}. Consequently, Steps 1, 2, and 3 remain unchanged. Details of the revised Step 4, now referred to as Step 4', are provided below.

Step 4': Estimation of $\alpha^0_{(j)}, \sigma^2_{u(j)},$ and τ^0_j

We adopt the mixture distribution for $\alpha^0_i - u_i$ as previously described. Assuming that the inefficiency terms come from $\mathcal{K} \geq 1$ distributions, we obtain an estimate of $\left(\alpha^0_{(1)}, \sigma^2_{u(1)}, ..., \alpha^0_{(\mathcal{K})}, \sigma^2_{u(\mathcal{K})}, \tau^0_1, ..., \tau^0_{\mathcal{K}-1}\right)$ using MLE as follows:

$$\begin{aligned}
&\left(\hat{\alpha}^0_{(1)}, \hat{\sigma}^2_{u(1)}, ..., \hat{\alpha}^0_{(\mathcal{K})}, \hat{\sigma}^2_{u(\mathcal{K})}, \hat{\tau}_1, ..., \hat{\tau}_{\mathcal{K}-1}\right) \\
&= \arg \max_{(s,\delta^2_u,\tau)} \sum_{k=1}^{\hat{K}} \sum_{i \in \hat{G}_{k|\hat{K}}} \log \tilde{f}\left(y_i \left| x_i; s_{(1)}, \delta^2_{u(1)}, ..., s_{(\mathcal{K})}, \delta^2_{u(\mathcal{K})}, \tau_1, ..., \tau_{\mathcal{K}-1}, \hat{\vartheta}_{(k|\hat{K})}\right.\right)
\end{aligned}$$

where $\tilde{f}$ is the likelihood function defined in (C.3), and we incorporate estimates from Step 3, as detailed in Section 2.3.

Step 5: Determination of the Distributional Structures of the Inefficiency Term

To determine the optimal number of mixtures, we introduce a new information criterion for this task:

$$\widetilde{\text{IC}}(\mathcal{K}, \tilde{\lambda}_{NT}) = -\sum_{k=1}^{\hat{K}} \sum_{i \in \hat{G}_{k|\hat{K}}} \log \tilde{f}\left(y_i \left| x_i; \hat{\alpha}^0_{(1)}, \hat{\sigma}^2_{u(1)}, ..., \hat{\alpha}^0_{(\mathcal{K})}, \hat{\sigma}^2_{u(\mathcal{K})}, \hat{\tau}_1, ..., \hat{\tau}_{\mathcal{K}-1}, \hat{\vartheta}_{(k|\hat{K})}\right.\right) + \mathcal{K}\tilde{\lambda}_{NT}, \tag{2.8}$$

where $\tilde{\lambda}_{NT}$ is a suitable penalty term, and the estimates are as obtained from Steps 3 and 4'.

The optimal number of mixtures is the minimizer of (2.8)

$$\hat{\mathcal{K}}(\tilde{\lambda}_{NT}) = \arg \min_{\mathcal{K}=1,2,...,\bar{\mathcal{K}}} \widetilde{\text{IC}}(\mathcal{K}, \tilde{\lambda}_{NT}),$$

and we write $\hat{\mathcal{K}}$ for short. Finally, the estimated parameters are

$$\left(\hat{\alpha}^0_{(1)}, \hat{\sigma}^2_{u(1)}, ..., \hat{\alpha}^0_{(\hat{\mathcal{K}})}, \hat{\sigma}^2_{u(\hat{\mathcal{K}})}, \hat{\tau}_1, ..., \hat{\tau}_{\hat{\mathcal{K}}-1}\right).$$

2.5 A Summary of the Estimation Procedure

The outline of the estimation procedure is as follows:

1. Conduct estimations of the frontiers for each firm as described in Step 1.

2. Apply the HAC algorithm using the individual estimations from Step 1 for $K = 1, 2, \ldots, \bar{K}$.

3. Use the information criterion in (2.7) from Step 3 to determine the optimal number of groups and group memberships. Obtain an estimate of the frontier with the determined groups.

4. Based on the group assignments from Step 3, perform joint estimation for the inefficiency term as in Step 4'.

5. Use the information criterion in (2.8) to determine the distribution of $\alpha_i^0 - u_i$, as in Step 5.

6. Collect results. The estimates of the frontiers and the distribution of the inefficiency term are derived from Steps 3 and 5, respectively.

3 Asymptotic Properties

We examine classification consistency in Section 3.1. Subsequently, we discuss the large sample properties of the post-classification estimators in Section 3.2.

3.1 Classification

Assumption 1. *The process* $\{(x_{it}', v_{it}), t = 1, \ldots, T\}$ *is strong mixing with a mixing coefficient* $\alpha(j)$ *that satisfies* $\alpha(j) \leq C_\alpha \rho^j$ *for some positive* C_α *and* $0 < \rho < 1$*, and this holds for* $i = 1, \ldots, N$.

Assumption 2. $\max_{1 \leq i \leq N, 1 \leq t \leq T} \mathrm{E}\|x_{it}\|^q \leq \bar{C}_x < \infty$, *and* $\max_{1 \leq i \leq N, 1 \leq t \leq T} \mathrm{E}|v_{it}|^q \leq \bar{C}_v < \infty$, *for some* $q > 4$.

Assumption 3. *Denote* $\tilde{x}_{it} \equiv (1, x_{it}')'$. *Let* $\mu_{\min}$ *and* $\mu_{\max}$ *denote the minimum and maximum eigenvalues of a matrix, respectively. There exist* $\underline{C}_{xx}$ *and* $\bar{C}_{xx}$ *with* $0 < \underline{C}_{xx} \leq \bar{C}_{xx} < \infty$ *such*

that

$$0 < \underline{C}_{xx} \leq \min_{1\leq i\leq N, 1\leq t\leq T} \mu_{\min}[\mathrm{E}(\tilde{x}_{it}\tilde{x}_{it}')] \leq \max_{1\leq i\leq N, 1\leq t\leq T} \mu_{\max}[\mathrm{E}(\tilde{x}_{it}\tilde{x}_{it}')] \leq \bar{C}_{xx} < \infty.$$

Assumption 4. *For $k = 1, 2, \ldots, K^*$, $\alpha^*_{(k)}(s)$, $\beta^*_{(k)1}(s)$, ..., and $\beta^*_{(k)p}(s)$ belong to $L^2[0,1]$ and are κ times continuously differentiable.*

Assumption 5. *There exists a positive $\underline{C}^*$, such that*

$$\min_{1\leq j\neq k\leq K^*} \left\{ \|\alpha^*_{(j)} - \alpha^*_{(k)}\| + \sum_{l=1}^{p} \|\beta^*_{(j)l} - \beta^*_{(k)l}\| + |\sigma^*_{v(j)} - \sigma^*_{v(k)}| \right\} \geq \underline{C}^* > 0,$$

and $\min_{1\leq k\leq K^*} \sigma^{*2}_{v(k)} > 0$.

Assumption 6. *(i) N can either be a finite, or divergent. If N is divergent, $N = O(T^C)$ for some positive C as $T \to \infty$. (ii) $m \to \infty$ as $T \to \infty$. $Nm^{q/2+2}(\log N)^{2q}/T^{q/2-1} \to 0$, with q in Assumption 2.*

Assumption 1 imposes a condition of weak dependence across t, noting that independence across i is not required for classifications. Assumption 2 requires that x and v have finite q-th moment. Assumption 3 is the classic full rank condition. Assumption 4 stipulates that the coefficients are κ-th order differentiable, a standard condition for nonparametric or semiparametric estimation. Assumption 5 requires that at least one of the coefficients, including the variance of v, must differ across groups.

Assumption 6 specifies that the moment conditions must be sufficiently large or that T grows fast enough. N can be fixed. If N diverges, it cannot be too fast, e.g., at the rate of $\exp(T)$. In the empirical application, $(N, T) = (466, 80)$ and $466 \approx 80^{1.4}$, thus any $C \geq 1.4$ works in (i). The most stringent requirements arise from the estimation of the "design" matrix $\frac{1}{T}Z_{im}'Z_{im}$ with diverging dimensions, used in $\hat{\tilde{\pi}}_i$ (see (2.6)); a similar condition was imposed in Chen (2019). We take a logarithm of all covariates before estimation, and q can be reasonably considered large, e.g., $q \geq 8$. If we set $m = T^{1/5}$, Assumption 6 is satisfied. We do not have the usual bias

and variance tradeoff here, as explained below. The results of Theorem 3.1 are underpinned by the uniform convergence of $\widehat{\widehat{\pi}}_i$ without any rate requirement. As a result, it is not necessary to consider the trade-off between the bias and variance of the estimates for this aspect, when deciding m. Of course, we do need $m \to \infty$ to ensure the uniform convergence. However, to achieve the optimal convergence rate for the post-classification estimates, this consideration becomes crucial, as reflected in Assumption 10 (ii) in the subsequent section. With the aforementioned technical conditions, we demonstrate the consistency of the classification.

Theorem 3.1. *Suppose Assumptions 1 through 6 hold. Then:*

(i) For any small positive ϵ*,*

$$\Pr\left(\max_{i=1,2,\ldots,N}\left\|\hat{\vartheta}_i - \vartheta_i\right\| > \epsilon\right) = o(1);$$

(ii) Assuming $K = K^*$*, denote the event*

$$\mathcal{M} \equiv \left\{\left(\hat{G}_{1|K^*}, \hat{G}_{2|K^*}, \ldots, \hat{G}_{K^*|K^*}\right) = \left(G_{1|K^*}, G_{2|K^*}, \ldots, G_{K^*|K^*}\right)\right\},$$

then $\Pr(\mathcal{M}) \to 1$.

Theorem 3.1 (i) establishes the uniform convergence of $\hat{\vartheta}_i$ for $i = 1, 2, \ldots, N$ provided $m \to \infty$. Building on this, Theorem 3.1 (ii) demonstrates that the probability of correct classification approaches 1, provided that $K = K^*$. In the next section, we will argue that the K we choose converges to K^* with probability approaching 1, and we discuss the asymptotic properties of the post-classification estimates.

We note that significantly fewer assumptions are required for consistency in classification compared to those needed for post-classification and determining the number of groups. For instance, we do not need independence or weak dependence across i, nor do we require specific distributional assumptions on v_{it} and u_i.

3.2 The Number of Groups and Post-Classification Estimator

In this section, we address the question regarding the choice of K and the post-classification estimation. We begin by presenting additional assumptions necessary for this analysis.

Assumption 7. (x_i, ε_i), *for* $i = 1, 2, \ldots, N$ *are independent across* i.

Assumption 8. $N_k \propto N$ *for each* $k = 1, 2, \ldots, K^*$.

Assumption 9. $v_{it} \overset{iid}{\sim} N\left(0, \sigma_{vi}^2\right)$ *across* t. *There exist an integer* $\mathcal{K}^* \geq 1$, *such that,*

$$\alpha_i^0 - u_i \overset{d}{\sim} \alpha_{(j)}^0 - \left|N\left(0, \sigma_{u(j)}^2\right)\right| \text{ with probability } \tau_j^0, j = 1, 2, ..., \mathcal{K}^*,$$

where $0 < \tau_j^0 < 1$ *and* $\sigma_{u(j)}^2 > C > 0$ *for* $j = 1, 2, ..., \mathcal{K}^*$, $\tau_1^0 + \tau_2^0 + ... + \tau_{\mathcal{K}^*}^0 = 1$, $\left(\alpha_{(j)}^0, \sigma_{u(j)}^2\right)$ *differ across* $j = 1, 2, ..., \mathcal{K}^*$. *Lastly, the sequences* $\{v_{it}\}_{t=1}^T$, u_i, *and* $\{x_{it}\}_{t=1}^T$ *are mutually independent.*

Assumption 10. *(i)* $T \propto N^{C_*}$ *for some positive* C_* *as* $N \to \infty$; *(ii)* $\underline{m} \to \infty$ *as* $T, N \to \infty$. *Additionally,* $\underline{m}/T \to 0$, $\underline{m}^{q/2+2}(\log N)^{2q}/(NT)^{q/2-1} \to 0$, *and* $NT/\underline{m}^{1+2\kappa} \to 0$; *(iii)* $C_* > 1/(2\kappa)$ *and* $(q-2)\kappa > 3$.

Assumption 7 further imposes independence across i. Assumption 8 says the number of members in each group is proportional to N. This condition is not necessary, but it facilitates expositions. Assumption 9 specifies the distributional conditions on the error terms. As explained in Section 2.4, these conditions are essential for the identification of σ_u^2, and common in the literature, (see, for e.g., Yao et al. (2019)). Assumption 10 places restrictions on the rate of growth of T relative to N, and the rate of growth of the tuning parameter $\underline{m}$. The condition in (iii) is set to ensure that the set of $\underline{m}$ that satisfies (ii) is not empty. We need $\underline{m}/T \to 0$ so that the "design" matrix $\mathrm{E}\left(\tilde{z}_{it}\tilde{z}_{it}'\right)$ is still well-behaved, as required in Lemma H.2. However, condition $\underline{m}/T \to 0$ can be restrictive when N is much larger than T. $\underline{m}^{q/2+2}\left(\log N\right)^{2q} \Big/ \left(NT\right)^{q/2-1} \to 0$ is assumed to ensure the sample version of the design matrix, namely $Q_{(k),zz}$ defined in (3.1), is of full rank with very

high probability. Note that the dimension of $Q_{(k),zz}$ is diverging, so the consistency of this matrix requires uniform convergence of all elements and hence this restriction. $NT/\underline{m}^{1+2\kappa} \to 0$ ensures the bias term is asymptotically negligible. In the special case where $\kappa \geq 2$, we need $C_* > 1/4$, and $(q-2)\kappa > 3$ is satisfied due to $q > 4$ in Assumption 2. One can then set, for example, $\underline{m} = (NT)^{1/4.8}$, which satisfies condition (ii).

We show the asymptotic properties of our estimators for the case where $\mathcal{K}^* \geq 2$. The case in which $\alpha_i^0 - u_i$ comes from a unique distribution is straightforward given this result.

Recall that $\mathbb{B}_{-0}^m(\tau_t)' \pi_{i0}^0$ and $\mathbb{B}^m(\tau_t)' \pi_{il}^0$ represent the approximations of $\alpha_i(\tau_t)$ and $\beta_{il}(\tau_t)$. For the coefficients on the frontiers, let $\theta(s) \equiv \left(\alpha(s), \beta(s)'\right)'$, and correspondingly $\hat{\theta}(s) = \left(\mathbb{B}_{-0}^{\underline{m}}(\tau_t)' \hat{\pi}_0, \mathbb{B}^{\underline{m}}(\tau_t)' \hat{\pi}_1, ..., \mathbb{B}^{\underline{m}}(\tau_t)' \hat{\pi}_p\right)'$. For the parameters in the distribution of $\alpha_i^0 - u_i$, we denote

$$\varrho^0 \equiv \left(\alpha_{(1)}^0, \sigma_{u(1)}^2, ..., \alpha_{(\mathcal{K}^*)}^0, \sigma_{u(\mathcal{K}^*)}^2, \tau_1^0, ..., \tau_{\mathcal{K}^*-1}^0\right),$$

and correspondingly

$$\hat{\varrho} = \left(\hat{\alpha}_{(1)}^0, \hat{\sigma}_{u(1)}^2, ..., \hat{\alpha}_{(\mathcal{K}^*)}^0, \hat{\sigma}_{u(\mathcal{K}^*)}^2, \hat{\tau}_1, ..., \hat{\tau}_{\mathcal{K}^*-1}\right).$$

The following notations are used to characterize the asymptotic distribution. Denote

$$\mathbb{M}_{\mathbb{B}}(s) \equiv \begin{pmatrix} \mathbb{B}_{-0}^{\underline{m}}(s)' & 0 & \cdots & 0 \\ 0 & \mathbb{B}^{\underline{m}}(s)' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbb{B}^{\underline{m}}(s)' \end{pmatrix}_{(p+1)\times(\underline{m}-1+\underline{m}p)},$$

$$Q_{(k),zz} = \frac{1}{N_k T} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \ddot{z}_{it} \ddot{z}_{it}', \tag{3.1}$$

and

$$\mathbb{S}_{(k)}(s) = \frac{\sigma_{v(k)}^{*2}}{\underline{m}} \mathbb{M}_{\mathbb{B}}(s) Q_{(k),zz}^{-1} \mathbb{M}_{\mathbb{B}}(s)',$$

noting that $\mathbb{S}_{(k)}(s)$ is positive definite with very high probability; which we show it in equation (H.7) of Appendix H. For notation convenience, write

$$\tilde{f}_{i(k)}(\varrho) \equiv \tilde{f}\left(y_i \left| x_i; \varrho, \vartheta_{(k|K^*)}\right.\right),$$

and

$$\mathbb{I} \equiv -\mathrm{E}\left[\frac{1}{N}\sum_{k=1}^{K^*}\sum_{i\in G_k}\frac{\partial^2}{\partial\varrho\partial\varrho'}\log\tilde{f}_{i(k)}(\varrho)\right]\Bigg|_{\varrho=\varrho^0},$$

with $\mathbb{I}^{1/2}$ denoting the matrix such that $\mathbb{I}^{1/2}\mathbb{I}^{1/2\prime} = \mathbb{I}$. $\mathbb{I}$ is a positive definite matrix with finite eigenvalues, as shown in Appendix G. As before, we show the asymptotic property of $\hat{\theta}, \hat{\sigma}_v^2$ and $\hat{\varrho}$ pretending that we know K^* and $\mathcal{K}^*$. We then show that $\hat{K}$ and $\hat{\mathcal{K}}$ converge to K^* and $\mathcal{K}^*$, respectively, with probability approaching 1.

Theorem 3.2. *Suppose Assumptions 1 through 10 hold,* $\hat{K}(\lambda_{NT}) = K^*$ *and* $\hat{\mathcal{K}}\left(\tilde{\lambda}_{NT}\right) = \mathcal{K}^*$. *Let* I_l *denote the* $l \times l$ *identity matrix. Then, for each* $k = 1, 2, \ldots, K^*$,

(i)

$$\sqrt{\frac{N_k T}{m}}\mathbb{S}_{(k)}^{-1/2}(s)\left(\hat{\theta}_{(k|K^*)}(s) - \theta^*_{(k)}(s)\right) \xrightarrow{d} N(0, I_{p+1});$$

(ii)

$$\sqrt{N_k T}\left(\hat{\sigma}^2_{v(k|K^*)} - \sigma^{*2}_{v(k)}\right) \xrightarrow{d} N\left(0, \mathrm{Var}(v_{it}^2 | i \in G_{k|K^*})\right);$$

(iii)

$$\sqrt{N}\mathbb{I}^{-1/2}(\hat{\varrho} - \varrho^0) \xrightarrow{d} N(0, I_{3\mathcal{K}^*-1}).$$

This theorem establishes the asymptotic properties of the post-classification estimators. As expected, $\hat{\theta}$ converges at a nonparametric rate, while $\hat{\sigma}_v^2$ converges at a parametric rate. The convergence rate of $\hat{\varrho}$ is $\sqrt{N}$ and does not depend on T. This result may appear odd, but there is a simple explanation. Note that ϱ collects only the parameters that govern the distribution of u_i. The best scenario of estimating ϱ is that we observe $u_1, u_2, ..., u_N$ directly, in which case the rate of convergence of $\hat{\varrho}$ is $\sqrt{N}$. In theory, the value of T does not impact the convergence rate of $\hat{\varrho}$.

However, in finite samples, large T can potentially ensure a more precise estimation of u_i, and thus can possibly improve the finite-sample performance of $\hat{\varrho}$.

In addition, the validity of the proposed information criteria relies on these properties, as they depend on the accuracy and consistency of the post-classification estimators, as demonstrated above. For example, λ_{NT} depends on both N and T (due to the rates of $\hat{\theta}$ and $\hat{\sigma}_v^2$), while $\tilde{\lambda}_{NT}$ depends only on N (due to the rate of $\hat{\varrho}$).

Proposition 3.3. *Suppose Assumptions 1 through 10 hold.*

(i) Select a value of λ_{NT} such that $(NT)^{-1/2}\lambda_{NT} \to \infty$ and $(NT)^{-1}\lambda_{NT} \to 0$. Then,

$$\Pr\left(\hat{K}\left(\lambda_{NT}\right) = K^*\right) \to 1.$$

(ii) Select a value of $\tilde{\lambda}_{NT}$ such that $\tilde{\lambda}_{NT} \to \infty$ and $N^{-1}\tilde{\lambda}_{NT} \to 0$. Then,

$$\Pr\left(\hat{\mathcal{K}}\left(\tilde{\lambda}_{NT}\right) = \mathcal{K}^*\right) \to 1.$$

Proposition 3.3 presents conditions under which the information criteria are valid, focusing on the tuning parameters λ_{NT} and $\tilde{\lambda}_{NT}$. As highlighted earlier, selecting the correct range for λ_{NT} and $\tilde{\lambda}_{NT}$ is crucial. In the subsequent section, we will evaluate specific values for λ_{NT} and $\tilde{\lambda}_{NT}$, identify those that perform well in simulations, and recommend practical choices.

4 Monte Carlo Simulations

4.1 Simulation Designs

Heterogeneity in panel SF models arises from various sources. Thus, we design three different Monte Carlo experiments that allow us to examine the finite-sample performance of the proposed method and its ability to identify sources of heterogeneity. In the first design, we study the classification for the case with heterogeneity from frontiers yet with constant variances of v_{it}. In the

second design, we study the case with heterogeneous variances of v_{it}, yet homogeneous frontiers. In the third design, we check the performance of our methods in a more general and much more complicated scenario. In all three designs, we consider two sub-cases where the term $\alpha^0 - u$ comes from either unique or from mixture distribution, which previous methods did not consider. Due to the similarity and the space constraint, we defer the description and simulation results of Designs 1 and 2 to Appendix F and only present Design 3 here.

Design 3: In our third design (consisting of DGP3U and DGP3M), we study the performance of our method in a setting similar to those in Yao et al. (2019), where there are three groups for both the frontiers and variances, with two regressors. The DGP is

$$y_{it} = \alpha_i^0 - u_i + \alpha_i(\tau_t) + x_{it1}\beta_{i1}(\tau_t) + x_{it2}\beta_{i2}(\tau_t) + v_{it},$$

where $x_{itl} \sim N(1, 0.5^2)$ for both regressors $l = 1, 2$. Group 1 frontiers and error term are specified as $\alpha_{(1)}(s) = -\frac{1}{1+3s} - \varpi_1$, $\beta_{(1)1}(s) = 2s^3$, $\beta_{(1)2}(s) = \ln(5s)$, $v_{it} \overset{iid}{\sim} N(0, \sigma^2_{v(1)})$ with $\sigma_{v(1)} = 0.75$ and ϖ_1 is a mean of $-\frac{1}{1+3s}$. Group 2 frontiers and error term are specified as $\alpha_{(2)}(s) = -\cos(4s) - \varpi_2$, $\beta_{(2)1}(s) = \sin(4s)$, $\beta_{(2)2}(s) = \ln(\frac{s}{1-s})$, $v_{it} \overset{iid}{\sim} N(0, \sigma^2_{v(2)})$ with $\sigma_{v(2)} = 1.25$ and ϖ_2 is a mean of $-\cos(4s)$. Group 3 frontiers and error term are specified as $\alpha_{(3)}(s) = 5s^2 - s + 1 - \varpi_3$, $\beta_{(3)1}(s) = \exp(-s) + \sin(5s)$, $\beta_{(3)2}(s) = -5\sin(s)\cos(5s) + 1$, $v_{it} \overset{iid}{\sim} N(0, \sigma^2_{v(3)})$, with $\sigma_{v(3)} = 1.25$ and ϖ_3 is a mean of $5s^2 - s + 1$. We consider two sub-cases of $\alpha^0 - u$, which we denote them as DGP3U and DGP3M. In DGP3U, $\alpha^0 - u$ comes from a unique distribution, with $\alpha^0 = 0.5$ and $u_i \overset{iid}{\sim} |N(0, \sigma_u^2)|$, where $\sigma_u = 1$. In DGP3M, we let $\alpha^0 - u$ to come from $\alpha^0_{(1)} - |N(0, \sigma^2_{u(1)})|$ with probability τ^0 and $\alpha^0_{(2)} - |N(0, \sigma^2_{u(2)})|$ with probability $1 - \tau^0$, where $\alpha^0_{(1)} = 1$, $\alpha^0_{(2)} = -1$, $\sigma_{u(1)} = 0.75$, $\sigma_{u(2)} = 1.25$ and $\tau^0 = 0.5$. It is important to note that the mixture structure of $\alpha^0 - u$ is independent of the grouping structure.

We evaluate the performance of each model and the case for any combination of $N = 100, 250,$ or 500 and $T = 50, 75,$ or 100. Thus, there are $3 \times 2 \times 9 = 54$ different cases. We assess the finite sample properties of our method with 500 MC replications.

Note $(N, T) = (466, 80)$ in the empirical application of the paper, so our simulations, including the recommended tuning parameters in the next section, offer meaningful and practical guidance.

4.2 Choices of Tuning Parameters

We set $m = \lfloor T^{1/5} \rfloor$, where $\lfloor \cdot \rfloor$ denotes the integer part and $\underline{m} = \lfloor (N_k T)^{1/4.8} \rfloor$ for each group k. The value of the two tuning parameters align with standard choices in the literature.

Theoretically, the valid ranges for λ_{NT} and $\tilde{\lambda}_{NT}$ are quite broad. Based on simulation evidence, we recommend setting $\lambda_{NT} = \left(c_\lambda \sqrt{NT} \log(NT)\right)/2$ and $\tilde{\lambda}_{NT} = \left(\tilde{c}_\lambda \sqrt{N} \log N\right)/8$, where c_λ and $\tilde{c}_\lambda$ are constants. These values of λ_{NT} and $\tilde{\lambda}_{NT}$ meet the conditions specified in Proposition 3.3. The constants c_λ and $\tilde{c}_\lambda$ serve as sensitivity parameters, over which we conduct sensitivity analyses for the choice of λ_{NT} and $\tilde{\lambda}_{NT}$. Specifically, we test values of c_λ and $\tilde{c}_\lambda$ in $\{3/2, 1, 3/4\}$, with $c_\lambda = \tilde{c}_\lambda = 1$ as the benchmark setting.

In finding the number of components in the mixture distribution, we restrict our attention to one or two components, i.e., $\mathcal{K} = 1, 2$. We find that the small-sample properties of mixtures with more than two components perform poorly in simulations. We conjecture that this is due to the complicated log-likelihood functions (see, e.g., equations (C.2) and (C.3)), which cannot effectively handle mixtures with more than two components. In Appendix B, we propose an alternative method to identify the mixture structure with potentially more than two components. We also conduct simulations to evaluate its finite-sample performance. The simulation results for this alternative method for allowing mixtures with more than two components suggest that the method performs well in determining the correct number of components, but the parameter estimates can be off. Although not perfect, these findings provide a foundation for future research in this direction.

4.3 Simulation Results

We report results for DGP3M, the most complex model, featuring three groups and a mixture distribution structure in $\alpha_i^0 - u_i$. Results for the remaining DGPs are reported in Appendix F.

We first report the performance of the IC in Step 3 for coefficient groups and Step 5 for the mixture distribution structure in Table 1 for the benchmark specification with $c_\lambda = \tilde{c}_\lambda = 1$. Additionally, Table 1 includes the classification errors, denoted as $\bar{\Pr}(\bar{F})$. It is defined as the average percentage of observations misclassified to other groups across 500 replications. The performance of the IC in Step 3 for selecting the correct number of groups, $K^* = 3$, is reasonable. For $N = 500$, the classification error in Step 3 is less than 1 percent for each $T = 50, 75, 100$. The performance of the Step 5 IC is also strong for DGP3M, choosing the correct specification (mixture distribution) with a probability close to 1. For DPG3U, where $\alpha_i^0 - u_i$ comes from a unique distribution, Table F.10 in Appendix F shows that the probability of Step 5 IC selecting the correct distribution (unique distribution) quickly approaches 1 as N increases. Sensitivity analyses for both ICs in Steps 3 and 5, shown in Tables F.11 - F.19, demonstrate that the results are robust to the selected range of tuning parameters.

We assess the accuracy of the estimates of $\{\sigma_v \text{s}, \alpha_1^0, \sigma_{u(1)}, \alpha_2^0, \sigma_{u(2)}, \tau^0\}$ using two measures: (i) bias (BIAS) and (ii) root mean squared errors (RMSE). The reported values in Table 2, obtained by averaging over 500 MC iterations, are reasonable.

Illustrated in Figures F.1, F.2 and F.3 in Appendix F are the estimates of time-varying frontiers for $(N, T) = (500, 50), (500, 75),$ and $(500, 100)$. Black solid lines depict the true time-varying frontier, dotted lines show the mean of the estimated grouped frontiers averaged over 500 MC iterations, and the gray shaded region depicts the 90th percentile of the estimates. It is clear from Figure F.1 that, while the mean over MC iterations is reasonably close to the true frontiers, the 90th percentile bands are wide, suggesting possible classification errors between neighboring groups. Figures F.2 and F.3 show that the accuracy of frontier grouping improves as T increases. This

is consistent with the theory developed, since the Step 2 classification using HAC is based on $\hat{\vartheta}_i = (\hat{\pi}_i', \hat{\sigma}_{vi})'$ obtained using T observations.

Table 1: Performance of ICs for DGP3M

(N,T)	$K=1$	$K=2$	$K=3$	$K=4$	$\bar{\Pr}(\bar{F})$	α^0-u uni	α^0-u mix
(100,50)	0.000	0.352	0.648	0.000	0.106	0.008	0.992
(100,75)	0.000	0.078	0.922	0.000	0.024	0.000	1.000
(100,100)	0.000	0.000	1.000	0.000	0.024	0.000	1.000
(250,50)	0.000	0.086	0.914	0.000	0.026	0.002	0.998
(250,75)	0.000	0.000	1.000	0.000	0.026	0.000	1.000
(250,100)	0.000	0.000	1.000	0.000	0.026	0.000	1.000
(500,50)	0.000	0.006	0.994	0.000	0.002	0.004	0.996
(500,75)	0.000	0.000	1.000	0.000	0.002	0.000	1.000
(500,100)	0.000	0.000	1.000	0.000	0.002	0.000	1.000

Note: Results for the baseline case $c_\lambda = \tilde{c}_\lambda = 1$. Reported numbers are probabilities across replications.

Table 2: BIAS and RMSE over 500 MC iterations for DGP3M

	$\hat{\sigma}_{v(1)}$		$\hat{\sigma}_{v(2)}$		$\hat{\sigma}_{v(3)}$		$\hat{\tau}$		$\hat{\alpha}^0_{(1)}$		$\hat{\sigma}_{u(1)}$		$\hat{\alpha}^0_{(2)}$		$\hat{\sigma}_{u(2)}$	
(N,T)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
(100,50)	0.268	0.463	0.127	0.215	0.367	0.594	0.021	0.029	0.057	0.074	0.116	0.152	0.120	0.157	0.135	0.171
(100,75)	0.082	0.214	0.074	0.174	0.154	0.372	0.015	0.027	0.042	0.052	0.103	0.145	0.100	0.131	0.126	0.162
(100,100)	0.024	0.099	0.027	0.093	0.053	0.196	0.013	0.018	0.040	0.052	0.090	0.116	0.091	0.118	0.108	0.137
(250,50)	0.204	0.373	0.138	0.243	0.327	0.562	0.013	0.018	0.036	0.049	0.077	0.107	0.080	0.102	0.089	0.112
(250,75)	0.034	0.129	0.035	0.116	0.069	0.241	0.009	0.012	0.026	0.033	0.062	0.079	0.062	0.076	0.070	0.089
(250,100)	0.005	0.033	0.008	0.032	0.014	0.064	0.008	0.011	0.024	0.030	0.057	0.075	0.059	0.074	0.076	0.095
(500,50)	0.145	0.307	0.110	0.218	0.242	0.484	0.011	0.013	0.026	0.035	0.056	0.074	0.053	0.067	0.069	0.087
(500,75)	0.009	0.064	0.009	0.044	0.018	0.100	0.007	0.009	0.018	0.022	0.042	0.055	0.042	0.053	0.053	0.068
(500,100)	0.002	0.003	0.004	0.005	0.007	0.009	0.006	0.008	0.017	0.021	0.041	0.054	0.040	0.053	0.056	0.070

5 Application to the U.S. Commercial Banking Sector

In this section, we apply the developed method for stochastic cost frontier model to analyze the cost efficiency of the U.S. large commercial banks in presence of a series of gradual deregulation that allowed banks to increase their capacity of operation. We use the same dataset used by Feng et al. (2017), and focus our analysis on a sample of banks that operate continuously over the period 1986 to 2005 (thereby mitigating the impact of entry and exit) with assets of at least $1 billion in 1986 dollars. Data supporting the findings of this study are available upon request. As briefly explained in Feng et al. (2017), the banking sector over this period saw a number of gradual deregulation that allowed banks to increase the capacity of operation. In particular, the exact timing of the deregulation varied at the state level, and it was not until June 1997 that banks were allowed to operate across states as a result of the Riegle-Neal Interstate Banking and Branching Efficiency Act of 1994.[5] Given this context, our method that allows us to group banks based on the time-varying frontiers is well suited to capture the effect of gradual deregulation, as well as to analyze the inefficiency of banks in presence of such deregulation.

To set the stage, let $i = 1, 2, \ldots, N$ denote the banks, $t = 1, 2, \ldots, T$ denote the time periods. The data is recorded in quarterly frequency, over 1986 to 2005, with $T = 80$ and consists of $N = 466$ banks. We assume that banks use three inputs to generate three outputs. Specifically, the inputs used are: (i) price of labor, W_{it1}, (ii) price of purchased funds, W_{it2}, and (iii) price of core deposits, W_{it3}. Generated outputs are: (i) consumer loans, Y_{it1}, (ii) non-consumer loans, Y_{it2}, consisting of industrial, commercial, and real estate loans, and (iii) securities, Y_{it3}, which includes all non-loan financial assets. Summary statistics of these variables are reported in Table F.20 in Appendix F.

We estimate a cost frontier $C(Y_{it}, W_{it})$. Accordingly, Y_{it} are output quantities (loan categories,

[5] See Feng et al. (2017) and Jayaratne and Strahan (1997) for more detailed discussion of the history of deregulation in the banking sector.

securities) and W_{itj} are input prices. In particular, W_{it2} is the *price of purchased funds*, not an output; loans are treated as outputs under the intermediation view of banking services. This mapping is consistent with cost duality and with our specification, in which the frontier uses *grouped, smoothly time-varying coefficients* to capture heterogeneous technological regimes under staggered state-level deregulation.

The particular variant of the panel SF model we study is a panel stochastic cost frontier model adapted from Greene (2005b):

$$\begin{aligned}\log c_{it}^{*} = \alpha_i^0 + u_i + \alpha_i(\tau_t) + \beta_{i1}(\tau_t)\log w_{it1} + \beta_{i2}(\tau_t)\log w_{it2} \\ + \beta_{i3}(\tau_t)\log y_{it1} + \beta_{i4}(\tau_t)\log y_{it2} + \beta_{i5}(\tau_t)\log y_{it3} + v_{it},\end{aligned} \tag{5.1}$$

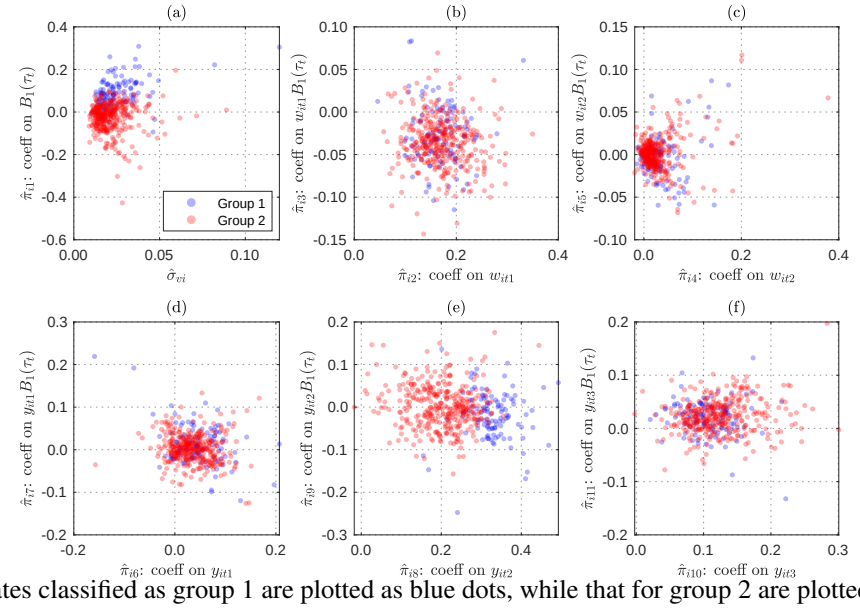
where linear homogeneity is imposed in input prices by the normalizations: $c_{it}^{*} = C_{it}/W_{it3}$, $w_{itl} = W_{itl}/W_{it3}$ for $l = 1, 2$ and $y_{itl} = Y_{itl}/W_{it3}$ for $l = 1, 2, 3$. The inefficiency term $u_i \geq 0$ enters the model positively as cost frontier models are derived from the dual cost minimization problem of the firm where the cost function is assumed to be Cobb-Douglas.

In a cost frontier, interest-rate conditions primarily operate through W_{it}, while local demand affects the output Y_{it}. Our specification already conditions on (Y_{it}, W_{it}) which are allowed to vary smoothly over time and across latent regimes. Adding rate or local controls would double-count channels captured by (Y_{it}, W_{it}).

We estimate the model in (5.1) using the method described in the previous section, setting the tuning parameters as in Section 4.2. We also check the sensitivity of parameter c_λ in Step 3 and $\tilde{c}_\lambda$ in Step 5 of the proposed method as in Section 4.2. Different values of c_λ and $\tilde{c}_\lambda$ deliver the same classification results.

As in the simulations, we set $\bar{K} = 4$. The information criteria in step 3 selects the optimal number of group for the banks to be two, splitting $N = 466$ banks into $(N_1, N_2) = (113, 353)$. Figure 1 depict the scatter plots of the elements in $\hat{\vartheta}_i = (\hat{\pi}_i', \hat{\sigma}_{vi})'$, that collects the parameters

Figure 1: Scatter plot of elements in $\hat{\vartheta}_i = (\hat{\pi}_i', \hat{\sigma}_{vi})'$

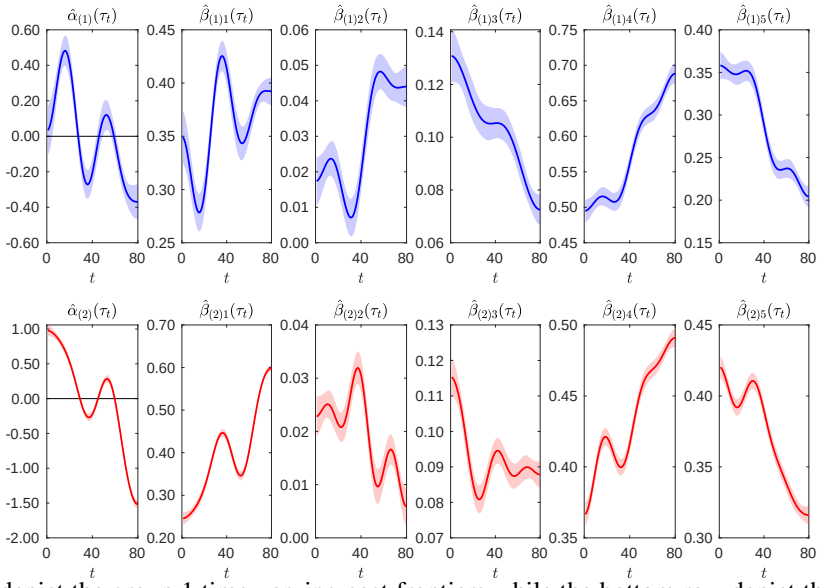


Note: Estimates classified as group 1 are plotted as blue dots, while that for group 2 are plotted as red dots.

obtained from individual level estimation in step 1 for classification in step 2. Note $m = \left\lfloor T^{1/5} \right\rfloor = 2$, so each $\hat{\vartheta}_i$ is a 12 by 1 vector. Individual estimates classified as groups 1 and 2 are depicted as blue and red dots, respectively. Panel (a) depicts the scatter plot of the estimates $\hat{\pi}_{i1}$ against $\hat{\vartheta}_{i12}$, while panels (b)–(f) depict the respective coefficients on the inputs/outputs (w_{itl}, y_{itl}) against $(w_{itl}B_1(\tau_t), y_{itl}B_1(\tau_t))$. We discuss what drives the classification in Appendix F.4.

Figure 2 depicts the frontiers. The top row depicts the time-varying frontiers of group 1, while the bottom row depicts that of group 2. Solid lines in blue and red are the point estimates for group 1 and group 2 respectively, and the shaded regions depict the 95% confidence interval. It is evident that there are substantial time-variations in the estimates, which may be a result of increasing the capacity of operation as a result of deregulation in the banking sector. Figure 3 shows the estimated economies of scale experienced by two groups of banks, $k = 1, 2$ defined by the inverse of the sum of elasticities of output, $1/(\hat{\beta}_{(k)3}(\tau_t) + \hat{\beta}_{(k)4}(\tau_t) + \hat{\beta}_{(k)5}(\tau_t)) - 1$. The estimates on economies of scale are comparable to the ones found in Greene (2005b) and suggest some considerable time-variations for both groups, with group 2 banks enjoying larger economies of scale.

Figure 2: Grouped Frontiers of the U.S. Large Commercial Banks



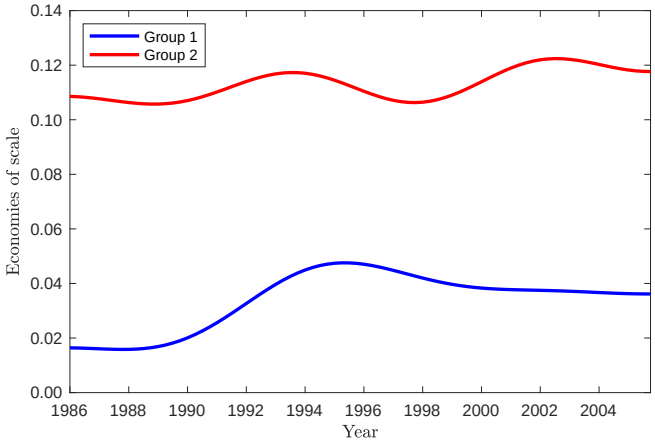
Note: Top row depict the group 1 time-varying cost frontiers while the bottom row depict that of group 2. Solid lines are the point estimates and the shaded regions are 95% confidence interval.

The results from Step 5 of the proposed method suggest that intercept and idiosyncratic random effects inefficiency terms, $\alpha^0 + u$, possess a mixture distribution structure. This result indicates that not only do frontiers form two distinct groups, but so do the level terms that represent the inefficiency of individual banks. The estimated values of the parameters along with the standard errors are presented in Table 3. The results suggest that there are no substantial differences in the standard deviation of random noise, $\hat{\sigma}_v$s, although there are significant differences in the standard deviation of the inefficiency terms.

As we briefly mentioned in Section 2.4, since α_i^0 differs across i, we cannot make a valid ranking of the inefficiencies. Luckily, $\hat{\alpha}^0_{(1)}$ and $\hat{\alpha}^0_{(2)}$ are not statistically different (by the likelihood-ratio test) and as such we can view them the same and construct a ranking of the inefficiencies. Recall that we were estimating cost frontiers. From the estimation results, we can view

$$\alpha_i^0 + u_i \overset{d}{\sim} \begin{cases} \hat{\alpha}^* + \left| N\left(0, \hat{\sigma}^2_{u(1)}\right) \right| & \text{with probability } \hat{\tau} \\ \hat{\alpha}^* + \left| N\left(0, \hat{\sigma}^2_{u(2)}\right) \right| & \text{with probability } 1 - \hat{\tau} \end{cases}, \tag{5.2}$$

Figure 3: Estimates of Economy of Scale



Note: Estimates of the economies of scale for each groups are calculated using the point estimates $\hat{\beta}_{(k)l}(\tau_t)$ for $l = 3, 4, 5$ and $k = 1, 2$.

where $\hat{\alpha}^* = \hat{\tau}\hat{\alpha}^0_{(1)} + (1 - \hat{\tau})\,\hat{\alpha}^0_{(2)}$. We compute $\mathrm{E}\left(\alpha_i^0 + \widehat{u_i|\varepsilon_{i1}, ..., \varepsilon_{iT}}\right)$ using (C.4). We compare the ranking in the homogeneous case where the frontiers and the variances of v_{it} are assumed the same across firms and the inefficiency term comes from one distribution. The result of top 60 is reported in Figure F.4 in Appendix F.5. We can see that the two rankings differ greatly after the top 3. This highlights the importance of classification to ensure valid inference of inefficiency term.

Table 3: Estimates of $\hat{\sigma}_v$s and $\hat{\varrho}$

$\hat{\sigma}_{v(1)}$	$\hat{\sigma}_{v(2)}$	$\hat{\tau}$	$\hat{\alpha}^0_{(1)}$	$\hat{\sigma}_{u(1)}$	$\hat{\alpha}^0_{(2)}$	$\hat{\sigma}_{u(2)}$
0.0862	0.0855	0.8748	0.0157	0.4426	0.6161	0.7756
(0.0041)	(0.0008)	(0.1017)	(0.3960)	(0.0362)	(0.1708)	(0.0235)

Note: Reported in parentheses are the standard errors.

6 Conclusion

In this paper, we develop a general framework for panel SF models with latent group structures. A natural concern is whether allowing for multiple frontiers weakens the interpretation of inef-

ficiency. Our results suggest the opposite. By accounting for latent technological regimes, we prevent unobserved heterogeneity from being mistakenly absorbed into the inefficiency term. Inefficiency in our framework is always measured relative to the appropriate group frontier. This distinction is crucial in empirical applications, such as our U.S. banking study, where ignoring heterogeneity would miscalculate inefficiency.

While it is possible that latent groups capture the influence of omitted inputs, we view this as a feature rather than a flaw: in practice, not all determinants of technology are observable. The latent group structure provides a statistically disciplined way to approximate these unobserved factors, offering a middle ground between a fully homogeneous frontier (too restrictive) and unrestricted firm-specific frontiers (too unstructured). In this sense, multiple frontiers should be interpreted as evidence of distinct technological environments, not as a failure to identify inefficiency.

Two extensions are worth mentioning. First, our framework cannot be directly generalized to endogenous cases where covariates, x are correlated with the error term, v. Extending the framework to accommodate endogeneity is an important direction for future work. Second, it would be valuable to explore a one-step HAC algorithm that avoids the use of information criteria, as proposed by Mugnier (2025).

References

AIGNER, D., C. A. K. LOVELL, AND P. SCHMIDT (1977): "Formulation and Estimation of Stochastic Frontier Production Function Models," *Journal of Econometrics,* 6, 21-37. 1

ANDO, T., AND J. BAI (2016): "Panel Data Models with Grouped Factor Structure under Unknown Group Membership," *Journal of Applied Econometrics,* 31, 163-191. 1

ATAK, A., T. YANG, Y. ZHANG, AND Q. ZHOU (2025): "Specification Tests for Time-Varying Coefficient Panel Data Models," *Econometric Theory*, 41 (1), 123-170. 2.2, A.1, H

BONHOMME, S., AND E. MANRESA (2015): "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 83, 1147-1184. 1

CHEN J. (2019): "Estimating Latent Group Structure in Time-Varying Coefficient Panel Data Models," *Econometrics Journal*, 22, 223-240. 1, 2.3, 3.1

CHEN, Y. Y., P. SCHMIDT, AND H. J. WANG (2014): "Consistent Estimation of the Fixed Effects Stochastic Frontier Model," Journal of Econometrics, 181(2) 65-76. 1, A.1, A.1

CHENG, M., S. WANG, L. XIA, AND X. ZHANG (2024): "Testing Specification of Distribution in Stochastic Frontier Analysis," *Journal of Econometrics,* 239. 2

COLOMBI, R., S. C. KUMBHAKAR, G. MARTINI, AND G. VITTADINI (2018): "Closed-skew Normality in Stochastic Frontiers with Individual Effects and Long/Short-run Efficiency," *Journal of Productivity Analysis,* 42, 123-136. 1

DONG, C., AND O. LINTON (2018): "Additive Nonparametric Models with Time Variable and Both Stationary and Nonstationary Regressors," *Journal of Econometrics,* 207, 212-236. 2.2, H

EVERITT, B. S., S. LANDAU, M. LEESE, AND D. STAHL (2011). Cluster Analysis. 5th ed., Wiley, Wiley Series in Probability and Statistics. 2.3, D

FENG, G., J. GAO, B. PENG, AND X. ZHANG (2017): "A Varying-Coefficient Panel Data Model with Fixed Effects: Theory and an Application to US commercial Banks," *Journal of Econometrics,* 6, 68-82. 5, 5

GALÁN, J. E., AND H. VEIGA, AND M. P. WIPER (2014): "Bayesian Estimation of Inefficiency Heterogeneity in Stochastic Frontier Models," *Journal of Productivity Analysis*, 42, 85-101. 1

GREENE W. (2005a): "Fixed and Random Effects in Stochastic Frontier Models," *Journal of Productivity Analysis*, 23, 7-32. 1, 2.1, A.1

GREENE W. (2005b): "Reconsidering Heterogeneity in Panel Data Estimators of the Stochastic Frontier Model," *Journal of Econometrics*, 126, 269-303. 1, 2, 2.4, 5, 5, A.1, C.3

HUANG, W., S. JIN, AND L. SU (2020): "Identifying Latent Grouped Patterns in Cointegrated Panels," *Econometric Theory*, 36(3), 410-456. 1

JAYARATNE, J., AND P. E. STRAHAN (1997): "The Benefits of Branching Deregulation," *Economic Policy Review,* 3(4), 13-29. 5

JONDROW, J., C. A. K. LOVELL, I. M. MATEROV, AND P. SCHMIDT (1982): "On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model," *Journal of Econometrics,* 19(2-3), 233-238. 1

KUMBHAKAR, S. C., G. LIEN, AND J. B. HARDAKER (2014): "Technical Efficiency in Competing Panel Data Models: A study of Norwegian Grain Farming," *Journal of Productivity Analysis,* 41, 321-337. 1

KUMBHAKAR, S. C., AND C. A. K. LOVELL(2000). Stochastic Frontier Analysis, Cambridge University Press. 1

KUMBHAKAR, S. C., C. PARMETER, AND V. ZELENYUK(2022): "Stochastic Frontier Analysis: Foundations and Advances I," *Handbook of Production Economics* ed. by S. C. Ray, R. G. Chambers, and S. C. Kumbhakar, Springer, Chap. 8, pp. 331-370. 1

LAI H. P., AND S. C. KUMBHAKAR (2023): "Panel Stochastic Frontier Model With Endogenous Inputs and Correlated Random Components," *Journal of Business & Economic Statistics*, 41:1, 80-96. 2, 2.1

LIN, C.C. AND S. NG (2012): "Estimation of Panel Data Models with Parameter Heterogeneity when Group Membership is Unknown," *Journal of Econometric Methods*, 1(1):42-55. 1

LOYO, J. A., AND T. BOOT (2024): "Grouped Heterogeneity in Linear Panel Data Models with Heterogeneous Error Variances," *Journal of Business & Economic Statistics*, 1-13. 1

MEEUSEN, W., AND J. VAN DEN BROECK (1977): "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error," *International Economic Review,* 18, 435-444. 1

MUGNIER, M. (2025): "A Simple and Computationally Trivial Estimator for Grouped Fixed Effects Models," *Journal of Econometrics*, 250, 106011. 6

PARK, B. U., AND L. SIMAR (1994): "Efficient Semiparametric Estimation in a Stochastic Frontier Model," *Journal of the American Statistical Association*, 89(427), 929-936. 2

SU, L., Z. SHI, AND P. C. B. PHILLIPS (2016): "Identifying Latent Structures in Panel Data," *Econometrica*, 84(6), 2215-2264. 1

SU, L., X. WANG, AND S. JIN (2019): "Sieve Estimation of Time-Varying Panel Data Models With Latent Structures," *Journal of Business & Economic Statistics*, 37(2), 334-349. 1, H

TSIONAS, E. G AND S. C. KUMBHAKAR (2014): "Firm Heterogeneity, Persistent and Transient Technical Inefficiency: A Generalized True Random-Effects Model," *Journal of Applied Econometrics*, 29(1), 110–132. 1, 2.1

TSIONAS, M., F. C. PARMETER, AND V. ZELENYUK (2023): "Bayesian artificial neural networks for frontier efficiency analysis," *Journal of Econometrics*, 236(2), 105491. 1

WANG, W. AND L. SU (2021): "Identifying Latent Group Structures in Nonlinear Panels," *Journal of Econometrics*, 220(2), 272-295. 1

YAO F., F. ZHANG, AND S. C. KUMBHAKAR (2019): "Semiparametric Smooth Coefficient Stochastic Frontier Model With Panel Data," *Journal of Business & Economic Statistics*, 37(3), 556-572. 1, 2, 2.1, 2.1, 3.2, 4.1, C.1

Online Appendix to

"Panel Stochastic Frontier Models with Latent Group Structures"

(NOT for Publication)

Additional Notation. For the deterministic series $\{a_n, b_n\}_{n=1}^{\infty}$, we denote $a_n \lesssim b_n$ if $\limsup_{n\to\infty} |a_n/b_n| \leq C$ for some constant C that does not depend on n, $a_n \gtrsim b_n$ if $b_n \lesssim a_n$, $a_n \ll b_n$ if $a_n = o(b_n)$, and $a_n \gg b_n$ if $b_n \ll a_n$. $\propto_P$ denotes proportional in probability, e.g., $x_n \propto_P y_n$ indicates that both $x_n = O_P(y_n)$ and $y_n = O_P(x_n)$ hold. A^c denotes the complement of A. C and M denote some positive constants that may vary from line to line.

A Other Stochastic Frontier Models

A.1 Fixed Effects Model with Distributional Conditions

We consider the fixed effects (FE) SF model when error terms are assumed to follow certain parametric distributions. The classification of the FE-SF model follows naturally from the methodology developed in the main body of the paper. Incorporating time-varying heterogeneous coefficients, the model proposed by Greene (2005a,b) and Chen et al. (2014) takes the form:

$$\begin{aligned} y_{it} &= \alpha_i^0 + \alpha_i(\tau_t) + x_{it}'\beta_i(\tau_t) + \varepsilon_{it} \\ &= \alpha_i^0 + \alpha_i(\tau_t) + x_{it}'\beta_i(\tau_t) + v_{it} - u_{it}, \end{aligned}$$

where $\varepsilon_{it} = v_{it} - u_{it}$, and the normalization condition $\int_0^1 \alpha_i(\tau_t)\, d\tau_t = 0$ is imposed. We note that this normalization is innocuous as in Atak et al. (2025), because the the constant (even varying acorss i) can be absorbed in α_i^0. The standard assumptions in these papers include:

$$v \perp u \perp (\alpha^0, x), \quad v \sim N(0, \sigma_v^2), \quad u \sim |N(0, \sigma_u^2)|,$$

with $\{v_{it}, u_{it}\}$ independent across time t. To address the incidental parameters problem, Chen et al. (2014) proposed the within MLE. Allowing for heterogeneous distributions across firms, we have:

$$v_{it} \sim N(0, \sigma_{vi}^2), \quad u_{it} \sim |N(0, \sigma_{ui}^2)|.$$

We next define the group structure. Specifically, we assume that there are K^* distinct groups of parameter sets, and each firm's parameters belong to one of these groups. Formally:

$$\{\alpha_i(\tau_t), \beta_i(\tau_t), \sigma_{vi}, \sigma_{ui}\} = \sum_{k=1}^{K^*} \left\{\alpha_{(k)}^*(\tau_t), \beta_{(k)}^*(\tau_t), \sigma_{v(k)}^*, \sigma_{u(k)}^*\right\} \cdot \mathbf{1}(i \in G_k). \tag{A.1}$$

Parameters across different groups are distinct, and each firm is uniquely assigned to one of the K^* groups.

The key distinction here is that we also classify firms based on σ_{ui}. The intuition is as follows: unlike the framework in the main body of the paper, we can consistently estimate σ_{ui} for each firm i individually, because we allow u_{it} to vary over time. This variation enables us to estimate σ_{ui} without pooling observations across firms. Thanks to this insight, we only need to apply Steps 1–3 in Section 2.3 for the FE model. Before detailing the procedure, we approximate the model using sieve expansions as in equation (2.5), leading to:

$$\begin{aligned} y_{it} &= \alpha_i^0 + \alpha_i(\tau_t) + \sum_{l=1}^{p} x_{itl}\beta_{il}(\tau_t) + \varepsilon_{it} \\ &\approx \alpha_i^0 + \mathbb{B}_{-0}^m(\tau_t)'\pi_{i0}^0 + \sum_{l=1}^{p} x_{itl}\mathbb{B}^m(\tau_t)'\pi_{il}^0 + \varepsilon_{it} \\ &= \alpha_i^0 + z_{it}'\pi_i^0 + \varepsilon_{it} = \alpha_i^0 + z_{it}'\pi_i^0 + v_{it} - u_{it}, \end{aligned} \tag{A.2}$$

where z_{it} collects all the basis function terms and their interactions with covariates. We now outline the estimation procedure.

Step 1*: Individual Estimation

We begin by applying the within transformation, and define the following notation:

$$\ddot{y}_{it} = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it}, \quad \ddot{z}_{it} = z_{it} - \frac{1}{T}\sum_{t=1}^{T} z_{it},$$

with $\ddot{v}_{it}$, $\ddot{u}_{it}$, and $\ddot{\varepsilon}_{it}$ defined analogously. The transformed model becomes:

$$\ddot{y}_{it} \approx \ddot{z}'_{it}\pi_i^0 + \ddot{\varepsilon}_{it} = \ddot{z}'_{it}\pi_i^0 + \ddot{v}_{it} - \ddot{u}_{it}.$$

Let $\ddot{\varepsilon}_i = (\ddot{\varepsilon}_{i1}, \ddot{\varepsilon}_{i2}, \ldots, \ddot{\varepsilon}_{i,T-1})'$ denote the vector of the first $T-1$ transformed residuals. Based on the results in Chen et al. (2014), $\ddot{\varepsilon}_i$ follows a closed skew-normal distribution:

$$\begin{aligned} \mathrm{CSN}_{T-1,T}\Bigg(& 0_{T-1}, (\sigma_{ui}^2 + \sigma_{vi}^2)\left(I_{T-1} - \frac{1}{T-1}\iota_{T-1}\iota'_{T-1}\right), \\ & - \frac{\sigma_{ui}/\sigma_{vi}}{\sqrt{\sigma_{ui}^2 + \sigma_{vi}^2}} \begin{pmatrix} I_{T-1} \\ -\iota'_{T-1} \end{pmatrix}, 0_T, I_T + \frac{\sigma_{ui}^2}{T\sigma_{vi}^2}\iota_T\iota'_T \Bigg), \end{aligned} \tag{A.3}$$

where 0_{T-1} is a $(T-1)\times 1$ vector of zeros, and ι_T is a $T\times 1$ vector of ones. The notation $\mathrm{CSN}_{p,q}$ denotes the closed skew-normal distribution with the following density for a p-dimensional random variable S:

$$f_{\mathrm{CNS}}(s) = C\,\phi_p(s;\mu,\Sigma)\,\Phi_q(D(s-\mu);\upsilon,\Delta),$$

where:

- $\phi_p(\cdot;\mu,\Sigma)$ is the density of a p-dimensional normal distribution,
- $\Phi_q(\cdot;\upsilon,\Delta)$ is the CDF of a q-dimensional normal distribution,
- $\mu \in \mathbb{R}^p$, $\Sigma \in \mathbb{R}^{p\times p}$,

- $D \in \mathbb{R}^{q \times p}$, $\upsilon \in \mathbb{R}^{q}$, and $\Delta \in \mathbb{R}^{q \times q}$.

Since $\ddot{\varepsilon}_{it} \approx \ddot{y}_{it} - \ddot{z}_{it}'\pi_i^0$, we can estimate the parameters $(\pi_i^0, \sigma_{ui}, \sigma_{vi})$ by substituting the expression for $\ddot{\varepsilon}_{it}$ into the density function in (A.3) and applying MLE. This yields consistent estimators $(\hat{\pi}_i^0, \hat{\sigma}_{ui}, \hat{\sigma}_{vi})$.

Step 2*: Classification

From Step 1*, we obtain individual estimates for each firm:

$$\left(\hat{\pi}_1^0, \hat{\sigma}_{u1}, \hat{\sigma}_{v1}\right), \left(\hat{\pi}_2^0, \hat{\sigma}_{u2}, \hat{\sigma}_{v2}\right), \ldots, \left(\hat{\pi}_N^0, \hat{\sigma}_{uN}, \hat{\sigma}_{vN}\right).$$

As in the original Step 2, given a prespecified number of groups K, we estimate the group membership structure. The estimated partition of the N firms is denoted by:

$$\left(\hat{G}_{1|K}, \hat{G}_{2|K}, \ldots, \hat{G}_{K|K}\right),$$

which forms a disjoint partition of the index set $\{1, 2, \ldots, N\}$.

Step 3*: Post-Classification Estimation and Determination of K^*

As in Step 3, we set the number of sieve terms to $\underline{m}$, which is substantially larger than m. We define the new set of regressors as $\underline{z}_{it}$:

$$\underline{z}_{it} = \left[\mathbb{B}_{-0}^{\underline{m}}(\tau_t)', (x_{it} \otimes \mathbb{B}^{\underline{m}}(\tau_t))'\right]',$$

and approximate $\alpha(\cdot)$ and $\beta(\cdot)$ accordingly.

Within each group, say $\hat{G}_{k|K}$ for $1 \leq k \leq K$, we perform post-classification estimation using MLE by maximizing the CSN density across all observations within the group. Specifically,

$$\left(\hat{\pi}_{(k|K)}, \hat{\sigma}^2_{u(k|K)}, \hat{\sigma}^2_{v(k|K)}\right) = \arg \max_{(\pi, \delta_u^2, \delta_v^2)} \sum_{i \in \hat{G}_{k|K}} \log f_{\text{CNS}}\left(\ddot{y}_i - \ddot{\underline{z}}_i\pi; \delta_u^2, \delta_v^2\right),$$

where $\ddot{y}_i$ and $\ddot{z}_i$ collect the first $T-1$ elements of $\ddot{y}_{it}$ and $\ddot{z}_{it}$, respectively, and $f_{\text{CNS}}(\cdot;\delta_u^2,\delta_v^2)$ denotes the CSN density evaluated at the specified variance parameters.

To determine the number of groups, we use the following information criterion:

$$\text{IC}_{\text{FE}}(K,\lambda_{NT}^{\text{FE}}) = -\sum_{k=1}^{K}\left\{\sum_{i\in\hat{G}_{k|K}}\log f_{\text{CNS}}\left(\ddot{y}_i-\ddot{z}_i\hat{\pi}_{(k|K)};\hat{\sigma}^2_{u(k|K)},\hat{\sigma}^2_{v(k|K)}\right)\right\}+\lambda_{NT}^{\text{FE}}K,$$

where λ_{NT}^{FE} is an appropriate penalty term.

The optimal number of groups is chosen as

$$\hat{K}(\lambda_{NT}) = \arg\min_{K=0,1,\ldots,\bar{K}}\text{IC}_{\text{FE}}(K,\lambda_{NT}),$$

for a suitably chosen upper bound $\bar{K}$. For simplicity, we denote this as $\hat{K}$. The final parameter estimates are then given by

$$\hat{\vartheta}_{(k|\hat{K})} = \left(\hat{\pi}_{(k|\hat{K})},\hat{\sigma}_{u(k|\hat{K})},\hat{\sigma}_{v(k|\hat{K})}\right), \quad k=1,2,\ldots,\hat{K}.$$

An investigation of the small sample properties of this procedure is beyond the scope of this paper and is left for future research.

A.2 Nonparametric Fixed Effects Model

One strand of the panel FE SF literature attempts to avoid imposing distributional assumptions on the error terms; see, e.g., Zhou et al. (2020). We take the framework in that paper as an example to demonstrate that our approach can be readily extended to such settings. The model remains the same as in the previous section:

$$\begin{aligned} y_{it} &= \alpha_i^0+\alpha_i(\tau_t)+x_{it}'\beta_i(\tau_t)+\varepsilon_{it} \\ &= \alpha_i^0+\alpha_i(\tau_t)+x_{it}'\beta_i(\tau_t)+v_{it}-u_{it}, \end{aligned}$$

with the constraint $\int_0^1\alpha_i(\tau_t)\,d\tau_t=0$. As in the last section, this normalization is innocuous due to the fixed effects α_i^0.

As noted in Zhou et al. (2020), the cost of this relaxation is that the covariates influencing the inefficiency term differ from those appearing in the frontier function. To accommodate this, we denote the covariate that solely affects the inefficiency term by h_{it} and, without loss of generality, assume that h_{it} is univariate to simplify notation and analysis. Formally, we impose the restriction:

$$\mathrm{E}(u_{it} \mid h_{it}, x_i) = \mathrm{E}(u_{it} \mid h_{it}) \equiv \varpi_i(h_{it}).$$

For the noise term v_{it}, we assume the usual conditional moment condition:

$$\mathrm{E}(v_{it} \mid h_{it}, x_i) = 0.$$

We normalize h_{it} so that its support is $[0, 1]$, which enables the use of orthogonal basis functions, as introduced in Section 2.2, to approximate $\varpi_i(h_{it})$. The key idea in Zhou et al. (2020) is to isolate $\varpi_i(h_{it})$ from the inefficiency term, yielding a new error term with zero conditional mean. Specifically,

$$\begin{aligned} y_{it} &= \alpha_i^0 + \alpha_i(\tau_t) + x_{it}'\beta_i(\tau_t) - \varpi_i(h_{it}) + v_{it} - [u_{it} - \varpi_i(h_{it})] \\ &\equiv \alpha_i^0 + \alpha_i(\tau_t) + x_{it}'\beta_i(\tau_t) - \varpi_i(h_{it}) + \varepsilon_{it}^*. \end{aligned}$$

To identify ϖ_i, we impose the additional normalization:

$$\int_0^1 \varpi_i(h)\,\mathrm{d}h = 0.$$

This normalization implies that inefficiency levels cannot be directly compared across firms, which is a limitation of the approach in Zhou et al. (2020). However, one can still analyze how inefficiency varies with changes in h. We refer readers to the original paper for further details. Under the above assumptions, the new error term ε_{it}^* satisfies:

$$\mathrm{E}(\varepsilon_{it}^* \mid h_{it}, x_i) = 0,$$

which behaves like a standard error term in fixed effects panel data models.

In the present setting, our goal is to estimate and classify both the frontier function and the conditional mean of the inefficiency term. The variances of the error components (v_{it} and ε_{it}^*)

are of secondary importance. This contrasts with earlier settings, where the variance of v_{it} was essential for the likelihood function in MLE for estimating inefficiency distribution. In our case, however, we directly estimate the conditional mean of the inefficiency $\mathrm{E}(u_{it} \mid h_{it})$ as $\varpi_i(h_{it})$. With this in mind, we define the group structure based solely on the frontier and the conditional mean of the inefficiency term. Specifically, we assume that there are K^* distinct groups of parameter sets, and each firm belongs to exactly one of these groups. Formally, we write:

$$\{\alpha_i(\tau_t), \beta_i(\tau_t), \varpi_i(\cdot)\} = \sum_{k=1}^{K^*} \left\{\alpha^*_{(k)}(\tau_t), \beta^*_{(k)}(\tau_t), \varpi^*_{(k)}(\cdot)\right\} \cdot \mathbf{1}(i \in G_k).$$

Each parameter set across the K^* groups is distinct, and each firm is uniquely assigned to one of these groups.

The approximation using orthogonal basis functions can be conducted in a similar manner. For each firm i, we have:

$$\begin{aligned}
y_{it} &= \alpha_i^0 + \alpha_i(\tau_t) + x'_{it}\beta_i(\tau_t) - \varpi_i(h_{it}) + \varepsilon^*_{it} \\
&\approx \alpha_i^0 + \mathbb{B}^m_{-0}(\tau_t)'\pi^0_{i0} + \sum_{l=1}^{p} x_{itl}\mathbb{B}^m(\tau_t)'\pi^0_{il} + \mathbb{B}^m_{-0}(h_{it})'\pi^0_{ip+1} + \varepsilon^*_{it} \\
&\equiv \alpha_i^0 + \left[\mathbb{B}^m_{-0}(\tau_t)', (x_{it} \otimes \mathbb{B}^m(\tau_t))', \mathbb{B}^m_{-0}(h_{it})'\right] \pi_i^0 + \varepsilon^*_{it} \\
&\equiv \alpha_i^0 + z'_{it}\pi_i^0 + \varepsilon^*_{it},
\end{aligned}$$

where we slightly abuse notation by reusing z_{it} and π_i^0, although they represent different quantities in this section. Specifically,

$$z_{it} \equiv \left[\mathbb{B}^m_{-0}(\tau_t)', (x_{it} \otimes \mathbb{B}^m(\tau_t))', \mathbb{B}^m_{-0}(h_{it})'\right]' \quad \text{and} \quad \pi_i^0 \equiv \left(\pi^{0\prime}_{i0}, \pi^{0\prime}_{i1}, \ldots, \pi^{0\prime}_{ip}, \pi^{0\prime}_{ip+1}\right)'.$$

With this setup, the classification and post-classification estimation become straightforward and follow the essentially same procedure as Steps 1–3 in Section 2.3.

Step 1*: Individual Estimation

Like Step 1^*, we begin by applying the within transformation for each firm i. We reuse following notation:

$$\ddot{y}_{it} = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it}, \quad \ddot{z}_{it} = z_{it} - \frac{1}{T}\sum_{t=1}^{T} z_{it},$$

with $\ddot{\varepsilon}^*_{it}$ defined analogously. The transformed model becomes:

$$\ddot{y}_{it} \approx \ddot{z}'_{it}\pi_i^0 + \ddot{\varepsilon}^*_{it}, t = 1, 2, ..., T.$$

We regress $\ddot{y}_{it}$ on $\ddot{z}_{it}$ for each fixed i using observations of $t = 1, 2, ..., T$.

Specifically,

$$\widehat{\pi}_i = \left(\sum_{t=1}^{T} \ddot{z}_{it}\ddot{z}'_{it}\right)^{-1} \left(\sum_{t=1}^{T} \ddot{z}_{it}\ddot{y}_{it}\right),$$

based on which we form groups.

Step 2*: Classification

From Step $1^\star$, we obtain individual estimates for each firm:

$$\widehat{\pi}_1, \widehat{\pi}_2, ..., \widehat{\pi}_N.$$

Given a pre-specified number of groups K, we estimate the group membership structure using the HAC algorithm. The estimated partition of the N firms is denoted by:

$$\left(\hat{G}_{1|K}, \hat{G}_{2|K}, \ldots, \hat{G}_{K|K}\right),$$

which forms a disjoint partition of the index set $\{1, 2, \ldots, N\}$.

Step 3*: Post-Classification Estimation and Determination of K^*

As in Step 3, we set the number of sieve terms to $\underline{m}$, which is substantially larger than m. We define the new set of regressors as $\underline{z}_{it}$:

$$\underline{z}_{it} = \left[\mathbb{B}^{\underline{m}}_{-0}(\tau_t)', \left(x_{it} \otimes \mathbb{B}^{\underline{m}}(\tau_t)\right)', \mathbb{B}^{\underline{m}}_{-0}\left(h_{it}\right)'\right]',$$

and approximate $\alpha(\cdot)$, $\beta(\cdot)$ and $\varpi(\cdot)$ accordingly.

Within each group, say $\hat{G}_{k|K}$ for $1 \leq k \leq K$, we perform the post-classification within estimation. Specifically,

$$\hat{\pi}_{(k|K)} = \left(\sum_{i \in \hat{G}_{k|K}} \sum_{t=1}^{T} \ddot{z}_{it} \ddot{z}_{it}' \right)^{-1} \left(\sum_{i \in \hat{G}_{k|K}} \sum_{t=1}^{T} \ddot{z}_{it} \ddot{y}_{it} \right).$$

To determine the number of groups, we use the following information criterion:

$$\begin{aligned} &\mathrm{IC}_{\mathrm{FENP}}(K, \lambda_{NT}^{\mathrm{FENP}}) \\ &= -\sum_{k=1}^{K} \left\{ \frac{N_k T}{2} \log \left[\frac{1}{(T-1) N_k} \sum_{i \in \hat{G}_{k|K}} \sum_{t=1}^{T} \left(\ddot{y}_{it} - \ddot{z}_{it}' \hat{\pi}_{(k|K)} \right)^2 \right] \right\} + \lambda_{NT}^{\mathrm{FENP}} K, \end{aligned}$$

where $\lambda_{NT}^{\mathrm{FENP}}$ is an appropriate penalty term.

The optimal number of groups is chosen as

$$\hat{K}(\lambda_{NT}^{\mathrm{FENP}}) = \arg \min_{K=0,1,\ldots,\bar{K}} \mathrm{IC}_{\mathrm{FENP}}(K, \lambda_{NT}^{\mathrm{FENP}}),$$

for a suitably chosen upper bound $\bar{K}$. For simplicity, we denote this as $\hat{K}$. The final parameter estimates are then given by

$$\hat{\pi}_{(k|\hat{K})}, \quad k = 1, 2, \ldots, \hat{K}.$$

The theoretical properties of the above procedure are relatively straightforward. However, complications may arise when h_{it} is multidimensional. In such cases, one may need to employ tensor product bases to approximate $\varpi(\cdot)$, which could result in a slower convergence rate. As an alternative, assuming an additive structure for $\varpi(\cdot)$ allows the convergence rate to remain unchanged. A rigorous investigation of the theoretical properties under these settings, as well as an analysis of the small-sample performance, is beyond the scope of this paper and is left for future research.

A.3 A Group Innocuous Normalization

As discussed in Section 2.1, the normalization $\int_0^1 \alpha_i(s)\, \mathrm{d}s = 0$ is not innocuous when $\alpha_i(s)$ varies across firms and possesses a group structure, because we need to assume that the levels of $\alpha_{(k)}^*(\tau_t)$

(that is, $\int_0^1 \alpha^*_{(k)}(s)\,\mathrm{d}s$) are the same across groups. In this section, we generalize the result in the main paper to the case in which we allow the levels to differ across groups but they remain the same within each group.

As before, we assume that there are K^* groups of specific parameters, and each firm's parameters belong to one of these groups:

$$\{\alpha_i(\tau_t), \beta_i(\tau_t), \sigma_{vi}\} = \sum_{k=1}^{K^*} \left\{\alpha^*_{(k)}(\tau_t), \beta^*_{(k)}(\tau_t), \sigma^*_{v(k)}\right\} \mathbf{1}(i \in G_k).$$

Note that our procedure is silent on the classification of the levels of $\alpha_i(\tau_t)$, so the classification is based solely on the time-varying part of $\alpha_i(\tau_t)$. As such, we titled the section "Group" innocuous, since it is not entirely innocuous. In other words, for all $i \in G_k$, we need to assume that $\int_0^1 \alpha_i(s)\,\mathrm{d}s = c_k$ (before normalization), which is identical within the group.

We recommend this procedure when, on average, each group contains at least a few hundred observations, due to the difficulty of uncovering the mixture structure with very small samples; see, e.g., Olson et al. (1980), Simar and Wilson (2010), and Christian et al. (2018).

A consequence of weakening the restriction on the levels is that the distribution of $\alpha_i^0 - u_i$ naturally differs across groups, because for group k, $\int_0^1 \alpha_i(s)\,\mathrm{d}s = c_k$ is absorbed into α_i^0 after normalization. As such, we assume that for observation i in group k, there exists a $\mathcal{K}^*_{(k)} \geq 1$ such that

$$\alpha_i^0 - u_i \overset{d}{\sim} \alpha^0_{(k)(j)} - \left|N\left(0, \sigma^2_{u(k)(j)}\right)\right| \quad \text{with probability } \tau^0_{(k)j}, \quad j = 1, 2, \ldots, \mathcal{K}^*_{(k)},$$

where $0 < \tau^0_{(k)j} < 1$ and $\sigma^2_{u(k)(j)} > C > 0$ for $j = 1, 2, \ldots, \mathcal{K}^*_{(k)}$, with $\tau^0_{(k)1} + \tau^0_{(k)2} + \ldots + \tau^0_{(k)\mathcal{K}^*_{(k)}} = 1$, and $\left(\alpha^0_{(k)(j)}, \sigma^2_{u(k)(j)}\right)$ differ across $j = 1, 2, \ldots, \mathcal{K}^*_{(k)}$. While the notation is rather tedious, the estimation remains straightforward. We simply estimate the distribution of $\alpha_i^0 - u_i$ for each group separately using the previous strategy.

We present the details of the procedure below. The generalization that allows the levels to differ across groups has no impact on the first three steps, as the intercept term is not involved.

Step 1$^\blacklozenge$: Individual Estimation

Same as Step 1.

Step 2$^\blacklozenge$: Classification

Same as Step 2.

Step 3$^\blacklozenge$: Post-Classification Estimation and Determination of K^*

Same as Step 3.

Step 4$^\blacklozenge$: Estimation of $\alpha^0_{(k)(j)}, \sigma^2_{u(k)(j)},$ and $\tau^0_{(k)j}$

For group k, assuming that the error term comes from $\mathcal{K}_{(k)}$ distributions, we obtain an estimate of $\left(\alpha^0_{(k)(1)}, \sigma^2_{u(k)(1)}, ..., \alpha^0_{(k)(\mathcal{K}_{(k)})}, \sigma^2_{u(k)(\mathcal{K}_{(k)})}, \tau^0_{(k)1}, ..., \tau^0_{(k)\mathcal{K}_{(k)}-1}\right)$ using MLE as follows:

$$\begin{aligned}&\left(\hat{\alpha}^0_{(k)(1)}, \hat{\sigma}^2_{u(k)(1)}, ..., \hat{\alpha}^0_{(k)(\mathcal{K}_{(k)})}, \hat{\sigma}^2_{u(k)(\mathcal{K}_{(k)})}, \hat{\tau}_{(k)1}, ..., \hat{\tau}_{(k)\mathcal{K}_{(k)}-1}\right)\\ =&\arg\max_{(s,\delta^2_u,\tau)} \sum_{i\in\hat{G}_{k|\hat{K}}} \log\tilde{f}\left(y_i \left| x_i; s_1, \delta^2_{u(1)}, ..., s_{\mathcal{K}_{(k)}}, \delta^2_{u\left(\mathcal{K}_{(k)}\right)}, \tau_1, ..., \tau_{\mathcal{K}_{(k)}-1}, \hat{\vartheta}_{(k|\hat{K})}\right.\right)\end{aligned}$$

where $\tilde{f}$ is the likelihood function, defined in (C.3), and we incorporate estimates from Step 3$^\blacklozenge$.

Collecting results from Steps 4$^\blacklozenge$, we proceed to Step 5$^\blacklozenge$ to determine the specification of $\alpha^0_i - u_i$ for each group.

Step 5$^\blacklozenge$: Determination of the Distributional Structures of the Inefficiency Term

The information criterion for group k is constructed as

$$\widetilde{\mathrm{IC}}_k(\mathcal{K}_{(k)}, \tilde{\lambda}_{NT}) = -\sum_{i\in\hat{G}_{k|\hat{K}}} \log\tilde{f}\left(y_i \left| x_i; s_1, \delta^2_{u(1)}, ..., s_{\mathcal{K}_{(k)}}, \delta^2_{u\left(\mathcal{K}_{(k)}\right)}, \tau_1, ..., \tau_{\mathcal{K}_{(k)}-1}, \hat{\vartheta}_{(k|\hat{K})}\right.\right) + \mathcal{K}_{(k)}\tilde{\lambda}_{NT},$$

where $\tilde{\lambda}_{NT}$ is a suitable penalty term and $k = 1, 2, ..., \hat{K}$. We take

$$\hat{\mathcal{K}}_{(k)}(\tilde{\lambda}_{NT}) = \arg\min_{\mathcal{K}_{(k)}=0,1,...,\bar{\mathcal{K}}} \widetilde{\mathrm{IC}}_k(\mathcal{K}_{(k)}, \tilde{\lambda}_{NT}),$$

and we write $\hat{\mathcal{K}}_{(k)}$ for short.

Collecting results across $\hat{K}$ groups, the estimated parameters are

$$\left\{\left(\hat{\alpha}^0_{(k)(1)}, \hat{\sigma}^2_{u(k)(1)}, ..., \hat{\alpha}^0_{(k)(\hat{\mathcal{K}}_{(k)})}, \hat{\sigma}^2_{u(k)(\hat{\mathcal{K}}_{(k)})}, \hat{\tau}_{(k)1}, ..., \hat{\tau}_{(k)\hat{\mathcal{K}}_{(k)}-1}\right)\right\}_{k=1}^{\hat{K}}.$$

The theoretical properties of the above procedure remain the same as those in the main body of the paper, provided that $N_k \propto N$ for $k = 1, 2, \ldots, K^*$. Before we conclude this section, we emphasize that we recommend this procedure only when practitioners have ample observations available, e.g., at least a few hundred for each group.

A.4 Latent Structure with Zero Inefficiency

Kumbhakar et al. (2013) and Rho and Schmidt (2015) proposed the zero inefficiency stochastic frontier (ZISF) model, in which the inefficiency term is exactly zero with a positive probability. Testing the validity of the ZISF model amounts to examining whether the variance of the inefficiency term is zero in one component of the mixture distribution. This leads to a nonstandard testing problem, as the null hypothesis places the parameter on the boundary of the parameter space. Consequently, standard inference methods such as the t-test may not be appropriate, and alternative testing procedures, as suggested by Kumbhakar et al. (2013), may be required. Compounding the challenge, Rho and Schmidt (2015) highlighted several identification issues inherent in this framework. Given these complications, it is worthwhile to explore this question from a different perspective.

We propose to address this problem using information criteria, following a similar approach to that used in the main body of the paper. To this end, we incorporate the ZISF assumption into our model. Naturally, this introduces an additional layer of complexity and difficulty. For clarity of exposition, we focus on a mixture distribution with at most two components. While extending to more components is conceptually straightforward, it may distract from the central focus of this

section. We adopt the model from the main body of the paper, maintaining the same group structure on the frontier. The only difference lies in the assumption regarding the distribution of $\alpha_i^0 - u_i$.

Specifically, we consider the following three specifications and propose an IC to choose one.

$$\text{Spec1: } \alpha_i^0 - u_i \overset{d}{\sim} \alpha^0 - N\left(0, \sigma_u^2\right),$$

$$\text{Spec2: } \alpha_i^0 - u_i \overset{d}{\sim} \begin{cases} \alpha_{(1)}^0 & \text{with probability } \tau^0 \\ \alpha_{(2)}^0 - N\left(0, \sigma_{u(2)}^2\right) & \text{with probability } 1 - \tau^0 \end{cases}, \text{ and}$$

$$\text{Spec3: } \alpha_i^0 - u_i \overset{d}{\sim} \begin{cases} \alpha_{(1)}^0 - N\left(0, \sigma_{u(1)}^2\right) & \text{with probability } \tau^0 \\ \alpha_{(2)}^0 - N\left(0, \sigma_{u(2)}^2\right) & \text{with probability } 1 - \tau^0 \end{cases},$$

where $0 < \tau^0 < 1$, $\sigma_u^2, \sigma_{u(1)}^2, \sigma_{u(2)}^2 > 0$, and $\left(\alpha_{(1)}^0, \sigma_{u(1)}^2\right) \neq \left(\alpha_{(2)}^0, \sigma_{u(2)}^2\right)$. Note that S2 is the assumption in the ZISF model. We do not consider the case when $\alpha_i^0 - u_i = \alpha^0$ because the fully efficient case is rare in practice.

We slightly abuse notation by using the same symbols for parameters in both Spec2 and Spec3. However, the context should make it clear which set of parameters is being referenced.

Recall that the only change in this setting is the assumption regarding the distribution of $\alpha_i^0 - u_i$. Consequently, the estimation of the frontier parameters proceeds exactly as before — specifically, following Steps 1 to 3 in Section 2.3. For clarity, we relabel these steps as Step $1^\blacktriangledown$, Step $2^\blacktriangledown$, and Step $3^\blacktriangledown$.

Steps $1^\blacktriangledown$, $2^\blacktriangledown$, and $3^\blacktriangledown$

Same as Steps 1 to 3.

Step 4$^{\blacktriangledown}$: Estimation of $\alpha^0, \sigma_u^2,$ and τ^0

For Spec1, (α^0, σ_u^2) can be estimated the same as the way in Step 4. That is,

$$\left(\hat{\alpha}^0, \hat{\sigma}_u^2\right) = \arg\max_{(s,\delta_u^2)} \sum_{k=1}^{\hat{K}} \sum_{i \in \hat{G}_{k|\hat{K}}} \log f\left(y_i \mid x_i; s, \delta_u^2, \hat{\vartheta}_{(k|\hat{K})}\right),$$

where $f\left(y_i \mid x_i; s, \delta_u^2, \hat{\vartheta}_{(k|\hat{K})}\right)$ is defined in (C.2), and we use the post-classification estimates $\hat{\vartheta}_{(k|\hat{K})}$ from Step 3$^{\blacktriangledown}$.

For Spec2, $\left(\alpha_{(1)}^0, \alpha_{(2)}^0, \sigma_{u(2)}^2, \tau^0\right)$ can be obtained as the restricted estimation in Step 4' (by forcing $\sigma_{u(1)}^2 = 0$) when $\mathcal{K} = 2$. That is

$$\left(\hat{\alpha}_{(1)}^0, \hat{\alpha}_{(2)}^0, \hat{\sigma}_{u(2)}^2, \hat{\tau}\right) = \arg\max_{(s,\delta_u^2,\tau)} \sum_{k=1}^{\hat{K}} \sum_{i \in \hat{G}_{k|\hat{K}}} \log \tilde{f}\left(y_i \left| x_i; s_1, 0, s_2, \delta_{u(2)}^2, \tau, \hat{\vartheta}_{(k|\hat{K})}\right.\right),$$

where $\tilde{f}$ is the likelihood function, defined in (C.3).

For Spec3, $\left(\alpha_{(1)}^0, \sigma_{u(1)}^2, \alpha_{(2)}^0, \sigma_{u(2)}^2, \tau^0\right)$ is the estimation in Step 4' when $\mathcal{K} = 2$. Thus,

$$\left(\hat{\alpha}_{(1)}^0, \hat{\sigma}_{u(1)}^2, \hat{\alpha}_{(2)}^0, \hat{\sigma}_{u(2)}^2, \hat{\tau}\right) = \arg\max_{(s,\delta_u^2,\tau)} \sum_{k=1}^{\hat{K}} \sum_{i \in \hat{G}_{k|\hat{K}}} \log \tilde{f}\left(y_i \left| x_i; s_1, \delta_{u(1)}^2, s_2, \delta_{u(2)}^2, \tau, \hat{\vartheta}_{(k|\hat{K})}\right.\right).$$

Collecting results from Step 4$^{\blacktriangledown}$, we proceed to Step 5$^{\blacktriangledown}$ to determine the specification of $\alpha_i^0 - u_i$.

Step 5$^{\blacktriangledown}$: Determination of the Distributional Structures of the Inefficiency Term

We calculate three IC values to determine the distribution of $\alpha_i^0 - u_i$.

We consider Spec1 and Spec3 first, because they are special cases of Step 4' when $\mathcal{K} = 1$ and $\mathcal{K} = 2$, respectively. Using this insight, we define the information criteria as follows:

For Spec1:

$$\widetilde{\text{IC}}_1\left(\tilde{\lambda}_{NT}\right) = -\sum_{k=1}^{\hat{K}} \sum_{i \in \hat{G}_{k|\hat{K}}} \log f\left(y_i \left| x_i; \hat{\alpha}^0, \hat{\sigma}_u^2 \hat{\vartheta}_{(k|\hat{K})}\right.\right) + \tilde{\lambda}_{NT}.$$

For Spec3:

$$\widetilde{\text{IC}}_3\left(\tilde{\lambda}_{NT}\right) = -\sum_{k=1}^{\hat{K}} \sum_{i \in \hat{G}_{k|\hat{K}}} \log \tilde{f}\left(y_i \left| x_i; \hat{\alpha}_{(1)}^0, \hat{\sigma}_{u(1)}^2, \hat{\alpha}_{(2)}^0, \hat{\sigma}_{u(2)}^2, \hat{\tau}, \hat{\vartheta}_{(k|\hat{K})}\right.\right) + 2\tilde{\lambda}_{NT}.$$

Both $\mathrm{IC}_1\left(\tilde{\lambda}_{NT}\right)$ and $\mathrm{IC}_3\left(\tilde{\lambda}_{NT}\right)$ are special cases of $\widetilde{\mathrm{IC}}(\mathcal{K}, \tilde{\lambda}_{NT})$ (defined in (2.8)) when $\mathcal{K} = 1$ and 2, respectively.

Note that Spec1, Spec2 and Spec3 contain 2, 4, and 5 parameters, respectively. Thus, the number of parameters increases by 3 when moving from Spec1 to Spec3, and by 2 when moving from Spec1 to Spec2. Based on this observation, we define:

$$\widetilde{\mathrm{IC}}_2\left(\tilde{\lambda}_{NT}\right) = -\sum_{k=1}^{\hat{K}} \sum_{i \in \hat{G}_{k|\hat{K}}} \log \tilde{f}\left(y_i \left| x_i; \hat{\alpha}^0_{(1)}, 0, \hat{\alpha}^0_{(2)}, \hat{\sigma}^2_{u(2)}, \hat{\tau}, \hat{\vartheta}_{(k|\hat{K})}\right.\right) + \frac{5}{3}\tilde{\lambda}_{NT}.$$

We then select the model that minimizes the information criterion:

$$\hat{l} = \arg\min_{l=1,2,3} \widetilde{\mathrm{IC}}_l\left(\tilde{\lambda}_{NT}\right),$$

and adopt Spec$\hat{l}$ accordingly.

A further investigation of this procedure is left for future.

B Allowing More Than Two Components in the Mixture

As discussed in the main body of the paper, estimating mixtures with more than two components, as in Step 4′, is numerically challenging. The difficulty arises because the original log-likelihood function is highly nonlinear and cannot be effectively optimized when the number of mixture components exceeds two.

The motivation for the alternative method stems from the observation that the original log-likelihood function is overly complicated, and we seek a numerically less demanding approach.

The procedure is as follows. For each $i = 1, 2, \ldots, N$, we first compute

$$\begin{aligned}
\widehat{\alpha_i^0 - u_i} &= \frac{1}{T}\sum_{t=1}^{T}\left(y_{it} - \underline{z}_{it}'\hat{\pi}_{(k|K)}\right), \quad \text{for } i \in \hat{G}_{k|K} \\
&= \alpha_i^0 - u_i + \frac{1}{T}\sum_{t=1}^{T} v_{it} + O_P\Big(\sqrt{\tfrac{m}{N_k T}}\Big) \\
&= \alpha_i^0 - u_i + O_P\Big(\tfrac{1}{\sqrt{T}} + \sqrt{\tfrac{m}{N_k T}}\Big) \\
&\approx \alpha_i^0 - u_i. \qquad \text{(B.1)}
\end{aligned}$$

Recall that for $\alpha_i^0 - u_i$ there exists an integer $\mathcal{K}^* \geq 1$ such that, with probability τ_j^0, it follows the distribution

$$\alpha_{(j)}^0 - |N(0, \sigma_{u(j)}^2)|, \quad j = 1, 2, \ldots, \mathcal{K}^*,$$

where $\left(\alpha_{(j)}^0, \sigma_{u(j)}^2\right)$, $j = 1, 2, \ldots, \mathcal{K}^*$, are distinct parameter pairs, with $0 < \tau_j^0 < 1$ for $j = 1, 2, \ldots, \mathcal{K}^* - 1$, and $1 - \tau_1^0 - \cdots - \tau_{\mathcal{K}^*-1}^0 > 0$.

The likelihood function of $\alpha_i^0 - u_i$ then takes the form

$$\begin{aligned}
&f_{\text{mix}}\Big(s \,\Big|\, \alpha_{(1)}^0, \sigma_{u(1)}^2, \ldots, \alpha_{(\mathcal{K}^*)}^0, \sigma_{u(\mathcal{K}^*)}^2, \tau_1^0, \ldots, \tau_{\mathcal{K}^*-1}^0\Big) \\
&\quad = \sum_{j=1}^{\mathcal{K}^*} \tau_j^0 \, \frac{2}{\sqrt{2\pi}} \exp\left\{-\frac{\left(s - \alpha_{(j)}^0\right)^2}{2\sigma_{u(j)}^2}\right\} \mathbf{1}\left(s \leq \alpha_{(j)}^0\right),
\end{aligned}$$

with $\tau_{\mathcal{K}^*}^0 = 1 - \tau_1^0 - \cdots - \tau_{\mathcal{K}^*-1}^0$. This likelihood is considerably simpler and more tractable than the one introduced in Step 4′.

We propose to estimate the mixture structure of $\alpha_i^0 - u_i$ by using $\widehat{\alpha_i^0 - u_i}$ as a surrogate and maximizing the likelihood function defined above. The information criterion can then be constructed analogously to before:

$$\text{IC}_{\text{mix}}(\mathcal{K}, \tilde{\lambda}_{NT}) = -\sum_{i=1}^{N} \log f_{\text{mix}}\left(\widehat{\alpha_i^0 - u_i} \,\middle|\, \hat{\alpha}_{(1)}^0, \hat{\sigma}_{u(1)}^2, \ldots, \hat{\alpha}_{(\mathcal{K})}^0, \hat{\sigma}_{u(\mathcal{K})}^2, \hat{\tau}_1^0, \ldots, \hat{\tau}_{\mathcal{K}-1}^0\right) + \mathcal{K}\tilde{\lambda}_{NT},$$

for a given number of mixture components $\mathcal{K}$.

Finally, we select the number of components by

$$\hat{\mathcal{K}}(\tilde{\lambda}_{NT}) = \arg \min_{\mathcal{K}=0,1,\ldots,\bar{\mathcal{K}}} \mathrm{IC}_{\mathrm{mix}}(\mathcal{K}, \tilde{\lambda}_{NT}),$$

and write $\hat{\mathcal{K}}$ for short. The corresponding parameter estimates are

$$\left(\hat{\alpha}^0_{(1)}, \hat{\sigma}^2_{u(1)}, \ldots, \hat{\alpha}^0_{(\hat{\mathcal{K}})}, \hat{\sigma}^2_{u(\hat{\mathcal{K}})}, \hat{\tau}_1, \ldots, \hat{\tau}_{\hat{\mathcal{K}}-1}\right).$$

Since $T \propto N$, it follows directly that $\tilde{\lambda}_{NT}$ in Proposition 3.3 remains valid in this setting. Therefore, in practice we adopt the recommended $\tilde{\lambda}_{NT}$ from Section 4.2.

We conduct simulations to evaluate the small sample properties of the above procedure. We study DGP 1M, 2M, and 3M, where the number of components in the mixture distribution is two. In addition, we consider DGP1T, DGP2T and DGP3T, where we allow the inefficiency terms to arise from the mixture of three distributions. DGPs 1T, 2T, and 3T share the same frontier and error distribution of v as DGPs 1M, 2M, and 3M, respectively. Specifically, for DGPs 1T, 2T and 3T, we let $\alpha^0 - u$ to come from $\alpha^0_{(1)} - |N(0, \sigma^2_{u(1)})|$ with probability τ_1, $\alpha^0_{(2)} - |N(0, \sigma^2_{u(2)})|$ with probability τ_2, and $\alpha^0_{(3)} - |N(0, \sigma^2_{u(3)})|$ with probability τ_3, where $\{\alpha^0_{(1)}, \alpha^0_{(2)}, \alpha^0_{(3)}\} = \{0.5, -1, 2\}$, $\{\sigma_{u(1)}, \sigma_{u(2)}, \sigma_{u(3)}\} = \{1, 1, 1\}$, and $\{\tau_1, \tau_2, \tau_3\} = \{0.3, 0.4, 0.3\}$.

Simulation results are reported in Appendix F.3. We find the following:

1. Using the previously recommended tuning parameters, the method performs reasonably well in identifying the correct number of components in the mixture for the original three DGPs (1M, 2M, and 3M) with two components, as well as three additional DGPs (1T, 2T, and 3T) with three components.

2. The parameter estimates converge at the rate of $\sqrt{T}$ when $N \geq T$. This is because we use $\widehat{\alpha^0_i - u_i}$ as observations, and $\widehat{\alpha^0_i - u_i}$ converges to $\alpha^0_i - u_i$ at the rate $\sqrt{T}$. In the simulations, we further set T smaller than N.

3. For fixed T, the estimates deteriorate as N increases for the following reasons:

(a) The theoretical convergence rate is $\sqrt{T}$, so increasing N does not yield improvements in theory.

(b) To illustrate the intuition, suppose the mixture has only one component with $\alpha_i^0 - u_i \overset{d}{\sim} \alpha^0 - |N(0, \sigma_u^2)|$. Continuing from (B.1), we obtain

$$\begin{aligned}\widehat{\alpha_i^0 - u_i} &= \alpha^0 - u_i + \frac{1}{T}\sum_{t=1}^{T} v_{it} + \left(\frac{1}{T}\sum_{t=1}^{T} \underset{\sim}{z}_{it}'\right)\left(\pi_{(k|K)}^0 - \hat{\pi}_{(k|K)}\right) + \frac{1}{T}\sum_{t=1}^{T} b_{i0}(\tau_t) \\ &\equiv \alpha^0 - u_i + \hat{\epsilon}_i,\end{aligned}$$

where $b_{i0}(\cdot)$ is the approximation bias term, and

$$\hat{\epsilon}_i \equiv \frac{1}{T}\sum_{t=1}^{T} v_{it} + \left(\frac{1}{T}\sum_{t=1}^{T} \underset{\sim}{z}_{it}'\right)\left(\pi_{(k|K)}^0 - \hat{\pi}_{(k|K)}\right) + \frac{1}{T}\sum_{t=1}^{T} b_{i0}(\tau_t).$$

If we treat $\widehat{\alpha_i^0 - u_i}$ as observations, then, analogous to the uniform distribution case, we have a closed-form solution for $\hat{\alpha}^0$:

$$\begin{aligned}\hat{\alpha}^0 &= \max_i \left\{\widehat{\alpha^0 - u_i}\right\} \\ &= \alpha^0 - \min_i \left\{u_i - \hat{\epsilon}_i\right\}.\end{aligned}$$

Since $\hat{\epsilon}_i$ can take negative values, the probability of observing extreme negative realizations of $\hat{\epsilon}_i$ increases with N, thereby reducing the accuracy of the estimate.

For these reasons, we recommend that practitioners use this approach only when the primary goal is to identify the number of mixture components or to model three or more components, and only when T is sufficiently large to ensure reliable estimation.

C The Approximation of the Likelihood Function

C.1 The Inefficiency Term with a Unique Distribution

We derive the likelihood function incorporating approximations of α and β. Using the last line of (2.5), we have

$$\begin{aligned} y_{it} &\approx \alpha_i^0 - u_i + z_{it}'\pi_i^0 + v_{it} \\ &= \alpha_i^0 + z_{it}'\pi_i^0 + \varepsilon_{it}. \end{aligned} \tag{C.1}$$

The primary distinction between the above approximation and (2.5) is that we do not separate u_i from ε_{it}.

We first derive the likelihood function when the distribution of $\alpha_i^0 - u_i$ is unique. We adopt the notation used in Yao et al. (2019) for the following presentation. Let

$$\sigma_i^2 = \sigma_{vi}^2 + T\sigma_u^2,$$

$$\rho_i = \sigma_u/\sigma_{vi},$$

$$\mu_{i*} = -\frac{\sigma_u^2}{\sigma_i^2}\sum_{t=1}^{T}\varepsilon_{it}, \quad \text{and}$$

$$\sigma_{i*}^2 = \frac{\sigma_u^2\sigma_{vi}^2}{\sigma_i^2},$$

then

$$\sigma_{vi}^2 = \frac{\sigma_i^2}{1+T\rho_i^2}.$$

We denote the density and cumulative distribution functions of a standard normal distribution as $\phi(\cdot)$ and $\Phi(\cdot)$, respectively. Following calculations similar to those in Yao et al. (2019), the density of $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$ is given by

$$f\left(\varepsilon_i; \sigma_u^2, \sigma_{vi}^2\right) = \frac{2}{\sigma_{vi}^{T-1}\sigma_i}\left[1 - \Phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)\right]\left[\prod_{t=1}^{T}\phi\left(\frac{\varepsilon_{it}}{\sigma_{vi}}\right)\right]\exp\left(\frac{1}{2}\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)^2\right),$$

which implies

$$\log f\left(\varepsilon_i;\sigma_u^2,\sigma_{vi}^2\right)=C-\frac{(T-1)}{2}\log\sigma_{vi}^2-\frac{1}{2}\log\left(\sigma_{vi}^2+T\sigma_u^2\right)$$
$$+\log\left[1-\Phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)\right]+\frac{1}{2}\left(\frac{\mu_{i*}}{\sigma_{i*}}\right)^2-\frac{\sum_{t=1}^T\varepsilon_{it}^2}{2\sigma_{vi}^2},$$

for some constant C that does not depend on the parameters to be estimated. Recall that $\vartheta_i=(\pi_i^{0\prime},\sigma_{vi}^2)'$. Using the approximation in (C.1), the log-likelihood density function for $y_i=(y_{i1},\ldots,y_{iT})'$ is given by

$$\log f\left(y_i\mid x_i;\alpha_i^0,\sigma_u^2,\vartheta_i\right)\approx C-\frac{(T-1)}{2}\log\sigma_{vi}^2-\frac{1}{2}\log\left(\sigma_{vi}^2+T\sigma_u^2\right) \tag{C.2}$$
$$+\log\left[1-\Phi\left(-\frac{\tilde{\mu}_{*i}}{\sigma_{*i}}\right)\right]+\frac{1}{2}\left(\frac{\tilde{\mu}_{*i}}{\sigma_{*i}}\right)^2-\frac{\sum_{t=1}^T\tilde{\varepsilon}_{it}^2}{2\sigma_{vi}^2},$$

with

$$\tilde{\varepsilon}_{it}=y_{it}-\alpha_i^0-z_{it}'\pi_i^0,\text{ and}$$
$$\tilde{\mu}_{i*}=-\frac{\sigma_u^2}{\sigma_i^2}\sum_{t=1}^T\tilde{\varepsilon}_{it}.$$

C.2 The Inefficiency Term with a Mixture Distribution

We now consider the case when $\alpha_i^0-u_i$ is distributed as $\alpha_{(j)}^0-\left|N\left(0,\sigma_{u(j)}^2\right)\right|$ with probability τ_j^0, $j=1,2,...,\mathcal{K}^*$. This log-likelihood function is denoted as $\log\tilde{f}$ and can be derived as:

$$\log\tilde{f}\left(y_i\left|x_i;\alpha_{(1)}^0,\sigma_{u(1)}^2,...,\alpha_{(\mathcal{K}^*)}^0,\sigma_{u(\mathcal{K}^*)}^2,\tau_1^0,...,\tau_{\mathcal{K}^*-1}^0,\vartheta_i\right.\right) \tag{C.3}$$
$$=\log\left[\tau_1^0 f\left(y_i\left|x_i;\alpha_{(1)}^0,\sigma_{u(1)}^2,\vartheta_i\right.\right)+\tau_2^0 f\left(y_i\left|x_i;\alpha_{(2)}^0,\sigma_{u(2)}^2,\vartheta_i\right.\right)+...\right.$$
$$\left.+\left(1-\tau_1^0-...-\tau_{\mathcal{K}^*-1}^0\right)f\left(y_i\left|x_i;\alpha_{(\mathcal{K}^*)}^0,\sigma_{u(\mathcal{K}^*)}^2,\vartheta_i\right.\right)\right],$$

where $f\left(y_i\mid x_i;\alpha^0,\sigma_u^2,\vartheta_i\right)$ is defined in (C.2). We note that $\tau_{\mathcal{K}^*}^0=1-\tau_1^0-...-\tau_{\mathcal{K}^*-1}^0$, so the last term in the above is equivalent to $\tau_{\mathcal{K}^*}^0 f\left(\cdot|\cdot\right)$.

C.3 Computation of the Inefficiency Term Post Estimation

In the case when α_i does not vary across i, we are able to compute the expectation of the inefficiency term. We take the case in the empirical application as an example. Other cases with more than two components in the mixture can be studied similarly.

Using (5.2) and $T^{-1}\sum_{t=1}^{T} v_{it} \overset{d}{\sim} N\left(0, \hat{\sigma}^2_{v(k)}/T\right)$ for $i \in \hat{G}_{k|K}$, Greene (2005b) implies that

$$\begin{aligned}
&\mathrm{E}\left(\alpha_i^0 + \widehat{u_i|\varepsilon_{i1}, ..., \varepsilon_{iT}}\right) \\
=&\hat{\alpha}^* + \hat{\tau}\left[\mu^*_{i(1)} + \sigma^*_{(1)}\frac{\phi\left(\mu^*_{i(1)}/\sigma^*_{(1)}\right)}{\Phi\left(\mu^*_{i(1)}/\sigma^*_{(1)}\right)}\right] + (1-\hat{\tau})\left[\mu^*_{i(2)} + \sigma^*_{(2)}\frac{\phi\left(\mu^*_{i(2)}/\sigma^*_{(2)}\right)}{\Phi\left(\mu^*_{i(2)}/\sigma^*_{(2)}\right)}\right], \qquad \text{(C.4)}
\end{aligned}$$

where

$$\mu^*_{i(j)} = \rho^2_{(j)}\left[\sum_{t=1}^{T}\left(y_{it} - \underset{\sim}{z}'_{it}\hat{\pi}_{(k|K)}\right) - \hat{\alpha}^*\right], \sigma^{*2}_{(j)} = \rho^2_{(j)}\hat{\sigma}^2_{v(k)},$$

$$\rho^2_{(j)} = \lambda^2_{(j)}\Big/\left(1 + T\lambda^2_{(j)}\right), \text{ and } \lambda_{(j)} = \hat{\sigma}_{u(j)}\Big/\hat{\sigma}_{v(k)},$$

for $j = 1, 2$, and $i \in \hat{G}_{k|K}$.

D HAC Method

In this section of the appendix, we describe the Hierarchical Agglomerative Clustering (HAC) method used as part of the proposed method. We largely adopt the description from Chapter 4 of Everitt et al. (2011).

HAC is a bottom-up clustering approach that starts by treating each data point as an individual cluster. At each step, the two "closest" clusters are merged to form a new cluster, and this process is repeated iteratively until a stopping rule is satisfied. The stopping rule may be a pre-specified number of clusters K, a cut-off distance threshold, or the completion of the full hierarchy where all observations are eventually merged into a single cluster. The result can be represented by a *dendrogram*, which is a binary tree structure showing how clusters are combined at each stage.

A crucial component of HAC is the definition of the *linkage criterion*, which determines how distances between clusters are computed during the iterative merging process. Common linkage criteria include:

- **Single linkage:** distance between two clusters is defined as the minimum distance between any pair of points across clusters.
- **Complete linkage:** distance is the maximum distance between any pair of points across clusters.
- **Average linkage:** distance is the average pairwise distance between all points across clusters.
- **Ward's method:** distance is defined as the increase in within-cluster variance resulting from a merge, which tends to produce compact and spherical clusters.

In this paper, we focus on Ward's method (Ward, 1963), which is widely regarded as producing more balanced clusters compared to single or complete linkage. At each iteration, Ward's method merges the two clusters whose union leads to the smallest possible increase in the total within-cluster variance (often called the "error sum of squares"). This variance-based criterion makes the procedure especially well-suited for applications where compactness and homogeneity within clusters are desired.

Formally, the distance metric in Ward's method for clusters A and B is given by

$$d_{AB} = \frac{|A||B|}{|A| + |B|} \|\bar{x}_A - \bar{x}_B\|^2,$$

where $|A|$ and $|B|$ denote the cluster sizes, and $\bar{x}_A, \bar{x}_B$ are the corresponding centroids. This expression is derived from the increment in within-cluster sum of squares that would occur if clusters A and B were merged. The algorithm therefore prioritizes merging clusters with centroids that are close to each other, scaled by their sizes.

In practice, HAC with Ward's method proceeds as follows. First, each data point is initialized as its own cluster. Second, the pairwise distances between all clusters are computed using Ward's criterion. Third, the two clusters with the smallest distance are merged. Fourth, the cluster distances are updated to reflect the new merged cluster. These steps are repeated until the stopping rule is satisfied.

The output of HAC provides a nested sequence of partitions, which can be cut at different levels depending on the desired granularity. This flexibility is useful in empirical applications where the "true" number of groups is not known ex ante, and different levels of aggregation can be explored.

E Main Proofs

The main proofs in this section are built on technical lemmas in Appendix H. We first present some well known results that are useful for the proofs in this appendix.

We let b_{il}, $l = 0, 1, ..., p$, denote the bias term from approximations. Specifically,

$$b_{i0}(s) = \alpha_i(s) - \mathbb{B}^m_{-0}(s)' \pi^0_{i0} \text{ and } b_{il}(s) = \beta_{il}(s) - \mathbb{B}^m(s)' \pi^0_{il}, \tag{E.1}$$

for $s \in [0, 1]$, and ξ_{it} collects the bias term in y_{it}:

$$\xi_{it} \equiv b_{i0}(\tau_t) + \sum_{l=1}^{p} x_{itl} b_{il}(\tau_t). \tag{E.2}$$

With this notation and (2.5),

$$y_{it} = \tilde{z}'_{it} \tilde{\pi}^0_i + \xi_{it} + v_{it}. \tag{E.3}$$

We know from Chen (2007) that

$$\sup_{s \in [0,1]} |b_{il}(s)| = O\left(m^{-\kappa}\right)$$

holds by Assumption 4. Since p is finite, clearly,

$$\max_{l=0,...,p} \sup_{s \in [0,1]} |b_{il}(s)| = O\left(m^{-\kappa}\right), \tag{E.4}$$

for $i = 1, ..., N$. Using similar logic on $\alpha_k^*(\tau_t)$ and the fact $K^*$is fixed, we can obtain

$$\max_{k=1,...,K^*} \max_{l=0,...,p} \sup_{s\in[0,1]} |b_{kl}^*(s)| = O\left(m^{-\kappa}\right), \tag{E.5}$$

where

$$b_{k0}^*(s) = \alpha_k^*(s) - \mathbb{B}_{-0}^m(s)' \pi_{i0}^{*0} \text{ and } b_{kl}^*(s) = \beta_{il}(s) - \mathbb{B}^m(s)' \pi_{il}^{*0} \text{ for } l = 1, ..., p.$$

Proof of Theorem 3.1. **(i)** is a direct result of Lemma H.5. To see that,

$$\begin{aligned}
\Pr\left(\max_{i=1,2,...,N} \left\|\hat{\vartheta}_i - \vartheta_i\right\| > \epsilon\right) &\leq \Pr\left(\max_{i=1,2,...,N} \left\|\hat{\pi}_i - \pi_i^0\right\| > \frac{\epsilon}{2}\right) + \Pr\left(\max_{i=1,2,...,N} \left\|\hat{\sigma}_{vi}^2 - \sigma_{vi}^2\right\| > \frac{\epsilon}{2}\right) \\
&\leq \Pr\left(\max_{i=1,2,...,N} \left\|\hat{\tilde{\pi}}_i - \tilde{\pi}_i^0\right\| > \frac{\epsilon}{2}\right) + \Pr\left(\max_{i=1,2,...,N} \left\|\hat{\sigma}_{vi}^2 - \sigma_{vi}^2\right\| > \frac{\epsilon}{2}\right) \\
&= o(1),
\end{aligned}$$

where the second line holds by the fact that $\hat{\pi}_i$ is a sub-vector of $\hat{\tilde{\pi}}_i$, and the last line applies the results in Lemma H.5.

(ii) Denote

$$\begin{aligned}
L_{ii'} &\equiv \sum_{l=0}^{p} \left\|\pi_{il}^0 - \pi_{i'l}^0\right\| + \left|\sigma_{vi} - \sigma_{vi'}\right|, \text{ and} \\
\hat{L}_{ii'} &\equiv \sum_{l=0}^{p} \left\|\hat{\pi}_{il} - \hat{\pi}_{i'l}\right\| + \left|\hat{\sigma}_{vi} - \hat{\sigma}_{vi'}\right|,
\end{aligned}$$

and

$$L_{jk}^* \equiv \sum_{l=0}^{p} \left\|\pi_{jl}^{*0} - \pi_{kl}^{*0}\right\| + \left|\sigma_{v(j)}^* - \sigma_{v(k)}^*\right|.$$

for $i, i' = 1, ..., n$ and $j, k = 1, ..., K^*$. We first claim that

$$\min_{1\leq j\neq k\leq K^*} L_{jk}^* \geq \frac{1}{2}\underline{C}^* \tag{E.6}$$

holds after some large T (and hence large m). We will show this claim at the end.

To show the result in (ii), it is equivalent to show that

$$\Pr\left(\max_{1\leq k\leq K^*} \max_{i,i'\in G_{k|K^*}} \hat{L}_{ii'} < \min_{1\leq j\neq k\leq K^*} \min_{i\in G_{j|K^*}, i'\in G_{k|K^*}} \hat{L}_{ii'}\right) = 1 - o(1).$$

When $i, i' \in G_{k|K^*}$, $L_{ii'} = 0$. Thus, the uniform convergence in **(i)** implies that, for $\epsilon = \underline{C}^*/6$,

$$\Pr\left(\max_{1\leq k\leq K^*}\max_{i,i'\in G_{k|K^*}} \hat{L}_{ii'} < \underline{C}^*/6\right) = 1 - o(1). \tag{E.7}$$

Denote this event as

$$A = \left\{\max_{1\leq k\leq K^*}\max_{i,i'\in G_{k|K^*}} \hat{L}_{ii'} < \underline{C}^*/6\right\}.$$

Conditional on this event A, the claim in (E.6) after some large T (and hence large m), the result in (E.7), and the triangular inequality imply that

$$\begin{aligned}
&\min_{1\leq j\neq k\leq K^*}\min_{i\in G_{j|K^*},i'\in G_{k|K^*}} \hat{L}_{ii'} \\
\geq &\min_{1\leq j\neq k\leq K^*} L^*_{jk} - 2\max_{i=1,\ldots,n}\left(\sum_{l=0}^{p}\left\|\hat{\pi}_{il} - \pi^0_{il}\right\| + |\hat{\sigma}_{vi} - \sigma_{vi}|\right) \\
\geq &\frac{\underline{C}^*}{2} - 2\cdot\frac{\underline{C}^*}{6} = \underline{C}^*/6 \\
> &\max_{1\leq k\leq K^*}\max_{i,i'\in G_{k|K^*}} \hat{L}_{ii'}.
\end{aligned}$$

Therefore, after some large T (and hence large m),

$$\begin{aligned}
&\Pr\left(\max_{1\leq k\leq K^*}\max_{i,i'\in G_{k|K^*}} \hat{L}_{ii'} < \min_{1\leq j\neq k\leq K^*}\min_{i\in G_{j|K^*},i'\in G_{k|K^*}} \hat{L}_{ii'}\right) \\
\geq &\Pr\left(\max_{1\leq k\leq K^*}\max_{i,i'\in G_{k|K^*}} \hat{L}_{ii'} < \min_{1\leq j\neq k\leq K^*}\min_{i\in G_{j|K^*},i'\in G_{k|K^*}} \hat{L}_{ii'}\middle| A\right)\Pr(A) \\
= &\Pr(A) = 1 - o(1),
\end{aligned}$$

as desired.

We now show the claim in (E.6). Notice that for any $\pi_1, \pi_2 \in \mathbb{R}^m$,

$$\begin{aligned}
\left\|\mathbb{B}^m(s)'\pi_1 - \mathbb{B}^m(s)'\pi_2\right\| &= \left[\int_0^1\left(\sum_{j=0}^{m-1} B_j(s)\pi_{1j} - \sum_{j=0}^{m-1} B_j(s)\pi_{2j}\right)^2 ds\right]^{1/2} \\
&= \left[\int_0^1\sum_{j=0}^{m-1} B_j(s)^2(\pi_{1j} - \pi_{2j})^2 ds\right]^{1/2} = \left[\sum_{j=0}^{m-1}(\pi_{1j} - \pi_{2j})^2\right]^{1/2} \\
&= \|\pi_1 - \pi_2\|,
\end{aligned} \tag{E.8}$$

where the second line holds by $\int_0^1 B_j(s) B_{j'}(s) ds = 0$ for $j \neq j'$,and the third line holds by $\int_0^1 B_j(s)^2 ds = 1$.

Using (E.8),

$$\begin{aligned}
L_{jk}^* &= \left\|\pi_{j0}^{*0} - \pi_{k0}^{*0}\right\| + \sum_{l=1}^{p} \left\|\pi_{jl}^{*0} - \pi_{kl}^{*0}\right\| + \left|\sigma_{v(j)}^* - \sigma_{v(k)}^*\right| \\
&= \left\|\mathbb{B}_{-0}^m(s)' \pi_{j0}^{*0} - \mathbb{B}_{-0}^m(s)' \pi_{k0}^{*0}\right\| + \sum_{l=1}^{p} \left\|\mathbb{B}^m(s)' \pi_{jl}^{*0} - \mathbb{B}^m(s)' \pi_{kl}^{*0}\right\| + \left|\sigma_{v(j)}^* - \sigma_{v(k)}^*\right| \\
&= \left\|\alpha_j^* - b_{j0}^*(s) - \alpha_k^* + b_{k0}^*(s)\right\| + \sum_{l=1}^{p} \left\|\beta_{jl}^*(s) - b_{jl}^*(s) - \beta_{kl}^*(s) + b_{jk}^*(s)\right\| + \left|\sigma_{v(j)}^* - \sigma_{v(k)}^*\right| \\
&\geq \left\|\alpha_j^* - \alpha_k^*\right\| - \left\|b_{j0}^*(s) - b_{k0}^*(s)\right\| + \sum_{l=1}^{p} \left[\left\|\beta_{jl}^* - \beta_{kl}^*\right\| - \left\|b_{jl}^*(s) - b_{jk}^*(s)\right\|\right] + \left|\sigma_{v(j)}^* - \sigma_{v(k)}^*\right| \\
&\geq \underline{C}^* - O\left(m^{-\kappa}\right),
\end{aligned}$$

where the fourth line holds by triangular inequality, and the last line holds by Assumption 5, the result in (E.5), and the fact that p is fixed. Using the result in (E.5) and the fact that K^* is fixed, then after some large T (and hence large m),

$$\min_{1 \leq j \neq k \leq K^*} L_{jk}^* = \min_{1 \leq j \neq k \leq K^*} \sum_{l=0}^{p} \left\|\pi_{jl}^{*0} - \pi_{kl}^{*0}\right\| + \left|\sigma_{v(j)}^* - \sigma_{v(k)}^*\right| \geq \frac{1}{2}\underline{C}^*,$$

as desired. □

Proof of Theorem 3.2. **(i)** Denote the event of correct classification as

$$\mathcal{M} = \left\{\left(\hat{G}_{1|K^*}, \hat{G}_{2|K^*}, \ldots, \hat{G}_{K^*|K^*}\right) = \left(G_{1|K^*}, G_{2|K^*}, \ldots, G_{K^*|K^*}\right)\right\}.$$

We first show the results conditional on the event $\mathcal{M}$.

For each $i \in G_{k|K^*}$

$$y_{it} = \alpha^0 - u_i + \underline{z}_{it}' \pi_{(k)}^{*0} + \xi_{it} + v_{it},$$

where similar to how π_i^0 is defined, $\pi_{(k)}^{*0}$ collects coefficients for the approximation of $\alpha_{(k)}^*(s)$ and $\beta_{(k)}^*(s)$. Thus

$$\ddot{y}_{it} = \ddot{\underline{z}}_{it}' \pi_{(k)}^{*0} + \ddot{\xi}_{it} + \ddot{v}_{it}. \tag{E.9}$$

Using (E.9),

$$\begin{aligned}\hat{\pi}^{*0}_{(k|K^*)} - \pi^{*0}_{(k)} &= \left(\sum_{i\in G_{k|K^*}}\sum_{t=1}^{T}\ddot{z}_{it}\ddot{z}'_{it}\right)^{-1}\left(\sum_{i\in G_{k|K^*}}\sum_{t=1}^{T}\ddot{z}_{it}\ddot{\xi}_{it}\right) \\ &+ \left(\sum_{i\in G_{k|K^*}}\sum_{t=1}^{T}\ddot{z}_{it}\ddot{z}'_{it}\right)^{-1}\left(\sum_{i\in G_{k|K^*}}\sum_{t=1}^{T}\ddot{z}_{it}\ddot{v}_{it}\right) \\ &\equiv A_{k1} + A_{k2}, \end{aligned} \tag{E.10}$$

where $\ddot{\xi}_{it}$ is the de-meaned ξ_{it} (defined in (E.2)) over t for observation i, and $\ddot{v}_{it}$ is similarly defined. With the decomposition in (E.10),

$$\begin{aligned}&\sqrt{N_kT/\underline{m}}\mathbb{S}^{-1/2}_{(k)}\left[\hat{\theta}_{(k|K^*)}(s) - \theta^*_{(k)}(s)\right] \\ &= \sqrt{N_kT/\underline{m}}\mathbb{S}^{-1/2}_{(k)}\mathbb{M}_{\mathbb{B}}(s)\left[\hat{\pi}^{*0}_{(k|K^*)} - \pi^{*0}_{(k)}\right] \\ &= \sqrt{N_kT/\underline{m}}\mathbb{S}^{-1/2}_{(k)}\mathbb{M}_{\mathbb{B}}(s)A_{k1} + \sqrt{N_kT/\underline{m}}\mathbb{S}^{-1/2}_{(k)}\mathbb{M}_{\mathbb{B}}(s)A_{k2} \\ &= o_P(1) + \sqrt{N_kT/\underline{m}}\mathbb{S}^{-1/2}_{(k)}\mathbb{M}_{\mathbb{B}}(s)A_{k2} \\ &\xrightarrow{d} N(0, I_{p+1}), \end{aligned} \tag{E.11}$$

where the fourth line uses the result in Lemma H.8 and $N_kT/\underline{m}^{1+2\kappa} \to 0$ imposed in Assumption 10, and the last line holds by evoking the Crámer-Wold device on the result in Lemma H.9.

To facilitate exposition, let $\omega_{NT} \equiv \sqrt{N_kT/\underline{m}}\left(\mathbb{S}^{-1/2}_{(k)}\right)'\left[\hat{\theta}_{(k|K^*)}(s) - \theta^*_{(k)}(s)\right]$. According to the definition of convergence in distribution, (E.11) implies that for any $\epsilon > 0$ and any $x \in R$, there exists a large N_1 such that for all $N > N_1$

$$\left|P(\omega_{NT} \leq x|\mathcal{M}) - \Phi(x)\right| < \frac{\epsilon}{3},$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal.

We now show the general results without conditional on $\mathcal{M}$. Theorem 3.1 implies that there exists a large N_2 such that for all $N > N_2$

$$P(\mathcal{M}) > 1 - \frac{\epsilon}{3}.$$

Thus, for all $N > \max(N_1, N_2)$ and any $x \in R$,

$$\begin{aligned}
|P(\omega_{NT} \leq x) - \Phi(x)| &= |P(\omega_{NT} \leq x | \mathcal{M}) P(\mathcal{M}) + P(\omega_{NT} \leq x | \mathcal{M}^c) P(\mathcal{M}^c) - \Phi(x)| \\
&\leq |[P(\omega_{NT} \leq x | \mathcal{M}) - \Phi(x)] P(\mathcal{M})| + [1 - P(\mathcal{M})] \Phi(x) \\
&+ P(\omega_{NT} \leq x | \mathcal{M}^c) P(\mathcal{M}^c) \\
&\leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon,
\end{aligned}$$

which is the desired result by the definition of convergence in distribution.

(ii) We first show the results conditional on the event $\mathcal{M}$. Using the representation in (E.9),

$$\begin{aligned}
&\hat{\sigma}^2_{v(k|K^*)} - \sigma^{*2}_{v(k)} \\
&= \frac{1}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \left(\ddot{y}_{it} - \ddot{\underline{z}}'_{it} \hat{\pi}_{(k)} \right)^2 - \sigma^{*2}_{v(k)} \\
&= \frac{1}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \left[\ddot{\underline{z}}'_{it} \left(\pi^{*0}_{(k)} - \hat{\pi}_{(k)} \right) + \ddot{\xi}_{it} + \ddot{v}_{it} \right]^2 - \sigma^{*2}_{v(k)} \\
&= \frac{1}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \ddot{v}^2_{it} - \sigma^{*2}_{v(k)} + \frac{1}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \left[\ddot{\underline{z}}'_{it} \left(\pi^{*0}_{(k)} - \hat{\pi}_{(k)} \right) \right]^2 \\
&+ \frac{1}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \ddot{\xi}^2_{it} + \frac{2}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \ddot{\underline{z}}'_{it} \left(\pi^{*0}_{(k)} - \hat{\pi}_{(k)} \right) \left(\ddot{\xi}_{it} + \ddot{v}_{it} \right) \\
&+ \frac{2}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \ddot{\xi}_{it} \ddot{v}_{it} \\
&\equiv A_{k1} + A_{k2} + A_{k3} + A_{k4} + A_{k5}.
\end{aligned}$$

We show that A_{k2}, A_{k3}, A_{k4}, and A_{k5} are asymptotically negligible by demonstrating the rates for A_{k2}, A_{k3}, A_{k4}, and A_{k5}. We then move to the asymptotic distribution of A_{k1}.

For A_{k2},

$$\begin{aligned} A_{k2} &= \frac{1}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \left[\ddot{\underline{z}}_{it}' \left(\pi_{(k)}^{*0} - \hat{\pi}_{(k)} \right) \right]^2 \\ &= \left(\pi_{(k)}^{*0} - \hat{\pi}_{(k)} \right)' \left(\frac{1}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \ddot{\underline{z}}_{it} \ddot{\underline{z}}_{it}' \right) \left(\pi_{(k)}^{*0} - \hat{\pi}_{(k)} \right) \\ &= O_P \left(\left\| \pi_{(k)}^{*0} - \hat{\pi}_{(k)} \right\|^2 \right) = O_P \left(\frac{\underline{m}}{N_k T} \right), \end{aligned} \tag{E.12}$$

where the last line holds by full rank condition implied by Lemmas H.6 and H.7, and the rate we show in **(i)**.

For A_{k3}, by the definition of ξ_{it}in (E.2), the rate in (E.5), and Assumption 10 (ii), we can obtain

$$A_{k3} = \frac{1}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \ddot{\xi}_{it}^2 = O_P \left(\underline{m}^{-2\kappa} \right) = o_P \left(\frac{\underline{m}}{N_k T} \right). \tag{E.13}$$

For A_{k4},

$$\begin{aligned} A_{k4} &= \frac{2}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \ddot{\underline{z}}_{it}' \left(\pi_{(k)}^{*0} - \hat{\pi}_{(k)} \right) \left(\ddot{\xi}_{it} + \ddot{v}_{it} \right) \\ &= \left(\pi_{(k)}^{*0} - \hat{\pi}_{(k)} \right)' \left[\frac{2}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \ddot{\underline{z}}_{it} \ddot{\xi}_{it} \right] + \left(\pi_{(k)}^{*0} - \hat{\pi}_{(k)} \right)' \frac{2}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \ddot{\underline{z}}_{it} \ddot{v}_{it} \\ &= O_P \left(\sqrt{\frac{\underline{m}}{N_k T}} \underline{m}^{-\kappa} \right) + O_P \left(\sqrt{\frac{\underline{m}}{N_k T}} \cdot \sqrt{\frac{1}{N_k T}} \right) = o_P \left(\frac{\underline{m}}{N_k T} \right), \end{aligned} \tag{E.14}$$

where for the third line we apply the result in Lemma H.8, the mixing condition across t in Assumption 1, and the independence across i in Assumption 7

For A_{k5}, again by the mixing condition across t, independence across i, and the rate in (E.5), we have

$$\text{Var} \left(\frac{2}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \ddot{\xi}_{it} \ddot{v}_{it} \right) \propto \frac{\underline{m}^{-2\kappa}}{NT}.$$

Using the Markov inequality, the above implies that

$$A_{k5} = O_P \left(\frac{\underline{m}^{-\kappa}}{\sqrt{NT}} \right) = o_P \left(\frac{\underline{m}}{N_k T} \right). \tag{E.15}$$

We turn to the leading term A_{k1} :

$$\begin{aligned}
A_{k1} &= \frac{1}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \ddot{v}_{it}^2 - \sigma_{v(k)}^2 \\
&= \frac{1}{N_k (T-1)} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \left(\ddot{v}_{it}^2 - \sigma_{v(k)}^2 \right) + \frac{1}{T-1} \sigma_{v(k)}^2 \\
&= \frac{T}{T-1} \left[\frac{1}{N_k T} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \left(\ddot{v}_{it}^2 - \sigma_{v(k)}^2 \right) \right] + \frac{1}{T-1} \sigma_{v(k)}^2 \\
&= \frac{T}{T-1} \left[\frac{1}{N_k T} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \left(v_{it}^2 - \sigma_{v(k)}^2 \right) \right] - \frac{T}{T-1} \frac{1}{N_k} \sum_{i \in G_{k|K^*}} \bar{v}_i^2 + \frac{1}{T-1} \sigma_{v(k)}^2 \\
&= \frac{T}{T-1} \left[\frac{1}{N_k T} \sum_{i \in G_{k|K^*}} \sum_{t=1}^{T} \left(v_{it}^2 - \sigma_{v(k)}^2 \right) \right] - \frac{1}{T-1} \frac{1}{N_k} \sum_{i \in G_{k|K^*}} \left(T \bar{v}_i^2 - \sigma_{v(k)}^2 \right) \\
&\equiv A_{k11} + A_{k12}.
\end{aligned}$$

By the i.i.d. assumption on v_{it},

$$\sqrt{N_k T} \cdot A_{k11} \xrightarrow{d} N \left(0, \text{Var} \left(v_{it}^2 \middle| i \in G_{k|K^*} \right) \right).$$

Note $\text{E}\left(T \bar{v}_i^2\right) = \sigma_{v(k)}^2$ and $\text{E}\left(T^2 \bar{v}_i^4\right) = O\left(1\right)$. By the independence across i and Markov's inequality,

$$A_{k12} = O_P \left(\frac{1}{T \sqrt{N_k}} \right),$$

which implies that

$$\sqrt{N_k T} \cdot A_{k12} = o_P \left(1\right).$$

Put the asymptotic distribution of A_{k11} and the rate of A_{k12} together, we obtain

$$\sqrt{N_k T} \cdot A_{k1} \xrightarrow{d} N \left(0, \text{Var} \left(v_{it}^2 \middle| i \in G_{k|K^*} \right) \right). \tag{E.16}$$

Equations (E.12), (E.13), (E.14), (E.15), and (E.16) imply that conditional on $\mathcal{M}$,

$$\sqrt{N_k T} \left(\hat{\sigma}_{v(k|K^*)}^2 - \sigma_{v(k)}^{*2} \right) \xrightarrow{d} N \left(0, \text{Var} \left(v_{it}^2 \middle| i \in G_{k|K^*} \right) \right).$$

The above result holds unconditionally, using a similar argument as in the proof of **(i)**.

(iii) We only need to show the result conditional on the event $\mathcal{M}$. After that, we can show the unconditional result using the same logic as in the proofs of **(i)** and **(ii)**. In the following, we assume that the event $\mathcal{M}$ happens.

In Appendix G.2, we show that the information matrix $\mathbb{I}$ and

$$\mathrm{E}\left[\sum_{k=1}^{K^*}\frac{N_k}{N}\cdot\left(\frac{\partial}{\partial\varrho}\log\tilde{f}_{i(k)}\left(\varrho\right)\right)\left(\frac{\partial}{\partial\varrho}\log\tilde{f}_{i(k)}\left(\varrho\right)\right)'\Big|_{\varrho=\varrho^0}\right]$$

are well behaved; that is, all elements in these two matrices are finite constants, not degenerated to 0, and they are positive definite.

Further, in **(i)** and **(ii)**, we have shown that $\hat{\theta}_{(k|K^*)}\left(s\right)$ and $\hat{\sigma}^2_{v(k|K^*)}$ converge to the trues at the rates of $\sqrt{\frac{NT}{m}}$ and $\sqrt{NT}$, respectively. Thus, the estimation errors of $\hat{\theta}_{(k|K^*)}\left(s\right)$ and $\hat{\sigma}^2_{v\left(k|\hat{K}\right)}$ have no impact on the convergence rate of $\hat{\varrho}$ and its asymptotic distribution, because $\hat{\varrho}$ supposedly converges to ϱ^0 at the rate of $\sqrt{N}$, which is slower than those of $\hat{\theta}_{(k|K^*)}\left(s\right)$ and $\hat{\sigma}^2_{v(k|K^*)}$.

By the independence across i, some standard analysis for the asymptotics of the MLE, see, e.g., Section 5.4.3 in Bickel and Doksum (2015), and the Slutsky's Theorem, the asymptotic variance of $\hat{\varrho}$ is $N^{-1}\mathbb{I}$, and, by the Lindeberg Central Limit Theorem,

$$\sqrt{N}\mathbb{I}^{-1/2}\left(\hat{\varrho}-\varrho\right)\xrightarrow{d}N\left(0,I_{3\mathcal{K}^*-1}\right).$$

□

Proof of Proposition 3.3. (i) This is a result direct from Lemmas H.10 and H.11.

(ii) This is a result from Lemma H.12. □

F DGP 1 and 2, Figures, Tables and Additional Discussions

F.1 DGPs 1 and 2

Design 1: In this design (consisting of DGP1U and DGP1M), we study the classification arising as a result of heterogeneity from frontiers. To this end, we specify the DGP as

$$y_{it} = \alpha_i^0 - u_i + \alpha_i(\tau_t) + x_{it}\beta_i(\tau_t) + v_{it},$$

and suppose there are two groups for frontiers with $x_{it} \sim N(1, 1^2)$. The error term v_{it} is generated from the same distribution $v_{it} \overset{iid}{\sim} N(0, \sigma_v^2)$, $\sigma_v = 1$ for both groups. Group 1 frontiers are specified as $\alpha_{(1)}(s) = 3F(s; 0.5, 0.1) - \varpi_1$, and $\beta_{(1)}(s) = 3[2s - 4s^2 + 2s^3 + F(s; 0.6, 0.1)$, where $F(s; \cdot, \cdot)$ denotes the logistic CDF and ϖ_1 is the mean of $3F(s; 0.5, 0.1)$, that is, $\int_0^1 3F(s; 0.5, 0.1)\mathrm{d}s$. Group 2 parameters are specified as $\alpha_{(2)}(s) = 3[2s - 6s^2 + 4s^3 + F(s; 0.7, 0.05)] - \varpi_2$, and $\beta_{(2)}(s) = 3[s - 3s^2 + 2s^3 + F(s; 0.7, 0.04)]$, ϖ_2 plays the same role as ϖ_1, and $\varpi_2 = \int_0^1 3[2s - 6s^2 + 4s^3 + F(s; 0.7, 0.05)]\mathrm{d}s$. In what we call this DGP1U, we consider a case where $\alpha^0 - u$ comes from a unique distribution, with $\alpha^0 = 0.5$ and $u_i \overset{iid}{\sim} |N(0, \sigma_u^2)|$, where $\sigma_u = 1$. We consider another DGP, which we call DGP1M, distinguished by letting $\alpha^0 - u$ to come from a mixture distribution. Specifically, we let $\alpha^0 - u$ to come from $\alpha_{(1)}^0 - |N(0, \sigma_{u(1)}^2)|$ with probability τ and $\alpha_{(2)}^0 - |N(0, \sigma_{u(2)}^2)|$ with probability $1 - \tau$, where $\alpha_{(1)}^0 = 1$, $\alpha_{(2)}^0 = -1$, $\sigma_{u(1)} = 0.75$, $\sigma_{u(2)} = 1.25$ and $\tau = 0.5$. It is important to note that the mixture structure of $\alpha^0 - u$ is independent of the grouping structure.

Design 2: In the second design (consisting of DGP2U and DGP2M), we study the case where there are two groups, but the classification is due to heterogeneity in variances. To this end, we adopt a similar setup as Design 1. Specifically, the DGP is

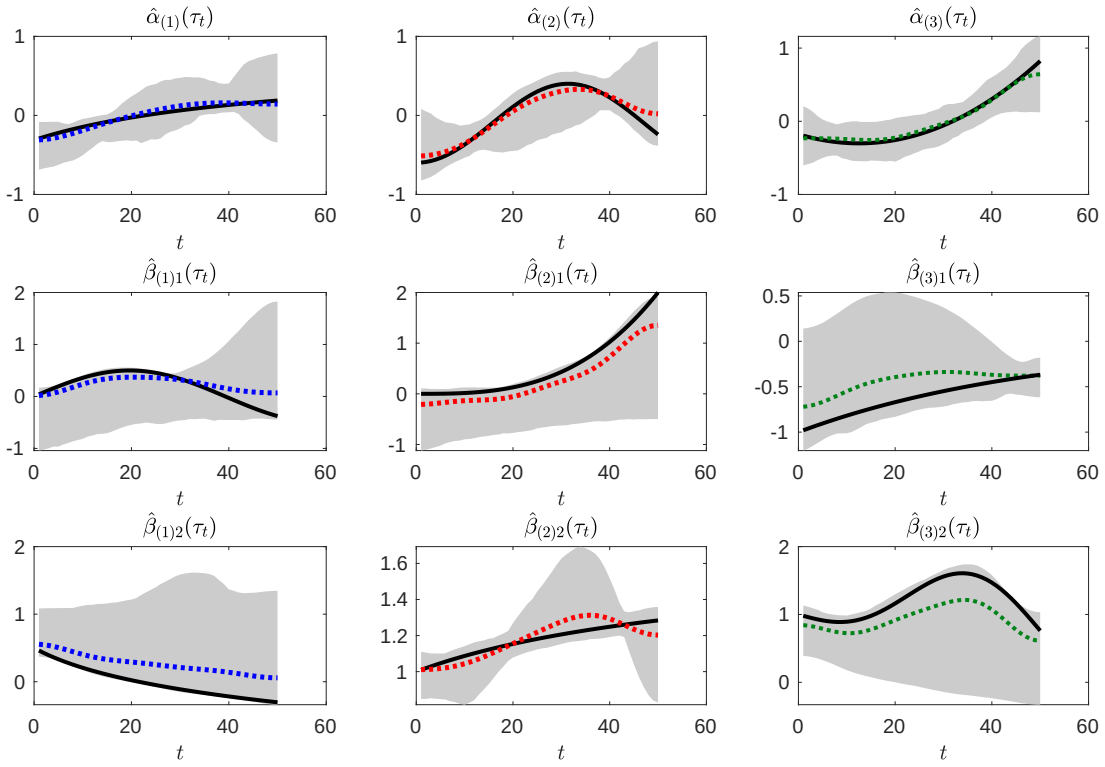
$$y_{it} = \alpha_i^0 - u_i + \alpha(\tau_t) + x_{it}\beta(\tau_t) + v_{it},$$

where $x_{it} \sim N(2, 0.75^2)$. The frontiers for both groups are specified as $\alpha(s) = \log s \sin(6s) - \varpi$,

and $\beta(s) = 7\sin(5s)\exp(-5s)$, where ϖ denotes the mean of $\log s \sin(6s)$. The error terms are generated from two distinct groups. Group 1 errors are generated from $v_{it} \overset{iid}{\sim} N(0, \sigma^2_{v(1)})$, while group 2 errors are generated from $v_{it} \overset{iid}{\sim} N(0, \sigma^2_{v(2)})$. Standard deviations are specified as $\sigma_{v(1)} = 0.5$ and $\sigma_{v(2)} = 1.5$. Similarly to Design 1, we consider two sub-cases where $\alpha^0 - u$ either comes from a unique (DGP2U) or mixture distribution (DGP2M), the same as in Design 1.

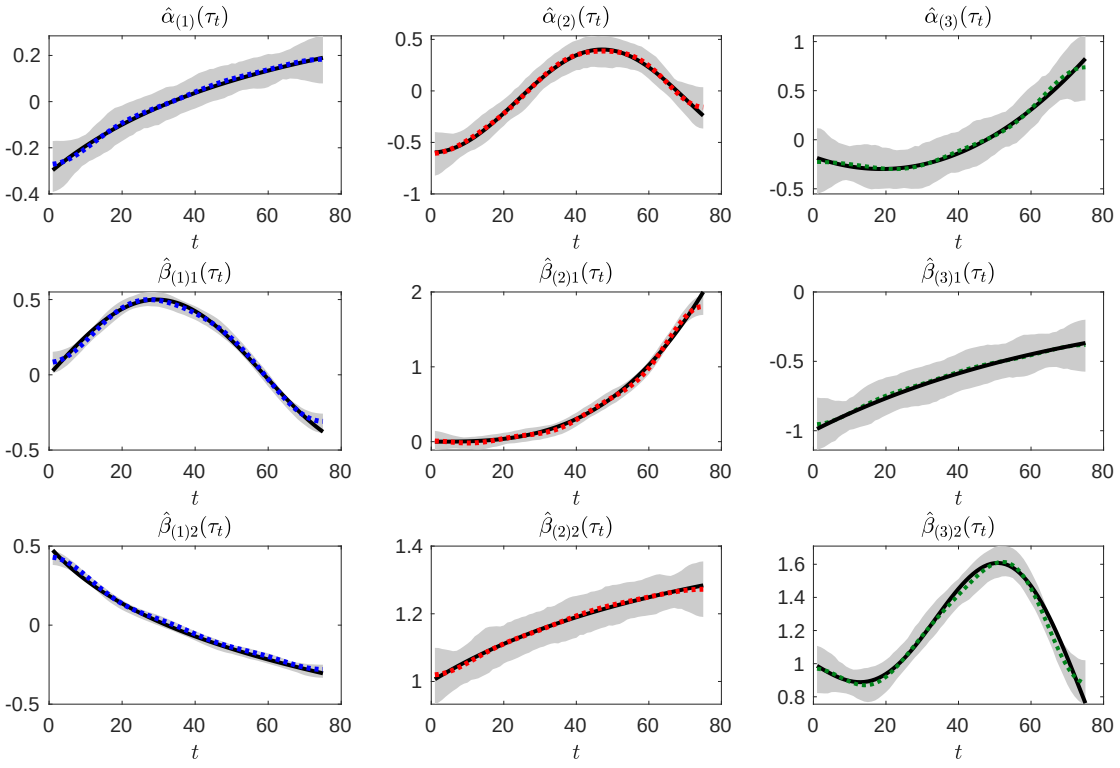
F.2 Additional Tables and Figures for Section 4 and Section 5

Figure F.1: Estimates of the grouped frontiers for DGP3M with $N = 500$ and $T = 50$



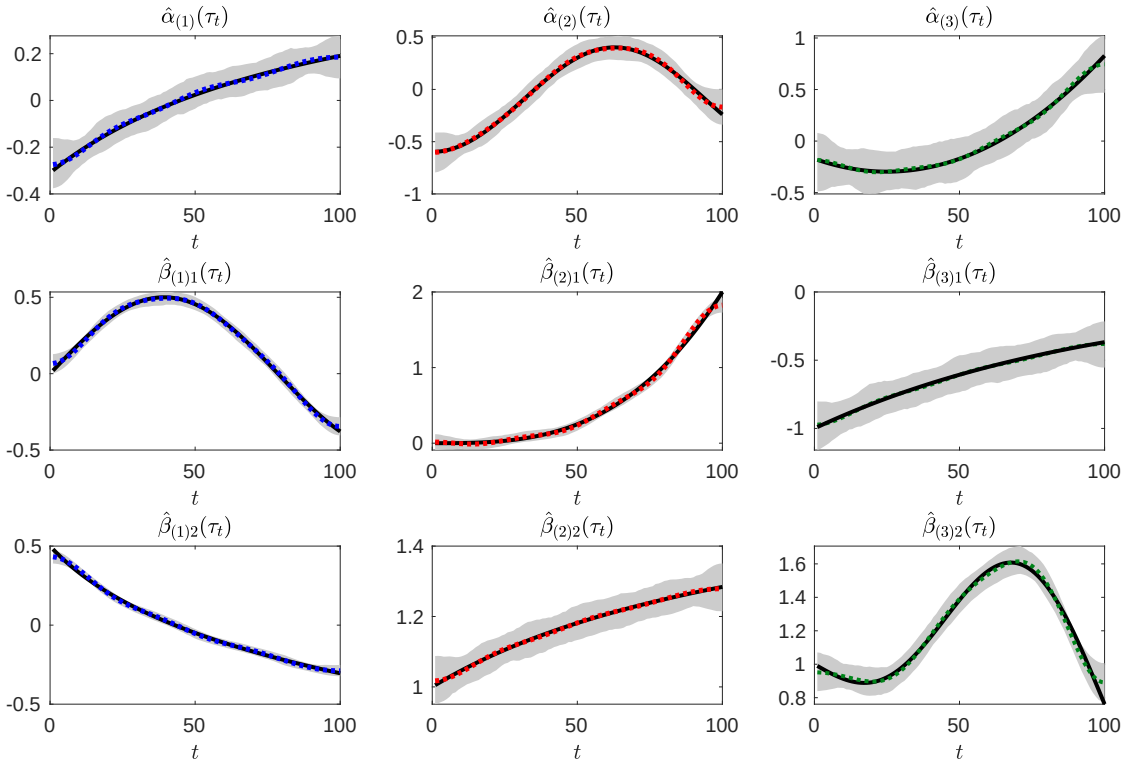
Note: Black solid line depict the true time-varying frontier, dotted lines depict the mean of the estimated grouped frontiers averaged over 500 MC iterations, and the grey shaded region depict the 90 percentile of the estimates. The gray area is wide, due to mis-classification errors.

Figure F.2: Estimates of the grouped frontiers for DGP3M with $N = 500$ and $T = 75$



Note: Black solid line depict the true time-varying frontier, dotted lines depict the mean of the estimated grouped frontiers averaged over 500 MC iterations, and the gray shaded region depict the 90 percentile of the estimates.

Figure F.3: Estimates of the grouped frontiers for DGP3M with $N = 500$ and $T = 100$



Note: Black solid line depict the true time-varying frontier, dotted lines depict the mean of the estimated grouped frontiers averaged over 500 MC iterations, and the gray shaded region depict the 90 percentile of the estimates.

Table F.1: BIAS and RMSE over 500 MC iterations for DGP1U

	$\hat{\sigma}_{v(1)}$		$\hat{\sigma}_{v(2)}$		$\hat{\alpha}^0$		$\hat{\sigma}_u$	
(N,T)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
(100,50)	0.015	0.020	0.034	0.038	0.046	0.058	0.066	0.082
(100,75)	0.014	0.019	0.033	0.036	0.039	0.049	0.062	0.077
(100,100)	0.012	0.017	0.034	0.037	0.036	0.046	0.063	0.081
(250,50)	0.008	0.011	0.017	0.019	0.035	0.043	0.042	0.053
(250,75)	0.007	0.009	0.016	0.018	0.027	0.034	0.042	0.052
(250,100)	0.005	0.007	0.011	0.013	0.023	0.030	0.042	0.052
(500,50)	0.005	0.007	0.012	0.013	0.030	0.035	0.029	0.036
(500,75)	0.005	0.006	0.011	0.012	0.023	0.028	0.028	0.035
(500,100)	0.004	0.005	0.007	0.008	0.019	0.024	0.029	0.035

Table F.2: BIAS and RMSE over 500 MC iterations for DGP1M

	$\hat{\sigma}_{v(1)}$		$\hat{\sigma}_{v(2)}$		$\hat{\tau}$		$\hat{\alpha}^0_{(1)}$		$\hat{\sigma}_{u(1)}$		$\hat{\alpha}^0_{(2)}$		$\hat{\sigma}_{u(2)}$	
(N,T)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
(100,50)	0.015	0.020	0.034	0.038	0.012	0.018	0.059	0.074	0.103	0.131	0.089	0.115	0.122	0.153
(100,75)	0.014	0.019	0.033	0.036	0.010	0.014	0.050	0.064	0.091	0.113	0.070	0.090	0.116	0.144
(100,100)	0.012	0.017	0.034	0.037	0.011	0.016	0.046	0.058	0.097	0.122	0.072	0.095	0.114	0.144
(250,50)	0.008	0.011	0.017	0.019	0.008	0.010	0.041	0.050	0.062	0.080	0.055	0.069	0.076	0.098
(250,75)	0.007	0.009	0.016	0.018	0.007	0.009	0.033	0.042	0.062	0.078	0.046	0.058	0.071	0.087
(250,100)	0.005	0.007	0.011	0.013	0.007	0.009	0.027	0.035	0.058	0.073	0.044	0.056	0.072	0.089
(500,50)	0.005	0.007	0.012	0.013	0.005	0.007	0.034	0.041	0.048	0.060	0.042	0.052	0.050	0.062
(500,75)	0.005	0.006	0.011	0.012	0.005	0.006	0.027	0.033	0.040	0.051	0.035	0.044	0.048	0.061
(500,100)	0.004	0.005	0.007	0.008	0.004	0.005	0.022	0.027	0.041	0.051	0.032	0.040	0.051	0.063

Table F.3: BIAS and RMSE over 500 MC iterations for DGP2U

	$\hat{\sigma}_{v(1)}$		$\hat{\sigma}_{v(2)}$		$\hat{\alpha}^0$		$\hat{\sigma}_u$	
(N,T)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
(100,50)	0.006	0.008	0.017	0.022	0.039	0.050	0.068	0.084
(100,75)	0.005	0.006	0.014	0.018	0.035	0.045	0.061	0.076
(100,100)	0.004	0.005	0.012	0.015	0.032	0.041	0.062	0.079
(250,50)	0.004	0.005	0.011	0.014	0.024	0.031	0.041	0.052
(250,75)	0.003	0.004	0.009	0.011	0.021	0.027	0.041	0.051
(250,100)	0.002	0.003	0.008	0.009	0.018	0.023	0.041	0.051
(500,50)	0.002	0.003	0.008	0.009	0.017	0.021	0.030	0.037
(500,75)	0.002	0.003	0.006	0.008	0.015	0.019	0.028	0.035
(500,100)	0.002	0.002	0.005	0.007	0.014	0.017	0.029	0.035

Table F.4: BIAS and RMSE over 500 MC iterations for DGP2M

	$\hat{\sigma}_{v(1)}$		$\hat{\sigma}_{v(2)}$		$\hat{\tau}$		$\hat{\alpha}^0_{(1)}$		$\hat{\sigma}_{u(1)}$		$\hat{\alpha}^0_{(2)}$		$\hat{\sigma}_{u(2)}$	
(N,T)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
(100,50)	0.006	0.008	0.017	0.022	0.019	0.027	0.040	0.050	0.105	0.139	0.125	0.160	0.133	0.166
(100,75)	0.005	0.006	0.014	0.018	0.014	0.019	0.034	0.044	0.089	0.115	0.105	0.132	0.121	0.150
(100,100)	0.004	0.005	0.012	0.015	0.015	0.021	0.032	0.041	0.098	0.126	0.100	0.129	0.121	0.153
(250,50)	0.004	0.005	0.011	0.014	0.014	0.018	0.025	0.031	0.071	0.092	0.084	0.106	0.082	0.106
(250,75)	0.003	0.004	0.009	0.011	0.012	0.015	0.021	0.026	0.067	0.087	0.072	0.090	0.079	0.098
(250,100)	0.002	0.003	0.008	0.009	0.010	0.013	0.017	0.022	0.063	0.080	0.070	0.086	0.079	0.097
(500,50)	0.002	0.003	0.008	0.009	0.014	0.016	0.018	0.022	0.061	0.077	0.063	0.079	0.058	0.073
(500,75)	0.002	0.003	0.006	0.008	0.009	0.011	0.015	0.019	0.048	0.061	0.052	0.065	0.054	0.068
(500,100)	0.002	0.002	0.005	0.007	0.008	0.010	0.013	0.017	0.048	0.062	0.051	0.063	0.055	0.070

Table F.5: BIAS and RMSE over 500 MC iterations for DGP3U

	$\hat{\sigma}_{v(1)}$		$\hat{\sigma}_{v(2)}$		$\hat{\sigma}_{v(3)}$		$\hat{\alpha}^0$		$\hat{\sigma}_u$	
(N,T)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
(100,50)	0.268	0.463	0.127	0.215	0.367	0.594	0.059	0.079	0.074	0.095
(100,75)	0.082	0.214	0.074	0.174	0.154	0.372	0.043	0.053	0.066	0.081
(100,100)	0.024	0.099	0.027	0.093	0.053	0.196	0.040	0.052	0.064	0.081
(250,50)	0.204	0.373	0.138	0.243	0.327	0.562	0.037	0.048	0.053	0.067
(250,75)	0.034	0.129	0.035	0.116	0.069	0.241	0.024	0.031	0.040	0.049
(250,100)	0.005	0.033	0.008	0.032	0.014	0.064	0.023	0.029	0.041	0.050
(500,50)	0.145	0.307	0.110	0.218	0.242	0.484	0.027	0.036	0.042	0.053
(500,75)	0.009	0.064	0.009	0.044	0.018	0.100	0.018	0.023	0.030	0.037
(500,100)	0.002	0.003	0.004	0.005	0.007	0.009	0.016	0.020	0.029	0.036

Table F.6: Performance of ICs for DGP1U

(N,T)	$K=1$	$K=2$	$K=3$	$K=4$	$\alpha^0 - u$ unique	$\alpha^0 - u$ mix
(100,50)	0.000	1.000	0.000	0.000	0.926	0.074
(100,75)	0.000	1.000	0.000	0.000	0.940	0.060
(100,100)	0.000	1.000	0.000	0.000	0.938	0.062
(250,50)	0.000	1.000	0.000	0.000	0.996	0.004
(250,75)	0.000	1.000	0.000	0.000	0.992	0.008
(250,100)	0.000	1.000	0.000	0.000	0.988	0.012
(500,50)	0.000	1.000	0.000	0.000	0.996	0.004
(500,75)	0.000	1.000	0.000	0.000	1.000	0.000
(500,100)	0.000	1.000	0.000	0.000	1.000	0.000

Note: Results for the baseline case $c_\lambda = \tilde{c}_\lambda = 1$. Reported numbers are probabilities across replications.

Table F.7: Performance of ICs for DGP1M

(N,T)	$K=1$	$K=2$	$K=3$	$K=4$	α^0-u unique	α^0-u mix
(100,50)	0.000	1.000	0.000	0.000	0.000	1.000
(100,75)	0.000	1.000	0.000	0.000	0.000	1.000
(100,100)	0.000	1.000	0.000	0.000	0.000	1.000
(250,50)	0.000	1.000	0.000	0.000	0.000	1.000
(250,75)	0.000	1.000	0.000	0.000	0.000	1.000
(250,100)	0.000	1.000	0.000	0.000	0.000	1.000
(500,50)	0.000	1.000	0.000	0.000	0.000	1.000
(500,75)	0.000	1.000	0.000	0.000	0.000	1.000
(500,100)	0.000	1.000	0.000	0.000	0.000	1.000

Note:Results for the baseline case $c_\lambda = \tilde{c}_\lambda = 1$. Reported numbers are probabilities across replications.

Table F.8: Performance of ICs for DGP2U

(N,T)	$K=1$	$K=2$	$K=3$	$K=4$	α^0-u unique	α^0-u mix
(100,50)	0.000	1.000	0.000	0.000	0.868	0.132
(100,75)	0.000	1.000	0.000	0.000	0.838	0.162
(100,100)	0.000	1.000	0.000	0.000	0.878	0.122
(250,50)	0.000	1.000	0.000	0.000	0.978	0.022
(250,75)	0.000	1.000	0.000	0.000	0.984	0.016
(250,100)	0.000	1.000	0.000	0.000	0.970	0.030
(500,50)	0.000	1.000	0.000	0.000	0.996	0.004
(500,75)	0.000	1.000	0.000	0.000	0.998	0.002
(500,100)	0.000	1.000	0.000	0.000	1.000	0.000

Note: Results for the baseline case $c_\lambda = \tilde{c}_\lambda = 1$. Reported numbers are probabilities across replications.

Table F.9: Performance of ICs for DGP2M

(N,T)	$K=1$	$K=2$	$K=3$	$K=4$	α^0-u unique	α^0-u mix
(100,50)	0.000	1.000	0.000	0.000	0.000	1.000
(100,75)	0.000	1.000	0.000	0.000	0.000	1.000
(100,100)	0.000	1.000	0.000	0.000	0.000	1.000
(250,50)	0.000	1.000	0.000	0.000	0.000	1.000
(250,75)	0.000	1.000	0.000	0.000	0.000	1.000
(250,100)	0.000	1.000	0.000	0.000	0.000	1.000
(500,50)	0.000	1.000	0.000	0.000	0.000	1.000
(500,75)	0.000	1.000	0.000	0.000	0.000	1.000
(500,100)	0.000	1.000	0.000	0.000	0.000	1.000

Note: Results for the baseline case $c_\lambda = \tilde{c}_\lambda = 1$. Reported numbers are probabilities across replications.

Table F.10: Performance of ICs for DGP3U

(N,T)	$K=1$	$K=2$	$K=3$	$K=4$	α^0-u unique	α^0-u mix
(100,50)	0.000	0.352	0.648	0.000	0.784	0.216
(100,75)	0.000	0.078	0.922	0.000	0.840	0.160
(100,100)	0.000	0.000	1.000	0.000	0.874	0.126
(250,50)	0.000	0.086	0.914	0.000	0.942	0.058
(250,75)	0.000	0.000	1.000	0.000	0.988	0.012
(250,100)	0.000	0.000	1.000	0.000	0.992	0.008
(500,50)	0.000	0.006	0.994	0.000	0.996	0.004
(500,75)	0.000	0.000	1.000	0.000	1.000	0.000
(500,100)	0.000	0.000	1.000	0.000	1.000	0.000

Note: Results for the baseline case $c_\lambda = \tilde{c}_\lambda = 1$. Reported numbers are probabilities across replications.

Table F.11: Sensitivity analysis for classification error in DGP1U and DGP1M

	$c_\lambda = 3/2$	$c_\lambda = 1$ (bench.)	$c_\lambda = 3/4$
(N, T)	$\bar{\text{Pr}}(F)$	$\bar{\text{Pr}}(F)$	$\bar{\text{Pr}}(F)$
(100,50)	0.144	0.144	0.144
(100,75)	0.152	0.152	0.152
(100,100)	0.140	0.140	0.140
(250,50)	0.140	0.140	0.140
(250,75)	0.102	0.102	0.102
(250,100)	0.068	0.068	0.068
(500,50)	0.070	0.070	0.070
(500,75)	0.068	0.068	0.068
(500,100)	0.034	0.034	0.034

Table F.12: Sensitivity analysis for classification error in DGP2U and DGP2M

	$c_\lambda = 3/2$	$c_\lambda = 1$ (bench.)	$c_\lambda = 3/4$
(N, T)	$\bar{\text{Pr}}(F)$	$\bar{\text{Pr}}(F)$	$\bar{\text{Pr}}(F)$
(100,50)	0.000	0.000	0.000
(100,75)	0.000	0.000	0.000
(100,100)	0.000	0.000	0.000
(250,50)	0.000	0.000	0.000
(250,75)	0.000	0.000	0.000
(250,100)	0.000	0.000	0.000
(500,50)	0.000	0.000	0.000
(500,75)	0.000	0.000	0.000
(500,100)	0.000	0.000	0.000

Table F.13: Sensitivity analysis for classification error in DGP3U and DGP3M

	$c_\lambda = 3/2$	$c_\lambda = 1$ (bench.)	$c_\lambda = 3/4$
(N,T)	$\bar{\text{Pr}}(F)$	$\bar{\text{Pr}}(F)$	$\bar{\text{Pr}}(F)$
(100,50)	0.122	0.106	0.198
(100,75)	0.046	0.024	0.137
(100,100)	0.012	0.024	0.039
(250,50)	0.108	0.026	0.294
(250,75)	0.037	0.026	0.060
(250,100)	0.005	0.026	0.005
(500,50)	0.178	0.002	0.228
(500,75)	0.012	0.002	0.012
(500,100)	0.001	0.002	0.001

Table F.14: Sensitivity analysis for $\alpha^0 - u$ mixture structure for DGP1U

	$\tilde{c}_\lambda = 3/2$		$\tilde{c}_\lambda = 1$ (bench.)		$\tilde{c}_\lambda = 3/4$	
(N,T)	unique	mix	unique	mix	unique	mix
(100,50)	1.000	0.000	0.994	0.006	0.912	0.088
(100,75)	1.000	0.000	0.992	0.008	0.908	0.092
(100,100)	1.000	0.000	0.996	0.004	0.934	0.066
(250,50)	1.000	0.000	1.000	0.000	0.990	0.010
(250,75)	1.000	0.000	1.000	0.000	0.986	0.014
(250,100)	1.000	0.000	1.000	0.000	0.996	0.004
(500,50)	1.000	0.000	1.000	0.000	1.000	0.000
(500,75)	1.000	0.000	1.000	0.000	0.998	0.002
(500,100)	1.000	0.000	1.000	0.000	1.000	0.000

Table F.15: Sensitivity analysis for $\alpha^0 - u$ mixture structure for DGP1M

	$\tilde{c}_\lambda = 3/2$		$\tilde{c}_\lambda = 1$ (bench.)		$\tilde{c}_\lambda = 3/4$	
(N,T)	unique	mix	unique	mix	unique	mix
(100,50)	0.244	0.756	0.006	0.994	0.000	1.000
(100,75)	0.210	0.790	0.000	1.000	0.000	1.000
(100,100)	0.142	0.858	0.002	0.998	0.000	1.000
(250,50)	0.018	0.982	0.000	1.000	0.000	1.000
(250,75)	0.010	0.990	0.000	1.000	0.000	1.000
(250,100)	0.002	0.998	0.000	1.000	0.000	1.000
(500,50)	0.000	1.000	0.000	1.000	0.000	1.000
(500,75)	0.000	1.000	0.000	1.000	0.000	1.000
(500,100)	0.000	1.000	0.000	1.000	0.000	1.000

Table F.16: Sensitivity analysis for $\alpha^0 - u$ mixture structure for DGP2U

	$\tilde{c}_\lambda = 3/2$		$\tilde{c}_\lambda = 1$ (bench.)		$\tilde{c}_\lambda = 3/4$	
(N,T)	unique	mix	unique	mix	unique	mix
(100,50)	1.000	0.000	0.986	0.014	0.844	0.156
(100,75)	1.000	0.000	0.986	0.014	0.844	0.156
(100,100)	1.000	0.000	0.994	0.006	0.854	0.146
(250,50)	1.000	0.000	1.000	0.000	0.978	0.022
(250,75)	1.000	0.000	1.000	0.000	0.978	0.022
(250,100)	1.000	0.000	1.000	0.000	0.992	0.008
(500,50)	1.000	0.000	1.000	0.000	1.000	0.000
(500,75)	1.000	0.000	1.000	0.000	1.000	0.000
(500,100)	1.000	0.000	1.000	0.000	1.000	0.000

Table F.17: Sensitivity analysis for $\alpha^0 - u$ mixture structure for DGP2M

	$\tilde{c}_\lambda = 3/2$		$\tilde{c}_\lambda = 1$ (bench.)		$\tilde{c}_\lambda = 3/4$	
(N,T)	unique	mix	unique	mix	unique	mix
(100,50)	0.360	0.640	0.028	0.972	0.000	1.000
(100,75)	0.304	0.696	0.010	0.990	0.000	1.000
(100,100)	0.210	0.790	0.008	0.992	0.000	1.000
(250,50)	0.094	0.906	0.000	1.000	0.000	1.000
(250,75)	0.052	0.948	0.000	1.000	0.000	1.000
(250,100)	0.022	0.978	0.000	1.000	0.000	1.000
(500,50)	0.006	0.994	0.000	1.000	0.000	1.000
(500,75)	0.000	1.000	0.000	1.000	0.000	1.000
(500,100)	0.000	1.000	0.000	1.000	0.000	1.000

Table F.18: Sensitivity analysis for $\alpha^0 - u$ mixture structure for DGP3U

	$\tilde{c}_\lambda = 3/2$		$\tilde{c}_\lambda = 1$ (bench.)		$\tilde{c}_\lambda = 3/4$	
(N,T)	unique	mix	unique	mix	unique	mix
(100,50)	1.000	0.000	0.968	0.032	0.778	0.222
(100,75)	1.000	0.000	0.970	0.030	0.810	0.190
(100,100)	1.000	0.000	0.992	0.008	0.852	0.148
(250,50)	1.000	0.000	0.992	0.008	0.948	0.052
(250,75)	1.000	0.000	1.000	0.000	0.976	0.024
(250,100)	1.000	0.000	1.000	0.000	0.976	0.024
(500,50)	1.000	0.000	1.000	0.000	0.994	0.006
(500,75)	1.000	0.000	1.000	0.000	1.000	0.000
(500,100)	1.000	0.000	1.000	0.000	1.000	0.000

Table F.19: Sensitivity analysis for $\alpha^0 - u$ mixture structure for DGP3M

	$\tilde{c}_\lambda = 3/2$		$\tilde{c}_\lambda = 1$ (bench.)		$\tilde{c}_\lambda = 3/4$	
(N, T)	unique	mix	unique	mix	unique	mix
(100,50)	0.496	0.504	0.064	0.936	0.006	0.994
(100,75)	0.328	0.672	0.018	0.982	0.002	0.998
(100,100)	0.258	0.742	0.016	0.984	0.000	1.000
(250,50)	0.148	0.852	0.010	0.990	0.006	0.994
(250,75)	0.038	0.962	0.000	1.000	0.000	1.000
(250,100)	0.028	0.972	0.000	1.000	0.000	1.000
(500,50)	0.006	0.994	0.002	0.998	0.002	0.998
(500,75)	0.000	1.000	0.000	1.000	0.000	1.000
(500,100)	0.000	1.000	0.000	1.000	0.000	1.000

Table F.20: Summary statistics of data used in the application

	MEAN	STD	MIN	MAX	$N \times T$
	Panel A: raw variables				
C_{it}	9.7974×10^4	7.0182×10^5	91.3791	2.2152×10^7	37280
W_{it1}	17.4185	9.7993	0.0913	252.9964	37280
W_{it2}	0.0073	0.0094	7.9946×10^{-7}	0.4393	37280
W_{it3}	0.02718	0.0200	6.1287×10^{-4}	0.4423	37280
Y_{it1}	2.1086×10^5	1.4001×10^6	4.2900	4.9625×10^7	37280
Y_{it2}	1.6894×10^6	1.0978×10^7	3020	3.1730×10^8	37280
Y_{it3}	1.3421×10^6	1.1142×10^7	2.4091×10^3	3.9273×10^8	37280
	Panel B: variables used in regressions				
$\log c_{it}^*$	12.8442	1.6370	9.1203	20.3854	37280
$\log w_{it1}$	6.5320	0.6060	2.8625	10.1837	37280
$\log w_{it2}$	-1.7023	1.0291	-8.4596	2.0101	37280
$\log y_{it1}$	9.8573	1.8283	1.4563	17.7200	37280
$\log y_{it2}$	11.8815	1.7312	8.0130	19.5754	37280
$\log y_{it3}$	11.6393	1.6402	7.7870	19.7886	37280

Note: Raw variables in Panel A are denominated in Millions of 1986 USD. Variables in Panel B have been divided by W_{it3} before taken logs.

F.3 Tables for Simulations in Appendix B

Table F.21: BIAS and RMSE over 500 MC Iterations for DGP1M

	$\hat{\alpha}^0_{(1)}$		$\hat{\sigma}_{u(1)}$		$\hat{\alpha}^0_{(2)}$		$\hat{\sigma}_{u(2)}$		$\hat{\tau}_1$	
(N,T)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
(100,50)	0.116	0.140	0.242	0.300	0.106	0.136	0.119	0.149	0.031	0.040
(100,75)	0.094	0.115	0.184	0.230	0.086	0.112	0.116	0.145	0.029	0.037
(100,100)	0.079	0.096	0.184	0.238	0.095	0.127	0.111	0.141	0.031	0.041
(250,50)	0.173	0.189	0.328	0.365	0.087	0.112	0.087	0.110	0.031	0.041
(250,75)	0.135	0.149	0.255	0.289	0.074	0.102	0.081	0.102	0.026	0.034
(250,100)	0.111	0.123	0.211	0.241	0.070	0.095	0.078	0.097	0.026	0.033
(500,50)	0.206	0.215	0.361	0.386	0.071	0.092	0.069	0.088	0.029	0.037
(500,75)	0.165	0.174	0.273	0.293	0.055	0.070	0.061	0.075	0.022	0.027
(500,100)	0.143	0.152	0.232	0.253	0.051	0.068	0.066	0.085	0.022	0.028

Table F.22: BIAS and RMSE over 500 MC iterations for DGP1T

	$\hat{\alpha}^0_{(1)}$		$\hat{\sigma}_{u(1)}$		$\hat{\alpha}^0_{(2)}$		$\hat{\sigma}_{u(2)}$		$\hat{\alpha}^0_{(3)}$		$\hat{\sigma}_{u(3)}$		$\hat{\tau}_1$		$\hat{\tau}_2$	
(N,T)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
(100,50)	0.095	0.123	0.240	0.312	0.080	0.103	0.104	0.133	0.093	0.117	0.117	0.145	0.038	0.047	0.038	0.048
(100,75)	0.084	0.109	0.222	0.282	0.071	0.095	0.106	0.137	0.078	0.099	0.108	0.133	0.037	0.048	0.037	0.047
(100,100)	0.083	0.109	0.223	0.282	0.069	0.091	0.107	0.135	0.071	0.092	0.102	0.128	0.037	0.048	0.035	0.046
(250,50)	0.073	0.094	0.235	0.293	0.066	0.088	0.074	0.096	0.125	0.148	0.147	0.173	0.033	0.042	0.038	0.048
(250,75)	0.066	0.085	0.229	0.287	0.056	0.075	0.075	0.096	0.098	0.119	0.118	0.142	0.033	0.042	0.035	0.045
(250,100)	0.057	0.074	0.199	0.257	0.056	0.073	0.072	0.093	0.082	0.100	0.103	0.127	0.030	0.037	0.031	0.040
(500,50)	0.060	0.075	0.263	0.317	0.057	0.073	0.059	0.077	0.163	0.179	0.183	0.200	0.030	0.039	0.043	0.052
(500,75)	0.050	0.066	0.237	0.290	0.048	0.060	0.058	0.073	0.128	0.143	0.148	0.168	0.028	0.036	0.037	0.045
(500,100)	0.050	0.065	0.217	0.270	0.042	0.053	0.052	0.067	0.110	0.125	0.128	0.148	0.027	0.034	0.033	0.042

Table F.23: BIAS and RMSE over 500 MC Iterations for DGP2M

	$\hat{\alpha}^0_{(1)}$		$\hat{\sigma}_{u(1)}$		$\hat{\alpha}^0_{(2)}$		$\hat{\sigma}_{u(2)}$		$\hat{\tau}_1$	
(N,T)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
(100,50)	0.059	0.075	0.168	0.221	0.125	0.159	0.123	0.155	0.036	0.046
(100,75)	0.047	0.059	0.137	0.181	0.106	0.134	0.126	0.154	0.030	0.039
(100,100)	0.040	0.050	0.132	0.180	0.096	0.125	0.113	0.144	0.032	0.042
(250,50)	0.082	0.094	0.196	0.231	0.102	0.128	0.104	0.127	0.033	0.040
(250,75)	0.061	0.071	0.158	0.191	0.088	0.112	0.096	0.120	0.029	0.036
(250,100)	0.048	0.056	0.144	0.176	0.075	0.095	0.090	0.113	0.029	0.036
(500,50)	0.104	0.111	0.221	0.245	0.088	0.110	0.087	0.108	0.027	0.034
(500,75)	0.078	0.086	0.175	0.197	0.069	0.088	0.087	0.108	0.026	0.032
(500,100)	0.066	0.072	0.154	0.173	0.061	0.076	0.086	0.106	0.025	0.030

Table F.24: BIAS and RMSE over 500 MC iterations for DGP2T

	$\hat{\alpha}^0_{(1)}$		$\hat{\sigma}_{u(1)}$		$\hat{\alpha}^0_{(2)}$		$\hat{\sigma}_{u(2)}$		$\hat{\alpha}^0_{(3)}$		$\hat{\sigma}_{u(3)}$		$\hat{\tau}_1$		$\hat{\tau}_2$	
(N,T)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
(100,50)	0.085	0.114	0.221	0.287	0.078	0.104	0.108	0.140	0.172	0.209	0.147	0.177	0.037	0.048	0.039	0.051
(100,75)	0.074	0.099	0.197	0.259	0.078	0.104	0.112	0.143	0.139	0.168	0.125	0.155	0.037	0.048	0.036	0.045
(100,100)	0.073	0.099	0.210	0.273	0.076	0.102	0.105	0.137	0.114	0.142	0.115	0.145	0.037	0.048	0.037	0.048
(250,50)	0.061	0.085	0.201	0.266	0.064	0.085	0.071	0.091	0.238	0.266	0.205	0.230	0.030	0.039	0.041	0.052
(250,75)	0.054	0.073	0.195	0.248	0.060	0.082	0.073	0.094	0.191	0.218	0.165	0.191	0.030	0.039	0.038	0.049
(250,100)	0.046	0.063	0.190	0.243	0.053	0.071	0.074	0.097	0.165	0.188	0.143	0.173	0.031	0.040	0.036	0.046
(500,50)	0.045	0.059	0.206	0.259	0.053	0.069	0.063	0.080	0.301	0.318	0.257	0.273	0.026	0.034	0.042	0.052
(500,75)	0.040	0.053	0.194	0.244	0.043	0.055	0.057	0.072	0.236	0.251	0.213	0.231	0.026	0.034	0.039	0.047
(500,100)	0.039	0.051	0.188	0.238	0.041	0.053	0.056	0.071	0.211	0.227	0.183	0.201	0.026	0.033	0.037	0.046

Table F.25: BIAS and RMSE over 500 MC iterations for DGP3M

	$\hat{\alpha}^0_{(1)}$		$\hat{\sigma}_{u(1)}$		$\hat{\alpha}^0_{(2)}$		$\hat{\sigma}_{u(2)}$		$\hat{\tau}_1$	
(N,T)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
(100,50)	0.127	0.173	0.247	0.335	0.135	0.172	0.139	0.176	0.037	0.050
(100,75)	0.083	0.109	0.179	0.231	0.106	0.136	0.123	0.154	0.030	0.040
(100,100)	0.069	0.088	0.148	0.194	0.102	0.131	0.109	0.139	0.030	0.039
(250,50)	0.173	0.203	0.316	0.371	0.106	0.136	0.108	0.135	0.034	0.045
(250,75)	0.113	0.133	0.222	0.253	0.081	0.104	0.083	0.104	0.029	0.036
(250,100)	0.089	0.107	0.175	0.203	0.075	0.095	0.089	0.112	0.026	0.032
(500,50)	0.206	0.225	0.335	0.370	0.083	0.108	0.101	0.125	0.031	0.041
(500,75)	0.139	0.151	0.237	0.260	0.063	0.079	0.073	0.094	0.024	0.030
(500,100)	0.116	0.127	0.207	0.226	0.059	0.076	0.070	0.089	0.023	0.028

Table F.26: BIAS and RMSE over 500 MC iterations for DGP3T

	$\hat{\alpha}^0_{(1)}$		$\hat{\sigma}_{u(1)}$		$\hat{\alpha}^0_{(2)}$		$\hat{\sigma}_{u(2)}$		$\hat{\alpha}^0_{(3)}$		$\hat{\sigma}_{u(3)}$		$\hat{\tau}_1$		$\hat{\tau}_2$	
(N,T)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
(100,50)	0.087	0.116	0.211	0.275	0.102	0.132	0.114	0.150	0.203	0.249	0.366	0.442	0.039	0.050	0.041	0.054
(100,75)	0.076	0.103	0.226	0.281	0.086	0.113	0.103	0.131	0.149	0.184	0.281	0.357	0.036	0.046	0.036	0.047
(100,100)	0.074	0.099	0.217	0.277	0.083	0.107	0.109	0.138	0.129	0.161	0.251	0.323	0.036	0.049	0.037	0.048
(250,50)	0.063	0.085	0.189	0.250	0.075	0.096	0.079	0.099	0.232	0.265	0.520	0.582	0.032	0.041	0.044	0.056
(250,75)	0.054	0.072	0.177	0.227	0.062	0.079	0.072	0.091	0.186	0.214	0.359	0.418	0.027	0.035	0.034	0.043
(250,100)	0.051	0.070	0.178	0.232	0.055	0.073	0.070	0.090	0.150	0.176	0.299	0.359	0.029	0.036	0.032	0.041
(500,50)	0.051	0.065	0.187	0.243	0.064	0.081	0.064	0.079	0.296	0.317	0.674	0.721	0.028	0.037	0.048	0.058
(500,75)	0.042	0.054	0.182	0.230	0.054	0.067	0.057	0.074	0.236	0.253	0.457	0.506	0.027	0.034	0.037	0.046
(500,100)	0.038	0.051	0.180	0.232	0.047	0.060	0.055	0.070	0.200	0.217	0.391	0.437	0.025	0.033	0.035	0.044

Table F.27: Sensitivity analysis for $\alpha^0 - u$ mixture structure for DGP1M

	$\tilde{c}_\lambda = 3/2$			$\tilde{c}_\lambda = 1$ (bench.)			$\tilde{c}_\lambda = 3/4$		
(N,T)	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$
(100,50)	0.114	0.876	0.010	0.024	0.886	0.090	0.008	0.808	0.184
(100,75)	0.050	0.938	0.012	0.006	0.946	0.048	0.000	0.882	0.118
(100,100)	0.048	0.946	0.006	0.004	0.948	0.048	0.000	0.900	0.100
(250,50)	0.012	0.888	0.100	0.002	0.738	0.260	0.002	0.614	0.384
(250,75)	0.012	0.938	0.050	0.000	0.800	0.200	0.000	0.698	0.302
(250,100)	0.006	0.960	0.034	0.000	0.870	0.130	0.000	0.774	0.226
(500,50)	0.000	0.766	0.234	0.000	0.492	0.508	0.000	0.348	0.652
(500,75)	0.000	0.842	0.158	0.000	0.646	0.354	0.000	0.498	0.502
(500,100)	0.000	0.906	0.094	0.000	0.714	0.286	0.000	0.576	0.424

Table F.28: Sensitivity analysis for $\alpha^0 - u$ mixture structure for DGP1T

	$\tilde{c}_\lambda = 3/2$				$\tilde{c}_\lambda = 1$ (bench.)				$\tilde{c}_\lambda = 3/4$			
(N,T)	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=4$	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=4$	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=4$
(100,50)	0.000	0.000	0.990	0.010	0.000	0.000	0.980	0.020	0.000	0.000	0.940	0.060
(100,75)	0.000	0.000	0.998	0.002	0.000	0.000	0.986	0.014	0.000	0.000	0.958	0.042
(100,100)	0.000	0.000	1.000	0.000	0.000	0.000	0.986	0.014	0.000	0.000	0.960	0.040
(250,50)	0.000	0.000	0.992	0.008	0.000	0.000	0.976	0.024	0.000	0.000	0.920	0.080
(250,75)	0.000	0.000	0.994	0.006	0.000	0.000	0.982	0.018	0.000	0.000	0.956	0.044
(250,100)	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.986	0.014
(500,50)	0.000	0.000	0.978	0.022	0.000	0.000	0.926	0.074	0.000	0.000	0.832	0.168
(500,75)	0.000	0.000	0.992	0.008	0.000	0.000	0.974	0.026	0.000	0.000	0.930	0.070
(500,100)	0.000	0.000	0.998	0.002	0.000	0.000	0.990	0.010	0.000	0.000	0.948	0.052

Table F.29: Sensitivity analysis for $\alpha^0 - u$ mixture structure for DGP2M

	$\tilde{c}_\lambda = 3/2$			$\tilde{c}_\lambda = 1$ (bench.)			$\tilde{c}_\lambda = 3/4$		
(N,T)	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$
(100,50)	0.124	0.874	0.002	0.014	0.962	0.024	0.002	0.924	0.074
(100,75)	0.068	0.930	0.002	0.004	0.978	0.018	0.000	0.946	0.054
(100,100)	0.046	0.950	0.004	0.004	0.986	0.010	0.002	0.966	0.032
(250,50)	0.020	0.978	0.002	0.002	0.954	0.044	0.000	0.878	0.122
(250,75)	0.012	0.982	0.006	0.000	0.966	0.034	0.000	0.932	0.068
(250,100)	0.002	0.998	0.000	0.000	0.992	0.008	0.000	0.958	0.042
(500,50)	0.002	0.978	0.020	0.000	0.900	0.100	0.000	0.778	0.222
(500,75)	0.000	0.994	0.006	0.000	0.934	0.066	0.000	0.862	0.138
(500,100)	0.000	0.994	0.006	0.000	0.966	0.034	0.000	0.890	0.110

Table F.30: Sensitivity analysis for $\alpha^0 - u$ mixture structure for DGP2T

	$\tilde{c}_\lambda = 3/2$				$\tilde{c}_\lambda = 1$ (bench.)				$\tilde{c}_\lambda = 3/4$			
(N,T)	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=4$	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=4$	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=4$
(100,50)	0.000	0.000	0.998	0.002	0.000	0.000	0.992	0.008	0.000	0.000	0.974	0.026
(100,75)	0.000	0.000	0.996	0.004	0.000	0.000	0.990	0.010	0.000	0.000	0.982	0.018
(100,100)	0.000	0.000	0.994	0.006	0.000	0.000	0.986	0.014	0.000	0.000	0.976	0.024
(250,50)	0.000	0.000	0.986	0.014	0.000	0.000	0.980	0.020	0.000	0.000	0.940	0.060
(250,75)	0.000	0.000	0.984	0.016	0.000	0.000	0.972	0.028	0.000	0.000	0.952	0.048
(250,100)	0.000	0.000	0.998	0.002	0.000	0.000	0.992	0.008	0.000	0.000	0.974	0.026
(500,50)	0.000	0.000	0.984	0.016	0.000	0.000	0.960	0.040	0.000	0.000	0.890	0.110
(500,75)	0.000	0.000	0.992	0.008	0.000	0.000	0.972	0.028	0.000	0.000	0.940	0.060
(500,100)	0.000	0.000	0.994	0.006	0.000	0.000	0.962	0.038	0.000	0.000	0.920	0.080

Table F.31: Sensitivity analysis for $\alpha^0 - u$ mixture structure for DGP3M

	$\tilde{c}_\lambda = 3/2$			$\tilde{c}_\lambda = 1$ (bench.)			$\tilde{c}_\lambda = 3/4$		
(N,T)	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$
(100,50)	0.346	0.640	0.014	0.166	0.764	0.070	0.100	0.782	0.118
(100,75)	0.148	0.842	0.010	0.054	0.892	0.054	0.024	0.888	0.088
(100,100)	0.070	0.922	0.008	0.010	0.964	0.026	0.002	0.924	0.074
(250,50)	0.152	0.772	0.076	0.052	0.708	0.240	0.034	0.584	0.382
(250,75)	0.008	0.954	0.038	0.000	0.880	0.120	0.000	0.782	0.218
(250,100)	0.010	0.970	0.020	0.000	0.910	0.090	0.000	0.834	0.166
(500,50)	0.050	0.706	0.244	0.030	0.536	0.434	0.024	0.358	0.618
(500,75)	0.002	0.898	0.100	0.000	0.710	0.290	0.000	0.576	0.424
(500,100)	0.000	0.942	0.058	0.000	0.812	0.188	0.000	0.660	0.340

Table F.32: Sensitivity analysis for $\alpha^0 - u$ mixture structure for DGP3T

	$\tilde{c}_\lambda = 3/2$				$\tilde{c}_\lambda = 1$ (bench.)				$\tilde{c}_\lambda = 3/4$			
(N,T)	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=4$	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=4$	$\mathcal{K}=1$	$\mathcal{K}=2$	$\mathcal{K}=3$	$\mathcal{K}=4$
(100,50)	0.000	0.000	0.998	0.002	0.000	0.000	0.992	0.008	0.000	0.000	0.974	0.026
(100,75)	0.000	0.000	0.996	0.004	0.000	0.000	0.990	0.010	0.000	0.000	0.982	0.018
(100,100)	0.000	0.000	0.994	0.006	0.000	0.000	0.986	0.014	0.000	0.000	0.976	0.024
(250,50)	0.000	0.000	0.986	0.014	0.000	0.000	0.980	0.020	0.000	0.000	0.940	0.060
(250,75)	0.000	0.000	0.984	0.016	0.000	0.000	0.972	0.028	0.000	0.000	0.952	0.048
(250,100)	0.000	0.000	0.998	0.002	0.000	0.000	0.992	0.008	0.000	0.000	0.974	0.026
(500,50)	0.000	0.000	0.984	0.016	0.000	0.000	0.960	0.040	0.000	0.000	0.890	0.110
(500,75)	0.000	0.000	0.992	0.008	0.000	0.000	0.972	0.028	0.000	0.000	0.940	0.060
(500,100)	0.000	0.000	0.994	0.006	0.000	0.000	0.962	0.038	0.000	0.000	0.920	0.080

F.4 Interpreting the Latent Groups

The classification is fully data–driven, but it is not completely a black box.

What drives the classification. As seen from Figure 1, the blue dots and red dots are well separated by a horizontal line in Panel (a) and by a vertical line in Panel (e), but are evenly mixed otherwise. Consequently, the classification is mostly driven by coefficient $\hat{\pi}_{i1}$ before $B_1(\tau_t)$ (time trend term in the intercept) and the coefficient $\hat{\pi}_{i8}$ before y_{it2} (level term before non-consumer loans).

What differs technologically. Figures 2 display the group–specific, time–varying coefficients $\{\widehat{\beta}_{k,\ell}(\tau)\}$ and intercepts $\widehat{\alpha}_k(\tau)$. Two facts are visible: (i) the levels/curvature of input elasticities differ across groups, and (ii) their time profiles (drifts) are not parallel. This is exactly the dimension our HAC step is designed to capture: firms in the same group share the same coefficient paths; firms in different groups load inputs and time differently (technology regimes).

We summarize certain statistics for each group below. These numbers indicate correlations only and do not suggest any causal relationship. We find that banks in group 1 are, on average, larger, more efficient (closer to their frontier), and exhibit less variation in inefficiency than banks in group 2.

What differs in performance. To separate technology from performance, we summarize the mean and standard deviation of $\widehat{\mathrm{E}}\left(\alpha_i^0 + u_i | \varepsilon_{i1}, ..., \varepsilon_{iT}\right)$ using (C.4), within each group. Group 1 has the mean (standard deviation) at 0.0923 (2.17×10^{-4}) with 113 observations, while Group 2 has its mean (standard deviation) at 0.1130 (0.0558) with 353 observations. The group means are statistically different at 5% level of significance using

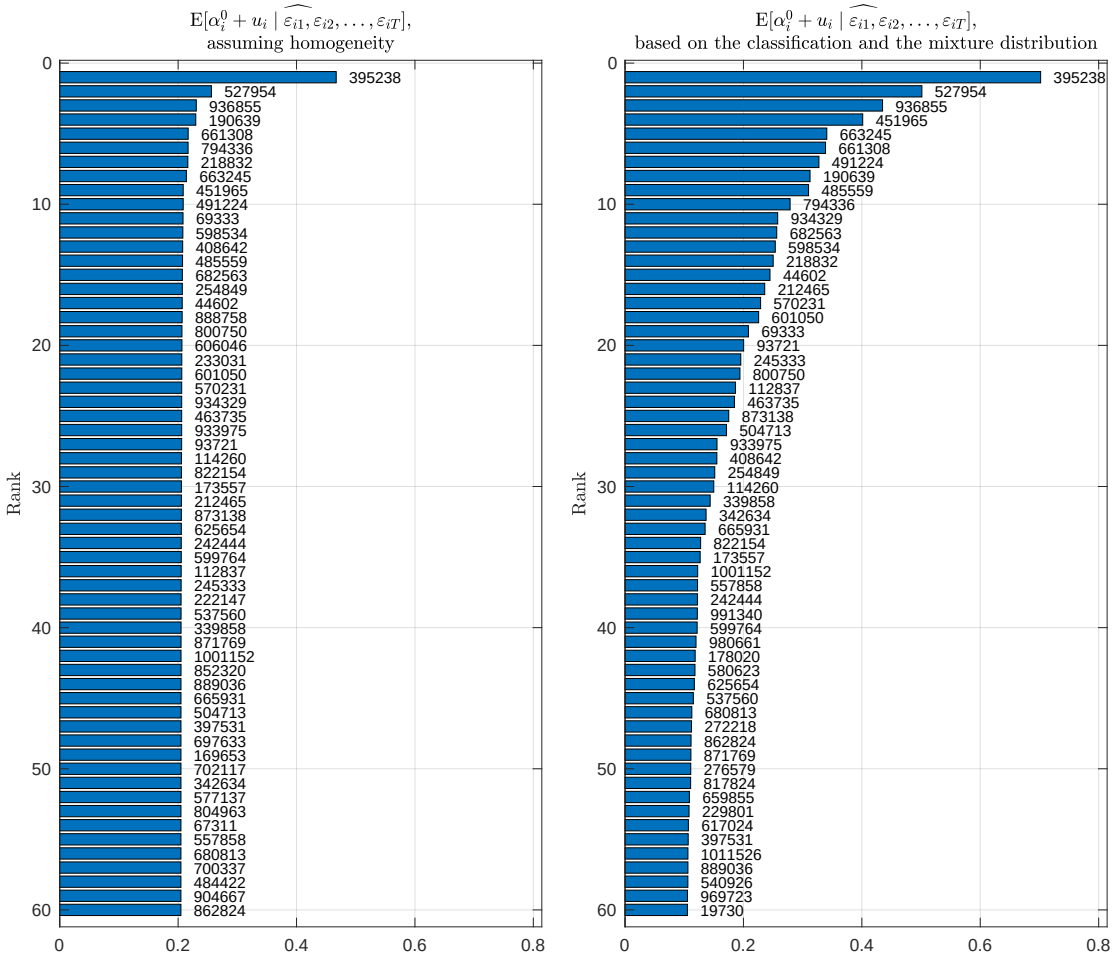
$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = -6.6179.$$

From the mean values, banks in Group 2 appear less efficient than those in Group 1. The standard deviations also indicate that inefficiency varies more within Group 2 than within Group 1.

Who is more often in each group. We compare bank sizes across the two groups and find the following. For group 1, the mean (standard deviation) of $\log(\text{size})$ is 13.4866 (1.8689) based on 113 observations. For group 2, the corresponding mean (standard deviation) is 12.4813 (1.5012) based on 353 observations. The difference in group means is statistically significant at the 5% level, with a t-statistic of 6.9721. Thus, on average, banks in group 1 are larger.

F.5 Additional Figure for the Empirical Application

Figure F.4: Inefficiency Estimates of 60 Banks in Descending Order



Note: Left panel shows the inefficiency estimates in the homogeneous case where all banks are treated as one group with uniquely distributed error term, while the right panel shows the inefficiency estimates in the heterogeneous case where banks are classified into groups and inefficiency estimates to possess a mixture distribution structure. The numbers to the right of each bar corresponds to their respective bank ID, "IDRSSD" from https://cdr.ffiec.gov/public/PWS/DownloadBulkData.aspx.

G On $\mathbb{I}$ in Theorem 3.2

Without loss of generality, we show the properties of $\mathbb{I}$ in the case when $\mathcal{K}^* = 2$ to simplify notation. The cases when $\mathcal{K}^* > 2$ can be similarly shown.

G.1 First and Second Order Derivatives

We use the notation in Appendix C. We assume α_i, β_i, and σ_{vi}^2 are known because they are not the parameters of interest in this appendix.

We first discuss the case when the distribution does not have a latent structure. The unknown parameters are (α^0, σ_u^2) in this case. Note

$$\varepsilon_{it} = y_{it} - \alpha^0 - \alpha(\tau_t) - x_{it}'\beta(\tau_t) = v_{it} - u_i,$$

$$f\left(y_i|\, x_i; \alpha^0, \sigma_u^2\right) = \frac{2}{\sigma_{vi}^{T-1}\sqrt{\sigma_{vi}^2 + T\sigma_u^2}}\left[1 - \Phi\left(\frac{\sum_{t=1}^T \varepsilon_{it}}{\sigma_{vi}\sqrt{\sigma_{vi}^2/\sigma_u^2 + T}}\right)\right]\left[\frac{1}{(2\pi)^{T/2}}\exp\left(-\frac{\sum_{t=1}^T \varepsilon_{it}^2}{2\sigma_{vi}^2}\right)\right]$$
$$\times \exp\left(\frac{1}{2}\frac{\left(\sum_{t=1}^T \varepsilon_{it}\right)^2}{\sigma_{vi}^2\left(\sigma_{vi}^2/\sigma_u^2 + T\right)}\right),$$

and

$$\log f\left(y_i|\, x_i; \alpha^0, \sigma_u^2\right) = C - \frac{(T-1)}{2}\log\sigma_{vi}^2 - \frac{1}{2}\log\left(\sigma_{vi}^2 + T\sigma_u^2\right)$$
$$+ \log\left[1 - \Phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)\right] + \frac{1}{2}\left(\frac{\mu_{i*}}{\sigma_{i*}}\right)^2 - \frac{\sum_{t=1}^T \varepsilon_{it}^2}{2\sigma_{vi}^2},$$

with $\sigma_i^2 = \sigma_{vi}^2 + T\sigma_u^2$, $\rho_i = \sigma_u/\sigma_{vi}$, $\mu_{i*} = -\sigma_u^2\sum_{t=1}^T \varepsilon_{it}/\sigma_i^2$, $\sigma_{i*}^2 = \sigma_u^2\sigma_{vi}^2/\sigma_i^2$, and a C that does not depend on parameters.

By some straightforward calculations,

$$\frac{\partial}{\partial \sigma_u^2}\log f\left(y_i|\, x_i;\alpha^0,\sigma_u^2\right)=-\frac{1}{2}\frac{T}{\sigma_{vi}^2+T\sigma_u^2}+\frac{\phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)}{1-\Phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)}\frac{-\sigma_{vi}}{2\sigma_u^4\left(\sigma_{vi}^2/\sigma_u^2+T\right)^{3/2}}\sum_{t=1}^{T}\varepsilon_{it}$$
$$+\frac{1}{2\sigma_u^4\left(\sigma_{vi}^2/\sigma_u^2+T\right)^2}\left(\sum_{t=1}^{T}\varepsilon_{it}\right)^2, \tag{G.1}$$

$$\frac{\partial}{\partial \alpha^0}\log f\left(y_i|\, x_i;\alpha^0,\sigma_u^2\right)=\frac{\phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)}{1-\Phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)}\frac{T}{\sigma_{vi}\sqrt{\sigma_{vi}^2/\sigma_u^2+T}}+\frac{1}{T^{-1}\sigma_{vi}^2+\sigma_u^2}\frac{1}{T}\left(\sum_{t=1}^{T}\varepsilon_{it}\right), \tag{G.2}$$

$$\frac{\partial^2}{\partial \sigma_u^2\partial \alpha^0}\log f\left(y_i|\, x_i;\alpha^0,\sigma_u^2\right)=g\left(\frac{\mu_{i*}}{\sigma_{i*}}\right)\frac{-T}{2\sigma_u^4\left(\sigma_{vi}^2/\sigma_u^2+T\right)^2}\sum_{t=1}^{T}\varepsilon_{it}+\frac{\phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)}{1-\Phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)}\frac{T\sigma_{vi}}{2\sigma_u^4\left(\sigma_{vi}^2/\sigma_u^2+T\right)^{3/2}}$$
$$-\frac{T}{\sigma_u^4\left(\sigma_{vi}^2/\sigma_u^2+T\right)^2}\left(\sum_{t=1}^{T}\varepsilon_{it}\right),$$

$$\frac{\partial^2}{\partial (\alpha^0)^2}\log f\left(y_i|\, x_i;\alpha^0,\sigma_u^2\right)=g\left(\frac{\mu_{i*}}{\sigma_{i*}}\right)\frac{T^2}{\sigma_{vi}^2\left(\sigma_{vi}^2/\sigma_u^2+T\right)}-\frac{1}{T^{-1}\sigma_{vi}^2+\sigma_u^2},$$

and

$$\frac{\partial^2}{\partial \left(\sigma_u^2\right)^2}\log f\left(y_i|\, x_i;\alpha^0,\sigma_u^2\right)=\frac{1}{2}\frac{T^2}{\left(\sigma_{vi}^2+T\sigma_u^2\right)^2}+g\left(\frac{\mu_{i*}}{\sigma_{i*}}\right)\frac{\sigma_{vi}^2}{4\sigma_u^8\left(\sigma_{vi}^2/\sigma_u^2+T\right)^3}\left(\sum_{t=1}^{T}\varepsilon_{it}\right)^2$$
$$+\frac{\phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)}{1-\Phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)}\frac{\left(\sigma_{vi}^3\sigma_u^{-4/3}+4T\sigma_{vi}\sigma_u^{2/3}\right)}{4\left(\sigma_{vi}^2\sigma_u^{2/3}+T\sigma_u^{8/3}\right)^{5/2}}\sum_{t=1}^{T}\varepsilon_{it}-\frac{T}{\left(\sigma_{vi}^2+T\sigma_u^2\right)^3}\left(\sum_{t=1}^{T}\varepsilon_{it}\right)^2,$$

where

$$g\left(\frac{\mu_{i*}}{\sigma_{i*}}\right)=\left.\frac{d}{ds}\frac{\phi\left(-s\right)}{1-\Phi\left(-s\right)}\right|_{s=\frac{\mu_{i*}}{\sigma_{i*}}}=\left.\frac{-\phi\left(-s\right)-\phi'\left(-s\right)\left[1-\phi\left(-s\right)\right]}{\left[1-\Phi\left(-s\right)\right]^2}\right|_{s=\frac{\mu_{i*}}{\sigma_{i*}}}.$$

In the case when the distribution possesses a latent structure, the log likelihood function becomes

$$\log\left[\tau f\left(y_i\left|x_i;\alpha_{(1)}^0,\sigma_{u(1)}^2\right.\right)+\left(1-\tau\right)f\left(y_i|\, x_i;\alpha_{(2)}^0,\sigma_{u(2)}^2\right)\right]\equiv\log\left[\tau f_{1i}+\left(1-\tau\right)f_{2i}\right].$$

Then,

$$\frac{\partial}{\partial \tau}\log\left[\tau f_{1i}+\left(1-\tau\right)f_{2i}\right]=\frac{f_{1i}-f_{2i}}{\tau f_{1i}+\left(1-\tau\right)f_{2i}},\text{ and}$$

$$\frac{\partial^2}{\partial\tau^2}\log\left[\tau f_{1i}+(1-\tau)f_{2i}\right]=-\frac{\left(f_{1i}-f_{2i}\right)^2}{\left[\tau f_{1i}+(1-\tau)f_{2i}\right]^2}.$$

The derivatives with respect to other arguments α^0 and σ_u^2 can be derived using the chain rule. For example,

$$\frac{\partial}{\partial\alpha^0_{(1)}}\log\left[\tau f_{1i}+(1-\tau)f_{2i}\right]=\frac{\tau\frac{\partial}{\partial\alpha^0_{(1)}}f_{1i}}{\tau f_{1i}+(1-\tau)f_{2i}}=\frac{\tau f_{1i}}{\tau f_{1i}+(1-\tau)f_{2i}}\cdot\frac{\partial}{\partial\alpha^0_{(1)}}\log f_{1i},\text{ and}$$

$$\frac{\partial}{\partial\sigma^2_{u(1)}}\log\left[\tau f_{1i}+(1-\tau)f_{2i}\right]=\frac{\tau\frac{\partial}{\partial\sigma^2_{u(1)}}f_{1i}}{\tau f_{1i}+(1-\tau)f_{2i}}=\frac{\tau f_{1i}}{\tau f_{1i}+(1-\tau)f_{2i}}\cdot\frac{\partial}{\partial\sigma^2_{u(1)}}\log f_{1i}.$$

G.2 Properties of $\mathbb{I}$

We verify that $\mathbb{I}$ behaves like a regular positive definite matrix in this section. The intuition of this result is as follows. Note $\tilde{f}_i(\varrho_0)$ is the density function for $\varepsilon_{it}=v_{it}-u_i$, and the parameters of interest are on the distribution of u_i only. In other words, v_{it} are nuisance for the estimation. Mathematically, one can see that the dominant components in $\frac{\partial}{\partial\varrho}\log\tilde{f}_i(\varrho)$ are functions of u_i and do not grow as $T\to\infty$. Thus, the convergence rate of $\hat{\varrho}$ is $\sqrt{N}$ based on this intuition.

Without loss of generality, we assume that there is no group structure for the frontiers and σ^2_{vi}, that is, $K^*=1$. We show the result by assuming $\alpha^0=\alpha^0_{(1)}=\alpha^0_{(2)}$, and $\sigma^2_{u(1)}\neq\sigma^2_{u(2)}$. The case with $\alpha^0_{(1)}\neq\alpha^0_{(2)}$ can be similarly handled but with more tedious discussions. We show the result by using the following identity

$$\begin{aligned}\mathbb{I}&=-\mathrm{E}\left[\frac{\partial^2}{\partial\varrho\partial\varrho'}\log\tilde{f}_i(\varrho)\right]\Bigg|_{\varrho=\varrho^0}\\&=\mathrm{E}\left[\frac{\partial}{\partial\varrho}\log\tilde{f}_i(\varrho)\cdot\frac{\partial}{\partial\varrho'}\log\tilde{f}_i(\varrho)\right]\Bigg|_{\varrho=\varrho^0}.\end{aligned}\tag{G.3}$$

From the forms of $\frac{\partial}{\partial\sigma^2_{uj}}\log\left[\tau f_1+(1-\tau)f_2\right]$, $\frac{\partial}{\partial\alpha^0_j}\log\left[\tau f_1+(1-\tau)f_2\right]$, $j=1,2$, and

$$\frac{\partial}{\partial\tau}\log\left[\tau f_1+(1-\tau)f_2\right]$$

in Appendix G.1, they are clearly not linear dependent. If T is fixed, $\mathbb{I}$ is just the information matrix for a regular likelihood function, and it is positive definite. Our analysis is complicated by

the fact that $T \to \infty$. The diagonal of $\mathbb{I}$ might explode as $T \to \infty$. Thus, it suffices to show that the diagonals of $\mathbb{I}$ behave like regular positive constants, and we show that in the below.

The following is useful for the analysis. At the true values of parameters, $\varepsilon_{it} = v_{it} - u_i$. So

$$\begin{aligned} T^{-1} \sum_{t=1}^{T} \varepsilon_{it} &= -u_i + T^{-1} \sum_{t=1}^{T} v_{it} \\ &= -u_i + O_P\left(T^{-1/2}\right). \end{aligned}$$

Since $-u_i$ is negative, $T^{-1} \sum_{t=1}^{T} \varepsilon_{it}$ is negative with very high probability. Another implication is $\sum_{t=1}^{T} \varepsilon_{it} \propto_P -T + O_P\left(T^{1/2}\right)$. Using the definition of $\frac{\mu_{i*}}{\sigma_{i*}}$,

$$\frac{\mu_{i*}}{\sigma_{i*}} = \frac{-1}{\sigma_{vi}\sqrt{\sigma_{vi}^2/\sigma_u^2 + T}} \sum_{t=1}^{T} \varepsilon_{it} \propto_P \sqrt{T} + O_P(1).$$

The above implies $\Phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right) = o_P(1)$,

$$\frac{\phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)}{1 - \Phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)} \propto_P \exp(-T), \quad \frac{\phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)}{1 - \Phi\left(-\frac{\mu_{i*}}{\sigma_{i*}}\right)} \sqrt{T} = o_P(1),$$

and

$$g\left(\frac{\mu_{i*}}{\sigma_{i*}}\right) \propto_P \exp(-T).$$

Applying the above for (G.2),

$$\frac{\partial}{\partial \alpha^0} \log f\left(y_i | x_i; \alpha^0, \sigma_u^2\right) = -u_i + E(u_i) + o_P(1).$$

Similarly (G.1) implies

$$\frac{\partial}{\partial \sigma_u^2} \log f\left(y_i | x_i; \alpha^0, \sigma_u^2\right) = \frac{u_i^2}{2\sigma_u^4} - \frac{1}{2\sigma_u^2} + o_P(1).$$

Some tedious yet straightforward calculations can yield

$$f\left(y_i \middle| x_i; \alpha^0, \sigma_{u(1)}^2\right) \propto_P f\left(y_i | x_i; \alpha^0, \sigma_{u(2)}^2\right),$$

and

$$\frac{\partial}{\partial\tau}\log\left[\tau f_{1i}+(1-\tau)f_{2i}\right]=\frac{f_{1i}-f_{2i}}{\tau f_{1i}+(1-\tau)f_{2i}}.$$
$$=\frac{1-\sqrt{\sigma^2_{u(1)}/\sigma^2_{u(2)}}\exp\left(\frac{1}{2}\left(\frac{1}{\sigma^2_{u(1)}}-\frac{1}{\sigma^2_{u(2)}}\right)u_i^2\right)}{\tau+(1-\tau)\sqrt{\sigma^2_{u(1)}/\sigma^2_{2u}}\exp\left(\frac{1}{2}\left(\frac{1}{\sigma^2_{u(1)}}-\frac{1}{\sigma^2_{u(2)}}\right)u_i^2\right)}\left(1+o_P\left(1\right)\right).$$

The variance of the leading terms in $\frac{\partial}{\partial\alpha^0}\log f$, $\frac{\partial}{\partial\sigma^2_u}\log f$, and $\frac{\partial}{\partial\tau}\log\left[\tau f_{1i}+(1-\tau)f_{2i}\right]$ are functions of u_i, and they are bounded and bounded away from zero.

Finally, since $f_{1i}\propto_P f_{2i}$, $\frac{\partial}{\partial\alpha^0_j}\log\left[\tau f_{1i}+(1-\tau)f_{2i}\right]$, $j=1,2$, enjoys similar properties as $\frac{\partial}{\partial\alpha^0}\log f$. This is also the case for $\frac{\partial}{\partial\sigma^2_{uj}}\log\left[\tau f_{1i}+(1-\tau)f_{2i}\right]$, $j=1,2$.

From the form of the derivatives, clearly, $\frac{\partial}{\partial\varrho_j}\log\left[\tau f_{1i}+(1-\tau)f_{2i}\right]$, $j=1,2,...,5$, are not linearly dependent. Together with the fact that the variance of them are bounded and bounded away from zero, then for any 5×1 vector a,

$$a'\,\mathrm{E}\left[\frac{\partial}{\partial\varrho}\log\tilde{f}_i\left(\varrho\right)\cdot\frac{\partial}{\partial\varrho'}\log\tilde{f}_i\left(\varrho\right)\right]\Bigg|_{\varrho=\varrho^0}a\propto\left\|a\right\|^2.$$

This implies that $\mathbb{I}$ must be positive definite with bounded eigenvalues, using the identity in (G.3).

H Technical Lemmas

We collect technical lemmas and their proofs in this section. For an easier reference, we present the bodies of those lemmas first, followed by proofs. We note the lemmas of probability bounds below are related to those developed in Atak et al. (2025).

Remark H.1 (Some inequalities)**.** We may apply the following inequalities in the proofs directly without referring them back to here. First

$$\Pr\left(X_1+X_2\geq C\right)\leq\Pr\left(X_1\geq\pi C\right)+\Pr\left(X_2\geq(1-\pi)C\right)$$

for any constant π. That is because $\{X_1 + X_2 \geq C\} \subseteq \{X_1 \geq \pi C\} \cup \{X_1 \geq (1-\pi) C\}$. Similarly, we have

$$\Pr\left(\sum_{i=1}^{n} X_i \geq C\right) \leq \sum_{i=1}^{n} \Pr\left(X_i \geq Cn^{-1}\right) \text{ and } \Pr\left(\max_{1\leq i\leq n} X_i \geq C\right) \leq \sum_{i=1}^{n} \Pr\left(X_i \geq C\right).$$

For any positive random variables X_1 and X_2 and positive constants C_1 and C_2,

$$\Pr\left(X_1 \cdot X_2 \geq C_1\right) \leq \Pr\left(X_1 \geq C_1/C_2\right) + \Pr\left(X_2 \geq C_2\right),$$

due to the fact that $\{X_1 \cdot X_2 \geq C_1\} \subseteq \{X_1 \geq C_1/C_2\} \cup \{X_2 \geq C_2\}$.

Lemma H.1. *Suppose e_{it} satisfies the mixing condition across t in Assumption 1 (ii) and is identically distributed across i. In addition $\max_t \mathrm{E}\left(|e_{it}|^{C_q}\right) < \infty, C_q > 2, T = N^C$ for some $C > 0$, $m \to \infty$ as $N \to \infty$, then*

(i) for any $\epsilon > 0$ and any positive $1 \leq \upsilon_{NT} \ll T^{1/2} (\log N)^{-2}$

$$\Pr\left(\max_{i=1,\ldots,N} \left|\frac{1}{T}\sum_{t=1}^{T} \left[e_{it} - \mathrm{E}\left(e_{it}\right)\right]\right| > \frac{\epsilon}{\upsilon_{NT}}\right) \lesssim \frac{NT\upsilon_{NT}^{C_q} (\log N)^{4C_q}}{T^{C_q}};$$

(ii) there exists a positive M, such that

$$\Pr\left(\max_{i=1,\ldots,N} \max_{l=0,1,\ldots,p} \left|\frac{1}{T}\sum_{t=1}^{T} e_{it} b_{il}\left(\tau_t\right)\right| > Mm^{-\kappa}\right) \lesssim \frac{NT (\log N)^{4C_q}}{T^{C_q}},$$

where $b_{il}(\tau_t)$ is defined in (E.1).

Lemma H.2. *Suppose that Assumption 3 holds and $m/T \to 0$. For any $\tilde{\pi} \in \mathbb{R}^{m(p+1)}$,*

$$\frac{C_{xx}}{2}\|\tilde{\pi}\| \leq \min_{1\leq i\leq N} \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\left(\tilde{\pi}' \tilde{z}_{it} \tilde{z}_{it}' \tilde{\pi}\right) \leq \max_{1\leq i\leq N} \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\left(\pi' \tilde{z}_{it} \tilde{z}_{it}' \pi\right) \leq 2\bar{C}_{xx}\|\tilde{\pi}\|,$$

after some large T.

Lemma H.3. *Suppose that Assumptions 1, 2, 3 and 6 (i) hold. Then*

(i) for any $\epsilon > 0$,

$$\Pr\left(\max_{1\leq i\leq N} \sup_{\tilde{\pi}\in\mathbb{R}^{m(p+1)}, \|\tilde{\pi}\|=1} \left|\frac{1}{T}\sum_{t=1}^{T} \tilde{\pi}' \tilde{z}_{it} \tilde{z}_{it}' \tilde{\pi} - \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\left(\tilde{\pi}' \tilde{z}_{it} \tilde{z}_{it}' \tilde{\pi}\right)\right| > \epsilon\right) \lesssim \frac{Nm^{q/2+2} (\log N)^{2q}}{T^{q/2-1}};$$

(ii) there exist some positive and finite $\underline{C}_{zz}$ *and* $\bar{C}_{zz}$ *such that*

$$\begin{aligned}&\Pr\left(\underline{C}_{zz} \leq \min_{1\leq i\leq N} \mu_{\min}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\tilde{z}_{it}'\right) \leq \max_{1\leq i\leq N} \mu_{\max}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\tilde{z}_{it}'\right) \leq \bar{C}_{zz}\right)\\ =&1-o\left(\frac{Nm^{q/2+2}\left(\log N\right)^{2q}}{T^{q/2-1}}\right).\end{aligned}$$

Lemma H.4. *Suppose Assumptions 1, 2, 4 and 6 (i) hold. Then*

(i) for any $\epsilon > 0$,

$$\Pr\left(\max_{i=1,\ldots,N}\left\|\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}v_{it}\right\| > \epsilon\right) \lesssim \frac{Nm^{q/4}\left(\log N\right)^{2q}}{T^{q/2-1}};$$

(ii) for the ξ_{it} *defined in (E.2),*

$$\Pr\left(\max_{i=1,\ldots,N}\left\|\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\xi_{it}\right\| > \epsilon\right) \lesssim \frac{Nm\left(\log N\right)^{2q}}{T^{q/2-1}}.$$

Lemma H.5. *Suppose that Assumptions 1, 2, 3, 4, and 6 hold. Then,*

(i) for any small positive ϵ,

$$\Pr\left(\max_{1\leq i\leq N}\left\|\hat{\tilde{\pi}}_i - \tilde{\pi}_i^0\right\| > \epsilon\right) = o\left(1\right);$$

(ii) for any small positive ϵ,

$$\Pr\left(\max_{1\leq i\leq N}\left\|\hat{\sigma}_{vi}^2 - \sigma_{vi}^2\right\| > \epsilon\right) = o\left(1\right).$$

Lemma H.6. *Suppose Assumption 3 holds and* $\underline{m}/T \to 0$. *Then after some large* T,

$$\frac{\underline{C}_{xx}}{2} \leq \min_{1\leq i\leq N} \mu_{\min}\left\{\mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\ddot{\underline{z}}_{it}\ddot{\underline{z}}_{it}'\right)\right\} \leq \max_{1\leq i\leq N} \mu_{\max}\left\{\mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\ddot{\underline{z}}_{it}\ddot{\underline{z}}_{it}'\right)\right\} \leq 2\bar{C}_{xx},$$

where $\underline{C}_{xx}$ *and* $\bar{C}_{xx}$ *are defined in Lemma H.2.*

Lemma H.7. *Suppose that Assumptions 1, 2, 3, 7 and 8 hold. Then for a fixed* k, $k = 1, 2, ..., K^*$, *and any* $\epsilon > 0$,

$$\Pr\left(\sup_{\pi\in\mathbb{R}^{\underline{m}-1+\underline{m}p},\|\pi\|=1}\left|\frac{1}{N_k}\sum_{i\in G_{k|K^*}}\left[\frac{1}{T}\sum_{t=1}^{T}\pi'\ddot{\underline{z}}_{it}\ddot{\underline{z}}_{it}'\pi - \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left(\pi'\ddot{\underline{z}}_{it}\ddot{\underline{z}}_{it}'\pi\right)\right]\right| > \epsilon\right) \lesssim \frac{\underline{m}^{q/2+2}\left(\log N\right)^{2q}}{\left(NT\right)^{q/2-1}};$$

Lemma H.8. *Suppose Assumptions 1, 2, 3, 4, 6, 7, 8, and 10 hold. Then, the term*

$$A_{k1} = \left(\sum_{i \in G_{k|K}} \sum_{t=1}^{T} \ddot{z}_{it} \ddot{z}_{it}' \right)^{-1} \left(\sum_{i \in G_{k|K}} \sum_{t=1}^{T} \ddot{z}_{it} \ddot{\xi}_{it} \right)$$

satisfies $\|A_{k1}\| = O_P\left(\underline{m}^{-\kappa}\right)$.

Lemma H.9. *Suppose Assumptions 1, 2, 3, 4, 6, 7, 8, 9 , and 10 hold. For*

$$A_{k2} = \left(\sum_{i \in G_{k|K}} \sum_{t=1}^{T} \ddot{z}_{it} \ddot{z}_{it}' \right)^{-1} \left(\sum_{i \in G_{k|K}} \sum_{t=1}^{T} \ddot{z}_{it} \ddot{v}_{it} \right)$$

and any $(p+1) \times 1$ *vector* a*, it satisfies*

$$\sqrt{\frac{N_k T}{\underline{m}}} a' \mathbb{M}_{\mathbb{B}}(s) A_{k2} \Bigg/ \left\{ a' \left[\frac{\sigma^2_{v(k)}}{\underline{m}} \mathbb{M}_{\mathbb{B}}(s) Q^{-1}_{(k),zz} \mathbb{M}_{\mathbb{B}}(s)' \right] a \right\}^{1/2} \xrightarrow{d} N(0,1).$$

Lemma H.10. *Suppose Assumptions 1 - 8 hold. In addition,* $\frac{\lambda_{NT}}{\sqrt{NT}} \to \infty$ *and* $\frac{\lambda_{NT}}{NT} \to 0$.

$$\Pr\left(IC(K^*, \lambda_{NT}) < IC(K, \lambda_{NT}), 1 \leq K \leq K^* - 1\right) \to 1.$$

Lemma H.11. *Suppose Assumptions 1 - 8 hold. In addition,* $\frac{\lambda_{NT}}{\sqrt{NT}} \to \infty$ *and* $\frac{\lambda_{NT}}{NT} \to 0$.

$$\Pr\left(IC(K^*, \lambda_{NT}) < IC(K, \lambda_{NT}), K^* + 1 \leq K \leq \bar{K}\right) \to 1.$$

Lemma H.12. *Suppose Assumptions 1 - 10 hold.* $\tilde{\lambda}_{NT}$ *satisfies* $\tilde{\lambda}_{NT} \to \infty$ *and* $\frac{\tilde{\lambda}_{NT}}{N} \to 0$.

$$\Pr\left(\hat{\mathcal{K}}\left(\tilde{\lambda}_{NT}\right) = \mathcal{K}^*\right) \to 1.$$

Proof of Lemma H.1. **(i)** Set $C_{NT} = v_{NT}^{-1} T (\log N)^{-4}$. Define $\mathbf{1}_{it} = \mathbf{1}(|e_{it}| < C_{NT})$ and $\bar{\mathbf{1}}_{it} = 1 - \mathbf{1}_{it}$. With it, further define $e_{it}^{(1)} = e_{it}\mathbf{1}_{it} - \mathrm{E}(e_{it}\mathbf{1}_{it})$, $e_{it}^{(2)} = e_{it}\bar{\mathbf{1}}_{it}$, and $e_{it}^{(3)} = \mathrm{E}\left(e_{it}\bar{\mathbf{1}}_{it}\right)$. Then

$$e_{it} - \mathrm{E}(e_{it}) = e_{it}^{(1)} + e_{it}^{(2)} + e_{it}^{(3)}.$$

Let $v_0 = \max_t \mathrm{Var}(e_{it}) + 2\sum_{s=t+1}^{T} \mathrm{Cov}(e_{it}, e_{is})$, which is finite by the mixing condition in Assumption 1.

Using Bernstein inequality for strong mixing processes (e.g., Theorem 2 in Merlevede et al. (2009)), there exists positive constants C_1 and C_2 such that

$$\begin{aligned}\Pr\left(\max_{i=1,\ldots,N}\left|\frac{1}{T}\sum_{t=1}^{T}e_{it}^{(1)}\right|>\frac{\epsilon}{3\upsilon_{NT}}\right)&\leq\sum_{i=1}^{N}\Pr\left(\left|\sum_{t=1}^{T}e_{it}^{(1)}\right|>\frac{T\epsilon}{3\upsilon_{NT}}\right)\\&\leq N\exp\left\{-\frac{C_1T^2\epsilon^2}{9T\upsilon_0\upsilon_{NT}^2+9C_{NT}^2\upsilon_{NT}^2+3T\epsilon\upsilon_{NT}C_{NT}\left(\log T\right)^2}\right\}\\&=N\exp\left\{-C_2\epsilon^2\left(\log N\right)^2\right\}=o\left(N^{-M}\right),\end{aligned}$$

for any large $M>0$, where the last line holds by the fact that $3T\epsilon\upsilon_{NT}C_{NT}\left(\log T\right)^2=3T^2\epsilon\left(\log N\right)^{-4}\left(\log T\right)^2$ is the dominant term in the denominator due to $\upsilon_{NT}\ll T^{1/2}\left(\log N\right)^{-2}$, and $\log T\propto\log N$.

We turn to $e_{it}^{(2)}$. Markov inequality implies that

$$\begin{aligned}\Pr\left(\max_{i=1,\ldots,N}\left|\frac{1}{T}\sum_{t=1}^{T}e_{it}^{(2)}\right|>\frac{\epsilon}{3}\right)&\leq\Pr\left(\max_{i=1,\ldots,N,t=1,\ldots,T}\left|e_{it}^{(2)}\right|>\frac{\epsilon}{3}\right)\\&\leq\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{C_{NT}^{C_q}}\mathrm{E}\left[\left|e_{it}\right|^{C_q}\mathbf{1}_{it}=\mathbf{1}(\left|e_{it}\right|\geq C_{NT})\right]\\&\lesssim\frac{NT\upsilon_{NT}^{C_q}\left(\log N\right)^{4C_q}}{T^{C_q}}.\end{aligned}$$

The last term can be bounded as

$$\frac{1}{T}\sum_{t=1}^{T}e_{it}^{(3)}\leq\max_{t=1,\ldots,T}\mathrm{E}\left(\left|e_{it}\right|\bar{\mathbf{1}}_{it}\right)=o\left(1\right)<\frac{\epsilon}{3},$$

uniformly for all i after some large T, where $o\left(1\right)$ holds because of the finite moment conditions on e_{it} over t, Markov inequality and the identical distribution across i. The above implies that

$$\Pr\left(\max_{i=1,\ldots,N}\left|\frac{1}{T}\sum_{t=1}^{T}e_{it}^{(3)}\right|>\frac{\epsilon}{3}\right)=0,$$

after some large T. Together, we have

$$\begin{aligned}&\Pr\left(\max_{i=1,\ldots,N}\left|\frac{1}{T}\sum_{t=1}^{T}\left[e_{it}-\mathrm{E}\left(e_{it}\right)\right]\right|>\epsilon\right)\\\leq&\Pr\left(\max_{i=1,\ldots,N}\left|\frac{1}{T}\sum_{t=1}^{T}e_{it}^{(1)}\right|>\frac{\epsilon}{3}\right)+\Pr\left(\max_{i=1,\ldots,N}\left|\frac{1}{T}\sum_{t=1}^{T}e_{it}^{(2)}\right|>\frac{\epsilon}{3}\right)+\Pr\left(\max_{i=1,\ldots,N}\left|\frac{1}{T}\sum_{t=1}^{T}e_{it}^{(3)}\right|>\frac{\epsilon}{3}\right)\\\lesssim&\frac{NT\upsilon_{NT}^{C_q}\left(\log N\right)^{4C_q}}{T^{C_q}}.\end{aligned}$$

(ii) Recall (E.4), we can find a positive $\bar{C}_b$ such that

$$\max_{l=0,...,p} \sup_{s\in[0,1]} |b_{il}(s)| \leq \bar{C}_b m^{-\kappa},$$

because p is finite. As a result,

$$\max_{i=1,...,N} \max_{l=0,1,...,p} \left| \frac{1}{T}\sum_{t=1}^{T} e_{it} b_{il}(\tau_t) \right| \leq \bar{C}_b m^{-\kappa} \max_{i=1,...,N} \frac{1}{T}\sum_{t=1}^{T} |e_{it}| .$$

Applying (i) by setting $\upsilon_{NT} = 1$ for series $|e_{it}|$ implies

$$\Pr\left(\max_{i=1,...,N} \left| \frac{1}{T}\sum_{t=1}^{T} |e_{it}| - \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}|e_{it}| \right| > \epsilon \right) \lesssim \frac{NT(\log N)^{4C_q}}{T^{C_q}}.$$

Since $\max_{i=1,...,N} \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}|e_{it}|$ is bounded and ϵ can be any arbitrary small positive number, we can set

$$M = 2\bar{C}_b \cdot \max_{i=1,...,N} \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}|e_{it}| ,$$

so that the desired result holds. □

Proof of Lemma H.2. Let $g_l(\tau_t) = \pi_l' \mathbb{B}^m(\tau_t)$, where π_l is $m \times 1$ and denotes the corresponding elements in $\tilde{\pi}$ for $x_{it,l} \otimes \mathbb{B}^m(\tau_t)$, $l = 0, 1, ..., p$.[6] Let $\mathbf{g}(\tau_t) \equiv (g_0(\tau_t), ..., g_p(\tau_t))'$. In other words, $g_l(\tau_t)$ is the approximation for $\beta_l(\tau_t)$ when π_l=π_l^0. Then $\frac{1}{T}\sum_{t=1}^{T} \mathrm{E}(\tilde{\pi}' \tilde{z}_{it} \tilde{z}_{it}' \tilde{\pi})$ can be written as

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}(\tilde{\pi}' \tilde{z}_{it} \tilde{z}_{it}' \tilde{\pi}) &= \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\left[\mathbf{g}(\tau_t)' \tilde{x}_{it} \tilde{x}_{it}' \mathbf{g}(\tau_t) \right] \\ &= \frac{1}{T}\sum_{t=1}^{T} \mathbf{g}(\tau_t)' \mathrm{E}(\tilde{x}_{it} \tilde{x}_{it}') \mathbf{g}(\tau_t) , \end{aligned}$$

By the rank condition in Assumption 3, the above implies

$$\begin{aligned} \underline{C}_{xx} \frac{1}{T}\sum_{t=1}^{T} \mathbf{g}(\tau_t)' \mathbf{g}(\tau_t) &\leq \min_{1\leq i\leq N} \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}(\tilde{\pi}' \tilde{z}_{it} \tilde{z}_{it}' \tilde{\pi}) \\ &\leq \max_{1\leq i\leq N} \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}(\tilde{\pi}' \tilde{z}_{it} \tilde{z}_{it}' \tilde{\pi}) \leq \bar{C}_{xx} \frac{1}{T}\sum_{t=1}^{T} \mathbf{g}(\tau_t)' \mathbf{g}(\tau_t) . \end{aligned} \tag{H.1}$$

[6] We abuse the notation a bit by letting $x_{it,0}$ denote 1.

Lemma A.4 in Dong and Linton (2018) implies that, $T^{-1}\sum_{t=1}^{T}\mathbb{B}^m(\tau_t)\mathbb{B}^m(\tau_t)' = I_m + O(m/T)$. Using it

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}g_l(\tau_t)^2 &= \frac{1}{T}\sum_{t=1}^{T}\tilde{\pi}_l'\mathbb{B}^m(\tau_t)\mathbb{B}^m(\tau_t)'\tilde{\pi}_l \\ &= \tilde{\pi}_l'\left[I_m + O(m/T)\right]\tilde{\pi}_l = \|\tilde{\pi}_l\|\left[1+o(1)\right],\end{aligned}$$

due to $m/T \to 0$. Substituting the above result into (H.1) yields the desired result by relaxing the lower and upper bounds to $\underline{C}_{xx}/2\,\|\tilde{\pi}_l\|$ and $2\bar{C}_{xx}\,\|\tilde{\pi}_l\|$, respectively. □

Proof of Lemma H.3. **(i)** is standard in the literature. Lemma 5 in Fan et al. (2011) and Lemma S1.4 in Su et al. (2024) show a similar result for variables with sub-exponential tails. We prove this lemma for $\mathrm{E}(|x_{it}|^q) < \infty$. Note that $\tilde{z}_{it} = \left[\mathbb{B}^m(\tau_t)', (x_{it}\otimes\mathbb{B}^m(\tau_t))'\right]$, a $[(p+1)m]\times 1$ vector, then

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\tilde{\pi}'\tilde{z}_{it}\tilde{z}_{it}'\tilde{\pi} &= \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{j=0}^{p}\sum_{l=0}^{p}\sum_{s=0}^{m-1}\sum_{s'=0}^{m-1}\pi_{js}\pi_{ls'}B_s(\tau_t)B_{s'}(\tau_t)x_{itj}x_{itl}\right] \\ &= \sum_{j=0}^{p}\sum_{l=0}^{p}\sum_{s=0}^{m-1}\sum_{s'=0}^{m-1}\left[\frac{1}{T}\sum_{t=1}^{T}\pi_{js}\pi_{ls'}B_s(\tau_t)B_{s'}(\tau_t)x_{itj}x_{itl}\right]\end{aligned}$$

where we let x_{it0} denote the constant 1. Using the above,

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\tilde{\pi}'\tilde{z}_{it}\tilde{z}_{it}'\tilde{\pi} - \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left(\tilde{\pi}'\tilde{z}_{it}\tilde{z}_{it}'\tilde{\pi}\right) \\ =&\sum_{j=0}^{p}\sum_{l=0}^{p}\sum_{s=0}^{m-1}\sum_{s'=0}^{m-1}\pi_{js}\pi_{ls'}\left\{\frac{1}{T}\sum_{t=1}^{T}B_s(\tau_t)B_{s'}(\tau_t)\left[x_{itj}x_{itl}-\mathrm{E}(x_{itj}x_{itl})\right]\right\} \\ \equiv&\sum_{j=0}^{p}\sum_{l=0}^{p}\sum_{s=0}^{m-1}\sum_{s'=0}^{m-1}\pi_{js}\pi_{ls'}I_{iss'},\end{aligned}$$

with

$$I_{iss'} \equiv \frac{1}{T}\sum_{t=1}^{T}B_s(\tau_t)B_{s'}(\tau_t)\left[x_{itj}x_{itl}-\mathrm{E}(x_{itj}x_{itl})\right].$$

We first assume the event $\left\{\max_{i=1,\ldots,N}\max_{s,s'=1,\ldots,m}|I_{iss'}| \le \frac{\epsilon}{mp^2}\right\}$ holds; its probability bound

will be shown later. Conditional on this event, for all $\|\tilde{\pi}\| = 1$,

$$\begin{aligned}
\max_{i=1,\ldots,N} \left| \sum_{j=0}^{p} \sum_{l=0}^{p} \sum_{s=0}^{m-1} \sum_{s'=0}^{m-1} \pi_{js} \pi_{ls'} I_{iss'} \right| &\leq \frac{\epsilon}{mp^2} \sum_{j=0}^{p} \sum_{l=0}^{p} \sum_{s=0}^{m-1} \sum_{s'=0}^{m-1} \left| \pi_{js} \pi_{ls'} \right| . \\
&\leq \frac{\epsilon}{mp^2} \sum_{j=0}^{p} \sum_{l=0}^{p} \sum_{s=0}^{m-1} \left[|\pi_{js}| \sum_{s'=0}^{m-1} |\pi_{ls'}| \right] \\
&\leq \frac{\epsilon}{mp^2} \sum_{j=0}^{p} \sum_{l=0}^{p} \sum_{s=0}^{m-1} |\pi_{js}| \left(\sum_{s'=0}^{m-1} \pi_{ls'}^2 \right)^{1/2} \left(\sum_{s'=0}^{m-1} 1^2 \right)^{1/2} \\
&= \frac{\epsilon}{\sqrt{m} p^2} \sum_{j=0}^{p} \sum_{l=0}^{p} \sum_{s=0}^{m-1} |\pi_{js}| \leq \epsilon,
\end{aligned}$$

where the third and fourth lines hold by Cauchy-Schwarz inequality. Thus

$$\begin{aligned}
&\Pr \left(\max_{i=1,\ldots,N} \sup_{\tilde{\pi} \in \mathbb{R}^{m(p+1)}, \|\tilde{\pi}\|=1} \left| \sum_{j=0}^{p} \sum_{l=0}^{p} \sum_{s=0}^{m-1} \sum_{s'=0}^{m-1} \pi_{js} \pi_{ls'} I_{iss'} \right| \leq \epsilon \right) \\
\geq &\Pr \left(\max_{i=1,\ldots,N} \max_{s,s'=1,\ldots,m} |I_{iss'}| \leq \frac{\epsilon}{mp^2} \right)
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\Pr \left(\max_{i=1,\ldots,N} \sup_{\tilde{\pi} \in \mathbb{R}^{m(p+1)}, \|\tilde{\pi}\|=1} \left| \sum_{j=0}^{p} \sum_{l=0}^{p} \sum_{s=0}^{m-1} \sum_{s'=0}^{m-1} \pi_{js} \pi_{ls'} I_{iss'} \right| > \epsilon \right) \\
< &\Pr \left(\max_{i=1,\ldots,N} \max_{s,s'=1,\ldots,m} |I_{iss'}| > \frac{\epsilon}{mp^2} \right) . \qquad \text{(H.2)}
\end{aligned}$$

Using the result in Lemma H.1 (i) (by setting $\upsilon_{NT} = m$), the moment condition on x (C_q in Lemma H.1 is $q/2$ in this case), and the fact that p is fixed, not hard to see that

$$\begin{aligned}
\Pr \left(\max_{i=1,\ldots,N} \max_{s,s'=1,\ldots,m} |I_{iss'}| > \frac{\epsilon}{mp^2} \right) &\leq \sum_{s,s'=1}^{m} \Pr \left(\max_{i=1,\ldots,N} |I_{iss'}| > \frac{\epsilon}{p^2} \right) \\
&\lesssim \sum_{s,s'=1}^{m} \frac{NTm^{q/2} (\log N)^{2q}}{T^{q/2}} \\
&\lesssim \frac{NTm^{q/2+2} (\log N)^{2q}}{T^{q/2}}.
\end{aligned}$$

Substitute it into (H.2), we obtain the desired result.

(ii) is an immediate result from Lemmas H.2 and the result (i).

Suppose the complement of the event in (i) holds with a small $\epsilon > 0$. For any $\tilde{\pi} \in \mathbb{R}^{m(p+1)}$ and $\|\tilde{\pi}\| = 1$,

$$\max_{1\leq i\leq N} \tilde{\pi}' \left(\frac{1}{T} \sum_{t=1}^{T} \tilde{z}_{it} \tilde{z}_{it}' \right) \tilde{\pi} \leq \max_{1\leq i\leq N} (1+\epsilon) \frac{1}{T} \sum_{t=1}^{T} \mathrm{E} \left(\tilde{\pi}' \tilde{z}_{it} \tilde{z}_{it}' \tilde{\pi} \right) \leq 2\bar{C}_{xx} (1+\epsilon).$$

Similarly,

$$\min_{1\leq i\leq N} \tilde{\pi}' \left(\frac{1}{T} \sum_{t=1}^{T} \tilde{z}_{it} \tilde{z}_{it}' \right) \tilde{\pi} \geq \min_{1\leq i\leq N} (1-\epsilon) \frac{1}{T} \sum_{t=1}^{T} \mathrm{E} \left(\tilde{\pi}' \tilde{z}_{it} \tilde{z}_{it}' \tilde{\pi} \right) \geq \frac{1}{2} \bar{C}_{xx} (1-\epsilon).$$

We show the result by setting $\underline{C}_{zz} = 1/2\bar{C}_{xx} (1-\epsilon)$ and $2\bar{C}_{xx} (1+\epsilon)$. □

Proof of Lemma H.4. **(i)** $\tilde{z}_{it} v_{it}$ is $(p+1)m \times 1$. Elements in $\frac{1}{T} \sum_{t=1}^{T} \tilde{z}_{it} v_{it}$ are

$$\frac{1}{T} \sum_{t=1}^{T} x_{itj} B_l (\tau_t) v_{it}, \text{ for } j = 0, ..., p \text{ and } l = 0, ..., m-1.$$

The idea is to obtain the probability bound of

$$\max_{i=1,...,N} \max_{j=0,...,p} \max_{l=0,...,m-1} \left| \frac{1}{T} \sum_{t=1}^{T} x_{itj} B_l (\tau_t) v_{it} \right| > \epsilon / \sqrt{(p+1)m},$$

with which we can derive the probability bound of the event $\max_{i=1,...,N} \left\| \frac{1}{T} \sum_{t=1}^{T} \tilde{z}_{it} v_{it} \right\| > \epsilon$, because the latter can be implied by the former. By the moment condition in Assumption 2, Lemma H.1 (i) implies that (by setting $\upsilon_{NT} = \sqrt{m}$ and $C_q = q/2$)

$$\begin{aligned}
&\Pr \left(\max_{i=1,...,N} \max_{j=0,...,p} \max_{l=0,...,m-1} \left| \frac{1}{T} \sum_{t=1}^{T} x_{itj} B_l (\tau_t) v_{it} \right| > \frac{\epsilon}{\sqrt{(p+1)m}} \right) \\
\leq &\sum_{l=0}^{m-1} \sum_{j=0}^{p} \Pr \left(\max_{i=1,...,N} \left| \frac{1}{T} \sum_{t=1}^{T} x_{itj} B_l (\tau_t) v_{it} \right| > \frac{\epsilon}{\sqrt{(p+1)m}} \right) \lesssim \frac{NTm^{q/4+1} (\log N)^{2q}}{T^{q/2}},
\end{aligned}$$

using the fact that p is a fixed constant. As a result

$$\begin{aligned}
&\Pr \left(\max_{i=1,...,N} \left\| \frac{1}{T} \sum_{t=1}^{T} \tilde{z}_{it} v_{it} \right\| > \epsilon \right) \\
\leq &\Pr \left(\max_{i=1,...,N} \max_{j=0,...,p} \max_{l=0,...,m-1} \left| \frac{1}{T} \sum_{t=1}^{T} x_{itj} B_l (\tau_t) v_{it} \right| > \frac{\epsilon}{\sqrt{(p+1)m}} \right) \\
\lesssim &\frac{NTm^{q/4+1} (\log N)^{2q}}{T^{q/2}}.
\end{aligned}$$

(ii) Similarly, $\tilde{z}_{it}\xi_{it}$ is $(p+1)m\times 1$. Elements in $\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\xi_{it}$ are

$$\frac{1}{T}\sum_{t=1}^{T}x_{itj}B_l\left(\tau_t\right)\xi_{it}=\frac{1}{T}\sum_{t=1}^{T}\sum_{j'=0}^{p}x_{itj}x_{itj'}B_l\left(\tau_t\right)b_{ij'}\left(\tau_t\right)$$

for $j=0,1,...,p$ and $l=0,1,...,m-1$. By the condition $\kappa\geq 1$ in Assumption 4, so that $m^{-\kappa}\ll m^{-1/2}$. Using Lemma H.1 (ii),

$$\begin{aligned}
&\Pr\left(\max_{i=1,...,N}\max_{j=0,...,p}\max_{l=0,...,m-1}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j'=0}^{p}x_{itj}x_{itj'}B_l\left(\tau_t\right)b_{ij'}\left(\tau_t\right)\right|>\frac{\epsilon}{\sqrt{(p+1)m}}\right)\\
\leq&\sum_{j=0}^{p}\sum_{l=0}^{m-1}\sum_{j'=0}^{p}\Pr\left(\max_{i=1,...,N}\left|\frac{1}{T}\sum_{t=1}^{T}x_{itj}x_{itj'}B_l\left(\tau_t\right)b_{ij'}\left(\tau_t\right)\right|>\frac{\epsilon}{(p+1)^{3/2}\sqrt{m}}\right)\\
\leq&\sum_{j=0}^{p}\sum_{l=0}^{m-1}\sum_{j'=0}^{p}\Pr\left(\max_{i=1,...,N}\left|\frac{1}{T}\sum_{t=1}^{T}x_{itj}x_{itj'}B_l\left(\tau_t\right)b_{ij'}\left(\tau_t\right)\right|>Mm^{-\kappa}\right)\\
\lesssim&\frac{NTm\left(\log N\right)^{2q}}{T^{q/2}}.
\end{aligned}$$

for an arbitrary large M after some large T, where the third line uses $m^{-\kappa}\ll m^{-1/2}$ and the last line holds by applying Lemma H.1 (ii) with $C_q=q/2$ and the fact that p is fixed.

We reach the desired result by

$$\begin{aligned}
&\Pr\left(\max_{i=1,...,N}\left\|\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\xi_{it}\right\|>\epsilon\right)\\
\leq&\Pr\left(\max_{i=1,...,N}\max_{l=0,...,m}\max_{j=0,...,p}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j'=0}^{p}x_{itj}x_{itj'}B_l\left(\tau_t\right)b_{ij'}\left(\tau_t\right)\right|>\frac{\epsilon}{\sqrt{(p+1)m}}\right).
\end{aligned}$$

□

Proof of Lemma H.5. **(i)** By (2.6) and (E.3), we substitute $y_{it}=\tilde{z}_{it}'\tilde{\pi}_i^0+\xi_{it}+v_{it}$ into $\hat{\tilde{\pi}}_i$, and obtain

$$\hat{\tilde{\pi}}_i-\tilde{\pi}_i^0=\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\tilde{z}_{it}'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\xi_{it}\right)+\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\tilde{z}_{it}'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}v_{it}\right).$$

Then

$$\left\|\hat{\tilde{\pi}}_i-\tilde{\pi}_i^0\right\|\leq\mu_{\min}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\tilde{z}_{it}'\right)\left[\left\|\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\xi_{it}\right\|+\left\|\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}v_{it}\right\|\right],$$

which implies

$$
\begin{aligned}
&\Pr\left(\max_{1\leq i\leq N}\left\|\hat{\tilde{\pi}}_i-\tilde{\pi}_i^0\right\|>\epsilon\right)\\
\leq&\Pr\left(\max_{1\leq i\leq N}\mu_{\min}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\tilde{z}_{it}'\right)>C_{zz}^{-1}\right)+\Pr\left(\max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\xi_{it}\right\|+\left\|\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}v_{it}\right\|>C_{zz}\epsilon\right)\\
\leq&\Pr\left(\max_{1\leq i\leq N}\mu_{\min}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\tilde{z}_{it}'\right)>C_{zz}^{-1}\right)+\Pr\left(\max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\xi_{it}\right\|>\frac{C_{zz}\epsilon}{2}\right)\\
&+\Pr\left(\max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}v_{it}\right\|>\frac{C_{zz}\epsilon}{2}\right)\\
=&o\left(1\right),
\end{aligned}
$$

by Lemmas H.3 and H.4, and the rates in Assumption 6 (ii).

(ii) Note that

$$
\begin{aligned}
\hat{\sigma}_{vi}^2-\sigma_{vi}^2&=\frac{1}{T-1}\sum_{t=1}^{T}\left(y_{it}-\tilde{z}_{it}'\hat{\tilde{\pi}}_i\right)^2-\sigma_{vi}^2\\
&=\frac{1}{T-1}\sum_{t=1}^{T}\left(\tilde{z}_{it}'\left(\tilde{\pi}_i^0-\hat{\tilde{\pi}}_i\right)+\xi_{it}+v_{it}\right)^2-\sigma_{vi}^2\\
&=\frac{1}{T-1}\sum_{t=1}^{T}v_{it}^2-\sigma_{vi}^2+\left(\tilde{\pi}_i^0-\hat{\tilde{\pi}}_i\right)'\left(\frac{1}{T-1}\sum_{t=1}^{T}\tilde{z}_{it}\tilde{z}_{it}'\right)\left(\tilde{\pi}_i^0-\hat{\tilde{\pi}}_i\right)+\frac{1}{T-1}\sum_{t=1}^{T}\xi_{it}^2\\
&+2\left[\frac{1}{T-1}\sum_{t=1}^{T}\left(\xi_{it}+v_{it}\right)z_{it}'\right]\left(\tilde{\pi}_i^0-\hat{\tilde{\pi}}_i\right)+2\frac{1}{T-1}\sum_{t=1}^{T}\xi_{it}v_{it}\\
&\equiv A_{i1}+A_{i2}+A_{i3}+A_{i4}+A_{i5}.
\end{aligned}
$$

The uniform convergence of $A_{i1}, A_{i2}, A_{i3}, A_{i4},$ and A_{i5} can be similarly shown as in Lemmas H.1, H.3, H.4 and part (i) of this lemma. We omit the proof for conciseness. Therefore

$$
\Pr\left(\max_{i=1,\ldots,N}\left|\hat{\sigma}_{vi}^2-\sigma_{vi}^2\right|>\epsilon\right)=o\left(1\right).
$$

□

Proof of Lemma H.6. Recall that Lemma H.2 shows that

$$
\frac{\underline{C}_{xx}}{2}\leq\mu_{\min}\left\{\mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\tilde{z}_{it}'\right)\right\}\leq\mu_{\max}\left\{\mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\tilde{z}_{it}'\right)\right\}\leq 2\bar{C}_{xx},
$$

after some large T. Note $\tilde{z}_{it} = (1, z_{it}')'$. Thus,

$$\mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\tilde{z}_{it}'\right) = \begin{pmatrix} 1 & \mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}'\right) \\ \mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}\right) & \mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}z_{it}'\right) \end{pmatrix}.$$

With the block representation of $\mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\tilde{z}_{it}'\right)$ and its full rank condition, the inverse of it can be calculated as (we only present its lower diagonal):

$$\begin{aligned}\left[\mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{z}_{it}\tilde{z}_{it}'\right)\right]^{-1} &= \begin{pmatrix} \cdot_{1\times 1} & \cdot_{1\times(mp+m-1)} \\ \cdot_{(mp+m-1)\times 1} & \left[\mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}z_{it}'\right) - \mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}\right)\mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}z_{it}'\right)\right]^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \cdot_{1\times 1} & \cdot_{1\times(mp+m-1)} \\ \cdot_{(mp+m-1)\times 1} & \left[\mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\ddot{z}_{it}\ddot{z}_{it}'\right)\right]^{-1} \end{pmatrix}.\end{aligned}$$

The above implies that $\mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\ddot{z}_{it}\ddot{z}_{it}'\right)$ must be of full rank, and

$$\frac{\bar{C}_{xx}^{-1}}{2} \leq \min_{1\leq i\leq N}\mu_{\min}\left\{\left[\mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\ddot{z}_{it}\ddot{z}_{it}'\right)\right]^{-1}\right\} \leq \max_{1\leq i\leq N}\mu_{\max}\left\{\left[\mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\ddot{z}_{it}\ddot{z}_{it}'\right)\right]^{-1}\right\} \leq 2\underline{C}_{xx}^{-1}.$$

We obtain the desired result by $\min_{1\leq i\leq N}\mu_{\min}\left(A_i^{-1}\right) = 1/\max_{1\leq i\leq N}\mu_{\max}\left(A_i\right)$ for any full rank matrices $\{A_i\}_{i=1}^{N}$. □

Proof of Lemma H.7. The proof here follows that of Lemma H.3. We similarly define

$$I_{ss'(k)} \equiv \frac{1}{N_k}\sum_{i\in G_{k|K^*}}\frac{1}{T}\sum_{t=1}^{T}\left\{B_s\left(\tau_t\right)B_{s'}\left(\tau_t\right)\left[x_{itj}x_{itl} - \mathrm{E}\left(x_{itj}x_{itl}\right)\right]\right\}.$$

Its probability bound is

$$\begin{aligned}\Pr\left(\max_{s,s'=1,\ldots,\underline{m}}\left|I_{ss'(k)}\right| > \frac{\epsilon}{\underline{m}p^2}\right) &\leq \sum_{s,s'=1}^{\underline{m}}\Pr\left(\left|I_{ss'(k)}\right| > \frac{\epsilon}{\underline{m}p^2}\right) \\ &\lesssim \frac{\underline{m}^{q/2+2}\left(\log N\right)^{2q}}{\left(NT\right)^{q/2-1}},\end{aligned}$$

where we use a similar argument as in the proof of (i) in Lemma H.1 and we set $\upsilon_{NT} = \underline{m}p^2$, $C_{NT} = \upsilon_{NT}^{-1}NT\left(\log N\right)^{-4}$, in addition, we use the assumptions that $T \propto N^C$ and $N_k \propto N$.

The desired result follows, using the same logic as in the proof of (i) in Lemma H.1. □

Proof of Lemma H.8. To analyze A_{k1}, we introduce some new notation:

$$\ddot{\underline{Z}}_{(k)} \equiv \left(\ddot{\underline{z}}_{11}, ..., \ddot{\underline{z}}_{1T}, ..., \ddot{\underline{z}}_{N_k1}, ..., \ddot{\underline{z}}_{N_kT}\right)',$$

a $N_kT \times (\underline{m} - 1 + \underline{m}p)$ matrix, where we abuse the notation by letting $\ddot{\underline{z}}_{i1}$ in $\ddot{\underline{Z}}_{(k)}$ denote all observations in $G_{k|K}$. Similarly

$$\ddot{\xi}_{(k)} \equiv \left(\ddot{\xi}_{11}, ..., \ddot{\xi}_{1T}, ..., \ddot{\xi}_{N_k1}, ..., \ddot{\xi}_{N_kT}\right)',$$

a $N_kT \times 1$ vector.

By Lemmas H.6 and H.7 on $\frac{1}{N_kT}\ddot{Z}'_{(k)}\ddot{Z}_{(k)}$, with very probability,

$$\mu_{\min}\left(\frac{1}{N_kT}\ddot{Z}'_{(k)}\ddot{Z}_{(k)}\right) = \mu_{\min}\left(\frac{1}{N_kT}\sum_{i\in G_{k|K^*}}\sum_{t=1}^{T}\ddot{\underline{z}}_{it}\ddot{\underline{z}}'_{it}\right) \geq \frac{\underline{C}_{xx}}{2} - \epsilon \geq \frac{\underline{C}_{xx}}{3}, \tag{H.3}$$

by setting a small enough ϵ, and similarly

$$\mu_{\max}\left(\frac{1}{N_kT}\ddot{Z}'_{(k)}\ddot{Z}_{(k)}\right) \leq 3\bar{C}_{xx}.$$

Thus, with very high probability,

$$\begin{aligned}
\|A_{k1}\| &= \left[\frac{1}{N_kT}\ddot{\xi}'_{(k)}\ddot{Z}_{(k)}\left(\frac{1}{N_kT}\ddot{Z}'_{(k)}\ddot{Z}_{(k)}\right)^{-1}\left(\frac{1}{N_kT}\ddot{Z}'_{(k)}\ddot{Z}_{(k)}\right)^{-1}\frac{1}{N_kT}\ddot{Z}'_{(k)}\ddot{\xi}_{(k)}\right]^{1/2} \\
&\leq 3\underline{C}_{xx}^{-1}\left[\frac{1}{N_kT}\ddot{\xi}'_{(k)}\left(\frac{1}{N_kT}\ddot{Z}_{(k)}\ddot{Z}'_{(k)}\right)\ddot{\xi}_{(k)}\right]^{1/2} \\
&\leq 3^{3/2}\underline{C}_{xx}^{-1}\bar{C}_{xx}^{1/2}\left[\frac{1}{N_kT}\ddot{\xi}'_{(k)}\ddot{\xi}_{(k)}\right]^{1/2} \\
&= 3^{3/2}\underline{C}_{xx}^{-1}\bar{C}_{xx}^{1/2}\left[\frac{1}{N_kT}\sum_{i\in G_{k|K}}\sum_{t=1}^{T}\ddot{\xi}_{it}^2\right]^{1/2}.
\end{aligned} \tag{H.4}$$

By the rate in (E.5),

$$\begin{aligned}
\frac{1}{N_kT}\sum_{i\in G_{k|K}}\sum_{t=1}^{T}\ddot{\xi}_{it}^2 &\leq \left[\frac{1}{N_kT}\sum_{i\in G_{k|K}}\sum_{t=1}^{T}\left(1 + \sum_{l=1}^{p}|x_{itl}|\right)^2\right]\cdot O\left(m^{-2\kappa}\right) \\
&= O_P\left(m^{-2\kappa}\right),
\end{aligned} \tag{H.5}$$

where the second line holds by the moment condition in Assumption 2. Substitute (H.5) back to (H.4), and we obtain the desired result. □

Proof of Lemma H.9. Recall that

$$\mathbb{M}_{\mathbb{B}}(s) = \begin{pmatrix} \mathbb{B}_{-0}^{\underline{m}}(s)' & 0 & \cdots & 0 \\ 0 & \mathbb{B}^{\underline{m}}(s)' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbb{B}^{\underline{m}}(s)' \end{pmatrix}_{(p+1)\times(\underline{m}-1+\underline{m}p)},$$

and

$$Q_{(k),zz} = \frac{1}{N_k T} \sum_{i \in G_{k|K}} \sum_{t=1}^{T} \ddot{\underline{z}}_{it} \ddot{\underline{z}}_{it}'.$$

For any $(p+1) \times 1$ vector a,

$$\begin{aligned} \sqrt{\frac{N_k T}{\underline{m}}} a' \mathbb{M}_{\mathbb{B}}(s) A_{k2} &= a' \mathbb{M}_{\mathbb{B}}(s) \left(\frac{1}{N_k T} \sum_{i \in G_{k|K}} \sum_{t=1}^{T} \ddot{\underline{z}}_{it} \ddot{\underline{z}}_{it}' \right)^{-1} \left(\frac{1}{\sqrt{N_k T \underline{m}}} \sum_{i \in G_{k|K}} \sum_{t=1}^{T} \ddot{\underline{z}}_{it} \ddot{v}_{it} \right) \\ &= a' \mathbb{M}_{\mathbb{B}}(s) Q_{(k),zz}^{-1} \left(\frac{1}{\sqrt{N_k T \underline{m}}} \sum_{i \in G_{k|K}} \sum_{t=1}^{T} \ddot{\underline{z}}_{it} v_{it} \right) \\ &= \frac{1}{\sqrt{N_k}} \sum_{i \in G_{k|K}} \left\{ \frac{1}{\sqrt{T \underline{m}}} a' \mathbb{M}_{\mathbb{B}}(s) Q_{(k),zz}^{-1} \sum_{t=1}^{T} \ddot{\underline{z}}_{it} v_{it} \right\}, \end{aligned} \tag{H.6}$$

where the second line uses $\sum_{t=1}^{T} \ddot{\underline{z}}_{it} \ddot{v}_{it} = \sum_{t=1}^{T} \ddot{\underline{z}}_{it} v_{it}$. By the i.i.d. assumption across i and t on v_{it} and its independence with x, the conditional variance of this term can be calculated as

$$\text{Var}\left(\sqrt{\frac{N_k T}{\underline{m}}} a' \mathbb{M}_{\mathbb{B}}(s) A_{k2} \middle| x_1, ..., x_N \right) = a' \left[\frac{\sigma_{v(k)}^2}{\underline{m}} \mathbb{M}_{\mathbb{B}}(s) Q_{(k),zz}^{-1} \mathbb{M}_{\mathbb{B}}(s)' \right] a. \tag{H.7}$$

The above is finite and proportional to $\|a\|^2$. To see it, (H.3) implies that with very high probability

$$\begin{aligned} \text{Var}\left(a' \mathbb{M}_{\mathbb{B}}(s) A_{k2} | x_1, ..., x_N \right) &\geq 3 \underline{C}_{xx}^{-1} \frac{\sigma_{v(k)}^2}{\underline{m}} a' \mathbb{M}_{\mathbb{B}}(s) \mathbb{M}_{\mathbb{B}}(s)' a \\ &= 3 \underline{C}_{xx}^{-1} \sigma_{v(k)}^2 \left[\frac{1}{\underline{m}} \mathbb{B}_{-0}^{\underline{m}}(s)' \mathbb{B}_{-0}^{\underline{m}}(s) a_0^2 + \frac{1}{\underline{m}} \mathbb{B}^{\underline{m}}(s)' \mathbb{B}^{\underline{m}}(s) \sum_{l=1}^{p} a_l^2 \right] \\ &\propto \sum_{l=0}^{p} a_l^2 = \|a\|^2, \end{aligned}$$

where we abuse the notation a bit by letting $a = (a_0, a_1, ..., a_p)'$. We can similarly verify the Lindeberg condition for the last term in (H.6) as in Lemma A.8 in Huang et al. (2004) or Lemma

A.8 in Su et al. (2019). We omit this verification process due to the similarity. We then apply the Lindeberg Central Limit Theorem, and obtain

$$\sqrt{\frac{N_k T}{m}} a' \mathbb{M}_{\mathbb{B}}(s) A_{k2} \Big/ \left\{ a' \left[\frac{\sigma^2_{v(k)}}{m} \mathbb{M}_{\mathbb{B}}(s) Q^{-1}_{(k),zz} \mathbb{M}_{\mathbb{B}}(s)' \right] a \right\}^{1/2} \xrightarrow{d} N(0,1).$$

□

Proof of Lemma H.10. We only show the result that

$$\Pr\left(\text{IC}(K^*, \lambda_{NT}) < \text{IC}(K^* - 1, \lambda_{NT})\right) \to 1.$$

Other cases are similar. Define an event

$$\mathcal{M}_1 = \{\text{Two groups merged and other groups corrected classified}\}.$$

By the nature of the HAC algorithm, and the proof in Theorem 3.1, we can similarly show that

$$\Pr(\mathcal{M}_1) \to 1.$$

Recall that

$$\mathcal{M} \equiv \left\{ \left(\hat{G}_{1|K^*}, \hat{G}_{2|K^*}, \ldots, \hat{G}_{K^*|K^*} \right) = \left(G_{1|K^*}, G_{2|K^*}, \ldots, G_{K^*|K^*} \right) \right\}.$$

We first show the result conditional on $\mathcal{M}$ and $\mathcal{M}_1$, then we show the result unconditionally. Without loss of generality, assume that observations in group $K^* - 1$ and K^* are merged. Thus,

$$\begin{aligned}
&\frac{\text{IC}(K^*, \lambda_{NT}) - \text{IC}(K^* - 1, \lambda_{NT})}{(N_{K^*-1} + N_{K^*}) T} \\
&= \frac{1}{(N_{K^*-1} + N_{K^*}) T} \left\{ \sum_{k=K^*-1}^{K^*} N_k T \log\left(\hat{\sigma}_{v(k|K^*)}\right) + \lambda_{NT} - (N_{K^*-1} + N_{K^*}) T \log\left(\hat{\sigma}_{v(K^*-1|K^*-1)}\right) \right\} \\
&= \frac{N_{K^*-1}}{N_{K^*-1} + N_{K^*}} \log\left(\hat{\sigma}_{v(K^*-1|K^*)}\right) + \frac{N_{K^*}}{N_{K^*-1} + N_{K^*}} \log\left(\hat{\sigma}_{v(K^*|K^*)}\right) - \log\left(\hat{\sigma}_{v(K^*-1|K^*-1)}\right) + o(1),
\end{aligned} \tag{H.8}$$

where the last line uses $\lambda_{NT} = o(NT)$ and $N_k \propto N$.

We claim the result below and we defer its proof to the end:

$$\hat{\sigma}^2_{v(K^*-1|K^*-1)} = \frac{N_{K^*-1}}{N_{K^*-1}+N_{K^*}}\left(\hat{\sigma}^2_{v(K^*-1|K^*)}+\Delta_1^2\right)+\frac{N_{K^*}}{N_{K^*-1}+N_{K^*}}\left(\hat{\sigma}^2_{v(K^*|K^*)}+\Delta_2^2\right)+o_P(1), \tag{H.9}$$

where

$$\Delta_1^2 \equiv \frac{1}{N_{K^*-1}T}\sum_{i\in G_{K^*-1|K^*}}\sum_{t=1}^{T}\left[\ddot{z}'_{it}\left(\pi^{0*}_{(K^*-1|K^*-1)}-\pi^{0*}_{(K^*-1|K^*)}\right)\right]^2, \text{ and}$$
$$\Delta_2^2 \equiv \frac{1}{N_{K^*}T}\sum_{i\in G_{K^*|K^*}}\sum_{t=1}^{T}\left[\ddot{z}'_{it}\left(\pi^{0*}_{(K^*-1|K^*-1)}-\pi^{0*}_{(K^*|K^*)}\right)\right]^2,$$

and ${\pi^{0*}}_{(K^*-1|K^*-1)}$ denotes the estimand of π for $G_{K^*-1|K^*}\cup G_{K^*|K^*}$.

Using (H.9) and the Jensen's inequality,

$$\begin{aligned}\log\hat{\sigma}^2_{v(K^*-1|K^*-1)} \geq{}& \frac{N_{K^*-1}}{N_{K^*-1}+N_{K^*}}\log\left(\hat{\sigma}^2_{v(K^*-1|K^*)}+\Delta_1^2\right)\\ &+\frac{N_{K^*}}{N_{K^*-1}+N_{K^*}}\log\left(\hat{\sigma}^2_{v(K^*|K^*)}+\Delta_2^2\right)+o_P(1).\end{aligned} \tag{H.10}$$

Since

$$\left\{\alpha^*_{(K^*-1)},\beta^*_{(K^*-1)},\sigma^{*2}_{v(K^*-1)}\right\}\neq\left\{\alpha^*_{(K^*)},\beta^*_{(K^*)},\sigma^{*2}_{v(K^*)}\right\},$$

we either have $\sigma^{*2}_{v(K^*-1)}\neq\sigma^{*2}_{v(K^*)}$ or $\left\{\alpha^*_{(K^*-1)},\beta^*_{(K^*-1)}\right\}\neq\left\{\alpha^*_{(K^*)},\beta^*_{(K^*)}\right\}$, so that either $\hat{\sigma}^2_{v(K^*-1|K^*)}\neq\hat{\sigma}^2_{v(K^*|K^*)}$ or $\max\{\Delta_1^2,\Delta_2^2\}>0$ holds with very high probability. Together with this, we can change "$\geq$" in (H.10) to "$>$", so that

$$\log\hat{\sigma}^2_{v(K^*-1|K^*-1)} > \frac{N_{K^*-1}}{N_{K^*-1}+N_{K^*}}\log\left(\hat{\sigma}^2_{v(K^*-1|K^*)}\right)+\frac{N_{K^*}}{N_{K^*-1}+N_{K^*}}\log\left(\hat{\sigma}^2_{v(K^*|K^*)}\right)+o_P(1) \tag{H.11}$$

Using (H.11), (H.8) implies that

$$\text{IC}(K^*,\lambda_{NT})-\text{IC}(K^*-1,\lambda_{NT})<0$$

holds with very high probability after some large N and T.

We have shown the desired result conditional on $\mathcal{M}$ and $\mathcal{M}_1$. Since

$$\Pr\left(\mathcal{M} \cap \mathcal{M}_1\right) \geq 1 - \Pr\left(\mathcal{M}^c\right) - \Pr\left(\mathcal{M}_1^c\right) \to 1,$$

the desired result then holds unconditionally.

We finish the proof by showing the claim in (H.9). Then as in the proof of (ii) in Theorem 3.2 (the decomposition of $\hat{\sigma}^2_{v(k|K^*)}$), we have

$$\begin{aligned}
&\sum_{i \in G_{K^*-1|K^*} \cup G_{K^*|K^*}} \sum_{t=1}^{T} \left(\ddot{y}_{it} - \ddot{\underline{z}}'_{it} \hat{\pi}_{(K^*-1|K^*-1)}\right)^2 \\
&= \sum_{i \in G_{K^*-1|K^*}} \sum_{t=1}^{T} \left[\ddot{\underline{z}}'_{it} \left(\hat{\pi}_{(K^*-1|K^*-1)} - \pi^{0*}_{(K^*-1|K^*-1)}\right) + \ddot{\underline{z}}'_{it} \left(\pi^{0*}_{(K^*-1|K^*-1)} - \pi^{0*}_{(K^*-1|K^*)}\right) + \ddot{\xi}_{it} + \ddot{v}_{it}\right]^2 \\
&\quad + \sum_{i \in \cup G_{K^*|K^*}} \sum_{t=1}^{T} \left[\ddot{\underline{z}}'_{it} \left(\hat{\pi}_{(K^*-1|K^*-1)} - \pi^{0*}_{(K^*-1|K^*-1)}\right) + \ddot{\underline{z}}'_{it} \left(\pi^{0*}_{(K^*-1|K^*-1)} - \pi^{0*}_{(K^*|K^*)}\right) + \ddot{\xi}_{it} + \ddot{v}_{it}\right]^2.
\end{aligned}$$

Using the decomposition above and by the same analysis for $\hat{\sigma}^2_{v(k|K^*)}$ in the proof of (ii) in Theorem 3.2, we can extract the leading term of the following:

$$\begin{aligned}
\hat{\sigma}^2_{v(K^*-1|K^*-1)} =& \frac{1}{\left(N_{K^*-1} + N_{K^*}\right)(T-1)} \sum_{i \in G_{K^*-1|K^*} \cup G_{K^*|K^*}} \sum_{t=1}^{T} \left(\ddot{y}_{it} - \ddot{\underline{z}}'_{it} \hat{\pi}_{(K^*-1|K^*-1)}\right)^2 \\
=& \frac{1}{\left(N_{K^*-1} + N_{K^*}\right)(T-1)} \sum_{i \in G_{K^*-1|K^*}} \sum_{i \in G_{K^*|K^*}} \sum_{t=1}^{T} \ddot{v}_{it}^2 + \Delta_1^2 + \Delta_2^2 + o_P(1) \\
=& \frac{N_{K^*-1}}{N_{K^*-1} + N_{K^*}} \hat{\sigma}^2_{v(K^*-1|K^*)} + \Delta_1^2 + \frac{N_{K^*}}{N_{K^*-1} + N_{K^*}} \hat{\sigma}^2_{v(K^*|K^*)} + \Delta_2^2 + o_P(1).
\end{aligned}$$

where the second and third lines repeatedly use the arguments for the analysis of $\hat{\sigma}^2_{v(k|K^*)}$. The above is the desired result. □

Proof of Lemma H.11. We only show the result for $K = K^* + 1$. Other cases are similar. Define an event

$$\mathcal{M}_2 = \{\text{One group separated into two groups and other groups corrected classified}\}.$$

By the nature of the HAC algorithm, and the proof in Theorem 3.1, we can similarly show that

$$\Pr\left(\mathcal{M}_2\right) \to 1.$$

We show the result conditional on $\mathcal{M}$ and $\mathcal{M}_2$. Without loss of generality, assume that observations in group K^* are separated into two groups, and the numbers of observations in groups K^* and K^*+1 are denoted as N_{K^*1} and N_{K^*2}, respectively. Clearly $N_{K^*} = N_{K^*1} + N_{K^*2}$. Since these two groups come from the same underlying group, the parameters of interest are the same. Using the rates in Theorem 3.2,

$$\hat{\sigma}^2_{v(K^*|K^*+1)} - \sigma^{*2}_{v(K^*)} = O_P\left(\frac{1}{\sqrt{N_{K^*1}T}}\right) \text{ and } \hat{\sigma}^2_{v(K^*+1|K^*+1)} - \sigma^{*2}_{v(K^*)} = O_P\left(\frac{1}{\sqrt{N_{K^*2}T}}\right).$$

Since $\hat{\sigma}^2_{v(K^*|K^*)} - \sigma^{*2}_{v(K^*)} = O_P\left(\frac{1}{\sqrt{N_{K^*}T}}\right)$, with the above, we have

$$\hat{\sigma}^2_{v(K^*|K^*+1)} - \hat{\sigma}^2_{v(K^*|K^*)} = O_P\left(\frac{1}{\sqrt{N_{K^*1}T}}\right) \text{ and } \hat{\sigma}^2_{v(K^*+1|K^*+1)} - \hat{\sigma}^2_{v(K^*|K^*)} = O_P\left(\frac{1}{\sqrt{N_{K^*2}T}}\right)$$

Then

$$\begin{aligned}
&\frac{\text{IC}\left(K^*+1, \lambda_{NT}\right) - \text{IC}\left(K^*, \lambda_{NT}\right)}{N_{K^*}T} \\
=&\frac{1}{N_{K^*}T}\left\{N_{K^*1}T\log\left(\hat{\sigma}_{v(K^*|K^*+1)}\right) + N_{K^*2}T\log\left(\hat{\sigma}_{v(K^*+1|K^*+1)}\right) + \lambda_{NT} \ -N_kT\log\left(\hat{\sigma}_{v(K^*|K^*)}\right)\right\} \\
=&\frac{N_{K^*1}}{N_{K^*}}\left[\log\left(\hat{\sigma}^2_{v(K^*|K^*+1)}\right) - \log\left(\hat{\sigma}^2_{v(K^*|K^*)}\right)\right] + \frac{N_{K^*2}}{N_{K^*}}\left[\log\left(\hat{\sigma}^2_{v(K^*+1|K^*+1)}\right) - \log\left(\hat{\sigma}^2_{v(K^*|K^*)}\right)\right] + \frac{\lambda_{NT}}{N_{K^*}T} \\
=&\frac{N_{K^*1}}{N_{K^*}}O_P\left(\frac{1}{\sqrt{N_{K^*1}T}}\right) + \frac{N_{K^*2}}{N_{K^*}}O_P\left(\frac{1}{\sqrt{N_{K^*2}T}}\right) + \frac{\lambda_{NT}}{N_{K^*}T} \\
=&\left(\sqrt{\frac{N_{K^*1}}{N_{K^*}}} + \sqrt{\frac{N_{K^*2}}{N_{K^*}}}\right)\cdot O_P\left(\frac{1}{\sqrt{N_{K^*}T}}\right) + \frac{\lambda_{NT}}{N_{K^*}T} \\
=&O_P\left(\frac{1}{\sqrt{N_{K^*}T}}\right) + \frac{\lambda_{NT}}{N_{K^*}T} > 0,
\end{aligned}$$

with very high probability, where the last line holds due to $\sqrt{\frac{N_{K^*1}}{N_{K^*}}} + \sqrt{\frac{N_{K^*2}}{N_{K^*}}} \leq 2$, $\lambda_{NT} \gg \sqrt{NT}$, and $N_{K^*} \propto N$.

We have shown the desired result conditional on $\mathcal{M}$ and $\mathcal{M}_2$. Since

$$\Pr\left(\mathcal{M} \cap \mathcal{M}_2\right) \geq 1 - \Pr\left(\mathcal{M}^c\right) - \Pr\left(\mathcal{M}_2^c\right) \to 1,$$

the desired result then holds unconditionally. □

Proof of Lemma H.12. Without loss of generality, we ignore the group structure of other parameters for the following discussion. Moreover, we assume that we know $(\alpha(\cdot), \beta(\cdot), \sigma_v^2)$. It is innocuous because $\hat{\vartheta}$ converges to the true value faster than $\sqrt{N}$ which is the convergence rate of $(\hat{\alpha}^0, \hat{\sigma}_u^2)$.

As in Lemmas H.10 and H.11, we only show that

$$\Pr\left(\widetilde{\mathrm{IC}}\left(\mathcal{K}^*, \tilde{\lambda}_{NT}\right) < \widetilde{\mathrm{IC}}\left(\mathcal{K}^* - 1, \tilde{\lambda}_{NT}\right)\right) \to 1,$$

and

$$\Pr\left(\widetilde{\mathrm{IC}}\left(\mathcal{K}^*, \tilde{\lambda}_{NT}\right) < \widetilde{\mathrm{IC}}\left(\mathcal{K}^* + 1, \tilde{\lambda}_{NT}\right)\right) \to 1.$$

As such, we divide the proof into two parts with Part 1 showing the first result and Part 2 for the other.

Part 1. We show in Appendix G.2 that the information matrix is a well-behaved positive definite matrix with diagonals being finite constants, even though $T \to \infty$. Recall that

$$\varrho^0 = \left(\alpha_{(1)}^0, \sigma_{u(1)}^2, ..., \alpha_{(\mathcal{K}^*)}^0, \sigma_{u(\mathcal{K}^*)}^2, \tau_1^0, ..., \tau_{\mathcal{K}^*-1}^0\right).$$

The true density function for y is

$$\tilde{f}\left(y \left| x; \varrho^0, \vartheta\right.\right),$$

where we abuse the notation by letting $\vartheta \equiv (\sigma_v^2, \alpha(\cdot), \beta(\cdot))$. By definition, ϱ^0 uniquely maximizes

$$\log\left[\tilde{f}\left(y \left| x; \varrho, \vartheta\right.\right)\right].$$

For a ϱ that is in a small neighborhood of ϱ^0,

$$\log\left[\tilde{f}\left(y \left| x; \varrho, \vartheta\right.\right)\right] - \log\left[\tilde{f}\left(y \left| x; \varrho^0, \vartheta\right.\right)\right] = -\left(\varrho - \varrho^0\right)' \mathbb{I}\left(\varrho - \varrho^0\right) + o\left(\left\|\varrho - \varrho^0\right\|^2\right).$$

As mentioned at the beginning, $\mathbb{I}$ is positive definite. Further, we restrict our attention to a compact support of ϱ, therefore, there exists a C such that

$$\log\left[\tilde{f}\left(y \left| x; \varrho, \vartheta\right.\right)\right] - \log\left[\tilde{f}\left(y \left| x; \varrho^0, \vartheta\right.\right)\right] \leq -C < 0, \tag{H.12}$$

for all ϱ outside of a small neighborhood of ϱ^0.

When we restrict the mixture distribution to come from $\mathcal{K}^* - 1$ distributions, clearly, the estimate converges to a parameter that is outside the small neighborhood of ϱ^0. By the consistency and inequality in (H.12),

$$\tilde{\text{IC}}\left(\mathcal{K}^*, \tilde{\lambda}_{NT}\right) - \tilde{\text{IC}}\left(\mathcal{K}^* - 1, \tilde{\lambda}_{NT}\right) \leq -C \cdot N + o_P\left(N\right) + \tilde{\lambda}_{NT}.$$

Further, $\tilde{\lambda}_{NT} \ll N$, the above implies the first desired result.

Part 2. When we force $\mathcal{K} = \mathcal{K}^* + 1$, denote

$$\varsigma^0 = \left(\alpha^0_{(1)}, \sigma^2_{u(1)}, ..., \alpha^0_{(\mathcal{K}^*+1)}, \sigma^2_{u(\mathcal{K}^*+1)}, \tau^0_1, ..., \tau^0_{\mathcal{K}^*}\right),$$

as the population parameter. It is not unique but we use the one that the estimated parameters converge to. ς^0 is essentially the same as ϱ^0 in terms of mixture distribution; mathematically,

$$\tilde{f}\left(y_i \left| x_i; \varsigma^0, \vartheta\right.\right) = \tilde{f}\left(y_i \left| x_i; \varrho^0, \vartheta\right.\right), \tag{H.13}$$

where as before $\vartheta = (\sigma^2_v, \alpha(\cdot), \beta(\cdot))$. To see that, on the one hand

$$\text{E}\left[\log \tilde{f}\left(y_i \left| x_i; \varsigma^0, \vartheta\right.\right)\right] \geq \text{E}\left[\log \tilde{f}\left(y_i \left| x_i; \varrho^0, \vartheta\right.\right)\right],$$

because ς^0 has more degrees of freedom than that of ϱ^0 . On the other hand,

$$\text{E}\left[\log \tilde{f}\left(y_i \left| x_i; \varsigma^0, \vartheta\right.\right)\right] \leq \text{E}\left[\log \tilde{f}\left(y_i \left| x_i; \varrho^0, \vartheta\right.\right)\right],$$

because $\tilde{f}\left(y_i \left| x_i; \varrho^0, \vartheta\right.\right)$ is the true underlying distribution. Therefore, (H.13) must hold.

Let $\hat{\varsigma}$ be the estimates when $\mathcal{K} = \mathcal{K}^* + 1$. Then, using (H.13),

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\log\tilde{f}\left(y_i\left|x_i;\hat{\varsigma},\vartheta\right.\right) - \frac{1}{N}\sum_{i=1}^{N}\log\tilde{f}\left(y_i\left|x_i;\hat{\varrho},\vartheta\right.\right) \\
=&\frac{1}{N}\sum_{i=1}^{N}\left[\log\tilde{f}\left(y_i\left|x_i;\hat{\varsigma},\vartheta\right.\right) - \log\tilde{f}\left(y_i\left|x_i;\varsigma^0,\vartheta\right.\right)\right] \\
&- \frac{1}{N}\sum_{i=1}^{N}\left[\log\tilde{f}\left(y_i\left|x_i;\hat{\varrho},\vartheta\right.\right) - \log\tilde{f}\left(y_i\left|x_i;\varrho^0,\vartheta\right.\right)\right] \\
=&\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\varsigma}\log\tilde{f}_i\Big|_{\varsigma=\varsigma^0}\right)'\left(\hat{\varsigma}-\varsigma^0\right) + O_P\left(\left\|\hat{\varsigma}-\varsigma^0\right\|^2\right) \\
&+\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\varrho}\log\tilde{f}_i\Big|_{\varrho=\varrho^0}\right)'\left(\hat{\varrho}-\varrho^0\right) + O_P\left(\left\|\hat{\varrho}-\varrho^0\right\|^2\right) \\
=&O_P\left(N^{-1}\right),
\end{aligned}$$

where the last line holds by the first order condition and the independence across N so that

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\varsigma}\log f_i\Big|_{\varsigma=\varsigma^0} = O_P\left(N^{-1/2}\right) \text{ and } \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\varrho}\log\tilde{f}_i\Big|_{\varrho=\varrho^0} = O_P\left(N^{-1/2}\right),$$

$\|\hat{\varsigma}-\varsigma^0\| = O_P\left(N^{-1/2}\right)$, and $\|\hat{\varrho}-\varrho^0\| = O_P\left(N^{-1/2}\right)$. Using the above,

$$\widetilde{\mathrm{IC}}\left(\mathcal{K}^*,\tilde{\lambda}_{NT}\right) - \widetilde{\mathrm{IC}}\left(\mathcal{K}^*+1,\tilde{\lambda}_{NT}\right) = -\tilde{\lambda}_{NT} + O_P\left(1\right).$$

Then

$$\Pr\left(\widetilde{\mathrm{IC}}\left(\mathcal{K}^*,\tilde{\lambda}_{NT}\right) < \widetilde{\mathrm{IC}}\left(\mathcal{K}^*+1,\tilde{\lambda}_{NT}\right)\right) \to 1,$$

due to $\tilde{\lambda}_{NT} \to \infty$. □

References

BICKEL, P. J., AND K. A. DOKSUM(2015). Mathematical Statistics: Basic Ideas and Selected Topics. Second Edition, Volume 1, Chapman & Hall/CRC Texts in Statistical Science. E

CHEN, X. (2007). Large Sample Sieve Estimation of Semi-Nonparametric Models. Handbook of Econometrics, Volume 6, Part B, 5549-5632. E

CHRISTIAN, M. H., H. MANNER, AND L. SIMAR (2018), "The 'Wrong Skewness' Problem in Stochastic Frontier Models: A New Approach," *Econometric Reviews*, 37(4), 380-400. A.3

FAN J., Y. FENG, AND R. SONG (2011): "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models," *Journal of the American Statistical Association*, 106(494), 544-557. H

HUANG J., C. O. WU, AND L. ZHOU (2004): "Polynomial Spline Estimation and Inference for Varying Coefficient Models with Longitudinal Data," *Statistica Sinica*, 14(3), 763-788. H

KUMBHAKAR, S. C., C. F. PARMETER, AND E. G. TSIONAS (2013): "A Zero Inefficiency Stochastic Frontier Model," *Journal of Econometrics*, 172(1), 66-76. A.4

MERLEVEDE, F., M. PELIGRAD, AND E. RIO (2009): "Bernstein Inequality and Moderate Deviations under Strong Mixing Conditions," IMS collections. *High Dimensional Probability* V., 273-292. H

OLSON, J. A., P. SCHMIDT, AND D. M. WALDMAN (1980): "A Monte Carlo Study of Estimators of the Stochastic Frontier Production Function," *Journal of Econometrics*, 13, 67–82. A.3

RHO, S., AND P. SCHMIDT (2015): "Are All Firms Inefficient?" *Journal of Productivity Analysis*, 43, 327-349. A.4

SIMAR, L., AND P. W. WILSON (2010): "Inference from Cross-Sectional Stochastic Frontier Models," *Econometric Reviews,* 29, 62–98. A.3

SU L., T.T. YANG, Y. ZHANG, AND Q. ZHOU (2024): "A One-Covariate-at-a-Time Method for Nonparametric Additive Models," *Econometric Reviews*, 37(2), 334-349. H

WARD, J. H. (1963): "Hierarchical Groupings to Optimize an Objective Function," *Journal of the American Statistical Association*, 58, 236-244. D

Zhou J., C. F. PARMETER, AND S. C. KUMBHAKAR (2020): "Nonparametric Estimation of the Determinants of Inefficiency in the Presence of Firm Heterogeneity," *European Journal of Operational Research*, 286, 1142–1152. 1, A.2