PENN INSTITUTE for ECONOMIC RESEARCH
UNIVERSITY of PENNSYLVANIA

The Ronald O. Perelman Center for Political Science and Economics (PCPSE)
133 South 36th Street
Philadelphia, PA 19104-6297

pier@econ.upenn.edu
http://economics.sas.upenn.edu/pier

PIER Working Paper
23-016

Clustering for Multi-Dimensional Heterogeneity with an Application to Production Function Estimation

XU CHENG
University of Pennsylvania

FRANK SCHORFHEIDE
University of Pennsylvania

PENG SHAO
Auburn University

September 19, 2023

Clustering for Multi-Dimensional Heterogeneity with an Application to Production Function Estimation*

Xu Cheng† Frank Schorfheide‡ Peng Shao§

This Version: September 19, 2023

Abstract

This paper studies the estimation of multi-dimensional heterogeneous parameters in a nonlinear panel data model with endogeneity. These heterogeneous parameters are modeled with group patterns. Through estimating multiple memberships for each unit, the proposed method is robust to sparse interactions; in other words, certain combinations of unobserved features are less common compared to other combinations. We estimate the memberships along with the group-specific and common parameters in a nonlinear GMM framework and derive their large sample properties. Finally, we apply this approach to the estimation of production function and re-evaluate the trajectory of the aggregate markup.

JEL CLASSIFICATION: C13, C23, D22, D24, E23

KEY WORDS: Clustering, GMM, K-mean, Panel Data, Production Function Estimation

*An earlier version of this paper circulated since 2019 under the title "Clustering for Multi-Dimensional Heterogeneity." We thank participants at various conferences and seminars for helpful comments. Schorfheide gratefully acknowledges financial support from the National Science Foundation under Grant SES 1851634.

†Department of Economics, University of Pennsylvania. E-mail: xucheng@upenn.edu.

‡Department of Economics, University of Pennsylvania. E-mail: schorf@upenn.edu.

§Department of Economics, Auburn University. E-mail: pzs0078@auburn.edu.

1 Introduction

Firms, individuals, and countries are heterogeneous in multiple dimensions. For example, firms can differ in their productivities, in their output elasticities of variable inputs (henceforth labor elasticities), and in their output elasticities of capital (henceforth capital elasticities). A flexible specification of the production function ideally allows for heterogeneity in all of these features. In practice, the key questions for estimation are how to specify a flexible yet parsimonious econometric model that is consistent with multi-dimensional unobserved heterogeneity in the data, and how to estimate these heterogenenous parameters in a nonlinear model with endogeneity.

Building on recent developments in modeling parameter heterogeneity through group (cluster) patterns, this paper (i) proposes a framework to assign multiple group memberships to each cross-sectional unit, where each group membership is determined by one particular characteristic of the unit, and (ii) estimate the memberships as well as group-specific and common parameters in a nonlinear generalized method of moments (GMM) framework. In the context of production function estimation, we consider a setting in which there are multiple (say, low, medium, and high) productivity, labor elasticity, and capital elasticity groups. A firm may, for instance, belong to the low productivity group, the medium labor elasticity group, and the high capital elasticity group.

We employ the K-mean type classification algorithm to estimate the cluster structure and study its asymptotic properties as in Bonhomme and Manresa (2015). We show that this algorithm leads to classification consistency of the multi-dimensional cluster structure and derive the asymptotic distribution of the group-specific and common parameters in a nonlinear panel data model with endogeneity. This nonlinear panel data analysis builds on some important theoretical results developed by Su, Shi, and Phillips (2016), who proposed a new clustering algorithm by shrinkage methods. The asymptotic results are obtained as N and T pass to infinite jointly, but allow T to grow much slower than N. Thus, they are compatible with relatively short panels with a large number of cross-sectional observations.

The traditional one-dimensional clustering approach is sensitive to *sparse interactions* among different features in the data. In the production function setting with three productivity, labor elasticity, and capital elasticity levels, respectively, there exists a total of $3^3 = 27$ possible combinations. A standard one-dimensional clustering approach would allocate the firms to 27 different groups and estimate three parameters for each group. Unfortunately, this approach requires a large number of observations from each of these 27 clusters, and

implicitly also requires that the cluster sizes are balanced. When the cluster sizes are imbalanced, we say the smaller clusters represent sparse interactions among the corresponding features.[1] In the presence of such sparse interactions, the standard one-dimensional clustering approach faces challenges to detect the relatively small clusters and to estimate their parameters accurately, evidenced by the violation of regularity conditions that guarantee their consistency.

In contrast, the multi-dimensional clustering approach proposed in this paper is robust to the shape of the joint distribution of these multiple features. It investigates the clusters in each dimension, e.g., productivity, labor elasticity, and capital elasticity in the production function context, separately, and only has to estimate $3 \cdot 3 = 9$ parameters in the running example.[2] Thus, inference with respect to cluster memberships and group-specific parameters is sharpened by pooling units that are homogeneous in one dimension, e.g., productivity, but heterogeneous in the other dimensions, e.g., labor and capital elasticities. This raises the number of observations available to estimate each parameter and makes the model more parsimonious by reducing the overall number of parameters.

We model the multi-dimensional heterogeneity symmetrically by assuming all heterogeneous parameters follow group patterns, with the added flexibility that they are associated with different memberships. Once the memberships are consistently estimated, we impose the estimated cluster structure and construct a pooled GMM criterion to obtain the more efficient two-step estimator. In this setup, all cluster-specific and common parameters are estimated at the $\sqrt{NT}$ rate with asymptotically unbiased normal distributions.

We conduct Monte Carlo simulations to illustrate the small sample properties of our multi-dimensional clustering estimator. The Monte Carlo designs are closely modeled after the empirical application, albeit they use a simplified version of the production function that abstracts from capital as an input. The main objective of the experiments is to examine the effect of unknown group membership on the precision of the parameter estimates, and to compare the accuracy of our proposed multi-dimensional clustering estimator to that of a conventional one-dimensional clustering estimator. We find that the multi-dimensional approach exhibits good small sample properties and generates sharper estimates than the one-dimensional clustering estimator.

[1] They resemble the weak factors in factor analysis.

[2] More generally, suppose there are m heterogeneous features, characterized by one parameter each, and k clusters for each feature. The one-dimensional approach involves k^m unknown parameters, whereas the multi-dimensional approach involves only km unknown parameters. Admittedly, the more parsimonious approach assumes a product structure of the unknown group-specific parameters in different dimensions.

We use the proposed multi-dimensional clustering technique to estimate firm-level production functions for a subset of two digit sectors defined by the North American Industry Classification System (NAICS). Within each two-digit sector, we allow for multi-dimensional group heterogeneity in terms of total factor productivity, and output elasticities with respect to variable inputs and capital. The production functions are estimated on panel data sets for publically traded firms. Using the approach of De Loecker and Warzynski (2012), we scale the estimated variable-input elasticities by the revenue-to-variable-cost ratio to obtain an approximation of firm-level markups. We then aggregate the firm-level markups to compute an aggregate markup and re-examine the rise of aggregate markups documented by De Loecker, Eeckhout, and Unger (2020). Overall, we conclude that in our setting allowing for group heterogeneity within two-digit NAICS sectors increases the estimated aggregate markup compared to the specification that imposes within-sector homogeneity.

Our paper is related to various strands of the literature. There is a large literature on cluster analysis of heterogeneity in panel data models. Hahn and Moon (2010) studied the incidental parameter problem when the fixed effect has a finite support. In practical estimation, the cluster membership could be known (e.g., Bester and Hansen, 2016) or estimated with various classification methods (e.g., Lin and Ng, 2012; Bonhomme and Manresa, 2015; Ando and Bai, 2016; Su, Shi, and Phillips, 2016; Ke, Li, and Zhang, 2016; Gu and Volgushev, 2019; Wang and Su, 2021; Chetverikov and Manresa, 2022; Krasnokutskaya, Song, and Tang, 2022; Zhang, 2023, among others). Similar to clusters, finite mixtures models can be fruitfully applied to model group-wise heterogeneity (e.g., Sun, 2005; Kasahara and Shimotsu, 2009; Henry, Kitamura, and Salanie, 2014). In a Bayesian setting, correlated random effects distributions modeled flexibly with Dirichlet process mixture priors can also capture forms of group heterogeneity, e.g., (e.g., Liu, 2023).

Our theoretical analysis of the K-mean classification builds on related results in Bonhomme and Manresa (2015). The main difference is that we estimate group memberships by a nonlinear GMM criterion instead of a linear least square criterion.[3] Our asymptotic analysis of the nonlinear GMM problem uses technical results from Su, Shi, and Phillips (2016), who developed a classifier-LASSO approach. They considered two panel settings: a nonlinear model without endogeneity and a linear model with instrumental variables. Liu, Shang, Zhang, and Zhou (2020) considered classification in a nonlinear M-estimation framework that allows for an over-specification of the number of clusters.

[3]We do not allow the parameters to be time-varying as in Bonhomme and Manresa (2015). We also do not consider the non-discrete population heterogeneity as in Bonhomme, Lamadon, and Manresa (2022).

A few recent papers have utilized the assignment of multiple memberships. For instance, Leng, Chen, and Wang (2023) applied the multi-dimensional clustering method to study quantile estimation of a linear model with multiple heterogeneous coefficients and two-way fixed effects. Cytrynbaum (2020) considered a blocked clusterwise regression with least squares estimation of a linear model with multiple heterogeneous coefficients.

There exists a large literature on production function estimation. Since we are using our production function estimates to compute markups using the approach of De Loecker and Warzynski (2012), we only highlight a three closely related papers: De Loecker, Eeckhout, and Unger (2020) and Demirer (2022) use the so-called proxy-variable approach to estimate production functions, based on the Compustat data set of publically traded US firms, that are assumed to be homogeneous within sectors. In contrast, we use a dynamic panel data approach that is robust to the identification problems discussed in Flynn, Gandhi, and Traina (2019) and allows for group heterogeneity. Kasahara, Schrimpf, and Suzuki (2023) studied unobserved heterogeneity in production function estimation in a finite-mixture framework and provided nonparametric identification results under a fixed number of time periods.

The remainder of the paper is organized as follows. Section 2 describes the model and the estimation procedure.Section 3 studies the theoretical properties of the proposed method. Section 4 compares the proposed multi-dimensional clustering method to the standard one-dimensional clustering method in a Monte Carlo simulation design for the production function estimation. The empirical analysis is presented in Section 5. Finally, Section 6 concludes. Proofs, derivations and additional results for the Monte Carlo experiment, and details on the construction of the data set used for the empirical analysis are collected in an Online Appendix.

Throughout the paper, we adopt the following notations. For vectors a, b, we use (a, b) to denote $(a', b')'$, unless the dimension is defined otherwise. Let $||A||$ denote the Frobenius norm of a matrix A. When A is symmetric, let $\mu_{\max}(A)$ and $\mu_{\min}(A)$ denote the largest and smallest eigenvalues of A. Let $1\{\cdot\}$ denote the indicator function.

2 The Nonlinear GMM Framework

We first present a framework of nonlinear GMM estimation of heterogeneous coefficients with multiple unknown group patterns. For the ease of presentation, we focus on a setup

with two group memberships. The generalization to more group memberships is straightforward, as done in the empirical application. Our model assumes that all the heterogeneous coefficients are group-specific, excluding individual-specific parameters. The model explicitly incorporates homogeneous parameters that are shared among all individuals. The model specification and estimation are presented in Section 2.1, a production function example is provided in Section 2.2, and implementation details are discussed in Section 2.3.

2.1 Model Specification and Estimation

We have panel data $\{w_{it} : i = 1, ..., N; t = 1, ..., T\}$ and use them to estimate the unknown parameters $\theta_i = (a_i, b_i, \lambda) \in R^{d_\alpha + d_\beta + d_\lambda}$ based on moment conditions. We envision N to be significantly larger than T such that it is difficult to obtain accurate estimates of θ_i based on the times series $\{w_{it} : t = 1, \ldots, T\}$ alone. We consider a parsimonious model of a_i and b_i with group patterns. Instead of assigning each individual i to one group membership and requiring (a_i, b_i) to depend on this membership, we allow each coefficient to have its own unknown membership. Let $g_i \in \{1, ..., n_g\}$ denote the membership for a_i and $h_i \in \{1, ..., n_h\}$ denote the membership for b_i. Then

$$a_i = \begin{cases} \alpha_1 & \text{if } g_i = 1 \\ \vdots & \vdots \\ \alpha_{n_g} & \text{if } g_i = n_g \end{cases} \quad \text{and} \quad b_i = \begin{cases} \beta_1 & \text{if } h_i = 1 \\ \vdots & \vdots \\ \beta_{n_h} & \text{if } h_i = n_h \end{cases}. \tag{2.1}$$

Let

$$\alpha = (\alpha_1, \ldots, \alpha_{n_g}) \in R^{d_\alpha d_{n_g}} \text{ and } \beta = (\beta_1, \ldots, \beta_{n_h}) \in R^{d_\beta d_{n_h}} \tag{2.2}$$

denote the group-specific values. We can write

$$a_i = \alpha(g_i) \quad \text{and} \quad b_i = \beta(h_i), \tag{2.3}$$

where $\alpha(g_i) = \alpha_{g_i}$ denotes the subvector of α associated with the g_i^{th} group, and similarly, $\beta(h_i) = \beta_{h_i}$ denotes the subvector of β associated with the h_i^{th} group. With the two-dimensional group pattens, the unknown parameters are

$$\theta = (\alpha, \beta, \lambda),\ G = (g_1, \ldots, g_N),\ H = (h_1, \ldots, h_N). \tag{2.4}$$

The parameter space is $(\theta, G, H) \in \overline{\Theta} \times \Gamma_G \times \Gamma_H$, where $\overline{\Theta} = A^{n_g} \times B^{n_h} \times \Lambda$, and Γ_G and Γ_H are sets of all possible partitions of $\{1, \ldots, N\}$ into n_g and n_h groups, respectively. We assume n_g and n_h are known for now. In practice, they can be selected by the quasi Bayesian information criterion (BIC) presented in (2.9) below.

We assume group patterns and moment conditions hold for the true values of the parameters. For each i, let g_i^0 and h_i^0 denote the true group memberships and $\theta_i^0 = (\alpha^0(g_i^0), \beta^0(h_i^0), \lambda^0)$ denote the true value for $\theta_i = (\alpha(g_i), \beta(h_i), \lambda)$. For some known finite-dimensional function $m(w_{it}; \cdot) \in R^{d_m}$, the following moment conditions hold:

$$\mathbb{E}\left[m\left(w_{it}; \theta_i^0\right)\right] = 0, \quad \text{for all } i \text{ and } t. \tag{2.5}$$

We consider the GMM estimator

$$\left(\widehat{\theta}, \widehat{G}, \widehat{H}\right) = \underset{(\theta,G,H)\in\overline{\Theta}\times\Gamma_G\times\Gamma_H}{\arg\min} \widehat{Q}(\theta, G, H), \quad \text{where} \tag{2.6}$$

$$\widehat{Q}(\theta, G, H) = N^{-1} \sum_{i=1}^{N} \widehat{Q}_i(\theta, g_i, h_i) \quad \text{and}$$

$$\widehat{Q}_i(\theta, g_i, h_i) = \left[T^{-1} \sum_{t=1}^{T} m\left(w_{it}; \alpha(g_i), \beta(h_i), \lambda\right)\right]' W_{iNT} \left[T^{-1} \sum_{t=1}^{T} m\left(w_{it}; \alpha(g_i), \beta(h_i), \lambda\right)\right]$$

for some weighting matrix W_{iNT}.[4] The individual GMM criterion $\widehat{Q}_i(\theta, g_i, h_i)$ is a quadratic form of the sample analog of the moment condition (2.5) for unit i, obtained by taking a time series average. The quadratic forms are then averaged across i to obtain the estimation objective function $\widehat{Q}(\theta, G, H)$.

Note that the criterion function $\widehat{Q}(\theta, G, H)$ is invariant to relabeling the group memberships in (θ, G, H). Without loss of generality, we assume $(\widehat{\theta}, \widehat{G}, \widehat{H})$ is already suitably relabeled such that we can show below in Section 3 that $\widehat{\alpha} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{n_g})$ and $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{n_h})$ are consistent estimators of $\alpha^0 = (\alpha_1^0, \ldots, \alpha_{n_g}^0)$ and $\beta^0 = (\beta_1^0, \ldots, \beta_{n_g}^0)$, respectively, and that the classification is consistent.

Given the group membership $\widehat{G}$ and $\widehat{H}$, we re-estimate $\theta^0 = (\alpha_1^0, \ldots, \alpha_{n_g}^0, \beta_1^0, \ldots, \beta_{n_h}^0, \lambda^0)$

[4] Fernandez-Val and Lee (2013), Su, Shi, and Phillips (2016), among others, use the same type of criterion in the presence of heterogeneous parameters for panel data.

in a second step by minimizing a pooled GMM criterion[5]

$$\widetilde{\theta} = \arg\min_{\theta \in \overline{\Theta}} \widetilde{Q}(\theta), \text{ where } \widetilde{Q}(\theta) = \widetilde{m}(\theta)' W_{NT} \widetilde{m}(\theta), \text{ with} \tag{2.7}$$
$$\widetilde{m}(\theta) = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} m\left(w_{it}; \alpha(\widehat{g}_i), \beta(\widehat{h}_i), \lambda\right),$$

and W_{NT} is a weighting matrix which could depend on $\widehat{G}$ and $\widehat{H}$. In a linear instrumental variable model with clustered coefficients, Su, Shi, and Phillips (2016) showed that a pooled two-step estimator like $\widetilde{\theta}$ is preferred to a one-step estimator obtained simultaneously with the classification algorithm, because the latter typically is less efficient and suffers from asymptotic bias. Our Monte Carlo simulations also confirm that the two-step estimator $\widetilde{\theta}$ has better finite-sample properties in a nonlinear GMM problem designed for the production function estimation.

The number of groups, n_g and n_h, can be determined based on a (quasi) BIC. Let $\widehat{\Omega}$ be a consistent estimator of the asymptotic covariance matrix Ω of $\widetilde{m}(\widetilde{\theta})$, derived in Section 3.2; see (3.15). The GMM criterion with the optimal weighting matrix is

$$\widetilde{Q}(n_g, n_h) = \widetilde{m}(\widetilde{\theta})' \widehat{\Omega}^{-1} \widetilde{m}(\widetilde{\theta}), \tag{2.8}$$

where we make it clear that $\widetilde{m}(\theta)$ and $\widehat{\Omega}$ are constructed with classification based on n_g and n_h groups for α and β, respectively. A BIC for the problem is

$$BIC(n_g, n_h) = \widetilde{Q}(n_g, n_h) + (n_g d_\alpha + n_h d_\beta + d_\lambda) \frac{\log(NT)}{NT}. \tag{2.9}$$

In practice, we choose (n_g, n_h) to minimize $BIC(n_g, n_h)$ with $1 \leq n_g \leq g_{\max}$ and $1 \leq n_h \leq h_{\max}$ for some user-selected upper bounds $g_{\max}$ and $h_{\max}$. In the current setup, $(NT)\,\widetilde{Q}(n_g, n_h)$ is an analog of the log-likelihood, and BIC in (2.9) is a natural choice for selecting the number of clusters.[6]

[5]The objective function corresponds to the averaged moment condition $N^{-1}\Sigma_{i=1}^{N}\mathbb{E}\left[m\left(w_{it}; \theta_i^0\right)\right] = 0$, which would not be suitable to estimate the group memberships.

[6]Besides BIC, a wide range of penalty coefficients can deliver model selection consistency, as shown by Bonhomme and Manresa (2015), Su, Shi, and Phillips (2016), and others for clustering problems. A testing procedure to determine the number of clusters is provided by Lu and Su (2017).

2.2 A Production Function Example

The following model will serve as a data generating process (DGP) for the Monte Carlo analysis in Section 4. A more elaborate version will be used for the empirical analysis in Section 5. The production function takes the form

$$y_{it} = b_i^0 v_{it} + \omega_{it} + \varepsilon_{it}, \tag{2.10}$$

where y_{it} is the observed log output and v_{it} represents log variable inputs (including labor, intermediate inputs, materials, etc), ω_{it} is an unobserved productivity shock that is known to the firm, and ε_{it} is a zero-mean unobserved output shock that is realized after the factor input has been chosen optimally to maximize profits. The productivity shock ω_{it} follows an AR(1) process

$$\omega_{it} = a_i^0 + \rho^0 \omega_{it-1} + \xi_{it}, \tag{2.11}$$

where the zero-mean innovation ξ_{it} is uncorrelated with input choices prior to period t. The output shock ε_{it} is uncorrelated with the variable input choice at period t and before.

Heterogeneity in the input elasticity b_i^0 is modeled with a group structure. Moreover, instead of relying on firm fixed effects, we also model productivity heterogeneity a_i^0 across firms with a separate group structure. In empirical applications that include the capital stock as factor of production (e.g., the model that we estimate in Section 5), using firm fixed effects often leads to unusually low capital elasticity estimates in applications; Ackerberg, Caves, and Frazer (2015). The production function literature theorizes that fixed effect estimators that rely on the within-transformation amplify the highly-persistent measurement error noise in capital stock. The GMM estimation of our group-specific parameters avoids this concern.

To control for the endogeneity of the production inputs when estimating elasticities, the literature has focused on two types of estimation strategies. The first strategy is based on an observable proxy variable that can be used to control for unobserved productivity and is known as the Olley and Pakes (1996) method. The second is the dynamic panel method that quasi-differences the production function and uses lagged input choices as instruments. The proxy variable approach relies on observables that are monotonic in the firm's productivity. Examples include capital investment, e.g., Olley and Pakes (1996), and material inputs, e.g., Levinsohn and Petrin (2003). The recent literature has moved toward material inputs as proxy, assuming that the capital stock, labor, and material inputs are observed separately to achieve identification; see Ackerberg, Caves, and Frazer (2015). Unfortunately, the data

set used for the empirical analysis does not contain material inputs as a separate firm-level variable. Thus, we use dynamic panel method in this example, which will be the basis for the Monte Carlo experiment in Section 4, and the empirical application in Section 5.

Let

$$\Delta y_{it}(\rho) = y_{it} - \rho y_{it-1},\ \Delta v_{it}(\rho) = v_{it} - \rho v_{it-1} \tag{2.12}$$

denote the differencing terms given the parameter ρ. Then we have

$$\Delta y_{it}(\rho^0) - a_i^0 - b_i^0 \Delta v_{it}(\rho^0) = \xi_{it} + \left(\varepsilon_{it} - \rho^0 \varepsilon_{it-1}\right). \tag{2.13}$$

Let z_{it} denote a vector comprising variable input choices prior to period t and the constant term. This ensures that z_{it} is uncorrelated with the right hand side of (2.13). We obtain the moment condition

$$\mathbb{E}\left[z_{it}\left(\Delta y_{it}(\rho^0) - a_i^0 - b_i^0 \Delta v_{it}(\rho^0)\right)\right] = 0. \tag{2.14}$$

With the two-dimensional group membership g_i and h_i for a_i and b_i, respectively, we have

$$\begin{aligned} m\left(w_{it}; \theta_i\right) &= z_{it}\left(\Delta y_{it}(\rho) - a_i - b_i \Delta v_{it}(\rho)\right), \\ a_i = \alpha(g_i), &\quad b_i = \beta(h_i). \end{aligned} \tag{2.15}$$

The common parameter is $\lambda = \rho$. This model will be used in the Monte Carlo experiment in Section 4. In the empirical estimation in Section 5, we consider a trans-log specification of the production function with capital, additional quadratic terms of the regressors, and a time trend added to (2.10).

2.3 Implementation Details

In practice, we compute the GMM estimator in (2.6) by Lloyd's Algorithm, as in Bonhomme and Manresa (2015), and generalize the iteration to multiple types of memberships. Given G and H, $\widehat{\theta}$ is a GMM estimator based on $\widehat{Q}(\theta, G, H)$. Given θ and H, we minimize the GMM criterion function with respect to G to obtain the group membership estimator $\widehat{G}$. After re-estimating θ and holding G fixed, the group memberships H are also determined by the GMM criterion function. In the subsequent description of the algorithm, M is a large number ensuring that the algorithm does not terminate after one iteration and ϵ is a number close to zero that characterizes the tolerance level for improvements in the objective function. The computation details is listed in algorithm 1.

Algorithm 1 Lloyd's Algorithm

Initialization, $s = 0$**:** Provide an initial guess $(\widehat{G}^{(0)}, \widehat{H}^{(0)})$. Let $\widehat{Q}^{(0)} = M$.
Iterations, $s = 1, 2, 3, \ldots$ **:**

1. Using the last iteration's estimate of group memberships $(\widehat{G}^{(s-1)}, \widehat{H}^{(s-1)})$, estimate the parameter θ: $\widehat{\theta} = \arg\min_{\theta \in \overline{\Theta}} \widehat{Q}\big(\theta, \widehat{G}^{(s-1)}, \widehat{H}^{(s-1)}\big)$.
2. For $i = 1, \ldots, N$, determine the g-group membership: $\widehat{g}_i^{(s)} = \arg\min_{g_i \in \{1,\ldots,n_g\}} \widehat{Q}_i\big(\widehat{\theta}, g_i, \widehat{h}_i^{(s-1)}\big)$.
3. Re-estimate the parameter θ: $\widehat{\theta}^{(s)} = \arg\min_{\theta \in \overline{\Theta}} \widehat{Q}\big(\theta, \widehat{G}^{(s)}, \widehat{H}^{(s-1)}\big)$.
4. For $i = 1, \ldots, N$, determine the h-group membership: $\widehat{h}_i^{(s)} = \arg\min_{h_i \in \{1,\ldots,n_h\}} \widehat{Q}_i\big(\widehat{\theta}, \widehat{g}_i^{(s)}, h_i\big)$.
5. Assess convergence: let $\widehat{Q}^{(s)} = \widehat{Q}\big(\theta^{(s)}, \widehat{G}^{(s)}, \widehat{H}^{(s)}\big)$ and stop if $\big|\widehat{Q}^{(s)} - \widehat{Q}^{(s-1)}\big| \leq \epsilon$.

To resolve the label indeterminacy, we assume that $\alpha_1 < \alpha_2 < \ldots < \alpha_{n_g}$ and $\beta_1 < \beta_2 < \ldots < \beta_{n_h}$. Because the objective function of the algorithm does not depend on the labeling, we relabel the groups such that the inequality restrictions are satisfied, after it has converged. For the Monte Carlo experiment in Section 4 we initialized the algorithm at the true membership indicators G^0 and H^0. For the empirical analysis we executed the optimization conditional on multiple starting values $(\widehat{G}^{(0)}, \widehat{H}^{(0)})$. One of the starting values was generated by using K-means clustering based on log output y_{it}, and the observed production inputs. The remaining starting points were generated by drawing membership indicators independently from a uniform distribution. The results reported in Section 5 are based on the optimization associated with the lowest value of the objective function.

3 Asymptotic Properties

In this section, we study the asymptotic properties of the nonlinear GMM estimator and classification of group memberships. Section 3.1 provides a set of assumptions and shows that the one-step GMM estimator $\widehat{\theta}$ in (2.6) is consistent on average across i and in estimating the group-specific parameters. However, to consistently estimate the parameters for each individual, we need the consistent classification results in Section 3.2. Section 3.2 also establishes the asymptotic distribution of the recommended two-step estimator $\widetilde{\theta}$ in (2.7). We verify all the general assumptions with specific conditions for the production function

estimation example. Among all the regularity assumptions, we highlight that the proposed method with multiple memberships is robust to sparse interactions across the groups for different coefficients.

3.1 Consistent Estimation under Sparse Interactions

First, we assume the following identification and regularity conditions.

Assumption ID. For any η, $\inf_N \min_{1\leq i\leq N} \inf_{||\theta_i-\theta_i^0||>\eta} ||\mathbb{E}[m(w_{it};\theta_i)]|| > 0$.

Assumption R. (i) $\{w_{it}, t = 1, 2, \ldots\}$ are independently distributed across i. For each i, $\{w_{it} : t = 1, 2, \ldots, \}$ is stationary strong mixing with mixing coefficients $\boldsymbol{\alpha}_i(\cdot)$, where $\boldsymbol{\alpha}(\cdot) = \sup_N \max_{1\leq i\leq N} \boldsymbol{\alpha}_i(\cdot)$ satisfies $\boldsymbol{\alpha}(\tau) \leq c_\alpha r^\tau$ for some $c_\alpha > 0$ and $r \in (0, 1)$.

(ii) The true value θ_i^0 lies in the interior of the convex compact set $\Theta = A \times B \times \Lambda$ for all i.

(iii) There exists a function $f(w_{it})$ such that $\sup_{\theta_i\in\Theta} ||m(w_{it};\theta_i)|| \leq f(w_{it})$ and $||m(w_{it},\theta_i) - m(w_{it},\overline{\theta}_i)|| \leq f(w_{it})||\theta_i - \overline{\theta}_i||$ for all $\theta_i, \overline{\theta}_i \in \Theta$. $\mathbb{E}|f(w_{it})|^q < \infty$ for some $q \geq 6$.

Assumption NT. $N = O(T^{q/2-1})$, where $q \geq 6$ is the constant in Assumption R(iii).

Assumption W. There exists nonrandom matrices W_i such that $\max_{1\leq i\leq N} \|W_{iNT} - W_i\| = o_p(1)$ and $\inf_N \min_{1\leq i\leq N} \mu_{\min}(W_i) > 0$ and $\sup_N \max_{1\leq i\leq N} \mu_{\max}(W_i) < \infty$.

These assumptions are comparable to Assumptions A1 and A2 of Su, Shi, and Phillips (2016). Assumption ID ensures that θ_i^0 is identifiable based on the moment condition (2.5). Assumption R and NT comprise various regularity conditions to guarantee that the sample moment function converges to the population moment function for each unit i uniformly over the parameter space Θ and over units $i = 1, \ldots, N$ at a desired rate. The observations are assumed to be cross-sectionally independent, and the temporal dependence is controlled by a mixing condition. The moment function $m(w_{it}, \theta_i)$ is uniformly (in $\theta_i \in \Theta$) bounded by the function $f(w_{it})$ and $f(w_{it})$ also serves as a Lipschitz bound. We assume that the q^{th} moment of $f(w_{it})$ exists, where $q \geq 6$.[7] The asymptotic results are obtained as N and T pass to infinite jointly. According to Assumption NT, the larger q, the more slowly the time series dimension of the panel can grow. Finally, Assumption W ensures that for each unit i the sequence of weight matrices is convergent.

[7] Alternatively, one can also impose tail condition on $f(w_{it})$ directly, as in Bonhomme and Manresa (2015) and Liu, Shang, Zhang, and Zhou (2020).

Under Assumptions R and NT, Su, Shi, and Phillips (2016) established in their Lemma S1.2(ii) the uniform convergence result

$$\max_{1\leq i\leq N}\mathbb{P}\left\{\sup_{\theta_i\in\Theta}\left\|T^{-1}\sum_{t=1}^{T}m\left(w_{it};\theta_i\right)-\mathbb{E}[m\left(w_{it};\theta_i\right)]\right\|\geq\eta\right\}=o(N^{-1}) \tag{3.1}$$

for any $\eta>0$, as $N,T\to\infty$. To show the classification consistency for the memberships among N units, the $o(N^{-1})$ rate is useful.

To consistently estimate of the group specific parameters $\alpha^0=(\alpha_1^0,\ldots,\alpha_{n_g}^0)$ and $\beta^0=(\beta_1^0,\ldots,\beta_{n_g}^0)$, Assumption S states that each group is well separated from the rest and each group size is a non-degenerate portion of the whole population.

Assumption S. (i) For all $g\neq\widetilde{g}$, $h\neq\widetilde{h}$, $||\alpha_g^0-\alpha_{\widetilde{g}}^0||^2>c$ and $||\beta_h^0-\beta_{\widetilde{h}}^0||^2>c$ for some $c>0$. (ii) $N^{-1}\Sigma_{i=1}^n 1\{g_i^0=g\}\to\pi_g>0$ and $N^{-1}\Sigma_{i=1}^n 1\{h_i^0=h\}\to\psi_h>0$ for all $g\in\{1,\ldots,n_g\}$ and $h\in\{1,\ldots,n_h\}$.

Assumption S(ii) allows for *sparse interactions* between two types, i.e.,

$$N^{-1}\sum_{i=1}^{N}1\{g_i=g\text{ and }h_i=h\}\to 0\quad\text{for some }(g,h). \tag{3.2}$$

One can handle the two-dimensional clustering model with the one-dimensional method by calling $\{i:g_i=g\text{ and }h_i=h\}$ a cluster. However, the standard one-dimensional method does not allow for sparse interactions, because the number of observations in this intersection cluster is too small compared to larger clusters. The two-dimensional clustering method solves this problem because we estimate $\alpha(g_i)$ with all observations that share the membership g_i, regardless of h_i. The same argument holds for the estimation of $\beta(h_i)$.

Lemma 3.1 *Suppose Assumptions ID, R, NT, W hold. Then,*

$$N^{-1}\sum_{i=1}^{N}\left(\widehat{\alpha}(\widehat{g}_i)-\alpha^0(g_i^0)\right)^2\to_p 0,\quad N^{-1}\sum_{i=1}^{N}\left(\widehat{\beta}(\widehat{h}_i)-\beta^0(h_i^0)\right)^2\to_p 0,\quad\widehat{\lambda}\to_p\lambda^0.$$

Lemma 3.2 *Suppose Assumptions ID, R, NT, W, and S hold. Then,*

$$\widehat{\theta}\to_p\theta^0,\quad\textit{i.e.,}\quad\widehat{\alpha}\to_p\alpha^0,\quad\widehat{\beta}\to_p\beta^0,\quad\widehat{\lambda}\to_p\lambda^0.$$

It is worth pointing out that $N^{-1}\Sigma_{i=1}^{N}(\widehat{\alpha}(\widehat{g}_i) - \alpha^0(g_i^0))^2$ in Lemma 3.1 and $||\widehat{\alpha} - \alpha^0||^2$ in Lemma 3.2 are two different measures between the estimator and the true value. The former is based on $\widehat{\alpha}(\widehat{g}_i)$, where the group membership $\widehat{g}_i$ could be possibly misclassified. The later $\widehat{\alpha}$ does not consider the group membership classification.

Production Function Example of Section 2.2 (Continued). To verify these assumptions (and Assumption E below), we assume the following conditions hold for the production function estimation example discussed above.
(i) $\{(v_{it}, \xi_{it}, \varepsilon_{it}) : t = 1, 2, \ldots\}$ are independently distributed over i. For each i, $\{(v_{it}, \xi_{it}, \varepsilon_{it}) : t = 1, 2, \ldots\}$ is stationary strong mixing that satisfies Assumption R(i). Note that $(\xi_{it}, \epsilon_{it})$ are exogenous shocks and one would typically assume that they are also i.i.d. across time. The variable input choice v_{it} is endogenous and depends indirectly on ξ_{it} through the AR(1) process ω_{it} and the production function parameters θ_i. It also depends on product demand and factor prices, which have not been explicitly modeled at this stage.
(ii) $\theta_i = (a_i, b_i, \rho) \in \Theta = A \times B \times [0, \overline{\rho}]$ for some $\overline{\rho} < 1$, where $A, B \in R$ are both convex and compact. The true value θ_i^0 is in the interior of Θ. The assumption $0 \leq \rho < 1$ ensures that ω_{it} is stationary which is necessary for v_{it} to satisfy the mixing assumption.
(iii) Let $e_{it} = \omega_{it} + \epsilon_{it}$ and define $x_{it}(\rho) = (1, \Delta v_{it}(\rho), e_{it-1})$. Identification in this model depends on

$$\mathbb{E}[m_\theta\left(w_{it}; \theta_i^0\right)] = -\mathbb{E}[z_{it}x_{it}'(\rho^0)]. \tag{3.3}$$

Thus, we require that $\mathbb{E}[z_{it}x_{it}'(\rho^0)]$ has full rank.
(iv) Consider the bound

$$\begin{aligned}
\|m(w_{it}; \theta_i)\| &\leq \|z_{it}y_{it}\| + |\rho| \cdot \|z_{it}y_{it-1}\| + |a_i| \cdot \|z_{it}\| + |b_i| \cdot \|z_{it}v_{it}\| + |\rho| \cdot |b_i| \cdot \|z_{it}v_{it-1}\|, \\
\|m_\theta(w_{it}; \theta_i)\| &\leq \|z_{it}\| + \cdot\|z_{it}v_{it}\| + |\rho| \cdot \|z_{it}v_{it-1}\| + \|z_{it}y_{it-1}\| + |b_i| \cdot \|z_{it}v_{it-1}\|.
\end{aligned} \tag{3.4}$$

Because Θ is compact, there exists a finite constant M such that we can define the bounding function $f(w_{it})$ as

$$f(w_{it}) = M\big(\|z_{it}\| + \|z_{it}y_{it}\| + \|z_{it}y_{it-1}\| + \|z_{it}v_{it}\| + \|z_{it}v_{it-1}\|\big). \tag{3.5}$$

(v) Let $d_{it} = (1, y_{it}, y_{it-1}, v_{it}, v_{it-1})$. Let q_* be the largest q such that $\mathbb{E}||z_{it}d_{it}'||^q \leq \infty$. Then Assumption NT with q replaced by q_* provides the slowest rate at which the time series dimension T can grow in relative to the cross-sectional dimension N. Note that the production function implicitly determines the moments of y_{it} as a function of the moments

of $(v_{it}, \xi_{it}, \varepsilon_{it})$.

(v) Consider $W_{iNT} = (T^{-1}\Sigma_{t=1}^{T} z_{it} z_{it}')^{-1}$. It corresponds to the optimal weighting matrix if the conditional variance of the shocks are constant over time, although it may vary across i. For this choice of W_{iNT}, Assumption W holds by (3.1) and the condition $\mathbb{E}[z_{it} z_{it}']$ has full rank and $\mathbb{E}||z_{it}||^2 < \infty$. □

3.2 Classification and Asymptotic Distribution

Given $\widehat{\theta}$, $\widehat{G}$ and $\widehat{H}$ are K-mean type estimators of the group memberships that minimize the nonlinear GMM criterion function $Q(\widehat{\theta}, G, H)$. Bonhomme and Manresa (2015) provide consistency of the K-mean type classification method based on linear least squares estimation. We extend such classification consistency to nonlinear GMM problems with endogeneity and incorporate multiple-dimensional classification. Before presenting the formal result, we first illustrate the intuition and key arguments. For the ease of notation in subsequent arguments, write

$$m_{it}(\theta, g, h) = m(w_{it}; \alpha(g), \beta(h), \lambda), \tag{3.6}$$

for any $g \in \{1, \ldots, n_g\}$ and $h \in \{1, \ldots, n_h\}$. Because $\widehat{\theta} \to_p \theta_0$, it is sufficient to consider $\widehat{\theta} \in N_\eta = \{\theta \in \Theta : ||\theta - \theta_0|| \leq \eta\}$ for some positive number η. Given $\widehat{\theta}$, for any $(g_i, h_i) \neq (g_i^0, h_i^0)$, we have

$$\mathbb{P}\left\{\widehat{g}_i = g_i, \widehat{h}_i = h_i\right\} \leq \mathbb{P}\left\{\widehat{Q}_i(\widehat{\theta}, g_i, h_i) < \widehat{Q}_i(\widehat{\theta}, g_i^0, h_i^0)\right\}. \tag{3.7}$$

Because the limit of the weighting matrix has bounded eigenvalues by Assumption W,

$$\begin{aligned} \widehat{Q}_i(\widehat{\theta}, g_i, h_i) &\geq c_1 \left\| \frac{1}{T} \sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i) \right\|^2, \\ \widehat{Q}_i(\widehat{\theta}, g_i^0, h_i^0) &\leq c_2 \left\| \frac{1}{T} \sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i^0, h_i^0) \right\|^2 \end{aligned} \tag{3.8}$$

for some positive constants c_1 and c_2, with probability approaching 1. To bound the probability of misspecifying the membership of i to (g_i, h_i), it is therefore sufficient to bound

$$P_{i,gh}(\widehat{\theta}) = P\left\{ c_1 \left\| \frac{1}{T} \sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i) \right\|^2 \leq c_2 \left\| \frac{1}{T} \sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i^0, h_i^0) \right\|^2 \right\}. \tag{3.9}$$

With a decomposition,

$$\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i) = \underbrace{\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i) - \mathbb{E}[m_{it}(\widehat{\theta}, g_i, h_i)]}_{\text{noise}} + \underbrace{\mathbb{E}[m_{it}(\widehat{\theta}, g_i, h_i)]}_{\text{signal}}, \tag{3.10}$$

where (i) the first term on the right hand side is an $o_p(1)$ *noise* term and (ii) the second term $\mathbb{E}[m_{it}(\widehat{\theta}, g, h)]$ is a *signal* term that is strictly positive. This positive signal for misspecified group memberships is ensured by the separability condition in Assumption S and the identification condition in Assumption ID. By a similar decomposition for $T^{-1}\Sigma_{t=1}^{T} m_{it}(\widehat{\theta}, g_i^0, h_i^0)$ as in (3.10), we can show that (i) the noise is also $o_p(1)$ and (ii) the signal term $\mathbb{E}[m_{it}(\widehat{\theta}, g_i^0, h_i^0)]$ is close to 0 with $\widehat{\theta} \in N_\eta$ because $\mathbb{E}[m_{it}(\theta^0, g_i^0, h_i^0)] = 0$. We can show that, under Assumption R and NT, the probability of the noise terms being larger than the positive signal term converges to 0 at rate $o(N^{-1})$ uniformly, following results as in (3.1). Therefore, we have $P_{i,gh}(\widehat{\theta})$ converges to 0 at rate $o(N^{-1})$ uniformly and the whole group can be classified consistently. The result is presented in Theorem 3.3 below and its formal proof is given in the Appendix.

Theorem 3.3 *Suppose Assumptions ID, R, NT, W, S hold.*

$$\mathbb{P}\left\{\widehat{G} = G^0 \text{ and } \widehat{H} = H^0\right\} \to 1 \text{ as } N, T \to \infty,$$

where $G^0 = \{g_1^0, \ldots, g_N^0\}$ *and* $H^0 = \{h_1^0, \ldots, h_N^0\}$ *are the true memberships.*

Under Theorem 3.3, the proposed estimator $\widetilde{\theta}$ has the same asymptotic distribution as the oracle estimator, which is defined analogous to $\widetilde{\theta}$ but imposing the true memberships G^0 and H^0. Thus, we derive the asymptotic distribution of $\widetilde{\theta}$ by studying this oracle estimator. We first look at the first order derivative of the moment conditions. We assume that the function $m(w_{it}, \cdot)$ is differentiable in all parameters. Define

$$m_\theta\left(w_{it}; \theta_i^0\right) = \left[\frac{\partial}{\partial \alpha} m\left(w_{it}; \theta_i^0\right) : \frac{\partial}{\partial \beta} m\left(w_{it}; \theta_i^0\right) : \frac{\partial}{\partial \lambda} m\left(w_{it}; \theta_i^0\right)\right] \in R^{d_m \times \left(d_\alpha n_g + d_\beta n_h + d_\lambda\right)}, \tag{3.11}$$

where

$$\begin{aligned}
\frac{\partial}{\partial \alpha} m\left(w_{it}; \theta_i^0\right) &= \left[\frac{\partial}{\partial \alpha_1} m\left(w_{it}; \theta_i^0\right) : \cdots : \frac{\partial}{\partial \alpha_{n_g}} m\left(w_{it}; \theta_i^0\right)\right] \in R^{d_m \times (d_\alpha n_g)}, \\
\frac{\partial}{\partial \beta} m\left(w_{it}; \theta_i^0\right) &= \left[\frac{\partial}{\partial \beta_1} m\left(w_{it}; \theta_i^0\right) : \cdots : \frac{\partial}{\partial \beta_{n_h}} m\left(w_{it}; \theta_i^0\right)\right] \in R^{d_m \times \left(d_\beta n_h\right)}.
\end{aligned} \tag{3.12}$$

Under the group structure, $m(w_{it}, \theta_i^0)$ does not depend on α_g for $g \neq g_i^0$ or β_h for $h \neq h_i^0$. Thus, we have

$$\begin{aligned}
\frac{\partial}{\partial \alpha_g} m\left(w_{it}; \theta_i^0\right) &= 1\left\{g_i^0 = g\right\} \frac{\partial}{\partial a_i} m\left(w_{it}; a_i^0, b_i^0, \lambda^0\right) \text{ for } g = 1, \ldots, n_g, \\
\frac{\partial}{\partial \beta_h} m\left(w_{it}; \theta_i^0\right) &= 1\left\{h_i^0 = h\right\} \frac{\partial}{\partial b_i} m\left(w_{it}; a_i^0, b_i^0, \lambda^0\right) \text{ for } h = 1, \ldots, n_h.
\end{aligned} \tag{3.13}$$

The Jacobian matrix is

$$J = \lim_{N \to \infty} N^{-1} \sum_{i=1}^{N} \mathbb{E}\left[m_\theta\left(w_{it}; \theta_i^0\right)\right]. \tag{3.14}$$

The covariance of the moment condition is

$$\Omega = \lim_{N \to \infty} \lim_{T \to \infty} N^{-1} \sum_{i=1}^{N} \Omega_{iT}(\theta_i^0), \text{ where} \tag{3.15}$$

$$\Omega_{i_T}(\theta_i^0) = T^{-1} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbb{E}\left[m\left(w_{it}; \theta_i^0\right) m\left(w_{is}; \theta_i^0\right)'\right]. \tag{3.16}$$

We add the following regularity condition to derive the distribution of $\widetilde{\theta}$.

Assumption E. (i) $m(w_{it}, \theta_i)$ is differentiable in $\theta_i \in \Theta$; J and Ω exist and both have full rank.

(ii) $W_{NT} \to_p W$ for some full rank matrix W as $N, T \to \infty$.

(iii) Assumption R(iii) holds with $m(w_{it}; \theta_i)$ replaced by $m_\theta(w_{it}; \theta_i)$ and Θ replaced by a neighborhood around θ^0.

Theorem 3.4 *Suppose Assumptions ID, R, NT, W, S, E hold. Then,*

$$\sqrt{NT}\left(\widetilde{\theta} - \theta^0\right) \to_d N(0, V), \text{ where } V = \left(J'WJ\right)^{-1} J'W\Omega W \left(J'WJ\right)^{-1}.$$

In the estimation, α_g only shows up in the moment function $m(w_{it}; \alpha(\widehat{g}_i), \beta(\widehat{h}_i), \lambda)$ if $\widehat{g}_i = g$, i.e., individuals whose coefficient a_i belong to the g^{th} group. However, the estimator $\widehat{\alpha}_g$ also depends on individuals in other groups through the joint estimation of β and λ.

Production Function Example of Section 2.2 (Continued). The estimators $(\widehat{\theta}, \widehat{G}, \widehat{H})$ and $\widetilde{\theta}$ require the choice of weighting matrices W_{iNT}, $i = 1, \ldots, N$, and W_{NT}, respectively.

A two-step GMM approach uses a preliminary weighting matrix in the first stage and a consistent estimate of the optimal weighting matrix in the second stage.

The optimal weighting matrix for $(\widehat{\theta}, \widehat{G}, \widehat{H})$ is given by $\Omega_{i_T}^{-1}(\theta_i^0)$; see (3.16). Recall from Section 2.2 that

$$m(w_{it}; \theta_i^0) = z_{it}(\xi_{it} + \epsilon_{it} - \rho^0 \epsilon_{it-1}) = z_{it} u_{it}. \tag{3.17}$$

The vector of instruments z_{it} typically comprises a constant and lags of the production inputs, e.g., $z_{it} = (1, v_{it-1}, v_{it-2}, \ldots)'$. For the instruments to be valid, z_{it} has to be uncorrelated with $(\xi_{it}, \epsilon_{it}, \epsilon_{it-1})$. Suppose that $s \leq t-2$, then

$$\begin{aligned} \mathbb{E}\left[m\left(w_{it}; \theta_i^0\right) m\left(w_{is}; \theta_i^0\right)'\right] &= \mathbb{E}\left[z_{it} z_{is}' (\xi_{it} + \epsilon_{it} - \rho^0 \epsilon_{it-1})(\xi_{is} + \epsilon_{is} - \rho^0 \epsilon_{is-1})\right] \\ &= \mathbb{E}\left[\mathbb{E}_s\left[z_{it}(\xi_{it} + \epsilon_{it} - \rho^0 \epsilon_{it-1})\right] z_{is}' (\xi_{is} + \epsilon_{is} - \rho^0 \epsilon_{is-1})\right] \\ &= 0. \end{aligned} \tag{3.18}$$

Now define

$$\widehat{u}_{it} = \Delta y_{it}(\widehat{\rho}) - \widehat{\alpha}(\widehat{g}_i) - \widehat{\beta}(\widehat{h}_i) \Delta v_{it}(\widehat{\rho}), \tag{3.19}$$

to obtain the following estimator for Ω_{iT}:

$$\widehat{\Omega}_{iT} = T^{-1} \sum_{t=1}^{T} z_{it} z_{it}' \widehat{u}_{it}^2 + \frac{1}{T} \sum_{t=2}^{T} \left(z_{it} z_{it-1}' + z_{it-1} z_{it}'\right) \widehat{u}_{it} \widehat{u}_{it-1}, \tag{3.20}$$

which allows us to set $W_{iNT} = \widehat{\Omega}_{iT}^{-1}$. The optimal weighting matrix for $\widetilde{\theta}$ is given by Ω^{-1}; see (3.15). The matrix Ω can be consistently estimated by

$$\widehat{\Omega} = N^{-1} \sum_{i=1}^{N} \widehat{\Omega}_{iT}, \tag{3.21}$$

and we let $W_{NT} = \widehat{\Omega}^{-1}$.

In order to obtain an estimate of the asymptotic covariance matrix V of $\widetilde{\theta}$ in Theorem 3.4 we require an estimate of the Jacobian J in (3.14), which is a function of $\mathbb{E}[m_\theta\left(w_{it}; \theta_i^0\right)]$, previously given in (3.3). Define $\widehat{e}_{it} = y_{it} - \widehat{\beta}(\widehat{h}_i) v_{it}$ such that

$$\widehat{J} = -(NT^{-1}) \sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} (1, \Delta v_{it}(\widehat{\rho}), \widehat{e}_{it-1})'. \tag{3.22}$$

In applications with less structure one can replace the previously derived estimators of $\widehat{\Omega}_{iT}$ and $\widehat{\Omega}$ by a heteroskedasticity and autocorrelation consistent (HAC) covariance estimator, see Newey and West (1987) and Andrews (1991). In our Monte Carlo experiment and the empirical application we use a set of instruments that lead to exact identification which means that there is no need to construct optimal weighting matrices. However, to conduct inference on $\widetilde{\theta}$ the estimators $\widehat{\Omega}$ and $\widehat{J}$ are still required. □

4 Monte Carlo Experiment

In the Monte Carlo experiment we repeatedly simulate observations on production inputs and outputs for a panel of firms, to assess the small sample properties of our proposed estimator. The Monte Carlo design is based on a simplified version of the production function estimated in the empirical analysis: we abstract from capital as a factor of production and use a log-linear functional form. We compare the accuracy of the multi-dimensional clustering estimator to that of an estimator that only clusters in a single dimension. The data generating process (DGP) and the GMM moment conditions are described in Section 4.1, the DGP parameterization and simulation design are described in Section 4.2, and the results are summarized in Section 4.3.

4.1 Data Generating Process and Moment Conditions

The production function is identical to the one used in Section 2.2. Log output as a function of the log variable inputs is given by (2.10) and the unobserved productivity shock ω_{it} evolves according to the AR(1) law of motion (2.11). We make the following distributional assumptions about the innovations:

$$\xi_{it} \overset{iid}{\sim} N(0, \sigma_\xi^2), \quad \epsilon_{it} \overset{iid}{\sim} N(0, \sigma_\epsilon^2) \tag{4.1}$$

with the understanding that ϵ_{it} is an *ex post* productivity shock that is realized after the firm makes its period t input decisions.

In order to generate a variable input series v_{it} that is internally consistent with the production function, we add more structure to the economic environment in which the firms

are operating and derive the profit-maximizing input choice. Each firm i faces a downward-sloping demand curve of the form

$$P_{it}(Q) = \exp(\eta_{it})Q^{-\kappa}, \tag{4.2}$$

where the demand shifter η_{it} evolves according to

$$\eta_{it} = d_i + \phi\eta_{it-1} + \nu_{it}, \quad d_i \overset{iid}{\sim} N(\mu_d, \sigma_d^2), \quad \nu_{it} \overset{iid}{\sim} N(0, \sigma_\nu^2). \tag{4.3}$$

The firm-specific intercept d_i generates heterogeneity in the average level of demand and the innovations ν_{it} trigger fluctuations over time.

The state variables for the firms' decision problem are the demand shifter η_{it} and the *ex ante* productivity process ω_{it}. Because the state variables are exogenous, conditional on knowing (η_{it}, ω_{it}) the firms solve the static problem

$$\begin{aligned} \max_{V_{it}, Q_{it}} \quad & \mathbb{E}^{\epsilon_{it}}\left[P_{it}(Q_{it})Q_{it} \,\middle|\, \eta_{it}, \omega_{it}\right] - V_{it} \\ \text{s.t.} \quad & Q_{it} = \exp(\omega_{it} + \epsilon_{it})V_{it}^{b_i}, \end{aligned} \tag{4.4}$$

where the expectation is taken over ϵ_{it} and $V_{it} = \exp(v_{it})$. The first-order condition for this optimization determines the law of motion for the variable input V_{it} as a function of the exogenous processes η_{it} and ω_{it}. We show in the Online Appendix that v_{it} follows an ARMA(2,1) process. To generate a single sample (y_{it}, v_{it}), $i = 1, \ldots, N$ and $t = 1, \ldots, T$, we simulate the model based on the law of motion of η_{it} in (4.3), the law of motion of ω_{it} in (2.11), the optimal variable input choice v_{it} provided in the Online Appendix, and the loglinear production function (2.10). The initial values η_{i0} and ω_{i0} are drawn from their respective stationary distributions.

The estimation proceeds as outlined in Section 2.2. The moment conditions are given by (2.14) with $z_{it} = (1, v_{it-1}, v_{it-2})$. Instrument validity follows from the model implication that factor inputs v_{it-h}, $h = 1, 2$, do not depend on *ex post* productivity shocks and are also independent of *ex ante* productivity shock innovations dated t and later. Because y_{it-1} is a function of v_{it-1} instrumental relevance is satisfied with respect to the regressor y_{it-1}. Moreover, because v_{it} follows an ARMA(2,1) process, the instruments are also correlated with the regressor $v_{it} - \rho^0 v_{it-1}$. The moment conditions exactly identify the parameter vector $\theta^0 = (\alpha^0, \beta^0, \rho^0)$.

Table 1: Parameterization of DGP

Para	Value
Variable cost elasticity b_i	$\{0.2, 0.5, 0.8\}$
Ex ante productivity: intercept a_i	$\{-6, -3, 0\}$
Ex ante productivity: AR coefficient ρ	0.64
Ex ante productivity: innovation std. dev. σ_ξ	1
Ex post productivity: innovation std. dev. σ_ϵ	0.01
Price elasticity of demand $-1/\kappa$	-3
Demand shifter: intercept mean μ_d	0.01
Demand shifter: intercept std. dev. σ_d	0.35
Demand shifter: AR coefficient ϕ	0.9
Demand shifter: innovation std. dev. σ_v	1

4.2 DGP Parameterization and Simulation Design

The parameterization of the DGP is partly based on features of the actual data that we are using in the empirical analysis. A summary of the parameter values is provided in Table 1. We consider three values each for the average level of productivity and the input elasticity: $a_i \in \{-6, -3, 0\}$ and $b_i \in \{0.2, 0.5, 0.8\}$. We let $\sigma_\xi = 1$. Thus, the standard 90% confidence intervals of

$$\frac{1}{T-1}\sum_{t=2}^{T}(\omega_{it} - \rho\omega_{it-1}) = a_i + \frac{1}{T-1}\sum_{t=2}^{T-1}\xi_{it} \tag{4.5}$$

overlap for different a_i types for $T = 2$ but not for $T = 10$. The negative price elasticity of demand is set to $1/\kappa = 3$.

We now turn to the calibration of the parameters ϕ and ρ. We set $\phi = 0.9$ and choose ρ to match the cross-sectional average of the first-order autocorrelation of the log variable input in the data set used in the empirical analysis, which is

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\widehat{\text{Cov}}_i(v_{it}, v_{it-1})}{\widehat{\mathbb{V}}_i(v_{it})} \approx 0.867. \tag{4.6}$$

We let $\sigma_\nu = 1$ and show in the Online Appendix how a model-implied formula for the autocorrelation coefficient can be solved for ρ, conditional on κ, ϕ, σ_ξ, and σ_ν. Based on this calculation, we set $\rho = 0.64$.

The standard deviation of the *ex post* productivity shock is set to $\sigma_\epsilon = 0.01$ such that the *ex ante* productivity shock dominates. Finally, we set $\mu_d = -0.01$, and $\sigma_d = 0.35$. Under

Table 2: Monte Carlo Designs - Group Sizes

	Design 1			Design 2			Design 3			Group Labels		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
α_1	100	100	100	40	130	130	220	40	40	SS	SM	SL
α_2	100	100	100	130	40	130	40	220	40	MS	MM	ML
α_3	100	100	100	130	130	40	40	40	220	LS	LM	LL

Notes: The table reports the cross-sectional sample size N for the various groups which are denoted by SS through LL, where S refers to "small", M to "medium", and L to "large" (in absolute value) a_i and b_i values; see Table 1. The α parameters determine the productivity and the β parameters determine variable cost elasticity. The time series dimensions considered in the Monte Carlo are $T \in \{4, 6, 8, 10, 12, 14, 16, 18, 20\}$.

Design 1 (see below) and $T = 20$, we are able to reproduce the empirical estimate based on our sample:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\widehat{\mathbb{V}}_i(v_{it})}{\widehat{\mathbb{V}}(v_{it})} \approx 0.109, \tag{4.7}$$

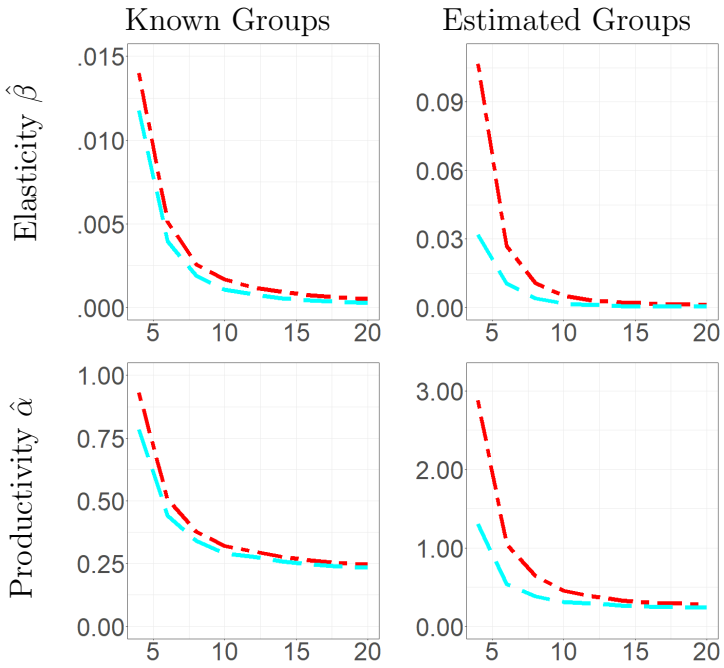
where $\widehat{\mathbb{V}}_i(\cdot)$ is the variance across time for unit i and $\widehat{\mathbb{V}}(\cdot)$ is the variance across i and t. For both the numerator and the denominator we demean variable costs for each i separately.

In terms of group size, we consider three designs which are summarized in Table 2. The total number of firms is always $N = 900$. The table reports the number of units in each combination of parameter groups. Under Design 1 all cells are of equal size. Under Design 2 the diagonal cells are sparsely populated relative to the off-diagonal cells and vice versa under Design 3. For each of these designs we report results for multiple choices of T, ranging from 4 to 20. Note that in Table 2, $\alpha_1 = 0$, $\alpha_2 = -3$, $\alpha_3 = -6$ are the three group-specific values for a_i and $\beta_1 = 0.2$, $\beta_2 = 0.5$, $\beta_3 = 0.8$ are the three group-specific values for b_i.

4.3 Monte Carlo Results

All results presented subsequently are based on $N_{sim} = 300$ Monte Carlo repetitions. We begin by examining the MSEs associated with the estimation of the group-specific parameters under Design 1. Recall that the variable input elasticity can take three group-specific values: $\beta_1 = 0.2$, $\beta_2 = 0.5$, and $\beta_3 = 0.8$. For each of these values we have an estimator $\hat{\beta}_j$. We use Monte Carlo averaging to approximate the MSE $\mathbb{E}[(\hat{\beta}_j - \beta_j)^2]$ and compute an average across j, weighted by the group size. The panels in the top row of Figure 1 show the MSE as a function of the time series dimension T of the panel for our proposed two-dimensional (2D)

Figure 1: MSE of Group-specific Parameters, Weighted by Group Size, Design 1

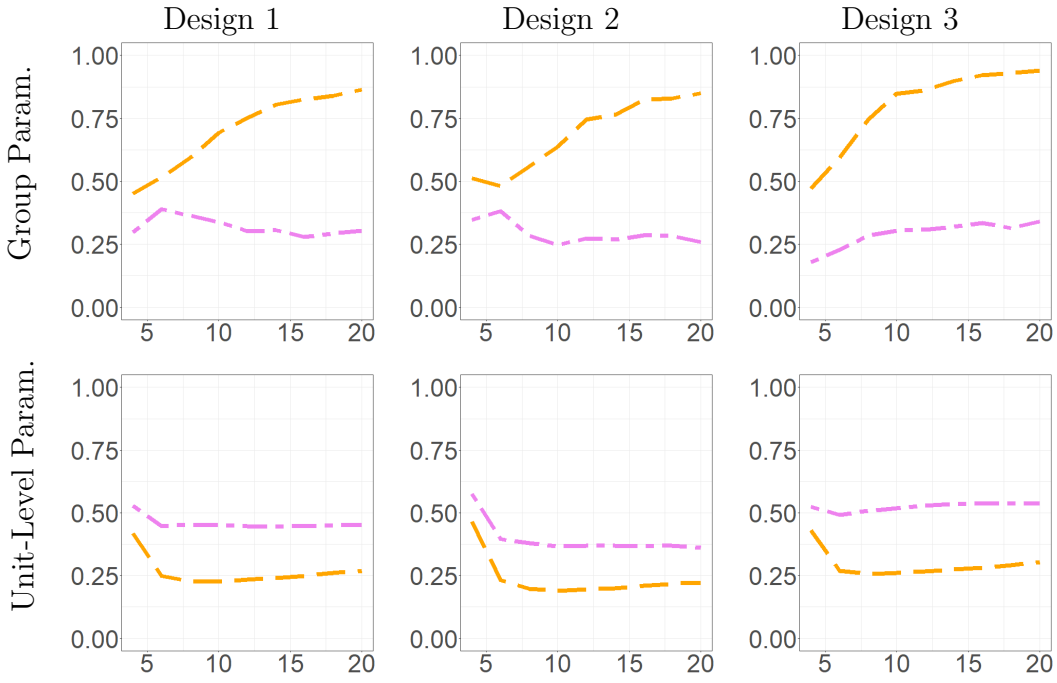


Notes: MSE associated with group-level parameter estimators averaged across groups, weighted by group size, as a function of T. Cyan dashed lines are 2D clustering. Red dashed-dotted lines are 1D clustering.

clustering estimator and the standard one-dimensional (1D) clustering estimator. MSEs for $\hat{\alpha}_j$ are displayed in the bottom row of the figure. The MSEs in all four panels are decreasing in the time-series dimension T of the panel. The larger T, the more precisely the unit-specific coefficients can be estimated by temporal averaging.

The results in the left column of Figure 1 are obtained under the assumption that the true group membership of each firm i is known. The 2D estimator is computed based on Step (c) of Algorithm 1. Due to the known group membership there is no need anymore to iterate. The 1D estimator partitions the sample of firms into nine bins, where each bin corresponds to a combination of β_j and α_k, delivering estimators $\hat{\beta}_{j|k}$ and $\hat{\alpha}_{k|j}$, $j = 1, 2, 3$ and $k = 1, 2, 3$. For the estimation of β_j, this leads to the MSEs $\mathbb{E}[(\hat{\beta}_{j|k} - \beta_j)^2]$, $k = 1, 2, 3$; and likewise for the estimation of α_k. The nine resulting MSEs are averaged by bin size. Because the groups under 2D clustering are larger than the bins under 1D clustering, the 2D estimator is associated with a smaller MSE. The MSE reductions for $\hat{\beta}$ range from 16 ($T = 4$) to 48 ($T = 20$) percent, whereas the MSE reductions for $\hat{\alpha}$ are between 5 ($T = 20$) and 16 percent ($T = 4$). For both parameters the MSE is dominated by a nonlinear GMM bias, which is reduced at a smaller rate through 2D clustering than the variance component.

Figure 2: MSE Ratios 2D/1D Clustering

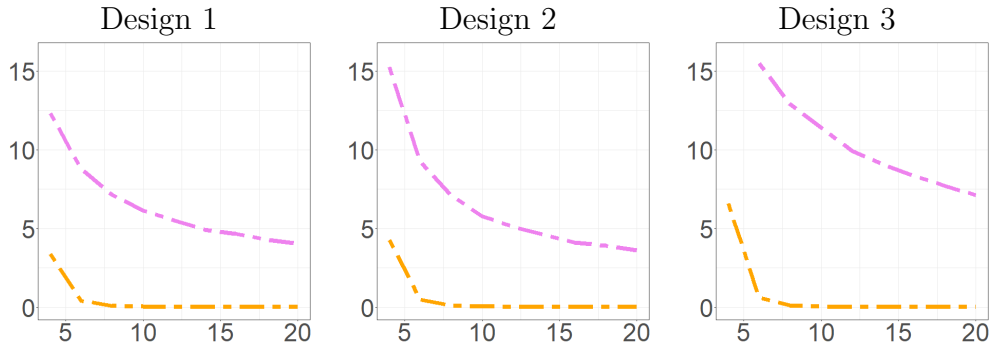


Notes: MSE ratios 2D/1D clustering. Magenta is elasticity β_i and b_i, and ocre is productivity α_i and a_i. Top row: MSEs for group-specific parameter estimates averaged across groups, weighted by group size. Bottom row: MSEs for unit-level parameter estimates averaged across i.

The panels in the right column of Figure 1 are obtained under the assumption that the group membership for each unit has to be estimated using the clustering algorithm. This drastically increases the MSEs and amplifies the benefit of multi-dimensional clustering. The top-left panel of Figure 2 shows the MSE ratio for 2D versus 1D clustering. A value below one indicates that 2D clustering yields a lower MSE than 1D clustering. For $T = 5$ this ratio is roughly 0.5 for the productivity values and 0.4 for the input elasticity values, implying an MSE reduction of 50 to 60 percent. For larger values of T the ratio for the productivity parameters stays roughly constant, whereas it increases to about 0.9 for the elasticity values. The percentage gain from multi-dimensional clustering is very similar across Monte Carlo designs.

The bottom panels of Figure 2 contain MSE ratios for the unit-level parameters, that is, in case of the elasticity, $N^{-1}\Sigma_{i=1}^{N}\mathbb{E}[(\hat{b}_i - b_i)^2]$. Here the 2D clustering leads to a larger MSE reduction for the productivity parameter (approximately 75%) than for the elasticity (approximately 50%). As a value of the sample size, the reduction stays fairly constant for values of T larger than 6.

Figure 3: Average Classification Errors (Percent)



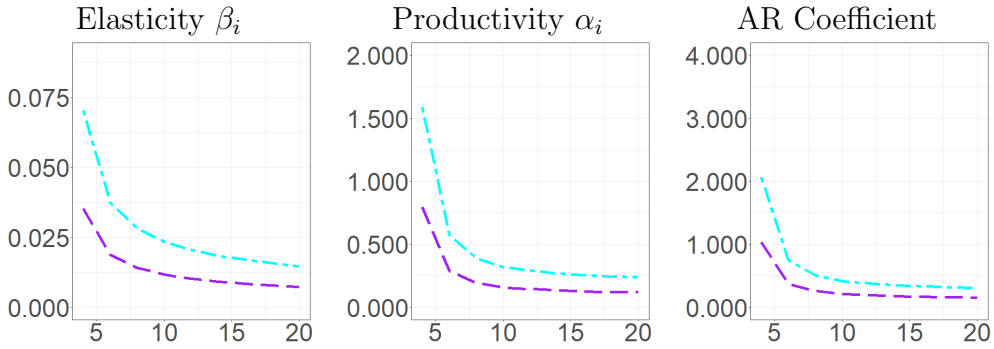
Notes: Magenta lines refer to elasticity b_i and ocre lines to productivity a_i classification errors (in percent) which are defined as the number of incorrectly classified firms divided by the total number of firms multiplied by 100.

Figure 3 shows the average classification errors. A firm i is counted as misclassified in terms of its elasticity, if the elasticity group membership is incorrectly estimated. The average classification errors are larger for the asymmetric Designs 2 and 3 than for the symmetric Design 1. Moreover, the classification errors are larger for the elasticities b_i than for the productivities a_i. We computed symmetric two-standard-deviation (of the 2D clustering estimator) intervals around the true b_i and a_i values. For the elasticity parameter values these bands overlap for $T = 4$ and $T = 20$. For the productivity values, on the other hand, there is no overlap, even for $T = 4$. The sharper estimates translate into smaller classification errors. Even though the classification errors depicted in Figure 3 are all below 15%, the MSE increase due to group membership estimation documented in Figure 1 is quite large. This indicates that the cost of misclassification is substantial. The parameter value differences generate a large bias in the estimates.

Thus far, we have focused on the performance of the one-step estimator, that jointly estimates the group membership indicators and model parameters (Algorithm 1). In Figure 4 we compare the MSE of the one-step estimator to the two-step-estimator, that re-estimates the model parameters conditional on the first-step membership estimates. The second-step estimation reduces the MSE by roughly 50%.

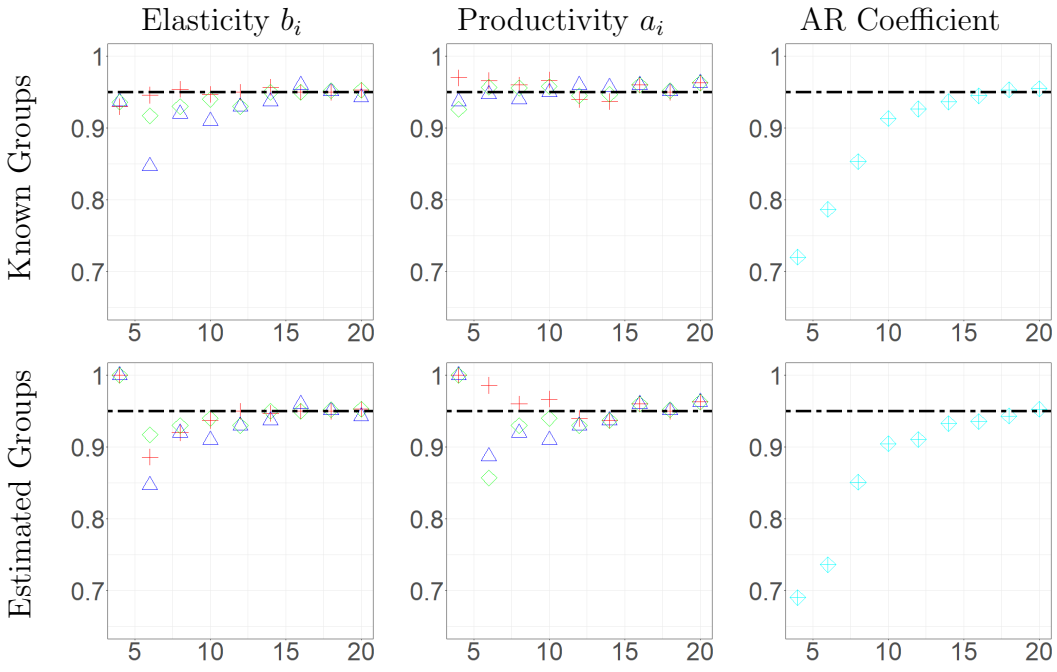
Finally, we turn from point to interval estimation and report coverage probabilities for 95% confidence intervals in Figure 5. The confidence intervals are based on the two-step estimator and its associated standard-error estimates. The horizontal line in each panel indicate the nominal coverage probability. The b_i and a_i panels show three different symbols,

Figure 4: One-Step versus Two-Step Estimation MSEs, Design 1



Notes: MSE associated with group-level parameter estimators averaged across groups, weighted by group size. MSE of homogeneous AR coefficient ρ. Cyan dashed-dotted lines are 2D one-step estimators. Purple dashed lines are 2D two-step estimators.

Figure 5: Coverage Probabilities



Notes: The horizontal lines indicate the nominal coverage probability of 95%. The three symbols in the left and center panels correspond to the three group-specific elasticity and productivity values.

corresponding to the three values that each parameter can take. The top row shows confidence intervals that are constructed under the assumption that the group membership is known. Deviations from the nominal coverage probability are due to small-sample behavior of the GMM estimator and unrelated to the clustering approach. For $T > 15$ the actual

coverage probability for all parameters is close to the nominal level. For smaller values of T the coverage probability of the AR coefficient and some of the elasticity values is less than the nominal level. When the true group memberships are replaced by the estimated group memberships, the discrepancy between actual and nominal coverage increases only slightly for $T < 15$, but essentially stays close to zero for $T \geq 15$.

Overall, the proposed 2D clustering estimator exhibits good small sample properties in the simulation experiment and sharper estimates than the 1D clustering approach.

5 Empirical Analysis

We now estimate firm-level trans-log production functions using our multi-dimensional clustering approach. Each firm is part of a sector s which we take to be a two-digit NAICS sector. The production function is given by

$$y_{it} = a_i + b_i k_{it} + c_i v_{it} + d_i v_{it}^2 + \zeta v_{it} k_{it} + \psi t + \omega_{it} + \epsilon_{it}, \tag{5.1}$$

where k_{it} is the capital stock and all variables are in logs. As in the Monte Carlo design of Section 4 we assume that the total factor productivity has two components. The component that is known to the firm *ex ante* follows the AR(1) process:

$$\omega_{it} = \rho \omega_{it-1} + \xi_{it}. \tag{5.2}$$

The *ex post* productivity shock ϵ_{it} is assumed to be independent over time. The intercept and the coefficients on capital, variable inputs, and squared variable inputs are group-specific. In addition to $\alpha(\cdot)$ and $\beta(\cdot)$, we define $\gamma(\cdot)$ and $\delta(\cdot)$ to characterize the group-specific values of c_i and d_i. Rather than introducing separate group structures for c_i and d_i, we assume that a third set of groups determines (c_i, d_i) jointly.

We use ℓ_i to indicate the group membership of unit i in the third dimension, $\widehat{L}$ to denote the collection of all memberships, and n_ℓ to denote the number of (c_i, d_i) groups. The interaction coefficient ζ and the time trend coefficient ψ are assumed to be homogeneous. As in Section 2, the production function is quasi-(ρ)-differenced to eliminate the serial correlation in ω_{it}. The GMM estimation is based on the instrument vector

$z_{it} = (1, k_{it}, k_{it-1}, v_{it-1}, v_{it-1}^2, v_{it-1}k_{it-1}, t)$ which leads to the moment conditions

$$\mathbb{E}[z_{it}(\epsilon_{it} - \rho\epsilon_{it-1} + \xi_{it})] = 0 \tag{5.3}$$

at the "true" parameter values.

Based on the estimated variable input elasticities we compute an estimate of the firms' markups. De Loecker and Warzynski (2012) show that if v_{it} induces no dynamic constraints in the firm's cost minimization problem and if the firm's capital is predetermined, then the markup can be expressed as a function of the revenue-to-variable-cost ratio

$$mu_{it} = \varphi_{it}\frac{p_{it}^y \exp[y_{it}]}{p_{it}^v \exp[v_{it}]}, \tag{5.4}$$

where p_{it}^y and p_{it}^v are firm-specific prices of the output and the variable input, respectively. Here, φ_{it} is the elasticity of output with respect to variable input. For the translog production function (5.1) this elasticity is given by

$$\varphi_{it} = c_i + 2d_i v_{it} + \zeta k_{it}. \tag{5.5}$$

5.1 Data Set, Model Specifications, and Estimation

As in Flynn, Gandhi, and Traina (2019) and De Loecker, Eeckhout, and Unger (2020), the firm-level data set is constructed from the Compustat Fundamentals (North America) database. We take a time period t to be one year. The firms' *Sales of Goods* and *Cost of Goods Sold* are used as output and variable input, respectively. The firms' capital stocks are calculated based on the perpetual inventory method using the *Net Property, Plant, and Equipment* series. Nominal variables are converted to real variables using the appropriate deflators. Our sample starts in 1999 and ends in 2019. Further details on data definitions, transformations, and subsample selection are provided in the Online Appendix.

The subsequent analysis is conducted for firms that are associated with the same two-digit NAICS sector. The estimation for each sector includes firms for which we have at least one observation between 1999 and 2019. There are 24 two-digit NAICS sectors. Following conventions in the literature on markup estimation, we exclude the following sectors from the subsequent analysis: Utilities (NAICS 22), Finance and Insurance (NAICS 52), Real Estate and Rental and Leasing (NAICS 53), Management of Companies and Enterprises (NAICS

Table 3: Two-Digit-Level Sectors Used in Estimation of Models with Group Heterogeneity

NAICS	Description
21	Mining, Quarrying, and Oil and Gas Extraction
23	Construction
31	Manufacturing (Food, Apparel, and other Consumer Goods)
32	Manufacturing (Paper, Wood, Petroleum, Chemical, and Non-Metallic Minerals Related)
33	Manufacturing (Furniture, Metal, Electronic, and Machinery Related)
42	Wholesale Trade
44	Retail Trade (Food, Apparel, Vehicles, and other Consumer Goods)
45	Retail Trade (Entertainment, Department Stores, Online, etc.)
48	Transportation
51	Information
54	Professional, Scientific, and Technical Services
56	Administrative and Support Services, etc.
62	Health Care and Social Assistance
72	Accommodation and Food Services

55), and Public Administration (NAICS 92). For these sectors, the cost minimization assumptions underlying (5.4) are not compelling. Five sectors (NAICS 11, 49, 61, 71, 81) have relatively few firms so that there are not enough observations in the cross section to estimate group-specific effects. We will estimate production functions for firms in these sectors by imposing homogeneity. The 14 sectors for which we estimate group-specific firm-level production functions are listed in Table 3.

5.2 Empirical Results

We begin with evidence of firm heterogeneity within two-digit industries, discuss estimation results for one of the manufacturing sectors (NAICS 33) in more detail, and then present summaries of the results across all sectors.

Model Selection. The first step of the empirical analysis is to determine the sector-specific degree of heterogeneity in the production function coefficients. To do so, we use the quasi-Bayesian information criterion introduced in (2.9). Because our panel is unbalanced we replace NT by the total number of observations in the sector-specific panel. We restrict the number of groups for a_i, b_i, and c_i to be identical, i.e., $n_g = n_h = n_\ell = n \leq 6$. Thus, for a model specification with n groups the number of parameters in the penalty term is $3n + 3$.

Table 4: Model Selection: BIC Values

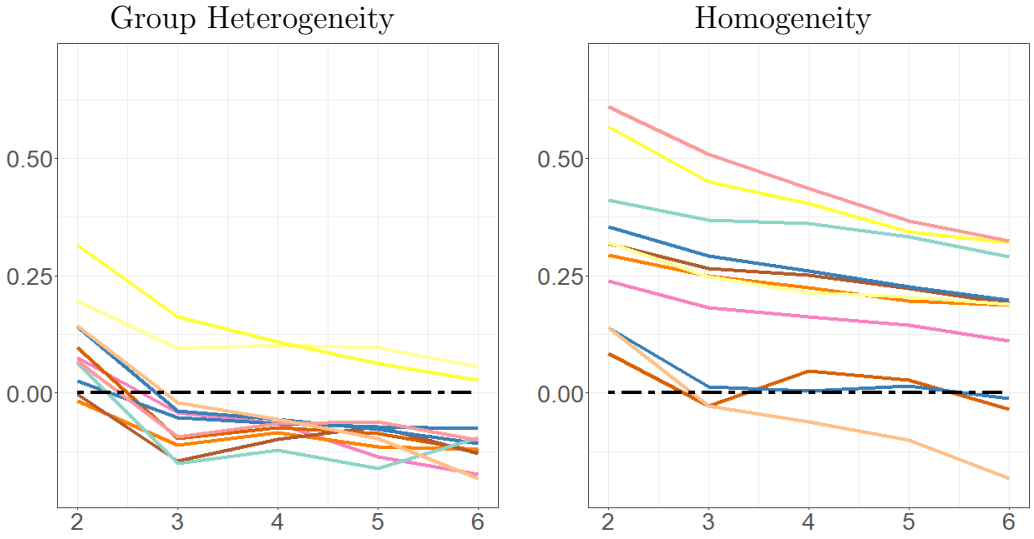
	Number of Groups n							
NAICS	1	2	3	4	5	6	$\hat{n}$	Nobs
21	188.66	126.28	**122.4**	122.75	122.95	130.12	3	544
23	31818.89	30701.5	28601.19	24809.54	10013.01	**7429.59**	6	139
31	1822.03	939.01	735.98	**513.06**	513.08	513.13	4	370
32	408.82	**180.91**	180.92	180.94	180.96	181.04	2	1428
33	178.14	121.11	**88.17**	88.2	88.22	88.21	3	2394
42	422.09	380.12	**261.89**	261.94	262.1	262.14	3	347
44	3974.77	3268.47	**2114.14**	2114.21	2114.3	2114.35	3	268
45	5106.68	4876.12	4351.22	3771.87	**3401.2**	3401.57	4	205
48	**305.722**	305.81	305.91	306.08	306.16	306.14	1	257
51	166.51	106.42	79.21	35.31	**21.12**	21.31	5	1651
54	**120.94**	121.02	121.08	121.14	122.21	122.11	1	576
56	159.81	115.33	**75.71**	75.8	75.87	75.92	3	265
62	321.03	248.76	160.34	**151.97**	152.12	152.21	4	224
72	421.26	348.82	**268.62**	268.7	268.86	268.97	3	245

Notes: All table entries are rounded. For each industry the minimal BIC value appears in bold face. The last column contains the selected number of groups.

Table 4 summarizes the results from applying the information criteria. The number of selected groups varies strongly across sectors, ranging from $\hat{n} = 1$ for NAICS 48 and 54 to $\hat{n} = n_{max} = 6$ for NAICS 23. In general, increasing the number of groups from $n = 1$ to the optimal value leads to a drastic reduction of BIC because the goodness of in-sample fit improves. Increasing n beyond the optimal value leads to comparatively small increases in the information criterion, which are mostly caused by the penalty term.

As a sanity check, we compute the residual autocorrelation functions (ACFs) for the estimated quasi-differenced sectoral production functions under the selected group heterogeneity specification and under a specification that imposes coefficient homogeneity. The results are plotted in Figure 6. If the model is correctly specified, the autocorrelation of order $h > 1$ should be zero. The left panel of the figure shows the ACF for the selected level of group heterogeneity. Each line corresponds to a different sector. The residual autocorrelations are generally close to zero. The right panel shows the ACF under the assumption that the coefficients are homogeneous. Here the residual autocorrelations are generally much higher, indicating misspecification. The autocorrelation patterns are consistent with the selection of multiple groups based on the BIC in Table 4.

Figure 6: Residual Autocorrelation in Quasi-differenced Production Function



Notes: The x-axis is the temporal shift. Each line corresponds to a sector.

Group Size Distribution. If the selected number of groups for each of heterogeneous parameter is n, then there exist n^3 parameter combinations (cells). Figure 7 provides some information about the number of firms associated with each parameter combination. We show Lorenz curves for each industry, indicating the fraction y of firms belonging the fraction x of smallest cells. For all but one industry, the cell size distribution is very skewed, meaning that there are many cells that contain very few observations. This implies that the alternative approach of using a one-dimensional clustering approach that assigns firms to the n^3 cells would have very little information to estimated the parameters of sparsely-populated cells.[8]

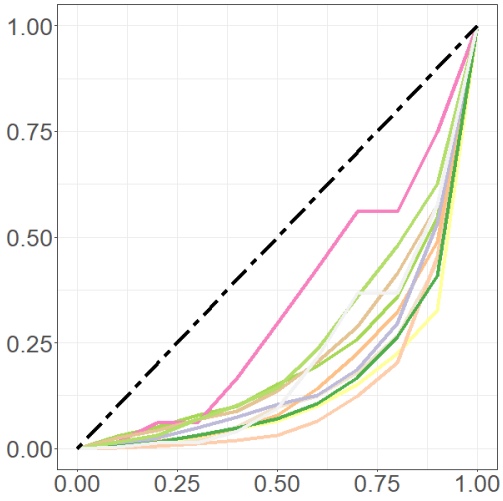
Parameter Estimates. Our empirical analysis generates a large amount of parameter estimates. Coefficient and standard error estimates for NAICS 33 are summarized in Table 5. There is a substantial difference in the level of productivity of Group 1, on the one hand, and Groups 2 and 3, on the other hand. The values for the capital coefficient b_i range from 0.139 to 0.491. Finally, there is also substantial heterogeneity in (c_i, d_i).

Elasticity and Markup Distribution Across Sectors. In the top-left panel of Figure 8 we plot quantiles of the cross-sectional distribution (across firms in the two-digit industries included in the analysis) of φ_{it}, the elasticity of output with respect to variable input, defined in (5.5).[9] The cross-sectional variation is due to the group-heterogeneity of the coefficients

[8]Granted, a selection criterion might elimate some of the sparsely-populated cells.

[9]To generate the top-row panels in Figure 8, we deleted firms with negative elasticity/markup estimates.

Figure 7: Cell Size Distribution Across Industries



Notes: Lorenz curves indicating the fraction y of firms belonging to the fraction x of smallest cells. Each solid hairline corresponds to a sector. Dashed-dotted line is the 45-degree line.

Table 5: Parameter Estimates NAICS 33

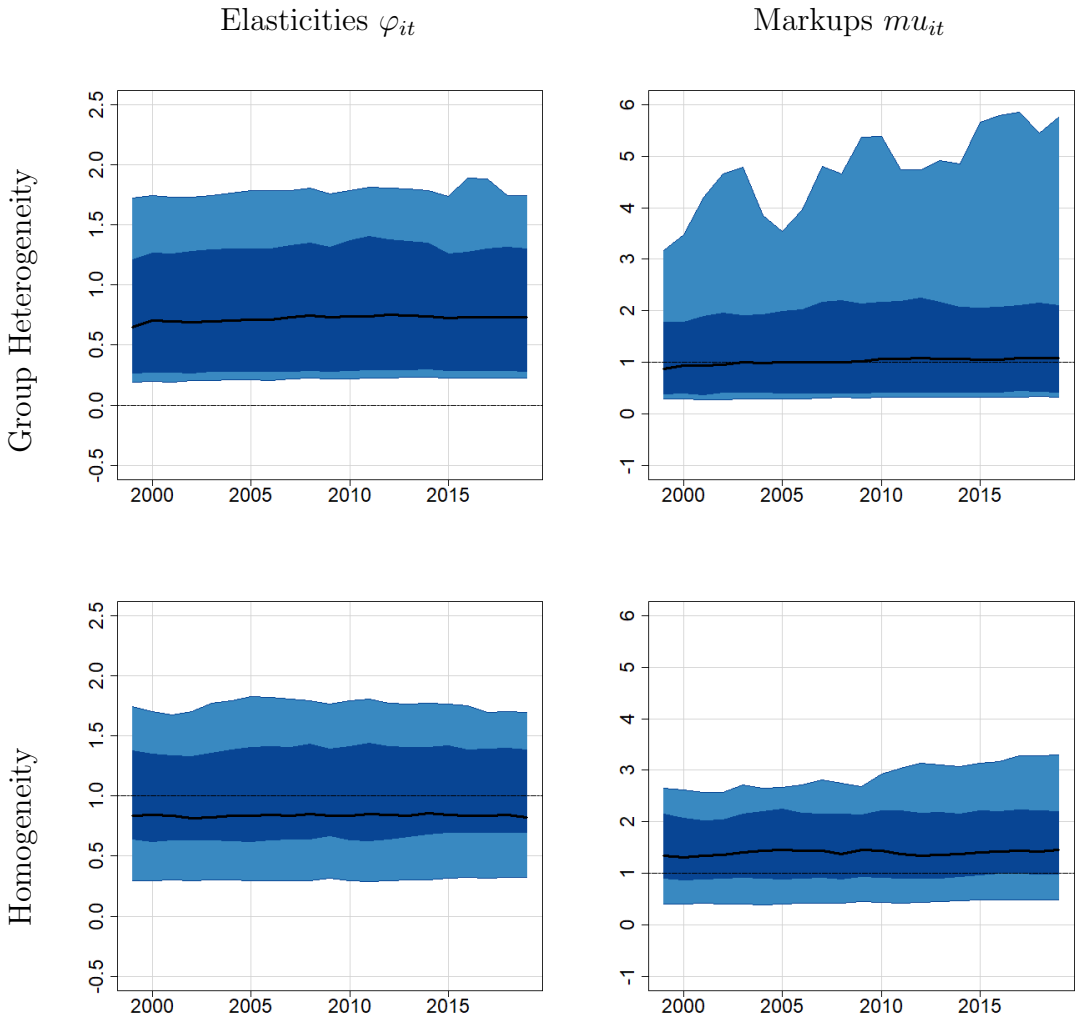
	Para. Set 1 Productivity	Para. Set 2 Capital	Para. Set 3 Var. [Input, Input^2]
Value 1	0.314 (0.133)	0.139 (0.099)	$\underset{(0.007)}{-1.875}, \underset{(0.003)}{0.009}$
Value 2	2.221 (0.240)	0.174 (0.026)	$\underset{(0.182)}{0.076}, \underset{(0.032)}{0.023}$
Value 3	2.270 (0.075)	0.491 (0.014)	$\underset{(0.004)}{0.412}, \underset{(0.001)}{0.525}$

(c_i, d_i) and the input choice heterogeneity (v_{it}, k_{it}). The time series variation in φ_{it} is solely due to fluctuations in (v_{it}, k_{it}). Most of the time series variation of the φ_{it} distribution is visible at the median which rises above 0.5 from 2007 to 2014 and is below 0.5 in the other years. The 10th and 90th percentiles, on the other hand, are fairly flat over time. For comparison, we plot in the bottom-left panel the φ_{it} elasticities based on estimates that impose parameter homogeneity within sector. The cross-sectional distribution of the φ_{it} values is more concentrated, because for each sector there is only a single parameter value. The median value of φ_{it} is approximately one and it is essentially time invariant.

The top-right panel of the figure shows the distribution of markups under group het-

We conjecture that for these firms the "variable" input is not flexible and the markup formula does not hold.

Figure 8: Distribution of Elasticities and Markups



Notes: The graphs depict the 10%, 25%, 50%, 75%, and 90% quantiles of the cross-sectional distributions of the estimated elasticities and markups across the firms in the two-digit sectors included in the analysis.

erogeneity. A value of 1 implies that the firm charges marginal costs. The median of the cross-sectional distribution is close to one. Most of the time variation is concentrated in the 90th percentile. Here mu_{it} increased from slightly over 3 to above 5. Because the 90th percentile of φ_{it} is flat, much of the increase is due to an increase in the revenue-to-variable-cost ratio for high φ_{it} elasticity firms. The markup distribution under parameter homogeneity, depicted in the bottom-right panel is less dispersed because of the lower variance of the φ_{it} distribution. At the median the markup is higher (around 50%) under homogeneity, which is consistent with the larger median (1.0 instead of approximately 0.5) of the φ_{it} distribution.

Aggregate Markup. In an influential paper, De Loecker, Eeckhout, and Unger (2020) documented that the aggregate markup has been steadily rising over the past six decades. The aggregate markup can be defined as the sales-share weighted average of firm-level markups:

$$mu_t = \sum_{i=1}^{N} \left(\frac{p_{it}^y \exp[y_{it}]}{\sum_{i=1}^{N} p_{it}^y \exp[y_{it}]} \right) mu_{it}. \tag{5.6}$$

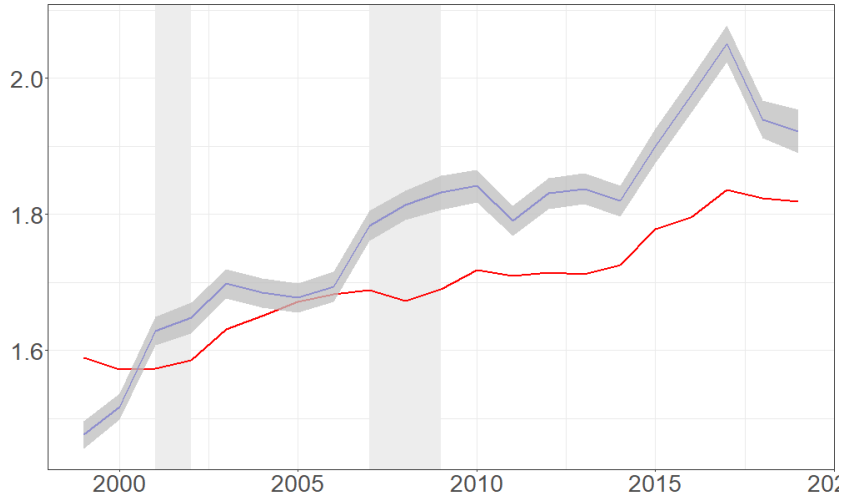
The authors estimated Cobb-Douglas production functions (in our notation $d_i = \zeta = 0$) for the two-digit NAICS sectors using five-year rolling windows to obtain the φ_{it}s. Endogeneity of the production inputs is handled through a variant of the proxy variable approach proposed by Levinsohn and Petrin (2003), which inverts the relationship between firms' input choices and productivity to replace ω_{it} in (5.1). De Loecker, Eeckhout, and Unger (2020) documented an increase of the aggregate markup from 1.25 in 1960 to 1.45 in the year 2000, and a subsequent rise to 1.6 by 2016.

Demirer (2022) considered a more general production function that also allows for labor-augmenting productivity. He extended the Olley and Pakes (1996) framework to multidimensional productivity and obtained markup estimates for US manufacturing that are generally lower than those obtained under the De Loecker, Eeckhout, and Unger (2020) production function specification. His markup estimate for the year 2000 is about 1.35 and it is only slightly higher in 2012.

De Loecker, Eeckhout, and Unger (2020) pointed out that Compustat does not report labor and material inputs separately. The two inputs are bundled together as cost of goods sold. The authors also highlight the difficulty to impute the material cost for the vast majority of firms, because Compustat reports wage bills for only a small percentage of firms. Thus, their main results use the sum of labor and material inputs as both a production function factor input and a proxy variable. This setup poses a challenge for the proxy variable approach to identify the production function, as shown in Gandhi, Navarro, and Rivers (2020). Demirer (2022) worked on a smaller subsample for which the wage bill information is available to impute the material inputs. We avoid this data issue by using the dynamic panel method over the proxy variable approach. However, we have to assume productivity is an autoregressive process instead of a more general Markovian process allowed by the proxy variable setup.

In Figure 9 we plot two aggregate markup time series based on our production function estimates: one is based on group heterogeneity in each sector and the other one based on

Figure 9: Evolution of Aggregate Markup



Notes: Aggregate markup is computed based on (5.6) with estimates obtained under group heterogeneity (blue line and grey bands) and within-sector homogeneity (red line).

within-sector homogeneity. The grey band captures 95% confidence intervals that reflect sampling uncertainty associated with the estimates of the production function parameters under group heterogeneity. Our estimates of φ_{it} are based on a single sample from 1999 to 2019. This longer sample facilitates the identification of coefficient heterogeneity. Despite the use of a single estimation sample, the aggregate markups are time varying for two reasons. First, the translog production function leads to time-varying input elasticities φ_{it}, see (5.5). Second, the firm-level markup depends on time-varying factor costs shares, see (5.4), and the aggregate markup depends on time-varying sales shared, see (5.6).

The group-heterogeneity elasticity estimates imply that the aggregate markup rises steadily from 1.54 in 2000 to 1.92 in 2015. It spikes above 2.08 in 2017 and then drops to approximately 1.95 in 2019. Under the assumption of within-sector homogeneity the rise in markup is less pronounced: in 2000 the markup is 1.57, slightly higher than under group heterogeneity, but it rises to only about 1.82 in 2019. We obtain a higher level of markup and a more rapid rise between 1999 and 2019 than De Loecker, Eeckhout, and Unger (2020) and Demirer (2022). In addition to using a somewhat different production function specification and allowing for more cross-sectional heterogeneity, we also use a different approach of controlling for endogeneity of the input choices: we quasi-difference the production function, whereas the other authors use the control function approach. Overall, we conclude that in our setting allowing for group heterogeneity within two-digit

NAICS sectors increases the estimated aggregate markup compared to the specification that imposes within-sector homogeneity.

6 Conclusion

Explicitly modeling and estimating heterogeneous parameters is an important development in the panel data literature. Our paper contributes to this literature by developing a nonlinear GMM framework that allows for multi-dimensional group heterogeneity. In this framework, each unit is associated with multiple groups, where each group is formed for a different unobserved characteristic of the unit. A feature of this approach is its robustness to sparse interactions of different characteristics. In the application, we cluster firms based on multiple unknown coefficients in a trans-log production function, which allows for heterogeneity in productivity and elasticities of output with respect to variable inputs and capital. In our application, we show that accounting for multi-dimensional group heterogeneity leads to higher estimates of the level and growth of aggregate markups than specifications that assume production technologies are homogeneous within two-digit NAICS sectors.

References

Ackerberg, D. A., K. Caves, and G. Frazer (2015): "Identification Properties of Recent Production Function Estimators," *Econometrica*, 83(6), 2411–2451.

Ando, T., and J. Bai (2016): "Panel Data Models with Grouped Factor Structure Under Unknown Group Membership," *Journal of Applied Econometrics*, 31(1), 163–191.

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59(3), 817–858.

Bester, C. A., and C. B. Hansen (2016): "Grouped effects estimators in fixed effects models," *Journal of Econometrics*, 190(1), 197 – 208.

Bonhomme, S., T. Lamadon, and E. Manresa (2022): "Discretizing Unobserved Heterogeneity," *Econometrica*, 90(2), 625–643.

Bonhomme, S., and E. Manresa (2015): "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 83(3), 1147–1184.

Chetverikov, D., and E. Manresa (2022): "Spectral and post-spectral estimators for grouped panel data models," *arXiv: 2212.13324*.

Cytrynbaum, M. (2020): "Blocked Clusterwise Regression," *arXiv: 2001.11130*.

De Loecker, J., J. Eeckhout, and G. Unger (2020): "The Rise of Market Power and the Macroeconomic Implications," *Quarterly Journal of Economics*, 135(2), 561–644.

De Loecker, J., and F. Warzynski (2012): "Markups and Firm-Level Export Status," *American Economic Review*, 102(6), 2437–2471.

Demirer, M. (2022): "Production Function Estimation with Factor-Augmenting Technology: An Application to Markups," *Manuscript, MIT Sloan*.

Fernandez-Val, I., and J. Lee (2013): "Panel Data Models with Nonadditive Unobserved Heterogeneity: Estimation and Inference," *Quantitative Economics*, 4(3), 453–481.

Flynn, Z., A. Gandhi, and J. Traina (2019): "Measuring Markups with Production Data," *SSRN Working Paper*, 3358472.

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

Gu, J., and S. Volgushev (2019): "Panel Data Quantile Regression with Grouped Fixed Effects," *Journal of Econometrics*, 213(1), 68 – 91.

Hahn, J., and H. R. Moon (2010): "Panel Data Models with Finite Number of Multiple Equilibria," *Econometric Theory*, 36(3), 863–881.

Henry, M., Y. Kitamura, and B. Salanie (2014): "Partial identification of finite mixtures in econometric models," *Quantitative Economics*, 5(1), 123–144.

Kasahara, H., P. Schrimpf, and M. Suzuki (2023): "Identification and Estimation of Production Function with Unobserved Heterogeneity," .

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

Ke, Y., J. Li, and W. Zhang (2016): "Structure identification in panel data analysis," *The Annals of Statistics*, 44(3), 1193 – 1233.

Krasnokutskaya, E., K. Song, and X. Tang (2022): "Estimating unobserved individual heterogeneity using pairwise comparisons," *Journal of Econometrics*, 226(2), 477–497.

Leng, X., H. Chen, and W. Wang (2023): "Multi-dimensional latent group structures with heterogeneous distributions," *Journal of Econometrics*, 233(1), 1–21.

Levinsohn, J., and A. Petrin (2003): "Estimating Production Functions Using Inputs to Control for Unobservables," *Review of Economic Studies*, 70(2), 317–341.

Lin, C., and S. Ng (2012): "Estimation of Panel Data Models with Parameter Heterogeneity when Group Membership is Unknown," *Applied Economics*, 1(1), 42–55.

Liu, L. (2023): "Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective," *Journal of Business & Economic Statistics*, 41(2), 349–363.

Liu, R., Z. Shang, Y. Zhang, and Q. Zhou (2020): "Identification and estimation in panel models with overspecified number of groups," *Journal of Econometrics*, 215(2), 574–590.

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

Newey, W. K., and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55(3), 703–708.

Olley, G. S., and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64(6), 1263–1297.

Su, L., Z. Shi, and P. C. Phillips (2016): "Identifying Latent Structures in Panel Data," *Econometrica*, 84(6), 2215–2264.

Sun, Y. X. (2005): "Estimation and Inference in Panel Structure Models," *Manuscript, University of California San Diego*.

Wang, W., and L. Su (2021): "Identifying latent group structures in nonlinear panels," *Journal of Econometrics*, 220(2), 272–295.

Zhang, B. (2023): "Incorporating Prior Knowledge of Latent Group Structure in Panel Data," *arXive Working Paper*, 2211.16714.

Online Appendix
Clustering for Multi-Dimensional Heterogeneity with an Application to Production Function Estimation

Xu Cheng, Frank Schorfheide, and Peng Shao

The Online Appendix consists of the following parts:

A. Proofs

B. Derivations and Additional Results for Monte Carlo

C. Data Construction for the Empirical Analysis

A Proofs

Proof of Lemma 3.1. Define the population criterion

$$\begin{aligned} Q_N(\theta, G, H) &= N^{-1}\sum_{i=1}^{N} Q_i(\theta, g_i, h_i), \text{ where} \\ Q_i(\theta, g_i, h_i) &= \mathbb{E}[m_{it}(\theta, g_i, h_i)]' W_i \mathbb{E}[m_{it}(\theta, g_i, h_i)]. \end{aligned} \tag{A.1}$$

Define $\delta_i(\theta, g_i, h_i) = \frac{1}{T}\Sigma_{t=1}^{T} m_{it}(\theta, g_i, h_i) - \mathbb{E}[m_{it}(\theta, g_i, h_i)]$. We can deduce from (3.1) that $\|\delta_i(\theta, g_i, h_i)\| = o_p(1)$ uniformly over i and $(\theta, g_i, h_i) \in \overline{\Theta} \times \Gamma_G \times \Gamma_H$, i.e., for any $\eta > 0$,

$$\begin{aligned} &\mathbb{P}\left\{\max_{1\le i\le N} \sup_{(\theta,g_i,h_i)\in\overline{\Theta}\times\Gamma_G\times\Gamma_H} \|\delta_i(\theta, g_i, h_i)\| \ge \eta\right\} \\ \le\ & \sum_{i=1}^{N} \mathbb{P}\left\{\sup_{(\theta,g_i,h_i)\in\overline{\Theta}\times\Gamma_G\times\Gamma_H} \|\delta_i(\theta, g_i, h_i)\| \ge \eta\right\} \\ \le\ & N \cdot \max_{1\le i\le N} \mathbb{P}\left\{\sup_{(\theta,g_i,h_i)\in\overline{\Theta}\times\Gamma_G\times\Gamma_H} \|\delta_i(\theta, g_i, h_i)\| \ge \eta\right\} \\ =\ & N \cdot o\left(\frac{1}{N}\right) = o(1). \end{aligned} \tag{A.2}$$

By Assumption W, $\|W_{iNT} - W_i\| = o_p(1)$ uniformly over i. Therefore, by the Slutsky's theorem, $|\widehat{Q}_i(\theta, g_i, h_i) - Q_i(\theta, g_i, h_i)| = o_p(1)$ uniformly over i and $(\theta, g_i, h_i) \in \overline{\Theta} \times \Gamma_G \times \Gamma_H$.

The uniform convergence of the individual criterion function implies convergence of the average criterion, i.e.,

$$\sup_{(\theta,G,H)\in\overline{\Theta}\times\Gamma_G\times\Gamma_H} |\widehat{Q}_N(\theta,G,H) - Q_N(\theta,G,H)| = o_p(1). \tag{A.3}$$

Define

$$d(\theta,G,H) = N^{-1}\sum_{i=1}^{N} d_i\left(\theta_i\right),$$

where

$$d_i(\theta_i) = \left(\alpha(g_i) - \alpha^0(g_i^0)\right)^2 + \left(\beta(h_i) - \beta^0(h_i^0)\right)^2 + ||\lambda - \lambda^0||^2.$$

We show that, for any $\delta > 0$, there exists $\varepsilon > 0$ such that for all N

$$\inf_{d(\theta,G,H)>\delta} Q_N(\theta,G,H) \geq \varepsilon. \tag{A.4}$$

Given that θ_i has a compact support Θ for all i, let $C = \sup_N \max_{1\leq i\leq N} \sup_{\theta_i\in\Theta} d_i(\theta_i) < \infty$. Let $S = \{i : d_i(\theta_i) > \delta/2\}$ and $N_S = \Sigma_{i=1}^N 1\{i \in S\}$. Note that $d_i(\theta_i) \leq C$ for $i \in S$ and $d_i(\theta_i) \leq \delta/2$ for $i \notin S$. Thus, $N_S C + (N - N_S)\delta/2 \geq N d(\theta,G,H) \geq N\delta$, which implies that $N_S \geq N\delta/(2C-\delta) > N\delta/(2C)$. Then,

$$\begin{aligned}\inf_{d(\theta,G,H)>\delta} Q_N(\theta,G,H) &\geq \inf_{d(\theta,G,H)>\delta} N^{-1}\sum_{i\in S} Q_i(\theta,g_i,h_i) \\ &\geq \frac{N_S}{N}\min_{i\in S} Q_i(\theta,g_i,h_i) \geq \frac{\delta}{2C}\varepsilon^*,\end{aligned} \tag{A.5}$$

where the last step holds because $\min_{i\in S} Q_i(\theta,g_i,h_i) \geq \varepsilon^*$ for some $\varepsilon^* > 0$ by Assumption ID and W. Thus, the identification condition for $Q_N(\theta,G,H)$ in (A.4) holds with $\varepsilon = \delta\varepsilon^*/(2C)$. This argument is analogous to that used to show Lemma A.1 of Liu, Shang, Zhang, and Zhou (2020).

Finally, we show the consistency result by combining (A.3) and (A.4). For any $\delta > 0$, there exists $\varepsilon > 0$, such that

$$\begin{aligned}\mathbb{P}\left\{d(\widehat{\theta},\widehat{G},\widehat{H}) > \delta\right\} \leq \mathbb{P}\left\{Q_N(\widehat{\theta},\widehat{G},\widehat{H}) \geq \varepsilon\right\} = \mathbb{P}\left\{d_a + d_b + d_c \geq \varepsilon\right\}, \text{ where} \\ d_a = Q_N(\widehat{\theta},\widehat{G},\widehat{H}) - \widehat{Q}_N(\widehat{\theta},\widehat{G},\widehat{H}), \\ d_b = \widehat{Q}_N(\widehat{\theta},\widehat{G},\widehat{H}) - \widehat{Q}_N(\theta^0,G^0,H^0), \\ d_c = \widehat{Q}_N(\theta^0,G^0,H^0) - Q_N(\theta^0,G^0,H^0).\end{aligned} \tag{A.6}$$

Because $d_b \leq 0$ by definition of the estimator, and both $d_a = o_p(1)$ and $d_c = o_p(1)$ by (A.3),

(A.6) implies that $\mathbb{P}\{d(\widehat{\theta}, \widehat{G}, \widehat{H}) > \delta\} \to 0$ for any $\delta > 0$. This completes the proof. □

Proof of Lemma 3.2. Given Lemma 3.1 and Assumption S, this Lemma follows from the same arguments used to show Lemma B.3 of Bonhomme and Manresa (2015). The arguments can be applied to α and β separately in our set-up. There is no need to take sample average over time here because our parameters are not time-varying. Lemma B.3 of Bonhomme and Manresa (2015) also shows how to relabel the groups and shows that this is a one-to-one mapping with probability approaching 1. □

Proof of Theorem 3.3. Let $E_W = 1\{\max_i \|W_{iNT} - W_i\| \leq \eta\}$ for some small positive constant η. Assumption W shows that $E_W = 1$ with probability approaching 1. Conditional on $E_W = 1$, for $(g_i, h_i) \neq (g_i^0, h_i^0)$, we have shown in (3.7)-(3.9) that

$$\begin{aligned}\mathbb{P}\left\{\widehat{g}_i = g_i, \widehat{h}_i = h_i\right\} &\leq \mathbb{P}\left\{\widehat{Q}_i(\widehat{\theta}, g_i, h_i) < \widehat{Q}_i(\widehat{\theta}, g_i^0, h_i^0)\right\} \\ &\leq \mathbb{P}\left\{c_1 \left\|\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i)\right\|^2 \leq c_2 \left\|\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i^0, h_i^0)\right\|^2\right\} \qquad \text{(A.7)}\end{aligned}$$

for constants $c_2 > c_1 > 0$. Using the decomposition in (3.10) and the triangle inequality,

$$\left\|\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i)\right\|^2 \geq \left\|b_i(\widehat{\theta}, g_i, h_i)\right\| - \left\|\delta_i(\widehat{\theta}, g_i, h_i)\right\|^2, \text{ where} \qquad \text{(A.8)}$$

$$\delta_i(\widehat{\theta}, g_i, h_i) = \frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i) - \mathbb{E}[m_{it}(\widehat{\theta}, g_i, h_i)],$$

$$b_i(\widehat{\theta}, g_i, h_i) = \mathbb{E}[m_{it}(\widehat{\theta}, g_i, h_i)].$$

By a similarly decomposition,

$$\left\|\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i^0, h_i^0)\right\|^2 \leq \left\|b_i(\widehat{\theta}, g_i^0, h_i^0)\right\| + \left\|\delta_i(\widehat{\theta}, g_i^0, h_i^0)\right\|^2. \qquad \text{(A.9)}$$

Below we analyze the four terms $\delta_i(\widehat{\theta}, g_i, h_i)$, $b_i(\widehat{\theta}, g_i, h_i)$, $\delta_i(\widehat{\theta}, g_i^0, h_i^0)$, $b_i(\widehat{\theta}, g_i^0, h_i^0)$.

For $\widehat{\theta} \in N_\eta = \{\theta \in \Theta : ||\theta - \theta_0||^2 \leq \eta^2\}$, we have

$$\begin{aligned}
\left\|b_i(\widehat{\theta}, g_i, h_i)\right\|^2 &= \left\|\mathbb{E}[m_{it}(\widehat{\theta}, g_i, h_i)] - \mathbb{E}[m_{it}(\theta^0, g_i^0, h_i^0)]\right\|^2 \\
&\geq b_{1,i}\left(\theta^0, g_i, h_i\right) - b_{2,i}\left(\widehat{\theta}, g_i, h_i\right), \text{ where} \\
b_{1,i}\left(\theta^0, g_i, h_i\right) &= \left\|\mathbb{E}[m_{it}\left(\theta^0, g_i, h_i\right)] - \mathbb{E}[m_{it}(\theta^0, g_i^0, h_i^0)]\right\|^2, \\
b_{2,i}\left(\widehat{\theta}, g_i, h_i\right) &= \left\|\mathbb{E}[m_{it}(\widehat{\theta}, g_i, h_i)] - \mathbb{E}[m_{it}(\theta^0, g_i, h_i)]\right\|^2,
\end{aligned} \tag{A.10}$$

where the first term $b_{1,i}(\theta^0, g_i, h_i)$ is due to misspecification of groups and the second term $b_{2,i}(\widehat{\theta}, g_i, h_i)$ is due to the estimation error between $\widehat{\theta}$ and θ^0. By Assumption ID and S, $b_{1,i}(\theta^0, g_i, h_i) \geq m_0$ for some $m_0 > 0$ for any $(g_i, h_i) \neq (g_i^0, h_i^0)$. By Assumption R(iii), $b_{2,i}(\widehat{\theta}, g_i, h_i) \leq M_0\eta^2$ for some $M_0 < \infty$. Therefore,

$$\left\|b_i(\widehat{\theta}, g_i, h_i)\right\|^2 \geq m_0 - M_0\eta^2. \tag{A.11}$$

Similarly, we have

$$\left\|b_i(\widehat{\theta}, g_i^0, h_i^0)\right\| = \left\|\mathbb{E}\left[m_{it}\left(\widehat{\theta}, g_i^0, h_i^0\right)\right] - \mathbb{E}[m_{it}(\theta^0, g_i^0, h_i^0)]\right\|^2 \leq M_0\eta^2. \tag{A.12}$$

Combining (A.7) with (A.8), (A.9), (A.11), (A.12), we obtain

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{g}_i = g_i, \widehat{h}_i = h_i\right\} \\
&\leq \quad \mathbb{P}\left\{c_1 m_0 - c_1 M_0\eta^2 - c_2 M_0\eta^2 \leq c_1\left\|\delta_i(\widehat{\theta}, g_i, h_i)\right\|^2 + c_2\left\|\delta_i(\widehat{\theta}, g_i^0, h_i^0)\right\|^2\right\}.
\end{aligned} \tag{A.13}$$

Take $\eta > 0$ small enough such that $s = c_1 m_0 - c_1 M_0\eta^2 - c_2 M_0\eta^2 > 0$. Note that $\delta_i(\widehat{\theta}, g_i, h_i)$ and $\delta_i(\widehat{\theta}, g_i^0, h_i^0)$ both are differences between sample mean and population mean. By Lemma S1.2(ii) of Su, Shi, and Phillips (2016),

$$\begin{aligned}
\max_{1\leq i\leq N} \mathbb{P}\left\{c_1\left\|\delta_i(\widehat{\theta}, g_i, h_i)\right\|^2 \geq s/2\right\} &= o(N^{-1}), \\
\max_{1\leq i\leq N} \mathbb{P}\left\{c_2\left\|\delta_i\left(\widehat{\theta}, g_i, h_i\right)\right\|^2 \geq s/2\right\} &= o(N^{-1}).
\end{aligned} \tag{A.14}$$

Therefore, for any $(g_i, h_i) \neq (g_i^0, h_i^0)$,

$$\max_{1\leq i\leq N} \mathbb{P}\left\{\widehat{g}_i = g_i, \widehat{h}_i = h_i\right\} = o(N^{-1}) \tag{A.15}$$

for $\widehat{\theta} \in N_\eta$. Because g_i and h_i both have finite support, we obtain

$$\max_{1\leq i\leq N} \mathbb{P}\left\{\widehat{g}_i \neq g_i^0, \widehat{h}_i \neq h_i^0\right\} = o(N^{-1}) \tag{A.16}$$

for $\widehat{\theta} \in N_\eta$. Finally, conditional on $\widehat{\theta} \in N_\eta$ and $E_W = 1$, we have

$$\begin{aligned}\mathbb{P}\left\{\widehat{G} = G^0 \text{ and } \widehat{H} = H^0\right\} &= 1 - \mathbb{P}\left\{1\left\{(\widehat{g}_i, \widehat{h}_i) \neq (g_i^0, h_i^0)\right\} \text{ for some } i\right\} \\ &\geq 1 - N \max_{1\leq i\leq N} \mathbb{P}\left\{(\widehat{g}_i, \widehat{h}_i) \neq (g_i^0, h_i^0)\right\} \to 1. \end{aligned} \tag{A.17}$$

By Lemma 3.2 and Assumption W, $\mathbb{P}\{\widehat{\theta} \in N_\eta\} \to 1$ and $\mathbb{P}\{E_W = 1\} \to 1$, which gives the desired result together with (A.17). □

Proof of Theorem 3.4. Because $\widehat{G} = G_0$ and $\widehat{H} = H_0$ with probability approaching 1, $\widetilde{\theta}$ has the same asymptotic distribution as the oracle estimator $\overline{\theta}$ that is obtained by assuming G_0 and H_0 are known, i.e.,

$$\begin{aligned}&\overline{\theta} = \arg\min_{\theta\in\overline{\Theta}} \overline{Q}(\theta), \text{ where} \\ &\overline{Q}(\theta) = \overline{m}(\theta)' W_{NT} \overline{m}(\theta) \text{ and } \overline{m}(\theta) = (NT)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} m\left(w_{it}; \alpha(g_i^0), \beta(h_i^0), \lambda\right). \end{aligned} \tag{A.18}$$

Now we derive the asymptotic distribution of $\overline{\theta}$. This is a standard GMM problem. By Assumption ID, E(ii), and (3.1), we have the typical identification and uniform convergence conditions for the consistency of $\overline{\theta}$. To get the asymptotic distribution, it is sufficient to show for some $\eta > 0$,

$$N^{-1} \sum_{i=1}^{N} \sup_{||\theta_i - \theta_i^0|| \leq \eta} \left\| T^{-1} \sum_{t=1}^{T} m_\theta\left(w_{it}; \theta_i\right) - E\left[m_\theta\left(w_{it}; \theta_i\right)\right] \right\| \to_p 0 \tag{A.19}$$

and

$$(NT)^{-1/2} \sum_{i=1}^{N}\sum_{t=1}^{T} m\left(w_{it}; \theta_i^0\right) \to_d N(0, \Omega) \tag{A.20}$$

as $N, T \to \infty$. The first result in (A.19) follows from the same arguments used to obtain (A.2), under Assumption R and E(iii). The second result in (A.20) follows from verifying a Lindeberg-Feller central limit theorem. Under the same set of conditions as those in this Theorem, part (ii) of the proof of Lemma S1.12 of Su, Shi, and Phillips (2016) provides the details of the verification; see page 29 of the Supplement to Su, Shi, and Phillips (2016). □

B Derivations and Additional Results for Monte Carlo

B.1 Derivations

First-order Conditions for V_{it}. Plugging (4.2) into (4.4), using the production function constraint to substitute out Q_{it}, and differentiating with respect to V_{it} yields the first-order condition

$$V_{it} = \left[(1-\kappa)b_i \exp(\eta_{it} + (1-\kappa)\omega_{it})\right]^{\frac{1}{1-(1-\kappa)b_i}} \exp\left(\frac{(1-\kappa)^2}{1-(1-\kappa)b_i}\frac{\sigma_\epsilon^2}{2}\right). \tag{A.1}$$

In turn, we can write the logged variable input as

$$v_{it} = \frac{1}{1-(1-\kappa)b_i}\left(\left(\eta_{it} - \frac{d_i}{1-\phi}\right) + (1-\kappa)\left(\omega_{it} - \frac{a_i}{1-\rho}\right)\right) + \tilde{a}_i, \tag{A.2}$$

where

$$\tilde{a}_i = \frac{1}{1-(1-\kappa)b_i}\left(\ln\left((1-\kappa)b_i\right) + (1-\kappa)^2\frac{\sigma_\epsilon^2}{2} + \frac{d_i}{1-\phi} + \frac{(1-\kappa)a_i}{1-\rho}\right).$$

Combining the two expressions we can eliminate some constants and obtain

$$v_{it} = \frac{1}{1-(1-\kappa)b_i}\left(\eta_{it} + (1-\kappa)\omega_{it} + \ln\left((1-\kappa)b_i\right) + (1-\kappa)^2\frac{\sigma_\epsilon^2}{2}\right), \tag{A.3}$$

The endogeneity of the marginal cost regressor v_{it} is apparent from its dependence on η_{it} and ω_{it}; see (A.3).

Because ω_{it} and η_{it} are AR(1) processes, the logged variable input v_{it} evolves according to an ARMA(2,1) process which we write as

$$v_{it} = (1-\psi_1-\psi_2)\tilde{a}_i + \psi_1 v_{it-1} + \psi_2 v_{it-2} + \zeta_{it} + \psi_3\zeta_{it-1}, \quad \zeta_{it} \overset{iid}{\sim} N(0,\sigma_\zeta^2). \tag{A.4}$$

The ARMA coefficients $(\psi_1, \psi_2, \psi_3, \sigma_\zeta^2)$ can be derived by multiplying (A.2) by the lag polynomials $(1-\phi L)$ and $(1-\rho L)$ associated with the AR(1) processes η_{it} and ω_{it} in (4.3) and (2.11). In particular,

$$\psi_1 = \phi + \rho \quad \text{and} \quad \psi_2 = -\phi\rho. \tag{A.5}$$

ψ_3 and σ_ζ^2 can be backed out from the equations

$$(1+\phi_3^2)\sigma_\zeta^2 = \frac{1}{\big(1-(1-\kappa)b_i\big)^2}\big((1+\rho^2)\sigma_\nu^2 + (1-\kappa)^2(1+\phi^2)\sigma_\xi^2\big), \tag{A.6}$$
$$\psi_3\sigma_\zeta^2 = -\frac{1}{\big(1-(1-\kappa)b_i\big)^2}\big(\rho\sigma_\nu^2 + (1-\kappa)^2\phi\sigma_\xi^2\big).$$

Calibration of ρ. Using (A.2), note that in our model

$$\mathbb{V}_i(v_{it}) = \frac{1}{\big(1-(1-\kappa)b_i\big)^2}\left(\frac{\sigma_\nu^2}{1-\phi^2} + (1-\kappa)^2\frac{\sigma_\xi^2}{1-\rho^2}\right), \tag{A.7}$$
$$\mathrm{Cov}_i(v_{it}, v_{it-1}) = \frac{1}{\big(1-(1-\kappa)b_i\big)^2}\left(\phi\frac{\sigma_\nu^2}{1-\phi^2} + (1-\kappa)^2\rho\frac{\sigma_\xi^2}{1-\rho^2}\right).$$

Thus, the first-order autocorrelation is given by a variance-weighted average of ϕ and ρ

$$\frac{\mathrm{Cov}_i(v_{it}, v_{it-1})}{\mathbb{V}_i(v_{it})} = \frac{\frac{\sigma_\nu^2}{1-\phi^2}}{\frac{\sigma_\nu^2}{1-\phi^2} + (1-\kappa)^2\frac{\sigma_\xi^2}{1-\rho^2}}\phi + \frac{(1-\kappa)^2\frac{\sigma_\xi^2}{1-\rho^2}}{\frac{\sigma_\nu^2}{1-\phi^2} + (1-\kappa)^2\frac{\sigma_\xi^2}{1-\rho^2}}\rho. \tag{A.8}$$

Given σ_ν, $\sigma_\xi = 1$, the autocorrelation coefficient on the left-hand side, and values for κ and ϕ, one can solve (A.8) for ρ.

Jacobian of GMM Objective Function. The population Jacobian matrix is given by

$$\frac{\partial M_i(\theta_i)}{\partial \theta_i'} = \begin{pmatrix} \mathbb{E}[y_{it-1} - b_i v_{it-1}] & \mathbb{E}[v_{it} - \rho v_{it-1}] & 1 \\ \mathbb{E}[v_{it-1}(y_{it-1} - b_i v_{it-1})] & \mathbb{E}[v_{it-1}(v_{it} - \rho v_{it-1})] & \mathbb{E}[v_{it-1}] \\ \mathbb{E}[v_{it-2}(y_{it-1} - b_i v_{it-1})] & \mathbb{E}[v_{it-2}(v_{it} - \rho v_{it-1})] & \mathbb{E}[v_{it-2}] \end{pmatrix}. \tag{A.9}$$

To obtain the entries for this matrix, first calculate

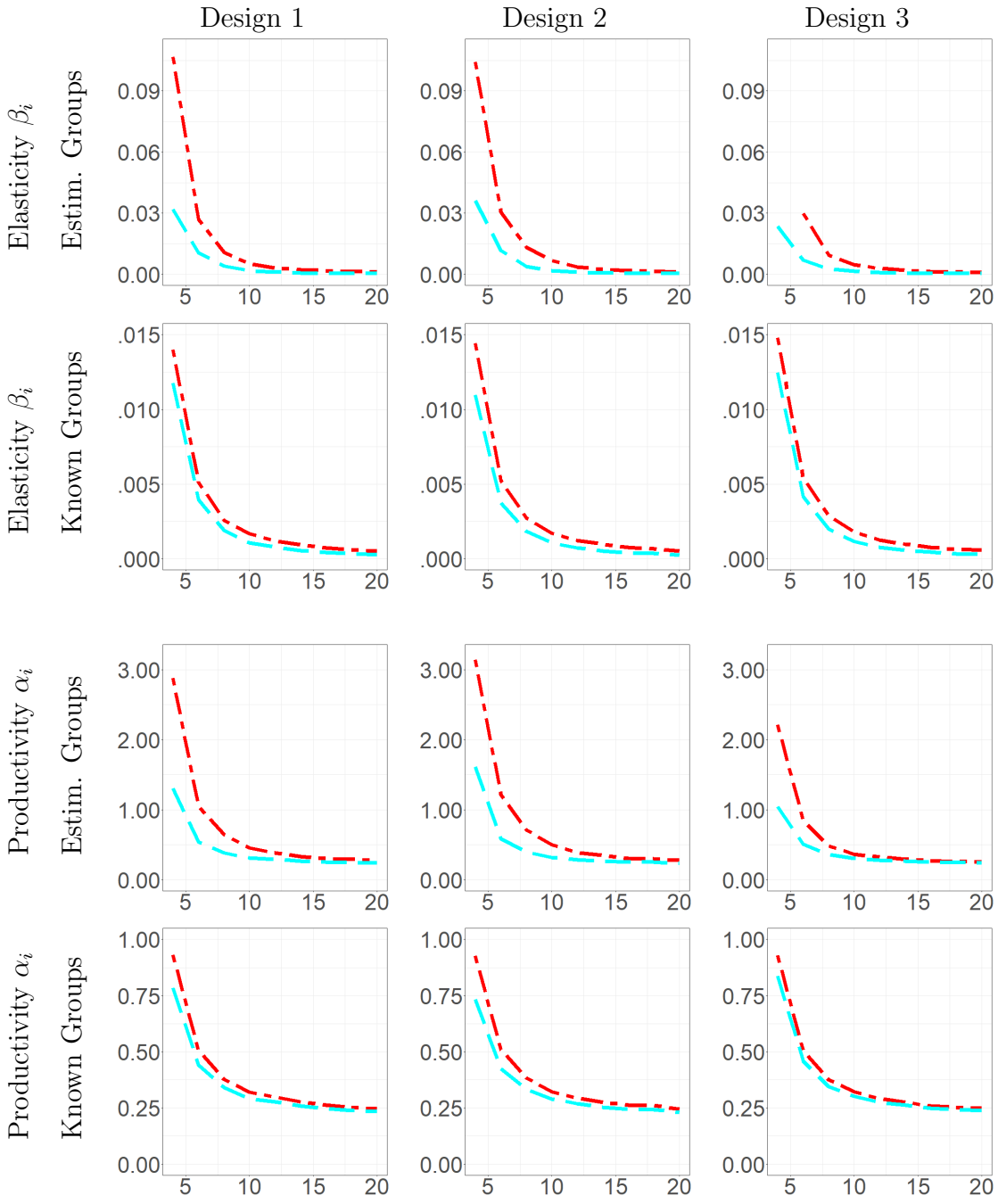
$$\begin{aligned}
\mathbb{E}[\eta_{it}] &= \frac{d_i}{1-\phi}, \\
\mathbb{E}[\omega_{it}] &= \frac{a_i}{1-\rho}, \\
\mathbb{E}[\omega_{it-1}^2], &= \mathbb{V}(\omega_{it-1}) + (\mathbb{E}[\omega_{it-1}])^2 = \frac{\sigma_\xi^2}{1-\rho^2} + \left(\frac{c_i}{1-\rho}\right)^2, \\
\mathbb{E}[\omega_{it-1}\omega_{it-2}] &= \rho\frac{\sigma_\xi^2}{1-\rho^2} + \left(\frac{a_i}{1-\rho}\right)^2, \\
\mathbb{E}[\eta_{it-1}^2] &= \mathbb{V}(\eta_{it-1}) + (\mathbb{E}[\eta_{it-1}])^2 = \frac{\sigma_\nu^2}{1-\phi^2} + \left(\frac{d_i}{1-\phi}\right)^2, \\
\mathbb{E}[\eta_{it}\eta_{it-1}] &= \phi\frac{\sigma_\nu^2}{1-\phi^2} + \left(\frac{d_i}{1-\phi}\right)^2.
\end{aligned} \tag{A.10}$$

For $\kappa = 1/3$ the entries of the Jacobian matrix are as follows:

$$\begin{aligned}
a_i^* &= \frac{3}{3-2b_i}\ln\left(\frac{2}{3}b_i\right) + \frac{2}{3-2b_i}\frac{\sigma_\epsilon^2}{3}, \\
\mathbb{E}[v_{it-1}] &= \mathbb{E}[v_{it-2}] = \tilde{a}_i = a_i^* + \frac{3}{3-2b_i}\left(\frac{d_i}{1-\phi} + \frac{2}{3}\frac{a_i}{1-\rho}\right), \\
\mathbb{E}[y_{it-1} - b_i v_{it-1}] &= \frac{a_i}{1-\rho}, \\
\mathbb{E}[v_{it} - \rho v_{it-1}] &= (1-\rho)\tilde{a}_i = (1-\rho)a_i^* + \frac{3}{3-2b_i}(d_i\frac{1-\rho}{1-\phi} + \frac{2}{3}a_i), \\
\mathbb{E}[v_{it-1}(y_{it-1} - b_i v_{it-1})] &= \frac{2}{3-2b_i}\left(\frac{\sigma_\xi^2}{1-\rho^2} + \left(\frac{a_i}{1-\rho}\right)^2\right) + \frac{a_i}{1-\rho}a_i^*, \\
\mathbb{E}[v_{it-2}(y_{it-1} - b_i v_{it-1})] &= \frac{2}{3-2b_i}\left[\rho\frac{\sigma_\xi^2}{1-\rho^2} + \left(\frac{a_i}{1-\rho}\right)^2\right] + \frac{a_i}{1-\rho}a_i^*, \\
\mathbb{E}[v_{it-1}(v_{it} - \rho v_{it-1})] &= \mathbb{E}[v_{it-1}]\left(\frac{3d_i + 2a_i}{3-2b_i} + (1-\rho)a_i^*\right) \\
&\quad + \frac{3}{3-2b_i}(\phi-\rho)\left(\frac{3}{3-2b_i}\left(\frac{\sigma_\nu^2}{1-\phi^2} + \left(\frac{d_i}{1-\phi}\right)^2\right) + \frac{d_i}{1-\phi}a_i^*\right), \\
\mathbb{E}[v_{it-2}(v_{it} - \rho v_{it-1})] &= \mathbb{E}[v_{it-2}]\left(\frac{3d_i + 2a_i}{3-2b_i} + (1-\rho)\,a_i^*\right) \\
&\quad + \frac{3}{3-2b_i}(\phi-\rho)\left(\frac{3}{3-2b_i}\left(\phi\frac{\sigma_\nu^2}{1-\phi^2} + \left(\frac{d_i}{1-\phi}\right)^2\right) + \frac{d_i}{1-\phi}a_i^*\right).
\end{aligned} \tag{A.11}$$

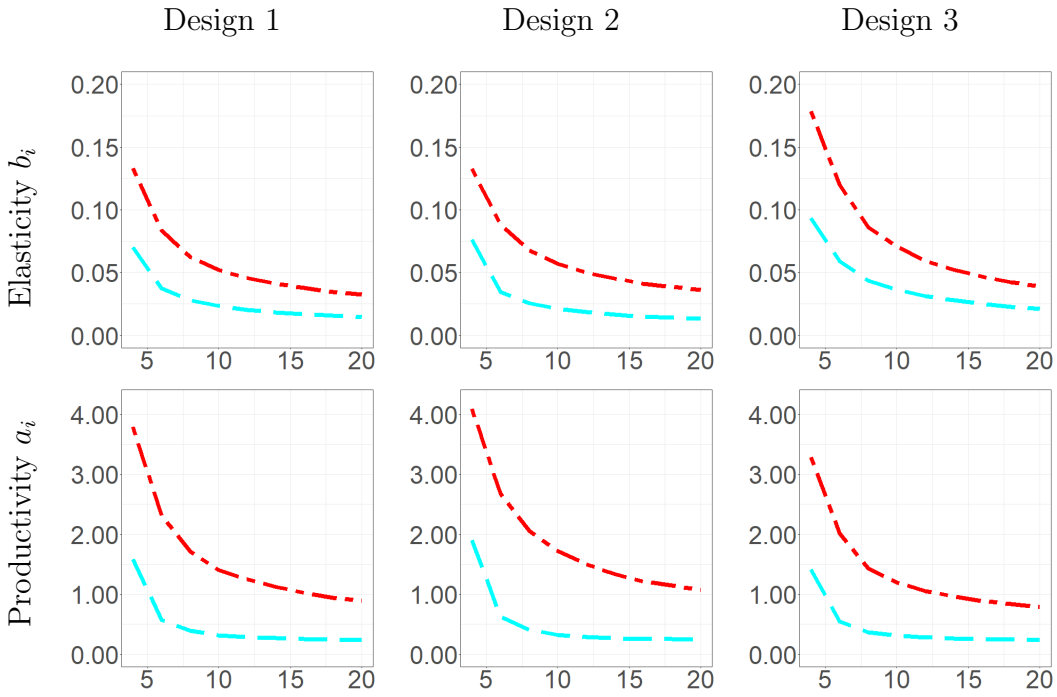
B.2 Additional Empirical Results

Figure A-1: MSE of Group-specific Coefficients, Weighted by Group Size, All Designs



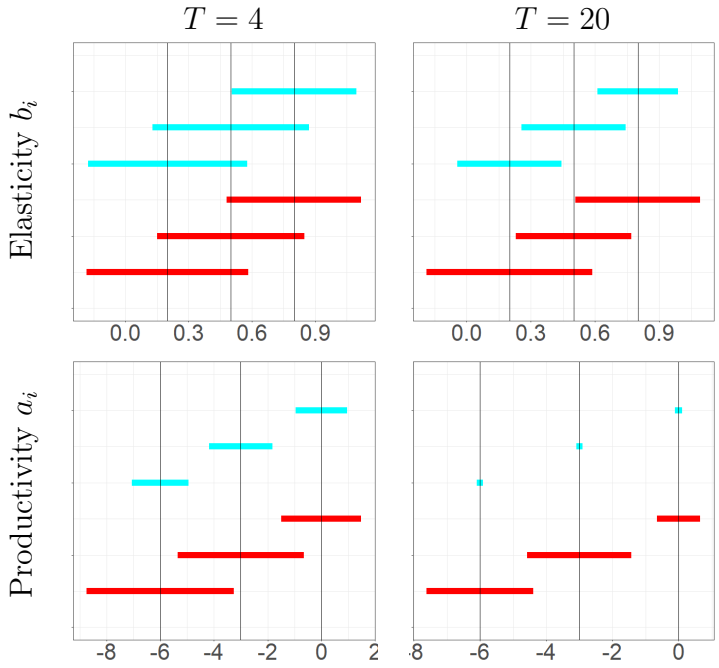
Notes: MSE associated with group-level parameter estimators averaged across groups, weighted by group size. Red dashed-dotted lines are 1D clustering. Cyan dashed lines are 2D clustering.

Figure A-2: MSE (Averaged Across i) of Unit-Level Coefficients, All Designs



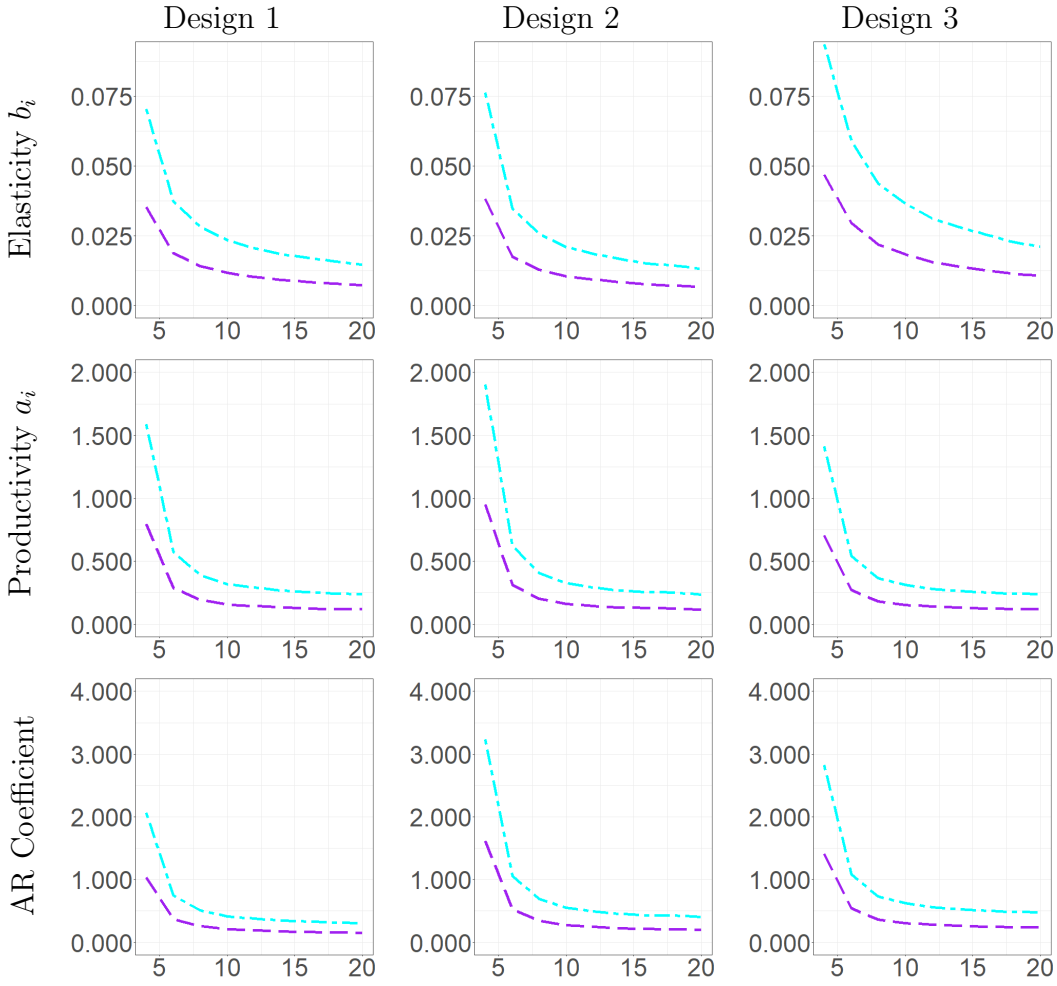
Notes: MSE associated with unit-specific estimators averaged across i. Red dashed-dotted lines are 1D clustering. Cyan dashed lines are 2D clustering.

Figure A-3: Estimation Precision



Notes: Horizontal bars represent true value ± 1.96 times the standard deviation of the estimator. The true values are $\{0.2, 0.5, 0.8\}$ for b_i and $\{-6, -3, 0\}$ for a_i. Red lines are 1D clustering and cyan lines are 2D clustering.

Figure A-4: One-Step versus Two-Step Estimation MSEs, All Designs



Notes: MSE associated with group-level parameter estimators averaged across groups, weighted by group size. MSE of homogeneous AR coefficient ρ. Cyan dashed-dotted lines are 2D one-step estimators. Purple dashed lines are 2D two-step estimators.

C Data Construction for the Empirical Analysis

The firm-level data set is constructed from the Compustat Fundamentals (North America) database. The time period t is one year. The sample starts in 1999 and ends in 2019. The variables in the raw data set include (using Compustat acronyms):

- **year** - Year (to determine period t)
- **conm** - Business Name (to determine firm index i)
- **NAICS** - North American Industry Classification System
- **sale** - Sales of Goods [millions of dollars]
- **cogs** - Cost of Goods Sold [millions of dollars]
- **ppent** - Property, Plant and Equipment - Total (Net) [millions of dollars]
- **ppegt** - Property, Plant and Equipment - Total (Gross) [millions of dollars]

In addition, we used the following series from the Bureau of Economic Analysis (BEA):

- **def_gdp** - GDP Deflator (2012 base year)
- **def_inv** - nonresidential fixed investment good deflator (2012 base year)

The following steps are executed to select and transform the raw data:

1. We removed the following sectors: Utilities (22); Finance and Insurance (52); Real Estate Rental and Leasing (53); Management of Companies and Enterprises (55); Public Administration (92).
2. We only kept US incorporated firms.
3. We removed firms listed in Canadian or Mexican Stock Exchange.
4. We removed observations having "NA" or non-positive **cogs** and **sale**.
5. We removed observations having non-positive **ppent** or **ppegt**.
6. We deflated **cogs**, **sale**, and **xsga** by **def_gdp**.

7. Capital stock construction by perpetual inventory method:

$$k_0 = \mathbf{ppegt}_t, \quad k_{t+1} = k_t + \frac{\mathbf{ppent}_{t+1} - \mathbf{ppent}_t}{\mathbf{def\_inv}_{t+1}}.$$

   Here we deflated changes in **ppent** by **def_inv**. Note that only flow, not stock variables can be deflated by a yearly deflator. Stock variables accumulate nominal prices over different years.

8. We deleted observations with negative capital k_t.

9. We defined variable input as

$$v_t = \frac{\mathbf{xopr}_t}{\mathbf{def\_gdp}_t}.$$

10. We defined production output as

$$y_t = \frac{\mathbf{sale}_t}{\mathbf{def\_gdp}_t}.$$

11. We deleted firms with gap years during their lifespan.

12. We deleted firms with a single period observation.

13. We deleted firms with no NAICS code.

14. We removed firms with acquisition **acq** greater than 5% of its total assets.