Clustering for Multi-Dimensional Heterogeneity

Xu Cheng* Frank Schorfheide Peng Shao
University of Pennsylvania

Preliminary Version: November 21, 2019

Abstract

This paper provides a new multi-dimensional clustering approach for unobserved heterogeneity in panel data models. Each unit is associated with multiple clusters. For example, a firm can belong to the high productivity group and the low output elasticity group. In contrast, the standard one-dimensional clustering approach would be based on separate groups for each productivity-elasticity pair. Our approach provides substantial gains in estimation accuracy when unobserved features have sparse interactions, e.g., there are only a few firms with high productivity and low output elasticity. We propose an estimator for the unobserved group memberships and the group-specific and common parameters in a nonlinear GMM framework and derive its large sample properties. In particular, we provide the first classification consistency result in a nonlinear GMM setup. We re-evaluate the rise of aggregate markup in De Loecker, Eeckhout, and Unger (2018) by replacing their sector-specific production functions with a cluster-based ones. We find that the upward trajectory persists, but the magnitude is less pronounced after accounting for multi-dimensional heterogeneity.

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

KEY WORDS: Clustering, GMM, Panel Data, Production Function Estimation.

*Correspondence: Department of Economics, Perelman Center for Political Science and Economics, University of Pennsylvania, 133 S. 36th St., Philadelphia, PA 19104-6297. Email: xucheng@upenn.edu (X. Cheng), schorf@upenn.edu (F. Schorfheide), pshao@sas.upenn.edu (P. Shao). 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.

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, and in their output elasticities of capital.[1] A flexible specification of the production function ideally allows for heterogeneity in all three of these features. For practical estimation, the key question is how to specify a flexible yet parsimonious and tractable econometric model that is consistent with such multi-dimensional unobserved heterogeneity in the data. In a panel data context, this paper proposes a framework to assign multiple cluster memberships to each cross-sectional unit, where each cluster membership is determined by one particular characteristic of the unit. We estimate the memberships as well as cluster-specific and common parameters in a nonlinear generalized method of moments (GMM) framework.

Recent years have seen increasing popularity of modeling heterogeneity through clusters. In panel data analysis, allowing each cross-sectional unit to have its own regression coefficient often leads to a large number of parameters and a poor estimation of them. Instead, researchers may divide the whole population into a finite-number of clusters and explore the commonality within and differences across clusters. The cluster membership could be known (Bester and Hansen (2016)) or estimated by machine learning methods (Lin and Ng (2012); Ando and Bai (2016); Bonhomme and Manresa (2015), BM hereafter; Su, Shi, and Phillips (2016), SSP hereafter). Similar to clusters, finite mixtures models can be used to model group-wise heterogeneity (Sun (2005); Kasahara and Shimotsu (2009)). Hahn and Moon (2010) provide economic foundations for fixed effects with a finite support. In a Bayesian setting, correlated random effects distributions modeled flexibly with Dirichlet process mixture priors can also capture forms of group heterogeneity (e.g., Liu (2018)). This paper contributes to the literature in various ways, discussed in the following paragraphs.

First, multiple clustering has the benefit of borrow strength among units that are homogeneous in one dimension but heterogeneous in other dimensions. By introducing multiple memberships, units in one cluster share some features but differ in other features. Existing methods are one-dimensional, giving only one membership to each unit and requiring units in a cluster to share all features. For example, in our empirical estimation of the production function, we pool all firms that share the same variable input elasticity together to estimate this common parameter, regardless of the other two features, i.e., productivity and capital

[1] Throughout the paper we will refer to these elasticities simply as variable input and capital elasticities, respectively.

elasticity. Yet, we allow for heterogeneity and cluster patterns in these other features. To fit the production example into the one-dimensional clustering framework, one would only assign firms to the same cluster if their production functions are identical in all dimensions. This results in much smaller cluster sizes and more cluster-specific parameters to estimate.

Second, multi-dimensional clustering is robust to sparse interactions among different features. To estimate cluster-specific parameters, we need a large number of observations from each group. The one-dimensional approach cuts the data finer by requiring all features to be the same in a cluster, making it possible that some cluster is much smaller than others. In the context of the production function example, the one-dimensional framework requires a large number of firms with high productivity and low output elasticity. The proposed multi-dimensional approach only requires a large number of highly productive firms and a large number of firms with low output elasticity separately.

Third, we establish classification consistency of the group membership in a nonlinear GMM framework. The group membership in each dimension is estimated by the K-means method. This theoretical analysis builds on the important classification consistency result in BM. The main difference is that the group memberships here are estimated by a nonlinear GMM criterion instead of a linear least square criterion with heterogenous intercept. We do not allow the parameters to be time-varying as in BM. To the best of our knowledge, SSP is the only paper that considered classification based on a GMM criterion. However, they restricted it to a linear IV model. Classification with other types of criteria are considered, for instance, by SSP, Liu, Schick, Shang, Zhang, and Zhou (2018), Gu and Volgushev (2019). The asymptotic results require both large N and large T, 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 number of clusters for each feature can be determined by a quasi-Bayesian information criterion. Homogeneity is a special case with one cluster.

Fourth, we derive the asymptotic distributions of the cluster-specific and common parameters. SSP model some parameters to be cluster-specific and some parameters to be unit-specific. The latter results in incidental parameter bias that is subsequently corrected. Different from their approach, we model the multi-dimensional heterogeneity symmetrically by assuming all heterogeneous parameters follow cluster patterns. The added flexibility is that different parameters are associated with different memberships. Once the memberships are consistently estimated, we impose the estimated memberships and construct a pooled GMM criterion. All cluster-specific and common parameters are estimated with $\sqrt{NT}$ rate.

We use the proposed multi-dimensional clustering technique to estimate firm-level Cobb-Douglas 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 a sequence of rolling 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 for each rolling sample and re-examine the rise of aggregate markups documented by De Loecker, Eeckhout, and Unger (2018). Our main finding is that the overall level of aggregate markup is lower and the rise in the markup between 1970 and 2016 is less pronounced once one accounts for group heterogeneity among publically-traded firms within two-digit NAICS sectors.

The remainder of the paper is organized as follows. Section 2 describes the model and the estimation procedure. Section 3 provides some key regularity conditions and shows consistency of the estimators. Section 4 starts with some heuristic arguments on classification of group memberships with a nonlinear GMM criterion Subsequently, we provide formal results on classification consistency and the asymptotic distribution of the GMM estimator based on a pooled criterion. Section compares the proposed multi-dimensional clustering to standard one-dimensional clustering in a Monte Carlo simulation. The empirical analysis is presented in Section 6. Finally, Section 7 concludes. Proofs, data definitions, and additional numerical results are relegated to 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 a symmetric, let $\mu_{\max}(A)$ and $\mu_{\min}(A)$ denote the largest and smallest eigenvalues of A. Let $1\{\cdot\}$ denote the indicator function. All asymptotic results are obtain as N and T pass to infinite jointly.

2 Model and Estimator

We have panel data $\{w_{it} : i = 1, ..., N; t = 1, ..., T\}$ and use them to estimate unknown parameters $\theta_i = (a_i, b_i, \lambda) \in A \times B \times \Lambda$ based on moment conditions. The parameter space A, B, Λ are subsets of $R^{d_a}, R^{d_b}, R^{d_\lambda}$, respectively. To study applications where N is

significantly larger than T, we provide a parsimonious model of a_i and b_i by two separate group patterns. Let $g_i \in \{1, ..., n_g\}$ denote the membership for a_i and $h_i \in \{1, ..., n_h\}$ denote the group membership for b_i. We have

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

Let

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

denote the group-specific values. We can write

$$a_i = \alpha(g_i) \text{ and } b_i = \beta(h_i), \tag{2}$$

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

$$\theta = (\alpha, \beta, \lambda),\ G = (g_1, ..., g_N),\ H = (h_1, ..., h_N). \tag{3}$$

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 $\{1, ..., N\}$ 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 Bayesian information criterion given 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 $\theta_i = (\alpha(g_i), \beta(h_i), \lambda)$. The moment condition is

$$M_i(\theta_i^0) = E\left[m\left(w_{it}; \theta_i^0\right)\right] = 0 \tag{4}$$

hold for all i and t. The GMM estimator is[2]

$$\left(\widehat{\theta}, \widehat{G}, \widehat{H}\right) = \underset{(\theta,G,H) \in \overline{\Theta} \times \Gamma_G \times \Gamma_H}{\arg\min} \widehat{Q}(\theta, G, H), \tag{5}$$

where

$$\widehat{Q}(\theta, G, H) = N^{-1} \sum_{i=1}^{N} \widehat{Q}_i(\theta, g_i, h_i), \tag{6}$$

[2] Fernandez-Val and Lee (2013) and SSP also use the same type of criterion in the presence of unit-specific parameters. In our case the unit-specific parameters are the group memberships.

and

$$\widehat{Q}_i(\theta, g_i, h_i) = \left[T^{-1}\sum_{t=1}^{T} m\left(w_{it}; \alpha(g_i), \beta(h_i), \lambda\right)\right]' W_{iNT} \left[T^{-1}\sum_{t=1}^{T} m\left(w_{it}; \alpha(g_i), \beta(h_i), \lambda\right)\right]. \tag{7}$$

for some finite-dimensional function $m(w_{it}; \cdot) \in R^{d_m}$ and weighting matrix W_{iNT}.

Application: Production Function Estimation. Consider a Cobb-Douglas production function

$$y_{it} = a_i^0 + b_i^0 v_{it} + c_i^0 k_{it} + \omega_{it} + \varepsilon_{it}, \tag{8}$$

where y_{it}, k_{it}, v_{it} are the observed log output, log capital input, and 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 an unobserved output shock that is realized after the factor inputs have been chosen. The productivity shock ω_{it} follows an AR(1) process

$$\omega_{it} = \rho^0 \omega_{it-1} + \xi_{it}, \tag{9}$$

where the innovation ξ_{it} is uncorrelated with input choices prior to period t. The output shock ε_{it} is uncorrelated with any input choices at period t and before. For a markup calculation following De Loecker and Warzynski (2012) and De Loecker, Eeckhout, and Unger (2018), the parameter of interest is the output elasticity of the variable input, i.e., b_i^0. The rest are nuisance parameters. As in these papers, we assume the capital input k_{it} is determined at period $t-1$ and firms choose the variable input v_{it} optimaly at period t.

Let

$$\Delta y_{it}(\rho) = y_{it} - \rho y_{it-1},\ \Delta k_{it}(\rho) = k_{it} - \rho k_{it-1},\ \Delta v_{it}(\rho) = v_{it} - \rho v_{it-1} \tag{10}$$

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

$$\Delta y_{it}(\rho^0) - a_i^0(1-\rho^0) - b_i^0 \Delta v_{it}(\rho^0) - c_i^0 \Delta k_{it}(\rho^0) = \xi_{it} + \left(\varepsilon_{it} - \rho^0 \varepsilon_{it-1}\right). \tag{11}$$

Let z_{it} denote a vector of capital and variable inputs choices prior to period t plus the constant term. In the empirical application in Section 6 we will use $z_{it} = (1, k_{it}, k_{it-1}, v_{it-1})'$. This ensures that z_{it} is uncorrelated to the right hand side of (11). We have the moment condition

$$E\left[z_{it}\left(\Delta y_{it}(\rho^0) - a_i^0(1-\rho^0) - b_i^0 \Delta v_{it}(\rho^0) - c_i^0 \Delta k_{it}(\rho^0)\right)\right] = 0. \tag{12}$$

For illustration purpose, we consider a model with two-dimensional group heterogeneity based on a_i and b_i and assume $c_i = c$ for all i. In this case, the common parameter is

$\lambda = (c, \rho)$. With the two-dimensional group membership g_i and h_i for a_i and b_i, respectively, we have

$$\begin{aligned} m\left(w_{it};\theta_i\right) &= z_{it}\left(\Delta y_{it}(\rho) - a_i(1-\rho) - b_i\Delta v_{it}(\rho) - c_i\Delta k_{it}(\rho)\right), \text{ where} \\ a_i &= \alpha(g_i),\ b_i = \beta(h_i),\ \text{and } c_i = c. \end{aligned} \tag{13}$$

In the empirical estimation, we allow for three-dimensional heterogeneity on a_i, b_i, c_i and set the common parameter $\lambda = \rho$. In this case, each firm i has three memberships and the model can be adjusted accordingly. □

In practice, we compute the GMM estimator in (5) by Lloyd's Algorithm. 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 to determine the group memberships $\widehat{G}$. 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 that ensures 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.

Algorithm 1 (Lloyd's Algorithm)

1. **Initialization,** $k = 0$**:** *Provide an initial guess* $(\widehat{G}^{(0)}, \widehat{H}^{(0)})$*. Let* $c = 0$ *and* $\widehat{Q}^{(0)} = M$*.*

2. **Iterations,** $s > 0$**:** *Until* $c = 1$ *execute the following steps:*

   (a) *Using the last iteration's estimate of group memberships* $(\widehat{G}^{(s-1)}, \widehat{H}^{(s-1)})$*, estimate the parameter* θ*:*

   $$\widehat{\theta} = \arg\min_{\theta\in\overline{\Theta}} \widehat{Q}\big(\theta, \widehat{G}^{(s-1)}, \widehat{H}^{(s-1)}\big).$$

   (b) *For* $i = 1, \ldots, N$*, determine the g-group membership:*

   $$\widehat{g}_i^{(s)} = \arg\min_{g_i\in\{1,\ldots,n_g\}} \widehat{Q}_i\big(\widehat{\theta}, g_i, \widehat{h}_i^{(s-1)}\big).$$

   (c) *Re-estimate the parameter* θ*:*

   $$\widehat{\theta}^{(s)} = \arg\min_{\theta\in\overline{\Theta}} \widehat{Q}\big(\theta, \widehat{G}^{(s)}, \widehat{H}^{(s-1)}\big).$$

   (d) *For* $i = 1, \ldots, N$*, determine the h-group membership:*

   $$\widehat{h}_i^{(s)} = \arg\min_{h_i\in\{1,\ldots,n_h\}} \widehat{Q}_i\big(\widehat{\theta}, \widehat{g}_i^{(s)}, h_i\big).$$

   (e) *Assess convergence: let* $\widehat{Q}^{(s)} = \widehat{Q}\big(\theta^{(s)}, \widehat{G}^{(s)}, \widehat{H}^{(s)}\big)$ *and set*

   $$c = 1\big\{\big|\widehat{Q}^{(s)} - \widehat{Q}^{(s-1)}\big| \le \epsilon\big\}.$$

3 Assumptions and Consistent Estimation

First, we assume the following identification condition and regularity conditions on the data generating process.

Assumption ID. For any η, $\min_{1\leq i\leq N}\inf_{||\theta_i-\theta_i^0||>\eta}||M_i(\theta_i)||>\varepsilon>0$.

Assumption R. (i) $\{w_{it}, t=1,2,...\}$ are i.i.d. across i. For each i, $\{w_{it}: t=1,2...,\}$ is stationary strong mixing with mixing coefficients $\alpha_i(\cdot)$, where $\alpha(\cdot)=\sup_i \alpha_i(\cdot)$ satisfies $\alpha(\tau)\leq c_\alpha r^\tau$ 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 $\Theta=\mathcal{A}\times\mathcal{B}\times\Lambda$ 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 $\theta_i,\overline{\theta}_i\in\Theta$. $E|f(w_{it})|^q<\infty$ for some $q\geq 6$.

Application (Continued). We assume the following conditions hold for the production function estimation. (i) $\{(v_{it},k_{it},\xi_{it},\varepsilon_{it}): t=1,...\}$ are i.i.d. over i. For each i, $\{(v_{it},k_{it},\xi_{it},\varepsilon_{it}): t=1,...\}$ is stationary strong mixing that satisfies Assumption R(i). $E(\varepsilon_{it})=0$, $E(\xi_{it})=0$, $E(\varepsilon_{it}k_{it-\tau})$ for $\tau\geq 0$, $E(\varepsilon_{it}v_{it-\tau})$ for $\tau\geq 1$. (ii) $\theta_i=(a_i,b_i,c_i,\rho)\in\Theta=\mathcal{A}\times\mathcal{B}\times\mathcal{C}\times[0,\overline{\rho}]$ for some $\overline{\rho}<1$, where $\mathcal{A},\mathcal{B},\mathcal{C}\in R$ are all convext and compact. The true value θ_i^0 is in the interior of Θ. (iii) Let $x_{it}(\rho)=(1,\Delta v_{it}(\rho),\Delta k_{it}(\rho),\omega_{it-1})'$. $\mu_{\min}(E[z_{it}x_{it}(\rho)]')\geq\delta$ for some $\delta>0$ for any $\rho\in[0,\overline{\rho}]$. (iv) Let $d_{it}=(1,y_{it},y_{it-1},v_{it},v_{it-1},k_{it},k_{it-1})$. For some $C<\infty$ and $q\geq 6$, $E||z_{it}d_{it}||^q\leq C$. Assumption ID and Assumption R hold for the production function example under conditions (i)-(iv). □

Assumption NT. $N^2=O(T^{q/2-1})$, where $q\geq 6$ is the constant in Assumption R1(iii).

Assumption NT allows N to be much larger than T, if the condition holds for a large q, which further translates to the moment condition in Assumption R1(iii). Alternatively, one can also impose tail condition on $f(w_{it})$ directly, as in BM.

Under Assumption R1 and NT, SSP establishes the uniform convergence result[3]

$$P\left\{\max_{1\leq i\leq N}\sup_{\theta_i\in\Theta}\left\|T^{-1}\sum_{t=1}^{T}m\left(w_{it};\theta_i\right)-E[m\left(w_{it};\theta_i\right)]\right\|\geq\eta\right\}=o(N^{-1}) \tag{14}$$

for any $\eta>0$, as $N,T\to\infty$. To establish the estimation consistency in Lemma 3.1 below, the convergence rate $o(N^{-1})$ can be replaced with $o(1)$ in (14). However, to subsequently show the K-mean classification consistency for the memberships, the $o(N^{-1})$ rate is necessary.

[3]See Lemma S1.2(iii) of SSP.

Assumption W. There exists nonrandom matrices W_i such that $\max_{1\leq i\leq N}\|W_{iNT}-W_i\|\to_p 0$ and $\min_{1\leq i\leq N}\mu_{\min}(W_i)=\underline{c}_W>0$ and $\max_i \mu_{\max}(W_i)=\overline{c}_W<\infty$.

Application (Continued). For the production function application, we can choose $W_{iNT}=(T^{-1}\sum_{t=1}^{T} z_{it}z_{it}')^{-1}$. 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 (14) and condition $E[z_iz_i']$ has full rank and $E||z_i||^2<\infty$. □

The following Lemma shows that the estimators are consistent on average.

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

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

Next, we consider estimation of the group specific parameters $\alpha^0=(\alpha_1^0,...,\alpha_{n_g}^0)$ and $\beta^0=(\beta_1^0,...,\beta_{n_g}^0)$. To this end, we add Assumption S, which 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 $c>0$.

(ii) $N^{-1}\sum_{i=1}^{n}1\{g_i^0=g\}\to\pi_g>0$ and $N^{-1}\sum_{i=1}^{n}1\{h_i^0=h\}\to\psi_h>0$ for all $g\in\{1,...,n_g\}$ and $h\in\{1,...,n_h\}$.

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

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

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, this one-dimensional method does not allow for sparse interactions, because the number of observations in this interaction is too small. 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 $\beta(h_i)$.

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 $\widehat{\alpha}=(\widehat{\alpha}_1,...,\widehat{\alpha}_{n_g})$ is a consistent estimator of $\alpha^0=(\alpha_1^0,...,\alpha_{n_g}^0)$ and $\widehat{\beta}=(\widehat{\beta}_1,...,\widehat{\beta}_{n_h})$ is a consistent estimator of $\beta^0=(\beta_1^0,...,\beta_{n_g}^0)$ below.

Lemma 3.2 *Under the assumptions for Lemma 3.1 and Assumption S,* $\widehat{\theta} \to_p \theta^0$, *i.e.,* $\widehat{\alpha} \to_p \alpha^0$, $\widehat{\beta} \to_p \beta^0$, $\widehat{\lambda} \to_p \lambda^0$.

It is worth pointing out that $N^{-1}\sum_{i=1}^{N}(\widehat{\alpha}(\widehat{g}_i) - \alpha^0(g_i^0))^2$ 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.

4 Classification and Asymptotic Distribution

Given $\widehat{\theta}$, $\widehat{G}$ and $\widehat{H}$ are K-mean estimators of the group memberships that minimize the non-linear GMM criterion function $Q(\widehat{\theta}, G, H)$. BM provide consistency of the K-mean clustering for linear least squares estimation. SSP study classification with the GMM criterion using a shrinkage procedure, but also restrict it to linear models. We extend classification consistency to nonlinear GMM problems and allow for multiple-dimensional K-mean methods.

Before presenting the formal result, we first illustrate the intuition and key arguments. For the ease of notation in subsequent arguments, write

$$m_{it}(\theta, g, h) = m(w_{it}; \alpha(g), \beta(h), \lambda), \tag{15}$$

for any $g \in \{1, ..., n_g\}$, $h \in \{1, ..., n_h\}$. 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

$$P\left\{\widehat{g}_i = g_i, \widehat{h}_i = h_i\right\} \leq P\left\{\widehat{Q}_i(\widehat{\theta}, g_i, h_i) < \widehat{Q}_i(\widehat{\theta}, g_i^0, h_i^0)\right\}. \tag{16}$$

By Assumption W,

$$\begin{aligned} \widehat{Q}_i(\widehat{\theta}, g_i, h_i) &\geq c_1 \left\| \frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i) \right\|^2, \\ \widehat{Q}_i(\widehat{\theta}, g_i^0, h_i^0) &\leq c_2 \left\| \frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i^0, h_i^0) \right\|^2 \end{aligned} \tag{17}$$

for some positive constants c_2 and c_1, 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

$$P_{i,gh}(\widehat{\theta}) = P\left\{ c_1 \left\| \frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i) \right\|^2 \leq c_2 \left\| \frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i^0, h_i^0) \right\|^2 \right\}. \tag{18}$$

With a decomposition,

$$\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i) = \left(\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i, h_i) - E[m_{it}(\widehat{\theta}, g_i, h_i)] \right) + E[m_{it}(\widehat{\theta}, g_i, h_i)], \tag{19}$$

where (i) the first term on the right hand side is a $o_p(1)$ *noise* term and (ii) the second term $E[m_{it}(\widehat{\theta}, g, h)]$ is a *signal* term that is strictly positive and bounded away from 0 conditional on $\widehat{\theta} \in N_\eta$ for η small enough. This positive signal for misspecified group is ensured by the separability condition in Assumption S and the identification condition in Assumption ID. By a similar decomposition for $T^{-1}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i^0, h_i^0)$, we can show that (i) the noise is also $o_p(1)$ and (ii) the signal term $E[m_{it}(\widehat{\theta}, g_i^0, h_i^0)]$ is arbitrarily small with $\widehat{\theta} \in N_\eta$ for η small enough because $E[m_{it}(\theta^0, g_i^0, h_i^0)] = 0$. 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 $o(N^{-1})$. Therefore, we have $P_{i,gh}(\widehat{\theta})$ converges to 0 at $o(N^{-1})$ rate and the who group can be classified consistently. The result is presented in the Theorem below and its formal proof is given in the Appendix.

Theorem 4.1 *Suppose Assumptions ID, R, NT, W, S hold.*

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

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

Next we study estimation of $\theta^0 = (\alpha_1^0, ..., \alpha_{n_g}^0, \beta_1^0, ..., \beta_{n_h}^0, \lambda^0)$. Given the group membership $\widehat{G}$ and $\widehat{H}$, we can estimate θ^0 by minimizing a pooled GMM criterion

$$\widetilde{\theta} = \arg\min_{\theta \in \overline{\Theta}} \widetilde{Q}(\theta), \text{ where } \widetilde{Q}(\theta) = \widetilde{m}(\theta)' W_{NT} \widetilde{m}(\theta), \tag{20}$$

with

$$\widetilde{m}(\theta) = (NT)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} m\left(w_{it}; \alpha(\widehat{g}_i), \beta(\widehat{h}_i), \lambda \right), \tag{21}$$

and W_{NT} is a weighting matrix which could depend on $\widehat{G}$ and $\widehat{H}$. In a linear instrumental variable model with heterogeneous coefficients, SSP show that the pooled estimator $\widetilde{\theta}$ is preferred to $\widehat{\theta}$ in (5) because $\widehat{\theta}$ typically is less efficient and suffers from asymptotic bias. Under Theorem 4.1, $\widetilde{\theta}$ 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 the 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

$$m_\theta\left(w_{it};\theta_i^0\right) = \left[\frac{\partial}{\partial\alpha}m\left(w_{it};\theta_i^0\right) : \frac{\partial}{\partial\beta}m\left(w_{it};\theta_i^0\right) : \frac{\partial}{\partial\lambda}m\left(w_{it};\theta_i^0\right)\right] \in R^{d_m\times\left(d_a n_g + d_\beta n_h + d_\lambda\right)}, \quad (22)$$

where

$$\begin{aligned}
\frac{\partial}{\partial\alpha}m\left(w_{it};\theta_i^0\right) &= \left[\frac{\partial}{\partial\alpha_1}m\left(w_{it};\theta_i^0\right) : \cdots : \frac{\partial}{\partial\alpha_{n_g}}m\left(w_{it};\theta_i^0\right)\right] \in R^{d_m\times(d_\alpha n_g)}, \\
\frac{\partial}{\partial\beta}m\left(w_{it};\theta_i^0\right) &= \left[\frac{\partial}{\partial\beta_1}m\left(w_{it};\theta_i^0\right) : \cdots : \frac{\partial}{\partial\beta_{n_h}}m\left(w_{it};\theta_i^0\right)\right] \in R^{d_m\times\left(d_\beta n_h\right)}. \quad (23)
\end{aligned}$$

Under the group structure, $m(w_{it}, \theta_i^0)$ do note depend on α_g for $g \neq g_i^0$ or β_h for $h \neq h_i^0$. Thus, we have

$$\begin{aligned}
\frac{\partial}{\partial\alpha_g}m\left(w_{it};\theta_i^0\right) &= 1\left\{g_i^0 = g\right\} m_\alpha(w_{it},\theta_i^0) \text{ for } g = 1, ..., n_g, \\
\frac{\partial}{\partial\beta_h}m\left(w_{it};\theta_i^0\right) &= 1\left\{h_i^0 = h\right\} m_\beta(w_{it},\theta_i^0) \text{ for } h = 1, ..., n_h, \quad (24)
\end{aligned}$$

where

$$\begin{aligned}
m_\alpha(w_{it},\theta_i) &= \frac{\partial}{\partial a_i}m\left(w_{it};a_i,b_i,\lambda\right) \in R^{d_m\times d_\alpha}, \\
m_\beta(w_{it},\theta_i) &= \frac{\partial}{\partial b_i}m\left(w_{it};a_i,b_i,\lambda\right) \in R^{d_m\times d_\beta}. \quad (25)
\end{aligned}$$

The Jacobian matrix is

$$J = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} E\left[m_\theta\left(w_{it};\theta_i^0\right)\right]. \quad (26)$$

The covariance of the moment condition is

$$\begin{aligned}
\Omega &= \lim_{N\to\infty}\lim_{T\to\infty} N^{-1}\sum_{i=1}^{N}\Omega_{iT}(\theta_{i,0}), \text{ where} \\
\Omega_{i_T}(\theta_i^0) &= T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T} E\left[m\left(w_{it};\theta_i^0\right) m\left(w_{is};\theta_i^0\right)'\right]. \quad (27)
\end{aligned}$$

These limits exist because the data is strong mixing over t, i.i.d. over i, and there is a finite-number of groups whose share converges to constants. We add the following regularity condition to derive the distribution of $\widetilde{\theta}$.

Assumption E. (i) J and Ω 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 4.2 *Suppose Assumptions 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 estimation of β and λ. This is different from the case of a one-dimensional clustering considered by linear GMM problem in SSP, where the estimator of cluster specific parameter only depends on individuals in that cluster.

Application (Continued). In this application, the Jacobian matrix is

$$J = E[m_\theta\left(w_{it}; \theta_i^0\right)] = -E[z_{it}((1-\rho^0), \Delta v_{it}(\rho^0), \Delta k_{it}(\rho^0), \omega_{it-1})] \tag{28}$$

which is full rank under condition (iii) for this example and $\rho_0 < 1$. Let $u_{it} = \xi_{it} + \left(\varepsilon_{it} - \rho^0 \varepsilon_{it-1}\right)$. The covariance matrix is

$$\Omega = \Sigma_{j=-\infty}^{\infty} \Gamma_j, \text{ where } \Gamma_j = E[z_{it} z'_{it-j} u_{it} u_{it-j}]. \tag{29}$$

We assume Ω is positive definite. In the first step, we use $W_{NT} = I_{d_m}$. In the second step, we use the optimal weighting matrix $W_{NT} = \widehat{\Omega}^{-1}$, where $\widehat{\Omega}$ is a heteroskedasticity and autocorrelation consistent (HAC) covariance estimator of Ω , see Newey and West (1987) and Andrews (1991). In the construction of the HAC estimator, we replace the expectation with the sample average over both i and t because this is for the pooled estimator. Similarly, we can get a consistent estimator of J by replacing the expectation with the sample average over both i and t and replacing ρ^0 with the pooled estimator $\widetilde{\rho}$. Assumption E(iii) holds under condition (iv) for this example, listed below Assumption R. □

The GMM criterion with the optimal weighting matrix is

$$\widetilde{Q}(n_g, n_h) = \widetilde{m}(\widetilde{\theta})'\widehat{\Omega}^{-1}\widetilde{m}(\widetilde{\theta}), \tag{30}$$

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 criterion for the problem is

$$BIC(n_g, n_h) = (NT)\,\widetilde{Q}(n_g, n_h) + \log(NT)(n_g d_\alpha + n_h d_\beta). \tag{31}$$

In practice, we can 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 $h_{\max}$. Besides the BIC criterion, a wide range of penalty can be derived for model selection consistency, as shown by BM and SSP for clusters and Bai and Ng (2002) and Cheng, Liao, and Schorfheide (2016) for factor models. Different from these papers, all parameter are estimated at the $\sqrt{NT}$ rate in this problem and the J statistic, i.e., $(NT)\,\widetilde{Q}(n_g, n_h)$, is a natural analog of the log-likelihood. Therefore, the BIC criterion in (31) is a natural choice for selecting the number of clusters. A formal testing procedure for n_g and n_h similar to that in Lu and Su (2016) is worth investigating but is beyond the scope of this paper.

5 Monte Carlo Experiment

We conduct a small Monte Carlo experiment to illustrate the difference between multi-dimensional and one-dimensional clustering in a simple location model. Let $w_{it} = (w_{1,it}, w_{2,it})'$. $\alpha(k)$, $\beta(l)$, $k, l \in \{1, 2\}$ are the parameters of interest and the group memberships are denoted by g_i and h_i. We assume that the following moment condition holds at the true parameter values:

$$E\big[w_{it} - (\alpha^0(g_i^0), \beta^0(h_i^0))'\big] = 0. \tag{32}$$

Defining $\theta = (\alpha(1), \alpha(2), \beta(1), \beta(2))'$ and $W_{iNT} = I$, where $I_{2\times 2}$, where $I_{2\times 2}$ identity matrix, we obtain

$$\begin{aligned} \widehat{Q}(\theta, g_i, h_i) &= \left(T^{-1}\sum_{t=1}^{T} w_{1,it} - \alpha(g_i)\right)^2 + \left(T^{-1}\sum_{t=1}^{T} w_{2,it} - \beta(h_i)\right)^2 \\ &= \big(\bar{w}_{1,i} - \alpha(g_i)\big)^2 + \big(\bar{w}_{2,i} - \beta(h_i)\big)^2, \end{aligned} \tag{33}$$

where $\bar{w}_{j,i}$ is the time series average of the $w_{j,it}$'s. Rather than modeling the law of motion of $w_{j,it}$ explicitly, we simply make distributional assumptions about the $\bar{w}_{j,i}$'s. For large

T, we expect the sample averages to be approximately normally distributed, which is why we are assuming a data generating process (DGP) of the following form (omitting the 0 superscripts)

$$\bar{w}_i = \begin{bmatrix} \bar{w}_{1,i} \\ \bar{w}_{2,i} \end{bmatrix} \sim N\left(\begin{bmatrix} \alpha(g_i) \\ \beta(h_i) \end{bmatrix}, \begin{bmatrix} \sigma^2(g_i,h_i) & 0 \\ 0 & \sigma^2(g_i,h_i) \end{bmatrix} \right), \quad g_i, h_i \in \{1,2\}. \tag{34}$$

We consider the following parameterization:

$$\big[(\alpha(k),\beta(l),\sigma^2(k,l))\big]_{k,l\in\{1,2\}} = \begin{bmatrix} (0.3,0.3,0.1) & (0.3,0.7,4.0) \\ (0.7,0.3,0.5) & (0.7,0.7,2.5) \end{bmatrix}. \tag{35}$$

The parameters $\alpha(k)$, $\beta(l)$ and the group memberships g_i and h_i are estimated based on the following objective function

$$\widehat{Q}(\theta,G,H) = N^{-1}\sum_{i=1}^{N}\big(\bar{w}_{1,i}-\alpha(g_i)\big)^2 + N^{-1}\sum_{i=1}^{N}\big(\bar{w}_{2,i}-\beta(h_i)\big)^2. \tag{36}$$

In our stylized DGP, the co-clustering algorithm determines the group memberships g_i from $\bar{w}_{1,i}$, whereas the group memberships h_i are determined from $\bar{w}_{2,i}$. The GMM estimator $(\widehat{\theta},\widehat{G},\widehat{H})$ has the following representation. There are cutoff points $\widehat{\alpha}_*$ and $\widehat{\beta}_*$ such that

$$\widehat{g}_i = \begin{cases} 1 & \text{if } \bar{w}_{1,i} < \widehat{\alpha}_* \\ 2 & \text{otherwise} \end{cases} \qquad \widehat{h}_i = \begin{cases} 1 & \text{if } \bar{w}_{2,i} < \widehat{\beta}_* \\ 2 & \text{otherwise} \end{cases}$$

and

$$\widehat{\alpha}(k) = \frac{N^{-1}\sum_{i=1}^{N}\bar{w}_{1,i}1\{\widehat{g}_i=k\}}{N^{-1}\sum_{i=1}^{N}1\{\widehat{g}_i=k\}}, \quad \widehat{\beta}(l) = \frac{N^{-1}\sum_{i=1}^{N}\bar{w}_{2,i}1\{\widehat{h}_i=l\}}{N^{-1}\sum_{i=1}^{N}1\{\widehat{h}_i=l\}}, \quad k,l\in\{1,2\}.$$

In this simple linear setting in which the estimators are sample averages, the GMM estimator $\widehat{\theta}$ is identical to pooled GMM estimator $\widetilde{\theta}$ in (20).

Under a single-dimensional clustering approach one would form four separate groups which we denote by $(1,1)$, $(1,2)$, $(2,1)$, and $(2,2)$. The parameters a_i and b_i could now take on four different values each. Accordingly, we write $a_i = \alpha_c(g_i,h_i)$ and $b_i = \beta_c(g_i,h_i)$. Here we use c subscript to indicate one-dimensional clustering. The resulting least squares objective function takes the form

$$\widehat{Q}_c(\theta_c,G,H) = N^{-1}\sum_{i=1}^{N}\big(\bar{w}_{1,i}-\alpha_c(g_i,h_i)\big)^2 + N^{-1}\sum_{i=1}^{N}\big(\bar{w}_{2,i}-\beta_c(g_i,h_i)\big)^2. \tag{37}$$

Table 1: Monte Carlo Results

	Membership			
	Known		Estimated	
Algorithm	MSE	Bias^2	MSE	Bias^2
Estimate of $\alpha(1)$				
2D Clustering (0.3, $\cdot$, $\cdot$)	.0082	0	0.401	0.320
1D Clustering (0.3, 0.3, 0.1)	.0008	0	0.686	0.183
1D Clustering (0.3, 0.7, 4.0)	.0320	0	2.853	1.814
Estimate of $\alpha(2)$				
2D Clustering (0.7, $\cdot$, $\cdot$)	.0060	0	1.220	1.127
1D Clustering (0.7, 0.3, 0.5)	.0040	0	1.397	1.007
1D Clustering (0.7, 0.7, 2.5)	.0200	0	1.399	0.743
Unit-level Estimate of a_i				
2D Clustering	.0071	0	0.875	.0416
1D Clustering	.0142	0	1.025	.0063

Notes: The results are based on $n_{sim} = 2,000$ samples of size $N = 500$ with 125 observations from each of the four groups.

It is now no longer additively separable because the $\alpha_c(\cdot)$ and $\beta_c(\cdot)$ functions depend on both g_i and h_i. The standard one-dimensional clustering algorithm divides the α-β plane into four sections. However, unlike in the case of the two-dimensional clustering, the boundaries of these segments are not simply given by the intersection of a horizontal and a vertical line.

We now generate samples $n_{sim} = 2,000$ samples of size $N = 500$ from the DGP in (34). Each sample has 125 observations from the four groups. The results are summarized in Table 1. We report mean-squared errors (MSE) and squared bias for $\widehat{\alpha}(\cdot)$ and $\widehat{a}_i$. The estimation error for a_i can be decomposed as follows:

$$\widehat{a}_i - a_i = \big(\widehat{\alpha}(g_i) - \alpha(g_i)\big) + \big(\widehat{\alpha}(\widehat{g}_i) - \widehat{\alpha}(g_i)\big).$$

The first term captures the error caused by the estimation of $\alpha(\cdot)$, assuming that the group memberships are known. In this case all estimates are unbiased. The resulting MSEs capture the estimation variance and are summarized in the second column of Table 1. They can be directly calculated by taking appropriate averages of $\sigma^2(g_i, h_i)$. For instance, for two-dimensional clustering we obtain an $\text{MSE}(\widehat{\alpha}(1))$ of $0.25 \cdot (\sigma^2(1,1)/125 + \sigma^2(1,2)/125)$. One-dimensional clustering generates two estimates $\text{MSE}(\widehat{\alpha}(1,l)) = \sigma^2(1,l)/125$, $l \in \{1,2\}$.

Because the two-dimensional clustering estimator is based on the identity weight matrix, under known group memberships it dominates the one-dimensional estimate based on the high-variance cluster, $\widehat{\alpha}(1,2)$, but not the estimate based on the low-variance cluster, $\widehat{\alpha}(1,1)$. The four groups have equal shares in the simulated samples. Therefore, the MSEs associated with the unit-level estimates of a_i are simply averages of the MSEs associated with the estimates of the various α's. Due to this averaging, the two-dimensional clustering dominates the one-dimensional clustering even if group memberships are known.

The results in the last two columns of Table 1 capture both the estimation error of the α's and the misclassification errors. If the group membership has to be estimated, the MSEs increase drastically. This is not surprising because the centers of the group-specific Normal distributions in (34) are close to each other relative to their respective variances. Roughly 50% of the observations are missclassified. While the bias component of the one-dimensional clustering estimator often exceeds that of the two-dimensional clustering procedure, the pooling of observations leads to a strong variance reduction so that in terms of MSEs the two-dimensional clustering estimator clearly dominates.

The parameterization of the DGP in (35) sets a very high bar for the clustering algorithms. Under the large (N,T) asymptotics the variance of $\bar{w}_{j,i}$ will shrink as $T \to \infty$. This makes it easier to detect the group memberships and the bias from the classification error would eventually vanish.

6 Empirical Analysis

Our empirical analysis re-examines the rise of aggregate markups documented by De Loecker, Eeckhout, and Unger (2018). Rising markups are a reflection of a decrease of competitiveness within sectors and can contribute to the observed fall of the labor share and increase in income inequality. We will show that allowing for multi-dimensional group heterogeneity within firms in two-digit NAICS sectors leads to a lower level of estimated markups and a smaller growth rate. Section 6.1 reviews the specification of the production function and the computation of the markups. The data set and the model specifications considered in the empirical analysis are described in Section 6.2. The empirical results are presented in Section 6.3.

6.1 Production Function and Markups

We will now estimate firm-level Cobb-Douglas production functions. Each firm is part of a sector d which we take to be a a two-digit NAICS sector. We follow the setup discussed in Section (2). The production function and the autoregressive law of motion for the unobserved productivity shock ω_{it} are given in (8) and (9), respectively. For convenience, we reproduce the equations:

$$y_{it} = a_i + b_i v_{it} + c_i k_{it} + \omega_{it} + \epsilon_{it}, \quad \omega_{it} = \rho \omega_{it-1} + \xi_{it}.$$

The GMM estimation is based on the moment conditions (12). Recall that the production function of is quasi-differenced to eliminate the serial correlation in ω_{it} and the vector of instruments is defined as $z_{it} = (1, k_{it}, k_{it-1}, v_{it-1})'$. We allow for group heterogeneity in a_i, b_i, and c_i. In addition to $\alpha(\cdot)$ and $\beta(\cdot)$, we define $\gamma(\cdot)$ to characterize the group-specific values of c_i. We use j_i to indicate group memberships for the third group and n_j to denote the number of groups.

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

$$mu_{it} = b_i \frac{p_{it}^y \exp[y_{it}]}{p_{it}^v \exp[v_{it}]}, \tag{38}$$

where p_{it}^y and p_{it}^v are firm-specific prices of the output and the variable input, respectively. Using market shares, we aggregate the firm-level markups to the sectoral level and the economy-wide level. Let $\mathcal{I}_t^d$ be the set of firms i that below to sector d, then the sector-level and the economy-wide markups are given by

$$mu_t^d = \sum_{i \in \mathcal{I}_t^d} \left(\frac{p_{it}^y \exp[y_{it}]}{\sum_{i \in \mathcal{I}_t^d} p_{it}^y \exp[y_{it}]} \right) mu_{it}, \quad mu_t = \sum_{i=1}^{N} \left(\frac{p_{it}^y \exp[y_{it}]}{\sum_{i=1}^{N} p_{it}^y \exp[y_{it}]} \right) mu_{it}. \tag{39}$$

6.2 Data Set, Model Specifications, and Estimation

As in De Loecker, Eeckhout, and Unger (2018) and Flynn, Gandhi, and Traina (2019), 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*

Table 2: 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

Equipment series. Nominal variables are converted to real variables using the appropriate deflators. Our sample starts in 1961 and ends in 2016. Further details on data definitions, transformations, and subsample selection are provided in the Online Appendix.

There are 22 two-digit NAICS sectors. We exclude the following sectors from the subsequent analysis: Finance and Insurance (NAICS 52), Real Estate and Rental and Leasing (NAICS 53), and Public Administration (NAICS 92). 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 2.

The subsequent analysis is conducted for firms that are associated with the same two-digit NAICS sector d. Hence, we drop the sector sub- and superscripts d if no ambiguity arises. We estimate the coefficients of the production function (8) for a sequence of rolling samples. The length of the rolling sample is $T = 10$ years. The first sample spans the period

from 1961 to 1970 whereas the last rolling sample ranges from 2007 to 2016. The estimation for sector d includes firms for which we have at least one observation between $t = 1, \ldots, T$. We set the number of groups for a_i, b_i, and c_i equal to $n_g = n_h = n_j = 3$.[4] We refer to the results obtained from our multi-dimensional clustering estimator implemented with Algorithm 1 as *estimated heterogeneity*. In addition, we consider two alternative estimators. The *homogeneity* estimator is based on imposing that all firms within a sector d use the same production function. This corresponds to $n_g = n_h = n_j = 1$. The *subsector heterogeneity* estimator assumes that the production functions differ across three-digit NAICS codes. Thus, it is based on a grouping determined by a statistical agency instead of an estimation criterion.

An important set in the empirical analysis is to determine the sector-specific degree of heterogeneity in the production function coefficients. To do so, we use a quasi-Bayesian information criterion introduced in (31). For sample τ and model specification m, we rewrite the criterion as

$$BIC_\tau(m) = S_\tau \widetilde{Q}_{\tau,m}(\widetilde{\theta}, \widehat{G}, \widehat{H}, \widehat{J}) + k_m \log S_\tau,$$

where k_m is the number of group-specific and homogeneous coefficients and S_τ is the total number of observations in each panel τ, accounting for the fact that the panel is unbalanced.[5] Currently, we have not yet implemented a full search over $1 \leq n_g, n_h, n_j \leq \bar{n}$. Thus, we will use the criterion to compare the three above-mentioned specifications: *estimated heterogeneity*, *homogeneity*, and *subsector heterogeneity*.

6.3 Empirical Results

We will begin with evidence of firm heterogeneity within two-digit industries, discuss estimation results for the manufacturing sector (NAICS 32) in more detail, and then present summaries of the results across all sectors and rolling samples.

Model Selection. Table 3 summarizes the results from applying the information criteria. Rather than computing the BIC for each period separately, we are averaging over multiple samples. Columns (2) to (4) of the table contain information about the complexity, measured in terms of number of parameters, of each of the specifications. Under *estimated heterogeneity* there are generally ten free parameters: three productivities $\alpha(\cdot)$, three variable input elasticities $\beta(\cdot)$, three capital elasticities $\gamma(\cdot)$, and the autoregressive coefficient ρ. For a few industries we use slightly more restrictive specifications. Under *homogeneity*, there are four

[4] There are three exceptions; see notes for Table 3.

[5] Under *subsector heterogeneity* we also estimate separate ρ's for each three-digit industry.

Table 3: Model Selection

	Complexity			Selection		
NAICS	Est.Het.	Homog.	Subsector	Est.Het.	Homog.	Subsector
21	10	4	16	X		
23	7	4	24	X		
31	10	4	24			X
32	10	4	28	X		
33	10	4	36			X
42	9	4	24			X
44	10	4	36		X	
45	10	4	16			X
48	10	4	36			X
51	10	4	40	X		
54	10	4	4	X		
56	9	4	8	X		
62	10	4	20	X		
72	10	4	8	X		

Notes: Due to data limitations we restricted the heteroneity in three industries. For 23 we use $n_g = 3$ (productivity), $n_h = 2$ (variable inputs), $n_j = 1$ (capital). For 42 we use $n_g = 3$ (productivity), $n_h = 2$ (variable inputs), $n_j = 2$ (capital). For 56 we use $n_g = 3$ (productivity), $n_h = 3$ (variable inputs), $n_2 = 1$ (capital).

parameters to estimate, and under *subsector heterogeneity* the number of parameters is four times the number of three-digit subsectors. For eight out of the fourteen sectors listed in the table, the *estimated heterogeneity* specification is preferred. For five sectors, the *subsector heterogeneity* specification attains the lowest BIC value. Because the *subsector* specification is more densly parameterized than the *estimated heterogeneity* specification, it is conceivable that the result would be overturned if we allow for additional groups.

Estimated Parameters and Groupings. Table 4 contains estimates of firm-specific productivities $\alpha(\cdot)$, variable input elasticities $\beta(\cdot)$, and capital elasticities $\gamma(\cdot)$. We consider five non-overlapping samples. Each of the samples features substantial heterogeneity in productivity. The heterogeneity in variable input and capital elasticities in the early samples, 1965-74 and 1975-84 is less pronounced. These periods feature only one or two, instead of three, distinct estimate of $\beta(\cdot)$ and $\gamma(\cdot)$. From 1985 onwards, the amount of heterogeneity

Table 4: 2007-2916 Parameter Estimates: Manufacturing (NAICS 32)

	Productivity			Variable Input			Capital		
Sample	$\widehat{\alpha}(1)$	$\widehat{\alpha}(2)$	$\widehat{\alpha}(3)$	$\widehat{\beta}(1)$	$\widehat{\beta}(2)$	$\widehat{\beta}(3)$	$\widehat{\gamma}(1)$	$\widehat{\gamma}(2)$	$\widehat{\gamma}(3)$
1965-1974	.012	.015	.024	0.74	0.82	0.82	0.12	0.15	0.15
1975-1984	.045	.058	.066	0.83	0.83	0.84	0.10	0.11	0.12
1985-1994	.016	.055	.109	0.58	0.63	0.67	0.32	0.39	0.44
1995-2004	-.032	-.001	.028	0.78	0.83	0.94	0.14	0.16	0.20
2005-2014	.010	.270	.941	0.32	0.58	0.59	0.13	0.27	0.38

Table 5: Group Sizes: Manufacturing (NAICS 32), 2007-2016 Estimates, 2016 Firms

Panel (1)

	Productivity			Variable Input			Capital		
	$\widehat{\alpha}(1)$	$\widehat{\alpha}(2)$	$\widehat{\alpha}(3)$	$\widehat{\beta}(1)$	$\widehat{\beta}(2)$	$\widehat{\beta}(3)$	$\widehat{\gamma}(1)$	$\widehat{\gamma}(2)$	$\widehat{\gamma}(3)$
Estimate	-.080	0.118	0.676	0.43	0.55	0.72	0.25	0.51	0.59
Members	180	177	5	153	47	162	142	67	153

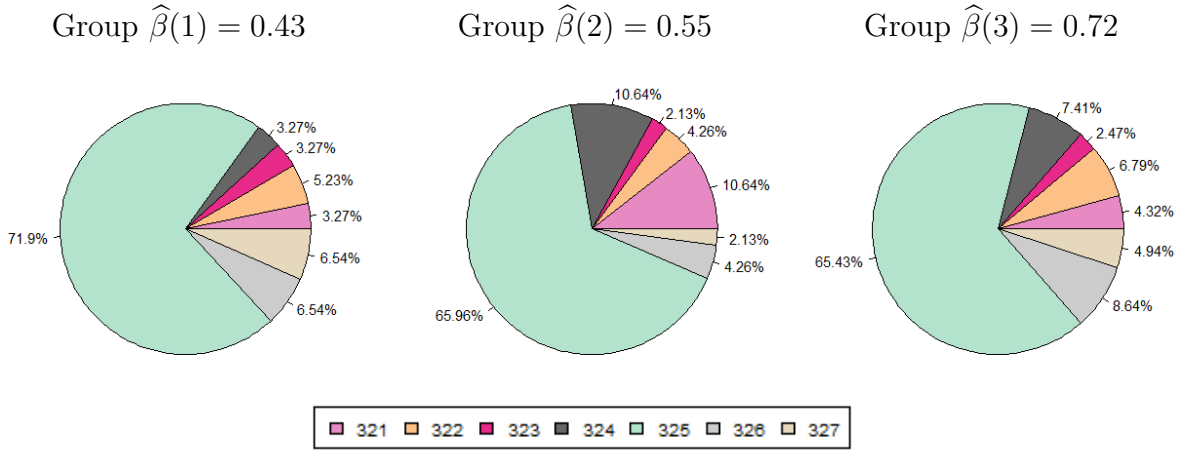
Panel (2)

	$\widehat{\alpha}(1)$			$\widehat{\alpha}(2)$			$\widehat{\alpha}(3)$		
	$\widehat{\beta}(1)$	$\widehat{\beta}(2)$	$\widehat{\beta}(3)$	$\widehat{\beta}(1)$	$\widehat{\beta}(2)$	$\widehat{\beta}(3)$	$\widehat{\beta}(1)$	$\widehat{\beta}(2)$	$\widehat{\beta}(3)$
$\widehat{\gamma}(1)$	12	1	8	89	19	11	1	0	1
$\widehat{\gamma}(2)$	3	9	27	19	0	9	0	0	0
$\widehat{\gamma}(3)$	19	15	86	9	3	18	1	0	2

appears to be increasing, as the parameter estimates for the three $\beta(\cdot)$ and $\gamma(\cdot)$ groups are quite different from each other.

The two panels of Table 5 provide information about the number of firms belonging to each of the groups. Here we focus on the 2007-16 sample estimates of parameters and group memberships. Because firms enter and exit the panels, we compute the number of group members for a particular year within the estimation sample, namely 2016. Panel (1) of the figure has the estimates of the group-specific parameters and the number of group members.

Figure 1: Group Composition: Manufacturing (NAICS 32), 2007-2016 Estimates, 2016 Firms



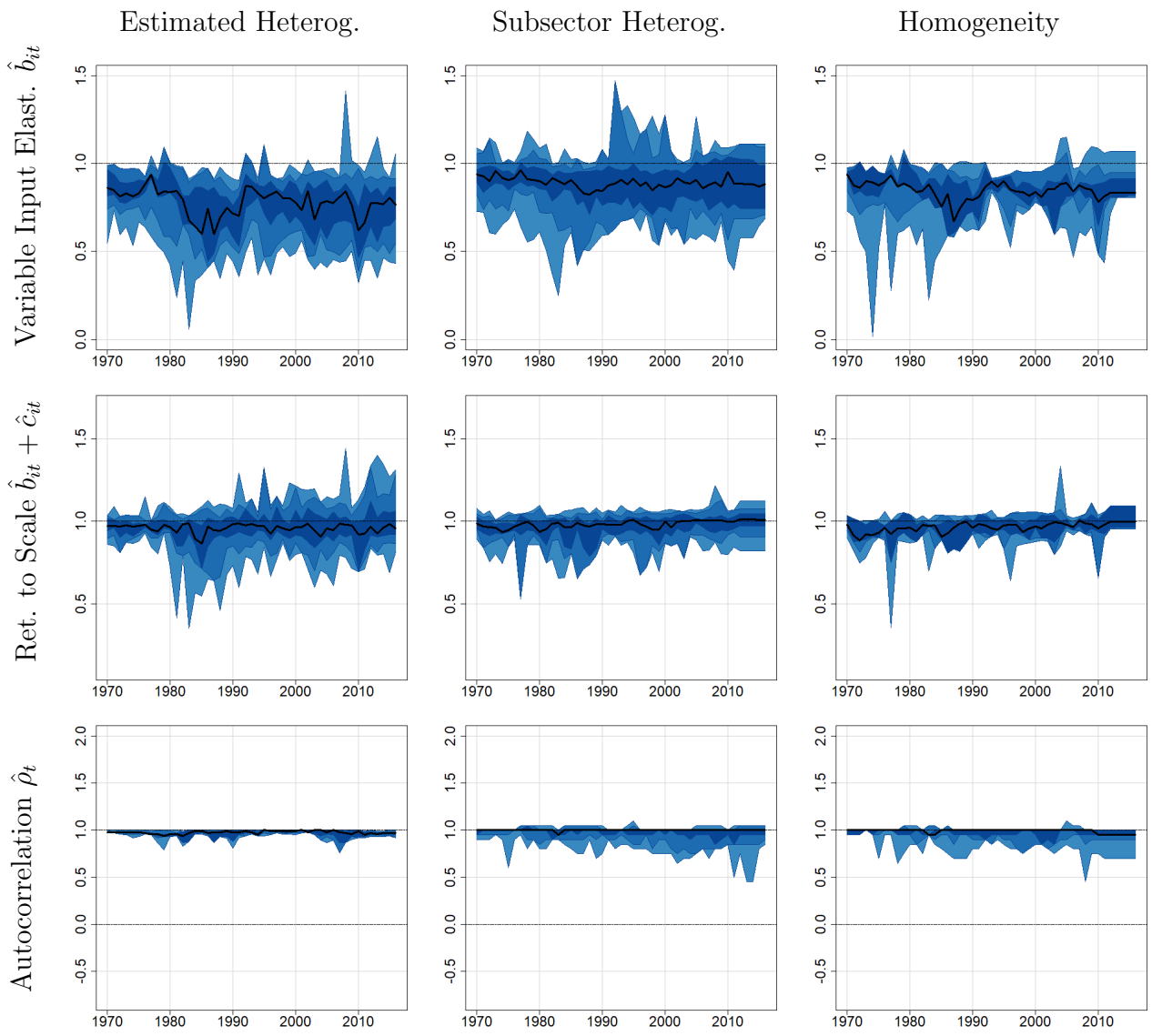
Notes: 321 = Wood Product Manufacturing, 322 = Paper Manufacturing, 323 = Printing and Related Support Activities, 324 = Petroleum and Coal Products Manufacturing, 325 = Chemical Manufacturing, 326 = Plastics and Rubber Products Manufacturing, 327 = Nonmetallic Mineral Product Manufacturing.

Except for the high-productivity group ($\hat{\alpha}(3) = 0.676$), which only has five members and capture probably some outliers in the sample, all other groups have a substantial number of observations, allowing us to sharply estimate the group-specific coefficients. Panel (2) reports the number of firms associated with the $3 * 3 * 3 = 27$ parameter combinations that can be formed based on the nine $\alpha(\cdot)$, $\beta(\cdot)$, and $\gamma(\cdot)$ estimates. The most striking feature is that the entries in the table are sparse, in the sense that many cells have less than 10 firms. As pointed out previously, there are very few high productivity firms. More interestingly, there are few firms with medium productivity, high capital elasticity and low or medium variable input elasticity. For these sparse configurations, a one-dimensional clustering strategy based on 27 groups would have been very inefficient. Our multi-dimensional approach allows us to "extrapolate" our estimates into these sparsely-populated cells.

Figure 1 depicts the composition of the three variable cost elasticity groups for 2016. Each segment of the pie chart corresponds to a different three-digit subsector of the Manufacturing sector 32. The figure shows that each of the 7 subsectors is represented in each group. In fact, the subsector shares are very similar across $\beta(\cdot)$ groups. Thus, the estimated classification is very different from the classification of the statistical agency.

Elasticity Estimates. The firm-specific markups depend on the elasticity estimates $\hat{b}_i$ and the average markup is a function of the distribution of the $\hat{b}_i$'s within and across industries; see (38) and (39). In the top row of Figure 2 we plot quantiles of the cross-sectional dis-

Figure 2: Quantiles of Estimated Elasticities Across Sectors



Notes: The graphs depicts the 10%, 25%, 50%, 75%, 90%, and 95% quantiles of the cross-sectional distributions of the estimated elasticities across all two-digit sectors included in the analysis.

tribution of the variable input elasticity estimates $\hat{b}_i$. The time series dimension of the plot traces out the sequence of rolling samples based on which we are estimating the production functions. The year on the x-axis corresponds to the midpoint (sixth observation) of each estimation sample. Because the our data set ends in 2016, the last five cross-sectional distributions for 2012 to 2016 are based on estimates from the 2007-16 sample.

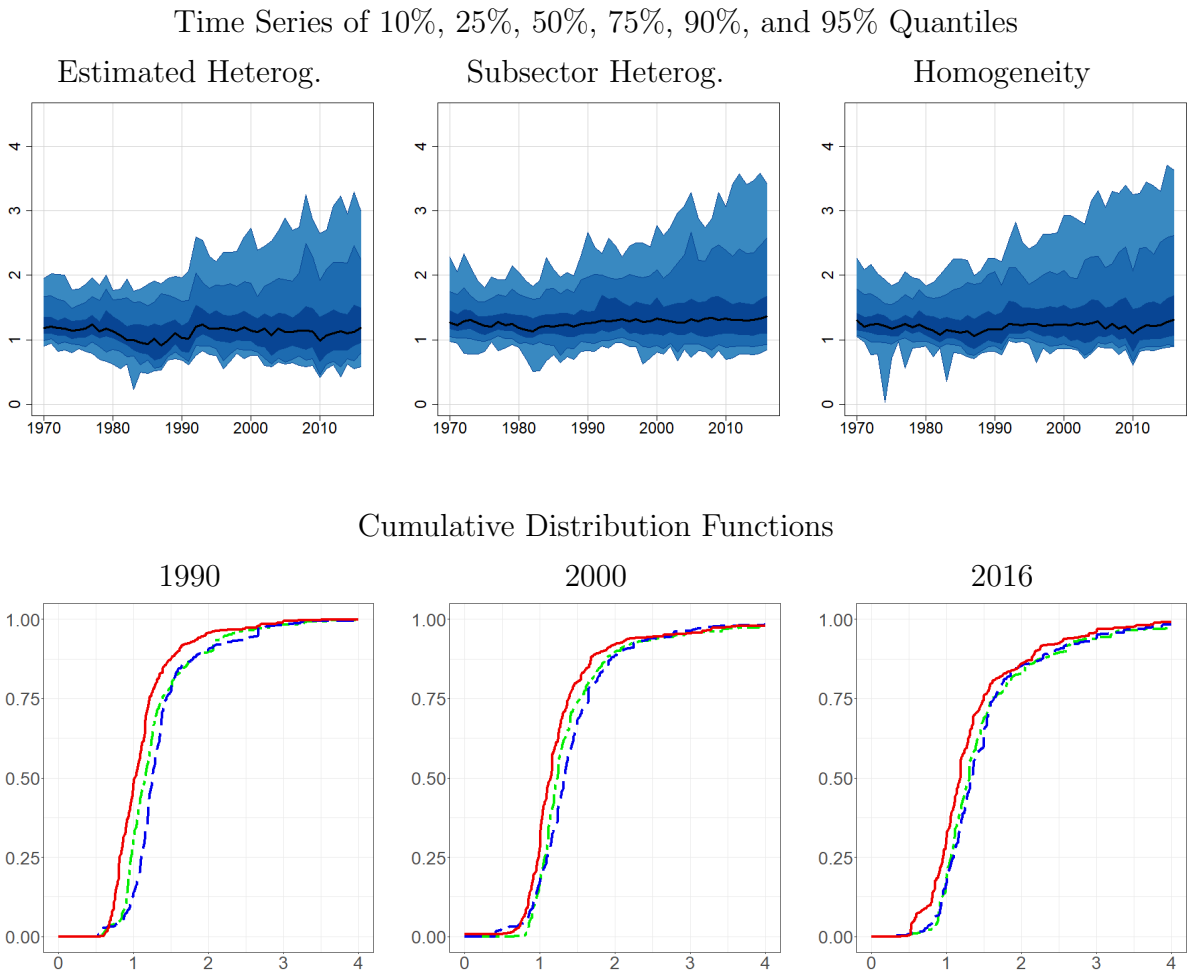
The $\hat{b}_i$ estimates are weighted by the market share of firm i in that particular year. Because market shares fluctuate over time and firms enter and exit, the distribution of parameter estimates between 2012 and 2016 varies, even though the underlying estimates $\hat{\alpha}(\cdot)$, $\hat{\beta}(\cdot)$, and $\hat{\gamma}(\cdot)$ are the same. The columns of subplots in the figure correspond to the three model specifications *estimated heterogeneity*, *subsector heterogeneity*, and *homogeneity*. Under *estimated heterogeneity* the $\hat{b}_i$ estimates are lower than under *subsector heterogeneity*. By construction the estimates that impose *homogeneity* are generally less dispersed because they are identical within sector.

The second row shows the evolution of the cross-sectional distribution of the returns to scale, $\hat{b}_{it} + \hat{c}_{it}$. The sequence of medians fluctuates slightly below one indicating that the median firm operates approximately with constant returns to scale. The dispersion of the returns to scale estimates is larger under *estimated* heterogeneity than under the other two specification. This is consistent with the interpretation that grouping firms incorrectly (or imposing homogeneity), leads to estimates that average over high and low population parameters and are not representative of the dispersion in the population. The last row of Figure 2 shows the quantiles of the autocorrelation estimates. Under *estimated* heterogeneity all $\hat{\rho}$'s are very close to one, whereas for the other two specifications the estimates in the bottom quantiles often fall below 0.8.

Markup Estimates. Figure 4 shows the cross-sectional distribution of estimated markups over time, weighted by the firms' market shares. The timing convention is the same as in Figure 2. Recall that the markups are obtained by scaling the $\hat{b}_i$'s by the revenue-to-variable-cost ratio; see (38). Because the elasticity estimates obtained from the *estimated heterogeneity* specification are lower than from the other two specifications, so are the markups. In the bottom panels we show empirical distribution functions for the years 1990, 2000, and 2016. The graphs indicate a clear stochastic dominance. In all three periods, the distribution function associated with *estimated heterogeneity* lies above the distribution functions obtained from the other two specifications, indicating that the estimated markups are lower.

Using (39) we now compute estimates of the average markup across the sectors considered in our analysis. The results are depicted in Figure 4. The main result is that the overall level of the aggregate markup is lower and the rise in the markup between 1970 and 2016 is less pronounced under *estimated heterogeneity*, than it is under *subsector heterogeneity* and *homogeneity*. Estimated slope coefficients from a simple deterministic time trend model imply that according to the *estimated heterogeneity* version markups have risen by approximately 0.4 percentages annually. Under the *homogeneity* specification the annual increase is

Figure 3: Distribution of Markups Across Sectors



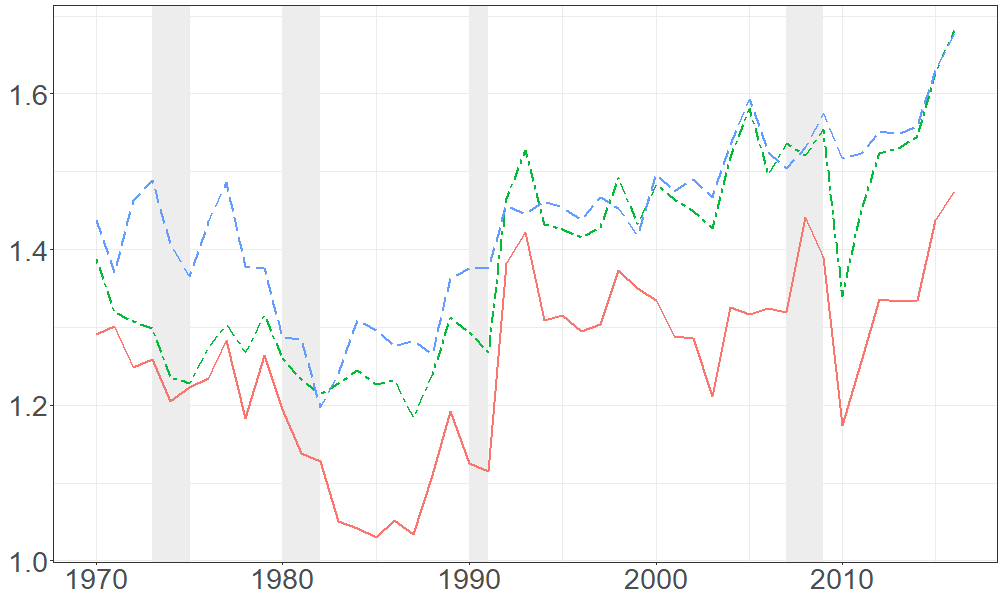
Notes: Top row: the graphs depict the evolution of the 10%, 25%, 50%, 75%, 90%, and 95% quantiles of the cross-sectional distributions of the estimated elasticities across all two-digit sectors included in the analysis. Bottom row: cumulative distribution functions for selected years based on estimated heterogeneity (red, solid), subsector heterogeneity (blue, dashed), homogeneity (green, dashed-dotted).

on average 0.7 percentages. Because our selection criterion prefers chooses the *estimated heterogeneity* specificaton for the majority of sectors, we regard the resulting markup estimates from this specification as more reliable.

7 Conclusion

Explicitly modeling and estimating heterogeneous parameters, as opposed to simply "differencing them out" and focusing exclusively on homogeneous parameters, is an important

Figure 4: Aggregate Markups



Notes: Estimated heterogeneity (red, solid), subsector heterogeneity (blue, dashed), homogeneity (green, dashed-dotted).

development in the panel data literature. Our paper contributes to this literature by developing a 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 characteristic of the unit. In the application, we clustered firms based on their productivity, and their elasticities of output with respect to variable inputs and capital. In our application we show that accounting for multi-dimensional group heterogeneity leads to lower estimates of the level and growth of aggregate markups than specifications that assume production technologies are homogeneous within two-digit NAICS sectors.

References

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.

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70(1), 191–221.

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

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

Cheng, X., Z. Liao, and F. Schorfheide (2016): "Shrinkage Estimation of High-Dimensional Factor Models with Structural Instabilities," *Review of Economic Studies*, 83(4), 1511–1543.

De Loecker, J., J. Eeckhout, and G. Unger (2018): "The Rise of Market Power and the Macroeconomic Implications," *Manuscript, KU Leuven, UPF Barcelona, and Harvard University*.

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

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.

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

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

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

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. (2018): "Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective," *arXiv preprint 1805.04178*.

Liu, R., A. Schick, Z. Shang, Y. Zhang, and Q. Zhou (2018): "Identification and estimation in panel models with overspecified number of groups," *Working Paper, Louisiana State University*, 2018-03.

Lu, X., and L. Su (2016): "Determining the Number of Groups in Latent Panel Structures with an Application to Income and Democracy," *Manuscript, Singapore Management University*.

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

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*.

Online Appendix for "Clustering for Multi-Dimensional Heterogeneity"

Xu Cheng, Frank Schorfheide, and Peng Shao

A Proofs

Proof of Lemma 3.1. Define the population criterion

$$\begin{aligned} Q_N(\theta, G, H) &= N^{-1}\sum_{i=1}^{N} Q_i(\theta, g_i, h_i), \text{ where} \\ Q_i(\theta, g_i, h_i) &= E[m\left(w_{it}; \alpha(g_i), \beta(h_i), \lambda\right)]' W_i E[m\left(w_{it}; \alpha(g_i), \beta(h_i), \lambda\right)]. \end{aligned} \tag{A.1}$$

By Assumption W and (14), we have the uniform convergence

$$\sup_{(\theta,G,H)\in\overline{\Theta}\times\Gamma_G\times\Gamma_H} |\widehat{Q}_N(\theta, G, H) - \widehat{Q}_N(\theta, G, H)| = o_p(1). \tag{A.2}$$

Define

$$\begin{aligned} d(\theta, G, H) &= N^{-1}\sum_{i=1}^{N} d_i\left(\theta_i\right), \text{ where} \\ d_i(\theta_i) &= \left(\alpha(g_i) - \alpha^0(g_i^0)\right)^2 + \left(\beta(h_i) - \beta^0(h_i^0)\right)^2 + ||\lambda - \lambda^0||^2. \end{aligned} \tag{A.3}$$

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

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

Given that θ_i has a compact support Θ for all i, let $C = \sup_i \sup_{\theta_i\in\Theta} d_i(\theta_i) < \infty$. Let $S = \{i : d_i(\theta_i) > \delta/2\}$ and $N_S = \sum_{i=1}^{N} 1\{i \in S\}$. 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,

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

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 $\varepsilon = \delta\varepsilon^*/(2C)$. Results in (A.5) is analogous to Lemma A.4 in Liu et al. (2018).

Finally, we show the consistency result by combining (A.2) and (A.4). For any $\delta > 0$, there exists $\varepsilon > 0$, such that

$$\begin{aligned} P\left\{d(\widehat{\theta},\widehat{G},\widehat{H}) > \delta\right\} &\leq P\left\{Q_N(\widehat{\theta},\widehat{G},\widehat{H}) \geq \varepsilon\right\} \\ &= P\left\{d_1 + d_2 + d_3 \geq \varepsilon\right\}, \end{aligned} \tag{A.6}$$

where

$$\begin{aligned} d_1 &= Q_N(\widehat{\theta},\widehat{G},\widehat{H}) - \widehat{Q}_N(\widehat{\theta},\widehat{G},\widehat{H}), \\ d_2 &= \widehat{Q}_N(\widehat{\theta},\widehat{G},\widehat{H}) - \widehat{Q}_N(\theta^0,G^0,H^0), \\ d_3 &= \widehat{Q}_N(\theta^0,G^0,H^0) - Q_N(\theta^0,G^0,H^0). \end{aligned} \tag{A.7}$$

Because $d_2 \leq 0$ by definition of the estimator and $d_1 = o_p(1)$ and $d_3 = o_p(1)$ by (A.2), (A.6) implies that $P\{d(\widehat{\theta},\widehat{G},\widehat{H}) > \delta\} \to 0$ 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 BM. The arguments can be applied to α and β separately in our set-up. There is no need to take sample average here because our parameters are not time-varying. Lemma B.3 also shows how to relabel the groups and shows that this is a one-to-one mapping with probability approaching 1. □

Proof of Theorem 4.1. Let $E_W = 1\{\max_i \|W_{iNT} - W_i\| \leq \eta\}$ for some small 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 (16)-(18) that

$$\begin{aligned} & P\left\{\widehat{g}_i = g_i, \widehat{h}_i = h_i\right\} \\ \leq\ & P\left\{\widehat{Q}_i(\widehat{\theta},g_i,h_i) < \widehat{Q}_i(\widehat{\theta},g_i^0,h_i^0)\right\} \\ \leq\ & P\left\{c_1\left\|\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta},g_i,h_i)\right\|^2 \leq c_2\left\|\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta},g_i^0,h_i^0)\right\|^2\right\} \end{aligned} \tag{A.8}$$

for constants $c_2 > c_1 > 0$. Using the decomposition in (19) and the triangle inequality,

$$\left\|\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta},g_i,h_i)\right\|^2 \geq \left\|b_i(\widehat{\theta},g_i,h_i)\right\| - \left\|\delta_i(\widehat{\theta},g_i,h_i)\right\|^2, \tag{A.9}$$

where

$$\begin{aligned} \delta_i(\widehat{\theta},g_i,h_i) &= \frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta},g_i,h_i) - E[m_{it}(\widehat{\theta},g_i,h_i)], \\ b_i(\widehat{\theta},g_i,h_i) &= E[m_{it}(\widehat{\theta},g_i,h_i)]. \end{aligned}$$

By a similarly decomposition,

$$\left\|\frac{1}{T}\sum_{t=1}^{T} m_{it}(\widehat{\theta}, g_i^0, h_i^0)\right\|^2 \leq \left\|b_i(\widehat{\theta}, g_i^0, h_i^0)\right\| + \left\|\delta_i(\widehat{\theta}, g_i^0, h_i^0)\right\|^2. \tag{A.10}$$

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

$$\begin{aligned} \left\|b_i(\widehat{\theta}, g_i, h_i)\right\|^2 &= \left\|E[m_{it}(\widehat{\theta}, g_i, h_i)] - E[m_{it}(\theta^0, g_i^0, h_i^0)]\right\|^2 \\ &\geq b_{1,i}\left(\alpha_g^0, \beta_h^0\right) - b_{2,i}\left(\alpha_g, \beta_h\right) \end{aligned} \tag{A.11}$$

where

$$\begin{aligned} b_{1,i}\left(\theta^0, g_i, h_i\right) &= \left\|E[m_{it}\left(\theta^0, g_i, h_i\right)] - E[m_{it}(\theta^0, g_i^0, h_i^0)]\right\|^2, \\ b_{2,i}\left(\widehat{\theta}, g_i, h_i\right) &= \left\|E[m_{it}(\widehat{\theta}, g_i, h_i)] - E[m_{it}(\theta^0, g_i, h_i)]\right\|^2, \end{aligned} \tag{A.12}$$

where the first term $b_{1,i}(\theta^0, g_i, h_i)$ is due to misspecification of group 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,

$$b_{1,i}\left(\theta^0, g_i, h_i\right) \geq m_0 \tag{A.13}$$

for some $m_0 > 0$ for any $(g_i, h_i) \neq (g_i^0, h_i^0)$. By Assumption R(iii),

$$b_{2,i}\left(\widehat{\theta}, g_i, h_i\right) \leq M_0\eta^2 \tag{A.14}$$

for some $M_0 < \infty$. Therefore,

$$\left\|b_i(\widehat{\theta}, g_i, h_i)\right\|^2 \geq m_0 - M_0\eta^2. \tag{A.15}$$

Similarly, we have

$$\begin{aligned} \left\|b_i(\widehat{\theta}, g_i^0, h_i^0)\right\| &= \left\|E[m_{it}\left(\widehat{\theta}, g_i^0, h_i^0\right)] - E[m_{it}(\theta^0, g_i^0, h_i^0)]\right\|^2 \\ &\leq M_0\eta^2. \end{aligned} \tag{A.16}$$

Combining (A.8) with (A.9), (A.10), (A.15), (A.16), we obtain

$$\begin{aligned} & P_{i,gh}(\widehat{\theta}) \\ \leq\ & P\left\{c_1 m_0 - c_1 M_0\eta^2 - c_2 M_0\eta^2 \leq c_1\left\|\delta_i(\widehat{\theta}, g_i, h_i)\right\|^2 + c_2\left\|\delta_i(\widehat{\theta}, g_i^0, h_i^0)\right\|^2\right\}. \end{aligned} \tag{A.17}$$

Take $\eta > 0$ small enough such that

$$s = c_1 m_0 - c_1 M_0 \eta^2 - c_2 M_0 \eta^2 > 0. \tag{A.18}$$

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. Under Assumption R,

$$\begin{aligned} \max_{1\le i\le N} P\left\{ c_1 \left\| \delta_i(\widehat{\theta}, g_i, h_i) \right\|^2 \ge s \right\} &= o(N^{-1}), \\ \max_{1\le i\le N} P\left\{ c_2 \left\| \delta_i\left(\widehat{\theta}, g_i, h_i\right) \right\|^2 \ge s \right\} &= o(N^{-1}), \end{aligned} \tag{A.19}$$

by Lemma S1.2(ii) of SSP. Therefore, for any $(g_i, h_i) \neq (g_i^0, h_i^0)$,

$$\max_{1\le i\le N} P\left\{ \widehat{g}_i = g_i, \widehat{h}_i = h_i \right\} = o(N^{-1}) \tag{A.20}$$

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

$$\max_{1\le i\le N} P\left\{ \widehat{g}_i \neq g_i^0, \widehat{h}_i \neq h_i^0 \right\} = o(N^{-1}) \tag{A.21}$$

for $\widehat{\theta} \in N_\eta$.

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

$$\begin{aligned} & P\left\{ \widehat{G} = G^0 \text{ and } \widehat{H} = H^0 \right\} \\ = \;& 1 - P\left\{ 1\left\{ (\widehat{g}_i, \widehat{h}_i) \neq (g_i^0, h_i^0) \right\} \text{ for some } i \right\} \\ \ge \;& 1 - N \max_{1\le i\le N} P\left\{ (\widehat{g}_i, \widehat{h}_i) \neq (g_i^0, h_i^0) \right\} \\ \to \;& 1. \end{aligned} \tag{A.22}$$

By Lemma 3.2 and Assumption W, $P\{\widehat{\theta} \in N_\eta\} \to 1$ and $P\{E_W = 1\} \to 1$, which gives the desirable result together with (A.22). □

Proof of Theorem 4.2. 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.,

$$\overline{\theta} = \arg\min_{\theta\in\overline{\Theta}} \overline{Q}(\theta), \text{ where } \overline{Q}(\theta) = \overline{m}(\theta)' W_{NT} \overline{m}(\theta), \tag{A.23}$$

with

$$\overline{m}(\theta) = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} m\left(w_{it}; \alpha(g_i^0), \beta(h_i^0), \lambda\right). \tag{A.24}$$

Now we derive the asymptotic distribution of $\overline{\theta}$. This is a standard GMM problem. By Assumption ID, E(ii), and (14), 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 $\eta > 0$,

$$N^{-1}\sum_{i=1}^{N}\sup_{||\theta_i-\theta_i^0||\leq\eta}\left\|T^{-1}\sum_{t=1}^{T}m_\theta\left(w_{it};\theta_i\right)-E\left[m_\theta\left(w_{it};\theta_i\right)\right]\right\|\to_p 0 \tag{A.25}$$

and

$$(NT)^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}m\left(w_{it};\theta_i^0\right)\to_d N(0,\Omega) \tag{A.26}$$

as $N,T\to\infty$. The first result in (A.25) follows from a uniform convergence over i, which is obtained by applying Lemma S1.2(iii) of SSP under Assumption R and E(iii). The second result in (A.26) follows from verifying a Lindeberg-Feller central limit theorem. Lemma S1.12 of SSP proves a result of the same form and provide the details of the verification, see p.29 of the Supplement to SSP. This completes the proof. □

Verification of Assumptions for the Production Function Example.

We first verify Assumption ID. For any $\theta_i=(a_i,b_i,c_i,\rho)$, we have

$$\Delta y_{it}(\rho)=a_i^0(1-\rho)+b_i^0v_{it}(\rho)+c_i^0k_{it}(\rho)+(\rho_0-\rho)\,\omega_{it-1}+\xi_{it}+\varepsilon_{it}-\rho\varepsilon_{it-1}, \tag{A.27}$$

and

$$\begin{aligned}&E\left[z_{it}\left(\Delta y_{it}(\rho)-a_i(1-\rho)-b_i\Delta v_{it}(\rho)-c_i\Delta k_{it}(\rho)\right)\right]\\ =\;&E\left[z_{it}\left(\left(a_i^0-a_i\right)(1-\rho)+\left(b_i^0-b_i\right)\Delta v_{it}(\rho)+\left(c_i^0-c_i\right)\Delta k_{it}(\rho)+(\rho_0-\rho)\,\omega_{it-1}+\ldots\right)\right]\end{aligned} \tag{A.28}$$

under condition (i). Assumption ID holds under $\mu_{\min}(E[z_{it}x_{it}(\rho)]')\geq\delta>0$ and $\rho<1$. Assumption R(i)-R(ii) holds automatically under conditions (i) and (ii). The first order derivative is

$$m_\theta\left(w_{it};\theta_i\right)=-z_{it}[(1-\rho),\Delta v_{it}(\rho),\Delta k_{it}(\rho),y_{it-1}-a_i-b_iv_{it-1}-ck_{it-1}]. \tag{A.29}$$

Assumption R(iii) and E(iii) holds under $E||z_{it}d_{it}||^q\leq C$. □