Factor income shares and input distortions in China *

Xiaoyue Zhang† Junjie Xia ‡

April 18, 2023

The Latest Version

Abstract

This paper quantifies the impact of input distortions on the labor and capital shares in the aggregate revenue of Chinese industry with heterogeneous productivity, technology, and demand elasticities. The input distortions include anything that prevents firms from using production factors at their market prices or from freely adjusting the usage. Using parameters estimated from firm-level data, we find that removing the input distortions in Chinese industry would raise the industrial labor share by 20 percentage points, and lower the industrial capital share by 1 percentage point. This implies that improving allocation efficiency through removing input distortions can create socioeconomic shocks such as a massive inflow of labor to industry. The distributional impact of input distortions also suggests that improving the allocation efficiency of inputs may not always be the solution to close the gap between developing and developed countries.

Keywords: Input distortions, distributional impact, factor income shares, latent market structures, heterogeneous demand elasticities, product differentiation

*We are grateful to Jaap H. Abbring, Christoph B. T. Walsh, and Jeffrey R. Campbell for their patience, support, and guidance throughout this project. This paper has benefited from insightful discussions with and comments from Daniel Xu, Has van Vlokhoven, Malik Çürük, Guiying Laura Wu, Chang-Tai Hsieh, Giuseppe Forte, Jo Van Biesebroeck, Jan De Loecker, Frank Verboven as well as participants at the Tilburg Structural Econometrics Group, Tilburg University GSS seminars, the IO seminar series at KU Leuven, the ENTER Jamboree, China International Conference in Macroeconomics, the IAAE annual conference, the AMES China meeting, China Center for Economic Research's Summer Institute, the NSE International Annual Workshop, and EEA-ESEM. This project is supported by CentER and IAAE Student Travel Grant. Xiaoyue is also thankful for the never-ending support and confidence from Edi Karni.

†Tilburg University, corresponding author, email: x.zhang_11@uvt.nl

‡Central University of Finance and Economics and Peking University, email: junjiexia@nsd.pku.edu.cn

1 Introduction

This paper estimates the impact of input distortions on the labor and the capital shares in the aggregate revenue of Chinese industry with heterogeneous productivity, technology, and demand elasticities. The input distortions include anything that prevents firms from using capital and labor at their market prices or from freely adjusting their usage such as favorable interest rates enjoyed by state-owned enterprises (SOEs) or frictions in adjusting labor. Restuccia and Rogerson (2008), Hsieh and Klenow (2009) (hereafter HK), and Zhang and Xia (2022) document that these distortions are pervasive, and can generate 30% to 50% decreases in total factor productivity (TFP) in large economies such as China. However, the input distortions can affect workers and the owners of capital differently. While the size of the efficiency gains from removing the input distortions is well understood, the distributional impact remains unclear. We study this distributional impact by estimating the changes in the labor and the capital shares when reallocating the factors to their more productive use. Using demand and production parameters estimated from firm-level data, we find that removing the input distortions in Chinese industry would raise the industrial labor share by 20 percentage points, and lower the industrial capital share by 1 percentage point in 2005 if the aggregate industrial labor and capital are constant. This implies that the significant reduction in input distortions due to the 1978 Reform and Opening-up in China together with strict restrictions on labor mobility across sectors may have caused the rapid increase in the industrial labor share after the 1978 Reform and Opening-up. The 20-percentage-point increase in the labor share also suggests that improving allocation efficiency through removing input distortions can create socioeconomic shocks such as a massive inflow of labor to industry. Combining the predicted factor share changes estimated in this paper with the predicted TFP gains estimated in Zhang and Xia (2022), removing the input distortions may reduce the returns to economic groups crucial for future growth, such as entrepreneurs or investors.

We build on HK, one of the most cited papers in estimating efficiency gains from removing input distortions. In HK's economy, firms are divided into nests with each nest being an industry. Firms inside each nest interact in monopolistic competition and face constant demand elasticities, which are assumed to be 3 for all the nests. We allow the possibility that industries can have a high-demand-elasticity nest and a low-demand-elasticity nest, and that each nest has its own demand elasticity. This modification captures the different levels of product differentiation across and within industries, which, according to Autor et al. (2020), have a large impact on the aggregate labor share. We include unexpected cost shocks that are realized after firms set prices and choose factor usage so that the distributions of

firms' markups are nondegenerate. Together with heterogeneous demand elasticities, the unexpected cost shocks rationalize variation in firms' markups. The latent nest structure within industries and the nest-specific demand elasticities will be estimated using firm-level data.

Instead of using production parameters of a benchmark economy, the common practice in the literature on misallocation, we use parameters estimated from firm-level data and show that using parameters from the US, the most commonly used benchmark economy, implies that, contrary to our results, the input distortions cause the underuse of capital and the overuse of labor in China, which reflects mainly the technology differences, i.e. American firms are specialized in more capital-intensive production while Chinese firms are more labor-intensive. Therefore, instead of an increase in the labor share and a decrease in the capital share found in our paper, the labor share would decrease by 3 percentage points and the capital share would increase by 31 percentage points. These differences in predicted changes of factor shares is due to the differences between American and Chinese firms' production elasticities. When using American firms' production elasticities to calibrate those of Chinese firms', technology differences contaminate the measured input distortions, which causes a systematic upward bias in measured capital distortions, and a systematic downward bias in measured labor distortions. These systematic biases in measured input distortions do not affect within-nest TFP gains (Zhang and Xia (2022)), a question that HK is interested in, because within-nest TFP gains depend on the dispersion of firm-level factor shares. Since the biases are the same for an input inside a nest, the biases do not affect the dispersion. However, this paper shows that the biases are carried over to the predicted factor-share changes and flip the results because both the dispersion and the levels of input distortions matter for factor-share changes.

We model input distortions as firm-specific wedges in factor prices, following Restuccia and Rogerson (2008). Positive wedges indicate that firms use less factors than in an economy free of input distortions. Negative wedges indicate higher usage. These distortions can be either monetary such as taxes and subsidies, or non-monetary such as limited or no access to the financial market or frictions in adjusting labor. We do not specify to whom the revenues collected through input distortions belong. These revenues are transferred as a lump sum to a representative consumer if the input distortions are monetary. When the input distortions are non-monetary, those revenues are retained in firms. Since the representative consumer owns all the firm, the revenues collected due to non-monetary input distortions are income of the representative consumer. Therefore, revenues collected via both monetary and non-monetary input distortions go to the representative consumer.

We use firms' industry categories and revenue-cost ratios observed in the data to identify

the latent nest structure and the nest-specific demand elasticities. In our model, firms' markups are determined by their demand elasticities and unexpected idiosyncratic cost shocks that are realized after firms set prices. The cost shocks follow a nest-specific normal distribution, and markups are measured using revenue-cost ratios, which are observed. When an industry contains two nests, the distribution of revenue-cost ratios for firms in this industry is a mixture of two normal distributions. Since each industry has either one or two nests, we use the homogeneity test for mixture normal distributions to infer whether an industry contains two nests or only one nest. We then estimate demand elasticities using maximum likelihood estimation (MLE).

We estimate the production parameters using the firm-level capital shares and labor shares in 2005. Production functions are assumed to be Cobb-Douglas. Apart from the Hicks-neutral firm-specific productivities, the production functions are the same within industries, i.e. constant production elasticities, because firms operating in the same industry have very similar production technology. The key identification assumption is that the modes of the capital and labor distortions within an industry are both zero. This means the modes of the labor and capital shares within each industry are their respective production elasticities. This is because under the assumption of constant production elasticities within an industry, and after controlling for demand elasticities, variation in labor and capital shares of firms from an industry is variation in labor and capital distortions. We can then estimate production elasticities and the input distortions. The merit of this zero-mode assumption is that we do not assume ex ante that input distortions are mean zero. This allows the possibility that all the firms in an industry have positive input distortions or similarly negative input distortions. In reality, this can happen when, for example, certain industries are deemed strategic, and therefore most or all the firms inside this industry receive favorable interest rates. Besides, this assumption allows the distribution of positive distortions to differ from that of negative distortions, and also allows the distributions to vary across industries. This captures the idea that the mechanism behind positive distortions can be different from that behind negative ones, and that the mechanism may vary across industries.

Methods of estimating production functions using firm-level panel data while taking into account endogeneity caused by unobserved productivity (Olley and Pakes (1996), Levinsohn and Petrin (2003), and Ackerberg et al. (2015)) cannot be applied to estimating the production parameters when firms face idiosyncratic input distortions because the crucial assumption of input usage as a monotonic function of productivity is violated. If input distortions are positively correlated with productivity, higher-productivity firms no longer hire more inputs. For the same reason, we cannot use these methods to calculate firm-level markups as the ratios between production elasticities and input shares.

Applying our model to the Chinese Annual Survey of Industry in 2005, we find that 90% of the industries are better modeled as having two nests. For these industries, the demand elasticities of the high-demand-elasticity nests are in most cases at least twice as large as those of the low-demand-elasticity nests. In 95% of these two-nest industries, there are more firms in the high-demand-elasticity nest than the low-demand-elasticity nest, suggesting that a high level of product differentiation is difficult to achieve and that only few firms succeed. When comparing across industries, estimated demand elasticities are lower in nests from industries that tend to have strong market power such as oil and gas extraction, tobacco manufacturing, and pharmaceutical manufacturing. Demand elasticities vary across nests with the top 10 percentile about three times lager than the bottom 10 percentile. On average, firms' behaviors display constant returns to scale but with large variation across industries. The distribution of capital distortions are more skewed to the left compared to that of labor distortions suggesting that capital is more frequently overused due to, for example, subsidies.

Removing all the input distortions would raise the labor share by 20 percentage points but reduce the capital share by 1 percentage point. Both the size of the input distortions and the heterogeneity across nests matter. If we would be in a world where there is only one representative industry and the demand elasticities within the industry are constant, or in other words only one representative nest, removing the input distortions would increase the labor share by 7 percentage points and reduce the capital share by 8 percentage points. The comovement between input distortions measured at the nest level and parameters describing nest heterogeneity, which includes production technologies, expenditure shares, and demand elasticities, explains the remaining two thirds of the labor share change and offsets the reduction in the capital share to 1 percentage point. An example of a strong comovement is that nests that would experience larger changes in factor shares also have a larger market expenditure, a higher demand for the factor due to its technology, or greater demand elasticities. Among these three dimensions of heterogeneity, we find that heterogeneity of production technologies has the largest impact on the predicted changes of factor shares.

The results in this paper offer a tentative explanation that the increase in Chinese industrial labor share in 1978-1995 may have been caused by the 1978 Reform and Opening-up (hereafter the 1978 Reform), the most important economic reform in modern-day China, together with strict restrictions on labor mobility. After thirty years of central planning and domination by the SOEs, the 1978 Reform initiated the transition to allocating production factors, especially in industry, based on market rules. This more market-based allocation systematically reduced input distortions, and consequently unleashed more than two decades of 2-digit growth. Meanwhile, it was still very difficult if not impossible for agricultural workers to enter industry, so the aggregate supply of industrial labor is by and large fixed in 1978-

1995. According to our results, removing the input distortions would increase the industrial labor share when the total industrial labor supply is fixed. During 1978-1995, Chinese industrial labor share increased from 35% to 49%. After major relaxation of the restrictions on labor mobility in the late 1990s, this increase slowed down. Although further studies are required before claiming that the reduction in input distortions indeed caused the increase in the industrial labor share in 1978-1995, our results suggest a possible link. Furthermore, if all the input distortions in industry are removed and China brings back its rigid control over labor mobility, our results indicate that a further increase up to 20 percentage points in the industrial labor share in 2005 is possible, which is higher than but comparable to the 14-percentage-point increase in 1978-1995.

The 1978 Reform coincides with the gradual relaxation of restrictions that prevent agricultural labor from migrating to industry but the main relaxation did not take place until the late 1990s. In spite of significant improvements, obstacles to labor mobility across sectors still exist in 2005, and some even exist nowadays such as difficult access to health care and, for the children of the migrating families, no or little access to public primary and middle schools. Predicting the influx of labor to industry after removing industrial input distortions is difficult because there is no good way to accurately measure the level of labor mobility across sectors. However, people can use the predicted changes in factor shares under the extreme case where the total supplies of factors are fixed to infer the possibility of a large-scale labor influx. In 1998-2005, input distortions in China declined by 15% (Hsieh and Klenow (2009)). At the same time, migrant workers leaving agriculture increased by 60% to 210 million, reaching 15% of the total population in 2005 (Ministry of Agriculture). Our results suggests that removing all the input distortions in China in 2005 may trigger another wave of migrant workers from agriculture that is at least comparable to the one observed in 1998-2005. Since the predicted 20-percentage-point increase is comparable to the 14-percentage-point increase in 1978-1995 and because the large influx of labor from rural to urban areas was triggered after restrictions in labor mobility were significantly relaxed, chances that removing the input distortions in 2005 causes mass migration between rural and urban areas are high. This mass migration can cause various socioeconomic issues such as overcrowded cities. While existing studies focus on the causes and the TFP costs of input distortions, the side effects of inducing a mass migration and its socioeconomic consequences have been ignored. This paper is the first to touch on this worrying large-scale socioeconomic shock by estimating the impact of input distortions on the industrial labor share. The factor shares and input distortions in the rest of this paper refer to industrial factor shares and input distortions in industry unless specified otherwise.

Recent studies on explaining why aggregate labor shares decline in the majority of coun-

tries in the world, and in the US in particular, motivate us to examine factor shares under rich firm and industry heterogeneity. In our model, the shares are affected by the size of the input distortions, the joint distribution of production technology, demand elasticities, productivity, and nests' expenditure shares. We draw inspirations from this literature and complement it by studying changes in labor shares combining insights from multiple studies. Piketty (2014) and Blanchard et al. (1997) show that distortions in the labor market are pervasive. If we consider input distortions as wedges that modify the relative prices of capital and labor, Karabarbounis and Neiman (2014) confirms the importance of input distortions in explaining the aggregate labor share. We interpret the impact of technological progress on labor shares studied by Blanchard et al. (1997), Karabarbounis and Neiman (2014), and Autor and Salomons (2018) as evidence that labor shares react to differences in technology, and therefore take into account heterogeneous technology and productivity. Dispersion in demand elasticities also has an impact especially when factors including both labor and capital are reallocated across firms with different markups (Basu (2019), Autor et al. (2020), De Loecker et al. (2020), and Hopenhayn et al. (2022)).[1] Elsby et al. (2013) points out that the aggregation structure may veil substantial changes at the industry level for US labor shares, while we find a similar offsetting effect for Chinese capital shares.

To understand what drives the factor-share changes at the aggregate level, we decompose the share changes into predicted changes in an economy with one representative nest and changes due to the covariance between input distortions and nest heterogeneity. The decomposition results in this paper provide more empirical evidence to enrich our understanding about how micro-level disturbances are aggregated to the macro-level changes, which has been examined in a different context by Olley and Pakes (1996) (efficiency gains identified in panel data), Edmond et al. (2019) (predicted gains from increasing competition), and Autor et al. (2020) (the decline in the US labor share).

Our study complements the literature on heterogeneous markups by estimating the heterogeneous levels of product differentiation and by offering a way to infer the latent market structure. The existing literature interprets variation in markups as variation in firms' productivity under incomplete pass-through and focuses on changes in markups over time (Atkeson and Burstein (2008), Edmond et al. (2015), and Burstein et al. (2020) use nested CES with oligopolies, Klenow and Willis (2016) uses Kimball preferences, Feenstra and Weinstein (2017) uses translog preferences, Mrázová et al. (2021) uses the Constant Revenue Elasticity of Marginal Revenue demand, and Haltiwanger et al. (2018) uses hyperbolic absolute risk aversion preference). In these frameworks, more productive firms have higher

[1] Koh et al. (2020), Karabarbounis and Neiman (2014), and Gutirrez and Piton (2020) show measurement issue explains a large share of the decline but disagree on whether the decline is purely a measurement issue.

markups. However, if varying markups reflect different levels of product differentiation, it is possible that less productive firms have higher markups because achieving a higher level of product differentiation costs resources that could otherwise be used for production and, consequently, these firms will be estimated as having lower productivity. In our firm-level data, variation in productivity can barely rationalize the variation in markups. Since our data is cross-sectional and contains a large variety of firms from mining, manufacturing, and public utilities, we argue that heterogeneous levels of product differentiation are the major source of markups variation and assume complete pass-through. Another assumption the existing literature makes is that the market structure is perfectly observed. However, the underlying market structure is usually not directly observed. Our paper takes this into account and infers the latent market structure under heterogeneous markups.

The remainder of the paper is organized as follows. We introduce the data set in Section 2 and describes our theoretical model and the decomposition framework in Section 3. We discuss our identification procedure in Section 4, and present our estimation results in Section 5. Section 6 concludes. The Appendix provides the derivation of theoretical results, details of estimation procedures, the identification steps, and a model extension that includes intangible assets.

2 Data

Our data source is the Chinese Annual Survey Data for Industries in 2005 collected by the National Bureau of Statistics of China.[2] This data set has been used by previous studies including HK, Song et al. (2011), and David and Venkateswaran (2019). It includes all the SOEs and nonstate firms with revenue above 5 million RMB ($600,000), which are in total about 230,000 firms. It covers manufacturing, mining, and public utilities. The industry classification used in this paper contains 523 industries.

The data set contains rich information on firm-level value-added, wage expenditure, net value of fixed assets, sales, and cost. When cleaning the data, we follow Brandt et al. (2012) to drop unreasonable observations accounting-wise, such as negative value added, negative debts, negative sales, et cetera. A full list of the types of observations dropped is provided in Appendix A. We calculate the net present value of depreciated real capital also following that paper. We trim the 1% tails of value added, labor and capital share of value added, revenue-cost ratio, capital, and labor. We do not trim the tails of profits because trimming the tails of revenue-cost ratio should already deal with abnormal profits. In the cleaned data, 15% of the firms have negative profits. Table 1 provides the summary statistics. Value

[2]We acquire the data through a data center at Peking University.

added is the amount of revenues after netting out expenditures other than capital and labor. K is the net present value of depreciated real capital and wL is labor expenditure.

TABLE 1: Summary Statistics of Cleaned Data (2005)

Statistic	N	Mean	Min	Pctl(25)	Median	Pctl(75)	Max
value added	229,061	13,814.46	122	2,517	5,377	13,250	277,908
K	229,061	16,366.41	83.76	1,620.23	4,211.66	12,151.88	515,954.20
wL	229,061	2,730.73	80	583	1,188	2,665	78,956
revenue	229,061	50,184.74	2	9,500	19,457	45,994	11,041,153
cost	229,061	43,075.61	1	7,935	16,481	39,072	10,757,115
profits	229,061	2,370.47	−292,087	72	480	1,815	415,879
revenue/cost	229,061	1.21	0.81	1.08	1.14	1.25	4.68
wL/value added	229,061	0.32	0.01	0.12	0.23	0.42	3.15
wL^c/value added	229,061	0.850	0.033	0.310	0.621	1.108	8.338

Notes: wL^c is the corrected labor share

One well-known limitation of this data is that the labor expenditure does not include the non-wage portion and that the aggregate labor share of value added is too low compared to the one inferred by the Chinese Input-and-Output Table and national accounts. HK scales up each firm's labor expenditure by the same proportion so that the aggregate labor share reaches 50%. Instead of assuming a 50% aggregate labor share, we correct the labor share so that the sum of the labor share, the capital share, the profits share, and the share collected through input distortions are 1. In theory, these shares should sum to 1, but, without correction, their sum is less than one. This gap should be the missing labor share and the input distortions collected from the missing labor share. We assume a constant ratio of unreported and reported labor shares across firms, similar to HK. After correction, the aggregate labor share is 52%.

3 Model

The economy is modeled as consisting of S industries whose names are s_1, s_2, ...,s_S. An industry is referred to as s when its name is not specified. Firms, labeled as i, inside an industry s have the same production elasticities α_s^K and α_s^L but differ in their Hicks-neutral productivity, A_i. This captures that firms operating in the same industry have similar production technology, for example the production technology of firms producing soap and liquid soap is very similar compared to firms producing bear or spirit drinks. Firm i produces product Y_i according to a Cobb-Douglas production using capital K_i and labor L_i:

$$Y_i = A_i K_i^{\alpha_s^K} L_i^{\alpha_s^L}$$

Demand is modeled as having nested constant elasticities of substitutions (nested CES). The first layer is industries s, and the second layer is the high- and low-demand-elasticity nests g inside industries which captures market fragmentation within industries. The elasticities of substitution across nests are 1 and within nest are $\epsilon_g > 1$. This means the total output in the economy $\mathcal{Y}$ is aggregated as a product of each nest's compound product $\bar{Y}_g$ (Equation (1)), and the compound product of nest g is a CES aggregation across all the firms' products Y_i in the nest g (Equation (2)). β_g is a parameter in the aggregation across nests, and it is nest g's expenditure share in equilibrium. The high-demand-elasticity nest of industry s is called $\bar{g}(s)$ and the low-demand-elasticity nest is called $\underline{g}(s)$. Their demand elasticities are $\epsilon_{\bar{g}(s)}$ and $\epsilon_{\underline{g}(s)}$ respectively, and $\epsilon_{\bar{g}(s)} > \epsilon_{\underline{g}(s)}$. Firms inside a nest g interact in monopolistic competition, and the set these firms form is $\mathcal{G}(g)$. It also possible that an industry has only one nest. In that case $\underline{g}(s) = \bar{g}(s)$ and $\epsilon_{\bar{g}(s)} = \epsilon_{\underline{g}(s)}$. Whether an industry has one nest or two nests will be estimated using data. This aggregation structure is the preference of a hypothetical representative consumer who represents the consumer tastes in this economy.

$$Y = \prod_{s\in\{s_1,\cdots,s_S\}} \prod_{g\in\{\bar{\mathbf{g}}(s),\underline{\mathbf{g}}(s)\}} Y_g^{\beta_g} \tag{1}$$

$$\bar{Y}_g = \left(\sum_{i\in\mathcal{G}(g)} Y_i^{\frac{\epsilon_g - 1}{\epsilon_g}} \right)^{\frac{\epsilon_g}{\epsilon_g - 1}} \tag{2}$$

We deviate from HK to allow each nest to have its own demand elasticities because different industries have different levels of product differentiation. For example, it is more difficult for soap and liquid soap firms to differentiate their products compared to cosmetics. Another modification we make is the additional layer of nests inside industries. This describes the different scopes of product differentiation among firms with similar production technology. For example, a Dior lipstick, a luxurious brand, is 15 times more expensive than a lipstick from Essence, a less-known beauty brand usually found in supermarkets because consumers are more willing to pay higher prices for a Dior lipstick, not because the cost is 15 times higher. Varying levels of product differentiation is taken into account because Autor et al. (2020) shows that heterogeneous demand elasticities have a large impact on the aggregate labor share.

3.1 The firms' problem

Firms maximize their profits Π_i by choosing the optimal prices P_i and their capital and labor usage K_i and L_i subject to the nested CES demand. The production factors capital and labor are hired from perfectly competitive factor markets with the market prices of capital and labor being R and w, but the factor markets contain firm-specific input distortions τ_i^K and τ_i^L that enter the profit functions as wedges multiplied with the market prices of the factors. In addition to capital and labor expenditure, the production cost also contains idiosyncratic cost shocks that are multiplied with the expenditure on capital and labor. These cost shocks are realized after firms set prices, and choose capital and labor, which gives us a nondegenerate distribution of firms' markups. Together with the heterogeneous demand elasticities, these cost shocks rationalize the markups distribution in the data and allow us to infer demand elasticities using data on firm-level markups. The cost shocks follow a nest-specific normal distribution $\mathcal{N}(-\frac{\sigma_g^2}{2}, \sigma_g^2)$. The mean is minus a half of the variance because we normalize $\mathbb{E}[e^{\delta_i}] = 1$. The cost shocks do not enter the observed labor and capital shares.

$$\Pi_i = P_i Y_i - (R(1+\tau_i^K)K_i + w(1+\tau_i^L)L_i)e^{\delta_i}$$

Since the cost shocks are unknown when firms choose K_i, L_i, and P_i, K_i, L_i, and P_i are chosen to maximize their expected profits:

$$\begin{aligned}\max_{K_i, L_i, P_i} \mathbb{E}[\Pi_i] =& P_i Y_i - (R(1+\tau_i^K)K_i + w(1+\tau_i^L)L_i) \\ &\text{s.t. the nested CES demand} \\ & Y_i = A_i K_i^{\alpha_s^K} L_i^{\alpha_s^L}\end{aligned}$$

The first-order conditions of the firms' problem demonstrate that firm-specific input distortions determines the gap between observed factor shares and the predicted factor shares under no input distortions. The unexpected cost shocks do not enter the observed labor and capital shares. Since the factor shares have to be positive, input distortions τ_i^K and τ_i^L have

to be larger than -1 by construction.

$$\underbrace{\log(1+\tau_i^L)}_{\text{input distortions}} = \underbrace{\log\left(\alpha_s^L \frac{\epsilon_g - 1}{\epsilon_g}\right)}_{\substack{\text{predicted shares} \\ \text{w/o input distortions}}} - \underbrace{\log\left(\frac{wL_i}{P_i Y_i}\right)}_{\text{observed shares}}$$

$$\underbrace{\log(1+\tau_i^K)}_{\text{input distortions}} = \underbrace{\log\left(\alpha_s^K \frac{\epsilon_g - 1}{\epsilon_g}\right)}_{\substack{\text{predicted shares} \\ \text{w/o input distortions}}} - \underbrace{\log\left(\frac{RK_i}{P_i Y_i}\right)}_{\text{observed shares}}$$

Firms carry out production at the realized production costs and sell their products at the prices set before cost shocks are realized. Therefore, firms' markups, which are the ratios between marginal prices and marginal costs, are the product of a function of the nest-specific demand elasticities and unexpected cost shocks:

$$\mu_i = \frac{\epsilon_g}{\epsilon_g - 1} e^{-\delta_i}$$

Derivations of the results in this section and other auxiliary variables such as marginal costs and the optimal prices are provided in Appendix B.

3.2 General equilibrium and the decomposition of the total revenue

This is a static model with neither entry nor exit. In general equilibrium, firms choose their optimal prices and factor usage facing the nested CES demand and the input distortions. The market prices of capital and labor clear both factor markets. The goods market is always cleared because firms set prices for their products based on their demand. The total supply of capital and labor, K and L, is assumed to be fixed.

The total revenue in this economy consists of the income earned by capital and labor, profits earned by firm owners, and the part collected through the input distortions:

$$\underbrace{\sum_i P_i Y_i}_{\text{the total revenue}} = \underbrace{wL + RK}_{\text{the total labor and capital income}} + \underbrace{\sum_i \Pi_i}_{\text{total profits}} + \underbrace{\sum_i (wL_i * \tau_i^L + RK_i * \tau_i^K)}_{\text{collected in the input distortions}}$$

The summation signs in this equation are over all the firms in the economy.

We do not specify to whom the part of the revenues collected via the input distortions belong because this is not a trivial question to answer, and the focus of this paper is to

estimate the changes in factor shares. In conclusion, we will discuss possible scenarios about the ownership of this part of the revenues and the implications of these scenarios with regard to whether removing the input distortions is always beneficial for an economy's growth perspective.

3.3 Aggregate factor shares and their predicted changes

Aggregate factor shares, $\frac{wL}{PY}$ and $\frac{RK}{PY}$, are the weighted sum of nest-level factor shares, $\sum_{i\in\mathcal{G}(g)}\frac{wL_i}{\bar{P}_g\bar{Y}_g}$ and $\sum_{i\in\mathcal{G}(g)}\frac{RK_i}{\bar{P}_g\bar{Y}_g}$ with the weights being each nests expenditure share $\frac{\bar{P}_g\bar{Y}_g}{PY}$. The first lines of Equation (3) and Equation (4) express the aggregate factor shares as the weighted sum but write $\sum_{i\in\mathcal{G}(g)}\frac{wL_i}{\bar{P}_g\bar{Y}_g}$ as $\sum_{i\in g}\frac{wL_i}{P_iY_i}\frac{P_iY_i}{\bar{P}_g\bar{Y}_g}$ and $\sum_{i\in\mathcal{G}(g)}\frac{RK_i}{\bar{P}_g\bar{Y}_g}$ as $\sum_{i\in g}\frac{RK_i}{P_iY_i}\frac{P_iY_i}{\bar{P}_g\bar{Y}_g}$. $\bar{P}_g$ is the price index of nest g and P is the price index for the entire economy: $\bar{P}_g\equiv\frac{\sum_{i\in\mathcal{G}(g)}P_iY_i}{\bar{Y}_g}$, and $P\equiv\frac{\sum_i P_iY_i}{Y}$.

$$\begin{aligned}\frac{wL}{PY}&=\sum_g\frac{\bar{P}_g\bar{Y}_g}{PY}\sum_{i\in\mathcal{G}(g)}\frac{wL_i}{P_iY_i}\frac{P_iY_i}{\bar{P}_g\bar{Y}_g}\\&=\sum_g\beta_g\underbrace{\alpha_g^L\frac{\epsilon_g-1}{\epsilon_g}\cdot\frac{1}{1+\bar{\tau}_g^L}}_{\text{nest-level labor shares}}\end{aligned}\tag{3}$$

$$\begin{aligned}\frac{RK}{PY}&=\sum_g\frac{\bar{P}_g\bar{Y}_g}{PY}\sum_{i\in\mathcal{G}(g)}\frac{RK_i}{P_iY_i}\frac{P_iY_i}{\bar{P}_g\bar{Y}_g}\\&=\sum_g\beta_g\underbrace{\alpha_g^K\frac{\epsilon_g-1}{\epsilon_g}\cdot\frac{1}{1+\bar{\tau}_g^K}}_{\text{nest-level capital shares}}\end{aligned}\tag{4}$$

where $\bar{\tau}_g^L$ and $\bar{\tau}_g^K$ are defined as:

$$\frac{1}{1+\bar{\tau}_g^L}\equiv\sum_{i\in\mathcal{G}(g)}\frac{1}{1+\tau_i^L}\frac{\left(\frac{A_i}{(1+\tau_i^L)^{\alpha_g^L}(1+\tau_i^K)^{\alpha_g^K}}\right)^{\frac{\epsilon_g-1}{\epsilon_g-(\epsilon_g-1)(\alpha_g^L+\alpha_g^K)}}}{\sum_{j\in\mathcal{G}(g)}\left(\frac{A_j}{(1+\tau_j^L)^{\alpha_g^L}(1+\tau_j^K)^{\alpha_g^K}}\right)^{\frac{\epsilon_g-1}{\epsilon_g-(\epsilon_g-1)(\alpha_g^L+\alpha_g^K)}}}\tag{5}$$

$$\frac{1}{1+\bar{\tau}_g^K}\equiv\sum_{i\in\mathcal{G}(g)}\frac{1}{1+\tau_i^K}\frac{\left(\frac{A_i}{(1+\tau_i^L)^{\alpha_g^L}(1+\tau_i^K)^{\alpha_g^K}}\right)^{\frac{\epsilon_g-1}{\epsilon_g-(\epsilon_g-1)(\alpha_g^L+\alpha_g^K)}}}{\sum_{j\in\mathcal{G}(g)}\left(\frac{A_j}{(1+\tau_j^L)^{\alpha_g^L}(1+\tau_j^K)^{\alpha_g^K}}\right)^{\frac{\epsilon_g-1}{\epsilon_g-(\epsilon_g-1)(\alpha_g^L+\alpha_g^K)}}}\tag{6}$$

In equilibrium, nests' expenditure shares are β_g. We can express nest-level factor shares

using model primitives and distortions, which are expressed in the second lines of Equation (3) and Equation (4). α_g^L and α_g^K are production elasticities in nest g. They equal the α_s^L and α_s^K of industry s to which nest g belongs. ϵ_g is the demand elasticities of nest g. $\frac{1}{1+\bar{\tau}_g^L}$ is a weighted average of $\frac{1}{1+\tau_i^L}$ where the weights are firms' equilibrium market shares in the nest g, i.e. $\frac{P_i Y_i}{P_g Y_g}$, which are the second factor in Equation (5) and Equation (6). The same applies to $\frac{1}{1+\bar{\tau}_g^K}$. $\bar{\tau}_g^L$ and $\bar{\tau}_g^K$ measure the gap between the observed nest-level factor shares and the predicted nest-level factor shares without the input distortions. In the equations below, $\bar{K}_g$ and $\bar{L}_g$ are the total capital and labor used in nest g.

$$\log\left(\frac{1}{1+\bar{\tau}_g^L}\right) = \underbrace{\log\left(\frac{w\bar{L}_g}{\bar{P}_g\bar{Y}_g}\right)}_{\text{observed shares}} - \underbrace{\log\left(\alpha_g^L\frac{\epsilon_g-1}{\epsilon_g}\right)}_{\substack{\text{predicted shares} \\ \text{w/o input distortions}}}$$

$$\log\left(\frac{1}{1+\bar{\tau}_g^K}\right) = \underbrace{\log\left(\frac{R\bar{K}_g}{\bar{P}_g\bar{Y}_g}\right)}_{\text{observed shares}} - \underbrace{\log\left(\alpha_g^K\frac{\epsilon_g-1}{\epsilon_g}\right)}_{\substack{\text{predicted shares} \\ \text{w/o input distortions}}}$$

$\bar{\tau}_g^L$ and $\bar{\tau}_g^K$ would become zero when we remove all the input distortions, i.e. $\tau_i^K = \tau_i^L = 0$. Therefore, the predicted aggregate factor shares without the input distortions, $\frac{w^*L}{P^*Y^*}$ and $\frac{R^*K}{P^*Y^*}$, are:

$$\frac{w^*L}{P^*Y^*} = \sum_g \beta_g \alpha_g^L \frac{\epsilon_g - 1}{\epsilon_g}$$
$$\frac{R^*K}{P^*Y^*} = \sum_g \beta_g \alpha_g^K \frac{\epsilon_g - 1}{\epsilon_g}$$

There is no $*$ for L and K because the aggregate capital, K, and the aggregate labor, L, are assumed to be constant. The predicted changes in factor shares are:

$$\frac{w^*L}{P^*Y^*} - \frac{wL}{PY} = \sum_g \underbrace{\beta_g \alpha_g^L \frac{\epsilon_g - 1}{\epsilon_g}}_{\text{weights}} \cdot \left[1 - \frac{1}{1+\bar{\tau}_g^L}\right] \tag{7}$$

$$\frac{R^*K}{P^*Y^*} - \frac{RK}{PY} = \sum_g \underbrace{\beta_g \alpha_g^K \frac{\epsilon_g - 1}{\epsilon_g}}_{\text{weights}} \cdot \left[1 - \frac{1}{1+\bar{\tau}_g^K}\right] \tag{8}$$

Derivations of these equations are in the Appendix **??**. Equation (7) and Equation (8)

allow us to interpret the changes in factor shares as the weighted sum of the differences between 1 and $\frac{1}{1+\bar{\tau}_g}$ with weights being a compound of production elasticities α_g, demand elasticities ϵ_g, and expenditure shares β_g. $\bar{\tau}_g$ is $\bar{\tau}_g^L$ in the predicted labor-share change and $\bar{\tau}_g^K$ in the predicted capital-share change. This means the predicted factor-share changes are determined by the value of $\bar{\tau}_g$, the value of the weights, and their covariance. Higher $\bar{\tau}_g$ means larger $1-\frac{1}{1+\bar{\tau}_g}$ and therefore larger increases or smaller decreases. Higher weights also play a role. The increase in the aggregate factor share would be larger if nests with larger τ_g have larger expenditure shares, larger demand elasticities, and larger demand for the factor due to its production technology.

To understand what part of the factor-share changes is caused by the covariance and to compare it to the factor-share changes in an economy with only one representative nest, i.e. all the firms have the same demand elasticities, production elasticities and belong to the same nest, we decompose the factor-share changes into three parts:

$$\begin{aligned}\frac{w^*L}{P^*Y^*}-\frac{wL}{PY} =& N_g\cdot \underbrace{\operatorname{cov}\left(\beta_g\alpha_g^L\frac{\epsilon_g-1}{\epsilon_g}, 1-\frac{1}{1+\bar{\tau}_g^L}\right)}_{\text{covariance between weights and input distortions}} + \\ & N_g\cdot \underbrace{\left(\overline{\beta_g\alpha_g^L\frac{\epsilon_g-1}{\epsilon_g}}-\alpha^L\frac{1}{N_g}\frac{\epsilon-1}{\epsilon}\right)}_{\text{co-movement in weights}}\overline{\left(1-\frac{1}{1+\bar{\tau}_g^L}\right)} + \underbrace{\alpha^L\frac{\epsilon-1}{\epsilon}\cdot\overline{\left(1-\frac{1}{1+\bar{\tau}_g^L}\right)}}_{\text{one representative nest}}\end{aligned} \tag{9}$$

$$\begin{aligned}\frac{R^*K}{P^*Y^*}-\frac{RK}{PY} =& N_g\cdot \underbrace{\operatorname{cov}\left(\beta_g\alpha_g^K\frac{\epsilon_g-1}{\epsilon_g}, 1-\frac{1}{1+\bar{\tau}_g^K}\right)}_{\text{covariance between weights and input distortions}} + \\ & N_g\cdot \underbrace{\left(\overline{\beta_g\alpha_g^K\frac{\epsilon_g-1}{\epsilon_g}}-\alpha^K\frac{1}{N_g}\frac{\epsilon-1}{\epsilon}\right)}_{\text{co-movement in weights}}\overline{\left(1-\frac{1}{1+\bar{\tau}_g^K}\right)} + \underbrace{\alpha^K\frac{\epsilon-1}{\epsilon}\cdot\overline{\left(1-\frac{1}{1+\bar{\tau}_g^K}\right)}}_{\text{one representative nest}}\end{aligned} \tag{10}$$

The first term on the right hand side of Equation (9) (or Equation (10)) measures factor-share changes due to the covariance between the weights and $\bar{\tau}_g^L$ (or $\bar{\tau}_g^K$) across nests. N_g is the total number of nests. A larger covariance means a larger increase in the factor shares. The first part of the second term captures the co-movement among the parameters in weights, i.e. β_g, α_g^L (or α_g^K), and ϵ_g. $\overline{\beta_g\alpha_g^L\frac{\epsilon_g-1}{\epsilon_g}}$ (or $\overline{\beta_g\alpha_g^K\frac{\epsilon_g-1}{\epsilon_g}}$) is the average of $\beta_g\alpha_g^L\frac{\epsilon_g-1}{\epsilon_g}$ (or $\beta_g\alpha_g^K\frac{\epsilon_g-1}{\epsilon_g}$) across nests. α^L (or α^K) and $\frac{\epsilon-1}{\epsilon}$ are the averages of α_g^L (or α_g^K) and $\frac{\epsilon_g-1}{\epsilon_g}$ across nests respectively. The more aligned the parameters are, the larger the changes in factor shares. The second part of the second term, $\overline{1-\frac{1}{1+\bar{\tau}_g^L}}$ (or $\overline{1-\frac{1}{1+\bar{\tau}_g^K}}$) is the average of $1-\frac{1}{1+\bar{\tau}_g^L}$ (or $1-\frac{1}{1+\bar{\tau}_g^K}$) across nests, which measures the average change in labor shares (or capital shares) across nests. Put the two parts together, the second term measures the

impact of co-movement in weights at the average level $\overline{1-\frac{1}{1+\bar{\tau}_g}}$. The last term describes how much the factor shares would change if there is only one nest whose production elasticities are α^L (or α^K) and demand elasticities are ϵ as define above.

The three terms may or may not have the same sign. For example, the covariance term is positive if nests with larger weights have higher $\bar{\tau}_g$, and this can still be the case if $\overline{1-\frac{1}{1+\bar{\tau}_g}}$ is negative, i.e. $\bar{\tau}_g$ is negative. In this case, the third term is negative while the first term is positive. Depending on the co-movement pattern of the parameters in the weights, the second term may be positive or negative. Therefore, depending on the joint distribution of $\bar{\tau}_g^L$, $\bar{\tau}_g^K$, β_g, α_g^L, α_g^K, and ϵ_g, the three terms in Equation (9) and Equation (10) can reinforce or offset each other's impact on the factor-share changes.

We are interested in the scenario of one representative nest because assuming away heterogeneity across industries and modeling an economy with one or only a couple of representative industries are common practice in the growth literature and other macroeconomics studies. Decomposing our empirical results based on Equation (9) and Equation (10) in Section 5 illustrates whether heterogeneity across nests and industries affects the predicted changes of factor shares and whether assuming a representative nest provides misleading results.

4 Identification and estimation

4.1 The latent market structure

We defined each market as a nest of firms that face the same demand elasticities and production elasticities but can differ on productivity and input distortions. We do not directly observe the nests and need to infer the latent nest structure using observed firm characteristics. We assume that firms belonging to the same nest always belong to the same industry. We then use revenue-cost ratios to measure firm-specific markups and then infer the latent nest structure within each industry using the distribution of revenue-cost ratios. According to our model, each industry can have one or two nests, and cost shocks in each nest follow a normal distribution $\mathcal{N}(-\frac{\sigma_g^2}{2}, \sigma_g^2)$, where σ_g is a parameter to be estimated. Therefore, we can use the homogeneity test for mixed normal distribution to infer whether an industry contains one nest or two nests. If an industry contains one nest, the distribution of firm-level markups in the industry is:

$$\log(\mu_i) \sim \mathcal{N}\left(\log\frac{\epsilon_g}{\epsilon_g - 1} - \frac{\sigma_g^2}{2}, \sigma_g^2\right) \text{ for } i \in \mathcal{S}(s)$$

where g indicate the industry s itself since there is only one nest in s, and $\mathcal{S}(s)$ is the set of all the firms in industry s. When there are two nests, the distribution becomes:

$$\log(\mu_i) \sim (1-w_s)\mathcal{N}\left(\log\frac{\epsilon_{\underline{g}(s)}}{\epsilon_{\underline{g}(s)}-1}, \sigma^2_{\underline{g}(s)}\right) + w_s\mathcal{N}\left(\log\frac{\epsilon_{\bar{g}(s)}}{\epsilon_{\bar{g}(s)}-1}, \sigma^2_{\bar{g}(s)}\right) \text{ for } i \in \mathcal{S}(s)$$

Firms' nest identity and each nest's demand elasticities are then estimated by maximizing each industry's likelihood of firm-level markups. Since it is difficult to find the global maximum of the likelihood function of a mixture distribution, we use both the quasi-Newton method and the Expectation-Maximization Algorithm to help find the global maximum. More details are provided in Appendix C.2.

4.2 Production elasticities and input distortions

We use observed firm-level value added, labor expenditure, the depreciated net value of capital, together with the market structure and demand elasticities estimated above to identify production elasticities and input distortions. We do not need to observe wage as we can directly observe wage expenditure but we do need to assume the annual market rental price of capital, R. We follow HK to assume $R = 0.1$.

Profit maximization gives firms' capital and labor expenditures as a function of production elasticities and input distortions:

$$\log\left(\frac{wL_i}{P_iY_i(\epsilon_g-1)/\epsilon_g}\right) = \log(\alpha_s^L) - \log(1+\tau_i^L)$$
$$\log\left(\frac{RK_i}{P_iY_i(\epsilon_g-1)/\epsilon_g}\right) = \log(\alpha_s^K) - \log(1+\tau_i^K)$$

We treat the left-hand side of the equations as known because ϵ_g is estimated in the previous step, R is set to 0.1, and the rest is directly observed. α_s^L and α_s^K can be interpreted as the location of the distribution of $\log\left(\frac{wL_i}{P_iY_i(\epsilon_g-1)/\epsilon_g}\right)$ and $\log\left(\frac{RK_i}{P_iY_i(\epsilon_g-1)/\epsilon_g}\right)$ while $\log(1+\tau_i^L)$ and $\log(1+\tau_i^K)$ determine the deviation from α_s^L and α_s^K. Since the mechanisms behind positive distortions may be very different from those behind negative distortions, we allow the distribution of positive τ_i^K and τ_i^L to differ from the distribution of negative ones for each industry, and we allow the probability of having positive distortions in an industry to be a free parameter. Distortions are independent and identically distributed within an industry and are independent across industries. Distortions on capital are independent from distortions on labor. Firm i in industry s has probability κ_s^K and κ_s^L of having a positive

capital distortion and labor distortion respectively:

$$\begin{cases} \mathbb{P}(\tau_i^K > 0 | i \in \mathcal{S}(s)) = & \kappa_s^K \\ \mathbb{P}(\tau_i^K \leq 0 | i \in \mathcal{S}(s)) = & 1 - \kappa_s^K \\ \mathbb{P}(\tau_i^L > 0 | i \in \mathcal{S}(s)) = & \kappa_s^L \\ \mathbb{P}(\tau_i^L \leq 0 | i \in \mathcal{S}(s)) = & 1 - \kappa_s^L \end{cases}$$

Conditional on $\tau_i^K > 0$, the distribution of τ_i^K is the right half of $2\mathcal{N}(0, \sigma_{s,+}^K)$; conditional on $\tau_i^K < 0$, the distribution is the left half of $2\mathcal{N}(0, \sigma_{s,-}^K)$. Similarly, conditional on $\tau_i^L > 0$, the distribution of τ_i^L is the right half of $2\mathcal{N}(0, \sigma_{s,+}^L)$; conditional on $\tau_i^L < 0$, the distribution is the left half of $2\mathcal{N}(0, \sigma_{s,-}^L)$. Under this setup, the distributions of firm i's input distortions without conditioning on the sign of the distortions are:

$$\begin{aligned} \log(\tau_i^K + 1) \sim & 2\kappa_s^K \mathbb{1}[\tau_i^K > 0] \cdot \mathcal{N}(0, \sigma_{s,+}^K) + (2 - 2\kappa_s^K)\mathbb{1}[\tau_i^K \leq 0] \cdot \mathcal{N}(0, \sigma_{s,-}^K) \\ \log(\tau_i^L + 1) \sim & 2\kappa_s^L \mathbb{1}[\tau_i^L > 0] \cdot \mathcal{N}(0, \sigma_{s,+}^L) + (2 - 2\kappa_s^L)\mathbb{1}[\tau_i^L \leq 0] \cdot \mathcal{N}(0, \sigma_{s,-}^L) \end{aligned}$$

Production elasticities α_s^K and α_s^L are estimated by maximizing the likelihood of the firm-level labor shares and capital shares. Input distortions are then calculated using the estimated production elasticities. The key identification assumption here is the modes of capital distortions and labor distortions of firms inside an industry are both zero. More technical details as well as the likelihood function of labor shares and capital shares are provided in Appendix C.3.

4.2.1 Problems with calibrated production elasticities

In HK, α_g^K and α_g^L are calibrated using American data, assuming that Chinese and American firms share the same technology. This is not a costly assumption for studying the predicted TFP gains because it is the dispersion of distortions within a market not the level of distortions that affect predicted TFP gains. Under the assumption that α_g^K and α_g^L are the same within a market, the dispersion of distortions is determined by the dispersion of labor shares and capital shares and is independent from the values that α_g^K and α_g^L take. The same argument applies for demand elasticities. Whether demand elasticities for each market are the same and what values they have do not affect the dispersion and therefore do not affect the part of TFP gains that directly result from distortions. They only affects the aggregate TFP gains by how the gains are aggregated.

The equation below demonstrates this. Following HK's notation, we define the revenue-

based TFP, i.e. TFPR at the firm level and the nest level as:

$$\begin{aligned}\text{TFPR}_i \equiv & P_i A_i = \frac{P_i Y_i}{K_i^{\alpha_s^K} L_i^{\alpha_s^L}} \\ \overline{\text{TFPR}}_g \equiv & \frac{\sum_{i \in \mathcal{G}(g)} P_i Y_i}{\bar{K}_g^{\alpha_s^K} \bar{L}_g^{\alpha_s^L}}\end{aligned}$$

where $\bar{K}_g$ and $\bar{L}_g$ are the total amount of capital and labor used in nest g. Then the market-level TFP is:

$$\begin{aligned}\overline{\text{TFP}}_g &= \left[\sum_{i \in \mathcal{G}(g)} \left(A_i \frac{\overline{\text{TFPR}}_g}{\text{TFPR}_i}\right)^{\epsilon_g - 1}\right]^{\frac{1}{\epsilon_g - 1}} \\ &= \left[\sum_{i \in \mathcal{G}(g)} \left(A_i \left(\underbrace{\frac{1}{1+\tau_i^L} \frac{1}{\sum_{i \in \mathcal{G}(g)} \frac{P_i Y_i}{\bar{P}_g \bar{Y}_g} \frac{1}{1+\tau_i^L}}}_{\text{labor distortions}}\right)^{\alpha_s^L} \cdot \right.\right. \\ &\quad \left.\left. \left(\underbrace{\frac{1}{1+\tau_i^K} \frac{1}{\sum_{i \in \mathcal{G}(g)} \frac{P_i Y_i}{\bar{P}_g \bar{Y}_g} \frac{1}{1+\tau_i^K}}}_{\text{capital distortions}}\right)^{\alpha_s^K} \left(\frac{P_i Y_i}{\bar{P}_g \bar{Y}_g}\right)^{\alpha_s^L + \alpha_s^K - 1}\right)^{\epsilon_g - 1}\right]^{\frac{1}{\epsilon_g - 1}} \\ &= \left[\sum_{i \in \mathcal{G}(g)} \left(A_i \left(\underbrace{\frac{w L_i / (P_i Y_i)}{w \bar{L}_g / (\bar{P}_g \bar{Y}_g)}}_{\text{labor distortions}}\right)^{\alpha_s^L} \left(\underbrace{\frac{R K_i / (P_i Y_i)}{R \bar{K}_g / (\bar{P}_g \bar{Y}_g)}}_{\text{capital distortions}}\right)^{\alpha_s^K} \left(\frac{P_i Y_i}{\bar{P}_g \bar{Y}_g}\right)^{\alpha_s^L + \alpha_s^K - 1}\right)^{\epsilon_g - 1}\right]^{\frac{1}{\epsilon_g - 1}}\end{aligned}$$

Here, $w\bar{L}_g = \sum_{i \in \mathcal{G}(g)} w L_i$ is the labor expenditure of the nest g. $R\bar{K}_g = \sum_{i \in \mathcal{G}(g)} R K_i$ is this nest's capital expenditure and $\bar{P}_g \bar{Y}_g = \sum_{i \in \mathcal{G}(g)} P_i Y_i$ is its total value added. The distortions enter the nest-level $\overline{\text{TFP}}_g$ as a ratio between input distortions and a weighted average of all the firm's input distortions in nest g, as shown in the second line of the above equation for $\overline{\text{TFP}}_g$. Therefore, the level of each firm's input distortions cancels out and only the dispersion remains. Plugging in the first-order conditions for firms, we can see in the third line that this dispersion in input distortions are, in fact, dispersion in labor shares and capital shares.

However, both the level and the dispersion of distortions matter for labor shares as we can see in Equation (3) and Equation (4). When increase the calibrated α_g^K and α_g^L from a

fairly low level, the sign of labor and capital share changes may flip from negative to positive because estimated distortions will increase from mostly negative to mostly positive. In fact, since American firms are typically understood as more capital-intensive and Chinese firms as more labor-intensive, calibrating α_g^K and α_g^L using American firms will bias the changes in the Chinese labor shares downwards and bias the changes in the Chinese capital shares upwards.

5 Results

In this section, we will first discuss the structural parameters estimated using our firm-level data and then present the predicted changes in factor shares. To explain the different behaviors of labor share and capital and how the nest-level heterogeneity affects the predict changes, we first decompose the changes using Equation (9) and Equation (10) and then use three experiments to investigate whether one of the three dimensions of heterogeneity across nests, i.e. production technology, demand elasticities, and expenditure shares, plays a big role than the others.

5.1 The nest structure and demand elasticities

Table 2 reports the distribution of firm counts at the industry level. The first row is for industries that are estimated as having one nest and the second row is for industries estimated as having two nests. There are in total 523 industries. About 90% of the industries (462 out of 523) are estimated as having two nests, and these industries are generally larger.

Table 2: Distribution of industry-level firm counts

	N	Mean	Min	Pctl(25)	Median	Pctl(75)	Max
One nest	61	23	2	6	15	27	237
Two nests	462	494	12	118	256	545	9,947

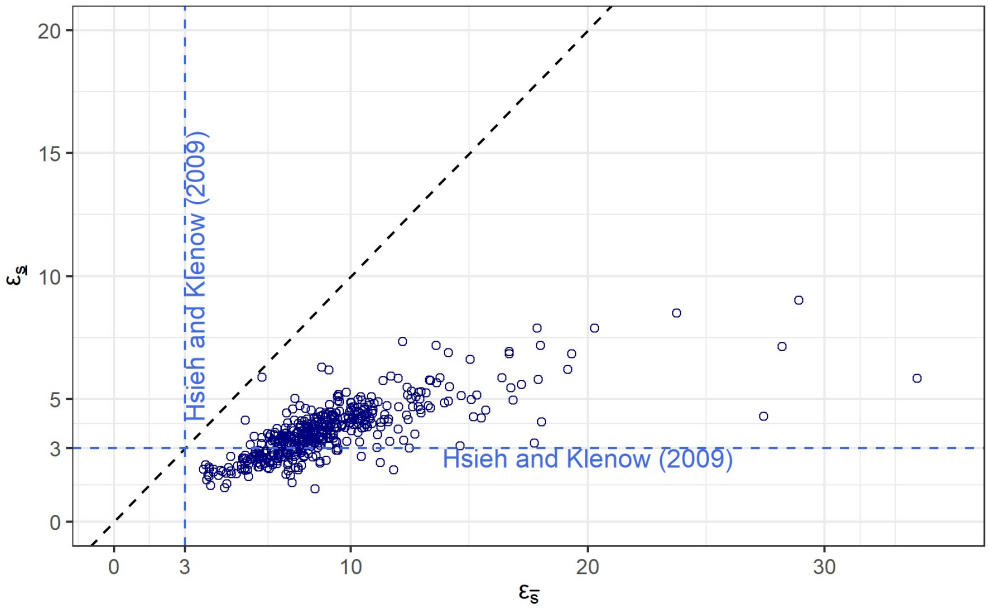
The first row in Table 3 shows the ex-ante probability of belonging to the high demand-elasticity nest. In 90% of the 462 industries with two nests, it is more likely to be in the high demand-elasticity nest. This suggests that achieving a high level of demand elasticities is difficult. The second row in Table 3 is the distribution of demand elasticities across 229,064 firms. The third and forth row respectively weight each firm by their costs and revenues. Depending on whether firms are weighted by their costs or revenues, the average demand elasticity is between 8.5 and 9.4 and the median is between 8.6 and 9.2. There is large

TABLE 3: Summary statistics of selected estimated parameters

	Mean	St. Dev.	Pctl(10)	Pctl(25)	Median	Pctl(75)	Pctl(90)
w_s	0.73	0.16	0.54	0.66	0.75	0.83	0.89
ϵ_g	8.49	3.26	3.99	6.50	8.57	10.27	12.85
ϵ_g (cost)[1]	9.37	3.51	4.50	7.44	9.20	10.91	14.16
ϵ_g (revenue)[2]	9.07	3.58	4.14	7.06	9.14	10.76	14.12

[1] The distribution is weighted by firms' costs.
[2] The distribution is weighted by firms' revenues.

FIGURE 1: Demand elasticities of industries with two nests

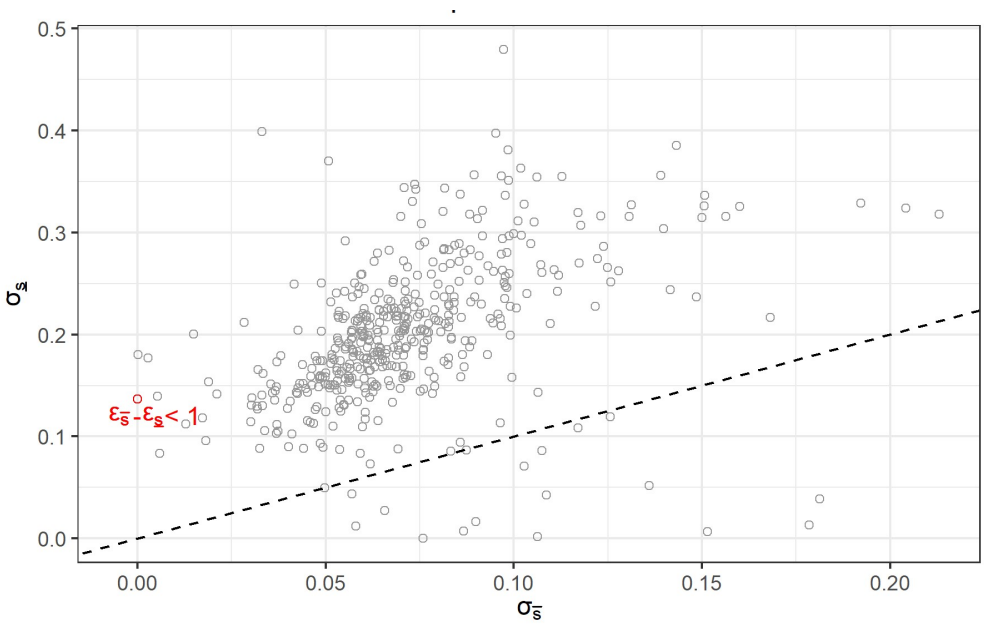


variation in demand elasticities, with the top 10 percentile about three times larger than the bottom 10 percentile.

This Figure 1 shows how demand elasticities vary within industries. Each point here is an industry that contains two nests. The x-axis is demand elasticities of the high-demand-elasticity nest, and the y-axis is that of the low-demand-elasticity nest. In most industries, the demand elasticities of $\bar{g}(s)$ are two times or more larger than those of $\underline{g}(s)$. The point which is very close to the 45 degree line is identified as having two nests because the standard deviations of the two nests' markups are far away enough from the 45 degree line (the red point in Figure 2).

Our estimated demand elasticities are in line with those found in literature. There are few markups estimates for Chinese firms in the literature, so we check our estimates

FIGURE 2: Standard deviations of component normal distributions



by comparing to existing estimates of American firms' markups. The cost-weighted average markups from our estimation are 1.15 which coincides with the 1.15 benchmark cost-weighted average markups in Edmond et al. (2019). It is also consistent with Baqaee and Farhi (2020)'s estimate when using the method developed by De Loecker and Warzynski (2012). De Loecker and Warzynski (2012) itself estimates average markups to be between 1.10 and 1.28, a range contains our estimates. In terms of sales-weighted average markups, ours is 1.17, which is below the estimates from De Loecker et al. (2020) whose sales-weighted average markups are 1.20 in 1980 and 1.60 in 2012. Our median markups are 1.24, a bit lower than the 1.30 median by Feenstra and Weinstein (2017). All these studies mentioned so far using American data. Compared to firms from developing countries, our 1.15 average is higher but not far from the 1.12 average markups found by Peters (2020) using Indonesian data.

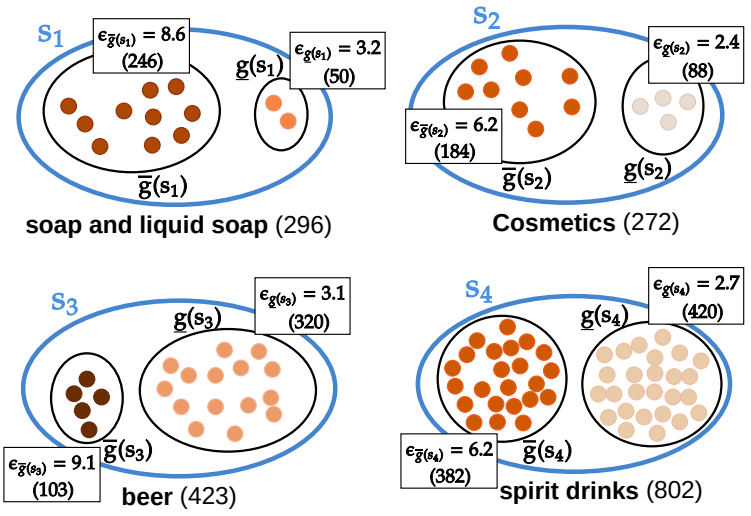
Since there are in total 523 industries and 985 nests, it is difficult to report all the estimated demand elasticities. So we calculate the average demand elasticities at the 2 digit industry level and report here a selection of the 2-digit industries. Oil and gas extraction, tobacco, and pharmaceutical industries have low demand elasticities. This is in line with the intuition that these industries usually have strong market power. The second observation is that although sometimes larger industries have higher demand elasticities, such as the textile industry or when comparing general and special-purpose equipment, it is not always the case. For example, the pharmaceutical industry contains 10 times more firms than the recycling industry but its demand elasticity is about 1/3 of that of the recycling industry.

TABLE 4: Average demand elasticities at the 2-digit industry level

Industry	ϵ	**Firm counts**
Oil and gas extraction	3.22	33
Agricultural and Sideline Food Processing	9.59	12275
Tobacco	3.97	102
Textile	11.31	20197
Pharmaceutical	4.82	4233
Rubber products	7.69	2693
General Equipment Manufacturing	7.51	18088
Special-Purpose Equipment Manufacturing	6.48	8923
Recycling and processing of waste resources and materials	14.36	347

Figure 3 displays the estimated latent nest structure inside an industry. The high-demand-elasticity nests contain more firms in about 95% of the industries that are estimated as having two nests, similar to the industry of soap and liquid soap and the industry of cosmetics illustrated in Figure 3. This indicates that a high level of product differentiation is difficult to achieve and that only few firms succeed. However, there are 5% of the industries where this is not the case. For example, in the beer industry, the low margin market segments are dominated by a small number of firms and the high margin market segment is full of specialized craft beer. Or in the spirit drinks industry where both segments contain a large amount of firms.

FIGURE 3: Four industries as examples



Notes: numbers in parentheses are firm counts in nests.

5.2 Production parameters and input distortions

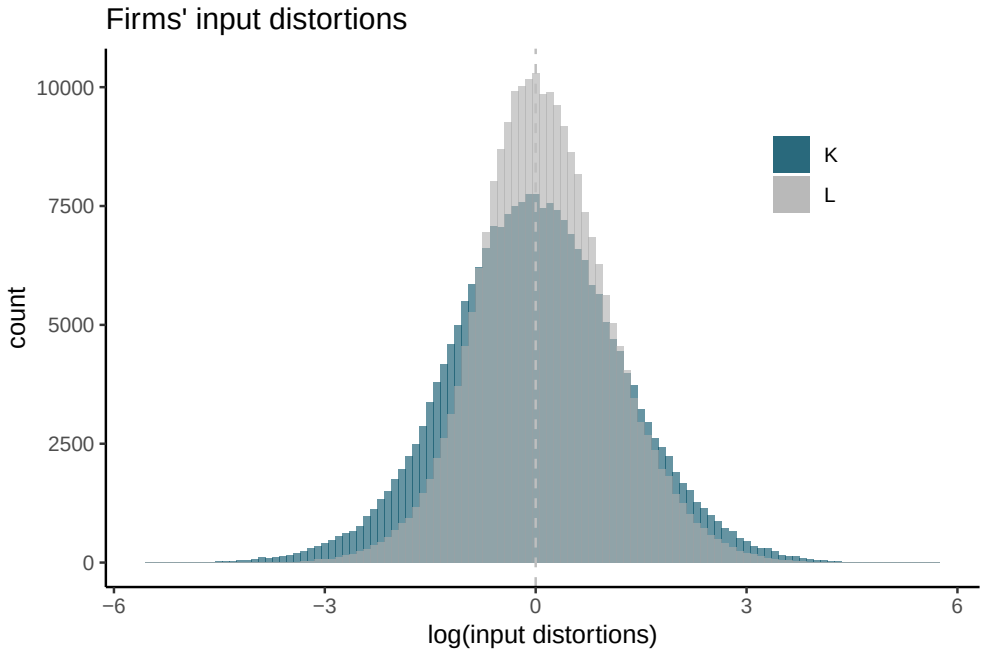
Table 5 reports the distribution of estimated returns to scale across firms. On average, the industrial firms have constant returns to scale but there is also larger variation across industries. Figure 4 displays the estimated input distortions. The modes of both distributions are 0. The capital distortions are more dispersed and more skewed to the left.

TABLE 5: Summary statistics of selected estimated parameters

	Mean	St. Dev.	Pctl(10)	Pctl(25)	Median	Pctl(75)	Pctl(90)
$\alpha_s^K + \alpha_s^L$	1.04	0.58	0.44	0.64	0.88	1.22	1.88
$\alpha_s^K + \alpha_s^L$(cost)[1]	0.96	0.55	0.36	0.53	0.82	1.16	1.83
$\alpha_s^K + \alpha_s^L$ (revenue)[2]	0.97	0.55	0.36	0.54	0.84	1.16	1.83

[1] The distribution is weighted by firms' costs.
[2] The distribution is weighted by firms' revenues.

FIGURE 4: Estimated input distortions $\log(\tau_i^L + 1)$ and $\log(\tau_i^K + 1)$



5.3 Predicted changes in factor shares

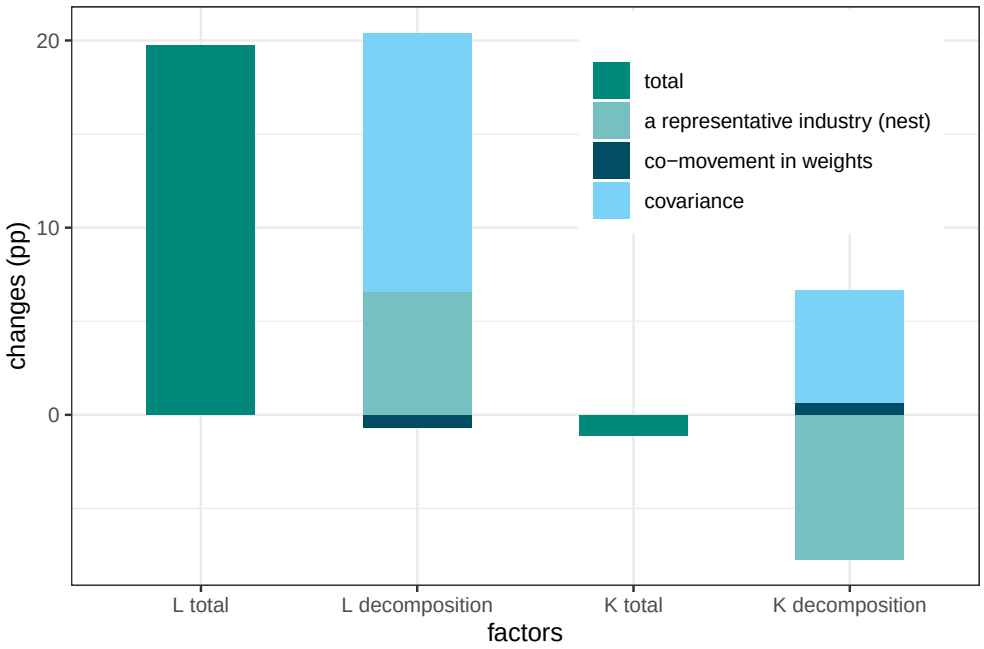
Removing input distortions increases the aggregate labor share by 20 percentage points and decreases the aggregate capital share 1 percentage point (Table 6). In Figure 5, we decompose the predicted factor-share changes according to Equation (9) and Equation (10).

TABLE 6: Labor and capital income shares and their predicted changes

	with input distortions (%)	w/o input distortions (%)	changes (pp)
wL/PY	52.37	72.10	19.73
RK /PY	11.86	10.77	-1.09

It shows that if we are in an economy with one representative nest, labor share increases by 7 percentage points and capital share decreases by 7 percentage points. The covariance between input distortions and the weights, i.e. heterogeneous production elasticities, demand elasticities, and expenditure shares explains 2/3 of the 20 percentage points total increase for labor but offsets most of the decrease in capital share in the one-representative-nest scenario. Comovements among heterogeneous production elasticities, demand elasticities, and expenditure shares have a very minor impact. This implies that going from an one-representative-nest economy to an economy with heterogeneous industries, the covariance between the weights and the input distortions causes the different magnitude of the predicted changes in the labor and capital shares. The capital share decreases while the labor share increases mainly because the capital distortions is more skewed to the left (Figure 4).

FIGURE 5: Decomposition of the predicted factor-share changes



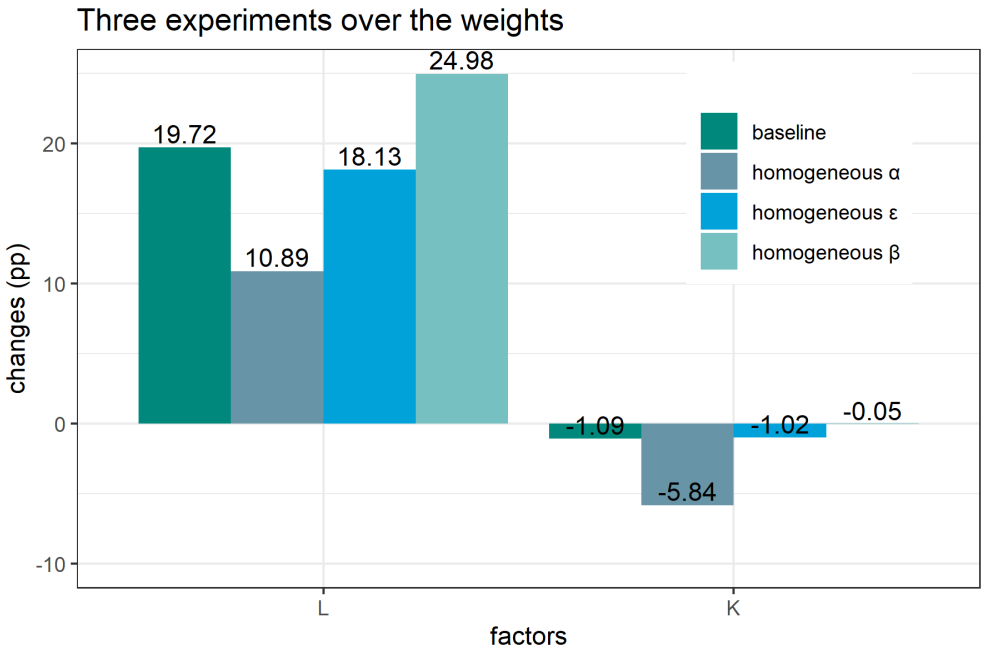
In addition to knowing that the heterogeneity across nests explains a large part of the predicted changes in factor shares, we also want to investigate whether some of the nest het-

erogeneity, i.e. production elasticities, demand elasticities, or expenditure shares, are more important in the affecting the magnitude of factor-share changes. According to Equation (7) and Equation (8), if a parameter has a strong correlation with $\bar{\tau}_g^L$ or $\bar{\tau}_g^K$, the heterogeneity which this parameter represent should have stronger effects on the predicted changes in factor shares. For example, if α_g^K has the strongest correlation with $\bar{\tau}_g^K$, production heterogeneity across nests has the strongest impact on the magnitude of the predicted changes in the capital share. Removing this heterogeneity will cause a big difference in the predicted changes of the capital share. To investigate which type of the heterogeneity is more important, we first report the correlation between the parameters and $1 - \frac{1}{1+\bar{\tau}_g^L}$ and between the parameters and $1 - \frac{1}{1+\bar{\tau}_g^K}$ (Table 7) and then run three experiments where we set each of these three parameters to their averages across firms one by one (Figure 6).

TABLE 7: Correlations between $\bar{\tau}_g$ and weights

	$\bar{\tau}_g^L$	$\bar{\tau}_g^K$
$(\epsilon_g - 1)/\epsilon_g$	0.14	0.02
α_s	0.68	0.62
β_g	-0.10	-0.09

FIGURE 6: Decomposition of factor share changes



Both Table 7 and Figure 6 show that heterogeneous technology is the key heterogeneity for both factors. They have the strongest correlation with $1 - \frac{1}{1+\bar{\tau}_g^L}$ and with $1 - \frac{1}{1+\bar{\tau}_g^K}$.

Removing technology heterogeneity brings the predicted factor-share changes closest to the scenario of one representative nest.

5.3.1 Biases when using calibrated production parameters

Using American firms' production elasticities changes both the sign and the magnitude of predicted factor-share changes, and implies that the labor share would decrease by 3 percentage points while the capital share would increase by about 31 percentage points (Table 8).[3] This is because American firms' are more capital intensive and less labor intensive. These technology differences bias the labor distortions downwards and the capital distortions upwards (the bottom figure in Figure 7) so that labor appears to be overused while capital appears to be heavily underused.

TABLE 8: Changes in labor and capital income shares using different α (pp)

	Benchmark	American α (HK)
wL/PY	19.73	-2.89
RK/PY	-1.09	31.38

5.4 Distributional impact when aggregate TFP gains are considered

When removing the input distortions, the aggregate TFP would also change. Using the predicted TFP gains estimated in Zhang and Xia (2022), Figure 8 reports the changes in the decomposition of the aggregate industrial revenue.

5.5 Input distortions and ownership types

Table 9 demonstrate the distribution of input distortions for domestic private firms and state-owned enterprises (SOEs). Although some domestic private firms face lower distortions, i.e. using too much capital and labor, and some SOEs face higher distortions, i.e. using too little capital and labor, the distortion distributions of domestic private firms first-order stochastically dominates that of SOEs', suggesting that SOEs are more likely to use too much capital and labor compared to domestic private firms. The large variation within both ownership types may result from a fuzzy connection between the ownership labels and their

[3]When using American firms' α, we follow HK to scale up Chinese labor shares so that the aggregate Chinese labor share is 50%, the same as in HK.

FIGURE 7: The distribution of input distortions τ_i compared to using calibrated production parameters

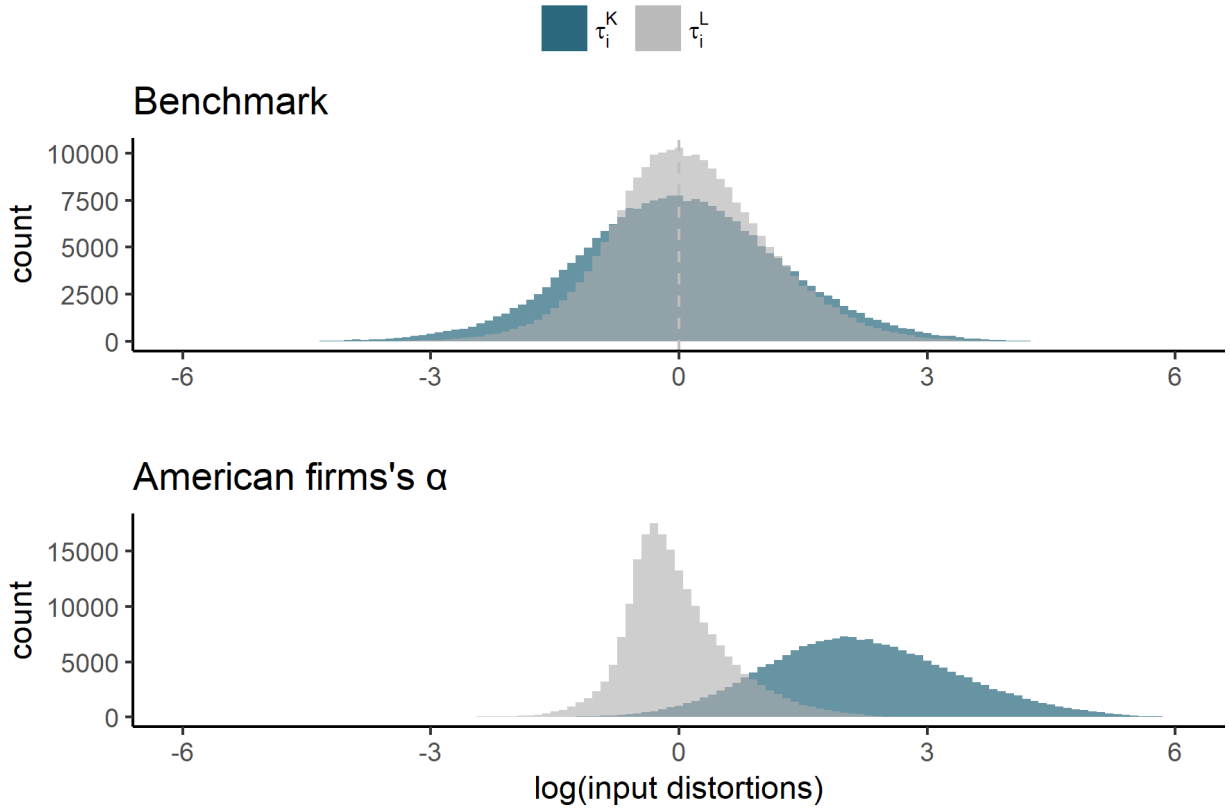
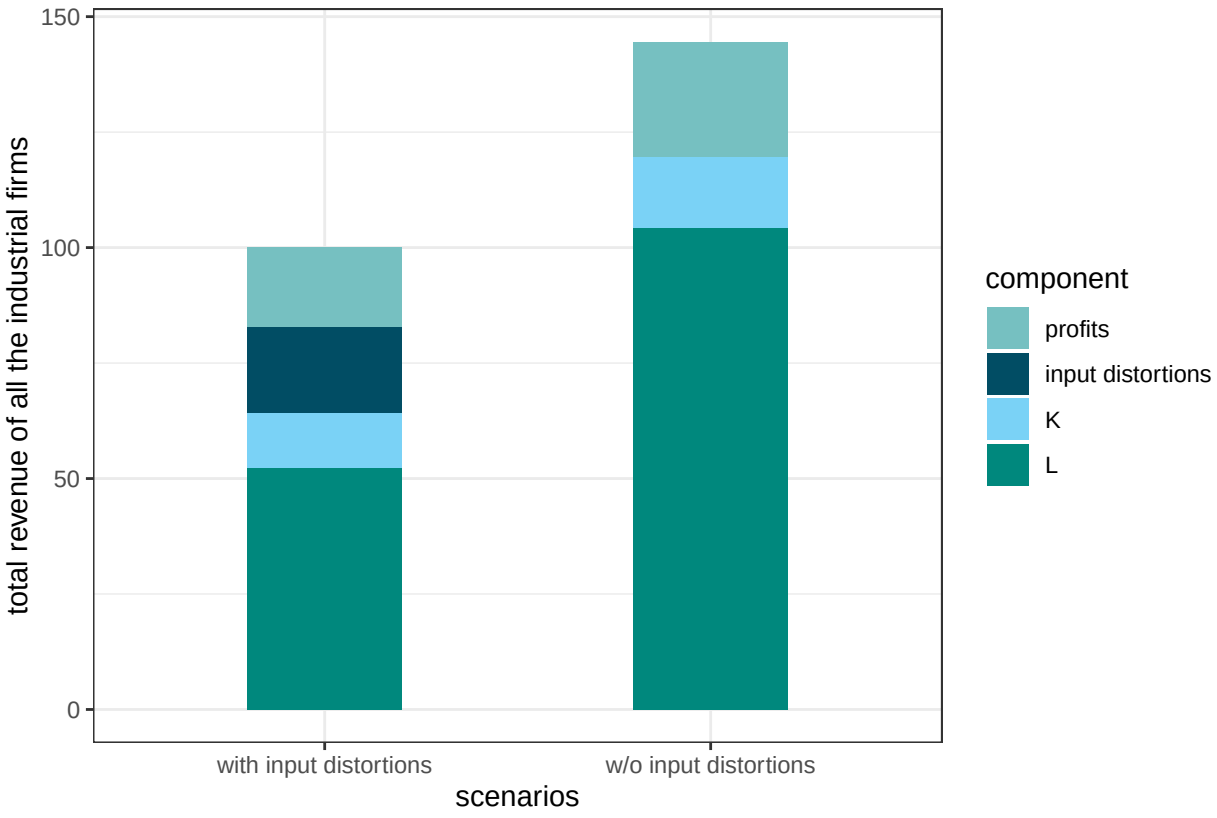


FIGURE 8: The distributional impact of removing the input distortions



business environment. Some domestic private firms may still enjoy favorable financial access because they used to be an SOE or some SOEs hold shares in them. Domestic private firms may receive financial support from central or local governments if they are deemed strategically important. Sometimes, the distinction between an SOE and a domestic private firm is not clear. Normally, there are two criteria of determining whether a firm is an SOE: its registration type and its major share holders. A firm can be labeled as an SOE, according to the first criterion, if it is registered as an SOE; it can also be called an SOE, based on the latter criterion, if its major shareholders are SOEs or some public agents. The same applies to domestic private firms. The two criteria generally agree except for some special cases where, for example, firms are labeled as SOEs under one criterion but not under the other. To remove this ambiguity, Table 9 keeps only those observations where the two criteria agree.

TABLE 9: Estimated distortions for different firm types

	firm type	N	Mean	Min	Pctl(25)	Median	Pctl(75)	Max
τ_i^K	domestic private	164396	1.36	-0.99	-0.50	0.08	1.41	305.22
	SOE	10600	0.41	-1.00	-0.73	-0.38	0.33	147.12
	all	174996	1.31	-1.00	-0.52	0.05	1.34	305.22
τ_i^L	domestic private	164396	0.94	-0.98	-0.35	0.16	1.18	54.49
	SOE	10600	0.33	-0.99	-0.53	-0.13	0.54	26.08
	all	174996	0.91	-0.99	-0.36	0.13	1.13	54.49

6 Conclusion

This paper analyzes the impact of removing input distortions on Chinese industrial factor shares with heterogeneous technology and demand elasticities and provides a novel way of estimating demand and production parameters using firm-level data instead of calibrating using a benchmark economy. In this section, we will discuss the socioeconomic implications of our results and how they connect to other patterns in the Chinese economy.

The predicted changes of factor shares estimated in this paper can bring about the discussion whether removing the input distortions is always beneficial to China's growth perspective. In this paper, we do not specify the exact source of the input distortions. However, adjustment frictions, limited or no access to factors, additional taxes or subsidies, or complicated expenditures related to bureaucracy are the common sources of input distortions. Different types of input distortions means the revenues collected through the input distortions belong to different groups in the economy. If they are mostly adjustment frictions or limited access to factors, the revenues collected via input distortions are retained in firms and are

part of the returns to entrepreneurs. In this cause, entrepreneurs' income would be 15% lower after removing the input distortions. If these distortions are additional fees charged by capital owners or intermediaries (like the input distortions in Liu (2019)), then the returns to capital owners would be 12% lower. If removing the input distortions reduces the returns to entrepreneurs or the owners of capital, there would be fewer new firms, fewer entrepreneurs, and less capital invested in industry.

The input distortions in China may have helped push down the labor price and possibly contributed to the decade-long 2-digit growth. In other words, the input distortions induce industrial firms to produce less and hire less labor, which keeps the industrial demand of labor lower than it otherwise would be. If China had managed to remove its input distortions in the late twentieth century so that its input allocation had been as efficient as the one in America, Chinese firms would have had to face a higher labor cost and starting a new business would have been less attractive. Since cheap labor has been crucial to China's rapid economic growth, removing the input distortions may not have been the solution to close the gap between China and the US, at least in the late twentieth century when China was still poor.

References

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

Atkeson, A. and A. Burstein (2008): "Pricing-to-Market, Trade Costs, and International Relative Prices," *The American Economic Review*, 98, 1998–2031, publisher: American Economic Association.

Autor, D., D. Dorn, L. F. Katz, C. Patterson, and J. Van Reenen (2020): "The Fall of the Labor Share and the Rise of Superstar Firms," *The Quarterly Journal of Economics*, 135, 645–709.

Autor, D. and A. Salomons (2018): "Is Automation Labor ShareDisplacing? Productivity Growth, Employment, and the Labor Share," *Brookings Papers on Economic Activity*, 1–63, publisher: Brookings Institution Press.

Baqaee, D. R. and E. Farhi (2020): "Productivity and Misallocation in General Equilibrium," *The Quarterly Journal of Economics*, 135, 105–163.

Basu, S. (2019): "Are Price-Cost Markups Rising in the United States? A Discussion of the Evidence," *Journal of Economic Perspectives*, 33, 3–22.

Benaglia, T., D. Chauveau, D. R. Hunter, and D. Young (2009): "mixtools: An R Package for Analyzing Finite Mixture Models," *Journal of Statistical Software*, 32, 1–29.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841.

Blanchard, O. J., W. D. Nordhaus, and E. S. Phelps (1997): "The Medium Run," *Brookings Papers on Economic Activity*, 1997, 89–158, publisher: Brookings Institution Press.

Bond, S., A. Hashemi, G. Kaplan, and P. Zoch (2021): "Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data," *Journal of Monetary Economics*, 121, 1–14.

Brandt, L., J. Van Biesebroeck, and Y. Zhang (2012): "Creative Accounting or Creative Destruction? Firm-Level Productivity Growth in Chinese Manufacturing," *Journal of Development Economics*, 97, 339–351.

Burstein, A., V. M. Carvalho, and B. Grassi (2020): "Bottom-up Markup Fluctuations," Working Paper 27958, National Bureau of Economic Research.

Chen, J. and P. Li (2009): "Hypothesis test for normal mixture models: The EM approach," *Ann. Statist.*, 37, arXiv: 0908.3428.

Claude J. P. Bélisle (1992): "Convergence Theorems for a Class of Simulated Annealing Algorithms on Rd," *Journal of Applied Probability*, 29, 885–895.

David, J. M. and V. Venkateswaran (2019): "The Sources of Capital Misallocation," *American Economic Review*, 109, 2531–2567.

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

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

Edmond, C., V. Midrigan, and D. Y. Xu (2015): "Competition, Markups, and the Gains from International Trade," *American Economic Review*, 105, 3183–3221.

——— (2019): "How Costly Are Markups," *working paper*.

Elsby, M. W. L., B. Hobijn, and A. ahin (2013): "The Decline of the U.S. Labor Share," *Brookings Papers on Economic Activity*, 1–52, publisher: Brookings Institution Press.

Feenstra, R. C. and D. E. Weinstein (2017): "Globalization, Markups, and US Welfare," *Journal of Political Economy*, 125, 1040–1074, publisher: The University of Chicago Press.

Gutirrez, G. and S. Piton (2020): "Revisiting the Global Decline of the (Non-housing) Labor Share," *American Economic Review: Insights*, 2, 321–38.

Haltiwanger, J., R. Kulick, and C. Syverson (2018): "Misallocation Measures: The Distortion That Ate the Residual," Working Paper 24199, National Bureau of Economic Research.

Henningsen, A. and O. Toomet (2011): "maxLik: A package for maximum likelihood estimation in R," *Computational Statistics*, 26, 443–458.

Hopenhayn, H., J. Neira, and R. Singhania (2022): "From Population Growth to Firm Demographics: Implications for Concentration, Entrepreneurship and the Labor Share," *Econometrica*, 90, 1879–1914, publisher: John Wiley & Sons, Ltd.

Hsieh, C.-T. and P. J. Klenow (2009): "Misallocation and Manufacturing TFP in China and India," *Quarterly Journal of Economics*, 46.

Karabarbounis, L. and B. Neiman (2014): "The Global Decline of the Labor Share*," *The Quarterly Journal of Economics*, 129, 61–103.

Klenow, P. J. and J. L. Willis (2016): "Real Rigidities and Nominal Price Changes," *Economica*, 83, 443–472.

Klette, T. J. and Z. Griliches (1996): "The Inconsistency of Common Scale Estimators When Output Prices Are Unobserved and Endogenous," *Journal of Applied Econometrics*, 11, 343–361.

Koh, D., R. Santaeullia-Llopis, and Y. Zheng (2020): "Labor Share Decline and Intellectual Property Products Capital," *Econometrica*, 88, 2609–2628, publisher: John

Wiley & Sons, Ltd.

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

Liu, E. (2019): "Industrial Policies in Production Networks," *The Quarterly Journal of Economics*, 134, 1883–1948.

McLachlan, G. and D. Peel (2004): *Finite Mixture Models*, Wiley Series in Probability and Statistics, Wiley.

Mrázová, M., J. P. Neary, and M. Parenti (2021): "Sales and Markup Dispersion: Theory and Empirics," *Econometrica*, 89, 1753–1788.

Nelder, J. A. and R. Mead (1965): "A Simplex Method for Function Minimization," *The Computer Journal*, 7, 308–313.

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

Peters, M. (2020): "Heterogeneous Markups, Growth, and Endogenous Misallocation," *Econometrica*, 88, 2037–2073.

Piketty, T. (2014): *Capital in the Twenty-First Century*, Harvard University Press.

Restuccia, D. and R. Rogerson (2008): "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments," *Review of Economic Dynamics*, 11, 707–720.

Ridder, M. D., B. Grassi, and G. Morzenti (2021): "The Hitchhikers Guide to Markup Estimation," Working Papers 677, IGIER , Bocconi University.

Song, Z., K. Storesletten, and F. Zilibotti (2011): "Growing Like China," *American Economic Review*, 101, 196–233.

Zhang, X. and J. Xia (2022): "Misallocation under Heterogeneous Markups," *working paper*.

Appendix

A Data

We drop observations with negative value added, negative wage expenditures, negative capital, negative total assets, negative account receivable, negative total debts, negative long-term debts, negative account payable, negative exports, negative sales, and negative costs. We also drop observations whose account receivable is larger than total assets, total debts larger than total assets, account payable larger than liquid debts, and profits larger than sales. If a firm's costs are missing but its sales and profits are observed, then its costs are sales minus profits. The survey reports firms' net value of capital and investment. To calculate depreciated net value of capital, we use perpetual annuity method following Brandt et al. (2012).

TABLE 10: Summary Statistics of Cleaned Data (1998-2009)

Statistic	N	Mean	Min	Pctl(25)	Median	Pctl(75)	Max
value added	1,767,623	12,891.55	122	2,328	4,952	12,210	277,956
K	1,767,623	18,502.50	83.61	1,745.29	4,644.34	13,889.94	515,969.30
wL	1,767,623	2,650.66	80	537	1,120	2,570	79,200
revenue	1,767,623	46,385.36	2	8,544	17,564	42,409	58,906,099
cost	1,767,623	41,004.20	1	7,540.5	15,546	37,503	57,460,589
profits	1,767,623	2,242.86	−531,161	49	404	1,583	546,835

B Derivation

For some given Y_i, firms' profits maximization problem can be formulated as, :

$$\begin{aligned} &\min_{K_i,L_i}(R(1+\tau_i^K)K_i + w(1+\tau_i^L)) \\ &\text{s.t. } A_i K_i^{\alpha_s^K} L_i^{\alpha_s^L} \geq Y_i \end{aligned}$$

Expected marginal cost is the Lagrange multiplier of its Lagrange function

$$\min_{K_i,L_i}(R(1+\tau_i^K)K_i + w(1+\tau_i^L)) - \lambda(A_i K_i^{\alpha_s^K} L_i^{\alpha_s^L} - Y_i)$$

Solving it gives expected marginal cost:

$$\mathbb{E}[MC(Y_i)] = \left(\frac{1}{A_i}\right)^{\frac{1}{\alpha_s^L+\alpha_s^K}} Y_i^{\frac{1-\alpha_s^L-\alpha_s^K}{\alpha_s^L+\alpha_s^K}} \left(\frac{R(1+\tau_i^K)}{\alpha_s^K}\right)^{\frac{\alpha_s^K}{\alpha_s^L+\alpha_s^K}} \left(\frac{w(1+\tau_i^L)}{\alpha_s^L}\right)^{\frac{\alpha_s^L}{\alpha_s^L+\alpha_s^K}}$$

The nested-CES demand implies that the optimal prices are always the expected marginal cost times $\epsilon_g/(\epsilon_g-1)$. Therefore, the optimal prices are:

$$P_i = \frac{\epsilon_g}{\epsilon_g-1} \cdot \underbrace{\left(\frac{1}{A_i}\right)^{\frac{1}{\alpha_s^L+\alpha_s^K}} Y_i^{\frac{1-\alpha_s^L-\alpha_s^K}{\alpha_s^L+\alpha_s^K}} \left(\frac{R(1+\tau_i^K)}{\alpha_s^K}\right)^{\frac{\alpha_s^K}{\alpha_s^L+\alpha_s^K}} \left(\frac{w(1+\tau_i^L)}{\alpha_s^L}\right)^{\frac{\alpha_s^L}{\alpha_s^L+\alpha_s^K}}}_{\text{expected marginal cost}}$$

Firms' profit maximization also gives:

$$\begin{aligned}
\frac{K_i}{P_gY_g} &= \frac{\epsilon_g-1}{\epsilon_g} \cdot \frac{\alpha_K^g}{(1+\tau_i^K)R} \cdot \frac{P_iY_i}{P_gY_g} \\
\frac{L_i}{P_gY_g} &= \frac{\epsilon_g-1}{\epsilon_g} \cdot \frac{\alpha_s^L}{(1+\tau_i^L)w} \cdot \frac{P_iY_i}{P_gY_g} \\
\frac{K_i}{P_iY_i} &= \frac{\epsilon_g-1}{\epsilon_g} \cdot \frac{\alpha_s^K}{(1+\tau_i^K)R} \\
\frac{L_i}{P_iY_i} &= \frac{\epsilon_g-1}{\epsilon_g} \cdot \frac{\alpha_s^L}{(1+\tau_i^L)w}
\end{aligned}$$

Nest, we derive the equilibrium market shares $\frac{P_iY_i}{P_gY_g}$. Using the optimal pricing rule, we can write the price ratio of two firms from the same nest as:

$$\frac{P_i}{P_j} = \left(\frac{A_j}{A_i}\right)^{\frac{1}{\alpha_s^L+\alpha_s^K}} \left(\frac{Y_i}{Y_j}\right)^{\frac{1}{\alpha_s^L+\alpha_s^K}-1} \left(\frac{1+\tau_i^K}{1+\tau_j^K}\right)^{\frac{\alpha^K}{\alpha_s^L+\alpha_s^K}} \left(\frac{1+\tau_i^L}{1+\tau_j^L}\right)^{\frac{\alpha^L}{\alpha_s^L+\alpha_s^K}}$$

Using demand side equation, $\frac{Y_i}{Y_j} = \left(\frac{P_i}{P_j}\right)^{-\epsilon_g}$, this can be rewritten as

$$\left(\frac{P_i}{P_j}\right)^{1+\epsilon_g\left(\frac{1}{\alpha_s^L+\alpha_s^K}-1\right)} = \left(\frac{A_j}{A_i}\right)^{\frac{1}{\alpha_s^L+\alpha_s^K}} \left(\frac{1+\tau_i^K}{1+\tau_j^K}\right)^{\frac{\alpha^K}{\alpha_s^L+\alpha_s^K}} \left(\frac{1+\tau_i^L}{1+\tau_j^L}\right)^{\frac{\alpha^L}{\alpha_s^L+\alpha_s^K}}$$

Demand side tells us, $\frac{P_iY_i}{P_jY_j} = \left(\frac{P_i}{P_j}\right)^{1-\epsilon}$, therefore

$$\frac{P_iY_i}{P_jY_j} = \left(\frac{A_j}{A_i}\right)^{\frac{1-\epsilon_g}{(1-\epsilon_g)(\alpha_s^L+\alpha_s^K)+\epsilon_g}} \left(\frac{1+\tau_i^K}{1+\tau_j^K}\right)^{\frac{\alpha^K(1-\epsilon_g)}{(1-\epsilon_g)(\alpha_s^L+\alpha_s^K)+\epsilon_g}} \left(\frac{1+\tau_i^L}{1+\tau_j^L}\right)^{\frac{\alpha^L(1-\epsilon_g)}{(1-\epsilon_g)(\alpha_s^L+\alpha_s^K)+\epsilon_g}}$$

Thus,

$$P_i Y_i \propto \left(\frac{1}{A_i}\right)^{\frac{1-\epsilon_g}{(1-\epsilon_g)(\alpha_s^L+\alpha_s^K)+\epsilon_g}} (1+\tau_i^K)^{\frac{\alpha^K(1-\epsilon_g)}{(1-\epsilon_g)(\alpha_s^L+\alpha_s^K)+\epsilon_g}} (1+\tau_i^L)^{\frac{\alpha^L(1-\epsilon_g)}{(1-\epsilon_g)(\alpha_s^L+\alpha_s^K)+\epsilon_g}} \equiv W_i$$

Hence,

$$\frac{P_i Y_i}{P_g Y_g} = \frac{W_i}{\sum_{j\in g} W_j}$$

C Identification and estimation

C.1 Calculating firm-level markups——the limitations and its remedies

Inferring markups without observed prices, physical production, and physical inputs is difficult. Generally, there are three methods for estimating markups: the demand approach, the production approach, and the accounting approach. Developed by Berry et al. (1995), the demand approach models consumers' choices among products and infers markups from parameters in consumers' utility functions. This method requires product prices, sales in units of products, and some observed characteristics of the products. The production approach measures markups as the ratio of production elasticities to cost share of a variable input (De Loecker and Warzynski (2012)). Although it does not require prices, applying it to markets with heterogeneous markups and heterogeneous production functions creates various problems when physical production and physical inputs are replaced by revenue production and input expenditure (See Bond et al. (2021) for detailed explanations. A brief discussion on this is offered below). The accounting approach does not require any econometric assumption apart from that the marginal cost equals the average cost. This approach only needs cost and revenue data.

We do not observe prices and units of products sold, so only the production approach and the accounting approach are feasible. In fact, these are the methods used by many papers that infer firm-level markups using similar data as ours, such as De Loecker and Warzynski (2012), Liu (2019), Autor et al. (2020), De Loecker et al. (2020) and Baqaee and Farhi (2020). While both approaches create bias in our model setup, we prefer using the accounting approach and then carefully check whether possible bias affects our results because apart from the measurement errors in the observed cost and revenue, there is only one source of bias, i.e. non-constant returns to scale.

Dealing with the bias in the production approach is a lot of more difficult if not completely unfeasible. There are four sources of bias in the production approach under our setup when

physical production and physical inputs are not observed and when firms have heterogeneous markups. The first one results from replacing production elasticities by revenue elasticities. If the revenue elasticities are consistently estimated, the estimated markups by the production approach should always be 1 (Bond et al. (2021)). Secondly, the assumption of variable input is very restrictive and it is almost impossible to find a truly variable input in data. Besides, the production approach also requires that the variable input do not affect demand and it can be common for inputs, such as labor inputs for marketing, to affect demand (Bond et al. (2021)). Most commonly used variable inputs are material and energy but we observe neither in our data. The last two sources are related to the consistency of estimated production elasticities using revenue data. In order to estimate production elasticities, the production approach needs to estimate production functions using Olley and Pakes (1996), Levinsohn and Petrin (2003), or Ackerberg et al. (2015). However, when revenue production is used in the place of physical production, Klette and Griliches (1996) demonstrates that heterogeneous markups can bias the estimated production elasticities downward. Last but no the least, even if one successfully corrects this bias by controlling for industry-level sales and prices, weak instruments can still plague the estimators (Bond et al. (2021)). Although Ridder et al. (2021) shows that estimated markups using revenue gives the correct dispersion but this requires using material as variable input. We only observe labor and capital. Since labor and capital are far from being variable, applying the production approach in our case is problematic.

C.2 Step 1: Identifying nest-related parameters and cost shocks' distribution parameters

We allow demand elasticities to differ within the finest industry category observed in our data. The hard part is we do not see which industry s contains two nests $\{\bar{g}(s), \underline{g}(s)\}$ and which contains only one nest, nor do we observe a firm's nest identify when an industry contains two nests.

The observed markup, i.e. a firm-level revenue-cost ratio, is a noisy indicator of a firm's nest identity:

$$\log(\mu_i + 1) = \underbrace{\log\left(\frac{\epsilon_g}{\epsilon_g - 1}\right)}_{\text{nest specific}} - \underbrace{\delta_i}_{\text{noises}}$$

To identify the existence of two nests, firms' nest identity, and demand elasticities, we assume δ_i follows a normal distribution with mean 0 within each nest:

$$\delta_{ig} \sim \mathcal{N}(-\sigma_g^2/2, \sigma_g)$$

The distribution variances σ_g differ across nests. When there is one nest inside an industry, the distribution of the logarithmic markups is:

$$\log(\mu_i + 1) \sim \mathcal{N}\left(\log \frac{\epsilon_s}{\epsilon_s - 1}, \sigma_{\epsilon_s}\right) \text{ for } i \in s$$

Since we only observe the pooled distribution of $\underline{g}(s)$ and $\bar{g}(s)$ when there are two nests, the distributions of $\log(\mu_i + 1)$ for all those industries follow a mixed normal distribution:

$$\log(\mu_i + 1) \sim (1 - w_s)\mathcal{N}\left(\log \frac{\epsilon_{\underline{s}}}{\epsilon_{\underline{s}} - 1}, \sigma_{\epsilon_{\underline{s}}}\right) + w_s\mathcal{N}\left(\log \frac{\epsilon_{\bar{s}}}{\epsilon_{\bar{s}} - 1}, \sigma_{\epsilon_{\bar{s}}}\right) \text{ for } i \in s$$

w_s is industry specific. It is the probability that a firm belongs to nest $\underline{g}(s)$ and $1 - w_s$ is the probability that a firm belongs to nest $\bar{g}(s)$ conditioned on the firm is from s. In other words, they are the ex-ante probabilities. They enter the density function as weights of the respective density components. After observing a firm's markups, the ex-post probability of a firm with markups $\mu_i + 1$ belonging to nest $\underline{g}(s)$ is:

$$\begin{aligned}
\mathbb{P}(i \in \underline{g}(s)|\mu_i) =& \frac{w_s f(\mu_i; \mu_{\underline{g}(s)}, \sigma_{\underline{g}(s)})}{w_s f(\mu_i; \mu_{\underline{g}(s)}, \sigma_{\underline{g}(s)}) + (1 - w_s) f(\mu_i; \mu_{\bar{g}(s)}, \sigma_{\bar{g}(s)})} \\
\mathbb{P}(i \in \bar{g}(s)|\mu_i) =& 1 - \mathbb{P}(i \in \underline{g}(s)|\mu_i)
\end{aligned}$$

$f(\mu_i; \mu_g, \sigma_g) = \frac{1}{\sigma_g}\phi\left(\frac{\log(1+\mu_i) - \log(1+\mu_g)}{\sigma_g}\right)$ for $g \in \{\bar{g}(s), \underline{g}(s)\}$ and $\phi(\cdot)$ is the density function of a standard normal distribution. $1 + \mu_g$ equals $\frac{\epsilon_g}{\epsilon_g - 1}$ for $g \in \{\underline{g}(s), \bar{g}(s)\}$. A firm belongs to $\underline{g}(s)$ if $\mathbb{P}(i \in \underline{g}(s)|\mu_i) > \mathbb{P}(i \in \bar{g}(s)|\mu_i)$, otherwise it belongs to $\bar{g}(s)$.

The log-likelihood of observing the data in an industry with mixture distribution is:

$$\ell\ell(\{\mu_i\}_{i \in s}|w_s, \mu_{\bar{g}(s)}, \mu_{\underline{g}(s)}, \sigma_{\bar{g}(s)}, \sigma_{\underline{g}(s)}) = \sum_{i \in s} \log(w_s f(\mu_i; \mu_{\underline{g}(s)}, \sigma_{\underline{g}(s)}) + (1 - w_s) f(\mu_i; \mu_{\bar{g}(s)}, \sigma_{\bar{g}(s)})) \tag{11}$$

The distribution for an industry with only one nest is a standard normal distribution:

$$\ell\ell(\{\mu_i\}_{i \in s}|\mu_s, \sigma_s) = \sum_{i \in s} \log(f(\mu_i; \mu_s, \sigma_s)) \tag{12}$$

We use the algorithm developed by Chen and Li (2009) to estimate whether an industry contains one nest or two nests. We then use the relevant likelihood function to estimate the demand elasticities and the latent nest structure, i.e. Equation (11) when there are two nests and Equation (12) when there is one nest.

Estimating mixture distribution is difficult because it is hard to identify overdispersion when two distribution components are close. To make our estimation robust, we use two algorithms for each industry from 50 starting values: the expected maximization (EM) algorithm (McLachlan and Peel (2004)) and a direct optimization of Equation (11). Both algorithms are sensitive to starting values because both objective functions contain numerous local maximums. Our simulation experiments show no guarantee which one works better (Appendix D.1), so including both increases the chance of finding or getting close to the global maximum.

Estimating the number of nests inside an industry is sensitive to outliers. If one nest's standard deviation is 100 times larger than the other nest in the same industry or when the weights of one nest is less than 5%, we treat the smaller nest as an outlier and drop observations in the nest. nests with only one observation are dropped as well. Test and estimation are implemented again after dropping all the outliers. More details on this are in the Appendix D.2.

C.3 Step 2: Identifying production elasticities and distortions

Profit maximization gives firms' capital and labor expenditures as a function of production elasticities and distortions:

$$\log\left(\frac{wL_i}{P_iY_i(\epsilon_g-1)/\epsilon_g}\right) = \log(\alpha_s^L) - \log(1+\tau_i^L)$$
$$\log\left(\frac{RK_i}{P_iY_i(\epsilon_g-1)/\epsilon_g}\right) = \log(\alpha_s^K) - \log(1+\tau_i^K)$$

We treat the left-hand side of the equations as known because ϵ_g is estimated in the previous step, R is set to 0.1, and the rest is directly observed. α_s^L and α_s^K can be interpreted as the location of the distribution of $\log\left(\frac{wL_i\mathbb{E}[e^{\delta_i}]}{P_iY_i(\epsilon_g-1)/\epsilon_g}\right)$ and $\log\left(\frac{RK_i\mathbb{E}[e^{\delta_i}]}{P_iY_i(\epsilon_g-1)/\epsilon_g}\right)$ while the variations in $\log(1+\tau_i^L)$ and $\log(1+\tau_i^K)$ determine the deviation from α_s^L and α_s^K. Since the mechanisms behind positive distortions may be very different from those behind negative distortions, we allow the distribution of positive τ_i^K and τ_i^L to differ from the distribution of negative ones for each industry, and we allow the probability of having positive distortions in an industry to be a free parameter. Distortions are independent and identically distributed within an industry and are independent across industries. Distortions on capital are independent from

distortions on labor.

$$\log(\tau_i^K + 1) \sim 2\kappa_s^K \mathbb{1}[\tau_i^K > 0] \cdot \mathcal{N}(0, \sigma_{s,+}^K) + (2 - 2\kappa_s^K)\mathbb{1}[\tau_i^K \leq 0] \cdot \mathcal{N}(0, \sigma_{s,-}^K)$$
$$\log(\tau_i^L + 1) \sim 2\kappa_s^L \mathbb{1}[\tau_i^L > 0] \cdot \mathcal{N}(0, \sigma_{s,+}^L) + (2 - 2\kappa_s^L)\mathbb{1}[\tau_i^L \leq 0] \cdot \mathcal{N}(0, \sigma_{s,-}^L)$$

The log-likelihood of observing P_iY_i, K_i, wL_i in industry s is the sum of the log-likelihood of $\{P_iY_i, K_i\}_{i\in s}$ and $\{P_iY_i, L_i\}_{i\in s}$:

$$\ell\ell(\{P_iY_i, K_i, wL_i\}_{i\in s}|\Theta_K, \Theta_L) = \sum_{i\in s} \ell\ell(P_iY_i, K_i|\Theta_K) + \ell\ell(P_iY_i, L_i|\Theta_L) \tag{13}$$

where

$$\ell\ell(P_iY_i, K_i|\Theta_K) = 2\kappa_s^K h\left(\theta_i^K; \log(\alpha_s^K), \sigma_{s,+}^K\right) \mathbb{1}\left[\frac{\alpha_s^K}{\theta_i^K} > 1\right] + (2 - 2\kappa_s^K) h\left(\theta_i^K; \log(\alpha_s^K), \sigma_{s,-}^K\right) \mathbb{1}\left[\frac{\alpha_s^K}{\theta_i^K} \leq 1\right]$$
$$\ell\ell(P_iY_i, L_i|\Theta_L) = 2\kappa_s^L h\left(\theta_i^L; \log(\alpha_s^L), \sigma_{s,+}^L\right) \mathbb{1}\left[\frac{\alpha_s^L}{\theta_i^L} > 1\right] + (2 - 2\kappa_s^L) h\left(\theta_i^L; \log(\alpha_s^L), \sigma_{s,-}^L\right) \mathbb{1}\left[\frac{\alpha_s^L}{\theta_i^L} \leq 1\right]$$

Θ_K indicates the parameters related to capital expenditure $\{\kappa_s^K, \alpha_s^K, \sigma_{s,+}^K, \sigma_{s,-}^K\}$, and Θ_L indicates the parameters related to labor expenditure $\{\kappa_s^L, \alpha_s^L, \sigma_{s,+}^L, \sigma_{s,-}^L\}$. $\mathbb{1}[\cdot]$ takes 1 if the statement inside is true and 0 otherwise. $h(\cdot; \log(\alpha), \sigma)$ is the log density function of a normal distribution with mean $\log(\alpha)$ and standard deviation σ. θ_i^K and θ_i^L are the log of capital and labor expenditure share corrected by expected markups and expected cost shocks. As mentioned above, θ_i^K and θ_i^L are treated as known.

$$\theta_i^K = \log\left(\frac{RK_i\mathbb{E}[e^{\delta_i}]}{P_iY_i(\epsilon_g - 1)/\epsilon_g}\right)$$
$$\theta_i^L = \log\left(\frac{wL_i\mathbb{E}[e^{\delta_i}]}{P_iY_i(\epsilon_g - 1)/\epsilon_g}\right)$$

The identification is a simple maximum likelihood estimation (MLE) except that the log-likelihood function is not differentiable with regard to α_s^K and α_s^L when $\frac{\alpha_s^K}{\theta_i^K} = 1$ and $\frac{\alpha_s^L}{\theta_i^L} = 1$. Standard optimization methods for this type of problem does not guarantee a global maximum. We propose a combination of grid searching and first-order conditions that guarantees global maximum under a mild restriction that α_s^K and α_s^L is in $(0, 1)$. It is mild because positive labor and capital expenditure requires α_s^K and α_s^L to be positive. α_s^K and α_s^L larger than 1 means firms have increasing returns to scale in one production factor while holding all the other factors constant. This rarely holds in reality. Since the objective function is continuously differentiable once α_s^K and α_s^L are fixed, we maximize the

log-likelihood function with respect to the rest parameters for each guess of α_s^K and α_s^L. We then pick the α_s^K and α_s^L which give the highest log-likelihood. Because the objective function is a linear summation of a capital part and a labor part, we can estimate Θ_K and Θ_L separately. In other words, instead of searching over a two-dimension unit square, we search over two independent one-dimension $(0, 1)$ intervals, which significantly speeds up the process.

From Equation (13), when α_s^K and α_s^L are fixed, estimator of the remaining parameters are:

$$\widehat{\kappa_s^K} = \frac{\sum_{i\in s} \mathbb{1}\left[\frac{\alpha_s^K P_i Y_i(\epsilon_i-1)/\epsilon_i}{RK_i\widehat{\mathbb{E}[e^{\delta_i}]}} > 1\right]}{N_s}$$

$$\widehat{\kappa_s^L} = \frac{\sum_{i\in s} \mathbb{1}\left[\frac{\alpha_s^L P_i Y_i(\epsilon_i-1)/\epsilon_i}{wL_i\widehat{\mathbb{E}[e^{\delta_i}]}} > 1\right]}{N_s}$$

$$\widehat{(\sigma_+^K)^2} = \frac{\sum_{i\in s} \mathbb{1}\left[\frac{\alpha_s^K P_i Y_i(\epsilon_i-1)/\epsilon_i}{RK_i\widehat{\mathbb{E}[e^{\delta_i}]}} > 1\right]\left(\log\left(\frac{\alpha_s^K P_i Y_i(\epsilon_i-1)/\epsilon_i}{RK_i\widehat{\mathbb{E}[e^{\delta_i}]}}\right)\right)^2}{\sum_{i\in s} \mathbb{1}\left[\frac{\alpha_s^K P_i Y_i(\epsilon_i-1)/\epsilon_i}{RK_i\widehat{\mathbb{E}[e^{\delta_i}]}} > 1\right]}$$

$$\widehat{(\sigma_{s,-}^K)^2} = \frac{\sum_{i\in s} \mathbb{1}\left[\frac{\alpha_s^K P_i Y_i(\epsilon_i-1)/\epsilon_i}{RK_i\widehat{\mathbb{E}[e^{\delta_i}]}} < 1\right]\left(\log\left(\frac{\alpha_s^K P_i Y_i(\epsilon_i-1)/\epsilon_i}{RK_i\widehat{\mathbb{E}[e^{\delta_i}]}}\right)\right)^2}{\sum_{i\in s} \mathbb{1}\left[\frac{\alpha_s^K P_i Y_i(\epsilon_i-1)/\epsilon_i}{RK_i\widehat{\mathbb{E}[e^{\delta_i}]}} < 1\right]}$$

$$\widehat{(\sigma_+^L)^2} = \frac{\sum_{i\in s} \mathbb{1}\left[\frac{\alpha_s^L P_i Y_i(\epsilon_i-1)/\epsilon_i}{wL_i\widehat{\mathbb{E}[e^{\delta_i}]}} > 1\right]\left(\log\left(\frac{\alpha_s^L P_i Y_i(\epsilon_i-1)/\epsilon_i}{wL_i\widehat{\mathbb{E}[e^{\delta_i}]}}\right)\right)^2}{\sum_{i\in s} \mathbb{1}\left[\frac{\alpha_s^L P_i Y_i(\epsilon_i-1)/\epsilon_i}{wL_i\widehat{\mathbb{E}[e^{\delta_i}]}} > 1\right]}$$

$$\widehat{(\sigma_{s,-}^L)^2} = \frac{\sum_{i\in s} \mathbb{1}\left[\frac{\alpha_s^L P_i Y_i(\epsilon_i-1)/\epsilon_i}{wL_i\widehat{\mathbb{E}[e^{\delta_i}]}} < 1\right]\left(\log\left(\frac{\alpha_s^L P_i Y_i(\epsilon_i-1)/\epsilon_i}{wL_i\widehat{\mathbb{E}[e^{\delta_i}]}}\right)\right)^2}{\sum_{i\in s} \mathbb{1}\left[\frac{\alpha_s^L P_i Y_i(\epsilon_i-1)/\epsilon_i}{wL_i\widehat{\mathbb{E}[e^{\delta_i}]}} < 1\right]}$$

The last four equations are from first-order conditions. The right-hand side are either observed or estimated in previous steps except for α_s^K and α_s^L. We calculate the log-likelihood of the capital part and labor part at each guess of α_s^K and α_s^L separately. $\hat{\alpha}_s^K$ and $\hat{\alpha}_s^L$ maximize

the log-likelihood.

$$\hat{\alpha}_s^K = \arg\max_{\alpha_s^K} \sum_{i \in s} \ell\ell(P_i Y_i, K_i | \hat{\kappa}_s^K, \alpha_s^K, \hat{\sigma}_{s,+}^K, \hat{\sigma}_{s,-}^K)$$
$$\hat{\alpha}_s^L = \arg\max_{\alpha_s^L} \sum_{i \in s} \ell\ell(P_i Y_i, L_i | \hat{\kappa}_s^L, \alpha_s^L, \hat{\sigma}_{s,+}^L, \hat{\sigma}_{s,-}^L)$$

where $(\hat{\kappa}_s^K, \hat{\sigma}_{s,+}^K, \hat{\sigma}_{s,-}^K, \hat{\kappa}_s^L, \hat{\sigma}_{s,+}^L, \hat{\sigma}_{s,-}^L)$ are determined as above for each α_s^K and α_s^L. $\hat{\alpha}_s^K$ and $\hat{\alpha}_s^L$ are determined using a grid search on two $(0, 1)$ intervals.

Using the estimated capital intensity $\hat{\alpha}_s^K$ and $\hat{\alpha}_s^L$, we can calculate the distortions faced by firm i:

$$1 + \hat{\tau}_i^L = \frac{\hat{\alpha}_s^L P_i Y_i (\hat{\epsilon}_g - 1)/\hat{\epsilon}_g}{w L_i \widehat{\mathbb{E}[e^{\delta_i}]}}$$
$$1 + \hat{\tau}_i^K = \frac{\hat{\alpha}_s^K P_i Y_i (\hat{\epsilon}_g - 1)/\hat{\epsilon}_g}{R K_i \widehat{\mathbb{E}[e^{\delta_i}]}}$$

D Mixture estimation and outliers

D.1 Mixture Estimation

The EM algorithm essentially searches for the fixed point of a function that is not a contraction mapping. It does not guarantee converging to the global maximum or minimum and it may not even converge at all. Existing optimizers can only ensure local maximum of Equation (11) in Section C.2, which contains a lot of local maximums. To improve the robustness of our estimators, we draw 50 triplets of random starting values for p, $\mu_{\bar{g}(s)}$, and $\mu_{\underline{g}(s)}$ in each industry s.

The random values of p are independent draws from a uniform distribution on (0,1). $\mu_{\bar{g}(s)}$ and $\mu_{\underline{g}(s)}$ are two independent draws from the interval three sample standard deviations away from the sample mean. We use the EM algorithm of Benaglia et al. (2009) developed for R. When optimizing the likelihood function directly, we use the optim() function in R with BFGS method. We pick BFGS, or quasi-Newton because it provides the best combination of speed and accuracy among all the available R optimizers that we are aware of.

Table 11, Table 12, and Table 13 compare 6 different methods' performance on simulated data. EM and BFGS are the ones we pick. NM is the method of Nelder and Mead (1965). SANN is a variant of simulated annealing (Claude J. P. Bélisle (1992)). NR and BHHH are from Henningsen and Toomet (2011), with NR referring to Newton-Raphson and BHHH to Berndt-Hall-Hall-Hausman.

We simulate two types of data to test how the algorithms works when the difficulties of

identification change. The first data is very hard to identify with equal mean of 1 and very close standard deviations $\sigma_{\bar{g}(s)} = 1$ and $\sigma_{\underline{g}(s)} = 1.5$. The weight p is 0.25. The second also has weight p equal to 0.25 but with means further apart relatively to standard deviations: $\mu_{\bar{g}(s)} = 0$, $\mu_{\underline{g}(s)} = 4$, $\sigma_{\bar{g}(s)} = 1$ and $\sigma_{\underline{g}(s)} = 2$.

TABLE 11: Estimates under different methods: 50 random starting values of p, $\mu_{\bar{g}(s)}$, $\mu_{\underline{g}(s)}$; sample size:200

true values	methods		p	$\mu_{\bar{g}(s)}$	$\mu_{\underline{g}(s)}$	$\sigma_{\bar{g}(s)}$	$\sigma_{\underline{g}(s)}$	nll
	EM	mixtools	0.4683282	0.2859527	1.591074	0.8251734	1.306893	329.2560831
	BFGS	optim	0.4679615	0.2856622	1.590424	0.8249206	1.306958	329.2560826
	NM	optim	0.5325401	1.589622	0.2852179	1.306917	0.8245313	329.2560842
(0.25,1,1,1,1.5)	SANN	optim	0.5399907	1.587284	0.2726049	1.308941	0.8179273	329.2592109
	NR	maxLik	0.4679779	0.285673	1.590456	0.8249318	1.306953	329.2560826
	BHHH	maxLik	0.4679779	0.285673	1.590456	0.8249318	1.306953	329.2560826
	EM	mixtools	0.1047498	-0.5673427	3.38516	0.4612274	2.161369	450.5243907
	BFGS	optim	0.8952547	3.385141	-0.5673413	2.161387	0.4612123	450.5243907
	NM	optim	0.8951295	3.385244	-0.5677538	2.160873	0.4611698	450.5244077
(0.25,0,4,1,2)	SANN	optim	0.1026536	-0.5735226	3.373923	0.4525442	2.161418	450.5287012
	NR	maxLik	0.1047467	-0.5673531	3.385145	0.4612135	2.161376	450.5243907
	BHHH	maxLik	0.1047467	-0.5673531	3.385145	0.4612135	2.161376	450.5243907

The maximum step when generating random starting values is 1 standard deviation.

TABLE 12: Standard deviation of estimates across the 50 starting values; sample size: 200

true values	methods		p	$\mu_{\bar{g}(s)}$	$\mu_{\underline{g}(s)}$	$\sigma_{\bar{g}(s)}$	$\sigma_{\underline{g}(s)}$	nll$*e8$
	EM	mixtools	1.38e-04	1.11e-04	2.44e-04	9.55e-05	2.36e-05	0.00e+00
	BFGS	optim	1.66e-01	3.99e+00	2.15e-01	3.73e+00	1.08e-02	1.53e+08
	NM	optim	1.21e-01	7.88e-01	5.02e-01	3.10e-01	1.70e-01	9.26e+07
(0.25,1,1,1,1.5)	SANN	optim	1.28e-01	6.81e-01	4.53e-01	5.10e-01	1.56e-01	9.26e+07
	NR	maxLik	5.17e-02	6.89e+01	2.05e-01	1.61e+01	7.92e-01	5.53e+08
	BHHH	maxLik	2.20e-01	1.57e+01	4.41e-01	3.01e+02	8.63e-02	1.42e+09
	EM	mixtools	2.88e-07	9.57e-07	1.35e-06	1.29e-06	6.99e-07	5.01e+00
	BFGS	optim	4.53e-02	1.21e+01	1.86e-01	1.12e+01	6.79e-01	3.03e+08
	NM	optim	9.74e-02	1.87e+00	4.29e-01	9.58e-01	1.31e-01	2.52e+08
(0.25,0,4,1,2)	SANN	optim	3.11e-02	1.47e+00	1.64e-01	1.33e+00	7.42e-02	2.01e+08
	NR	maxLik	2.21e-01	2.48e+02	3.06e-01	8.58e+01	3.01e-01	2.27e+08
	BHHH	maxLik	1.25e-01	4.33e+02	7.76e-01	1.70e+02	1.97e-01	1.59e+09

Normalization: $p < 1 - p$

TABLE 13: Execution time for one starting value (in seconds)

mixtools package	optim package			maxLik package	
EM	BFGS	NM	SAANN	NR	BHHH
0.06	0.01	0.02	1.34	0.18	4.93

Using 50 random starting values, all methods generate similar results apart from the lack of identification of the components' names. In spite of sample bias, BFGS, NR, and BHHH are slightly better at finding the minimum as they produce the lowers negative log-likelihood (nll). EM also does well when components' means are away from each other. A closer look tells us EM produces a lot less variations in the negative log likelihood (nll) across random starting values, suggesting if the number of random starting values is not large, it is safer to use EM than BFGS, NR or BHHH. BFGS, NR and BHHH perform better when there are a large number of starting values but may lead to estimates far away from the global minimum when starting values are few. The execution time for one starting value shows BFGS is the fastest. Although the simulation data favors BFGS, BFGS performs badly on industry "1753" in our data. Therefore, we use both EM and BFGS in our estimation.

D.2 Outliers

We drop nests that contain only one observations. We also drop nests whose standard deviation is 1/100 of the other nest in the same industry and its weight is less than 5%. This drops 8 observations from 8 industries, i.e. all the dropped nests turn out to contain only one observation. After dropping these outliers, we rerun the test of mixture and re-estimate the parameters accordingly.

E Identification issue of correcting the biases in inferred markups

E.1 Cobb-Douglas production function

Integrating over the marginal cost function and divide it by production gives:

$$\mathrm{AC}_i = r_s \mathrm{MC}_i$$

where AC_i is the average cost, MC_i is the marginal cost, and r is the returns to scale, i.e. $r_s = \alpha_s^L + \alpha_s^K$. The revenue-cost ratio is:

$$\log\left(\frac{P_i Y_i}{Y_i \mathrm{AC}_i}\right) = \log\left(\frac{\epsilon_g}{\epsilon_g - 1}\right) - \log(r_s) + \log\left(\mathbb{E}[e^{\delta_i}]\right) - \delta_i$$

When there is one nest, its distribution is:

$$\log\left(\frac{P_iY_i}{Y_i\text{AC}_i}\right) \sim \mathcal{N}\left(\log\frac{\epsilon_s}{\epsilon_s-1} - \log(r), \sigma_{\epsilon_s}\right) \text{ for } i \in s$$

when there are two nests, its distribution is:

$$\log\left(\frac{P_iY_i}{Y_i\text{AC}_i}\right) \sim w_s\mathcal{N}\left(\log\frac{\epsilon_{\underline{s}}}{\epsilon_{\underline{s}}-1} - \log(r), \sigma_{\epsilon_{\underline{s}}}\right) + (1-w_s)\mathcal{N}\left(\log\frac{\epsilon_{\bar{s}}}{\epsilon_{\bar{s}}-1} - \log(r), \sigma_{\epsilon_{\bar{s}}}\right) \text{ for } i \in s$$

Denote $\Xi \equiv \frac{\epsilon_g}{\epsilon_g-1}\frac{1}{r_s} = \frac{\epsilon_g}{\epsilon_g-1}\frac{1}{\alpha_s^L+\alpha_s^K}$. Our second estimation step can still estimate the mean but instead of directly estimating the demand elasticities, we can only estimate $\log(\Xi)$, denoted as $\widehat{\log(\Xi)}$.

In the third step, we use these equations:

$$\log\left(\frac{wL_i\mathbb{E}[e^{\delta_i}]}{P_iY_i}\right) = \log(\alpha_s^L) - \log\frac{\epsilon_g}{\epsilon_g-1} - \log(1+\tau_i^L)$$
$$\log\left(\frac{RK_i\mathbb{E}[e^{\delta_i}]}{P_iY_i}\right) = \log(\alpha_s^K) - \log\frac{\epsilon_g}{\epsilon_g-1} - \log(1+\tau_i^K)$$

We denote $\Xi^L \equiv \frac{\epsilon_g}{\epsilon_g-1}\frac{1}{\alpha_s^L}$ and $\Xi^K \equiv \frac{\epsilon_g}{\epsilon_g-1}\frac{1}{\alpha_s^K}$. The third step estimation gives: $\widehat{\log(\Xi^L)}$ and $\widehat{\log(\Xi^K)}$. If we estimate the parameters simultaneously, we need to solve the following equation for $\hat{\epsilon}_g$, $\hat{\alpha}_s^L$ and $\hat{\alpha}_s^K$:. We denote them as

$$\widehat{\Xi} \equiv \frac{\hat{\epsilon_s}}{\hat{\epsilon_s}-1}\frac{1}{\hat{\alpha}_s^L+\hat{\alpha}_s^K} \tag{14}$$
$$\widehat{\Xi}^L \equiv \frac{\hat{\epsilon_s}}{\hat{\epsilon_s}-1}\frac{1}{\hat{\alpha}_s^L} \tag{15}$$
$$\widehat{\Xi}^K \equiv \frac{\hat{\epsilon_s}}{\hat{\epsilon_s}-1}\frac{1}{\hat{\alpha}_s^K} \tag{16}$$

Although we have three equations for three unknowns, but the assumption of CES demand and Cobb-Douglas production function render one of the three equations redundant. If we know the true value of Ξ, Ξ^L, and Ξ^K, then we must have $\Xi^L + \Xi^K = \Xi$. Therefore, only two of these three equations contain useful information about the parameters. The extra information brought by the third one is only about the difference between the sample analogues and the true values. It is not possible to identify two equations for three unknowns. If we increase $\frac{\epsilon_g}{\epsilon_g-1}$ by a factor of ϕ, we can keep the equations hold by increase α_s^L and α_s^K by ϕ.

However, using the estimators from our model, one can still ignore this identification

issue and implement the correction of markups using the returns to scale estimated from the third step. This process will not converge to consistent estimators. In fact, whether it converges or not only depends on whether the absolute value of $\frac{\hat{X}i}{\hat{\Xi}^L+\hat{\Xi}^K}$ is larger than 1. As discussed above, if our estimated Ξ, Ξ^L, and Ξ^K equal their the true values, the returns to scale estimated in our third step should be 1.

If we start with a guess of ϵ_g, denoted as $\hat{\epsilon}^0$. Use Equation (16) and Equation (15), we get estimates for α_s^K and α_s^L, denoted as $\hat{\alpha}_K^1$ and $\hat{\alpha}_L^1$:

$$\hat{\alpha}_K^1 = \widehat{\Xi}^K * \frac{\epsilon^0}{\epsilon^0 - 1}$$
$$\hat{\alpha}_L^1 = \widehat{\Xi}^L * \frac{\epsilon^0}{\epsilon^0 - 1}$$

Use Equation (14), we update $\hat{\epsilon}^0$ to $\hat{\epsilon}^1$:

$$1 - \frac{1}{\hat{\epsilon}^1} = \frac{\widehat{\Xi}}{\widehat{\Xi}^L + \widehat{\Xi}^K}(1 - \frac{1}{\hat{\epsilon}^0})$$

If $|\frac{\widehat{\Xi}}{\Xi^L+\Xi^K}| < 1$, then we will converge to the unique fixed point $1 - \frac{1}{\hat{\epsilon}} = 0$. However, if we know the true value, we must have $\Xi = \Xi^L + \Xi^K$, which means any point is a fixed point. We can not identify the parameters. One can also see this by noticing Equation (14), Equation (16) and Equation (15) are in fact only two equations. Any two of these equation can derive the third one. If we increase $\frac{\epsilon_g}{\epsilon_g - 1}$ by a factor of k, we can keep the equations hold by increase α_s^L and α_s^K by k.

If we ignore this issue and still update estimation this way, the updating is possible not because it is not a fixed point but because we do not observe the true value of $\frac{\Xi}{\Xi^L+\Xi^K}$. Depending on the difference between estimation and the true value, $1-\frac{1}{\hat{\epsilon}}$ may either converge to 0 or to infinity. It contains no meaningful information about demand elasticities. Such identification problem also means simultaneous estimating all the parameters won't work neither.

E.2 More general production function: homogeneous of degree r

This problem remains as long as we can only use revenue-cost ratio to infer markups and when production function is homogeneity of degree r. For simplicity of demonstration, we omit firm or nest subscripts, distortions, and cost shocks. We omit distortions and cost shocks because we only need to show using revenue-cost ratio, we can only identify $\widehat{\log \Xi}$. Using the sum of labor and capital expenditure share, we can also only identify $\widehat{\log \Xi}$. Hence,

once we use the labor share and the capital share, the information contained in revenue-cost ratio is redundant for parameter estimation. We are then left with only two equations. The first-order condition of profits maximization gives:

$$\frac{\epsilon-1}{\epsilon}PF_1 = r$$
$$\frac{\epsilon-1}{\epsilon}PF_2 = w$$

where $F_1 = \frac{\partial F(K,L)}{\partial K}$ and $F_2 = \frac{\partial F(K,L)}{\partial L}$. Due to homogeneity of degree r, $rF(K,L) = KF_1 + LF_2$. Combine the F.O.C.:

$$rK + wL = \frac{\epsilon-1}{\epsilon}P(F_1K + F_2L) = \frac{r\epsilon-1}{\epsilon}PF(K,L)$$

Hence

$$\frac{rK+wL}{PF(K,L)} = \log(r) - \log\frac{\epsilon}{\epsilon-1}$$

We next need to show under this more general production function, we still haveAC $= r$MC It is easy to show that if for production level Y, K^* and L^* are the optimal capital and laboe, then for any factor $\gamma > 0$, the optimal capital and labor for producing $\gamma^r Y$ are γK^* and γL^*. We denote the optimal amount of capital and labor for the first unit of output as θ_K and θ_L. For any level of production, we can write it as

$$Y = F(Y^{1/r}\theta_K, Y^{1/r}\theta_L)$$

Its cost under the optimal capital and labor choices is

$$c = Y^{1/r}\theta_K R + Y^{1/r}\theta_L w$$

Differentiate cost with respect to Y:

$$\frac{dc}{dY} = \frac{1}{r}\frac{c}{Y}$$

Hence AC $= r$MC. Therefore,

$$\log\left(\frac{P_iY_i}{Y_i\text{AC}_i}\right) = \log\left(\frac{\epsilon}{\epsilon-1}\right) - \log(r)$$

F A model with intangible assets

Our structural estimation of returns to scale is on average 0.7 which appears to cause concerns over inferring markups using revenue-cost ratios. In fact, the seemingly inconsistency is resolved if we use a more complete model where both tangible and intangible assets are included. Capital in our main results contains only tangible assets. However, production does require intangible assets. A constant-returns-to-scale can appear decreasing returns to scale if we do not include the intangible assets. In this section, we will show that the TFP gains we find comes from equalizing the marginal revenue of labor and tangible assets while treating intangible asset as a state variable.

Denote the intangible assets of firm i as N_i which is taken as given when the firm maximize its profits at time t. We treat N_i as a state variable because it is a lot more difficult to adjust intangible assets in one period. One may take into account today's choice on future value of intangible of intangible assets but doing so requires another project of dynamic model. To keep things simple, we shut down the dynamic part and treat N_i as given. The production function is then:

$$Y_i = A_i K_i^{\alpha_s^K} L_i^{\alpha_s^L} N_i^{\alpha_s^N}$$

Since N_i is fixed, we can rewrite the production function:

$$Y_i = \tilde{A}_i K_i^{\alpha_s^K} L_i^{\alpha_s^L}$$

where $\tilde{A}_i = A_i N_i^{\alpha_s^N}$. Replacing A_i by $\tilde{A}_i$, all the other results are the same as those in Section 3.