Factor income shares and input distortions in China [*]

Xiaoyue Zhang[†] Junjie Xia [‡]

November 21, 2022

The Latest Version

Abstract

This paper quantifies the impact of input distortions on the labor and capital shares in Chinese aggregate income, where the economy is allowed to have rich variations in productivity, technology, and demand elasticities. Under the nested constant elasticity of substitution demand, we decompose the share changes into changes due to within-nest distortions and between-nest distortions. We use firm-level data to estimate industry-specific production elasticities, nest-specific demand elasticities when the nest structure is not perfectly observed, and firm-specific input distortions. We find that removing input distortions in China will raise the labor share by 7 percentage points, and lower the capital share by 1 percentage point. Removing between-nest distortions explains two-third of the increase in the labor share, but offsets 90% of the decrease in the capital share caused by removing within-nest distortions.

Keywords: Misallocation, productivity, input distortions, factor income shares, latent market structure

[*] 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 aggregate income. The input distortions include any institutional and market imperfections that cause capital and labor to be allocated not according to their market values. 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. Using parameters estimated from firm-level data, we quantify the impact on the labor and the capital shares when removing input distortions in an economy with rich variations in productivity, technology, and demand elasticities. We find that removing input distortions in China raises the labor share by 7 percentage points and reduces the capital share by 1 percentage point. While both the size of the distortions and the structure of the economy matter, the latter explains two-thirds of the increase for labor but offsets 90% of the decline for capital caused by the former.

It is generally accepted that China has experienced rapid TFP growth in the early twenty-first century, and one-third of the growth can be explained by reductions in input distortions (HK). However, little is known about the distortions' impact on the Chinese labor share. In spite of the heated discussion about a global decline in the labor share, whether the Chinese labor share declines during that period is the subject of debate (Karabarbounis and Neiman (2014), Gutirrez and Piton (2020)).

Recent studies on explaining why aggregate labor shares decline in the majority of countries in the world, and in the US in particular, motivate us to examine labor shares as well as capital shares in a complicated system. In this system, the shares are affected by the size of the distortions, the joint distribution of technology, demand elasticities, and productivity, as well as how different markets inside an economy are aggregated for a hypothetical representative consumer, namely the aggregation structure. Piketty (2014) and Blanchard et al. (1997) show that distortions in the labor market matter. 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. 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. The offsetting occurs in our system because when holding the industry-level distortions constant, different aggregation structures, and different joint distributions of demand elasticities and technology can raise or reduce the changes in factor shares or even flip their signs.

We model input distortions as wedges in factor prices that firms face following Restuccia and Rogerson (2008). Positive wedges indicate that firms face higher prices for capital and labor, and therefore use less capital and labor than in an economy free of input distortions. Negative wedges indicate lower prices and 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. These firm-specific distortions are usually asymmetric, which is crucial for the structure of the economy to have a significant influence on the aggregate shares.

We offer a way to decompose the factor-share changes into two parts: the part due to the size of the distortions at the market level, and the part due to the structure of the economy which includes the aggregation structure, heterogeneous technology, and heterogeneous demand elasticities. More specifically, we borrow the model from HK where a market is defined as a cluster or a nest of firms who face the same demand elasticities and have the same production elasticities with regard to inputs, i.e. the same technology, but can differ on productivity and input distortions.[2] In essence, under the nested constant elasticity of substitution (nested-CES) demand as in HK, changes in aggregate factor shares are a weighted sum of within-nest distortions. The within-nest distortions are the change in nest-level factor shares from removing input distortions within a nest and are our measure of the nest-level distortions. They are affected by the level and dispersion of firm-specific distortions within a nest, as well as their interactions with firm-level productivity. The weights are a composite of demand elasticities, production elasticities, and parameters for cross-nest aggregation. It determines how reallocation across nests affects aggregate factor shares. We define this reallocation as caused by between-nest distortions. This within-and-between decomposition follows the spirit of Olley and Pakes (1996), where the aggregate allocation efficiency is decomposed into reallocation among plants within a firm and reallocation between firms. Similar to them, changes in aggregate factor shares due to the within-nest distortions are the unweighted sum of nest-level changes in factor shares and the remaining changes result

[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.

[2] In Hsieh and Klenow (2009), all the markets are assumed to have the same demand elasticities. We allow demand elasticities to differ across markets.

from between-nest distortions. The unweighted sum corresponds to an artificial economy where the weights for each nest are the same, which means that there is no variation in demand elasticities or technology and that nests are treated equally in aggregation.

The structure of the economy may amplify or offset the changes from within-nest distortions. For example, when within-nest distortions raise factor shares, removing between-nest distortions will amplify the increase if factors are reallocated to larger markets with higher demand elasticities and a higher demand for the factors, but will offset the increase if the reallocation is in the opposite direction. The direction of reallocation can be either way for given within-nest distortions because it is determined by how within-nest distortions interact with technology, demand elasticities, and aggregation parameters.

One challenge is to infer from data what firms belong to the same nest. HK assumes each industry as a nest. We first document evidence suggesting that treating each industry as a nest with constant demand elasticities omits useful information about markups. Second, we propose to treat each nest as a latent cluster and to use observed firm characteristics to infer the latent clusters as well as their demand elasticities. Since the demand elasticities within each nest are constant under nested-CES demand, unless we treat each firm as a nest, there will always be some variation in markups that cannot be explained by the variation of demand elasticities. The remaining variation is then modeled using unexpected cost shocks and price rigidity.

Another challenge is to identify production elasticities and firm-specific input distortions. In HK, production elasticities are calibrated using American firms' production elasticities. This means the differences between American firms' and Chinese firms' production elasticities are treated as input distortions. In fact, US firms are more capital-intensive and Chinese firms are more labor-intensive. Calibration based on American firms' parameters suggests that most Chinese firms use too much labor and too little capital. Therefore, instead of an increasing labor share and a decreasing capital share, we find a 3% decrease in the labor share and a 30% increase in the capital share after removing the input distortions. The primary reason behind this is a systematic upward bias in capital distortions, and a systematic downward bias in labor distortions due to the calibration. These systematic biases are not costly when talking about within-nest TFP gains, a question that HK is interested in because within-market TFP gains depend on the dispersion of firm-level factor shares, not the level. However, these biases will appear in the predicted factor share changes and flip the results. Besides, methods of estimating production function 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 here because the crucial assumption of input use as a monotonic function of productivity is violated. If input distortions are positively correlated with productivity, higher-productivity firms no longer hire more inputs.

The difficulty in estimating production elasticities is that observed capital and labor income shares are affected by demand elasticities, input distortions, and production elasticities. While demand elasticities are estimated during the identification of the latent nests, we still need to estimate production elasticities and distortions. Without any parametric restrictions, this model is not identified. We disentangle production elasticities from input distortions by allowing a flexible distribution of distortions, and use the assumption that production elasticities are constant within markets. When production elasticities are the same within a market, variation in observed firm-level capital and labor income shares reflects distortions after controlling for differences in demand elasticities. The key assumption about the distortion distributions is that distributions reach their modes when distortions are small or zero. It is flexible because it 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 the mechanism may vary across industries. Furthermore, an industry's probability of having positive distortions is a free parameter that is industry specific, so that we do not need to assume a ratio of positive distortions in an industry ex ante. This means neither the mean nor any percentile of the distortions within an industry is set ex ante. In the framework where firms inside an industry have the same technology, the assumption that modes are reached when input distortions are small gives us a conservative estimation of the magnitude of the distortions.

Applying our model to the Chinese Annual Survey of Manufacturing in 2005, we find that 90% of the industries are better modeled as having more than one nest. In terms of input distortions, the distortions on capital are more often negative than on labor suggesting that capital is perhaps more frequently subsidized. The aggregate labor share increases by about 7%, and only one-third of the increase results from the within-nest distortions. The structure of the economy explains the remaining two-thirds. This happens because labor is in general reallocated to larger markets with higher demand for labor and low markups. The capital share will decrease by 1% after removing the input distortions. However, this does not mean that capital distortions are smaller in magnitude. Without the resources reallocated across markets, the decrease will be almost 18% reflecting the negative within-nest distortions for capital in general. Similar to labor, capital is also reallocated to larger markets with higher capital demand, which offsets 90% of the decline in the capital share. A weak pattern of capital reallocated to markets with higher markups raises the decline by a small 0.06 percentage point.

To the best of our knowledge, this is the first paper that studies the impact of removing firm-specific input distortions on the aggregate factor income shares in a rich model with an intricate aggregation structure and heterogeneous demand elasticities, technology, and productivity. Authors working on other types of rich systems are also interested in analyzing the impact of different channels. Autor et al. (2020) decomposes changes in the labor share, and shows that the observed decline in the US labor share is mainly due to reallocation across firms with different markups instead of a decline in the unweighted mean of labor shares. We find a limited impact from reallocating resources across firms with different demand elasticities but a sizable change due to the other channels mentioned above. Edmond et al. (2019) decomposes the predicted gains from increasing competition and finds that the net gains may be little even though each channel has a significant impact. This result is similar to our decomposition of the predicted capital share changes, where the net change is almost 0 but the unweighted change is about -17%.

Our study contributes to the literature on heterogeneous markups by offering a way to model demand with heterogeneous markups when the market is not observed. Existing studies model heterogeneous markups by introducing endogenous markups, such as nested CES with oligopolies (Atkeson and Burstein (2008), Edmond et al. (2015), and Burstein et al. (2020)), Kimball preferences (Klenow and Willis (2016)), translog preferences (Feenstra and Weinstein (2017)), the Constant Revenue Elasticity of Marginal Revenue demand (Mrázová et al. (2021)), and hyperbolic absolute risk aversion preference (Haltiwanger et al. (2018)). These studies explicitly model different levels of pass-through but have to rely on accurate information on the market structure. We assume complete pass-through but do not require accurate knowledge about the market structure.

The remainder of the paper is organized as follows. We introduce the data set in Section 2 and provide empirical evidence that multiple markets are likely to exist in an industry. Section 3 describes our theoretical model and the decomposition framework. 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 reason for multiple identification steps, and a model extension that includes intangible assets.

2 Data

Our data source is the Chinese Annual Survey of Manufacturing (ASM) collected by the National Bureau of Statistics of China and we only use the year 2005.[3] This data set has been

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

used by previous studies including HK, Song et al. (2011), and David and Venkateswaran (2019). It includes all the nonstate firms with revenue above 5 million RMB (approximately $\$600,000$) and all the state-owned enterprises (SOEs).

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 debt, negative sales, et cetera. A full list of the types of observations dropped is provided in the 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, there are 15% of the firms with negative profits. Table 1 provides the summary statistics. Value added is the amount of revenues after netting out expenditures other than capital and labor, or in other words, it is the sum of profits, and the expenditure of 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	229,241	13,814.46	122	2,517	5,377	13,250	277,908
K	229,241	16,366.41	83.76	1,620.23	4,211.66	12,151.88	515,954.20
wL	229,241	2,730.73	80	583	1,188	2,665	78,956
revenue	229,241	50,184.74	2	9,500	19,457	45,994	11,041,153
cost	229,241	43,075.61	1	7,935	16,481	39,072	10,757,115
profits	229,241	2,370.47	−292,087	72	480	1,815	415,879
revenue/cost	229,241	1.21	0.81	1.08	1.14	1.25	4.68
wL/value added	229,241	0.32	0.01	0.12	0.23	0.42	3.15

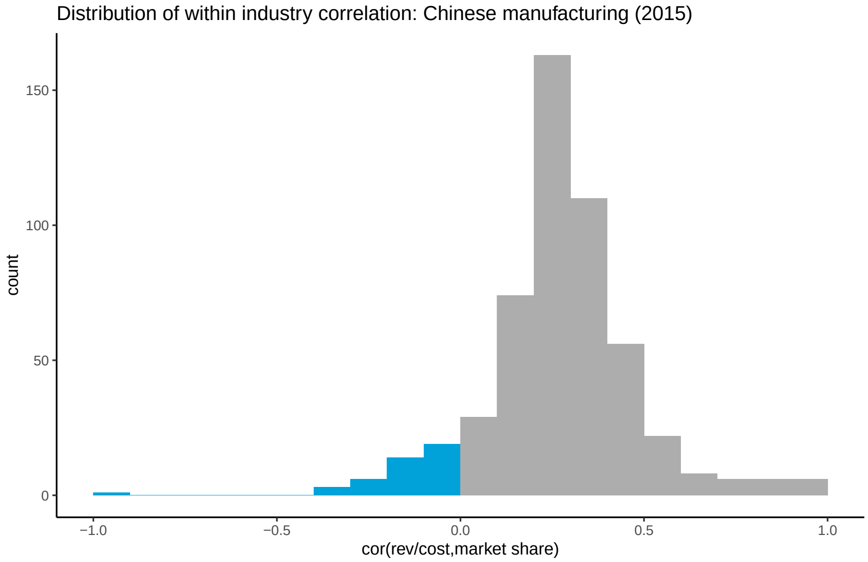
One well-known limitation of this data is that the labor expenditure does not include the non-wage portion and that the aggregate labor share of value added is too low compared to the one inferred by the Chinese Input-and-Output Table and national accounts. HK is aware of this issue and scales up each firm's labor expenditure by the same proportion so that the aggregate labor share reaches 50%. We run our estimation using unscaled labor expenditure because the under-reporting seems more severe in large firms than small firms. In fact, very large firms are concentrated in the area with very low labor expenditure while smaller firms are more spread out. Figure 7 in the Appendix A demonstrates this pattern with more details. This discrepancy is intuitive because larger firms are more capable of providing non-wage labor income. Although the aggregate labor expenditure share in this data is around 20%, the average of all the firms' labor expenditure share is 32% after the data cleaning described above. Because firms of all sizes receive equal weights in our estimation, scaling up all the firms' labor expenditure by the same proportion overestimates labor expenditure, and our estimators reflect the unweighted 32% average more than the weighted 20% average. In other words, such uniform scale-up introduces new biases.

It is possible that using unscaled labor shares underestimates production elasticities of labor and consequently creates a downward bias. If firms in the same industry underreport a same proportion of their labor income, the underreporting does not bias our estimated distortions because the downward biases in estimated production elasticities cancel out the missing proportion in labor income. Nonetheless, the downward biases in labor production elasticities can cause a downward bias in the impact from removing between-nest distortions, and therefore give a conservative estimation of the impact.

2.1 An industry is not necessarily a market

Intuitively, the finest industry categories provided in manufacturing data are usually too broad to be considered as one market. Two examples of industry categories in our data are plastic shoes manufacturing, and automobile and aircraft tire manufacturing. The tire market for automobiles should be different from the one for aircraft. The market for generic plastic shoes may also differ from the one for special purposes. Although it is hard to directly observe whether an industry contains multiple markets, we can infer evidence about it from the correlations between market shares and revenue-cost ratios within each industries. The theory behind is that the correlation of markups and productivity within a market should reflect the level and the type of pass-throughs in the market. Positive correlations indicate incomplete pass-throughs, negative correlations reflect more-than-complete pass-throughs, and zero correlations suggest complete pass-throughs. It is generally believed that more-than-complete pass-throughs are rare in practice. If we use the revenue-cost ratios as a measure for markups and market shares as an approximation for productivity, a fair share of industries with negative correlations between market shares and revenue-cost ratios suggests that industries may contain multiple markets. In our data, 8% of the industries exhibit negative correlations as shown by the blue part of Figure 1. We consider this 8% as too high to be explained by more-than-complete pass-throughs, and allow multiple markets to exist inside an industry.

Figure 1: Correlations between revenue-cost ratios and market shares at the industry level



3 Model

We use a standard monopolistic competition model where firms not only differ on productivity but also on demand elasticities, production elasticities, and returns to scale. It is an extension of HK, relaxing their restrictions on demand elasticities and production elasticities.

We follow HK to characterize demand side structure using a final good producer who combines products Y_s from S industries using a Cobb-Douglas aggregator and sells the final good in a perfectly competitive market to a representative consumer:

$$\mathcal{Y} = \prod_{s=1}^{S} Y_s^{\beta_s} \text{ , where } \sum_{s=1}^{S} \beta_s = 1$$

Cost minimization gives:

$$P_s Y_s = \beta_s \mathcal{P} \mathcal{Y}$$

Y_s is the compound product of industry s and P_s is its price. $\mathcal{P} = \prod_{s=1}^{S} \left(\frac{P_s}{\beta_s} \right)^{1/\beta_s}$ is the price of the final good and is set to 1 because the final good is a numeraire. Unlike HK, each industry faces its own demand elasticities and within each industry there is a possibility of having a high-demand-elasticity nest and a low-demand-elasticity nest. When there are two nests, $\bar{s}$ and $\underline{s}$, inside an industry, the industry compound product is written as:

$$Y_s = Y_{\bar{s}}^{\gamma_s} Y_{\underline{s}}^{1-\gamma_s}$$

and

$$Y_g = \left(\sum_{i \in g} Y_i^{\frac{\epsilon_g - 1}{\epsilon_g}} \right)^{\frac{\epsilon_g}{\epsilon_g - 1}} \text{, where } \begin{cases} g \in \{\bar{s}, \underline{s}\} \text{ , if two nests inside s} \\ g = s \text{ , if one nest inside s} \end{cases}$$

Firms are denoted by index i. Firms in $\bar{s}$ face higher demand elasticities, thus lower markups, than firms in $\underline{s}$, i.e. $\epsilon_{\bar{s}} > \epsilon_{\underline{s}}$. An intuition behind high-demand-elasticity versus low-demand-elasticity nests is established brands versus lesser-known brands. Alternatively, it can also be firms capable of producing products for special purposes versus those producing generic ones. In the rest of this paper, a nest refers to the nest $\bar{s}$ or $\underline{s}$ when there are two nests inside s or to industry s itself when there is only one nest; an industry always refer to an industry s. Loosely speaking, our nests are comparable to industries in HK when deriving most results in this section.

Our production function is a Cobb-Douglas function with non-constant returns to scale:

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

Unlike HK, $\alpha_s^k + \alpha_s^L$ does not have to be 1. Notice, α_s^K and α_s^L change with s but not within s. In other words, α_s^K and α_s^L are the same for firms from $\bar{s}$ and $\underline{s}$.

We denote distortions that change the marginal revenues of capital and labor for firm i as τ_i^k and τ_i^L. Firms with limited access to capital have larger τ_i^K, while those enjoying cheap financial credits have lower τ_i^K; similarly, firms that use permanent labor contracts, such as state-owned enterprises, have higher τ_i^L, as they normally cannot reduce labor inputs easily. τ_i^K and τ_i^L can be negative when firms' rental cost is below market rental rates, but τ_i^K and τ_i^L must be greater than -1 so that labor and capital expenditure is positive. The market rental rates of capital and labor are R and w. The aggregate supply of capital and labor is fixed. τ_i^K and τ_i^L include non-monetary obstacles such as lack of access to the capital and labor market as well as as monetary ones such as subsidies or taxes. The part of τ_i^K and τ_i^L which are taxes and subsidies are given back to or taken away from the representative consumer as a lump-sum transfer.

Firms face idiosyncratic cost shocks δ_i, which are realized after choosing their prices, labor, and capital. The cost shocks are proportional to but not part of capital and labor expenditure. They are the part of the production costs which is not foreseen by firms when setting prices. In the empirical exercises, they help explain the excessive variation in the aggregate firm-level costs, which include the expenditure on materials and energy in addition to capital and labor. Therefore, the cost shocks can be understood as unexpected fluctuations in materials and energy costs. When making production choices, firms maximize expected profits:

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

which gives the standard pricing rule where prices are proportional to the expected marginal cost:[4]

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

Derivation of optimal prices and other key variables in this section are provided in the Appendix B. Optimal prices are a function of production because firms are not constant returns to scale, and marginal costs depend on production. After the cost shocks, firms carry out their production as planned and earn their profits:

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

No entry or exit is allowed. The markups predicted by our model are the ratios between the optimal prices and the realized marginal cost, denoted as $\mu_i + 1$:

$$\mu_i + 1 = \frac{\epsilon_g}{\epsilon_g - 1}\frac{\mathbb{E}[e^{\delta_i}]}{e^{\delta_i}}$$

There are two sources of variations in firms' markups: different demand elasticities and idiosyncratic cost shocks. The latter enlarges the range of theoretical markups from greater than 1 in a standard CES model to greater than 0, and allows large variations in markups, while remaining tractable compared to an arbitrary nested-CES model that may need to treat each firm as a nest. It is an important feature of our model because our data suggests a large variation in markups, which is difficult to fit with a tractable standard nested-CES model. Besides, 15% of the firms have negative profits implying markups less than 1. These negative profits cannot be explained by a static model without cost shocks. The first part of $\mu_i + 1$ is the same as the one in a standard CES model, and has to be larger than 1 because $\epsilon_g > 1$. The second part can take any positive values, and can be arbitrarily close to 0, if a cost shock δ_i goes to positive infinity. Firms with highly unfavorable cost shocks may have markups below 1 and earn negative profits.

The gaps between observed income shares for capital and labor and the ratio of production elasticities and markups are determined by the input distortions via the first-order conditions

[4] For the existence of optimal pricing rule, we assume $\epsilon_g > 1$

of firms' profit maximization:[5]

$$\frac{wL_i\mathbb{E}[e^{\delta_i}]}{P_iY_i(\epsilon_g-1)/\epsilon_g} = \frac{\alpha_s^L}{1+\tau_i^L} \tag{1}$$

$$\frac{RK_i\mathbb{E}[e^{\delta_i}]}{P_iY_i(\epsilon_g-1)/\epsilon_g} = \frac{\alpha_s^K}{1+\tau_i^K} \tag{2}$$

The demand side is very simple. A representative consumer owns all the capital and labor. The payment for cost shocks also goes to the consumer as a lump sum. The economy reaches a general equilibrium where all the firms and the representative consumer solve their own optimization problems and market rental prices of labor and capital clear markets.

3.1 Aggregate labor and capital shares

Using the first order conditions of firms' profit maximization, i.e. Equation (1) and (2), and the structure of demand, the aggregate labor share and capital share are:

$$\begin{aligned}\frac{wL}{PY} &= \sum_g \frac{P_gY_g}{PY}\sum_{i\in g}\frac{wL_i}{P_iY_i}\frac{P_iY_i}{P_gY_g} \\ &= \sum_g \underbrace{\beta_g\alpha_g^L\frac{\epsilon_g-1}{\epsilon_g\mathbb{E}_g[e^{\delta_i}]}}_{\text{between-nest distortions}} \cdot \underbrace{\frac{1}{1+\bar{\tau}_i^L}}_{\text{within-nest distortions}} \end{aligned} \tag{3}$$

$$\begin{aligned}\frac{RK}{PY} &= \sum_g \frac{P_gY_g}{PY}\sum_{i\in g}\frac{RK_i}{P_iY_i}\frac{P_iY_i}{P_gY_g} \\ &= \sum_g \underbrace{\beta_g\alpha_g^K\frac{\epsilon_g-1}{\epsilon_g\mathbb{E}_g[e^{\delta_i}]}}_{\text{between-nest distortions}} \cdot \underbrace{\frac{1}{1+\bar{\tau}_i^K}}_{\text{within-nest distortions}} \end{aligned} \tag{4}$$

[5] Notice combining these two equations gives the optimal pricing rule: $P_i = \frac{P_iY_i}{Y_i} = \frac{P_iY_i}{A_iK_i^{\alpha_s^K}L_i^{\alpha_s^L}} = \left(\frac{P_iY_i}{K_i}\right)^{\alpha_s^K}\left(\frac{P_iY_i}{L_i}\right)^{\alpha_s^L}\frac{(P_iY_i)^{1-\alpha_s^K-\alpha_s^L}}{A_i}$.

Rearrange the equation gives: $P_i = \left(\frac{P_iY_i}{K_i}\right)^{\frac{\alpha_s^K}{\alpha_s^K+\alpha_s^L}}\left(\frac{P_iY_i}{L_i}\right)^{\frac{\alpha_s^L}{\alpha_s^K+\alpha_s^L}} Y_i^{\frac{1-\alpha_s^K-\alpha_s^L}{\alpha_s^K+\alpha_s^L}}\left(\frac{1}{A_i}\right)^{\frac{1}{\alpha_s^K+\alpha_s^L}} = \frac{\epsilon_g}{\epsilon_g-1}\left(\frac{1}{A_i}\right)^{\frac{1}{\alpha_s^L+\alpha_s^K}} Y_i^{\frac{1-\alpha_s^L-\alpha_s^K}{\alpha_s^L+\alpha_s^K}}\left(\frac{R(1+\tau_i^K)}{\alpha_s^K}\right)^{\frac{\alpha_s^K}{\alpha_s^L+\alpha_s^K}}\left(\frac{w(1+\tau_i^L)}{\alpha_s^L}\right)^{\frac{\alpha_s^L}{\alpha_s^L+\alpha_s^K}}\mathbb{E}[e^{\delta_i}]$.

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

$$\frac{1}{1+\bar{\tau}_g^L} \equiv \sum_{i\in g} \frac{1}{1+\tau_i^L} \frac{\left(\frac{A_i}{(1+\tau_i^L)^{\alpha_g^L}(1+\tau_i^K)^{\alpha_g^K}}\right)^{\frac{\epsilon_g-1}{\epsilon_g-(\epsilon_g-1)(\alpha_g^L+\alpha_g^K)}}}{\sum_{i\in g}\left(\frac{A_i}{(1+\tau_i^L)^{\alpha_g^L}(1+\tau_i^K)^{\alpha_g^K}}\right)^{\frac{\epsilon_g-1}{\epsilon_g-(\epsilon_g-1)(\alpha_g^L+\alpha_g^K)}}}$$

$$\frac{1}{1+\bar{\tau}_g^K} \equiv \sum_{i\in g} \frac{1}{1+\tau_i^K} \frac{\left(\frac{A_i}{(1+\tau_i^L)^{\alpha_g^L}(1+\tau_i^K)^{\alpha_g^K}}\right)^{\frac{\epsilon_g-1}{\epsilon_g-(\epsilon_g-1)(\alpha_g^L+\alpha_g^K)}}}{\sum_{i\in g}\left(\frac{A_i}{(1+\tau_i^L)^{\alpha_g^L}(1+\tau_i^K)^{\alpha_g^K}}\right)^{\frac{\epsilon_g-1}{\epsilon_g-(\epsilon_g-1)(\alpha_g^L+\alpha_g^K)}}}$$

Derivation of these equations are in the Appendix B.3. We decompose changes in the aggregate labor and capital shares into the part due to within-nest distortions and between-nest distortions. The within-nest distortions are $\bar{\tau}_g^L$ and $\bar{\tau}_g^K$, which affect how resources are reallocated within nests. Between nest-distortions describe how resources are reallocated across nests, and how cross-nest reallocation affect the aggregate shares.

The within-nest distortions are determined by the joint distribution of firm-level distortions and productivity. Parameters such as returns to scale and demand elasticities also affect the nest-level distortions because substitutability and the economies of scale matter when reallocating resources across heterogeneous firms. Intuitively, if all the firms inside an nest have the same input distortions, $\bar{\tau}_g^L$ and $\bar{\tau}_g^K$ are the common distortions which make the within-nest distortions, i.e the second term in Equation (3) and Equation (4), in this case equal to those where firms differ on their input distortions. In other word, $\bar{\tau}_g^L$ and $\bar{\tau}_g^K$ indicate the size of the nest-level distortions.

Removing all the input distortions will collapse the within-nest distortions to 1, i.e. the distortions no longer affect the factor shares. Using the notation of $\bar{\tau}_g^L$ and $\bar{\tau}_g^K$, the change in the labor and capital shares from removing input distortions can be expressed as:

$$\frac{w^*L}{P^*Y^*} - \frac{wL}{PY} = \sum_g \beta_g \alpha_g^L \frac{\epsilon_g - 1}{\epsilon_g \mathbb{E}_g[e^{\delta_i}]}\left(1 - \frac{1}{1+\bar{\tau}_g^L}\right)$$

$$\frac{R^*K}{P^*Y^*} - \frac{RK}{PY} = \sum_g \beta_g \alpha_g^K \frac{\epsilon_g - 1}{\epsilon_g \mathbb{E}_g[e^{\delta_i}]}\left(1 - \frac{1}{1+\bar{\tau}_g^K}\right)$$

where $\frac{w^*L}{P^*Y^*}$ and $\frac{R^*K}{P^*Y^*}$ are the aggregate labor and capital shares without input distortions. There is no $*$ for L and K because aggregate capital, K, and aggregate labor, L, are assumed to be constant. The decomposition is:

$$\begin{aligned}
\frac{w^*L}{P^*Y^*} - \frac{wL}{PY} =& \underbrace{\alpha^L \frac{\epsilon - 1}{\epsilon \mathbb{E}[e^{\delta_i}]} \frac{1}{N} \sum_g \left(1 - \frac{1}{1 + \bar{\tau}^L_g}\right)}_{\text{due to within-nest distortions}} \\
&+ \underbrace{\sum_g \left(\beta_g \alpha^L_g \frac{\epsilon_g - 1}{\epsilon_g \mathbb{E}_g[e^{\delta_i}]} - \frac{1}{N} \alpha^L \frac{\epsilon - 1}{\epsilon \mathbb{E}[e^{\delta_i}]}\right) \left(1 - \frac{1}{1 + \bar{\tau}^L_g}\right)}_{\text{due to between-nest distortions}} \\
\frac{R^*K}{P^*Y^*} - \frac{RK}{PY} =& \underbrace{\alpha^K \frac{\epsilon - 1}{\epsilon \mathbb{E}[e^{\delta_i}]} \frac{1}{N} \sum_g \left(1 - \frac{1}{1 + \bar{\tau}^K_g}\right)}_{\text{due to within-nest distortions}} \\
&+ \underbrace{\sum_g \left(\beta_g \alpha^K_g \frac{\epsilon_g - 1}{\epsilon_g \mathbb{E}_g[e^{\delta_i}]} - \frac{1}{N} \alpha^K \frac{\epsilon - 1}{\epsilon \mathbb{E}[e^{\delta_i}]}\right) \left(1 - \frac{1}{1 + \bar{\tau}^K_g}\right)}_{\text{due to between-nest distortions}}
\end{aligned}$$

where $\alpha^L \equiv \bar{\alpha}^L_g$, $\alpha^K \equiv \bar{\alpha}^K_g$, $\frac{\epsilon-1}{\epsilon\mathbb{E}[e^{\delta_i}]} \equiv \overline{\frac{\epsilon_g-1}{\epsilon_g\mathbb{E}_g[e^{\delta_i}]}}$. The upper-bar denotes the sample average. Holding the first part constant, we can see the joint distribution of $1 - \frac{1}{1+\bar{\tau}^m_g}$ and $\beta_g \alpha^m_g \frac{\epsilon_g-1}{\epsilon_g\mathbb{E}_g[e^{\delta_i}]} - \frac{1}{N}\alpha^m \frac{\epsilon-1}{\epsilon\mathbb{E}[e^{\delta_i}]}$ for $m \in \{L, K\}$ decides the value and the sign of the second term, the impact of between-nest distortions. When the second part dominants, the predicted labor and capital share changes are mainly driven by the structure of the economy instead of the size of the distortions. This means, even if the size of input distortions predicts that the income shares should increase, if all the resources are reallocated to smaller markets (small β_g) with lower demand elasticities (lower $\frac{\epsilon_g-1}{\epsilon_g\mathbb{E}_g[e^{\delta_i}]}$) and a lower demand for those resources (lower α^L_g or α^K_g), the net change on income shares will be close to zero or even negative.

4 Identification

In HK, α^K_g and α^L_g are calibrated using American data. 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^K_g and α^L_g 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^K_g and α^L_g 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 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 market level and the firm level as:

$$\begin{aligned}
\text{TFPR}_i \equiv & P_i A_i = \frac{P_i Y_i}{K_i^{\alpha_s^K} L_i^{\alpha_s^L}} \\
\text{TFPR}_g \equiv & \frac{\sum_{i \in g} P_i Y_i}{K_g^{\alpha_s^K} L_g^{\alpha_s^L}}
\end{aligned}$$

Then the market-level TFP is:

$$\begin{aligned}
\text{TFP}_g &= \left[\sum_{i \in g} \left(A_i \frac{\text{TFPR}_g}{\text{TFPR}_i} \right)^{\epsilon_g - 1} \right]^{\frac{1}{\epsilon_g - 1}} \\
&= \left[\sum_{i \in g} \left(A_i \left(\underbrace{\frac{1}{1+\tau_i^L} \frac{1}{\sum_{i \in g} \frac{P_i Y_i}{P_g Y_g} \frac{1}{1+\tau_i^L}}}_{\text{labor distortions}} \right)^{\alpha_s^L} \left(\underbrace{\frac{1}{1+\tau_i^K} \frac{1}{\sum_{i \in g} \frac{P_i Y_i}{P_g Y_g} \frac{1}{1+\tau_i^K}}}_{\text{capital distortions}} \right)^{\alpha_s^K} \left(\frac{P_i Y_i}{P_g Y_g} \right)^{\alpha_s^L + \alpha_s^K - 1} \right)^{\epsilon_g - 1} \right]^{\frac{1}{\epsilon_g - 1}} \\
&= \left[\sum_{i \in g} \left(A_i \left(\underbrace{\frac{wL_i/(P_i Y_i)}{wL_g/(P_g Y_g)}}_{\text{labor distortions}} \right)^{\alpha_s^L} \left(\underbrace{\frac{RK_i/(P_i Y_i)}{RK_g/(P_g Y_g)}}_{\text{capital distortions}} \right)^{\alpha_s^K} \left(\frac{P_i Y_i}{P_g Y_g} \right)^{\alpha_s^L + \alpha_s^K - 1} \right)^{\epsilon_g - 1} \right]^{\frac{1}{\epsilon_g - 1}}
\end{aligned}$$

Here, $wL_g = \sum_{i \in g} wL_i$ is the labor expenditure of the nest g. $RK_g = \sum_{i \in g} Rk_i$ is this nest's capital expenditure and $P_g Y_g = \sum_{i \in g} P_i Y_i$ is its total value added. The distortions enter the nest-level TFP_g as a ratio between firm-level distortions and a weighted average of the firm-level distortions in a market, as shown in the second line of the above equation for TFP_g. Therefore, the level of each firm's distortions cancels out and only the dispersion remains. Plug in the first-order conditions for firms, we can see in the third line that this dispersion in distortions are dispersion in labor shares and capital shares.

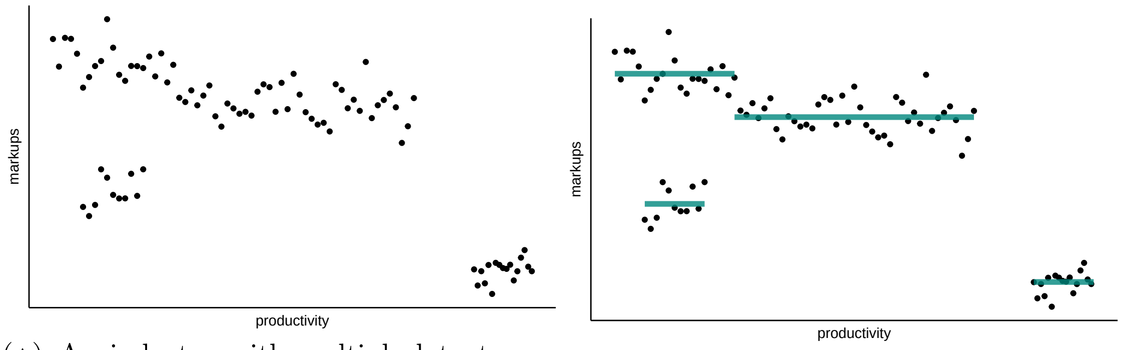
However, both the level and the dispersion of distortions matter for labor shares as we can see in Equation (3) and Equation (4). When increase the calibrated α_g^K and α_g^L from a fairly low level, the sign of labor and capital share changes may flip from negative to positive because estimated distortions will increase from mostly negative to mostly positive. In fact, since American firms are typically understood as more capita-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.

4.1 Identify the latent market structure

We define each market as a cluster of firms that face the same demand elasticities and production elasticities but can differ on productivity and input distortions. We do not directly observe clusters and need to infer the market structure using observed firm characteristics. In this project, we assume that firms belonging to the same market always belong to the same industry. We then use revenue-cost ratios to measure firm-specific markups and then infer the cluster structure within each industry using the distribution of revenue-cost ratios. Although we only use industry categories and revenue-cost ratios to infer clusters, the data contains more firm characteristics that can be used to improve our clustering inference or even relax the assumption that firms from the same market have to be in the same industry. We leave a refined clustering inference to future works. Another simplification we impose is each industry can have at most two clusters. This simplifies the hypothesis test procedure on the number of clusters in each industry. After assuming cost shocks in each cluster follow a normal distribution with mean 0, firms' cluster identity and each cluster's demand elasticities are then estimated by maximizing the mixture distribution of revenue-cost ratios. More details are provided in the Appendix.

Figure 2 demonstrate how to infer latent markets using a hypothetical industry. In this hypothetical industry, there are 5 latent nests and each has its own demand elasticities. We use latent nests to cluster firms with similar markups together while markups are measured using revenue-cost ratios. Figure 2a plots its firms' markups and productivity without knowing the latent nests. Figure 2b demonstrate how firms with similar markups are clustered together to form nests.

Figure 2: A hypothetical industry: markups vs. productivity



(a) An industry with multiple latent markets

(b) Fitted by latent nests

4.2 Identify 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 market rental price of capital. We follow HK to assume $R = 0.1$.

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

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

We treat the left-hand side of the equations as known because ϵ_g and $\mathbb{E}[e^{\delta_i}]$ are 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.

$$\log(\tau_i^K+1) \sim \begin{cases} 2\kappa_s^K\mathcal{N}(0,\sigma_{s,+}^K) \text{ , if } \tau_i^K > 0 \\ (2-2\kappa_s^K)\mathcal{N}(0,\sigma_{s,-}^K) \text{ , if } \tau_i^K < 0 \end{cases}$$

$$\log(\tau_i^L+1) \sim \begin{cases} 2\kappa_s^L\mathcal{N}(0,\sigma_{s,+}^L) \text{ , if } \tau_i^L > 0 \\ (2-2\kappa_s^L)\mathcal{N}(0,\sigma_{s,-}^L) \text{ , if } \tau_i^L < 0 \end{cases}$$

Production elasticities α_s^K and α_s^L are estimated by maximizing the likelihood of observing the labor shares and capital shares. Input distortions are then calculated using the estimated production elasticities. More technical details are provided in the Appendix.

5 Results

5.1 Market structure

Table 2 reports the distribution of the number of firms within industries for industries that are estimated as having one nest and as having two nets separately. Whether an industry is estimated as having one nest shows the estimated latent market structure. Although we assume each industry can have at most two latent nests, our parameters are still informative about whether an industry should be considered as having multiple nests. Whenever the industry is estimated as having two nests, this should be interpreted as the industry should have multiple nests. We can interpret the results this way because the number of nests inside an industry is identified via hypothesis test. We first test having only one nest against having more than one nest. Under our assumption of having at most two nests, we only need one hypothesis test for each industry. When we reject the null of having only one test, then we estimate the industry as having two nests. If we remove the assumption of having at most two nets, then we need to continue the test of having two nests against having more than two nests. We will stop either when reaching the maximum number of nests allowed or when we do not reject the null hypothesis. Therefore, Table 2 shows us about 90% of the industries (462 out of 523) are better estimated as having more than one nest. Those that should be estimated as having more than one nest usuallly contain more firms.

Table 2: Distribution of firm counts for industries

two types	N	Mean	Min	Pctl(25)	Median	Pctl(75)	Max
No	61	23	2	6	15	27	237
Yes	462	494	12	118	256	544.500	$9,947$

Table 3 reports the other estimated parameters. The one related to the latent market structure is ex-ante $P_g[\bar{s}]$. It is the probability that a firm belongs to the lower-markup nest in an industry. The bottom quartile of ex-ante $P_g[\bar{s}]$ is about 60%. This suggests Chinese firms in 2005 are mostly in the lower-markup nest because This coincides the idea that Chinese manufacturing firms are mostly lower-value firms on the value chain.

5.2 Demand elasticities

σ_g in Table 3 is the estimated demand elasticicities. There is a larger variation with the top 10 percentile being about 4 times larger than the bottom 10 percentile. There is also a large difference in demand elasticities within industries. Figure 3 plots the demand elasticities of

Table 3: Market-level summary statistics of estimates

	N	Mean	St. Dev.	Pctl(10)	Pctl(25)	Median	Pctl(75)	Pctl(90)
$\frac{\sigma_g}{\sigma_g - 1}$	985	1.30	0.25	1.11	1.14	1.22	1.39	1.57
σ_g	985	6.33	3.64	2.77	3.59	5.45	8.32	10.48
$\mathbb{E}_g[e^{\delta_i}]$	985	1.01	0.02	1	1	1.01	1.02	1.03
ex-ante $P_g[\bar{s}]$	928	0.66	0.22	0.27	0.59	0.73	0.82	0.88
α_g^K	523	0.16	0.17	0.04	0.06	0.09	0.19	0.36
α_g^L	523	0.39	0.23	0.13	0.21	0.33	0.57	0.76
RTS	523	0.55	0.31	0.22	0.32	0.48	0.75	0.95

the nests that are in the same industry on the two axis. If the demand elasticities are similar inside industries, points should be close to the dashed line which is the 45 degree line.

Our estimated demand elasticities are in line with those found in literature. There is little markups estimation for Chinese firms in literature, so we check our estimates by comparing to American markups estimated by existing studies. 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.

Markups indicate how much market power a firm has and how much concentration there is in a market. Therefore, our markups should be positive correlated with indicators about market concentration. Figure 4a shows how expected markups at the industry level correlated with the number of firms in an industry. More firms usually indicate less concentration and hence lower markups. Figure 4a confirms this correlation. It also shows when an industry type has a lot of SOEs relative to the total number of firms in the industry, it is more likely to deviate from this negative linear correlation. A similar pattern remains when we look at the market share of SOEs in an industry type (Figure 4b). Besides, we compare our industry-level average markups to the Herfindahl indexes in Figure 5 and find that our markups increase with the Herfindahl indexes.

Figure 3: Differences of demand elasticities within industries

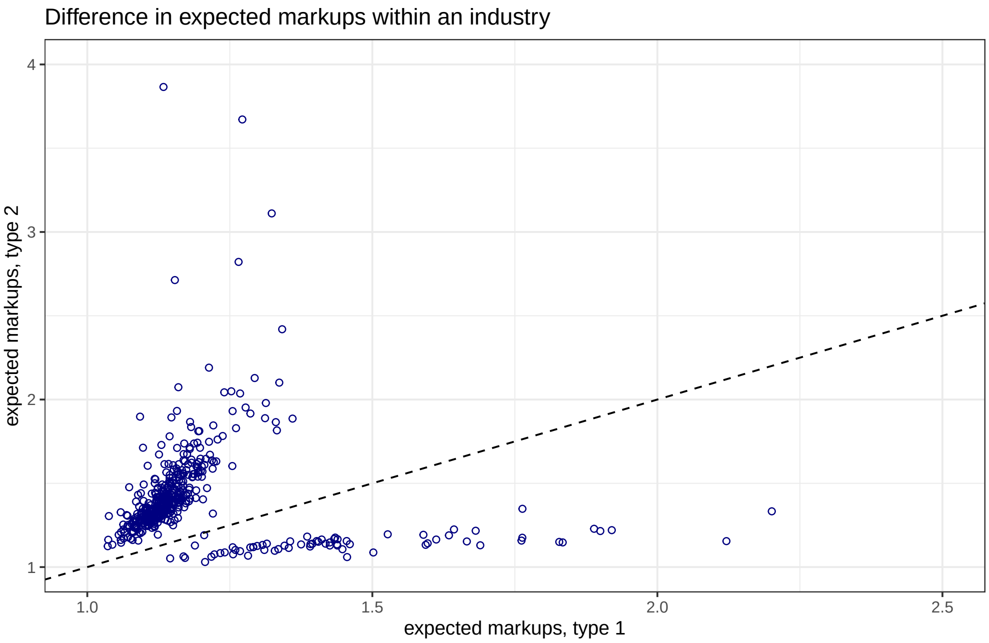
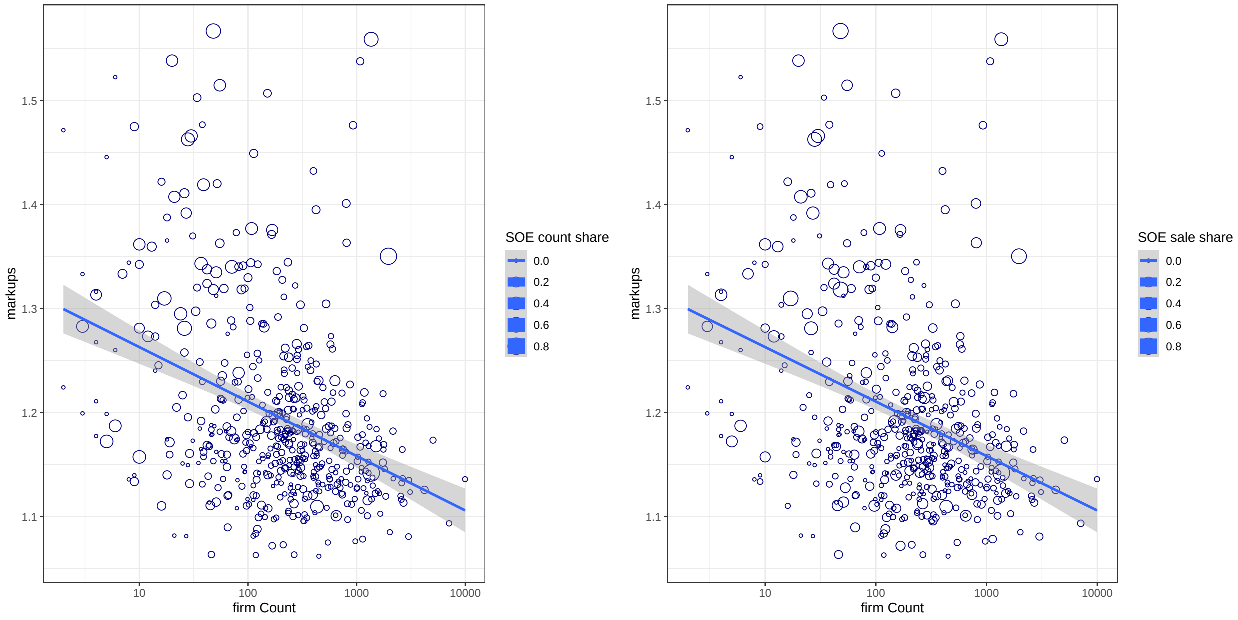


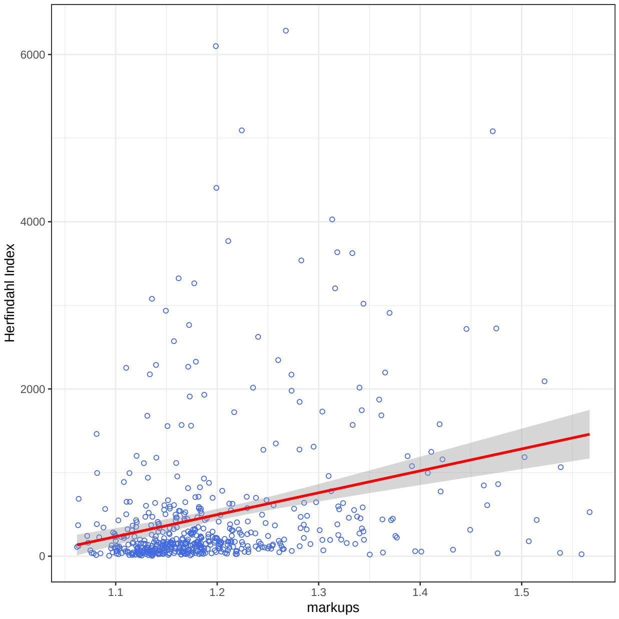
Figure 4: Relation between industry-level firm counts, expected markups, and shares of SOEs



(a) Shares of SOEs as firm-count shares (b) Shares of SOEs as sales shares of SOEs

Notes: industry-level expected markups are cost-weighted average

Figure 5: Relation between market-level expected markups, and the Herfindahl index



Notes: market-level expected markups are cost-weighted average

5.3 Changes in the aggregate labor and capital shares

Removing input distortions increases the aggregate labor share by 7% and decreases the aggregate capital share 1% (Table 7a). We use our decomposition to explain what drives these changes and why the labor and capital share behave differently.

Table 4b shows that removing within-nest distortions increases the labor share by 2.23% and and reduces the capital share by 17%. This indicates that 5.21% increase in the labor share is due to between-nest distortions. For capital, between-nest distortions actually raises capital shares by 16.59% which cancels out most of the decrease in the capital share. The decomposition is summarized in Table 5. The labor share increases when removing within-nest distortions because the input distortions for labor are mostly positive (the top figure in Figure 6) suggesting too little demand for labor due to the distortions. For capital, its input distortions are mostly negative (the top figure in Figure 6) creating too high a demand for capital. When we remove the input distortions, within-nest distortions consequently raise the labor share and lower the capital share.

The between-nest distortions raise both the labor share and capital share because nests with higher distortions have higher weights (the correlations in the first row of Table 6 are positive). Further decomposition tells us differences in technology (α_g^L and α_g^K) is the main reason and heterogeneous demand elasticities play a minor role. Variation in technology explains almost half of the increase in the labor share (3.79% out of 45.21%) and about 80%

(A) Benchmark

	observed	predicted	change
L	19.76	27.2	7.44
K	11.86	10.77	-1.09
L+K	31.62	37.97	6.35

(B) Nest-level τ fixed & homo α, β, and σ

	observed	predicted	change
L	24.69	26.92	2.23
K	44.57	26.92	-17.65
L+K	69.26	53.85	-15.42

(C) Nest-level τ fixed & homo ϵ

	observed	predicted	change
L	18.54	25.43	6.88
K	11.29	10.26	-1.03
L+K	29.83	35.69	5.85

(D) Nest-level τ fixed & homo α

	observed	predicted	change
L	25.55	29.20	3.65
K	42.63	29.20	-13.43
L+K	68.18	58.40	-9.78

of the increase in the capital share (12.34% out of 16.59%) due to between nest distortions. Differences in demand elasticities contribute only about 0.5 percentage point.

TABLE 5: Decomposition of the labor and capital income share changes (%)

	within-nest	between-nest	total
L	2.23	5.21	7.44
K	-17.65	16.59	-1.09
L+K	-15.42	21.77	6.35

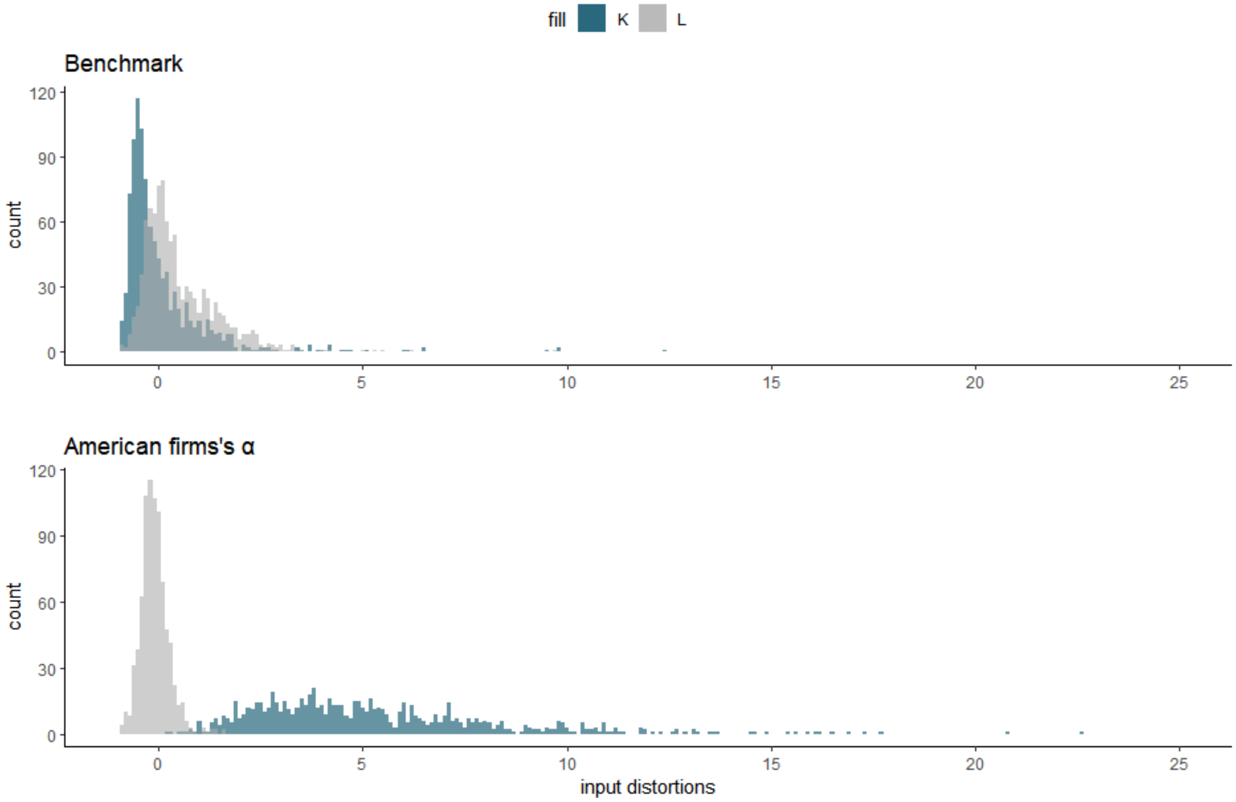
TABLE 6: Correlations behind the cross-nest reallocation

correlations	L	K
$cor(\bar{\tau}_g, \beta_g\alpha_g\frac{\epsilon_g-1}{\epsilon_g\mathbb{E}_g[e^{\delta_i}]})$	0.05	0.04
$cor(\bar{\tau}_g, \beta_g\alpha_g)$	0.05	0.04
$cor(\epsilon_g/(\epsilon_g-1), \text{input}^*_g/\text{input}_g)$	-0.09	0.02
$cor(\alpha^L_g, \text{input}^*_g/\text{input}_g)$	0.68	0.62

Table 6 shows that technology plays a big role because it is strongly correlated with reallocation of inputs across nests which results from the strong correlation with within-nest distortions. Demand elasticities is weakly correlated with reallocation across nests and has an ignorable impact of the correlation with within-nest distortions.

Using American firms production elasticities flip the sign of factor income share changes. The labor share will in this case decrease by almost 3% while the capital share will increase

Figure 6: The distribution of within-nest distortions



by about 31% (Tale 7b).[6] This is because American firms' are more capital intensive and less labor intensive. This technology differences bias the labor distortions downwards and the capital distortions upwards (the bottom figure in Figure 6) so that labor appears to be overused while capital appears to be heavily underused.

TABLE 7: Labor and capital income share using different α (%)

(A) Benchmark

	observed	predicted	change
L	19.76	27.2	7.44
K	11.86	10.77	-1.09
L+K	31.62	37.97	6.35

(B) American α (HK)

	observed	predicted	change
L	50	47.11	-2.89
K	6.73	38.11	31.38
L+K	56.73	85.21	28.48

5.4 Input distortions and ownership types

Our estimated distortions suggest that SOEs are more likely to use more capital and labor compared to domestic private firms. Although there are some domestic private firms facing lower distortions, i.e. using relatively more capital and labor than most SOEs, and there are SOEs facing higher distortions, i.e. using relatively less capital and labor than most domestic private firms, the distortion distribution of domestic private firms first-order stochastically dominates that of SOEs (Table 8). The large variation within both ownership types may result from a fuzzy connection between the ownership labels and their business environment. Some domestic private firms may still enjoy favorite 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 government if they are deemed as strategically important by the government. 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 8 keeps only those observations where the two criteria agree.

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

Table 8: Estimated distortions for different firm types

	firm type	N	Mean	Min	Pctl(25)	Median	Pctl(75)	Max
τ_i^K	domestic priv	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 priv	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 factor income shares, i.e. the income shares of capital and labor in a rich economy structure with heterogeneous productivity, technology, input distortions, and demand elasticities. Under a nested CES demand, we decompose the impact into changes due to within-nest distortions and between-nest distortions. We then use firm-level survey data to estimate industry-specific production elasticities, nest-specific demand elasticities for latent nests, and firm-specific input distortions. Our findings confirm that changes in the aggregate factor shares may veil large variation at the nests level. Depending on the structure of economy, the between-nest distortions may amplify or offset the impact of within-nest distortions. Our results imply that Chinese firms in general use too much capital and too little labor, though it varies across firms with different ownership types.

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

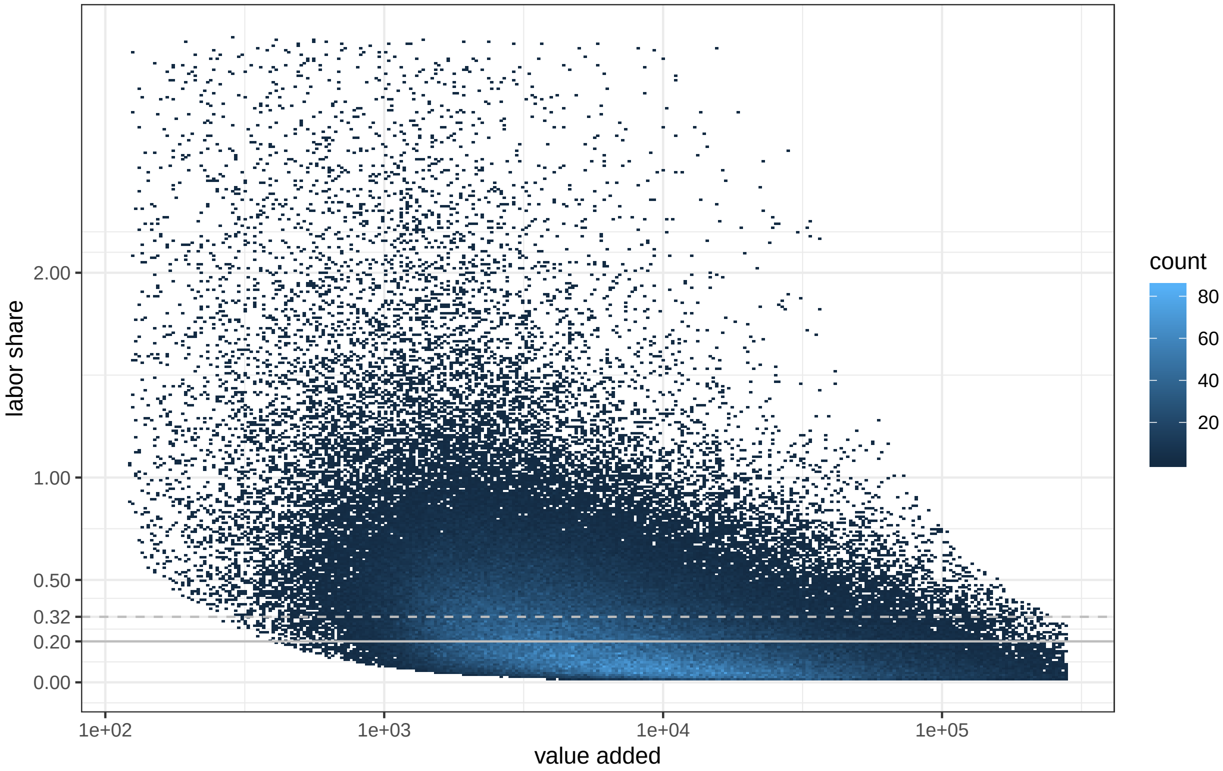
We drop unreasonable observations accounting-wise including observations with negative value added, negative wage expenditure, negative capital, negative total asset, negative account receivable, negative total debt, negative long-term debt, negative account payable, negative export, negative sales, and negative cost. We also drop observations whose account receivable is larger than total asset, total debt larger than total asset, account payable larger than liquid debt, and profits larger than sales. If a firm's cost is missing but its sales and profits are observed, then its cost is 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.

TABLE 9: 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

Figure 7 plots the unscaled wage share against firms' value added. The labor shares of firms with value added above $1e5$ are mostly below the aggregate labor share 0.19 while smaller firms are more spread out. Firms are more concentrated in the domain of low labor share as value added increases.

Figure 7: Joint distribution of labor income share and firm sizes (2005)



B Derivation

We first derive the formula for predicted TFP gains which is an extended version of HK that does not impose constant returns to scale. The results derived in predicted TFP gains are than used to estimate changes in aggregate labor and capital shares.

B.1 Predicted TFP gains

We first show how to derive the optimal prices. The optimal prices are always the expected marginal cost times $\epsilon_g/(\epsilon_g - 1)$. For some given Y_i, firms' profits maximization problem can be formulated as, :

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

Expected marginal cost is the Lagrange multiplier of its Lagrange function

$$\min_{K_i,L_i}(R(1+\tau_i^K)K_i + w(1+\tau_i^L))\mathbb{E}[e^{\delta_i}] - \lambda(A_i K_i^{\alpha_s^K} L_i^{\alpha_s^L} - Y_i)$$

Solving it gives expected marginal cost:

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

and optimal prices:

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

The type-level TFP as a weighted sum of firm-level TFP is the same as the one in HK because the expression only requires the type-level aggregator to be CES:

$$\begin{aligned}
\text{TFP}_g =& \text{TFPR}_g \cdot \frac{1}{P_g} \\
=& \text{TFPR}_g \cdot \left(\sum_{i\in g} P_i^{1-\epsilon_g}\right)^{1/(\epsilon_g-1))} \\
=& \text{TFPR}_g \cdot \left(\sum_{i\in g} \left(\frac{A_i}{\text{TFPR}_i}\right)^{\epsilon_g-1}\right)^{1/(\epsilon_g-1))} \\
=& \left(\sum_{i\in g} \left(A_i \cdot \frac{\text{TFPR}_g}{\text{TFPR}_i}\right)^{\epsilon_g-1}\right)^{\frac{1}{\epsilon_g-1}}
\end{aligned}$$

From the definition of TFPR:

$$\begin{aligned}
\text{TFPR}_g =& \left(\frac{P_g Y_g}{K_g}\right)^{\alpha_s^K} \left(\frac{P_g Y_g}{L_g}\right)^{\alpha_s^L} (P_g Y_g)^{1-\alpha_s^K-\alpha_s^L} \\
\text{TFPR}_i =& \left(\frac{P_i Y_i}{K_i}\right)^{\alpha_s^K} \left(\frac{P_i Y_i}{L_i}\right)^{\alpha_s^L} (P_i Y_i)^{1-\alpha_s^K-\alpha_s^L}
\end{aligned}$$

Firms' profit maximization also gives:

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

Plug these into TFPR_i and TFPR_g:

$$\begin{aligned}
\text{TFPR}_i &= \left(\frac{\epsilon_g - 1}{\epsilon_g} \cdot \frac{\alpha_s^K}{(1+\tau_i^K)R}\right)^{-\alpha_s^K} \left(\frac{\epsilon_g - 1}{\epsilon_g} \cdot \frac{\alpha_s^L}{(1+\tau_i^L)w}\right)^{-\alpha_s^L} \cdot (P_iY_i)^{1-\alpha_s^K-\alpha_s^L} \\
&= \underbrace{(1+\tau_i^K)^{\alpha_s^K}(1+\tau_i^L)^{\alpha_L}\left(\frac{R}{\alpha_s^K}\right)^{\alpha_s^K}\left(\frac{w}{\alpha_s^L}\right)^{\alpha_s^L}\left(\frac{\epsilon_g}{\epsilon_g - 1}\right)^{\alpha_s^K+\alpha_s^L}}_{\text{Same as CRS}}(P_iY_i)^{1-\alpha_s^K-\alpha_s^L} \\
\text{TFPR}_g &= \left(\sum_{i\in g}\frac{\epsilon_g - 1}{\epsilon_g} \cdot \frac{\alpha_s^K}{(1+\tau_i^K)R} \cdot \frac{P_iY_i}{P_gY_g}\right)^{-\alpha_s^K} \left(\sum_{i\in g}\frac{\epsilon_g - 1}{\epsilon_g} \cdot \frac{\alpha_s^L}{(1+\tau_i^L)w} \cdot \frac{P_iY_i}{P_gY_g}\right)^{-\alpha_s^L} \cdot (P_gY_g)^{1-\alpha_s^K-\alpha_s^L} \\
&= \underbrace{\left(\sum_{i\in g}\frac{1}{1+\tau_i^K} \cdot \frac{P_iY_i}{P_gY_g}\right)^{-\alpha_s^K}\left(\sum_{i\in g}\frac{1}{1+\tau_i^L} \cdot \frac{P_iY_i}{P_gY_g}\right)^{-\alpha_s^L}\left(\frac{R}{\alpha_s^K}\right)^{\alpha_s^K}\left(\frac{w}{\alpha_s^L}\right)^{\alpha_s^L}\left(\frac{\epsilon_g}{\epsilon_g - 1}\right)^{\alpha_s^K+\alpha_s^L}}_{\text{Same as CRS}} \\
&\quad \cdot (P_gY_g)^{1-\alpha_s^K-\alpha_s^L}
\end{aligned}$$

In the code, we use an equivalent but easier formula because K_g and wL_g are observed. Follow HK, we define:

$$\begin{aligned}
\text{MPK}_g &\equiv \sum_{i\in g}\frac{P_iY_i}{P_gY_g(1+\tau_i^K)} = \frac{\epsilon_g}{\epsilon_g - 1} \cdot \frac{R}{\alpha_s^K} \cdot \frac{K_g}{P_gY_g} \\
\text{MPL}_g &\equiv \sum_{i\in g}\frac{P_iY_i}{P_gY_g(1+\tau_i^L)} = \frac{\epsilon_g}{\epsilon_g - 1} \cdot \frac{w}{\alpha_s^L} \cdot \frac{L_g}{P_gY_g}
\end{aligned}$$

Then we can write:

$$\frac{\text{TFPR}_i}{\text{TFPR}_g} = \underbrace{(1+\tau_i^K)^{\alpha_s^K}(1+\tau_i^L)^{\alpha_s^L}\text{MPK}_g^{\alpha_s^K}\text{MPL}_g^{\alpha_s^L}}_{\text{Same as CRS}}\left(\frac{P_iY_i}{P_gY_g}\right)^{1-\alpha_K-\alpha_L}$$

Set τ_i^K and τ_i^L to 0 gives:

$$\text{TFP}_g^* = \left(\sum_{i \in g} \left(A_i \cdot \left(\frac{P_g^* Y_g^*}{P_i^* Y_i^*} \right)^{1-\alpha_K-\alpha_L} \right)^{\epsilon_g - 1} \right)^{\frac{1}{\epsilon_g - 1}}$$

Since firms inside the same g has the same demand elasticities and expected cost shocks, for any firm i and j from the same g:

$$\frac{Y_i^*}{Y_j^*} = \left(\frac{P_i^*}{P_j^*} \right)^{-\epsilon_g} = \left(\frac{\left(\frac{1}{A_i}\right)^{1/(\alpha_s^L+\alpha_s^K)} \left(Y_i^*\right)^{\frac{1-\alpha_s^L-\alpha_s^K}{\alpha_s^L+\alpha_s^K}}}{\left(\frac{1}{A_j}\right)^{1/(\alpha_s^L+\alpha_s^K)} \left(Y_j^*\right)^{\frac{1-\alpha_s^L-\alpha_s^K}{\alpha_s^L+\alpha_s^K}}} \right)^{-\epsilon_g}$$

The first equation is due to the demand structure and the second equation simply plug in the expression of optimal prices. Solve for Y_i^*/Y_j^*:

$$\frac{Y_i^*}{Y_j^*} = \left(\frac{A_i}{A_j} \right)^{\frac{\epsilon_g}{\epsilon_g + (\alpha_s^L+\alpha_s^K)(1-\epsilon_g)}}$$

From the demand structure, $\frac{P_i Y_i}{P_g Y_g} = \left(\frac{P_i}{P_g}\right)^{1-\epsilon}$ which also means $\frac{Y_i}{Y_g} = \left(\frac{P_i}{P_g}\right)^{-\epsilon}$, thus

$$\frac{P_i Y_i}{P_j Y_j} = \left(\frac{P_i}{P_j} \right)^{1-\epsilon} = \left(\frac{P_i}{P_j} \right)^{-\epsilon \cdot \frac{1-\epsilon}{-\epsilon}} = \left(\frac{Y_i}{Y_j} \right)^{\frac{\epsilon-1}{\epsilon}}$$

Hence,

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

which can be easily written as:

$$\frac{P_i^* Y_i^*}{P_g^* Y_g^*} = \frac{A_i^{\frac{\epsilon_g - 1}{\epsilon_g + (1-\epsilon_g)(\alpha_L+\alpha_K)}}}{\sum_{i \in g} A_i^{\frac{\epsilon_g - 1}{\epsilon_g + (1-\epsilon_g)(\alpha_L+\alpha_K)}}}$$

B.2 TFP gains under homogeneous demand elasticities with known primitives

This sections provides derivations of formulas used when calculating TFP gains in the counterfactual scenario of homogeneous demand elasticities while keep all the other primitives the same as those estimated by our preferred model. This requires we first solve for the equilibrium of the economy given those primitives and than find predicted TFP gains when removing distortions. The formula of type-level TFP and TFPR ratio is the same as the one in Section 3

$$\text{TFP}_g = \left(\sum_{i \in g} \left(A_i \cdot \frac{\text{TFPR}_g}{\text{TFPR}_i} \right)^{\epsilon_g - 1} \right)^{\frac{1}{\epsilon_g - 1}}$$

$$\frac{\text{TFPR}_i}{\text{TFPR}_g} = \underbrace{(1+\tau_i^K)^{\alpha_s^K}(1+\tau_i^L)^{\alpha_s^L} \left(\sum_{i \in g} \frac{P_i Y_i}{P_g Y_g (1+\tau_i^K)} \right)^{\alpha_s^K} \left(\sum_{i \in g} \frac{P_i Y_i}{P_g Y_g (1+\tau_i^L)} \right)^{\alpha_s^L}}_{\text{Same as CRS}} \cdot \left(\frac{P_i Y_i}{P_g Y_g} \right)^{1-\alpha_s^K-\alpha_s^L}$$

Because we have known primitives, A_i, ϵ_g, τ_i^K, τ_i^L, α_s^K, and α_s^L are known. $\frac{P_i Y_i}{P_g Y_g}$ is the equilibrium sales share determined by those primitives and is the only unknown. Using the optimal pricing rule, we can write the price ratio of two firms from the same type 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, $\frac{P_i Y_i}{P_j Y_j} = \left(\frac{P_i}{P_j} \right)^{1-\epsilon}$, therefore

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

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,

$$\frac{P_i Y_i}{P_g Y_g} = \frac{W_i}{\sum_{j \in g} W_j} \tag{5}$$

B.3 Changes in factor income shares

The nested CES demand structure gives the following formula for the aggregate factor income shares:

$$\frac{wL}{PY} = \sum_{g} \frac{P_g Y_g}{PY} \sum_{i \in g} \frac{wL_i}{P_i Y_i} \frac{P_i Y_i}{P_g Y_g}$$
$$\frac{RK}{PY} = \sum_{g} \frac{P_g Y_g}{PY} \sum_{i \in g} \frac{RK_i}{P_i Y_i} \frac{P_i Y_i}{P_g Y_g}$$

We know that $\frac{P_g Y_g}{PY} = \beta_g$ and

$$\frac{wL_i}{P_i Y_i} = \frac{\alpha_s^L}{1+\tau_i^L} \cdot \frac{\epsilon_g - 1}{\epsilon_g \mathbb{E}[e^{\delta_i}]}$$
$$\frac{RK_i}{P_i Y_i} = \frac{\alpha_s^K}{1+\tau_i^K} \cdot \frac{\epsilon_g - 1}{\epsilon_g \mathbb{E}[e^{\delta_i}]}$$

Plug in Equation (5) and use the definition of W_i, we get Equation (3) and Equation (4).

C Identification

C.1 Step 1: calculate 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 2: Identify type-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 types $\{\bar{s}, \underline{s}\}$ and which contains only one type, nor do we observe a firm's type when two types are possible in an industry.

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

$$\log(\mu_i + 1) = \underbrace{\log\left(\frac{\epsilon_g}{\epsilon_g - 1}\right)}_{\text{indicator of type}} + \underbrace{\log\left(\mathbb{E}[e^{\delta_i}]\right) - \delta_i}_{\text{noises}}$$

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

$$\delta_{ig} \sim \mathcal{N}(0, \sigma_g)$$

The distribution variances σ_g differ across types. When there is no type 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 $\underline{s}$ and $\bar{s}$ when there are two types, the distributions of $\log(\mu_i + 1)$ for all those industries follow a mixed normal distribution:

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

w_s is the probability that a firm is type $\underline{s}$ and $1 - w_s$ is the probability that a firm is type $\bar{s}$ 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 $\mu_i + 1$ being type $\underline{s}$ is:

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

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

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

$$\ell\ell(\{\mu_i\}_{i\in s}|w_s, \mu_{\bar{s}}, \mu_{\underline{s}}, \sigma_{\bar{s}}, \sigma_{\underline{s}}) = \sum_{i\in s} \log(w_s f(\mu_i; \mu_{\underline{s}}, \sigma_{\underline{s}}) + (1-w_s) f(\mu_i; \mu_{\bar{s}}, \sigma_{\bar{s}})) \tag{6}$$

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

$$\ell\ell(\{\mu_i\}_{i\in s}|\mu_s, \sigma_s) = \sum_{i\in s} \log(f(\mu_i; \mu_s, \sigma_s)) \tag{7}$$

We use the EM test developed by Chen and Li (2009) to test a mixture of two distributions versus one. If the test rejects the null hypothesis of no mixture for an industry, we estimate it using Equation (6); otherwise, we use Equation (7).

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 (6). 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, so including both increases the chance of finding or getting close to the global maximum. As our contribution is not on mixture estimation, details are provided in the Appendix D.1.

Tests on whether an industry contains two types are sensitive to outliers. If one type's standard deviation is 100 times larger than the other type in the same industry or when the weights of one type is less than 5%, we treat the smaller type as an outlier and drop observations in the type. Types 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 3: Identify production elasticities and distortions

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

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

We treat the left-hand side of the equations as known because ϵ_g and $\mathbb{E}[e^{\delta_i}]$ are 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.

$$\log(\tau_i^K+1) \sim \begin{cases} 2\kappa_s^K\mathcal{N}(0,\sigma_{s,+}^K) \text{ , if } \tau_i^K>0 \\ (2-2\kappa_s^K)\mathcal{N}(0,\sigma_{s,-}^K) \text{ , if } \tau_i^K<0 \end{cases}$$

$$\log(\tau_i^L+1) \sim \begin{cases} 2\kappa_s^L\mathcal{N}(0,\sigma_{s,+}^L) \text{ , if } \tau_i^L>0 \\ (2-2\kappa_s^L)\mathcal{N}(0,\sigma_{s,-}^L) \text{ , if } \tau_i^L<0 \end{cases}$$

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 $\{P_iY_i, L_i\}_{i\in s}$:

$$\ell\ell(\{P_iY_i,K_i,wL_i\}_{i\in s}|\Theta_K,\Theta_L)=\sum_{i\in s}\ell\ell(P_iY_i,K_i|\Theta_K)+\ell\ell(P_iY_i,L_i|\Theta_L) \tag{8}$$

where

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

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.

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

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

$$
\begin{aligned}
\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]}
\end{aligned}
$$

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.

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

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:

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

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

D Mixture estimation, outliers, and unreasonable demand elasticities

D.1 Mixture Estimation

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 (6) 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, $\mu_{\bar{s}}$, and $\mu_{\underline{s}}$ in each industry s.

The random valuess of p are independent draws from a uniform distribution on (0,1). $\mu_{\bar{s}}$ and $\mu_{\underline{s}}$ 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 10, Table 11, and Table 12 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 $\sigma_1 = 1$ and $\sigma_2 = 1.5$. 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: $\mu_1 = 0$, $\mu_2 = 4$, $\sigma_1 = 1$ and $\sigma_2 = 2$.

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

Table 10: Estimates under different methods: 50 random starting values of p, μ_1, μ_2; sample size:200

true values	methods		p	μ_1	μ_2	σ_1	σ_2	nll
(0.25,1,1,1,1.5)	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
	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
(0.25,0,4,1,2)	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
	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 11: Standard deviation of estimates across the 50 starting values; sample size: 200

true values	methods		p	μ_1	μ_2	σ_1	σ_2	nll$*e8$
(0.25,1,1,1,1.5)	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
	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
(0.25,0,4,1,2)	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
	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 12: 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

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

D.3 Demand elasticities when each industry has only one type

When all the industries have only one type, the distribution of markups is a normal distribution:

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

Table 13 provides the summary statistics under this specification.

Table 13: Unweighted summary statistics of estimates not allowing for types inside industries

	N	Mean	St. Dev.	Pctl(10)	Pctl(25)	Median	Pctl(75)	Pctl(90)
MarkupsSNoGrouping	523	1.21	0.09	1.13	1.15	1.19	1.24	1.33
sigmaSNoGrouping	523	6.42	1.96	3.99	5.11	6.31	7.52	8.62
expCostShockNoGrouping	523	1.02	0.01	1.01	1.01	1.01	1.02	1.03

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 type, its distribution is:

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

when there are two types, its distribution is:

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

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:

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

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

$$\widehat{\Xi} \equiv \frac{\hat{\epsilon}_s}{\hat{\epsilon}_s - 1} \frac{1}{\hat{\alpha}_s^L + \hat{\alpha}_s^K} \tag{9}$$

$$\widehat{\Xi}^L \equiv \frac{\hat{\epsilon}_s}{\hat{\epsilon}_s - 1} \frac{1}{\hat{\alpha}_s^L} \tag{10}$$

$$\widehat{\Xi}^K \equiv \frac{\hat{\epsilon}_s}{\hat{\epsilon}_s - 1} \frac{1}{\hat{\alpha}_s^K} \tag{11}$$

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 (11) and Equation (10), 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 (9), 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 (9), Equation (11) and Equation (10) 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 type 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:

$$\frac{\epsilon - 1}{\epsilon} P F_1 = r$$
$$\frac{\epsilon - 1}{\epsilon} P F_2 = w$$

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.:

$$rK + wL = \frac{\epsilon - 1}{\epsilon} P(F_1 K + F_2 L) = \frac{r\epsilon - 1}{\epsilon} PF(K, L)$$

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 = rMC. Therefore,

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

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.