Misallocation under Heterogeneous Markups and Non-Constant Returns to Scale*

Xiaoyue Zhang† Junjie Xia ‡

May 20, 2022

The Latest Version

Abstract

Predicted TFP gains under Hsieh and Klenow (2009)'s framework are sensitive to demand elasticities and returns to scale, but simultaneously estimating them is difficult. We solve this problem by developing a framework allowing for an arbitrary distribution of firm-level markups and use microdata to estimate industry-specific production elasticities, within-industry type-specific demand elasticities when types are not observed, and firm-specific distortions. We apply our model to 2005 Chinese firm-level data and find that the predicted Total Factor Productivity (TFP) gains are 44% which is half of the previous findings. While the variation in markups does not affect predicted TFP gains, it lowers the predicted increase in labor income share by one-third, suggesting lower gains to average workers due to heterogeneous markups.

Keywords: Misallocation, heterogeneous markups, non-constant returns to scale, TFP gains, labor income share

*We are grateful to Jaap Abbring and Christoph Walsh for their patience, support and guidance throughout this project. This paper has benefited from insightful discussions with and comments from Jeffrey Campbell, Daniel Xu, Has van Vlokhoven, Malik Çürük, participants at the Tilburg Structural Econometrics Group, and Tilburg University GSS seminars. Xiaoyue is also thankful for the never-ending support and confidence from Edi Karni.

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

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

1 Introduction

Standard competitive-market theory predicts that equalizing the marginal revenue of production factors across firms brings efficiency gains (Melitz (2003), Restuccia and Rogerson (2008)). This implies that the large variation in marginal revenues in developing countries is an important source of their lower economic performance compared to developed countries (Hsieh and Klenow (2009)). However, recent trends of rising wage inequality and rising aggregate markups raise concerns over whether equalizing marginal revenues will hurt average working people. When all the firms have the same markups, reallocating capital and labor from high- to low-marginal-revenue firms brings efficiency gains and raises the labor income share. However, when firms have heterogeneous markups, the labor income share may not increase as much or even decrease if the majority of production factors are reallocated to high-markup firms. Besides, firms' returns to scale affect the predicted gains. Apart from some recent studies, many existing misallocation models assume constant returns to scale. A growing literature finds empirical evidence for large variation in the returns to scale (Chirinko and Fazzari (1994), Basu and Fernald (1997), Gao and Kehrig (2016), Lafortune et al. (2021)). Unfortunately, it is not clear whether and how the constant-returns-to-scale assumption affects predicted efficiency gains.

Our paper demonstrates that predicted total factor productivity (TFP) gains under Hsieh and Klenow (2009)'s framework (hereafter HK) are highly sensitive to demand elasticities and returns to scale. Simultaneously estimating them is difficult. Our primary contribution is to develop an empirical framework allowing for an arbitrary distribution of firm-level markups and use microdata to estimate industry-specific production elasticities, within-industry type-specific demand elasticities when types are not observed, and firm-specific distortions. In our model, distortions cause the variation of the marginal revenues of capital and labor. Our framework does not impose constant returns to scale and is able to fit the large variation in firm-level markups. Therefore, it is robust to errors in the measurement of distortions caused by heterogeneous demand elasticities and is robust to biases in predicted efficiency gains that appear when efficiency gains are from equalizing revenue-based total factor productivity, i.e. TFPR (HK). This equalizing-TFPR approach is valid only under constant returns to scale and constant demand elasticities which, however, are often violated in empirical work. Moreover, the predicted changes in the labor income share using our framework can offer insights into the welfare impact of heterogeneous markups on labor.

Directly measuring the source of misallocation is difficult and sometimes impossible when the source is unclear or when multiple sources are at work. Following HK's idea of identifying the effect of misallocation without specifying the underlying sources, we use dispersion in

the marginal revenues of labor and capital to measure misallocation. A key feature of our model is that we do not rely on production elasticities of firms in other countries. HK uses American firms' production elasticities as the production elasticities of Chinese and Indian firms when estimating firm-specific distortions. Instead, we simultaneously estimate production elasticities and distortions using data on the labor and capital shares under a flexible distribution of distortions.

The difficulty in estimating production elasticities is that observed capital and labor income shares are affected by demand elasticities, distortions, and production elasticities. While demand elasticities are estimated in a separate step beforehand, we still need to estimate production elasticities and distortions. Without any parametric restrictions, this model is not identified. We disentangle production elasticities from distortions by allowing a flexible distribution of distortions and imposing constant production elasticities within industries.[1] When production elasticities are the same within an industry, variation in observed firm-level capital and labor income shares reflects distortions after controlling for differences in markups. Some firms may hire too much capital or labor so that the marginal revenue is lower than the rental rate of capital and labor, which indicates negative distortions. In contrast, other firms may hire too little capital or labor, exhibiting higher marginal revenues and positive distortions. Intuitively, positive distortions happen when firms face obstacles in acquiring labor or capital, and negative distortions take place when firms enjoy subsidies or favorable access to financial and labor markets. This idea of negative and positive distortions is also used by Restuccia and Rogerson (2008) to model misallocation. The advantage of our flexible distribution of distortions is that 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 completely 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.

In addition, accommodating the large variation of markups is challenging under a constant-elasticity-of-substitution (CES) demand because it requires constant demand elasticities and therefore constant markups. Nested CES allows demand elasticities and markups to vary across nests but not within nests. An arbitrary nested CES is not tractable because it is not identifiable without parametric restrictions. In general, there are two channels through which firms' markups vary: different demand elasticities and unexpected cost shocks that

[1]The latter is a standard assumption in the literature of misallocation, while the former is a convenient and practical feature of our estimation strategy.

change firms' realized marginal costs. The latter occurs when firms face price rigidity so that they can not modify prices when marginal costs change. We overcome the challenge of accommodating heterogeneous markups using a combination of nested CES demand and unexpected idiosyncratic cost shocks. We solve the tractability issue by specifying the finest industry category observed as parent nests and use the model to estimate the number of unobserved types within an industry. The remaining variation of markups is explained by cost shocks. Using cost shocks to capture excessive variation in markups under CES demand framework is employed in Atkeson and Burstein (2008). Its firms observe cost shocks when they set prices, and cost shocks are used to explain excessive variation in prices among domestic markets and different foreign markets through varying market structures. However, in Atkeson and Burstein (2008), firms in the same sector face the same cost shocks, which means cost shocks can not help explain markup variation within a sector. Our cost shocks are realized after firms set price and are firm specific. This allows us to accommodate any markup variation within an industry which is the modeling counterpart of a sector in Atkeson and Burstein (2008).

We do not ex ante impose a positive correlation between markups and market shares that results from a positive correlation between markups and productivity. Although studies on trade and domestic markets using American firms (Bernard et al. (2003), Atkeson and Burstein (2008), De Loecker and Warzynski (2012), and Edmond et al. (2019)) and Gupta (2021) using Indian firms assume or support this positive correlation, our data give no direct empirical evidence for it. In fact, we find a negative correlation between markups and sales. One explanation is that this positive correlation is more likely to occur in a market-based economy such as the American economy while the Chinese economy contains many regulations and distortions, such as entry barriers, a lack of market-based allocation of financial credit, and a significant role for State-Owned Enterprises (SOEs). Removing SOEs attenuates the negative correlation, which hints at the possibility of finding a positive correlation if we can create a sample of Chinese firms whose environment is more market based. Another explanation is that the positive correlation between markups and market shares under nested CES demand exists among oligopolies (Atkeson and Burstein (2008)). However, 80% of the industries in our data contain more than 50 firms and the top decile contains more than 1,000 firms. Furthermore, our data do not support the positive correlation even for industries with less than 25 firms and after dropping all SOEs. Finally, the missing positive correlation may result from distortions in capital and labor market which distort observed sizes or market shares. If high-markup firms tend to have larger distortions while low-markup firms tend to have smaller or even negative distortions, we would not be able to observe a positive correlation even if in a distortion-free economy exists. Therefore,

it is unclear whether Chinese firms follow this pattern. Given this, we argue that it is better not to impose any correlation ex ante and to use a framework that accommodates arbitrary correlations. This motivates our decision on not to use demand structures that offer 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 CREMR demand (Mrázová et al. (2021)), and hyperbolic absolute risk aversion preference (Haltiwanger et al. (2018)).

The cost of not imposing any correlation in our framework is that we can not measure the deadweight loss caused by heterogeneous markups because the market outcome always aligns with that of the social planner. The welfare implication on the labor income share comes merely from reallocating resources among firms with different markups and different firm-level labor shares. There are several recent studies that measure the deadweight loss caused by heterogeneous markups, including the theoretical foundations provided by Dhingra and Morrow (2019) and Mrázová et al. (2021), empirical studies by Liang (2021) and Gupta (2021) in the context of Indian firms, and the impact of trade on the deadweight loss by Edmond et al. (2015), Feenstra and Weinstein (2017), and Baqaee et al. (2020). To the best of our knowledge, all the existing structural models that capture the deadweight loss require imposing a given correlation between firm sizes and markups ex ante.[2]

In spite of this, our framework offers a way of applying the nested CES framework when firm-level markups vary. One motivation of the other demand framework mentioned above is that applying CES means constant demand elasticities and markups. The other framework allows varying markups and can, therefore, fit the data more closely (Mrázová et al. (2021)). Our framework shows that one can benefit from the tractability and simplicity of the CES demand while still allow the large variation of markups.

We apply our model to Chinese data on the year 2005 and find that predicted TFP gains from reallocating resources within industries are 43.9%, which is less than the 86.6% predicted by HK. Applying our model to other years is straightforward and we find similar results when using data on 2001.[3] When taking into account the gains from reallocating resources across industries, TFP gains are 50.6%. The labor income share increases by 7.4 percentage points to 27.2%. In a counterfactual scenario of homogeneous demand elasticities where all other primitives are the same as our estimators, predicted TFP gains are similar but the increase in labor income share is higher. More specifically, when demand elasticities are 8.5, the average of our estimated demand elasticities, the labor income share increases

[2] Baqaee et al. (2020) uses a generalized Kimball preference but still requires that markups increase as a firm grows.

[3] 2001 is chosen because HK also reports 2001, which makes comparison convenient. We do not pick 2004 which is also reported in HK because the data quality in 2001 is better.

by 11.4 percentage points while total TFP gains are 51.8% almost the same as the 50.6% under heterogeneous demand elasticities. This indicates changes in TFP gains are not always informative about changes in the labor income share. Failing to account for heterogeneous markups may miss the impact on average workers. Furthermore, our estimated distortions suggest that SOEs are more likely to overuse labor and capital compared to domestic private firms. This is consistent with the fact that SOEs tend to have more favorable financial access but have difficulties in reducing labor costs, as they use more permanent labor contracts.

Relaxing the assumptions imposed by HK provides empirical evidence on the importance of using parameters estimated from microdata. If HK's low demand elasticities are replaced with the average of our estimates, 8.5, predicted TFP gains from reallocating resources within industries rise to 362.3%. They drop slightly but are still 298.6% when allowing heterogeneous demand elasticities. These numbers shrink to 63.8% and 59.2% once use our estimated production elasticities. This pattern calls attention to the sensitivity of predicted TFP gains to these parameters and suggests the importance of using our estimation method.

Our paper can be seen as a generalization of HK. In fact, our method can be applied to situations where the assumptions required by the original HK method are violated, i.e. heterogeneous markups and sloped marginal cost curves. Replacing our estimated production elasticities and demand elasticities by those assumed in HK reproduces its result and its requirement of equalizing TFPR under no distortion.

Haltiwanger et al. (2018) argues that variation in TFPR can reflect demand shifters instead of misallocation and, when shifting along a sloped marginal cost curve, efficiency rather than distortions. Furthermore, if demand elasticities differ across firms, distortions measured by HK are a mixture of distortions and markups. We find large dispersion in observed revenue-cost ratios. If this large dispersion is driven by differences in markups, the distortions measured by HK contain variations in markups. We also find returns to scale vary across industries which suggests some variation in TFPR should reflect efficiency. We deal with these biases by directly modeling demand as having heterogeneous demand elasticities and production functions as non-constant returns to scale. This disentangles the part of variation in marginal revenues due to markups from the part due to distortions. Unlike HK, TFPR is not equalized when distortions are removed and the TFPR variation in our no-distortion equilibrium indeed reflects changes in marginal cost and markups. Both Haltiwanger et al. (2018) and our paper deal with the implication of HK's assumptions and their impact on predicted TFP gains. However, we adopt a completely different approach in relaxing the assumptions. Haltiwanger et al. (2018) uses the hyperbolic absolute risk aversion utility function to relax the constant-markup assumption. This means they also imposes a positive correlation between market shares and markups. Besides, our method

can be used when price and quantity data are not available while Haltiwanger et al. (2018) requires those data.

Our paper complements recent studies on exploring sources of TFPR variation other than distortions. Our method takes into account the impact of varying returns to scale and demand elasticities on TFPR variation and disentangles the part of variation due to distortions from the part due to varying returns to scale and varying markups. Bils et al. (2020) focuses on the part of TFPR variation caused by measurement errors and proposes a way to correct additive measurement errors in revenue and input. David and Venkateswaran (2019) points out adjustment costs and information frictions can also cause dispersion in the ratio of value-added and capital and proposes a framework to disentangle different sources. Due to the link between TFPR and the ratio of value-added and capital, it is possible that adjustment costs and information frictions also contribute to TFPR variation.

Similar to our paper, Ruzic and Ho (2021) estimates parameters of heterogeneous demand elasticities and production elasticities using micro data. However, its method requires constant demand elasticities and markups within industries so that the variation of demand elasticities within industries can still contaminate its measures of distortions. It also requires that distortions have zero mean while we do not impose this restriction. Its model does not allow for firms to have negative profits but there are 15% of firms in our data experiencing negative profits. Dropping all these firms may leave out useful information. Our model can use this information because it allows negative profits. Lastly, its method can only talk about predicted TFP gains within industries because it calculates relative distortions whereas our method can calculate both within and cross industries TFP gains.

Our paper also contributes to the discussion whether constant returns to scale is consistent with firms' empirical production decisions. While many studies assume constant returns to scale, a growing literature finds empirical evidence against it, and large variation in industry-level returns to scale have been documented (Chirinko and Fazzari (1994), Basu and Fernald (1997), Gao and Kehrig (2016), Lafortune et al. (2021)). Our structural analysis confirms returns to scale vary significantly across industries. To check the robustness of our structural estimation of returns to scale, we implement a reduced form analysis using Klette and Griliches (1996) which relies on different modelling assumptions. Both our structural and reduced-form analyses demonstrate on average decreasing returns to scale. This may result from under-reported labor expenditures (non-wage labor expenditure is not included) and variation of the ratio between wage and non-wage labor expenditure across firms (without such variation, our reduced-form analysis should be able to correct it). It can also be caused by other unobserved production factors such as intangible assets. The fact that we use microdata can be another reason as Basu and Fernald (1997) points out returns

to scale is sensitive to aggregation levels. Earlier research on returns to scale mainly uses data aggregated at the level of broadly defined industries. In fact, Lafortune et al. (2021) also finds decreasing returns to scale using an industry-city panel data in the US.

Finally, our paper is related to recent studies on the welfare implications of markups. This literature covers a variety of topics, such as the trend of rising aggregate markups (Basu (2019), Autor et al. (2020), and De Loecker et al. (2020)), the interaction between trade and markups (Edmond et al. (2015) and Feenstra and Weinstein (2017)), endogenous markups in the context of creative destruction (Peters (2020)), and channels of welfare loss (Edmond et al. (2019)). Although we do not explicitly measure welfare changes, our paper improves the understanding of the welfare impact of heterogeneous markups and misallocation on labor. A lower increase in the labor income share under heterogeneous markups suggests a smaller gain for workers.

The remainder of the paper is organized as follows. We introduce the data set in Section 2 and conduct a reduced-form analysis in Section 3. Section 4 describes our theoretical model. We discuss our identification procedure in Section 5 and presents our estimation results in Section 6. Section 7 concludes the paper. Appendix provides the derivation of theoretical results, details of estimation procedures, a complete list of TFP gains when relaxing assumptions in HK, and results from using Chinese data on 2001.

2 Data

Our data source is the Chinese Annual Survey of Manufacturing (ASM, 1998-2009) collected by the National Bureau of Statistics of China.[4] This data set has been used by previous studies including HK, Song et al. (2011), and David and Venkateswaran (2019). It is a census containing above-scale non-state firms (firms with more than 5 million RMB, about $\$600,000$, in revenue) and all the state-owned enterprises (SOEs). For demonstration, our structural analysis uses the year 2005 but it should be easily applied to other years. We do the same structural analysis for the year 2001 in Appendix G. The reduced-form analysis on returns to scale uses 1998-2009.

The data set contains rich information on firm-level value-added, wage expenditure, net value of fixed asset, 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, etc. A full list of the types of observations dropped is provided in Appendix A. 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

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

the tails of revenue-cost ratio should already deal with abnormal profits. Since the focus of this research is on structural analysis and we use the year 2005 for the structural analysis in our main text, we provide here the summary statistics of the year 2005 and refer to Appendix A for the summary statistics of the entire data. In both the year 2005 and the entire data, there are 15% of the firms with negative profits.

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 its labor expenditure does not include the non-wage part and the aggregate labor share of value added is too low compared to the one inferred by 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 5 in Appendix A demonstrates this pattern with more details. This makes sense intuitively 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. Since 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 more the unweighted 32% average rather than the weighted 20% average. In other words, such uniform scale-up introduces new biases.

It is possible that using unscaled labor share underestimates production elasticities of labor and therefore underestimate returns to scale on average. We use a reduced form analysis to check our structural estimators. In fact, later sections show the reduced-form estimators by and large coincide with the structural estimators. We consider this as a favorable piece of evidence because the reduced form analysis is completely different from our structural model so that there is no guarantee the two results coincide. Moreover, scaling-up all the firms'

labor shares by the same proportion does not change the reduced-form estimators. Although the ideal but infeasible practice is to find the non-wage part for each firm, the evidence leads us to believe using the observed wage expenditure is the best feasible option for estimating production elasticities and predicting TFP gains.

3 Reduced-form analysis

Before carrying out structural analysis, we fist do a classical production estimation to show what the reduced-form analysis tells us about returns to scale. Simple linear regression on firms' expenditure share also demonstrate correlation between firm types and distortions.

Our reduced form analysis on average returns to scale closely follows Klette and Griliches (1996). We denote r_{it} as value added deflated by two-digit industry price index P_{st}. We deflate at this level because only this level's price index is available. A CES demand implies:

$$y_{it} - y_{st} = -\epsilon(p_{it} - p_{st}) + u_{it}^d$$

ϵ is demand elasticities. y_{it}, p_{it}, y_{st}, and p_{st} are the logarithm of firm-level production and prices, and industry-level production and pries. u_{it}^d is a demand shock. This reduced form requires demand elasticities to be constant across industries and over time. In Appendix B, we allow demand elasticities to vary across industries but still constant over time by applying this analysis to each 2-digit industry. Since $r_{it} = y_{it} + p_{it} - p_{st}$, the demand side requires deflated firm-level price $p_{it} - p_{st}$ to satisfy the following function of deflated value added r_{it} and industry-level production y_{st}:

$$p_{it} - p_{st} = \frac{r_{it} - y_{st} - u_{it}^d}{1 - \epsilon} \tag{1}$$

A firm's production function in logarithm is:

$$y_{it} = \log(A_i) + \alpha^K k_{it} + \alpha^L l_{it} + u_{it}^y$$

k_{it} and l_{it} are the logarithm of capital and labor usage. u_{it}^y is a supply shock. Production elasticities α^K and α^L are required to be constant across industries and over time. Appendix B allows them to vary across industries. The logarithm of deflated revenues using the supply-side structure is simply $y_{it} + p_{it} - p_{st}$. Combine this with the production function and Equation 1:

$$r_{it} = \beta_0 + \beta_1 k_{it} + \beta_2 l_{it} + \beta_3 y_{st} + v_{it} \tag{2}$$

where $\beta_0 = \frac{\epsilon-1}{\epsilon}\log(A_i)$, $\beta_1 = \frac{\epsilon-1}{\epsilon}\alpha^K$, $\beta_2 = \frac{\epsilon-1}{\epsilon}\alpha^L$, and $\beta_3 = \frac{1}{\epsilon}$. v_{it} is a combination of demand shocks and supply shocks, i.e. $\frac{\epsilon-1}{\epsilon}u^y_{it} + \frac{u^d_{it}}{\epsilon}$. Equation 2 is the regression used in this reduced form analysis. The estimated demand elasticities and returns to scale are:

$$\hat{\epsilon} = \frac{1}{\hat{\beta}_3}$$
$$\hat{\alpha}^K + \hat{\alpha}^L = \frac{\hat{\beta}_1 + \hat{\beta}_2}{1 - \hat{\beta}_3}$$

Following Klette and Griliches (1996), we estimate Equation 2 using first differences with a constant. The longer the panel data, the easier it is for identification. We therefore use all the years available, i.e. 1998-2009. The constant captures the possibility of a constant growth rate in deflated revenue. Due to concerns over endogeneity as explained in Klette and Griliches (1996), we also do an IV estimation using the second-order differences of labor and capital as instruments. Results of both OLS and IV estimation are reported in Table 2. Similar to Klette and Griliches (1996), our OLS estimator and IV estimator produce similar results. Our IV estimation infers returns to scale to be 0.61 $((0.406 + 0.156)/(1 - 0.077))$ and demand elasticities to be 12.92 $(1/0.077)$ while OLS infers returns to scale to be 0.53 $((0.343 + 0.134)/(1 - 0.104))$ and demand elasticities 9.62 $(1/0.104)$. The Hausman test rejects the null of no endogeneity, so we prefer our IV estimators. One thing worth noticing is that the estimators are not affected if we scale up every firms' labor share by a common proportion.

We also look at the relation between ownership type and production factor usage. For easier comparison to the distortions estimated from our structural model in later sections, we use only the year 2005. The benchmark in Table 3 is domestic private firms. It shows the SOEs on average have higher labor expenditure share and capital share than domestic private firms after controlling for industry fixed effects. SOEs also tend to use disproportionally more capital. This echoes the distortion patterns identified by our structural analysis. There are also foreign firms in the data but only the coefficients on SOEs are reported here as it is of the most interest.

TABLE 2: Reduced-form estimation of returns to scale

	Dependent variable	
	r_{it}	
	OLS	IV
	(1)	(2)
l_{it}	0.343^{***}	0.406^{***}
	(0.001)	(0.002)
k_{it}	0.134^{***}	0.156^{***}
	(0.001)	(0.002)
y_{st}	0.104^{***}	0.077^{***}
	(0.003)	(0.003)
constant	0.086^{***}	0.048^{***}
	(0.001)	(0.001)
Observations	1,182,562	815,546
R^2	0.099	0.070
Adjusted R^2	0.099	0.070
Residual Std. Error	0.660 (df = 1182558)	0.615 (df = 815542)
F Statistic	$43{,}471.580^{***}$ (df = 3; 1182558)	

Note: *p<0.1; **p<0.05; ***p<0.01

r_{it} is deflated firm-level value added, VA_s is industry s's aggregate VA.

l_{it} is deflated observed labor expenditure

TABLE 3: Relationship between income shares and ownership (2005)

	Dependent variable		
	log(wL/PYsi)	log(K/PYsi)	log(wl/K)
	(1)	(2)	(3)
SOEs	0.487***	0.730***	−0.244***
	(0.010)	(0.013)	(0.012)
Observations	229,416	229,416	229,416
R^2	0.020	0.020	0.002
Adjusted R^2	0.017	0.017	−0.0003
F Statistic (df = 3; 228890)	1,522.401***	1,521.787***	152.746***

Note: *p<0.1; **p<0.05; ***p<0.01
Industry fixed effect is included.

4 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 type and a low-demand-elasticity type. When there are two types, $\bar{s}$ 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

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

Firms in $\bar{s}$ face higher demand elasticities, thus lower markups, than firms in $\underline{s}$, i.e. $\epsilon_{\bar{s}} > \epsilon_{\underline{s}}$. The markups decided by demand elasticities are the expected markups as they are also affected by cost shocks, which will be explained in details below. An intuition behind high-demand-elasticity versus low-demand-elasticity types 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 type refers to the type $\bar{s}$ or $\underline{s}$ when there are two types inside s or to industry s itself when there is only one type; an industry always refer to an industry s. Loosely speaking, our types 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 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 SOEs, have higher τ_i^L as they normally can not 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 are collected by some agent or government and are transferred back to the representative consumer as a lump sum. We do not allow government to run deficit:

$$\sum_s \sum_g \sum_{i \in g} \tau_i^K K_i + \tau_i^L L_i \geq 0$$

However, this condition turns out to be not binding in our empirical study and is ignored when estimating all the other parameters.

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. One can think of it as, for example, uncertainty in managerial cost which is higher as firms hire more capital and labor. 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:[5]

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

Derivation of optimal prices and other key variables in this section are provided in Appendix C. Optimal prices are a function of production because firms are not constant returns to scale, and marginal cost depends 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 larger variations in markups using less parameters than regular nested-CES models. 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 regular nested-CES model. 15% of the firms have negative profits implying markups less than 1. These firms' behavior cannot be explained by a static model without cost shocks. The first part of $\mu_i + 1$ is the same as the standard ones 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.

The gaps between production elasticities and observed labor and capital income shares

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

pin down the distortions[6]:

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

The demand side is very simple. A representative consumer owns all the capital and labor, receiving as a lump sum the government income $\sum_s\sum_g\sum_{i\in g}\tau_i^K K_i + \tau_i^L L_i$. 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.

The aggregate TFP gains can be decomposed into two parts: gains from reallocation within types and gains from reallocation across types.[7]

$$\text{TFP gains} = \frac{\mathcal{Y}^*}{\mathcal{Y}} = \prod_g \underbrace{\left[\frac{\text{TFP}_g^*}{\text{TFP}_g}\right]^{\beta_g}}_{\text{gains within types}} \cdot \underbrace{\left[\left(\frac{L_g^*}{L_g}\right)^{\alpha_g^L}\left(\frac{K_g^*}{K_g}\right)^{\alpha_g^K}\right]^{\beta_g}}_{\text{gains across types}}$$

The type-level TFP, i.e. TFP_g is defined as $\frac{Y_g}{K_g^{\alpha_g^K}L_g^{\alpha_g^L}}$ and becomes TFP_g^* when distortions are 0. L_g and K_g are capital and labor used in g with distortions while L_g^* and K_g^* are without distortions. The first equation holds because the supply of total capital and labor is fixed.

The firm-level and type-level TFPR are defined as:

$$\text{TFPR}_i \equiv P_iA_i = \frac{P_iY_i}{K_i^{\alpha_s^K}L_i^{\alpha_s^L}}$$
$$\text{TFPR}_g \equiv \frac{\sum_{i\in g}P_iY_i}{K_g^{\alpha_s^K}L_g^{\alpha_s^L}}$$

The type-level TFP is also a weighted sum of firm-level TFP or A_i, which is the same as

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

[7] HK only has the within-industry gains.

HK.

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

We follow HK to calculate A_i using $\frac{(P_i Y_i)^{\epsilon_g/(\epsilon_g-1)}}{K_i^{\alpha_s^K}(wL_i)^{\alpha_s^L}}$ because we only need TFP ratios between firms belonging to the same type. However, the ratio $\text{TFPR}_i/\text{TFPR}_g$ and consequently TFP_g and TFP_g^* are different from HK due to non-constant returns to scale.

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

The last term in $\text{TFPR}_i/\text{TFPR}_g$ disappears under constant returns to scale. The rest is the same as the one in HK after replacing our notation of distortions by theirs.

The formula of TFP_g^* is:

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

where

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

Larger firms receive higher weights in type-level TFP if they are increasing returns to scale but receive lower weights if they are decreasing returns to scale. Different from HK, TFPR under no distortion is not equalized within an type unless $\alpha_s^K + \alpha_s^L = 1$ and these variations reflect efficiency, as argued by Haltiwanger et al. (2018). Demand elasticities ϵ_g, firm-level productivity A_i, and returns to scale $\alpha_s^L + \alpha_s^K$ affect the no-distortion TFPR ratio, which links the TFPR ratio to variation of markups and the marginal cost:

$$\frac{\text{TFPR}_i^*}{\text{TFPR}_g^*} = \left(\frac{P_i^* Y_i^*}{P_g^* Y_g^*} \right)^{1-\alpha_s^K-\alpha_s^L} = \left(\frac{A_i^{\frac{\epsilon_g - 1}{\epsilon_g + (1-\epsilon_g)(\alpha_s^L + \alpha_s^K)}}}{\sum_{i \in g} A_i^{\frac{\epsilon_g - 1}{\epsilon_g + (1-\epsilon_g)(\alpha_s^L + \alpha_s^K)}}} \right)^{1-\alpha_s^K-\alpha_s^L}$$

Calculating the gains across types is the same as calculating the ratios between type-level

labor and capital usage before and after the reallocation. The ratios can be written as:

$$\frac{L_g^*}{L_g} = \frac{w^* L_g^* / (w^* L)}{w L_g / (wL)}$$
$$\frac{K_g^*}{K_g} = \frac{K_g^* / K}{K_g / K}$$

Since $wL_g/(wL)$ and K_g/K are directly observed, we only need to calculate $w^*L_g^*/(w^*L)$ and K_g^*/K:

$$\frac{w^* L_g^*}{w^* L} = \frac{\beta_g \cdot \frac{\alpha_s^L}{\epsilon_g/(\epsilon_g - 1)\mathbb{E}[e^{\delta_i}]}}{\sum_g \beta_g \cdot \frac{\alpha_s^L}{\epsilon_g/(\epsilon_g - 1)\mathbb{E}[e^{\delta_i}]}}$$
$$\frac{K_g^*}{K} = \frac{\beta_g \cdot \frac{\alpha_s^K}{\epsilon_g/(\epsilon_g - 1)\mathbb{E}[e^{\delta_i}]}}{\sum_g \beta_g \cdot \frac{\alpha_s^K}{\epsilon_g/(\epsilon_g - 1)\mathbb{E}[e^{\delta_i}]}}$$

Using the formulas above, aggregate TFP gains can be calculated once the parameters involved are identified.

5 Identification

We use observed firm-level value added, labor expenditure, the depreciated net value of capital, total cost, and sales to identify all the required parameters after imposing structural assumptions about cost shocks and 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 $R = 0.1$. The value of R affects TFP gains only via estimated production elasticities. In other words, if production elasticities are known, as it is the case in HK, changing the value of R does not change TFP gains as R does not affect allocation across firms. Our identification involves three steps and are explained in the following three subsections.

5.1 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.

5.2 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:

$$\ell\ell(\{\mu_i\}_{i\in s}|w_s, \mu_{\bar{s}}, \mu_{\underline{s}}, \sigma_{\bar{s}}, \sigma_{\underline{s}}) = \sum_{i\in s} \log(w_s f(\mu_i; \mu_{\underline{s}}, \sigma_{\underline{s}}) + (1 - w_s) f(\mu_i; \mu_{\bar{s}}, \sigma_{\bar{s}})) \tag{3}$$

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

$$\ell\ell(\{\mu_i\}_{i\in s}|\mu_s, \sigma_s) = \sum_{i\in s} \log(f(\mu_i; \mu_s, \sigma_s)) \tag{4}$$

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 (3); otherwise, we use Equation (4).

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 (3). 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 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 Appendix D.2.

5.3 Step 3: Identify production elasticities and distortions

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

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

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.

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

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

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

$$\ell\ell(\{P_iY_i, K_i, wL_i\}_{i\in s}|\Theta_K, \Theta_L) = \sum_{i\in s} \ell\ell(P_iY_i, K_i|\Theta_K) + \ell\ell(P_iY_i, L_i|\Theta_L) \tag{5}$$

where

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

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

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

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

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

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

6 Results

In this section, we first report our estimators and discuss their possible biases. We then show our main results and finish the section with robustness checks.

6.1 Estimated parameters

Before going to the main results, we first present our estimators. Table 4 shows 462 industries are estimated as a mixture of two normal distributions and 61 as a single normal distribution. Industry containing two types tend to be larger in terms of firm counts. The middle row of Table 5 demonstrates the ex-ante probability of being a high-demand-elasticity type in general exceeds that of being a low-demand-elasticity type, which means in an industry with different demand elasticities, normally there are more firms facing higher demand elasticities.

TABLE 4: Distribution of firm counts for industries containing 1 type and 2 types

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 5: Type-level summary statistics of estimates allowing for types inside industries

	N	Mean	St. Dev.	Pctl(10)	Pctl(25)	Median	Pctl(75)	Pctl(90)
$\mathbb{E}_g[\mu_i + 1]$	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
$\mathbb{E}_g[e_i^\delta]$	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_K	523	0.16	0.17	0.04	0.06	0.09	0.19	0.36
α_L	523	0.39	0.23	0.13	0.21	0.33	0.57	0.76
scale	523	0.55	0.31	0.22	0.32	0.48	0.75	0.95

Table 5 also reports the summary statistics of estimated markups, demand elasticities, and expected cost shocks at the type level and estimated production elasticities and returns to scale at the industry level, treating each type or industry as having the same size. We document the large variation of industry-level production elasticities and returns to scale. The average returns to scale of industries are 0.55 and the average of all the firms is 0.47, i.e. weight each industry by its firm counts. Demand elasticities also vary across types, with the top 10 percentile more than 3 times larger than the bottom 10 percentile. The average

demand elasticities of types are 6.33 while the average of all the firms is 8.49. We use the latter in our counterfactual case of homogeneous markups and in the comparison to HK because it better reflects the average of our data for our production elasticities estimation. The top 10 percentile of markups is about 40% higher than the bottom 10 percentile of markups.

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.

The estimated average returns to scale and demand elasticities from our reduced form analysis in Section 3 are 0.61 and 10.87. The former is a bit higher than the industry-level average of returns to scale 0.55 and more far away from our firms-level average 0.47. The latter is higher than our type-level average of demand elasticities 6.33, and also higher but closer to the firm-level average 8.49.

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 6). 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 6 keeps only those observations where the two criteria agree.

TABLE 6: 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

6.2 The inferred markups: biased or not

The results in Table 5 show that the average returns to scale is about 0.6. If this is the true returns to scale, this suggests more than 80% of the firms in our data set price below marginal cost. This also means if we correct the markups inferred in the first identification step, our estimated markups would be very different from those in the literature listed above. In fact, one should not use the returns to scale estimated from the third step to correct the markups inferred in the first step because simultaneously identify returns to scale and markups using only revenue data and no physical production data is not possible. We provide a formal proof for this in Appendix H and show that the same problem exists as long as the production function is homogeneous of degree r for any positive number r. A similar finding is also discussed in Bond et al. (2021). If one ignores the identification issue and uses the returns to scale estimated in the third step to correct the markups inferred in the first step, updating can still take place but the update happens only when the sample analogues differ from their true values. If we have the entire population and our model correctly specifies the data generating process, the markups inferred from the first step should always gives constant returns to scale in the third step.

In fact, we should look at the production model as a simplification of a richer model where firms' use labor, tangible assets, and intangible assets. When maximizing profits, firms take intangible assets as given and choose the optimal labor and tangible assets. The capital we observed in the data is the tangible assets and we do not observe intangible assets.

Therefore, the sum of α_s^K and α_s^L is only part of the returns to scale. Therefore, firms may still be close to constant returns to scale when the estimated $\alpha_s^L + \alpha_s^K$ is below 1. We treat intangible assets as given because intangible asset such as knowledge and experience are more difficult to adjust than labor and tangible assets. Appendix I shows that the procedures of estimating TFP gains from equalizing the marginal revenue of labor and tangible assets in this extended model is the same as the one presented above and our predicted TFP gains should be interpreted as the gains from equalizing the marginal revenue of labor and tangible assets.

People familiar with this data may argue that the unobserved non-wage labor share leads to the low $\alpha_s^L + \alpha_s^K$ and may prefer following HK to scale up the observed labor share or using the observed number of employees. We agree that the unobserved non-wage labor is indeed a problem for anyone using this data but the two methods proposed are unlikely good solutions. If these methods can correct estimated $\alpha_s^L + \alpha_s^K$, reduced-form analysis using scaled wage expenditure or the number of employees should also give higher estimated returns to scale. However, we find the estimated returns to scale in both cases are around 0.6.

Given the fact that directly using revenue-cost ratio provides estimated markups in line with those in literature and our estimated $\alpha_s^L + \alpha_s^K$ is only part of the returns to scale, we prefer to not correcting the inferred markups in our first step at all. Even if our estimators contains some bias, our robustness check shows that the size of possible bias does not matter to our main results (Section 6.6).

6.3 Markups and sizes

When discussing the correlation between markups and sizes, existing research usually use sales to measure sizes. Our project also follows this practice. However, we find mixed evidence about whether there is a positive correlation between markups and sales. Following Edmond et al. (2019), we define relative sales as those normalized by the unweighted industry-level average, relative expected markups as expected markups normalized by the cost-weighted industry-level average, and relative revenue-cost ratios as revenue-cost ratios normalized by the cost-weighted industry-level average. The revenue-cost ratios are directly observed in our data and are treated as the realized markups, denoted as $\mu_i + 1$. The expected markups $\mathbb{E}_g[\mu_i + 1]$ are estimated by our model. If there is a positive correlation between markups and sales, we should see larger firms more likely to belong to the higher-markup type of an industry and to have higher relative expected markups. We should also see larger firms have higher relative revenue-cost ratio. However, the first two columns of

Table 7 demonstrate negative correlations.

TABLE 7: Correlation between relative sales and relative markups

	Dependent variable			
	$\ln(\mathbb{E}_g[\mu_i+1])$	$\ln(\mu_i+1)$	$\ln(\mathbb{E}_g[\mu_i+1])$	$\ln(\mu_i+1)$
	full sample		no SOEs	
	(1)	(2)	(3)	(4)
ln(sales)	-0.003^{***}	-0.005^{***}	-0.002^{***}	-0.004^{***}
	(0.0001)	(0.0003)	(0.0001)	(0.0003)
Constant	0.005^{***}	0.009^{***}	0.004^{***}	0.008^{***}
	(0.0002)	(0.0004)	(0.0002)	(0.0004)
Observations	229,410	229,410	217,835	217,835
R^2	0.001	0.001	0.001	0.001
Adjusted R^2	0.001	0.001	0.001	0.001
Residual Std. Error	0.078 (df = 229408)	0.153 (df = 229408)	0.076 (df = 217833)	0.148 (df = 217833)
F Statistic	318.612^{***} (df = 1; 229408)	259.458^{***} (df = 1; 229408)	194.828^{***} (df = 1; 217833)	165.349^{***} (df = 1; 217833)

Note: *p<0.1; **p<0.05; ***p<0.01

All the variables are in relative values, i.e. they are normalized by industry-type averages. $\ln(\mathbb{E}_g[\mu_i+1])$ is expected markups in log. $\ln(\mu_i+1)$ is the log of revenue-cost ratio.

TABLE 8: Correlation between relative sales and relative markups for industries with firm counts ≤ 25

	Dependent variable			
	$\ln(\mathbb{E}_g[\mu_i+1])$	$\ln(\mu_i+1)$	$\ln(\mathbb{E}_g[\mu_i+1])$	$\ln(\mu_i+1)$
	small industries		small industries and no SOEs	
	(1)	(2)	(3)	(4)
ln(sales)	-0.008^{***}	-0.023^{***}	-0.007^{***}	-0.022^{***}
	(0.003)	(0.004)	(0.003)	(0.005)
Constant	0.127^{***}	0.180^{***}	0.128^{***}	0.180^{***}
	(0.004)	(0.006)	(0.004)	(0.007)
Observations	2,652	2,652	2,397	2,397
R^2	0.004	0.011	0.003	0.009
Adjusted R^2	0.004	0.010	0.002	0.008
Residual Std. Error	0.158 (df = 2650)	0.273 (df = 2650)	0.155 (df = 2395)	0.269 (df = 2395)
F Statistic	10.858^{***} (df = 1; 2650)	28.277^{***} (df = 1; 2650)	6.836^{***} (df = 1; 2395)	21.050^{***} (df = 1; 2395)

Note: *p<0.1; **p<0.05; ***p<0.01

All the variables are in relative values, i.e. they are normalized by industry-type averages. $\ln(\mathbb{E}_g[\mu_i+1])$ is expected markups in log. $\ln(\mu_i+1)$ is the log of revenue-cost ratio.

There are multiple explanations for our different results. One possibility is that the positive correlation is more likely in a market-based economy but Chinese economy experience various distortions, such as entry barriers and entry permissions, the lack of market-based allocation of financial credits, the significant roles for State-Owned Enterprises (SOEs), et cetera. In fact, most studies on the positive correlation uses American firms (Bernard et al. (2003), Atkeson and Burstein (2008), De Loecker and Warzynski (2012), and Edmond et al.

(2015), Edmond et al. (2019)) [8] and American economy is more market-based than China. In the last two columns of Table 7, we drop all the SOEs in our data. The magnitude of the negative correlation is smaller but still significant. Suggesting SOEs contribute to some of the negative correlations. SOEs are only one part of all the possible distortions in China and our results in the later section demonstrates that the observed firm ownership is an informative but noisy indicator of whether a firm behaves like a typical SOE. It is, therefore, not surprising that dropping all the SOEs does not provide significant positive correlations.

Another explanation to the missing positive correlation in Table 7 is that this positive correlation may be more salient in industries where firms act as oligopolies, or in other words, only a few firms are interacting. This is also the assumption that generates the positive correlation in Atkeson and Burstein (2008) and Edmond et al. (2015). In Table 8, we check whether industries with less than 25 firms demonstrate a positive correlation. Similar to Table 7, the first two columns include SOEs while the last two drop them. Different from our expectation, the negative correlations become more salient and dropping SOEs attenuate it slightly. This result looks puzzling. Small industries seem to experience more market interruptions not captured by the presence of SOEs. Perhaps, entry permission imposed by the government artificially create some small industries so that small industries deviate more from market equilibrium. Similar results remain when looking at industries with less than 20 firms or 30 firms.

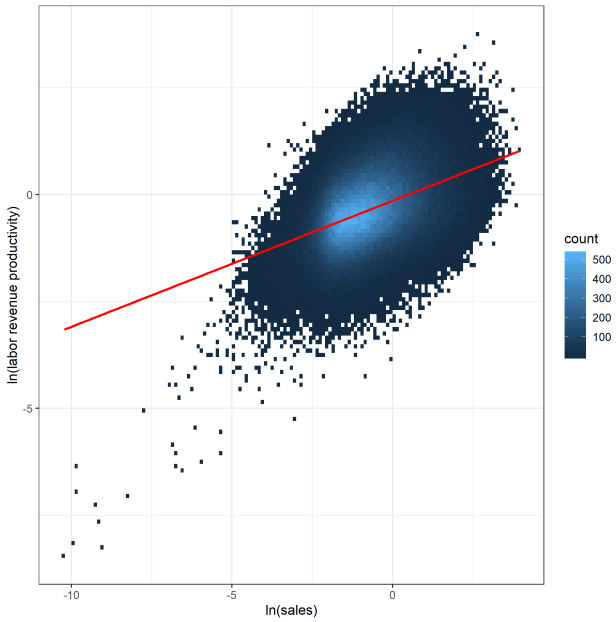
The third explanation is the observed sizes are distorted due to capital and labor distortions. If high-markup firms face larger distortions while low-markup firms face lower or even negative distortions, we won't be able to observe the positive correlation even if it exists in a distortion-free market. Since we do not find evidence to support the positive correlation, we favor not imposing any ex-ante correlation between productivity and markups and let the data tells us whether a larger firm belongs to a high-markup type of an industry. Interestingly, in spite of no restrictions, our model seems to successfully tease off part of the negative correlation and treat some part of the revenue-cost ratio which is negatively correlated with sales as cost shocks because the coefficients for $\ln(\mathbb{E}_g[\mu_i + 1])$ are smaller in magnitude than those for $\ln(\mu_i + 1)$.

Another interesting finding is that although markups and sales are negatively correlated, we do observe a positive correlation between the relative labor revenue productivity (hereafter labor productivity) and the relative sales (Figure 1). This is used by Edmond et al. (2019) to identify the parameters which determine the positive correlations between markups and productivity because if markups does not vary with productivity, then labor productivity

[8] To the best of our knowledge, there is only one paper, Gupta (2021), using data from developing countries that support the positive correlation. There it uses Indian data.

and markups also do not vary with sales. Following Edmond et al. (2019), relative labor productivity is defined as labor productivity normalized by the average at the industry level and labor productivity is sales divided by labor expenditures. It may look contradicting that our data demonstrates a positive correlations between the relative labor productivity and the relative sales but a negative correlation between the relative markups and the relative sales. However, when firms are decreasing returns to scale, higher sales do not translate into higher profits. The positive correlation we see in Figure 1 may simply due to the fact that, holding labor expenditures constant, higher sales create higher labor productivity because the numerator increases while the denominator does not change.

FIGURE 1: Joint density of relative sales and relative labor revenue productivity



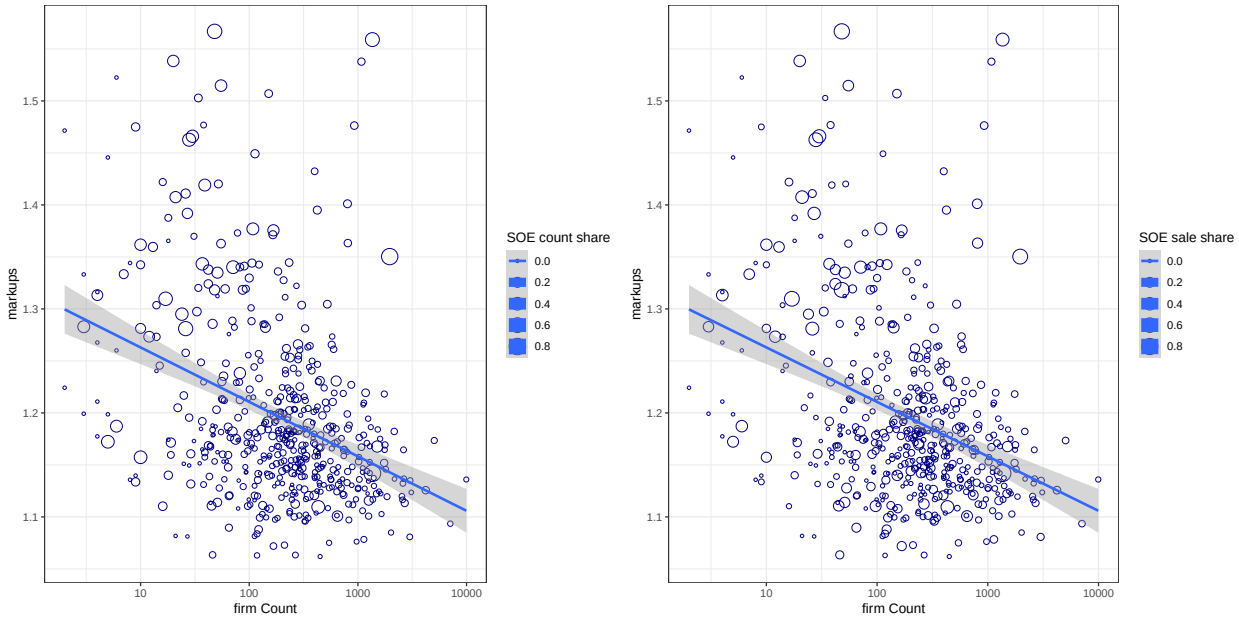
Notes: sales and relative labor revenue productivity are in their relative values, i.e. they are normalized industry-type average.

6.4 Markups and market concentration

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 2a 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 2a 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 2b). Besides, we compare our industry-level average markups to the Herfindahl indexes in Figure 3 and find that our markups increase with the Herfindahl indexes.

FIGURE 2: 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

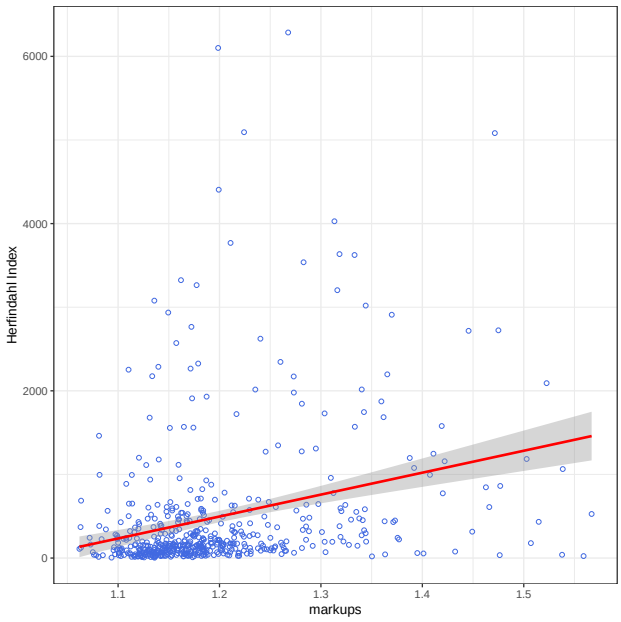
Notes: industry-level expected markups are cost-weighted average

6.5 TFP gains and income share changes

We use our estimated production elasticities and demand elasticities to show how the aggregate TFP gains change in HK's framework when relaxing its assumptions. To replace the production elasticities and demand elasticities in HK by our estimates, we use industry code to link two data sets. Cost is not provided in HK, so we can not link type-level estimates to their data. Instead, we estimate demand elasticities using our data as if only one type exists in an industry, i.e. no demand elasticities variation in any industry, and link them to HK's data. The estimators are less dispersed than but similar to those of our preferred model. Summary statistics of these estimators are provided in Appendix D.3.

As shown in Table 9, when we increase the demand elasticities in HK from 3 to 8.5, TFP gains jump from 87% to 362% and remains around 300% when introducing heterogeneity in demand elasticities. The number drops significantly once we introduce non-constant returns

FIGURE 3: Relation between industry-level expected markups, and the Herfindahl index



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

TABLE 9: Within-type TFP gains in China (2005) comparison across models

Data	α	σ	TFP gains (%)
HK	calibrated using US firms (HK)	3	86.6
HK	calibrated using US firms (HK)	8.5	362.3
HK	calibrated using US firms (HK)	heterogeneous (one type)	298.6
HK	Our estimators	3	51.5
HK	Our estimators	8.5	63.8
HK	Our estimators	heterogeneous (one type)	59.2
Our	Our estimators	8.5	46.3
Our	Our estimators	heterogeneous (two types)	43.9

to scale and is 59% when demand elasticities are estimated from micro data. Doing the same exercise using our data produces lower TFP gains but such pattern remains (Appendix F.2 provides a complete comparison between the two data and it demonstrates this pattern). Our data is a newer version of the ASM and the two data produce different results. Appendix F.1 compares both versions to the macro variables published in China Statistical Yearbooks. Our data matches the macro variables better than HK's.

TABLE 10: TFP gains in China (2005)

within industry (%)	across industry (%)	total (%)
43.9	4.7	50.6

Equalizing the marginal revenues of labor and capital generates 51% increase of aggregate TFP in China in 2005. Reallocation within types and across types raise aggregate TFP by 44% and 5% respectively. This reallocation involves large changes in type-level labor and capital usage and the changes differ across types. As shown in Table 11, more than half of the types reduce their capital and labor usage while some types' capital and labor are 10 and 7 times larger.

TABLE 11: Changes in type-level labor and capital

Statistic	Mean	Min	Pctl(25)	Median	Pctl(75)	Max
$\frac{l^*}{l}$	1.03	0.08	0.62	0.86	1.32	6.91
$\frac{k^*}{k}$	1.12	0.05	0.51	0.82	1.33	10.49

Reallocation causes aggregate labor and capital income share to increase by 6 percentage points (Table 12a). The observed labor income share is higher than capital income share and is predicted to have a larger increase. Labor income share increases from 20% to 27%, up by 7 percentage points while capital income share stays around 11% and drops by about 1 percentage point. The labor income share increases more and the capital income share increases when we keep all the other primitives but set the demand elasticities to be 8.5, the average of our estimated demand elasticities (Table 12b). The increase in the total labor and capital income share becomes 12 percentage points, with an increase of 11 percentage points and 0.1 percentage points respectively for labor and capital. The increase in aggregate labor and capital income share almost doubles under the homogeneous-markups case.

TFP gains does not change much under homogeneous markups. They are more sensitive to the level of average markups rather than the variations. Keeping all the other primitives the same as those estimated from our model, when demand elasticities are all equal to

TABLE 12: Labor and capital income share (%)

(A) Heterogeneous markups

	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) Homogeneous markups=8.5

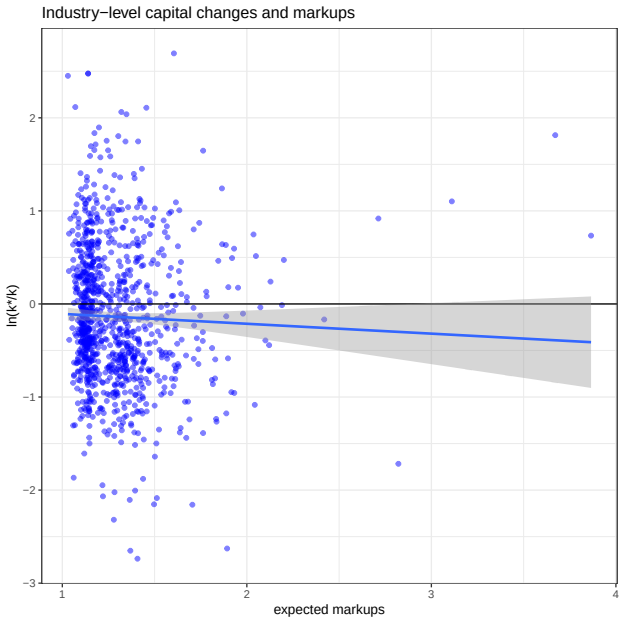
	observed	predicted	change
L	20.72	32.15	11.44
K	12.85	12.98	0.13
L+K	33.56	45.13	11.57

8.5, within-industry gains are 43.2%, less than 1 percentage point lower than those under heterogeneous markups. Taking into account the gains of reallocation across industries, TFP gains under homogeneous demand elasticities are 51.8%, about 1 percentage point higher than those under heterogeneous demand elasticities, meaning removing the variations in expected markups slightly increases aggregate TFP gains. However, if we not only remove the variations in demand elasticities and also reduce the demand elasticities to 3, within-industry TFP gains become 37.4% and the aggregate TFP gains are 44.2%. Notice, this is different from estimating our model parameters assuming demand elasticities to be 3 as the one shown in Table 22 in Appendix F.2 because there the primitives are different from the ones produced by our preferred model specification.

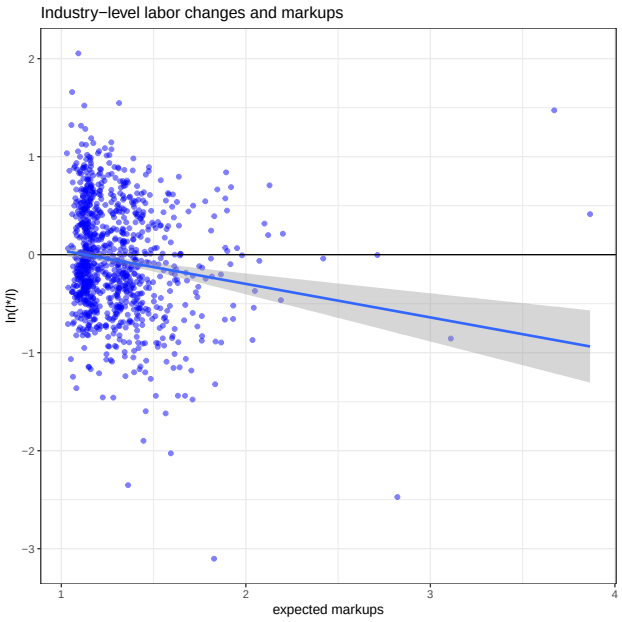
Labor income shares increase because the reallocation from high-markup to low-markup firms dominates the reallocation in the other direction. Since firms inside the same industry type have the same demand elasticities and the same production elasticities, we look at changes at the type level instead of the firm level. Although the pattern is similar in both figures, there is a significant negative correlation in Figure 4b but not in Figure 4a. The difference between the changes of capital and labor income share is then exaggerated by whether the increases or decreases in labor or capital usage take place in types with higher labor or capital production elasticities, because income shares are more sensitive to usage changes in types with larger production elasticities. Figure 4c and Figure 4d show higher production elasticities are usually associated with a larger increase in usage. It also demonstrates that most types dwell in the region where $\alpha_K < 0.3$ whereas they are more spread out for α_L. Therefore, labor income share is more sensitive to increases in labor usage which amplifies the negative correlation we observed in Figure 4b and creates a larger increase in labor income share.

FIGURE 4: Changes in capital and labor usage for industry types with different markups and production elasticities

(A) K vs. markups

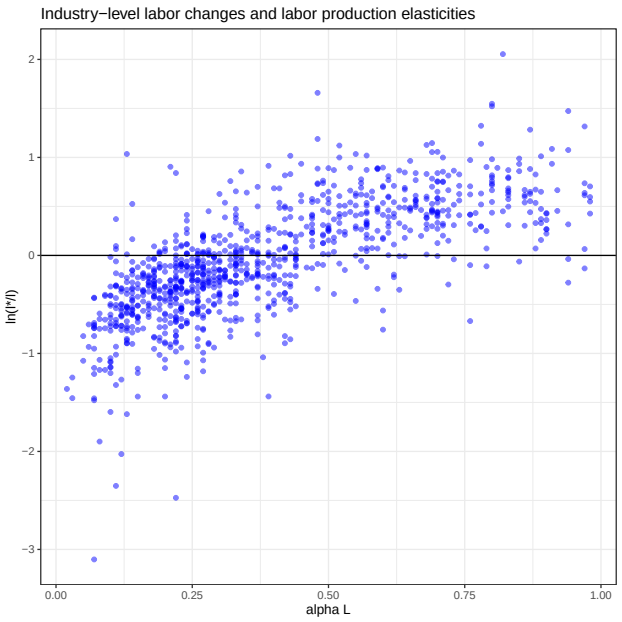


(B) L vs. markups



(C) K vs. α_K

(D) L vs. α_L



6.6 Robustness checks

This subsection checks whether our main results are sensitive to possible biases in our estimated demand elasticities and labor production elasticities. The underlying assumption here is the average demand elasticities and labor production elasticities obtained in our reduced form analysis are close enough to the true values. The first three rows of Table 13 scale up our estimated demand elasticities so that the average equals 12.90, the demand elasticities obtained from our reduced form analysis. We carry out the scaling for three types of average: value-added based average, sales-based average, and cost-based average. The average of our estimators before the scaling is reported in the fifth column. The fourth till sixth row scale up our estimated labor production elasticities so that the average is 0.5, the value imposed by HK. The seventh till ninth row scale them to 0.44, the labor production elasticities from our reduced form analysis while the last three rows replace the labor expenditure in our reduced form analysis by the number of employees.

The last column reports the predicted within-industry TFP gains. Unlike in Table 9, our predicted TFP gains are fairly stable, ranging between 43.3% and 56.7%. Therefore, we believe the possible biases in our estimators are probably not crucial for our main results.

Table 13: Within-type TFP gains in China (2005): robustness analysis

var of interest	target	target source	mean type	unscaled mean	TFP gains (%)
σ	12.90	RF main	va-based	8.48	46.55
σ	12.90	RF main	sales-based	9.07	46.63
σ	12.90	RF main	cost-based	9.37	46.72
α_L	0.50	HK's guess	va-based	0.32	48.40
α_L	0.50	HK's guess	sales-based	0.32	43.29
α_L	0.50	HK's guess	cost-based	0.32	43.81
α_L	0.44	RF main	va-based	0.32	55.72
α_L	0.44	RF main	sales-based	0.32	56.72
α_L	0.44	RF main	cost-based	0.32	53.29
α_L	0.46	RF using L	va-based	0.32	48.80
α_L	0.46	RF using L	sales-based	0.32	48.06
α_L	0.46	RF using L	cost-based	0.32	48.12

Notes: var of interest indicates on which variables the robustness analysis is carried out.
RF main means estimates from reduced form analysis.
RF using L is the reduced form estimates using the number of employees.
unscaled mean is the mean from structural estimates.

7 Conclusion

Measuring the TFP costs of misallocation has generated great interest, especially following Hsieh and Klenow (2009) (hereafter HK), but less attention is paid to the impact of misallocation on income shares. We first document the sensitivity of the results in HK to the assumed markups and constant returns to scale. We then propose a model where its parameters can be estimated using micro data. Our model is robust to several types of misspecification and to measurement errors caused by varying demand elasticities and non-constant returns to scale. Markup estimates from our model are consistent with the existing literature and with indicators of industry concentration. Using our model, we find that the gains in the labor share are dampened when markups are allowed to be heterogeneous, which suggests that distributional and welfare concerns come with the efficiency gains from equalizing the marginal revenues. The business environment for SOEs seems to be different from that of domestic private firms due to a systematic difference in capital and labor distortions. However, the variation within either ownership types is so large that some domestic private firms' labor and capital usage very much resembles a typical SOE and vice versa. Research that treats SOEs differently from domestic private firms may need to be careful with the highly noisy ownership labels observed.

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.

Baqaee, D., E. Farhi, and K. Sangani (2020): "The Darwinian Returns to Scale," Working Paper 27139, National Bureau of Economic Research.

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.

Basu, S. and J. G. Fernald (1997): "Returns to Scale in U.S. Production: Estimates and Implications," *Journal of Political Economy*, 105, 249–283.

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.

Bernard, A. B., J. Eaton, J. B. Jensen, and S. Kortum (2003): "Plants and Productivity in International Trade," *American Economic Review*, 93, 1268–1290.

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

Bils, M., P. J. Klenow, and C. Ruane (2020): "Misallocation or Mismeasurement?" Working Paper 26711, National Bureau of Economic Research.

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.

——— (2014): "Challenges of Working with the Chinese NBS Firm-Level Data," *China Economic Review*, 30, 339–352.

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.

Chirinko, R. S. and S. M. Fazzari (1994): "Economic Fluctuations, Market Power, and Returns to Scale: Evidence from Firm-Level Data," *Journal of Applied Econometrics*, 9, 47–69.

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.

Dhingra, S. and J. Morrow (2019): "Monopolistic Competition and Optimum Product Diversity under Firm Heterogeneity," *Journal of Political Economy*, 127, 196–232.

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*.

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.

Gao, W. and M. Kehrig (2016): "Returns to Scale, Productivity and Competition: Empirical Evidence from U.S. Manufacturing and Construction Establishments," *SSRN Journal*.

Gupta, A. (2021): "Demand for Quality, Variable Markups and Misallocation: Evidence from India," *working paper*.

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.

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

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.

Lafortune, J., E. G. Lewis, J. Pablo Martínez, and J. Tessada (2021): "Changing Returns to Scale in Manufacturing 1880-1930: The Rise of (Skilled) Labor?" NBER working paper 28633.

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

Liang, Y. (2021): "Misallocations and Markups: Evidence from Indian Manufacturing," *working paper*, 59.

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.

Melitz, M. J. (2003): "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 71, 1695–1725.

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.

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.

Ruzic, D. and S.-J. Ho (2021): "Returns to Scale, Productivity Measurement, and Trends in U.S. Manufacturing Misallocation," *The Review of Economics and Statistics*, 1–47.

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

Appendix

A Data

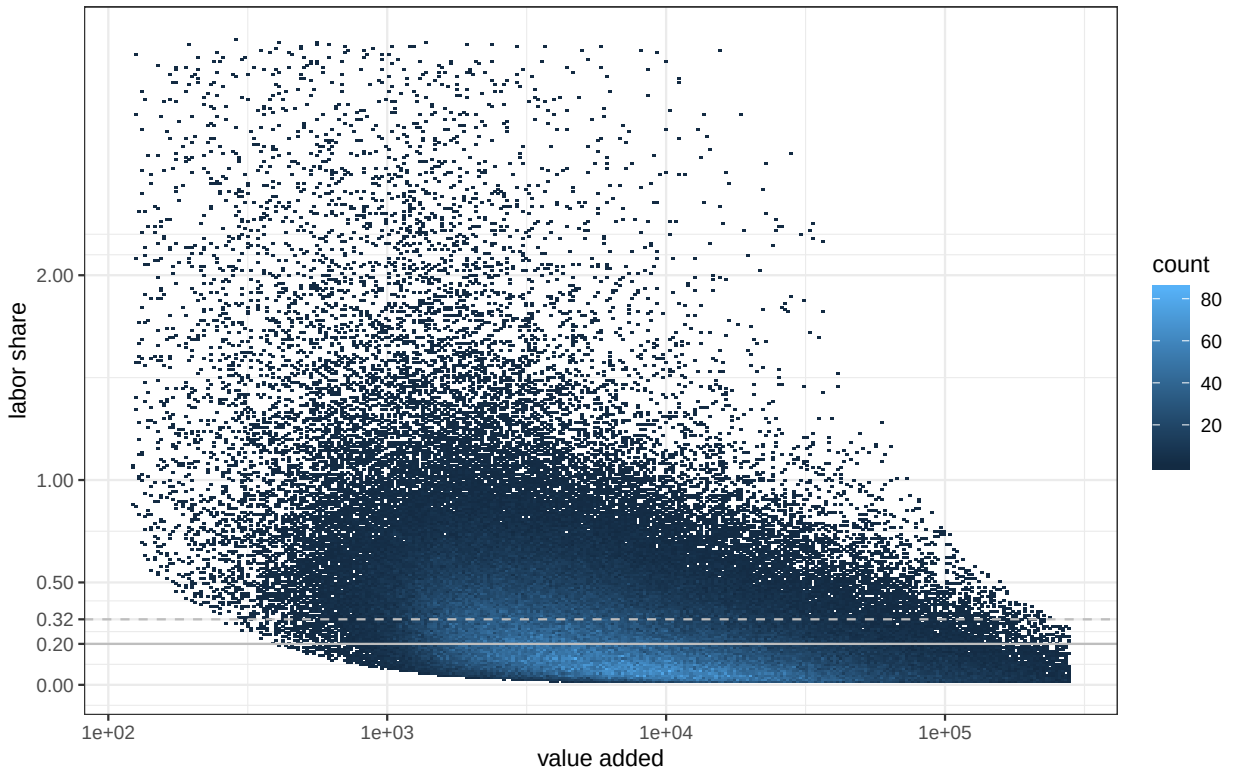
We drop unreasonable observations accounting-wise including observations with negative value added, negative wage expenditure, negative capital, negative total asset, negative account receivable, negative total debt, negative long-term debt, negative account payable, negative export, negative sales, 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 14: 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 5 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 5: Joint distribution of labor income share and firm sizes (2005)



B Reduced form analysis of returns to scale

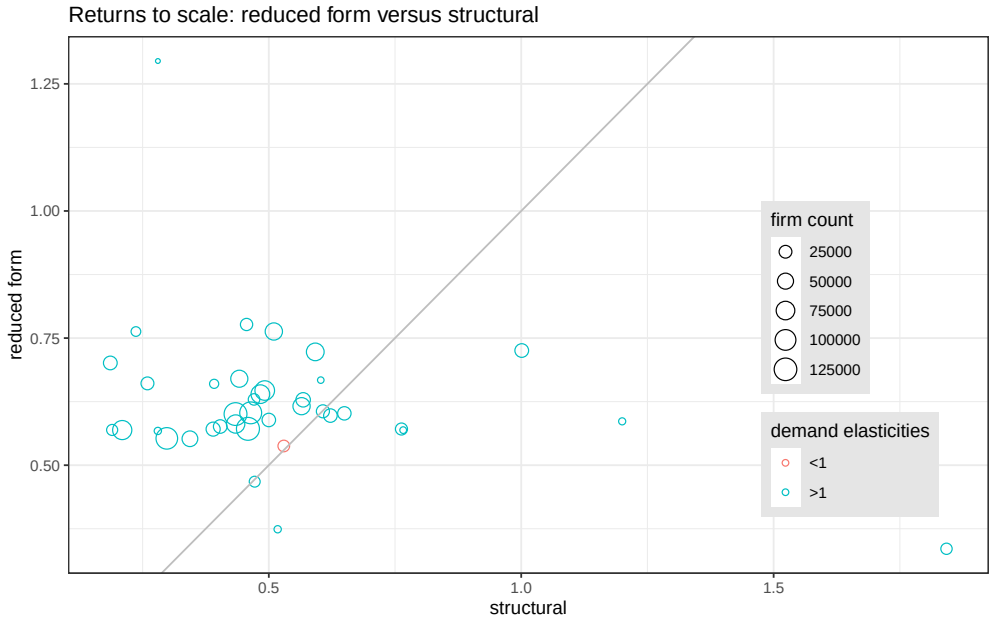
In this section, we follow Klette and Griliches (1996) more closerly by estimating Equation (2) in Section 3 for each 2-digit industry as Klette and Griliches (1996) also estimates it industry by industry. There are in total 38 2-digit industries in our data but we may need to be careful interpreting the results for smaller industries. The reduced-form analysis requires at least three consecutive observations for each firm but the bottom 10% industries contain less than 2500 observations and the smallest industry has only 133 observations (15).

TABLE 15: Sample sizes of 2-digit industries

Statistic	N	Mean	Min	Pctl(25)	Median	Pctl(75)	Max
firm count	38	40,778	133	15,792.5	29,957.5	61,011.8	130,007

We compare our reduced-form estimated returns to scale to those from our structural model. The structural ones are the average of returns to scale across firms in a 2-digit industry. For most larger industries, the two estimators are close or even the same (Figure 6). Those industries' whose reduced-form estimators differ a lot from their structural estimators

FIGURE 6: Returns to scale: 2-digit industries with a sample size above the first quartile



are all very small industries. However, one may need to be cautious in applying this analysis to all the industries as the reduced-form demand elasticities for 1 industry is below 1. This may be due to some endogeneity not dealt with by our instruments, or it can be model misspecification. The CES demand assumed in Klette and Griliches (1996) requires firms with higher prices have lower market shares. However, if the demand causes that firms charging higher prices have a higher production share, we may observe positive β_1 and β_2 but negative β_3. This coincides with what happens for this industry. Its β_3 is negative, and β_1 and β_2 are positive.

C Derivation of 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, :

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

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:

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

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

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:

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

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:

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

C.1 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 4

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

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

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,

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

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 (3) in Section 5.2, 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 16, Table 17, and Table 18 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 (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

TABLE 16: 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
	EM	mixtools	0.4683282	0.2859527	1.591074	0.8251734	1.306893	329.2560831
	BFGS	optim	0.4679615	0.2856622	1.590424	0.8249206	1.306958	329.2560826
	NM	optim	0.5325401	1.589622	0.2852179	1.306917	0.8245313	329.2560842
(0.25,1,1,1,1.5)	SANN	optim	0.5399907	1.587284	0.2726049	1.308941	0.8179273	329.2592109
	NR	maxLik	0.4679779	0.285673	1.590456	0.8249318	1.306953	329.2560826
	BHHH	maxLik	0.4679779	0.285673	1.590456	0.8249318	1.306953	329.2560826
	EM	mixtools	0.1047498	-0.5673427	3.38516	0.4612274	2.161369	450.5243907
	BFGS	optim	0.8952547	3.385141	-0.5673413	2.161387	0.4612123	450.5243907
	NM	optim	0.8951295	3.385244	-0.5677538	2.160873	0.4611698	450.5244077
(0.25,0,4,1,2)	SANN	optim	0.1026536	-0.5735226	3.373923	0.4525442	2.161418	450.5287012
	NR	maxLik	0.1047467	-0.5673531	3.385145	0.4612135	2.161376	450.5243907
	BHHH	maxLik	0.1047467	-0.5673531	3.385145	0.4612135	2.161376	450.5243907

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

TABLE 17: 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$*e8$
	EM	mixtools	1.38e-04	1.11e-04	2.44e-04	9.55e-05	2.36e-05	0.00e+00
	BFGS	optim	1.66e-01	3.99e+00	2.15e-01	3.73e+00	1.08e-02	1.53e+08
	NM	optim	1.21e-01	7.88e-01	5.02e-01	3.10e-01	1.70e-01	9.26e+07
(0.25,1,1,1,1.5)	SANN	optim	1.28e-01	6.81e-01	4.53e-01	5.10e-01	1.56e-01	9.26e+07
	NR	maxLik	5.17e-02	6.89e+01	2.05e-01	1.61e+01	7.92e-01	5.53e+08
	BHHH	maxLik	2.20e-01	1.57e+01	4.41e-01	3.01e+02	8.63e-02	1.42e+09
	EM	mixtools	2.88e-07	9.57e-07	1.35e-06	1.29e-06	6.99e-07	5.01e+00
	BFGS	optim	4.53e-02	1.21e+01	1.86e-01	1.12e+01	6.79e-01	3.03e+08
	NM	optim	9.74e-02	1.87e+00	4.29e-01	9.58e-01	1.31e-01	2.52e+08
(0.25,0,4,1,2)	SANN	optim	3.11e-02	1.47e+00	1.64e-01	1.33e+00	7.42e-02	2.01e+08
	NR	maxLik	2.21e-01	2.48e+02	3.06e-01	8.58e+01	3.01e-01	2.27e+08
	BHHH	maxLik	1.25e-01	4.33e+02	7.76e-01	1.70e+02	1.97e-01	1.59e+09

Normalization: $p < 1 - p$

TABLE 18: 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

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 19 provides the summary statistics under this specification.

TABLE 19: 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 Identify production elasticities and distortions

From Equation (5), when α_s^K and α_s^L are fixed, estimator of the remaining parameters are:

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

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

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

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

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

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

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

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

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

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

F Comparison to HK

HK requires two key assumption: demand elasticities equals 3 and constant returns to scale. In this section, we relax these assumptions one by one to show how TFP gains react. Before doing this exercise, we first investigate the difference between our data and HK's. The evidence favors using our data.

F.1 Data versions

Both HK and we use the ASM data but ours is a newer version acquired via a data center at Peking University. Table 20 and Table 21 show how much the aggregates of the two ASM data deviate from the counterpart macro variables published in China Statistical Yearbooks (CSY) reported as percentage shares of those variables in CSY. We report 1998-2008 for our data because other years are not used in this paper. HK only have 1998-2005 so Table 21 only reports these years. The differences between our data and CSY are mostly around or below 2% while those between HK's data and CSY are around $10-20\%$. Our data contains around $0.05-0.1\%$ more firms than CSY in each year except for 2004 and 2008 while HK's data contains around 20% less firms in 1998-2002 and around 10% less in 2003-2005.

TABLE 20: My data statistics in comparison with China Statistical Yearbook: ratio (%)

Year	Number of firms	Sales	Output	Value added	Employment	Net value of fixed assets	Export	profits
1998	0.05	0.41	0.38	0.41	-8.56	1.48	0.58	-2.76
1999	0.04	0.94	1.02	0.92	0.46	-2.21	1.19	0.20
2000	0.06	0.54	0.51	0.45	0.39	-1.31	0.11	0.09
2001	0.06	0.89	1.26	1.14	0.54	-1.50	0.81	1.91
2002	0.07	0.84	0.83	0.83	0.37	-1.57	0.16	0.64
2003	0.10	1.80	1.78	1.88	1.00	-1.41	1.59	2.32
2004	-0.54	0.78	0.74	5.20	0.98	-2.72	1.06	1.95
2005	0.09	1.24	1.22	1.30	1.14	-2.76	1.17	1.39
2006	0.12	1.38	1.18	1.12	0.64	-3.01	3.05	1.23
2007	0.13	1.95	1.64	2.14	1.52	-3.06	1.96	3.48
2008	-3.30	-0.74	-1.40		-2.74	-6.02	-0.46	-1.28

Notes: all the variables are from from the latest available yearbook issue.
Export data of China Statistical Yearbook is from Brandt et al. (2014).

F.2 Relax assumptions imposed in HK

The first three rows of Table 22 relax the assumption of demand elasticities σ. The third row uses our estimated industry-level σ by matching our estimated σ to the relevant industry in HK using the four-digit industry code. We don't allow demand elasticities to differ within four-digit industries here because this requires cost and sales data which is not available in

TABLE 21: HK's data statistics in comparison with China Statistical Yearbook: ratio (%)

Year	Number of firms	Sales	Value added	Employment	Net value of fixed assets	Export
1998	-27.87	-14.74	-20.19	-23.24	-16.69	-19.12
1999	-26.09	-12.22	-18.62	-19.84	-15.23	-14.09
2000	-23.14	-9.04	-12.81	-20.92	-2.53	-10.47
2001	-22.17	-10.44	-13.85	-19.17	-2.57	-11.41
2002	-19.00	-8.19	-11.54	-14.64	0.15	-9.13
2003	-14.15	-5.95	-5.96	-10.24	2.14	-6.34
2004	-9.44	-4.95	-10.34	33.13	-2.94	
2005	-8.40	-4.63	-12.66	-6.19	-2.74	-4.75

Notes: all the variables are from from the latest available yearbook issue.
Export data of China Statistical Yearbook is from Brandt et al. (2014).

TABLE 22: Within-type TFP gains in China (2005) comparison across models (a complete version)

Data	α	σ	TFP gains (%)
HK	calibrated using US firms (HK)	3	86.6
HK	calibrated using US firms (HK)	8.5	362.3
HK	calibrated using US firms (HK)	heterogeneous (one type)	298.6
HK	Our estimators	3	51.5
HK	Our estimators	8.5	63.8
HK	Our estimators	heterogeneous (one type)	59.2
Our	calibrated using US firms (HK)	3	116.1
Our	calibrated using US firms (HK)	8.5	419.9
Our	calibrated using US firms (HK)	heterogeneous (one type)	349
Our	calibrated using US firms (HK)	heterogeneous (two types)	358.7
Our	Our estimators	3	37.9
Our	Our estimators	8.5	46.3
Our	Our estimators	heterogeneous (one type)	49.4
Our	Our estimators	heterogeneous (two types)	43.9

HK. As shown in the highlighted first row, imposing the assumptions in HK to our model and using its data give its result. The second row sets σ to 8.4, the average of our estimated demand elasticities. The third row introduces heterogeneous demand elasticities. TFP gains quadruple when σ increases from 3 to 8.4 and then decline slightly but are still around 300% when moving to heterogeneous demand elasticities.

The fourth till sixth rows relax the assumption of constant returns to scale by replacing HK's production elasticities with my estimated production elasticities. The fourth row maintains the assumption of $\sigma = 3$ while the sixth row allows heterogeneous markups. When allowing non-constant returns to scale, TFP gains drop sharply to 59% and 74% respectively for $\sigma = 3$ and for the heterogeneous σ.

The second part of the table uses our data. Row 7 applies HK's method to our data and present higher TFP gains, 116%. Row 8-10 maintain constant returns to scale but employ different demand elasticities. To facilitate easier comparison to the first part of the table, we also include the scenarios where demand elasticities are constant in industries. Similar to the cases of using HK's data, TFP gains increase significantly when using our estimated demand elasticities. Once constant returns to scale is relaxed, predicted TFP gains drop from more than 350% to around 57%.

G Apply the analysis to the year 2001 for robustness check

In the main results, we use the year 2005 to show how our structural model works, but our model can be easily applied to other years. In this section we pick the year 2001 to demonstrate the robustness of our results. This year is picked because HK also reports its results for the year 2001. We do not do our analysis for 1998 which is also reported in HK because 1998 does not provide enough data needed for our structural analysis.

TABLE 23: Ex-ante probability of belonging to $\bar{s}$ in industries with mixture distribution (2001)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
prob of belonging to $\bar{s}$	453	0.71	0.18	0.03	0.62	0.74	0.84	0.99

Similar to the year 2005, most firms belongs to the low-markup group (Table 23) and industries containing two types also tend to be larger than those only containing one type (Table 24). Table 25 show the summary statistics of our structural estimators. The estima-

TABLE 24: Summary statistics on industry sizes (2001)

from an s with mixture	Number of industries	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
No	131	2	7.500	19	26.779	32.500	356
Yes	453	10	66	154	282.362	329	5,984

tors are also similar to those for 2005 except that the estimated average scale is 0.74, higher than that of 2005.

TABLE 25: Type-level summary statistics of estimates allowing types inside industries (2001)

	N	Mean	St. Dev.	Pctl(10)	Pctl(25)	Median	Pctl(75)	Pctl(90)
$\mathbb{E}_g[\mu_i + 1]$	1,036	1.31	0.26	1.10	1.14	1.22	1.40	1.58
σ_g	1,036	6.47	7.02	2.74	3.53	5.48	8.25	10.96
$\mathbb{E}_g[e_i^\delta]$	1,036	1.01	0.02	1	1	1.01	1.02	1.03
α_K	552	0.29	0.24	0.06	0.10	0.20	0.39	0.67
α_L	552	0.46	0.25	0.15	0.24	0.40	0.68	0.83
scale	552	0.74	0.37	0.31	0.44	0.70	0.98	1.24

Within-industry TFP gains are 70%, lower than the 96% by HK (Table 26). When taking into account gains from reallocating across industries, it becomes 80%. We also see a larger increase of labor income share and TFP gains when removing variations in expected markups, as shown in Table 27, Table 28, and Table 29.

TABLE 26: TFP gains in China (2001)

within industry (%)	across industry (%)	total (%)
66.9	8.0	80.3

TABLE 27: Labor and capital income share (%) (2001)

	observed	predicted	change
L	22.31	36.03	13.72
K	20.89	20.94	0.05
L+K	43.2	56.97	13.77

Table 30 present the same pattern we see before whether SOEs tend to have small τ_K and τ_L, meaning they tend to use too much capital and labor compared to the allocation when

TABLE 28: Labor and capital income share: $\sigma = 7.9$ (%) (2001)

	observed	predicted	change
L	24.15	43.26	19.11
K	22.28	25.44	3.16
L+K	46.43	68.70	22.27

TABLE 29: TFP gains in China, $\sigma = 7.9$ (2001)

within industry (%)	across industry (%)	total (%)
71.1	10.8	89.6

equalizing the marginal revenues of capital and labor. However, there are large variations within both ownership. Some domestic private firms appear to behave like an SOE and vice versa.

TABLE 30: Estimated distortions for different firm types (2001)

	firm type	N	Mean	Min	Pctl(25)	Median	Pctl(75)	Max
τ_K	domestic priv	80308	2.35	-1.00	-0.38	0.39	2.14	366.10
	SOE	19025	0.58	-0.99	-0.71	-0.35	0.43	171.22
	all	99333	2.01	-1.00	-0.48	0.20	1.77	366.10
τ_L	domestic priv	80308	1.44	-0.98	-0.23	0.42	1.67	65.87
	SOE	19025	0.51	-0.97	-0.48	-0.06	0.65	57.82
	all	99333	1.26	-0.98	-0.29	0.30	1.46	65.87

H Identification issue of correcting the biases in inferred markups

H.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:

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

When there is one type, its distribution is:

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

when there are two types, its distribution is:

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

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

In the third step, we use these equations:

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

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

$$\widehat{\Xi} \equiv \frac{\hat{\epsilon_s}}{\hat{\epsilon_s} - 1}\frac{1}{\hat{\alpha}_s^L + \hat{\alpha}_s^K} \tag{6}$$
$$\widehat{\Xi}^L \equiv \frac{\hat{\epsilon_s}}{\hat{\epsilon_s} - 1}\frac{1}{\hat{\alpha}_s^L} \tag{7}$$
$$\widehat{\Xi}^K \equiv \frac{\hat{\epsilon_s}}{\hat{\epsilon_s} - 1}\frac{1}{\hat{\alpha}_s^K} \tag{8}$$

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 (8) and Equation (7), 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 (6), 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 (6), Equation (8) and Equation (7) 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.

H.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:

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

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

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

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 $= r$MC. Therefore,

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

I 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 4.

J Robustness

J.1 Use the number of employees instead of labor expenditure

Reduced form analysis:

Average returns to scale is 0.66 ((0.41+0.17)/(1-0.11)) and average demand elasticities are 8.80 (1/0.11).

TABLE 31: Reduced-form estimation of returns to scale when using the number of employees

	Dependent varaible	
	r_{it}	
	OLS	IV
	(1)	(2)
l_{it}	0.322^{***}	0.410^{***}
	(0.002)	(0.003)
k_{it}	0.151^{***}	0.173^{***}
	(0.001)	(0.002)
y_{st}	0.130^{***}	0.114^{***}
	(0.003)	(0.003)
constant	0.117^{***}	0.080^{***}
	(0.001)	(0.001)
Observations	1,186,861	819,923
R^2	0.056	0.042
Adjusted R^2	0.056	0.042
Residual Std. Error	0.686 (df = 1186857)	0.635 (df = 819919)
F Statistic	$23{,}537.110^{***}$ (df = 3; 1186857)	

Note: *p<0.1; **p<0.05; ***p<0.01

r_{it} is deflated firm-level value added, VA_s is industry s's aggregate VA.