Unraveling Firms: Demand, Productivity and Markups Heterogeneity[*]

Emanuele Forlani[†] Ralf Martin[‡] Giordano Mion[§] Mirabelle Muûls[¶]

First draft: 16th June 2015. This draft: May 20, 2016.

Abstract

We develop a new econometric framework that simultaneously allows recovering heterogeneity in demand, TFP and markups across firms while leaving the correlation among the three unrestricted. We accomplish this by explicitly introducing demand heterogeneity and by systematically exploiting assumptions that are implicit in previous firm-level productivity estimation approaches. We use Belgian firms production data to quantify TFP, demand and markups and show how they are correlated with each other, across time and with measures obtained from other approaches. We also show to what extent our three dimensions of heterogeneity allow us to gain deeper and sharper insights on two key firm-level outcomes: export status and size.

Keywords: Demand, Productivity, Markups, Production function estimation, Export status, Firm size
JEL Classification: D24, L11, L25, F14

[*] We thank Andrea Ariu, Andrew Bernard, Ingo Borchert, Rosario Crino, Jan De Loecker, Catherine Fuss, Marc Melitz, Luca David Opromolla, Chad Syverson, Ilke Van Beveren and John Van Reenen as well as participants at various seminars and workshops and an anonymous referee from the National Bank of Belgium. Most of this research has been carried out while Giordano Mion was affiliated to the National Bank of Belgium. The views expressed in this paper are our own and do not necessarily reflect those of the National Bank of Belgium

[†] Department of Economics and Management, University of Pavia, Pavia, Italy; emanuele.forlani@eco.unipv.it; website: https://sites.google.com/site/emanueleforlani79/home. Also affiliated with the Centro Studi Luca d'Agliano, Milan, Italy.

[‡] Imperial College Business School, Imperial College, London, UK; r.martin@imperial.ac.uk; website: http://www.imperial.ac.uk/people/r.martin. Also affiliated with the CEP, London, UK; and the Grantham Institute, London, UK.

[§] Department of Economics, University of Sussex, Falmer, UK; g.mion@sussex.ac.uk; website: https://sites.google.com/site/giordanomionhp/. Also affiliated with the CEP, London, UK; CEPR, London, UK; and CESifo, Munich, Germany.

[¶] Imperial College Business School and Grantham Institute, Imperial College, London, UK; m.muuls@imperial.ac.uk; website: http://www.imperial.ac.uk/people/m.muuls. Also affiliated with the CEP, London, UK.

1 Introduction

Economists are interested in estimating firm-level productivity in a range of fields. These estimates are often used as inputs in a number of applications such as the firm size distribution, firm survival and growth, self-selection of firms into export status and the extensive and intensive margins of trade to name a few. In the literature, the most commonly used approach to estimate productivity involves estimating a production function by regressing output quantity on input quantity and using the resulting residual shock as a productivity index typically referred to as Total Factor Productivity (TFP). This raises at least three issues.[1]

First, most studies do not have output quantity data available at the firm-level so that regressions are fitted using revenue data, i.e., price times quantity. Such revenue-based measures of TFP look quite different from quantity-based ones (Foster et al., 2008). Second, a well known issue is the endogeneity of production factors used as explanatory variables (Olley and Pakes, 1996). Third and more importantly, firms could be heterogeneous in dimensions other than technical efficiency. In this respect the IO literature on demand systems (Ackerberg et al., 2007) points to substantial heterogeneity in markups and consumers' willingness to pay for the products sold by different firms. For example, the presence of vertical and horizontal product differentiation means that firms selling otherwise similar products face rather different demands. At the same time market power variations, due to product quality or technical efficiency, could substantially affect the markup that firms can charge. Moreover, markups could also vary because of factors unrelated to either quality or efficiency; e.g. some firms could have a better market access than others because of spatial differentiation and trade costs. Being able to account for these different dimensions and their interconnections is important for several reasons.

First of all it is crucial in order to correctly measure TFP. In this respect higher measured TFP is typically seen as welfare improving. However, conventional measures of TFP conflate actual TFP with demand and markups heterogeneity which may lead to different welfare implications. Second, being able to actually quantify dimensions other than TFP matters from both a welfare and a policy point of view. From a welfare perspective it is, for example, of great value to assess the impact on firm markups of a trade integration episode or

[1] There are at least two other important issues related to TFP estimation. The first one is that many firms are actually multi-product and so the problem of how to assign inputs to specific outputs needs to be addressed. In our empirical analysis we focus, for better comparability with previous studies, on single-product firms. However, in deriving our theoretical framework, we provide a multi-product firms extension that could be readily implemented. The second issue is that input quantity data at the firm-level is typically not available and input expenditure is used instead. We do not deal here with this second issue. Future research will expand in this direction. See De Loecker et al. (2016) for a joint treatment of input price bias and multi-product firms. See also Atalay (2014) for a quantification of the input price bias and Grieco et al. (2014) for a parsimonious methodology to deal with such a bias.

market size expansion. Some recent theoretical papers have indeed revisited the relationship between market size, markups and welfare and questioned the pervasiveness of the so called "pro-competitive effects" (Dhingra and Morrow, 2012 and Zhelobodko et al., 2012). Furthermore, being able to disentangle quality from efficiency is important for policy matters and in particular to understand where the competitiveness of a firm or an industry comes from and then target interventions accordingly.

This paper's contribution is to address these issues in a more comprehensive way than the existing literature. In doing so we combine some key elements of the productivity estimation and demand system literatures and propose a framework that simultaneously allows recovering productivity, demand, and markups heterogeneity across firms, while leaving the correlation among the three unrestricted. Our framework is rich enough to allow for multi-product firms, alternative hypotheses on preferences and market structure as well as on the production function and processes for productivity and demand. It is also amendable to include endogenous firm actions affecting TFP like the choice of its hierarchical structure as outlined in Caliendo et al. (2015). At the same time, our framework is parsimonious enough to allow retrieving productivity, demand, and markups heterogeneity with relatively little information compared to demand systems models.[2] These features provide a wide scope of applications to our framework.

In line with the demand systems literature we model demand heterogeneity as shocks potentially correlated with TFP shocks and shifting demand in a way that is complementary to heterogeneity in markups. Within our framework such shocks can be interpreted as a measure of quality of a firm's products and so in what follows we refer to demand heterogeneity, demand shocks and quality interchangeably. By simultaneously dealing with productivity, demand, and markup heterogeneity we improve upon existing productivity studies that either ignore one or several dimensions of heterogeneity or address some of the issues but in a fairly restrictive way. For example, Klette and Griliches (1996) recover TFP while having some demand shocks in the background but impose homogenous markups across firms. Other approaches, such as De Loecker and Warzynski (2012), De Loecker et al. (2016), and Dobbelaere and Mairesse (2013) estimate both TFP and markups but, again, keep demand shocks in the background. More specifically, demand shocks are treated as something to be simply controlled for by means of (often unavailable and restrictive) proxies rather than something to be quantified. Papers such as Foster et al. (2008) measure both TFP and demand shocks but impose constant markups and zero correlation between demand and productivity shocks.

[2] Demand system models have very rich structures and allow for consumer and product specific elasticities of demand. However, they require detailed information on product and consumer characteristics as well as suitable instruments (like cost shifters) for identification. See, for example, Berry et al. (2004) and Roberts et al. (2016) The high data requirements of these models are such that their application is usually limited to specific industries and contexts. By contrast, our simpler and more parsimonious framework only requires information on product prices, quantities and inputs and does not need any additional instrument.

We apply our econometric framework to Belgian manufacturing firms and use information on both the quantity and the value of production[3] over the period 1996-2007 to quantify the baseline version of our model on single-product firms for four selected industries.[4] We first document that demand shocks display at least as much variability across firms as productivity shocks. We further show that productivity and demand shocks are very strongly and negatively correlated in each of the four industries we consider. This finding is suggestive of a trade-off between the quality of a firm's products and their production cost as suggested in Ackerberg et al. (2007). Consider, for example, the car industry where there is the coexistence of manufacturers (like Nissan) producing many cars for a given amount of inputs (high productivity) and manufacturers (like Mercedes) producing less cars for a given amount of inputs (low productivity). To be more specific one of the most productive car plants in Europe is the Nissan factory located in Sunderland in the UK. In terms of sheer productivity measured as cars per employee it is nearly 100% more productive than a state of the art Mercedes plant near Rastatt in Germany. However, this hardly reflects a problem with the Mercedes plant. Rather, Mercedes and Nissan face quite different demands which leads to different prices as well as different markups. Both plants are profitable and perhaps generate a very similar revenue-based productivity. Yet, their business model is rather different.

Another pattern worth noting is that differences in markups across firms are reasonably well explained (in terms of R^2) by differences in demand and productivity. However, there remains a considerable amount of unexplained variation. Our framework allows for markups that are either entirely determined by productivity and demand shocks or are entirely independent from these other shocks - or a combination of the previous two cases. In this respect our findings suggest that, although closely related to demand and productivity, markups are far from being a residual dimension of heterogeneity. We also show how revenue-based productivity can be decomposed into the three dimensions of heterogeneity and find that variation in revenue TFP is actually attributable mainly to variation in demand and markups across firms rather than in quantity TFP. Demand shocks typically explain more variation than markups.

We finally assess how and to what extent our three dimensions of heterogeneity allow us to gain deeper and sharper insights into two key firm-level outcomes: export status and size. We start by showing that the usual positive correlation between revenue-based TFP and export status holds in our data. The availability of physical quantity data allows us to

[3] When only revenue data is available one can still identify markups as well as a composite of demand and TFP shocks. With firm-level output data one can, in addition, distinguish between TFP and demand shocks. See Martin (2014).

[4] We choose not to analyse multi-product firms in this paper and focus on single-product firms to allow a direct comparison with other approaches and so better highlight our methodological contribution. We could have used the model developed in Appendix E to estimate demand, TFP and markups for all Belgian manufacturing firms but in doing so we would not have had the chance to compare our analysis and results to, for example, De Loecker and Warzynski (2012), Foster et al. (2008) and Olley and Pakes (1996). Given our focus on single-product firms we select those industries populated by many such firms

also look at the correlation between quantity-based TFP and export status. We find this correlation to be positive, providing support to the mainstream theoretical framework based on differences across firms in term of their ability to turn inputs into output. Yet, within our framework we can go even further and ask whether - and how - demand and markups heterogeneity also matters and how it interacts with heterogeneity in productivity. First, we confirm that the positive correlation between quantity-based TFP and firm export status is robust to the inclusion of demand and markups heterogeneity. Second, we find that demand is more important than productivity in drawing the line between exporting and non-exporting firms. We also find that exporting firms typically sell higher quality goods and charge lower markups. When considering firm size, as measured by the number of employees, we show that the positive correlation between revenue-based TFP and size commonly found in the literature equally holds in our data. We also find that the positive correlation between quantity-based TFP and size is robust to including demand shocks and markups heterogeneity. However, demand and markups heterogeneity are as important as productivity in understanding why some firms are larger than others.

Our paper is related to the literature on firm TFP measurement on which Olley and Pakes (1996) has had a deep impact. The key endogeneity issue addressed in Olley and Pakes (1996) is omitted variables: the firm observes and takes decisions based on productivity shocks that are unobservable to the econometrician. Yet, the econometrician observes firm decisions (investments) that do not impact productivity today and that can (under certain conditions) be used as a proxy for productivity shocks. This proxy variable approach to tackle the issue of unobservable productivity shocks has been further developed in Levinsohn and Petrin (2003) and Ackerberg et al. (2016) and represents the current dominant framework.

In a recent paper De Loecker and Warzynski (2012) have extended this framework by explicitly allowing for another dimension of heterogeneity in the model, namely firm-specific markups, while providing an estimation strategy to separately identify productivity and markups. Building on Hall (1986) they show that, for a variety of market structures, there is a simple relationship between markups, the output elasticity of a variable inputs and the share of that input's expenditure in total sales. This simple relationship will feature as well in our model and allows readily computing firm-level markups from estimates of the parameters of the production function. In order to estimate the production function, they build on Ackerberg et al. (2016) and use an additional proxy for markups heterogeneity (firm export status) in the control function. De Loecker et al. (2016) build on a similar model and add information on the output price and market share of a firm to the standard control function.

Besides the well-know, yet little explored, issue of monotonicity and invertibility common to all proxy variable approaches both De Loecker and Warzynski (2012) and De Loecker et al. (2016) face an extra challenge in estimating the production function; namely that the set of additional controls represent a sufficient statistic for unobservables other than productivity driving variable inputs demand. Within our framework such unobservables would be represented by demand shocks and markups. In this respect the use of, for example, output prices and market shares as sufficient statistics imposes implicit and unclear assumptions about preferences and especially market structure. By contrast, we do not build on the proxy variable approach and, while taking advantage of standard hypotheses in the TFP estimation literature (cost minimization, a markov process for TFP and the presence of variable and predetermined inputs), we impose clear and explicit assumptions about preferences and market structure under which demand, markups and productivity can be all consistently quantified. In our baseline specification we propose a monopolistic competition approach featuring generalized CES preferences. We further show in Appendices A and B that our approach can be generalised to more general representative consumer preferences and random utility models as well as alternative forms of imperfect competition like, for example, the oligopolistic framework developed in Atkeson and Burstein (2008) and further refined by Hottman et al. (2016) in their analysis of multi-product firms.

Our interest in demand shocks is common to both De Loecker (2011) and Foster et al. (2008). De Loecker (2011) introduces demand shocks in a revenue-based production function model while relying on standard CES preferences and a common markup across varieties. This allows substituting for prices and getting a tractable expression for firm revenue as a function of inputs, TFP and demand shocks. Compared to our framework, De Loecker (2011) only requires revenue data but does not allow for different markups across varieties while needing some adequate proxies for demand shocks. By contrast, Foster et al. (2008) use data on both the quantity and the value of a firm's production in order to disentangle productivity shocks from demand shocks. More specifically, they focus on homogeneous goods and recover production function coefficients from industry average cost shares. They subsequently estimate a demand system featuring demand shocks measured as regression residuals and instrument firm price with firm TFP. Therefore, the identifying assumption allowing them to disentangle productivity shocks from demand shocks is that they are uncorrelated. In our framework we do not impose such an assumption and find productivity and demand shocks to be very strongly correlated with each other. This suggests that, at least in our data, Foster et al. (2008) assumption of a zero correlation is severely violated.

The rest of the paper is organized as follows. Section 2 provides our baseline econometric model and estimation procedure. We present our data in Section 3 while Section 4 contains estimation results as well as some descriptive statistics and correlations. We compare our measures with measures obtained from other approaches in Section 5 while in Section 6 we show how our framework can be used to get fresh insights into two key firm-level outcomes: export status and size. Section 7 concludes. Finally, in the Appendix we show how to extend our analysis to multi-product firm, more general preferences, forms of imperfect competition other than monopolistic competition as well as to a wider set of production functions and processes for productivity and demand shocks.

2 The MULAMA model and estimation procedure

We label our model MULAMA because of the names we give to the 3 heterogeneities we allow for: markups **MU**, demand **LAM**bda and productivity **A**. We provide the baseline model and estimation procedure in Sections 2.1 and 2.2 respectively. Section 2.3 contains highlights about the various extensions to the model we consider in the Appendix.

2.1 The Model

2.1.1 Production

Consider a Cobb-Douglas production technology[5] with 3 production factors: labour (L), materials (M) and capital (K). Labour and materials are variable inputs free of adjustment costs while capital is a dynamic-input who is fixed (predetermined) in the short-run.[6] We assume firms are single-product.[7] They minimize costs and take the prices of labour (W_L) and materials (W_M) as given. Consequently, at any given point in time, each firm i is dealing with the following short-run cost minimization problem:[8]

$$\min_{L_i, M_i} \{L_i W_L + M_i W_M\} \text{ s.t. } Q_i = A_i L_i^{\alpha_L} M_i^{\alpha_M} K_i^{\gamma - \alpha_M - \alpha_L}.$$

where A_i is an idiosyncratic productivity shock observable to the firm but not the econometrician characterized in Section 2.2. First order conditions to this problem imply that:

$$W_x = \chi_i \frac{Q_i}{X_{xi}} \alpha_x \tag{1}$$

for $x \in \{L, M\}$ where X_{xi} is either the amount of labour or of materials used by firm i and

[5] We do not need to assume constant returns to scale (γ=1). It is also relatively straightforward to adapt the model to more general production technologies. See Appendix C for more details.

[6] Labour does not need to be fully adjustable in the short-run. What we do need it at least one variable and one predetermined input. The assumption on capital is standard in the TFP estimation literature. As described in Ackerberg et al. (2016) capital is often assumed to be a dynamic input subject to an investment process with the period t capital stock of the firm actually determined at period t-1. Intuitively, the restriction behind this assumption is that it takes a full period for new capital to be ordered, delivered, and installed.

[7] We relax this assumption in Appendix E.

[8] To simplify notation we do not use time indices unless needed to avoid ambiguity. We also ignore components that are constant across firms in a given time period as they will be controlled for by time dummies.

χ_i is a Lagrange multiplier. Once solved for χ_i[9] we can write the short-run cost function as:

$$C_i = \chi_i \frac{Q_i}{W_L} \alpha_L W_L + \chi_i \frac{Q_i}{W_M} \alpha_M W_M = \chi_i Q_i \left(\alpha_L + \alpha_M\right)$$

$$= \left(\frac{Q_i}{A_i}\right)^{\frac{1}{\alpha_L+\alpha_M}} \left(\frac{W_L}{\alpha_L}\right)^{\frac{\alpha_L}{\alpha_L+\alpha_M}} \left(\frac{W_M}{\alpha_M}\right)^{\frac{\alpha_M}{\alpha_L+\alpha_M}} K_i^{1-\frac{\gamma}{\alpha_M+\alpha_L}} \left(\alpha_L + \alpha_M\right). \tag{2}$$

Marginal cost thus satisfies the following property:

$$\frac{\partial C_i}{\partial Q_i} = \frac{1}{\alpha_L + \alpha_M} \frac{C_i}{Q_i}. \tag{3}$$

2.1.2 Demand and market structure

We consider a monopolistically competitive industry[10] populated by a continuum of firms each producing one variety of a differentiated good. Each firm faces an idiosyncratic demand for its own variety and maximises profits while taking market aggregates as given. Demand heterogeneity across firms is characterized by a measure of consumers' willingness to pay for a particular product (Λ_i) that is observable to the firm but not the econometrician. We further allow firms to face different elasticities of demand.

We impose that Λ_i enters preferences in such a way that demand for product i satisfies the property:

$$\frac{\partial lnP_i}{\partial ln\Lambda_i} = \frac{\partial lnP_i}{\partial lnQ_i} + 1, \tag{4}$$

where $\frac{\partial lnP_i}{\partial lnQ_i} \equiv -\frac{1}{\eta_i}$ and η_i is the elasticity of demand. As will be better appreciated later, (4) is a very useful property allowing us to write firm revenue as a simple function of the markup, Λ_i and quantity.

Note that, for example, (4) is automatically satisfied if preferences over varieties of a representative consumer have a direct utility representation and the consumer pays a price P_i to consume quantity Q_i who enters utility as $\tilde{Q}_i = \Lambda_i Q_i$. In this sense Λ_i is a measure of vertical differentiation or quality.[11] We characterize the stochastic process driving Λ_i

[9] $\chi_i = Q_i^{\frac{1}{\alpha_L+\alpha_M}-1} A_i^{-\frac{1}{\alpha_L+\alpha_M}} \left(\frac{W_L}{\alpha_L}\right)^{\frac{\alpha_L}{\alpha_M+\alpha_L}} \left(\frac{W_M}{\alpha_M}\right)^{\frac{\alpha_M}{\alpha_M+\alpha_L}} K_i^{1-\frac{\gamma}{\alpha_M+\alpha_L}}$.

[10] We show in Appendix B how our approach can be generalised to alternative forms of imperfect competition like, for example, the framework developed in Atkeson and Burstein (2008).

[11] As discussed in Di Comite et al. (2014) clear definitions of horizontal and vertical differentiation until now only exist in discrete choice models with indivisible varieties and with consumers making mutually exclusive choices. Many discrete choice models actually incorporate both types of differentiation (Anderson et al., 1992). In contrast, a clear distinction between horizontal (taste) and vertical (quality) differentiation is to a great extent absent in models where consumers have a love-for variety and purchase many products in different quantities. As in Di Comite et al. (2014) our model has features including both horizontal and vertical differentiation, which Di Comite et al. (2014) refer to as "verti-zontal".

(as well as the one for A_i) in Section 2.2. In Appendix A, we provide more insights on the interpretation of Λ_i and characterize a class of representative consumer preferences and random utility models generating demands that satisfy (4).

A simple but flexible case satisfying (4) is the generalized CES preferences structure introduced by Spence (1976)[12] that we adopt throughout this Section. In our baseline specification a representative consumer demand is thus obtained from the following problem:

$$\max_{Q}\left\{\int_{i\in I}\frac{\eta_i}{\eta_i-1}\left(\Lambda_i Q_i\right)^{\frac{\eta_i-1}{\eta_i}}\mathrm{d}i\right\}\text{ s.t. }\int_i P_i Q_i \mathrm{d}i = B$$

where B is the budget, Q is a vector with elements Q_i and the set of varieties is denoted by I. The first order condition to this problem implies:

$$P_i\kappa = \Lambda_i^{\frac{\eta_i-1}{\eta_i}} Q_i^{-\frac{1}{\eta_i}} \tag{5}$$

where κ is a Lagrange multiplier. Re-arranging suggests that firm-level demand is:

$$Q_i = P_i^{-\eta_i}\Lambda_i^{\eta_i-1}\kappa^{-\eta_i}.$$

Profit maximization of firm i then requires:

$$P_i = \mu_i\frac{\partial C_i}{\partial Q_i} \tag{6}$$

where the markup of firm i is simply a function of the elasticity of demand: $\mu_i = \frac{\eta_i}{\eta_i-1}$.

2.1.3 Some key properties

Here, we derive some useful properties that we will use in Section 2.2 to manipulate equations. First, from (1), (2), (3) and (6) we have:

$$\alpha_x = \frac{X_{xi}W_x}{\chi_i Q_i} = \frac{X_{xi}W_x}{\frac{C_i}{\alpha_L+\alpha_M}} = \frac{X_{xi}W_x}{\frac{P_i}{\mu_i}Q_i}.$$

Therefore,

$$\frac{\alpha_x}{\mu_i} = \frac{X_{xi}W_x}{P_i Q_i} \equiv s_{xi} \tag{7}$$

where s_{xi} is the expenditure share of factor $x \in \{L, M\}$ in firm i revenue. This means that, for example, materials' expenditure share is equal to the the output elasticity of materials (the constant α_M in our Cobb-Douglas benchmark case) divided by the markup μ_i.[13] This is

[12] See Appendix A for further details.

[13] See Hall (1986) and De Loecker and Warzynski (2012) for a more general derivation of this property.

a very convenient property delivering a simple way to measure markups:

$$\mu_i = \frac{\alpha_M}{s_{Mi}}. \tag{8}$$

From (8) it is clear that the markup of a firm will be a scaling of the inverse of its materials' expenditure share. Such a share is typically observable in the data and does not require any estimation. However, we do need to estimate α_M in order to measure markups level.[14]

As already anticipated, the way we introduce demand heterogeneity in (4) is also very useful. From now onwards we denote with small case the log of a variable (for example λ denotes the natural logarithm of Λ). Note that using $\mu_i = \frac{\eta_i}{\eta_i - 1}$ as well as (5) we can write log revenue, up to an innocuous constant, as:[15]

$$r_i = q_i + p_i = \frac{1}{\mu_i}\left(q_i + \lambda_i\right). \tag{9}$$

Equation (9) thus shows we can rewrite the revenue equation as a simple function of the markup as well as of the quantity and the demand shock. By substituting q_i with the formula of the Cobb-Douglas we can transform this further as:

$$r_i = \frac{\alpha_L}{\mu_i}\left(l_i - k_i\right) + \frac{\alpha_M}{\mu_i}\left(m_i - k_i\right) + \frac{\gamma}{\mu_i}k_i + \frac{1}{\mu_i}\left(a_i + \lambda_i\right).$$

Furthermore, by using (7) and (8), we finally get:

$$LHS_i \equiv \frac{r_i - s_{Li}\left(l_i - k_i\right) - s_{Mi}\left(m_i - k_i\right)}{s_{Mi}} = \frac{\gamma}{\alpha_M}k_i + \frac{1}{\alpha_M}\left(a_i + \lambda_i\right). \tag{10}$$

There are two important features of (10). First, the entire left-hand side (LHS_i) is made up of variables that are fully observable. Second, on the right hand side we have some key parameters (γ and α_M), log capital k_i (which is given for a firm in the short-run) and two unobservable endogenous variables (that are known to the firm and drive its choices of inputs and pricing while being unobservable to the econometrician) entering linearly and with the same coefficient. By imposing enough structure to the process driving a_i and λ_i, which we do next, (10) will allow us to estimate some key parameters.

[14] Note that in what follows we do not need to specify the stochastic process driving μ_i nor spell out a specific relationship with A_i and Λ_i. Our framework allows for markups that are either entirely determined by productivity and demand shocks or are entirely independent from these other shocks or a combination of the previous two cases. See Appendix A for further details.

[15] Equation (9) holds as an equality in the generalized CES preferences structure we consider here. In other cases it holds as a local approximation. See Appendix A for further details.

2.2 Estimation Procedure

We now use the time index and assume, as it is typically done in models featuring unobservable productivity shocks, that a_{it} evolves over time as stochastic Markov processes. We further assume that λ_{it} can also be described by a Markov process:[16]

$$
\begin{aligned}
a_{it} &= \phi_a \; a_{it-1} + \nu_{ait} \\
\lambda_{it} &= \phi_\lambda \lambda_{it-1} + \nu_{\lambda it}
\end{aligned}
\qquad (11)
$$

where ν_{ait} and $\nu_{\lambda it}$ can be correlated with each other.

Note that so far our identifying assumptions (cost minimization, a markov process for TFP and the presence of variable and predetermined inputs) are very similar to those made in the TFP estimation literature and in particular to those of De Loecker and Warzynski (2012). This implies that the optimal expenditures in materials m_{it} and labour l_{it} can be expressed as a function of the predetermined input (capital) and the 3 state variables (productivity, markups and demand shocks in t). Provided monotonicity with respect to a_{it} applies and that suitable proxies for μ_{it} and λ_{it} are available one could invert the two expenditure functions and use the two-step proxy variable approach described in Ackerberg et al. (2016) to estimate the parameters of the production function and recover productivity. However, the use of specific proxies for markups and demand shocks (like output prices and market shares) amounts making implicit and sometimes unclear assumptions about demand and market structure. More specifically both output prices and market shares will in general depend on the full set of variables (k_{it}, a_{it}, λ_{it} and μ_{it}) and it is not guaranteed that: (i) one can express m_{it} as a deterministic function of capital, productivity, output price and market share; (ii) invertibility of this specific function with respect to a_{it} holds. By contrast we no not build on the proxy variable approach and impose clear and explicit assumptions about preferences and market structure under which productivity, as well as demand and markups, can be consistently quantified.

Turning back to our framework before substituting (11) into (10) we need to find a convenient way to express a_{it-1} and λ_{it-1}. By using (8) and (9) we have:

[16] In Section 4.3 we provide evidence supporting the Markov process assumption for λ_{it}. For simplicity we assume here that both the a_{it} and λ_{it} processes are linear, i.e., together they are a VAR(1) process. We allow for more complex non-linear Markov processes as well as for firm fixed effects and measurement error in capital in Appendix D. One could also build on Doraszelski and Jaumandreu (2013) to enrich our model and draw an explicit link between investments in various forms of R&D (like product and process) and the productivity and quality process we use here. Indeed, our model shares quite a few features and assumptions with the one developed in Doraszelski and Jaumandreu (2013).

$$\lambda_{it-1} = r_{it-1}\mu_{it-1} - q_{it-1} = r_{it-1}\frac{\alpha_M}{s_{Mit-1}} - q_{it-1}. \tag{12}$$

At the same time plugging (12) into (10) and re-arranging yields:

$$a_{it-1} = \alpha_M LHS_{it-1} - \gamma k_{it-1} - \left(r_{it-1}\frac{\alpha_M}{s_{Mit-1}} - q_{it-1} \right). \tag{13}$$

Finally, by substituting (11) to (13) into (10) we obtain:

$$\begin{aligned} LHS_{it} &= \frac{\gamma}{\alpha_M}k_{it} + \phi_a LHS_{it-1} - \phi_a \frac{\gamma}{\alpha_M}k_{it-1} \\ &+ (\phi_\lambda - \phi_a)\left(\frac{r_{it-1}}{s_{Mit-1}} - \frac{1}{\alpha_M}q_{it-1} \right) + \frac{1}{\alpha_M}\left(\nu_{ait} + \nu_{\lambda it}\right). \end{aligned} \tag{14}$$

Equation (14) is key because it allows identifying two key parameters: $\frac{\gamma}{\alpha_M} \equiv \beta$ and ϕ_a. As will become clearer later on, it turns out that we do not actually need to estimate all of the model parameters to get measures of productivity shocks, demand shocks and markups. Indeed β, ϕ_a and γ are sufficient. Using the revenue equation (14) we get estimates for β and ϕ_a. Using the additional information provided by the quantity equation described below, along with estimates $\hat{\beta}$ and $\hat{\phi}_a$, we will in turn be able to estimate the returns to scale parameter γ.

There are various way of estimating (14) and here we use perhaps the simplest one. More specifically, we rewrite (14) as the following linear regression:

$$LHS_{it} = b_1 z_{1it} + b_2 z_{2it} + b_3 z_{3it} + b_4 z_{4it} + b_5 z_{5it} + u_{it} \tag{15}$$

where $z_{1it}{=}k_{it}$, $z_{2it}{=}LHS_{it-1}$, $z_{3it}{=}k_{it-1}$, $z_{4it}{=}\frac{r_{it-1}}{s_{Mit-1}}$, $z_{5it}{=}q_{it-1}$, $u_{it}{=}\frac{1}{\alpha_M}\left(\nu_{ait} + \nu_{\lambda it}\right)$ as well as $b_1{=}\beta$, $b_2{=}\phi_a$, $b_3{=}-\phi_a\beta$, $b_4{=}(\phi_\lambda - \phi_a)$ and $b_5{=}-\left(\phi_\lambda - \phi_a\right)\frac{1}{\alpha_M}$. Given our assumptions, the error term u_{it} in (15) is uncorrelated with all of the regressors. Therefore (15) can be estimated via simple OLS. After doing this we set $\hat{\beta}{=}\hat{b}_1$ and $\hat{\phi}_a{=}\hat{b}_2$ and do not exploit parameters' constraints in the estimation.[17]

We now turn to estimating γ. Equation (7) implies $\alpha_L = \mu_{it}s_{Lit}$ and $\alpha_M = \mu_{it}s_{Mit}$. Firm log output q_{it} can thus be written as:

[17] This means that, for example, we do not exploit the non-linear constraint b_3=-$b_1 b_2$. We can certainly do this at the cost of using non-linear least squares. Furthermore, by exploiting parameters' constraints we could actually also estimate α_M, and so γ, from (15) without need for further estimations. However, identification of α_M from (15) rests on the reduced form parameter $(\phi_\lambda - \phi_a)$ being different from zero. In unreported results we generally fail to reject the hypothesis that $(\phi_\lambda - \phi_a)$ is equal to zero. By complementing the estimation with a second stage quantity equation we avoid these issues while at the same time $(\phi_\lambda - \phi_a)$ being zero does not affect identification of γ in the second stage.

$$q_{it} = \mu_{it} s_{Lit} \left(l_{it} - k_{it} \right) + \mu_{it} s_{Mit} \left(m_{it} - k_{it} \right) + \gamma k_{it} + a_{it}. \tag{16}$$

Further using (8) as well as the fact that $\alpha_M = \frac{\gamma}{\beta}$ we get:

$$q_{it} = \frac{\gamma}{\hat{\beta} s_{Mit}} s_{Lit} \left(l_{it} - k_{it} \right) + \frac{\gamma}{\hat{\beta}} \left(m_{it} - k_{it} \right) + \gamma k_{it} + a_{it} \tag{17}$$

where we replace β with $\hat{\beta}$. Finally, using (11) to substitute for a_{it} and (13) to substitute for a_{it-1} we obtain:

$$\begin{aligned} q_{it} &= \frac{\gamma}{\hat{\beta}} \frac{s_{Lit}}{s_{Mit}} \left(l_{it} - k_{it} \right) + \frac{\gamma}{\hat{\beta}} \left(m_{it} - k_{it} \right) + \gamma k_{it} \\ &+ \hat{\phi}_a \frac{\gamma}{\hat{\beta}} LHS_{it-1} - \hat{\phi}_a \gamma k_{it-1} - \hat{\phi}_a \left(r_{it-1} \frac{\gamma}{\hat{\beta} s_{Mit-1}} - q_{it-1} \right) + \nu_{ait}. \end{aligned} \tag{18}$$

Note that the only unobservable in (18) is the white noise term ν_{ait} while the only parameter left un-identified is the scale parameter γ. However, this time we cannot proceed with least squares because ν_{ait} is correlated with the regressors and in particular with l_{it}, s_{Lit}, m_{it} and s_{Mit}. Indeed, ν_{ait} affects a_{it} and so affects input choices, pricing and revenues. Nonetheless we can, for example, identify γ from the assumption that capital is predetermined using the following zero moment condition:

$$E \left\{ \nu_{ait} k_{it} \right\} = 0$$

applied to equation (18). We can implement this restriction in a linear regression framework by writing (18) as:

$$\overline{LHS}_{it} = b_6 z_{6it} + \nu_{ait} \tag{19}$$

where:

$$\overline{LHS}_{it} = q_{it} - \hat{\phi}_a q_{it-1}$$

$$z_{6it} = \frac{1}{\hat{\beta}} \frac{s_{Lit}}{s_{Mit}} \left(l_{it} - k_{it} \right) + \frac{1}{\hat{\beta}} \left(m_{it} - k_{it} \right) + k_{it} + \frac{\hat{\phi}_a}{\hat{\beta}} LHS_{it-1} - \hat{\phi}_a k_{it-1} - r_{it-1} \frac{\hat{\phi}_a}{\hat{\beta} s_{Mit-1}}$$

as well as b_6=γ and z_{6it} is instrumented with k_{it}. We set $\hat{\gamma}$=$\hat{b}_6$ and are in turn able to identify productivity shocks, demand shocks and markups:

$$\hat{a}_{it} = q_{it} - \frac{\hat{\gamma}}{\hat{\beta}} \frac{s_{Lit}}{s_{Mit}} (l_{it} - k_{it}) - \frac{\hat{\gamma}}{\hat{\beta}} (m_{it} - k_{it}) - \hat{\gamma} k_{it} \tag{20}$$

$$\hat{\mu}_{it} = \frac{\hat{\gamma}}{\hat{\beta} s_{Mit}} \tag{21}$$

$$\hat{\lambda}_{it} = \frac{\hat{\gamma}}{\hat{\beta} s_{Mit}} r_{it} - q_{it}. \tag{22}$$

Finally, standard errors of $\hat{\beta}$, $\hat{\phi}_a$ and $\hat{\gamma}$ can be obtained via bootstrapping by re-sampling residuals in regressions (15) and (19).

2.3 Extensions

In Appendices A to E we show how to extend our framework to more general preferences, to forms of imperfect competition other than monopolistic competition, to a wider set of production functions and processes for productivity and demand shocks as well as to multi-product firms.

In Appendix A we show there are various ways to extend our model and estimation methodology to preferences other than the baseline generalized CES case. More specifically, within the representative consumer framework, we identify a class of preferences under which the key equation (9) holds as a first-order linear approximation and then show how to modify the estimation procedure accordingly. The key intuition is that quantities Q_i should enter utility as $\tilde{Q}_i = \Lambda_i Q_i$ for (9) to hold as a local approximation. This can be easily accommodated for preferences that have a direct utility representation. A complementary approach, not involving any local approximation, is instead to fully specify a preference structure and work out the corresponding algebra for the estimation equations. We provide a fully worked-out example based on Gaussian preferences. Last but not least, we introduce a class of random utility models (discrete/continuous choice models) delivering demand systems that are compatible with our framework. These discrete/continuous choice models represent a generalisation of standard discrete choice models including, for example, the Multinomial Logit. They are obtained from a random utility framework in which consumers not only choose one alternative amongst many but also how much to consume of a particular good.[18]

Appendix B shows how our framework can be extended beyond the scope of monopolistic competition. More specifically, we show how to frame it in terms of the model developed in Atkeson and Burstein (2008) and further refined by Hottman et al. (2016) in their analysis of multi-product firms. Atkeson and Burstein (2008) provide two versions of their model. One is based on quantity competition while the other is based on price competition. In both cases,

[18] See Nocke and Schutz (2016) for further details.

firms are large enough to perceive their own impact on overall price indices. It turns out that both versions can be accommodated within our analysis. Indeed, our definition of quality can be incorporated into Atkeson and Burstein (2008) by noticing that everything works as if a firm was selling quantity $\tilde{Q}$ while charging a price $\tilde{P}$ where $\tilde{Q} = Q\Lambda$ and $\tilde{P} = P/\Lambda$.

Appendix C is devoted to extend the analysis to more flexible production functions. In particular, we consider the (homogenous) translog form and show how to modify the estimation procedure accordingly. In Appendix D we instead provide a number of examples of richer processes for a and λ that can be dealt with. More specifically, we consider: (i) a non-linear Markov process for a and λ; (ii) the presence of time-invariant unobserved heterogeneity. We show the former case complicates the algebra but not the underlying structure of the problem while the latter case can be accommodated quite easily. We also explain how to handle measurement error in capital within our framework.

We consider multi-product firms in Appendix E. There are several issues related to multi-product firms. We focus on the issue of the assignment of inputs to outputs. Produced quantities and generated revenues may be observable for individual products in some databases. However, in many instances information on inputs used for a specific product is not available for multi-product firms. We propose an extension of our baseline model and procedure to solve the problem of assigning inputs to outputs for multi-product firms. In doing so we assume, as in De Loecker et al. (2016), there is a limited role for economies (or diseconomies) of scope on the cost side. However, contrary to De Loecker et al. (2016), we do not impose multi-product firms to be characterized by a common productivity across the different products they produce. We also allow for firm-product-time specific markups but impose demand shocks to be common across products within a firm. This corresponds to a setting where firms can be distinguished into those consistently selling high quality products and those consistently selling low quality products. Yet firms are allowed to be more or less efficient in the production of a specific product and charge different markups. The framework we propose is consistent with a monopolistically competitive market structure where firms ignore cannibalisation effects. The model could be potentially enriched to cope with such cannibalisation effects by building, for example, on the demand and market structure developed in Hottman et al. (2016).

3 Data

Our primary data consists of firm-level production data for Belgian manufacturing firms coming from the Prodcom database and provided by the National Bank of Belgium. Prodcom is a monthly survey of industrial production established by Eurostat for all EU countries in order to improve the comparability of production statistics across the EU by the use of a common product nomenclature called Prodcom (8-digit codes whose first four digits come from NACE codes). Prodcom covers production of broad sectors C and D of NACE Rev. 1.1 (Mining and quarrying and manufacturing), except for sections 10 (Mining of coal and lignite), 11 (Extraction of crude petroleum and natural gas) and 23 (Manufacture of coke and refined petroleum products). During our sample period, each Belgian firm with 10 employees or more - or with a revenue greater than a certain threshold in a given year - had to fill out the survey.[19] Firms in the survey cover more than 90% of Belgian manufacturing production and the raw data is aggregated from the plant-level to the firm-level.

This gives us a sample of about 7,000 firms a year over the period of 1995 to 2007. Data is organised by product-year-month-firm. We use information on quantity (the unit of measurement depending on the specific product) and value (Euros) of production sold. We aggregate the data at the firm-year-product level. The same data has been previously used in Bernard et al. (2012b) in their analysis of carry along trade as well as by De Loecker et al. (2014) for their study of the links between international competition and firm performance.

We also make use of more standard balance sheet data to get information on firms' inputs. We build on annual firm accounts from the National Bank of Belgium. For this study, we selected those companies that filed a full-format or abbreviated balance sheet between 1996 and 2007 and with at least one full-time equivalent employee. The resulting dataset has been previously used in Behrens et al. (2013), Mion and Zhu (2013) and Muûls and Pisu (2009) and is representative of the Belgian economy. It includes information on FTE employment, material costs, capital stock and turnover. There are more than 15,000 manufacturing firms per year displaying non-missing values for these variables.

Besides, we use standard EU-type micro trade data at the product-country-firm-month level over the period 1995-2008 provided by the National Bank of Belgium. From this data we simple borrow information on firm export status. The data has been previously used in Behrens et al. (2013), Mion and Zhu (2013) and Muûls (2015) among others. The three datasets are matched by the unique firm VAT identifier

We focus on the period 1996-2007 for which all three datasets are available and during which there has not been any major change in data collection and data nomenclatures (such as the NACE nomenclature and 4-digits level Prodcom codes, etc.).[20] We choose not to analyse multi-product firms in this paper and focus on single-product firms to allow a direct comparison with other approaches and so better highlight our methodological contribution. We could have used the model developed in Appendix E to estimate demand, TFP and

[19] Rules are somewhat different for other EU countries. In particular some EU countries only surveyed firms with 20 or more employees. The 10 employees threshold has been recently increased to 20 in Belgium as well.

[20] As reported in, for example, Bernard et al. (2012b) there have been quite a few changes in 8-digit level Prodcom codes during our sample period. Yet these changes occurred within 4-digit level Prodcom codes and are thus not problematic in our analysis. Indeed the first 4 digits of Prodcom codes come from the NACE nomenclature who remained virtually unchanged over our time span.

markups for all Belgian manufacturing firms but in doing so we would not have had the chance to compare our analysis and results to, for example, De Loecker and Warzynski (2012), Foster et al. (2008) and Olley and Pakes (1996).

As in previous studies using either revenue or quantity data our estimations are run at a more aggregate level (labelled as "industry" here) rather than at the finest available classification (8-digits products). However, this means we need to aggregate products that are sometimes quite different from each other. This is a problem also faced by other studies using physical production data including Foster et al. (2008) and Dhyne et al. (2014) and for which, to the best of our knowledge, the common practice is to find the right balance between the number of observations and level of disaggregation and simply sum quantities across products within a firm.[21] We improve on the existing practice by using the average (across firms) log price for each product j as a weight in the aggregation. We fully spell out in Appendix F the assumptions that make this approach meaningful. We further note that our results are qualitatively and, to a large extent also quantitatively, not affected by this choice.

Given that products belonging to a given NACE 3 digits code may have different units of measurement, we define an industry as a NACE 3 digit-unit of measurement pair. This means there are roughly 200 industries in the dataset. The following cleaning and restrictions are applied to focus on single-product firms and to have many observations:

- Consider only firm-year observations for which the value and quantity of production for all products (8-digit) are recorded.

- Consider only firm-year observations for which employment, materials, sales and capital are available.

- Aggregate production data at the 3 digit-unit of measurement level. See Appendix F for further details.

- Create for each firm-year the production value shares of its different 3 digit-unit of measurement products and keep a firm-year couple only if $> 95\%$ of production value is within a given 3 digit-unit of measurement: single-product firms.

- Apply small trimmings (1% up and down) based on capital intensity, share of intermediates in revenues and unit prices.

- Consider only industries with more than 80 firms in each year.

[21] De Loecker et al. (2016) also run productivity estimations at an aggregate two-digit level but use information on detailed product prices to devise an estimation strategy that avoids the issue of aggregating quantities.

In doing so, we end up with four industries on which we will focus our empirical analysis:

1. NACE 151: "Production, processing and preserving of meat and meat products"
2. NACE 212: "Manufacture of articles of paper and paperboard"
3. NACE 266: "Manufacture of articles of concrete, plaster and cement"
4. NACE 361: "Manufacture of furniture"

It is worth noting is that the industry with NACE 266 is not "Ready mixed concrete" but rather "Manufacture of articles of concrete, plaster and cement". The former industry has been the object of numerous studies (Syverson, 2004; Foster et al., 2008) and is considered a rather homogeneous good which is at best differentiated in terms of the geographic location of firms. "Manufacture of articles of concrete, plaster and cement" is a quite different industry, featuring products that are far from being homogeneous.

4 Results

In this Section we provide a number of descriptive statistics about our estimations, our measures of productivity shocks, demand shocks and markups and examine how the three dimensions of heterogeneity correlate with each other in a cross section as well as across time.

4.1 Descriptive statistics

Our main contribution is to provide a framework that simultaneously allows recovering heterogeneity in demand, TFP and markups. Heterogeneity in demand is a particularly novel dimension of our analysis and, before going into any estimations, we believe there is value in showing how much such heterogeneity is present in the raw data. This is accomplished in Figure 1 where we show, for each of the four industries in our analysis, the plot of log price and log quantity stemming from the raw data. It is made clear that firms can sell very different quantities even though they charge the same price (different values on the X-axis for a given value of Y-axis). This is not per se evidence of heterogeneity in demand because such a feature might be, for example, generated with Cournot competition among firms with different costs but facing the same overall demand. Yet Figure 1 also points to firms selling similar quantities while charging very different prices (different values on the Y-axis for a given value of X-axis) which is more revealing of demand heterogeneity. Also note that differences are remarkably large considering we are using log prices and quantities.

Insert Figure 1 about here.

Turning to estimations, Table 1 shows the mean and standard deviation of μ, a and λ across firms in each of the four industries. Average markups range from 1.214 for "Manufacture of articles of concrete, plaster and cement" to 1.411 for "Manufacture of furniture". Magnitudes are comparable to those obtained by De Loecker and Warzynski (2012) under different specifications. Though, the standard deviation is relatively small as compared to the 0.5 reported in De Loecker and Warzynski (2012). This implies that the vast majority of our firm-level markups is indeed above the economically meaningful threshold of one. This is more evident in Figure 2 where we provide the density distribution of μ along with the mean (red vertical line).

Insert Table 1 and Figure 2 about here.

As for productivity shocks and demand shocks the mean is not of much importance per se. The standard deviation is instead meaningful with the one of a being considerably larger than the 0.26 reported in Foster et al. (2008) for physical TFP. Yet, it has to be considered that Foster et al. (2008) focus on industries characterized by rather homogeneous products for which it is reasonable to expect less TFP variability across firms. As far as demand shocks are concerned the standard deviations are instead in line with the 1.16 figure reported in Foster et al. (2008). Interestingly, in our analysis demand shocks display at least as much variability as productivity shocks. This can be further appreciated in Figure 3 where we provide the (centered) density distributions of λ and a. This is an important finding suggesting that heterogeneity in demand is a key component of firm idiosyncracies being at least as sizeable as heterogeneity in productivity.

Insert Figure 3 about here.

Finally, Table 1 provides our production function estimates for the coefficients of materials (α_M), labour (α_L) and scale (γ) along with bootstrapped standard errors (200 replications). Estimates are in line with previous findings in the literature. In particular, there is evidence for moderate increasing returns to scale for quantity-based production functions which is in line with the findings of De Loecker (2011). Furthermore, bootstrapped standard errors suggest our estimates are overall rather precise.

4.2 Cross-sectional correlations

Table 2 reports cross-sectional correlations between μ, λ and a as well as log price p. The Table provides several insights.

Insert Table 2 about here.

The first thing worth noting is that the correlation between demand and productivity shocks is far from being zero. A zero correlation is the identification hypothesis for demand shocks in Foster et al. (2008). Here, we find that productivity shocks a are very strongly and negatively correlated with demand shocks λ in each of the four industries we consider. This finding is suggestive of a trade-off between the quality of a firm's products (as measured by λ) and their production cost (as measured by a) as suggested in Ackerberg et al. (2007). Consider, for example, the car industry where there is the co-existence of manufacturers (like Nissan) producing many cars for a given amount of inputs (high a) and manufacturers (like Mercedes) producing much less cars for a given amount of inputs (low a). At the same time, however, Mercedes produces cars of a higher quality in that, if the two cars were priced the same, Mercedes would sell many more cars (higher λ). To be a bit more more specific one of the most productive car plants in Europe is the Nissan factory located in Sunderland in the UK. In terms of sheer productivity measured as cars per employee it is nearly 100% more productive than a state of the art Mercedes plant near Rastatt in Germany. However, this hardly reflects a problem with the Mercedes plant. Rather, Mercedes and Nissan face very different demands which leads to different prices as well different markups. Both plants are profitable and perhaps generate a very similar revenue productivity (we will come back to this issue below). Yet, their business model is quite different.

The presence of a negative relationship between λ and a can be rationalized in several ways. For example, it could be the outcome of firms optimally differentiating themselves in the quality-cost space and/or what is left after selection has taken place and only firms with high enough a and/or high enough λ survive. The negative relationship we find is far from being perfect, and so there are indeed firms in the data who have both high (low) λ and a. Yet, the presence of a negative correlation is a first order feature of the data in our sample. This can be further appreciated in Figure 4 where we plot the (centered) values of λ and a in each of the four industries. The strength of the linear relationship is quite apparent from Figure 4. Furthermore, paying attention to scaling in the two axes, suggests a regression coefficient of -1, i.e., if the products of a firm are twice as valuable to consumers than the products of another firm (the former firm has a λ twice as big as the latter) they will be (on average) twice as costly to produce (the latter firm has an a twice as big as the former).

Insert Figure 4 about here.

The second thing worth noting is that markups are reasonably well correlated with demand shocks. More specifically, we find that firms selling higher quality goods charge higher markups. The relationship between markups and a is instead much weaker and depends on the specific sector considered. Table 3 offers further insights into the relationship of markups with demand and productivity shocks. In a model in which the fundamental driver of heterogeneity in demand across firms is only λ (like in the Generalized Quadratic Utility case we consider in Appendix A) we would expect markups μ to vary across firms only to the extent that λ, a and capital (with the latter two determining marginal costs) vary across firms.

Insert Table 3 about here.

In Table 3 we regress μ on a, λ and capital k. Differences in markups across firms are reasonably well explained (in terms of R^2) by differences in demand shocks, productivity shocks and the capital stock. However, there is a considerable amount of unexplained heterogeneity. Therefore, the higher flexibility of the Generalised CES as compared to the Generalized Quadratic Utility seems to be needed in order to capture markups heterogeneity across firms. Furthermore, our results indicate, to the extent that a comparison can be made, that firms with higher productivity charge ceteris paribus, i.e., controlling for λ and k, higher markups. This is in line with preferences featuring increasing relative love for variety (from which pro-competitive effects can be rationalised) and the presence of market distortions such that the market leads to too little selection with respect to the social optimum.[22]

One last thing to note about Table 2 is the extremely strong correlation between log prices p and TFP a ranging from -0.916 to -0.956. This finding is in line with evidence reported in Foster et al. (2008) and was one of the grounds for their choice to instrument prices with TFP. Yet the correlation between prices and demand shocks is also very strong ranging from 0.691 to 0.926. The correlation with markups is instead much weaker ranging from not being significant to 0.213.

4.3 Correlations across time and predictive power

Numerous studies on productivity report a high degree of persistency across time while Foster et al. (2008) document a similar behavior for their measure of demand shocks. While based on a different approach and data type our analysis confirms these findings. Table 4 reports estimations. In each case we regress a, λ and μ on their respective time lag. Both a and λ are characterized by a high degree of time persistency with autoregressive coefficients being around 0.9 and an R^2 of 0.8 or above. This evidence supports our choice of a simple Markov process for a and, most importantly, λ. Turning to markups they are relatively less persistent with the autoregressive coefficient scoring around 0.8 and an R^2 of 0.7.

Insert Table 4 about here.

Before moving on to the next section, to a more systematic comparison of our measures of heterogeneity with those obtained from other methodologies, we end this section providing a feeling of their predictive power. This is accomplished in Tables 5 to 7 where we regress, respectively, log price p, log quantity q and log revenue r on a, λ, μ and log capital k. In our model p, q and r should ultimately be a (non-linear) function of the primitives of the model: a, λ, μ and k. Though being an approximation the linear regression gives us, by means of

[22] See Dhingra and Morrow (2012) and Zhelobodko et al. (2012) for further details.

the R^2, a feeling about the predictive power of the model. An inspection of the three tables reveals our model scores extremely well for log prices with R^2 of around 0.95. As for log quantities, results are more modest attaining R^2 of about 0.65. Considering the last Table we have R^2s slightly below those of q which is not surprising given the results of the previous two Tables and the fact that $r{=}p{+}q$.

Insert Tables 5 to 7 about here.

5 Comparison to other methodologies

5.1 TFP

In order to gain insights into how and to what extent our methodology to measure TFP differs from other approaches we have computed additional TFP estimates based on:

- The GMM version of Levinsohn and Petrin (2003) incorporating the Ackerberg et al. (2016) correction as implemented in De Loecker and Warzynski (2012): DLW
- A GMM version of Olley and Pakes (1996): OP
- Industry costs shares as in Foster et al. (2008): FHS
- Ordinary Least Squares: OLS

For each case we have computed a revenue-based TFP (using revenue as a measure of output) and a quantity-based TFP (using physical quantity as a measure of output).

Our results suggest the following. First, there is considerable difference between revenue-based and quantity-based TFP. This has already been documented in Foster et al. (2008) using FHS productivity estimates. Our findings extend this result to a broader set of TFP measurement approaches while pointing to more substantial differences. For example, Foster et al. (2008) report a correlation between the two TFP of 0.64.[23] We instead find the following correlations (across all industries while subtracting the mean) between revenue-based and quantity-based TFP:[24]

- DLW, quantity and revenue based: 0.380***
- OP, quantity and revenue based: 0.0929***

[23] The closest (to our) revenue TFP measure used in Foster et al. (2008), is what they label "Traditional TFP".

[24] As already noted earlier, Foster et al. (2008) focus on industries characterized by rather homogeneous products for which it is reasonable to expect less differences in prices and so a closer relationship between revenue-based and quantity-based TFP.

- FHS, quantity and revenue based: 0.0863***
- OLS, quantity and revenue based: 0.0921***
  *** p<0.01, ** p<0.05, * p<0.1

Second, the correlations (across all sectors while demeaning) between our quantity TFP measure a and quantity TFP measures computed with other methods are:

- DLW, quantity based: 0.866***
- OP, quantity based: 0.948***
- FHS, quantity based: 0.935***
- OLS, quantity based: 0.948***
  *** p<0.01, ** p<0.05, * p<0.1

Therefore, the key message is that having quantity TFP is the key thing. The specific methodology is certainly important and our approach has the advantage of allowing TFP measurement within a framework where both markups and demand heterogeneity are present and potentially correlated with TFP shocks. However this is, at least in our data, a second order problem. This lends support to Syverson (2011) in that: *"the inherent variation in establishment- or firm-level microdata is typically so large as to swamp any small measurement-induced differences in productivity metrics."*

Where our framework delivers its full potential and ultimately provides an important contribution is not just in getting the TFP "more right" than in other methodologies but rather in allowing to unravel many dimensions of heterogeneity potentially correlated with each other. These dimensions of heterogeneity, namely TFP shocks, demand shocks and markups, can in turn be used to get richer and deeper insights into important questions. Consider, for example, revenue TFP. Revenue TFP could be defined in our framework as $TFP_{it}^{R} \equiv r_{it} - \bar{q}_{it}$ where $\bar{q}_{it} \equiv q_{it} - a_{it} = \alpha_L \left(l_{it} - k_{it}\right) + \alpha_M \left(m_{it} - k_{it}\right) + \gamma k_{it}$. By further using Equation (9) and substituting we get:

$$TFP_{it}^{R} = \frac{1}{\mu_{it}} \left(a_{it} + \lambda_{it}\right) + \frac{1 - \mu_{it}}{\mu_{it}} \bar{q}_{it} \tag{23}$$

So TFP_{it}^{R} is a function of a, λ, μ and production scale. Tables 8 to 11 provide insights into the usefulness of this approach. In these four tables we regress DLW, OP, FHS and OLS revenue based productivities on a, λ and μ while reporting Beta coefficients. These tables are meant to highlight how, putting aside the issue of getting the revenue TFP "more right", differences in measured TFP are a mixture of underlying differences in physical TFP, demand and markups. Interestingly, in most instances variation in revenue TFP is actually attributable mainly to variation in demand across firms (λ and μ) rather than in quantity TFP. At the same time λ typically explains more variation than μ.

Insert Tables 8 to 11 about here.

5.2 Markups

De Loecker and Warzynski (2012) provide the first fully-fledged framework to compute markups at the firm-level. We share with them a few assumptions and so the question of how our markups compare to theirs arises naturally.

In both our framework and theirs, markups are obtained as the ratio of the estimated output elasticity of a variable input, free of adjustment costs (materials in our case with constant elasticity α_M), to the share of that input's expenditure in total sales. Therefore, provided both methods deliver the same estimate for α_M, markups will be identical. Even if the α_M were different, the correlation between the two sets of markups would be actually one. Only to the extent that getting the level of markups (and so the value of α_M) right is important to the analysis they would thus be different. In this respect our methodology has the advantage of not requiring suitable proxies for markups and demand shocks (nor the related implicit assumptions about demand and market structure) while allowing us to quantify both of these heterogeneities.

A substantial difference between the two methodologies arises when unobservable (to the firm) productivity shocks enter into the analysis. De Loecker and Warzynski (2012) propose a correction to the markups formula to take into account such shocks. When applying their correction in our data we get a (significant at the 1%) correlation (across all sectors while demeaning) between the two sets of markups of only 0.0633. The difference is clearly substantial and calls for a serious evaluation of both the importance of productivity shocks unobservable to the firm as well as the capacity of the proxy variable approach to separate observable and unobservable (to the firm) shocks.

To gain further insights, Table 12 provides average markups across the four industries computed with our methodology (μ), DLW quantity-based productivity (DLW1) and DLW quantity-based productivity with correction for unobservable (to the firm) productivity shocks (DLW2). Table 12 shows such averages are reasonably similar in all industries with the exception of industry 361 where DLW2 markup is substantially higher.

Insert Table 12 about here.

5.3 Demand shocks

In their seminal paper, Foster et al. (2008) use production data of US manufacturing firms, containing information on both value and physical quantity, to estimate quantity-based TFP as well as demand shocks. They measure demand shocks as the residual of a regression where log quantity is regressed on log price and the latter is instrumented with TFP obtained using industry costs shares to measure production function parameters (FHS TFP). The key identifying assumption in their framework is thus that productivity shocks are uncorrelated with demand shocks. We instead assume that demand shocks follow a Markov process while not imposing restrictions on their correlation with TFP shocks.

In light of our framework, the Foster et al. (2008) approach is problematic for at least two reasons:

1. Markups are heterogeneous across firms: this means that the log price coefficient in their regression should be firm-specific. Within our framework we do not need to estimate those firm-specific coefficients because, based on our assumptions, they equal $-\eta_{it} = -\frac{\mu_{it}}{\mu_{it}-1}$.
2. Demand shocks are strongly correlated with productivity shocks: this means that their IV strategy would not work in our data. Within our framework we do not need to take a stand on the correlation between demand and productivity shocks. Equation (9) provides us with sufficient means to measure demand shocks once we have estimated TFP and markups.

In order to gain insights into the differences between the two approaches we have followed Foster et al. (2008) and computed demand shocks as the residual of a regression where log quantity is regressed on log price and the latter is instrumented with FHS TFP.[25] Figure 5 shows a plot of λ and FHS demand shocks for our four industries. Though positively correlated (correlations in Table 13 range from 0.231 to 0.414) the two sets of demand shocks are clearly quite different and can potentially lead to completely different conclusions when used to answer a specific research question.

Insert Figure 5 and Table 13 about here.

To be fair, our λ does not precisely correspond to the definition of demand shocks in Foster et al. (2008). Nevertheless, we can still define demand shocks as the residual component of

[25] Foster et al. (2008) also control for a set of demand shifters, including a set of year dummies as well as the average income in the plant's local market m where local markets are defined based on Bureau of Economic Analysis' Economic Areas. We also include in our regressions a full battery of year dummies. Yet, given the small size of Belgium we did not include any control for the plant's local market income. Our IV estimations, available upon request, deliver highly (1%) significant coefficients for the log price coefficient in all four industries (point estimates are -1.4189, -1.4327, -.8850 and -1.1524 for industries 151, 212, 266 and 361 respectively).

model where log quantity is regressed over log price within our framework. From (5) we have $q_{it} = -\eta_{it}p_{it} + (\eta_{it} - 1)\lambda_{it} - \eta_{it} \ln \kappa_t$ and the residual component is thus $(\eta_{it} - 1)\lambda_{it} - \eta_{it} \ln \kappa_t$ rather than simply λ_{it}. We do not observe $\ln \kappa_t$ but we do observe $q_{it} + \eta_{it}p_{it} = (\eta_{it} - 1)\lambda_{it} - \eta_{it} \ln \kappa_t$. Figure 6 shows a plot of our residual demand shocks (as measured by $q_{it} + \eta_{it}p_{it}$) and FHS demand shocks. Again, though being positively correlated (the even smaller correlations in Table 14 now range from 0.058 to 0.299) the two sets of demand shocks are quite different.

Insert Figure 6 and Table 14 about here.

In sum, we believe our methodology is to be preferred to measure demand shocks because it is actually more flexible and explicit than Foster et al. (2008) while having the same data requirements.

6 An application to export status and firm size

6.1 Export status

Exporting firms are typically found to be more productive than non-exporters. The empirical evidence is vast (see Bernard et al., 2012a) while at the same time there are good reasons to believe the direction of causality goes from productivity to export status via a self-selection mechanism (Bernard and Jensen, 1999).[26] This mechanism has been first fully spelled out in Melitz (2003) and has been the basis of a very prolific and influential theoretical and empirical literature (Helpman et al., 2015).

Yet empirical evidence is based only on revenue-based measures of productivity and, as seen in Section 5.1, these measures are a mixture of actual physical TFP, demand shocks and markups. On the theory side, the mainstream approach relies on one dimension of heterogeneity across firms only (TFP), with heterogeneity in demand not receiving much attention.[27]. Yet, Melitz (2003) acknowledges that in his framework higher productivity can be either considered as producing a symmetric variety at lower marginal cost or producing a higher quality variety at equal cost. In what follows, we build on the three measures of heterogeneity to offer novel and sharper insights on the relationship between firm export status, productivity and demand.

We first start by showing that the usual positive correlation between revenue-based TFP and firm export status holds in our data. Tables 15 and 16 report results (Beta coefficients) of

[26] Interestingly, in a recent paper Garcia-Marin and Voigtländer (2014) offer some new insights into the question of learning by exporting. More specifically, they find among new Chilean exporters a decrease in marginal costs and a stable pattern in revenue-based productivity due to offsetting price changes.

[27] Some noticeable exceptions include Verhoogen (2008), Fajgelbaum et al. (2011) and Feenstra and Romalis (2014)

a linear estimation where the export status of a firm is regressed on its OP and DLW revenue-based TFP respectively. Both sets of estimations convey the same message. Irrespective of the TFP measure used and industry we find a positive correlation between revenue-based TFP and firm export status. The availability of physical quantity data allows us to go one step further and look at the correlation between quantity-based TFP and export status. This is done in Tables 17 and 18 where we report results (Beta coefficients) of a linear estimation, where firm export status is regressed on its OP and DLW quantity-based TFP respectively. Tables 17 and 18 indicate that the positive correlation between export status and revenue-based TFP extends to quantity-based TFP. This provides support to the mainstream theoretical framework based on differences across firms in term of their ability to produce at a lower marginal cost.

Insert Tables 15 to 18 about here.

Yet, within our framework we can go even further and ask whether and how demand and markups heterogeneity also matters and how it interacts with heterogeneity in productivity. In this respect it is important to note that our goal here is neither to draw any causal relationships nor to develop a fully fledged model of export participation with three underlying dimensions of heterogeneity but rather to uncover correlations that might be useful to further theoretical contributions.

Turning to the data in Table 19 we consider export status regressed on a, λ and μ while reporting Beta coefficients. On the one hand Table 19 confirms that the positive correlation between quantity-based TFP (a) and firms' export status is overall[28] robust to the inclusion of demand shocks and markups heterogeneity. On the other hand, Table 19 expands the horizon of the analysis by pointing to the importance of demand heterogeneity in the understanding of why some firm exports and others do not. Beta coefficients for μ and especially for λ are in general larger in magnitude than those of a. This suggests that demand heterogeneity is more important than differences in underlying physical productivity to draw the line between exporting and non-exporting firms. At the same time, the coefficients' signs indicate that exporters typically sell higher quality goods and charge lower markups. The first finding is quite intuitive while the second can be, for example, rationalized by the fact that exporters absorb part of the trade costs on their exports so charging, everything else equal, a lower markup.

Insert Table 19 about here.

[28] We fail to find a significant relationship in the furniture industry. This might be due to the very high correlation (-0.910) between a and λ in this industry.

6.2 Firm size

Larger firms are typically found to be more productive than smaller ones when using different measures of productivity. The empirical evidence is abundant and encompasses both developed (Van Ark and Monnikhof, 1996) and developing (Van Biesebroeck, 2005) countries. At the same time there are several models consistent with this relationship being, on average, true in a cross section of firms (Jovanovic, 1982; Lucas Jr, 1978). Yet on the empirical side only revenue based on measures of productivity have been used so far, meaning that it is not clear whether the positive correlation stems from physical TFP and/or demand shocks and/or markups. This mirrors the situation on the theory side, where the underlying differences across firms are typically in terms of their ability to turn inputs into output and not much in terms of the demand they face - meaning that the distinction between revenue-based and quantity-based productivity is to a large extent immaterial.

We start by showing that the usual positive correlation between revenue-based TFP and firm size (as measured by the log number of employees) holds in our data. Tables 20 and 21 report results (Beta coefficients) of a linear estimation where the log number of employees of a firm is regressed on its OP and DLW revenue-based TFP respectively. Both sets of estimations convey the same message. Irrespective of the TFP measure used and industry we find a positive correlation between revenue-based TFP and firm size. The availability of physical quantity data allows to go one step further and look at the correlation between quantity-based TFP and firm size. This is done in Tables 22 and 23 where we report results (Beta coefficients) of a linear estimation where the log number of employees of a firm is regressed on its OP and DLW quantity-based TFP respectively. Tables 22 and 23 indicate that the positive correlation between firm size and revenue-based TFP extends to quantity-based TFP. This provides support to the mainstream theoretical framework based on differences across firms in terms of their ability to turn inputs into output.

Insert Tables 20 to 23 about here.

As in the export status case we are in the position of assessing whether demand and markups heterogeneity also matters and how it interacts with heterogeneity in productivity. Parallel to that our goal here is again neither to draw any causal relationships nor to develop any fully fledged model but rather document correlations that might be useful to future research.

This is achieved in Table 24 where we consider firms' size regressed on a, λ and μ while reporting Beta coefficients. On the one hand Table 24 confirms that the positive correlation between quantity-based TFP (a) and firm size is robust to the inclusion of demand shocks and markups heterogeneity. On the other hand, Table 24 broadens the spectrum of the analysis by pointing to the importance of demand heterogeneity in the understanding of why some firm are larger than others. Beta coefficients for λ and μ are in general smaller in magnitude than those of a. Yet, their combined effect is comparable to the one of quantity-based TFP while coefficient signs indicate that larger firms typically sell higher quality goods and charge lower markups. As in the case of export status the first finding is somewhat intuitive; given two firms with the same TFP the one with the higher demand shock will be larger. The second finding is less intuitive and is perhaps related to the fact that most exporters can be found among larger firms.

Insert Table 24 about here.

7 Conclusions

We provide a novel framework that simultaneously allows recovering heterogeneity in productivity, demand, and markups across firms while leaving the correlation among the three unrestricted. We accomplish this by explicitly introducing demand heterogeneity and by systematically exploiting assumptions that are implicit in previous firm-level productivity estimation approaches. We apply our econometric framework to Belgian manufacturing firms and quantify productivity, markups and demand shocks for four industries. We show how these shocks are correlated among them, across time as well as with measures obtained from other approaches. We finally assess how and to what extent our three dimensions of heterogeneity allow us to gain deeper and sharper insights on two key firm-level outcomes: export status and size. Our takeaway message is that heterogeneity in demand and markups is quantitatively as sizeable as heterogeneity in productivity and by some measure even more important than the latter in drawing the line between small and large firms, between firms with higher or lower revenue TFP and between exporters and non-exporters.

Our methodology is rich enough to be applied to markets where products have some features of both horizontal and vertical differentiation. It allows for multi-product firms, alternative hypotheses on preferences and market structure as well as on the production function and processes for productivity and demand. At the same time, our framework is parsimonious enough to allow retrieving productivity, demand, and markups heterogeneity with relatively little information compared to other demand systems models. It also builds upon firm-level data on physical production that is becoming increasingly available to researchers (Belgium, Brazil, Chile, Denmark, France, India, UK and the US to name a few countries). Both elements provide a wide scope of applications to our framework. At the same time our approach does not simply allow recovering a consistent measure of TFP while having some other heterogeneities in the background. It also enables measuring all of the three heterogeneities we consider and potentially confront them with many research questions.

Our analysis has policy implications both at the micro and macro level. At the micro level it makes a big difference to know that some firms or industries lack in competitiveness because of poor physical TFP (due for example to low expenditure in process R&D) or poor product quality (due for example to low expenditure in product R&D). At the macro level our framework allows analyzing aggregate revenue productivity cycles, such as the severe downturn of EU countries' revenue productivity since the financial crisis, not only in terms of changes in some underlying production capacity of the economy, but also as changes in markups and demand.

References

Ackerberg, D., Benkard, C. L., Berry, S., and Pakes, A. (2007). Econometric tools for analyzing market outcomes. *Handbook of econometrics*, 6:4171–4276.

Ackerberg, D., Caves, K., and Frazer, G. (2016). Identification properties of recent production function estimators. *Econometrica*, Forthcoming.

Anderson, S. P., De Palma, A., and Thisse, J. F. (1992). *Discrete choice theory of product differentiation*. MIT press.

Atalay, E. (2014). Materials prices and productivity. *Journal of the European Economic Association*, 12(3):575–611.

Atkeson, A. and Burstein, A. (2008). Pricing-to-market, trade costs, and international relative prices. *The American Economic Review*, 98(5):1998–2031.

Behrens, K., Corcos, G., and Mion, G. (2013). Trade crisis? what trade crisis? *Review of economics and statistics*, 95(2):702–709.

Behrens, K., Mion, G., Murata, Y., and Südekum, J. (2014). Trade, wages, and productivity. *International Economic Review*, 55(4):1305–1348.

Bernard, A., Jensen, B., Redding, S., and Schott, P. (2012a). The empirics of firm heterogeneity and international trade. *Annual Review of Economics*, 4:283–313.

Bernard, A. B., Blanchard, E. J., Van Beveren, I., and Vandenbussche, H. Y. (2012b). Carry-along trade. NBER Working Papers 18246.

Bernard, A. B. and Jensen, J. B. (1999). Exceptional exporter performance: Cause, effect, or both? *Journal of International Economics*, 47(1):1–25.

Berry, S., Levinsohn, J., and Pakes, A. (2004). Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of Political Economy*, 112(1):68–105.

Caliendo, L., Mion, G., Opromolla, L. D., and Rossi-Hansberg, E. (2015). Productivity and organization in portuguese firms. NBER Working Papers 21811.

De Loecker, J. (2011). Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity. *Econometrica*, 79(5):1407–1451.

De Loecker, J., Fuss, C., and Van Biesebroeck, J. (2014). International competition and firm performance: Evidence from belgium. NBB Working Papers 269.

De Loecker, J., Goldberg, P. K., Khandelwal, A. K., and Pavcnik, N. (2016). Prices, markups and trade reform. *Econometrica*, Forthcoming.

De Loecker, J. and Warzynski, F. (2012). Markups and firm-level export status. *American Economic Review*, 102(6):2437–71.

Dhingra, S. and Morrow, J. (2012). Monopolistic competition and optimum product diversity under firm heterogeneity. *Mimeo*.

Dhyne, E., Petrin, A., Smeets, V., and Warzynski, F. (2014). Import competition, productivity, and multi-product firms. NBB Working Papers 268.

Di Comite, F., Thisse, J.-F., and Vandenbussche, H. (2014). Verti-zontal differentiation in export markets. *Journal of International Economics*, 93(1):50–66.

Dobbelaere, S. and Mairesse, J. (2013). Panel data estimates of the production function and product and labor market imperfections. *Journal of Applied Econometrics*, 28(1):1–46.

Doraszelski, U. and Jaumandreu, J. (2013). R&d and productivity: Estimating endogenous productivity. *The Review of Economic Studies*, 80(4):1338–1383.

Fajgelbaum, P., Grossman, G. M., and Helpman, E. (2011). Income distribution, product quality, and international trade. *Journal of Political Economy*, 119(4):721–765.

Feenstra, R. C. (2003). A homothetic utility function for monopolistic competition models, without constant price elasticity. *Economics Letters*, 78(1):79–86.

Feenstra, R. C. and Romalis, J. (2014). International prices and endogenous quality. *The Quarterly Journal of Economics*, 129(2):477–527.

Foster, L., Haltiwanger, J., and Syverson, C. (2008). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review*, 98(1):394–425.

Garcia-Marin, A. and Voigtländer, N. (2014). Exporting and plant-level efficiency gains: It's in the measure. NBER Working Papers 19033.

Grieco, P., Li, S., and Zhang, H. (2014). Production function estimation with unobserved input price dispersion. Technical report, Working paper, Pennsylvania State University, State College.

Hall, R. E. (1986). Market structure and macroeconomic fluctuations. *Brookings papers on economic activity*, 2:285–338.

Helpman, E., Rogoff, K., and Gopinath, G. (2015). *Handbook of international economics*, volume 4. Elsevier, Forthcoming.

Hottman, C., Redding, S. J., and Weinstein, D. E. (2016). Quantifying the sources of firm heterogeneity. *Quarterly Journal of Economics*, Forthcoming.

Jovanovic, B. (1982). Selection and the evolution of industry. *Econometrica*, pages 649–670.

Klette, T. J. and Griliches, Z. (1996). The inconsistency of common scale estimators when output prices are unobserved and endogenous. *Journal of Applied Econometrics*, 11(4):343–361.

Levinsohn, J. and Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies*, 70(2):317–341.

Lucas Jr, R. E. (1978). On the size distribution of business firms. *The Bell Journal of Economics*, pages 508–523.

Martin, R. (2014). Firm level production function estimation with firm specific productivity, demand and market power. *Mimeo*.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, 71(6):1695–1725.

Mion, G. and Zhu, L. (2013). Import competition from and offshoring to china: A curse or blessing for firms? *Journal of International Economics*, 89(1):202–215.

Muûls, M. (2015). Exporters, importers and credit constraints. *Journal of International Economics*, 95(2):333 – 343.

Muûls, M. and Pisu, M. (2009). Imports and Exports at the Level of the Firm: Evidence from Belgium. *The World Economy*, 32(5):692–734.

Nocke, V. and Schutz, N. (2016). Multiproduct-firm oligopoly: An aggregative games approach. University of Mannheim Mimeo.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):l263–l297.

Roberts, M. J., Xu, D. Y., Fan, X., and Zhang, S. (2016). The role of firm factors in demand, cost, and export market selection for chinese footwear producers. *Mimeo*.

Rodríguez-López, J. A. (2011). Prices and exchange rates: A theory of disconnect. *The Review of Economic Studies*, 78(3):1135–1177.

Spence, M. (1976). Product selection, fixed costs, and monopolistic competition. *The Review of Economic Studies*, 43(2):217–235.

Syverson, C. (2004). Market structure and productivity: A concrete example. *Journal of Political Economy*, 112(6):1181–1222.

Syverson, C. (2011). What determines productivity? *Journal of Economic Literature*, 49(2):326–365.

Van Ark, B. and Monnikhof, E. (1996). Size distribution of output and employment: a data set for manufacturing industries in five oecd countries, 1960s-1990. Technical report, OECD Publishing.

Van Biesebroeck, J. (2005). Firm size matters: Growth and productivity growth in african manufacturing. *Economic Development and Cultural Change*, 53(3):545–583.

Verhoogen, E. A. (2008). Trade, quality upgrading, and wage inequality in the mexican manufacturing sector. *The Quarterly Journal of Economics*, 123(2):489–530.

Wooldridge, J. M. (2009). On estimating firm-level production functions using proxy variables to control for unobservables. *Economics Letters*, 104(3):112–114.

Zhelobodko, E., Kokovin, S., Parenti, M., and Thisse, J.-F. (2012). Monopolistic competition: Beyond the constant elasticity of substitution. *Econometrica*, 80(6):2765–2784.

Table 1: Mean and Standard Deviation

Industry	μ	λ	a	α_M	α_L	γ
			Mean			
151	1.257	-13.230	14.000	0.976	0.170	1.109
212	1.304	-13.050	14.120	0.864	0.271	1.146
266	1.214	-13.190	14.160	0.813	0.238	1.068
361	1.411	-13.320	14.310	0.828	0.413	1.273
			Standard Deviation			
151	0.174	0.614	0.503	0.063*	0.011*	0.039*
212	0.229	0.768	0.591	0.048*	0.015*	0.032*
266	0.235	0.825	0.630	0.060*	0.018*	0.051*
361	0.313	1.163	1.160	0.033*	0.017*	0.036*

Notes: * indicates bootstrapped standard errors (200 replications).

Table 2: Correlations

		151				212	
	μ	λ	a		μ	λ	a
μ	1			μ	1		
λ	0.417***	1		λ	0.608***	1	
a	0.187***	-0.691***	1	a	-0.063	-0.663***	1
p	0.020	0.742***	-0.941***	p	0.213***	0.691***	-0.916***
		266				361	
	μ	λ	a		μ	λ	a
μ	1			μ	1		
λ	0.611***	1		λ	0.072**	1	
a	-0.115***	-0.767***	1	a	-0.088***	-0.910***	1
p	0.143***	0.791***	-0.956***	p	0.077**	0.926***	-0.940***

Notes: *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Regression of markup μ on a, λ and log capital k

Industry	151	212	266	361
a	.3369***	.2579***	.332***	.0155
	(.015)	(.0314)	(.0141)	(.0356)
λ	.3259***	.3297***	.374***	.0369
	(.0128)	(.014)	(.0089)	(.0355)
k	-.0368***	-.0447***	-.0259***	-.0801***
	(.0044)	(.0055)	(.005)	(.006)
# Obs	1235	770	1402	1566
R^2	.6625	.6491	.6961	.1112

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Regression of a, λ and μ on their time lag

Industry	151	212	266	361
		a		
Lag a	.9175***	.9477***	.9138***	.8372***
	(.0224)	(.0211)	(.0172)	(.0202)
R^2	.8535	.8925	.8825	.7342
		λ		
Lag λ	.8736***	.8944***	.9169***	.8231***
	(.0238)	(.0246)	(.0204)	(.0212)
R^2	.8135	.8058	.8396	.7096
		μ		
Lag μ	.8013***	.7949***	.8493***	.8743***
	(.0309)	(.0264)	(.019)	(.0225)
R^2	.6869	.7244	.7381	.7338

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Regression of log price p on a, λ and μ and log capital k

Industry	151	212	266	361
a	-.8891***	-.6858***	-.6195***	-.4877***
	(.0364)	(.0571)	(.0243)	(.0284)
λ	.0692*	.2005***	.286***	.48***
	(.0346)	(.0508)	(.0235)	(.0268)
μ	.4421***	-.0371	-.4334***	-.1729***
	(.0881)	(.1275)	(.0564)	(.0513)
k	-.0729***	-.1229***	-.0674***	-.1351***
	(.003)	(.0045)	(.0031)	(.0071)
# Obs	1235	770	1402	1566
R^2	.9496	.9357	.9415	.932

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 6: Regression of log quantity q on a, λ and μ and log capital k

Industry	151	212	266	361
a	1.828***	1.074***	1.36***	.5458***
	(.1563)	(.1975)	(.102)	(.0676)
λ	.9423***	.5549**	.7749***	-.3654***
	(.1543)	(.1724)	(.1002)	(.0665)
μ	-4.388***	-1.958***	-1.881***	-.7247***
	(.4769)	(.4129)	(.2218)	(.0991)
k	.5814***	.7576***	.4271***	.6378***
	(.0223)	(.029)	(.0198)	(.0217)
# Obs	1235	770	1402	1566
R^2	.6477	.7426	.5734	.6887

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 7: Regression of log revenue r on a, λ and μ and log capital k

Industry	151	212	266	361
a	.9393***	.388**	.741***	.0581
	(.1392)	(.1474)	(.0836)	(.0535)
λ	1.012***	.7554***	1.061***	.1146*
	(.1353)	(.1332)	(.0803)	(.051)
μ	-3.946***	-1.995***	-2.315***	-.8976***
	(.4266)	(.3276)	(.1818)	(.0673)
k	.5085***	.6346***	.3596***	.5028***
	(.0218)	(.022)	(.0189)	(.0198)
# Obs	1235	770	1402	1566
R^2	.6223	.733	.555	.5435

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 8: DLW TFP revenue based regressed on a, λ and μ: BETA COEFFICIENTS

Industry	151	212	266	361
a	1.236***	.9342***	.9488***	.9733***
	(.0197)	(.0305)	(.0171)	(.0167)
λ	1.561***	1.448***	1.525***	1.063***
	(.02)	(.0265)	(.0183)	(.0158)
μ	-.6635***	-.485***	-.279***	-.2395***
	(.0514)	(.063)	(.0427)	(.0226)
# Obs	1233	769	1402	1561
R^2	0.558	0.619	0.639	0.263

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 9: OP TFP revenue based regressed on a, λ and μ: BETA COEFFICIENTS

Industry	151	212	266	361
a	.4062***	.3995***	.6407***	.6387***
	(.011)	(.0124)	(.0101)	(.0079)
λ	.3678***	.461***	.821***	.6581***
	(.011)	(.0111)	(.0095)	(.0077)
μ	.4246***	.4022***	.1584**	.541***
	(.0282)	(.0283)	(.0214)	(.0094)
# Obs	1235	770	1402	1566
R^2	0.476	0.496	0.447	0.366

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 10: FHS TFP revenue based regressed on a, λ and μ: BETA COEFFICIENTS

Industry	151	212	266	361
a	.254**	.1858	.3797***	.2327**
	(.0155)	(.0217)	(.0205)	(.0098)
λ	.2235*	.2986**	.5891***	.2656**
	(.0156)	(.0184)	(.0194)	(.0097)
μ	.2115**	.2716***	.0241	.4876***
	(.044)	(.047)	(.0468)	(.0127)
# Obs	1235	770	1402	1566
R^2	0.159	0.223	0.168	0.250

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 11: OLS TFP revenue based regressed on a, λ and μ: BETA COEFFICIENTS

Industry	151	212	266	361
a	.4153***	.3812***	.6245***	.6486***
	(.0106)	(.0109)	(.0102)	(.0077)
λ	.3693***	.4339***	.783***	.6583***
	(.0107)	(.0093)	(.0095)	(.0072)
μ	.4255***	.4519***	.1813***	.572***
	(.0284)	(.0234)	(.0232)	(.0085)
# Obs	1235	770	1402	1566
R^2	0.464	0.523	0.429	0.392

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 12: Comparison of average markups between our methodology and DLW

Industry	μ	markup DLW1	markup DLW2
151	1.257	1.283	1.464
212	1.304	1.050	1.189
266	1.214	1.304	1.533
361	1.411	1.324	2.582

Table 13: Correlation between λ and FHS demand shocks

Industry	correlation
151	0.294***
212	0.414***
266	0.238***
361	0.231***

Notes: *** p<0.01, ** p<0.05, * p<0.1.

Table 14: Correlation between our residual demand shocks and FHS demand shocks

Industry	correlation
151	0.217***
212	0.299***
266	0.058
361	0.162***

Notes: *** p<0.01, ** p<0.05, * p<0.1.

Table 15: Export status regressed on OP revenue-based TFP: BETA COEFFICIENTS

Industry	151	212	266	361
OP TFP Revenue	.4428***	.3651***	.2075***	.4219***
	(.0128)	(.0202)	(.0232)	(.0153)
# Obs	1235	770	1402	1566
R^2	0.223	0.161	0.0512	0.206

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 16: Export status regressed on DLW revenue-based TFP: BETA COEFFICIENTS

Industry	151	212	266	361
DLW TFP revenue	.4249***	.4094***	.4072***	.4378***
	(.0705)	(.0603)	(.0533)	(.0319)
# Obs	1233	769	1402	1561
R^2	0.209	0.196	0.172	0.219

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 17: Export status regressed on OP quantity-based TFP: BETA COEFFICIENTS

Industry	151	212	266	361
OP TFP quantity	.3646***	.2247***	.0359	.2706***
	(.0134)	(.0172)	(.0164)	(.0079)
# Obs	1235	770	1402	1566
R^2	0.161	0.0804	0.0105	0.102

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 18: Export status regressed on DLW quantity-based TFP: BETA COEFFICIENTS

Industry	151	212	266	361
DLW TFP quantity	.2649***	.2572***	.0665*	.3248***
	(.0256)	(.0243)	(.0215)	(.0083)
# Obs	1233	769	1402	1561
R^2	0.0995	0.0961	0.0136	0.134

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 19: Export status regressed on a, λ and μ: BETA COEFFICIENTS

Industry	151	212	266	361
a	.5184***	.3714***	.4297***	.108
	(.0523)	(.0388)	(.0407)	(.0243)
λ	.6062***	.6067***	.7451***	.0205
	(.0481)	(.0364)	(.0402)	(.0244)
μ	-.4762***	-.3782***	-.2456***	-.221***
	(.1321)	(.1027)	(.0889)	(.0393)
# Obs	1235	770	1402	1566
R^2	0.126	0.119	0.117	0.0889

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 20: Log number of employees regressed on OP revenue-based TFP: BETA COEFFICIENTS

Industry	151	212	266	361
OP TFP Revenue	.5882***	.7237***	.557***	.6863***
	(.036)	(.0523)	(.0456)	(.0383)
# Obs	1235	770	1402	1566
R^2	0.348	0.524	0.316	0.470

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 21: Log number of employees regressed on DLW revenue-based TFP: BETA COEFFICIENTS

Industry	151	212	266	361
DLW TFP revenue	.7855***	.7664***	.7994***	.8514***
	(.1171)	(.107)	(.0789)	(.0505)
# Obs	1233	769	1402	1561
R^2	0.620	0.590	0.640	0.719

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 22: Log number of employees regressed on OP quantity-based TFP: BETA COEFFICIENTS

Industry	151	212	266	361
OP TFP quantity	.353***	.3565***	.2492***	.212***
	(.0314)	(.0487)	(.0386)	(.0153)
# Obs	1235	770	1402	1566
R^2	0.131	0.135	0.0751	0.0461

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 23: Log number of employees regressed on DLW quantity-based TFP: BETA COEFFICIENTS

Industry	151	212	266	361
DLW TFP quantity	.3285***	.3737***	.1384***	.3854***
	(.0461)	(.0589)	(.0398)	(.0169)
# Obs	1233	769	1402	1561
R^2	0.113	0.148	0.0329	0.149

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 24: Log number of employees regressed on a, λ and μ: BETA COEFFICIENTS

Industry	151	212	266	361
a	1.11***	1.021***	.9931***	.341***
	(.1236)	(.1523)	(.0764)	(.0427)
λ	.7612***	.4728***	.4961***	.2496***
	(.1339)	(.1643)	(.0752)	(.0439)
μ	-.5233***	-.5128***	-.2275***	-.2184***
	(.275)	(.3911)	(.1655)	(.0636)
# Obs	1235	770	1402	1566
R^2	0.291	0.290	0.289	0.0708

Notes: Time dummies are included in estimations but are not reported here. Bootstrapped standard errors in parenthesis (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Figure 1: How Important is Demand Heterogeneity? Plot of Log Price and Log Quantity

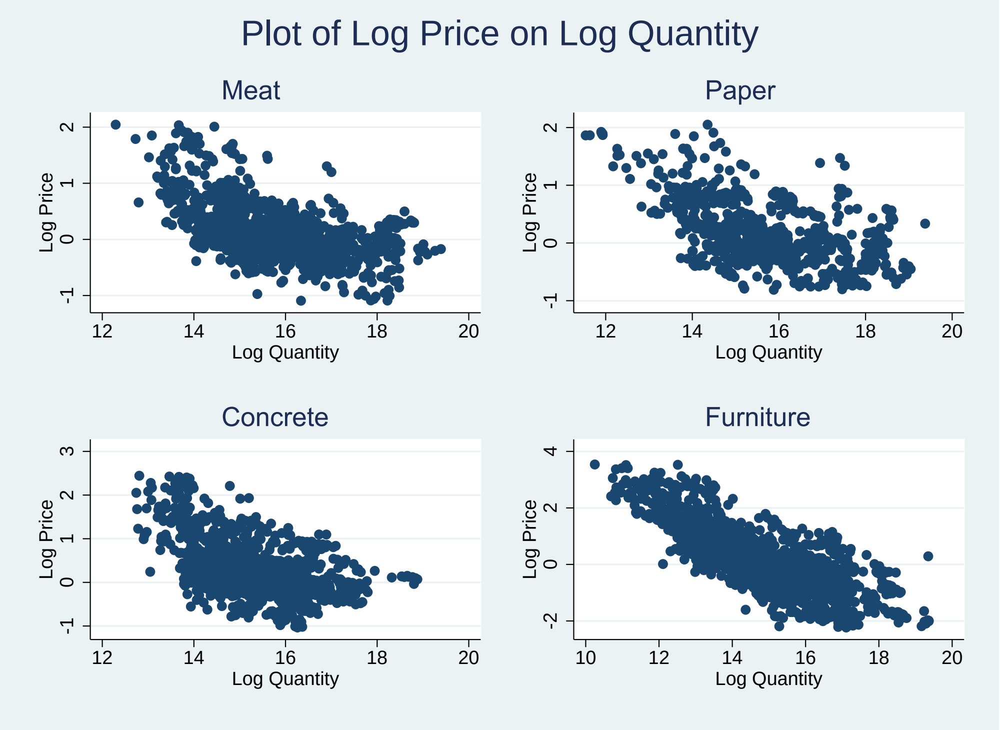
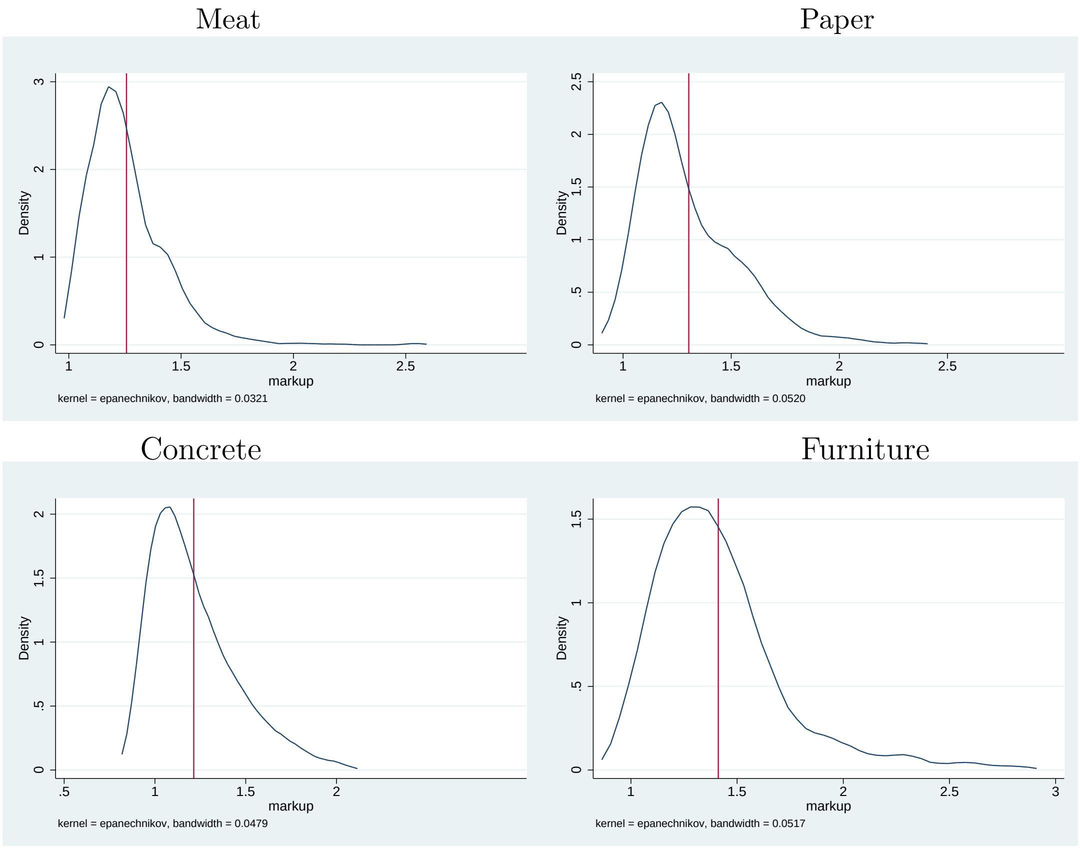


Figure 2: Density of μ. Red vertical line corresponds to the mean markup



Meat

Paper

Concrete

Furniture

Figure 3: Density of λ and a

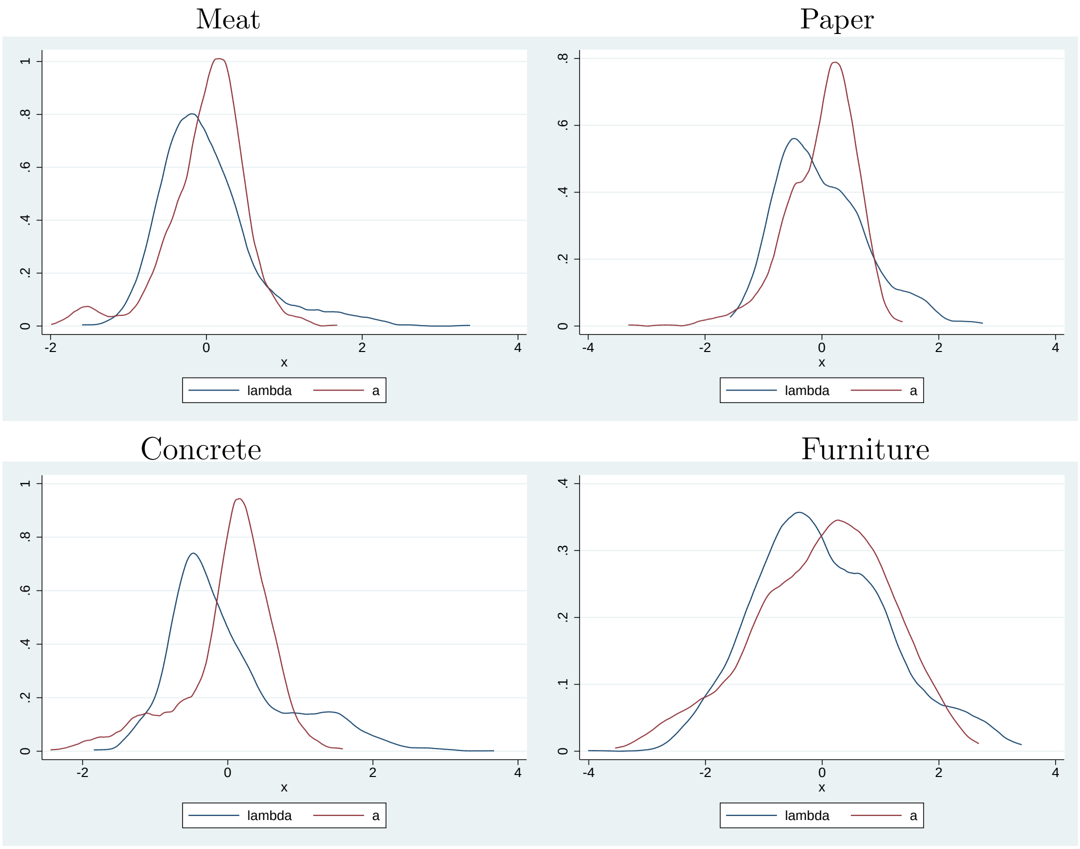
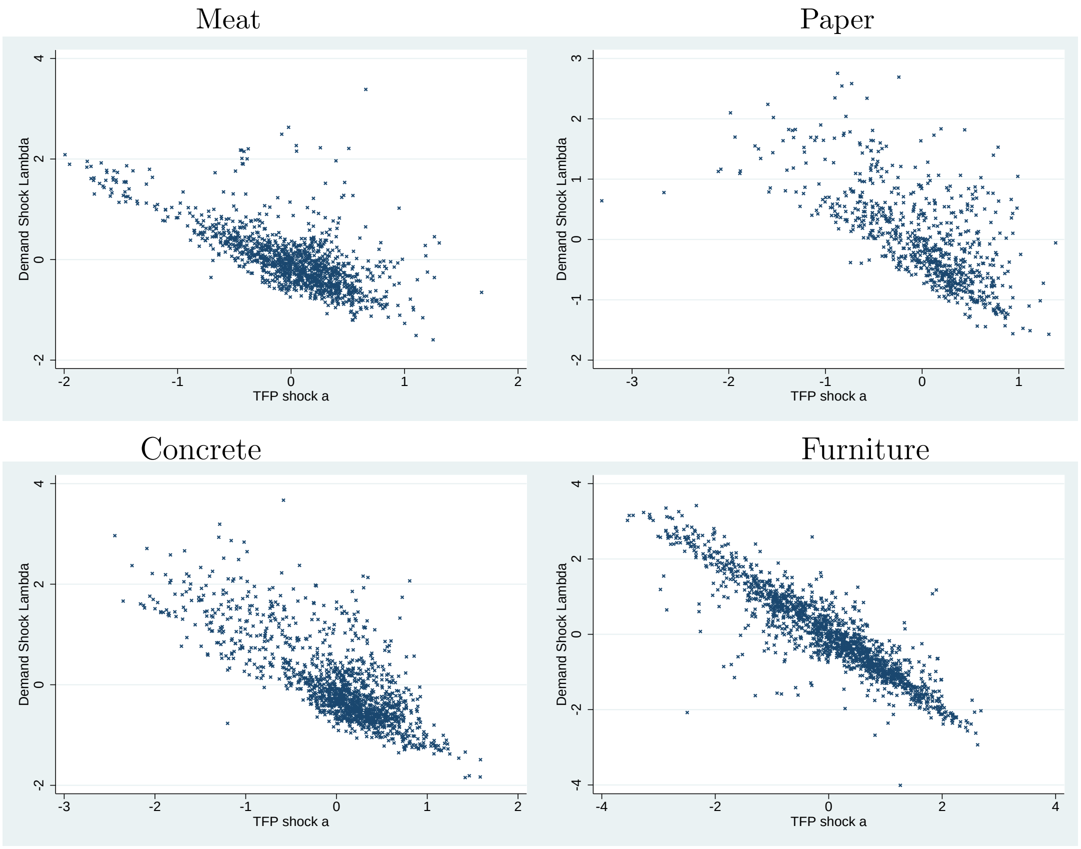


Figure 4: Plot of λ and a



Meat

Paper

Concrete

Furniture

Figure 5: Plot of λ and FHS demand shocks

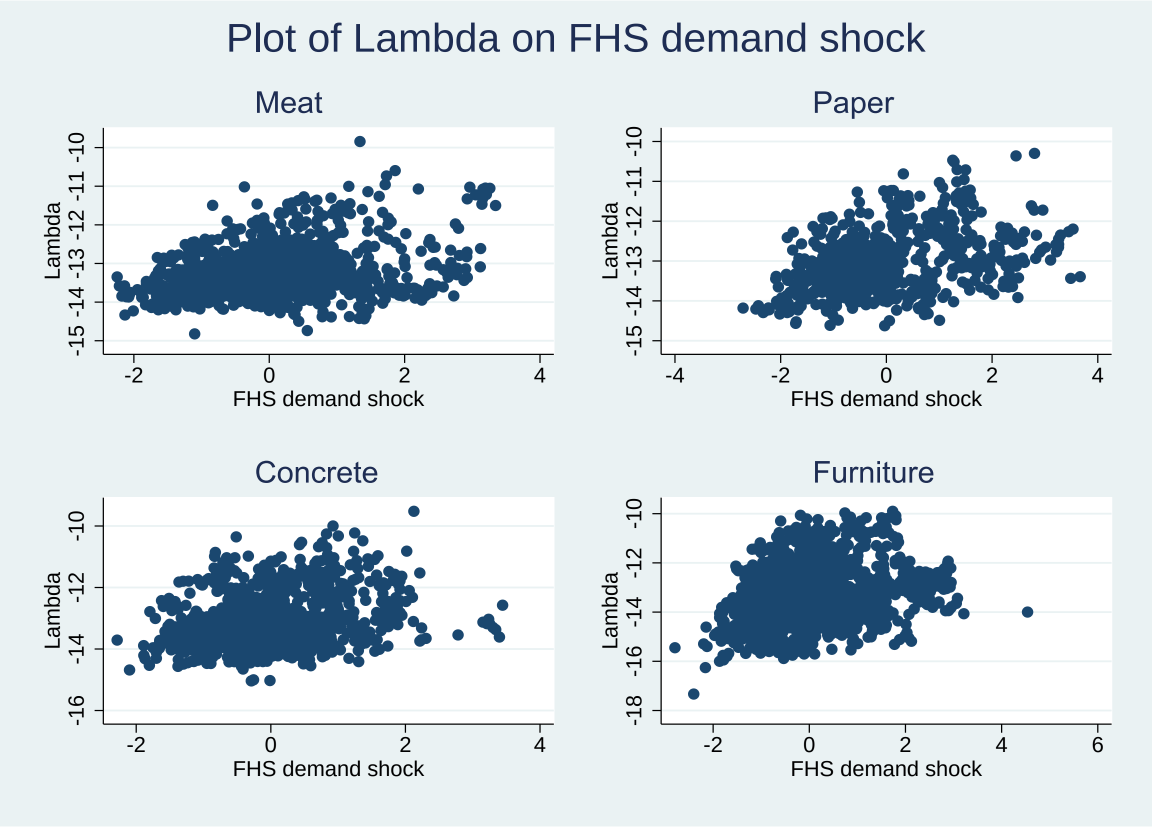
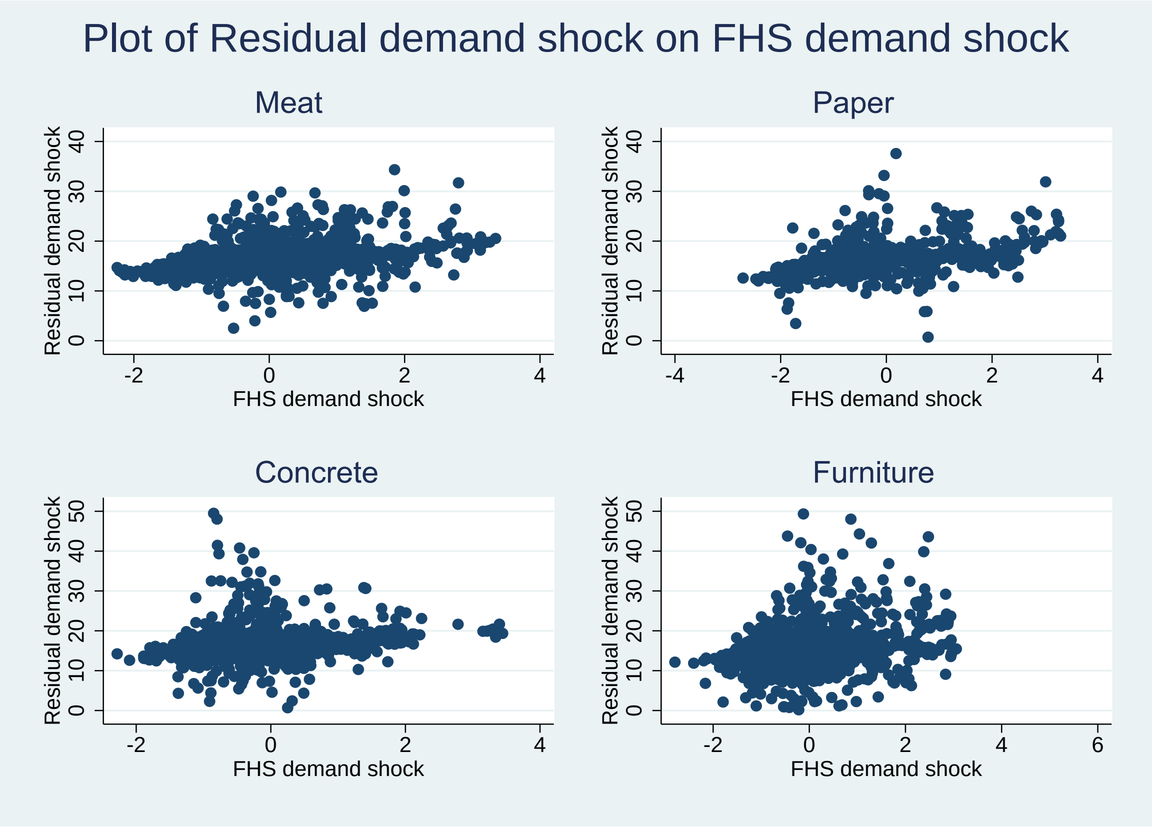


Figure 6: Plot of our residual demand shocks and FHS demand shocks



Appendix

In Appendices A to E we show how to extend our framework to more general preferences, forms of imperfect competition other than monopolistic competition, to a wider set of production functions and processes for productivity and demand shocks as well as to multi-product firms. In Appendix F we instead spell out the assumptions allowing us to deal with the issue of aggregation.

A Preferences

There are various ways, within the representative consumer framework, to extend our model and estimation methodology to preferences other than the baseline generalized CES case. One way is to identify a class of preferences under which the key equation (9) holds as a first-order linear approximation and then to modify the estimation procedure accordingly. Another possibility is to fully specify an alternative preference structure and work out the corresponding algebra for the estimation equations. We discuss the former case in Section A.1 while the latter case is presented in Section A.2. In Section A.3 we instead characterize a class of random utility models and provide conditions under which our model can be applied.

A.1 First-order linear approximation

The key property we want preferences to satisfy is that the elasticity of prices with respect to output quantity differs from the elasticity of prices with respect to the demand shock by one: $\frac{\partial p_i}{\partial \lambda_i} = \frac{\partial p_i}{\partial q_i} + 1$. There are different ways of achieving this. One way is to start from direct utility. Consider a representative consumer who maximises a differentiable utility function $U(.)$ subject to budget B:

$$\max_{Q} \left\{ U\left(\tilde{Q}\right) \right\} \text{ s.t. } \int_i P_i Q_i \mathrm{d}i - B = 0 \tag{A-1}$$

where $\tilde{Q}$ is a vector of elements $\Lambda_i Q_i$. Therefore, while the representative consumer chooses quantities Q, these quantities enter into the utility function as $\tilde{Q}$ and Λ_i can be interpreted as a measure of quality for variety i. For example, in the symmetric (with respect to $\tilde{Q}$) varieties case, the representative consumer would be indifferent between having one more unit of a variety with $\Lambda_i = \overline{\Lambda}$ or $\overline{\Lambda}$ more units of a a variety with $\Lambda_i = 1$.

The first order conditions of the utility maximization problem imply:

$$\frac{\partial U}{\partial Q_i} = \frac{\partial U}{\partial \tilde{Q}_i}\frac{\partial \tilde{Q}_i}{\partial Q_i} = \frac{\partial U}{\partial \tilde{Q}_i}\Lambda_i = \kappa P_i$$

where κ is a Lagrange multiplier and $\frac{\partial \tilde{Q}_i}{\partial Q_i} = \Lambda_i$. Taking logs we have:

$$\ln \frac{\partial U}{\partial \tilde{Q}_i} + \lambda_i = \ln \kappa + p_i. \tag{A-2}$$

Solving all of these conditions would give us demand functions for all varieties including that of firm i. However, even if we knew the exact form of $U(.)$, this might be tricky to work out. Nonetheless, equation (A-2) already tells us a lot about the shape of such demand functions. On the one hand, differentiating both sides with respect to q_i yields:

$$\frac{\partial p_i}{\partial q_i} = \frac{\partial \ln \frac{\partial U}{\partial \tilde{Q}_i}}{\partial q_i} = \frac{\partial \ln \frac{\partial U}{\partial \tilde{Q}_i}}{\partial \tilde{q}_i} \frac{\partial \tilde{q}_i}{\partial q_i} = \frac{\partial \ln \frac{\partial U}{\partial \tilde{Q}_i}}{\partial \tilde{q}_i} \tag{A-3}$$

where $\frac{\partial \tilde{q}_i}{\partial q_i} = 1$ and $\frac{\partial p_i}{\partial q_i} \equiv -\frac{1}{\eta_i}$. On the other hand, keeping in mind that $\frac{\partial \tilde{q}_i}{\partial \lambda_i} = 1$ differentiation of both sides with respect to λ_i gives:

$$\frac{\partial p_i}{\partial \lambda_i} = \frac{\partial \ln \frac{\partial U}{\partial \tilde{Q}_i}}{\partial \lambda_i} + 1 = \frac{\partial \ln \frac{\partial U}{\partial \tilde{Q}_i}}{\partial \tilde{q}_i} \frac{\partial \tilde{q}_i}{\partial \lambda_i} + 1 = \frac{\partial \ln \frac{\partial U}{\partial \tilde{Q}_i}}{\partial \tilde{q}_i} + 1 = 1 - \frac{1}{\eta_i}$$

i.e., the elasticity of the price with respect to quantity differs from the elasticity of the price with respect to the demand shock by one. This is the key property needed in our framework.

Let us now consider the implications of these results. Using equation (A-2) we can write log revenue r_i (up to a constant) as:

$$r_i = p_i + q_i = \ln \frac{\partial U}{\partial \tilde{Q}_i} + \lambda_i + q_i = \ln \frac{\partial U}{\partial \tilde{Q}_i} + \tilde{q}_i.$$

Differentiating both sides with respect to $\tilde{q}_i$ and making use of equation (A-3) we have:

$$\frac{\partial r_i}{\partial \tilde{q}_i} = \frac{\partial \ln \frac{\partial U}{\partial \tilde{Q}_i}}{\partial \tilde{q}_i} + 1 = -\frac{1}{\eta_i} + 1 = \frac{1}{\mu_i}$$

and so we finally get:

$$\Delta r_i \approx \frac{1}{\mu_i} \Delta \tilde{q}_i = \frac{1}{\mu_i} \Delta(q_i + \lambda_i). \tag{A-4}$$

Therefore, for any preferences structure that can be used to model monopolistic competition and that can be described by a well-behaved differentiable direct utility we can, starting from the baseline formulation $U(Q)$, introduce quality in such a way that equation (A-4) is satisfied. The advantage of equation (A-4) is that it can be directly used for estimations without the need to explicitly solve for the demand functions of the different varieties.

One interesting example is the Generalized CES (Spence, 1976):

$$U(\tilde{Q}) = \int_{i \in I} a_i \left(\tilde{Q}_i\right)^{b_i} \mathrm{d}i = \int_{i \in I} a_i \Lambda_i^{b_i} \left(Q_i\right)^{b_i} \mathrm{d}i$$

where $b_i = 1 - \frac{1}{\eta_i}$. If we further impose $a_i = \frac{\eta_i}{\eta_i - 1}$ not only equation (A-4) holds but we actually get equation (9): $r_i = \frac{1}{\mu_i}(q_i + \lambda_i)$. This is our benchmark case. Other examples of preferences falling within our class include the CARA preferences used in Behrens et al. (2014) as well as the Translog preferences featuring in Feenstra (2003) and Rodríguez-López (2011). Contrary to CARA preferences in the case of the Translog there is no closed-form expression for the utility function and the demand system is derived directly from the expenditure function. Yet the utility function does exists and it is well behaved and so it falls within our class. A workable Translog framework can be readily obtained by considering that everything works as if firms sell quantities $\tilde{Q}_{ij} = Q_{ij}\Lambda_{ij}$ while charging prices $\tilde{P}_{ij} = P_{ij}/\Lambda_{ij}$. Also note that $\tilde{P}_{ij}\tilde{Q}_{ij} = P_{ij}Q_{ij}$ by construction and so total consumer expenditure E in terms of $\tilde{Q}_{ij}$ is the same as in terms of Q_{ij}. Therefore, one simply needs to substitute log prices $\tilde{p}_{ij}$ with $p_i - \lambda_i$ in the Translog expenditure function.

Another way of getting the key property we need is to start from demand functions and work out some constraints. Consider, for example, the Generalised Quadratic Utility (Di Comite et al., 2014):

$$U(Q) = \int_{i \in I} a_i Q(i)\mathrm{d}i - \frac{1}{2}\int_{i \in I} b_i \left[Q(i)\right]^2 \mathrm{d}i - \frac{c_i}{2}\left[\int_{i \in I} Q(i)\mathrm{d}i\right]^2 + Q_0$$

where Q_0 is a numraire good. Because of the presence of a numraire good the Generalised Quadratic Utility does not fit the above described framework. Yet, one can easily derive the inverse demand function in this case and impose constraints on a_i, b_i and c_i such that $\frac{\partial p_i}{\partial \lambda_i} = \frac{\partial p_i}{\partial q_i} + 1$. This is obtained by setting $a_i = a_1\Lambda_i$, $b_i = b_1(\Lambda_i)^2$ and $c_i = c_1\Lambda_i$, where a_1, b_1 and c_1 are positive constants. The inverse demand will thus be:

$$P_i = d_1\Lambda_i - b_1\Lambda_i^2 Q_i,$$

where $d_1 = \left(a_1 - c_1\overline{Q}\right)$ and $\overline{Q} = \int_{i \in I} Q(i)$. The demand shock enters the demand function in order to ensure that $\frac{\partial p_i}{\partial \lambda_i} = -\frac{b_1 Q_i \Lambda_i^2}{P_i} + 1 = \frac{\partial p_i}{\partial q_i} + 1$. Therefore:

$$\frac{\partial r_i}{\partial q_i} = \frac{\partial q_i}{\partial q_i} + \frac{\partial p_i}{\partial q_i} = 1 + \frac{\partial p_i}{\partial q_i} = \frac{1}{\mu_i} = \frac{\partial p_i}{\partial \lambda_i} = \frac{\partial q_i}{\partial \lambda_i} + \frac{\partial p_i}{\partial \lambda_i} = \frac{\partial r_i}{\partial \lambda_i}$$

and we obtain from this equation (A-4):

$$\Delta r_i \approx \frac{1}{\mu_i}\left(\Delta q_i + \Delta\lambda_i\right) = \frac{1}{\mu_i}\Delta\left(q_i + \lambda_i\right) = \frac{1}{\mu_i}\Delta\tilde{q}_i.$$

Moving to the estimation strategy one needs to decide which type of local approximation

(Δ) to use. One possibility is consider differences in log revenue (as well as in other variables needed in the estimation) between a reference firm r and any other firm i. This approach has the advantage of allowing us to measure demand, productivity ad markups for all firms as deviations with respect to firm r. However, the drawback is that for firms that are very different from r the first-order approximation might not be very satisfactory. Yet, one could additionally invoke the mean value theorem and consider average derivatives, i.e., the average markup and consequently average revenue shares of variable factors between firm r and firm i, to improve the quality of the approximation. An alternative approach, that we fully develop below, consists instead in employing time changes. In this respect our findings based on the benchmark generalized CES case show a strong time persistence of the model's fundamentals: productivity shocks, demand shocks and markups. The small time changes should provide a reasonably good base for our first-order approximation. However, the drawback of this approach is that we can only measure time changes of demand shocks within a firm. As for productivity and markups we can instead measure their level for all firms and years.

In what follows we use, for example, $\Delta r_{it}{=}r_{it}-r_{it-1}$ for revenue. Equations (7) and (8) still hold in this broader setting and so, by proceeding as before, we get from equation (A-4), which is the equivalent of equation (9), to the following expression:

$$\widehat{LHS}_{it} \equiv \frac{\Delta r_{it} - s_{Lit}\left(\Delta l_{it} - \Delta k_{it}\right) - s_{Mit}\left(\Delta m_{it} - \Delta k_{it}\right)}{s_{Mit}} = \frac{\gamma}{\alpha_M}\Delta k_{it} + \frac{1}{\alpha_M}\left(\Delta a_{it} + \Delta\lambda_{it}\right) \tag{A-5}$$

which is the equivalent of equation (10). Combining equations (8) and (A-4) we then have:

$$\Delta\lambda_{it-1} = \Delta r_{it-1}\frac{\alpha_M}{s_{Mit-1}} - \Delta q_{it-1}$$

while plugging equations (8) and (A-4) into equation (A-5) and re-arranging yields:

$$\Delta a_{it-1} = \alpha_M \widehat{LHS}_{it-1} - \gamma\Delta k_{it-1} - \left(\Delta r_{it-1}\frac{\alpha_M}{s_{Mit-1}} - \Delta q_{it-1}\right). \tag{A-6}$$

Finally, by substituting the last two expressions as well as equation (11) into equation (A-5) we obtain:

$$\begin{aligned} \widehat{LHS}_{it} &= \frac{\gamma}{\alpha_M}\Delta k_{it} + \phi_a \widehat{LHS}_{it-1} - \phi_a\frac{\gamma}{\alpha_M}\Delta k_{it-1} \\ &+ (\phi_\lambda - \phi_a)\left(\frac{\Delta r_{it-1}}{s_{Mit-1}} - \frac{1}{\alpha_M}\Delta q_{it-1}\right) + \frac{1}{\alpha_M}\left(\Delta\nu_{ait} + \nu_{\Delta\lambda it}\right) \end{aligned} \tag{A-7}$$

which is the equivalent of equation (14). As in the case of equation (14), equation (A-7) can be estimated using a linear regression setting and an appropriate change in variables. However, contrary to equation (14), we cannot simply use OLS here because, for example, ν_{ait-1} (which is included in $\Delta\nu_{ait}$) is correlated with $\widehat{LHS}_{it-1}$, $\frac{\Delta r_{it-1}}{s_{Mit-1}}$ and Δq_{it-1}. A simple way to solve this issue is to instrument $\widehat{LHS}_{it-1}$, $\frac{\Delta r_{it-1}}{s_{Mit-1}}$ and Δq_{it-1} with their respective time lags. At the same time one might worry that capital at time t reacts to demand and productivity shocks in $t-1$. In this case one might also instrument Δk_{it} with, for example, Δk_{it-2}.

Estimation of equation (A-7) delivers $\hat{\beta}$ and $\hat{\phi}_a$ that can be used to get an estimate of γ along the lines of what we show in Section 2. To do so, however, we first need to time-difference the Cobb-Douglas production constraint:

$$\Delta q_{it} = \alpha_L \left(\Delta l_{it} - \Delta k_{it}\right) + \alpha_M \left(\Delta m_{it} - \Delta k_{it}\right) + \gamma \Delta k_{it} + \Delta a_{it}$$

and make use of equations (7), (8), (11) and (A-6) to get to:

$$\begin{aligned} \Delta q_{it} &= \frac{\gamma}{\hat{\beta}} \frac{s_{Lit}}{s_{Mit}} \left(\Delta l_{it} - \Delta k_{it}\right) + \frac{\gamma}{\hat{\beta}} \left(\Delta m_{it} - \Delta k_{it}\right) + \gamma \Delta k_{it} \\ &+ \hat{\phi}_a \frac{\gamma}{\hat{\beta}} \widehat{LHS}_{it-1} - \hat{\phi}_a \gamma \Delta k_{it-1} - \hat{\phi}_a \left(\Delta r_{it-1} \frac{\gamma}{\hat{\beta} s_{Mit-1}} - \Delta q_{it-1}\right) + \Delta\nu_{ait}. \quad \text{(A-8)} \end{aligned}$$

Equation (A-8) is the equivalent of equation (18) and can be manipulated in a similar fashion to estimate γ via a linear regression where the only covariate is instrumented with, for example, Δk_{it-1} based on the moment condition $E\left\{\Delta\nu_{ait}\Delta k_{it-1}\right\} = 0$. Productivity shocks and markups can in turn be computed as before:

$$\hat{a}_{it} = q_{it} - \frac{\hat{\gamma}}{\hat{\beta}} \frac{s_{Lit}}{s_{Mit}} \left(l_{it} - k_{it}\right) - \frac{\hat{\gamma}}{\hat{\beta}} \left(m_{it} - k_{it}\right) - \hat{\gamma} k_{it}$$

$$\hat{\mu}_{it} = \frac{\hat{\gamma}}{\hat{\beta} s_{Mit}}.$$

However, as far as demand shocks are concerned, only their change over time within a firm can be measured:

$$\Delta\lambda_{it} = \frac{\hat{\gamma}}{\hat{\beta} s_{Mit}} \Delta r_{it} - \Delta q_{it}.$$

A.2 Exact procedure with non log-linear demand

Here we discuss how our model can be extended, without resorting to any linear approximation, by fully specifying an alternative preference structure. We work out the corresponding algebra for the estimation equations. We are particularly interested in a flexible structure leading to a non-log linear demand. One reason why one might choose such a structure is that it allows for markups varying with equilibrium quantity. Specifically, we look here at an additively separable utility function shaped like the Gaussian CDF:[29]

$$U(\tilde{q}) = \int_{i \in I} \Phi\left(\tilde{q}_i, \beta_0, \beta_1, \beta_2\right) \mathrm{d}i$$

where $\Phi\left(\cdot\right)$ is the Gaussian cdf, i.e.,

$$\Phi\left(\tilde{q}_i\right) = u\left(\tilde{q}_i\right) = \int_{-\infty}^{\tilde{q}_i} \phi\left(\tilde{q}_i\right) \mathrm{d}\tau.$$

The inverse demand function for firm i at time t is consequently:

$$\frac{\phi\left(\tilde{q}_{it}\right)}{Q_{it}\kappa_t} = P_{it}$$

where $\phi\left(\tilde{q}_{it}\right)$ is the Gaussian PDF. In what follows we set $\phi\left(\tilde{q}_{it}\right) = \exp\left(-\beta_2^2\tilde{q}_{it}^3 + \beta_1\tilde{q}_{it} + \beta_0\right)$ but could have equally used more or less involved formulations.

The first thing to note is that the Gaussian utility as specified above implies a downward sloping demand curve for $\beta_1 < 1$. Indeed:

$$\frac{\partial p_{it}}{\partial q_{it}} = -3\beta_2^2\tilde{q}_{it}^2 + \beta_1 - 1 < 0 \text{ for } \beta_1 < 1$$

Moving forward, we ignore terms that are constant across firms and have:

$$\frac{\partial \ln \phi}{\partial \tilde{q}_{it}} = -3\beta_2^2\tilde{q}_{it}^2 + \beta_1 = \frac{\partial p_{it}}{\partial q_{it}} + 1 = \frac{1}{\mu_{it}},$$

as well as:

$$r_{it} = \ln \phi\left(\tilde{q}_{it}\right).$$

Using the definition $\chi_{it} = s_{Lit}\left(l_{it} - k_{it}\right) + s_{Mit}\left(m_{it} - k_{it}\right)$ we obtain:

$$\frac{1}{3}\frac{\partial \ln \phi}{\partial \tilde{q}_{it}}\tilde{q}_{it} = -\beta_2^2\tilde{q}_{it}^3 + \frac{1}{3}\beta_1\tilde{q}_{it} = r_{it} - \frac{2}{3}\beta_1\tilde{q}_{it}$$

[29] See Berhold (1973) for further discussion of the Gaussian CDF as a utility function.

$$\Rightarrow r_{it} = \left(\frac{1}{3}\frac{\partial \ln \phi}{\partial \tilde{q}_{it}} + \frac{2}{3}\beta_1\right)\tilde{q}_{it} = \left(\frac{1}{3}\frac{1}{\mu_{it}} + \frac{2}{3}\beta_1\right)\tilde{q}_{it}$$

$$= \left(\frac{1}{3} + \frac{2}{3}\beta_1\mu_{it}\right)\chi_{it} + \left(\frac{1}{3}\frac{1}{\mu_{it}\beta_1} + \frac{2}{3}\right)\beta_1\gamma k_{it} + \left(\frac{1}{3}\frac{1}{\mu_{it}\beta_1} + \frac{2}{3}\right)\beta_1(a_{it} + \lambda_{it})$$

$$\Rightarrow \frac{r_{it}}{\frac{1}{3}\frac{1}{\mu_{it}\beta_1} + \frac{2}{3}} = \frac{\frac{1}{3} + \frac{2}{3}\beta_1\mu_{it}}{\frac{1}{3}\frac{1}{\mu_{it}\beta_1} + \frac{2}{3}}\chi_{it} + \beta_1\gamma k_{it} + \beta_1(a_{it} + \lambda_{it}).$$

Further note that: $\frac{\frac{1}{3}+\frac{2}{3}\beta_1\mu_{it}}{\frac{1}{3}\frac{1}{\mu_{it}\beta_1}+\frac{2}{3}} = \frac{1+2\beta_1\mu_{it}}{\frac{1}{\mu_{it}\beta_1}+2} = \beta_1\mu_{it} = \beta_1\frac{\alpha_M}{s_{Mit}}$. Hence we can write:

$$\frac{r_{it}}{\frac{1}{3}\frac{s_{Mit}}{\beta_1\alpha_M} + \frac{2}{3}} = \beta_1\alpha_M\frac{\chi_{it}}{s_{Mit}} + \beta_1\gamma k_{it} + \beta_1(a_{it} + \lambda_{it}),$$

which implies:

$$r_{it} = \left(\frac{1}{3}\frac{1}{\mu_{it}} + \frac{2}{3}\beta_1\right)(q_{it} + \lambda_{it}.) \tag{A-9}$$

Equation (A-9) is the equivalent of equation (9). Building on the same logic utilized for equations (12) and (13) one finally gets:

$$\begin{aligned}\widetilde{LHS}_{it} &= \beta_1\gamma k_{it} + \phi_a\widetilde{LHS}_{it-1} - \phi_a\gamma\beta_1 k_{it-1} \\ &+ \frac{(\phi_\lambda - \phi_a)}{\beta_1}\left(\frac{r_{it-1}}{\frac{1}{3}\frac{s_{Mit-1}}{\beta_1\alpha_M} + \frac{2}{3}}\right) - (\phi_\lambda - \phi_a)\beta_1 q_{it-1} + \beta_1(\nu_{ait} + \nu_{\lambda it}).\end{aligned} \tag{A-10}$$

where $\widetilde{LHS}_{it} = \frac{r_{it}}{\frac{1}{3}\frac{s_{Mit}}{\beta_1\alpha_M}+\frac{2}{3}} - \beta_1\alpha_M\frac{\chi_{it}}{s_{Mit}}$.

Estimation of the various parameters in equation (A-10) can be carried by non-linear GMM, as in De Loecker and Warzynski (2012) for example, by considering that the error term $u_{it} = \beta_1(\nu_{ait} + \nu_{\lambda it})$ is a function of some data as well as of the parameters and by building on the following moment conditions: $E[k_{it}u_{it}] = E[k_{it-1}u_{it}] = E[m_{it-1}u_{it}] = E[l_{it-1}u_{it}] = E[q_{it-1}u_{it}] = E[r_{it-1}u_{it}] = 0$. Parallel to the generalized CES case one can avoid exploiting parameters' constraints and extract some reduced-form parameters including $\beta_1\alpha_M$ and $\beta_1\gamma$ as well as ϕ_a. In the very same way we recover γ in the generalized CES case via a second step estimation based on the quantity equation, we can, by using estimates of $\beta_1\alpha_M$, $\beta_1\gamma$ and ϕ_a, write the quantity equation as a linear expression involving only one unknown parameter (β_1) and one right-hand side variable. Therefore, we can use a simple IV strategy based on the moment condition $E[k_{it}\nu_{ait}] = 0$ to identify β_1 and so α_M, γ and ultimately productivity, demand and markup shocks.

Notice that

$$\frac{1}{\mu_{it}} = -3\beta_2^2\tilde{q}_{it}^2 + \beta_1$$

Hence

$$\frac{\partial\left(-3\beta_2^2\tilde{q}_{it}^2 + \beta_1\right)}{\partial q_{it}} = -6\beta_2^2\tilde{q}_{it}$$

while in the simple log-linear form the markup does not depend on equilibrium quantity.

A.3 Discrete/continuous choice models

Discrete/continuous choice models represent a generalisation of standard discrete choice models including, for example, the Multinomial Logit. They are obtained from a random utility framework in which consumers not only choose one alternative amongst many but also how much to consume of a particular good. In what follows we borrow the terminology from Nocke and Schutz (2016) to which the reader is referred for more details.

Let I be the set of differentiated products. A discrete/continuous choice model is a collection of functions of individual prices $\{H_i(P_i)\}_{i \in I}$, where, for every i in I

- H_i is C^3 from $\mathbb{R}_{++}$ to $\mathbb{R}_{++}$
- $V_i^{'} < 0$, $V_i^{''} \geq 0$

where $V_i \equiv \log(H_i)$ is an indirect sub-utility function in a world in which only product i and the outside good, good 0, are available.

The consumer makes choices as follows. He first observes idiosyncratic random components $\{\varepsilon_i\}_{i \in I}$ that are iid type-1 extreme-value as well as prices $\{P_i\}_{i \in I}$. He then chooses only one product and, if he chooses product i, he consumes $D_i(P_i) = -V_i^{'}(P_i)$ units of that product (Roy's identity) and uses the rest of his income to consume the outside good. In doing so he receives indirect utility $Y + V_i(P_i) + \varepsilon_i$. The consumer chooses the product i that maximizes indirect utility, i.e., $i \in \arg\max_{j \in I}\{Y + V_j(P_j) + \varepsilon_j\}$.

By Holman and Marley's theorem the probability of choosing alternative i is $\mathbb{P}_i(P) = \frac{e^{V_i(P_i)}}{\sum_{j \in I} e^{V_j(P_j)}}$ where P is the vector of prices while the expected demand for product i is given by:

$$Q_i(P) = \frac{-H_i^{'}(P_i)}{\sum_{j \in I} H_j(P_j)}. \tag{A-11}$$

The demand system (A-11) satisfies the IIA property and the basic Multinomial Logit is obtained as a special case by setting $H_i(P_i) = e^{-P_i}$. The basic Multinomial Logit can be enriched by introducing a measure of quality ($H_i(P_i) = e^{\Lambda_i - P_i}$) but it would not in general satisfy (4). This is due to the fact that the consumer only chooses one unit of a given product and so our concept of quality does not fit within this framework. (A-11) allows to go beyond the Multionomail Logit and, by relaxing the assumption of unit consumption, (A-11) can be used to characterize preferences satisfying (4). For example, an equivalent to generalized CES preferences can be obtained from (A-11) by setting $H_i(P_i) = \Lambda_i^{\eta_i - 1} P_i^{1-\eta_i}$. In this case we have:

$$Q_i(P) = \frac{(\eta_i - 1)\Lambda_i^{\eta_i - 1} P_i^{-\eta_i}}{\sum_{j \in I} \Lambda_j^{\eta_j - 1} P_j^{1-\eta_j}}.$$

In the limit case of monopolistic competition (the set I is large enough) the denominator is fixed for an individual firm and so it can be readily verified (4) holds. More broadly (A-11) can be used within the monopolistic competition market structure to generate demand systems compatible with (4) by an appropriate choice of $H_i(P_i)$. As far as other forms of imperfect competition are concerned Nocke and Schutz (2016) build on (A-11) to develop an oligopoly Bertrand competition model featuring product heterogeneity in quality Λ_i that can be casted within our framework. The game they consider is aggregative and they are able to establish conditions for both existence and uniqueness of a price equilibrium. We provide a more explicit example of an imperfectly competitive framework featuring strategic interactions that is compatible with (4) in the next Section.

B Other forms of imperfect competition

Our framework can be extended beyond the scope of monopolistic competition. In what follows we show how to frame it in terms of the model developed in Atkeson and Burstein (2008) and further refined by Hottman et al. (2016) in their analysis of multi-product firms.

Atkeson and Burstein (2008) provide two versions of their model. One is based on quantity competition while the other is based on price competition. In both cases firms are large enough to perceive their impact on overall price indices. More specifically a finite number of single-product firms operates within each industry j where preferences are characterized by a CES demand with parameter σ_j. Final consumption is produced by a competitive firm using the output of a continuum of industries y_j for $j \in [0, 1]$ as inputs subject to a CES production function with parameter σ and $1 < \sigma < \sigma_j$. Contrary to the monopolistic competition case a firm i operating in industry j does recognize here that sectoral prices and quantities vary when that firm changes its quantity or price. For example, in the quantity competition case Atkeson and Burstein (2008) show that firms charge a markup $\mu_{ij} = \eta_{ij}/(\eta_{ij} - 1)$ where η_{ij} is the perceived elasticity of demand given by:

$$\eta_{ij}(s_{ij}) = \left[\frac{1}{\sigma_j}(1 - s_{ij}) + \frac{1}{\sigma}s_{ij}\right]^{-1} \tag{B-1}$$

and $s_{ij} = P_{ij}Q_{ij}/\sum_i P_{ij}Q_{ij}$ is the market share of firm i in its sector j. This case falls under the scenarios considered in Hall (1986) and so equation (7) applies. It is readily verified that equation (4) also holds. The proof is straightforward and simply requires changing notation using $\tilde{Q}_{ij} = Q_{ij}\Lambda_{ij}$ instead of Q_{ij} and $\tilde{P}_{ij} = P_{ij}/\Lambda_{ij}$ instead of P_{ij} everywhere in the model. Note that $\tilde{P}_{ij}\tilde{Q}_{ij} = P_{ij}Q_{ij}$ and so $\tilde{s}_{ij} = s_{ij}$. Indeed everything works as if firm i was selling quantity $\tilde{Q}_{ij}$ while charging a price $\tilde{P}_{ij}$. With regards to profit maximization conditions it is straightforward to show that, for given Λ_{ij}, the elasticity of $\tilde{P}_{ij}$ with respect to $\tilde{Q}_{ij}$, given by $-1/\eta_{ij}(s_{ij})$, is the same as the elasticity of P_{ij} with respect to Q_{ij} as well as the the elasticity of $\tilde{P}_{ij}$ with respect to Λ_{ij}. Yet the elasticity of P_{ij} with respect to Λ_{ij} is different. Given $p_{ij} = \tilde{p}_{ij} + \lambda_{ij}$ we have:

$$\frac{\partial p_{ij}}{\partial \lambda_{ij}} = \frac{\partial \tilde{p}_{ij}}{\partial \lambda_{ij}} + \frac{\partial \lambda_{ij}}{\partial \lambda_{ij}} = \frac{\partial p_{ij}}{\partial q_{ij}} + 1,$$

i.e., equation (4) holds. The proof for the price competition case in Atkeson and Burstein (2008) follows the same logic.

C More general production functions

Here we show how we can introduce more flexible production functions. In particular we look at a (homogenous) translog form, i.e., our production function takes the form:

$$q_{it} = a_{it} + \sum_{X \in \{M,L,K\}} \left[\alpha_X \ln X_{it} + \frac{1}{2}\alpha_{XX} \ln (X_{it})^2\right] + \alpha_{MK} \ln M_{it} \ln K_{it} + \alpha_{ML} \ln M_{it} \ln L_{it} + \alpha_{LK} \ln L_{it} \ln K_{it}.$$

Note that,

$$\frac{\partial q_{it}}{\partial m_{it}} = \alpha_M + \alpha_{MM}m_{it} + \alpha_{MK}k_{it} + \alpha_{ML}l_{it}$$

$$\frac{\partial q_{it}}{\partial l_{it}} = \alpha_L + \alpha_{LL}l_{it} + \alpha_{LK}k_{it} + \alpha_{ML}m_{it}$$

$$\gamma - \frac{\partial q_{it}}{\partial m_{it}} - \frac{\partial q_{it}}{\partial l_{it}} = \alpha_K + \alpha_{KK}k_{it} + \alpha_{MK}m_{it} + \alpha_{LK}l_{it}$$

where the last equation follows from the homogeneity assumption (as before γ represents the returns to scale).

We also have that,

$$\frac{\partial q_{it}}{\partial m_{it}} m_{it} + \frac{\partial q_{it}}{\partial l_{it}} l_{it} + \left(\gamma - \frac{\partial q_{it}}{\partial m_{it}} - \frac{\partial q_{it}}{\partial l_{it}}\right) k_{it}$$

$$= \alpha_M m_{it} + \alpha_L l_{it} + \alpha_K k_{it} + \alpha_{MM} m_{it}^2 + \alpha_{LL} l_{it}^2 + \alpha_{KK} k_{it}^2 + 2\alpha_{MK} k_{it} m_{it} + 2\alpha_{ML} m_{it} l_{it} + 2\alpha_{LK} l_{it} k_{it}$$

$$= q_{it} - a_{it} + \frac{1}{2}\alpha_{MM} m_{it}^2 + \frac{1}{2}\alpha_{LL} l_{it}^2 + \frac{1}{2}\alpha_{KK} k_{it}^2 + \alpha_{MK} k_{it} m_{it} + \alpha_{ML} m_{it} l_{it} + \alpha_{LK} l_{it} k_{it}$$

and

$$\begin{aligned} q_{it} &= \frac{\partial q_{it}}{\partial m_{it}} m_{it} + \frac{\partial q_{it}}{\partial l_{it}} l_{it} + \left(\gamma - \frac{\partial q_{it}}{\partial m_{it}} - \frac{\partial q_{it}}{\partial l_{it}}\right) k_{it} \\ &\quad -\frac{1}{2}\alpha_{MM} m_{it}^2 - \frac{1}{2}\alpha_{LL} l_{it}^2 - \frac{1}{2}\alpha_{KK} k_{it}^2 - \alpha_{MK} k_{it} m_{it} - \alpha_{ML} m_{it} l_{it} - \alpha_{LK} l_{it} k_{it} + a_{it}. \end{aligned}$$

From the first order conditions,

$$s_{Mit}\mu_{it} = \frac{\partial q_{it}}{\partial m_{it}},\ s_{Lit}\mu_{it} = \frac{\partial q_{it}}{\partial l_{it}}$$

holds, so that

$$\begin{aligned} q_{it} &= s_{Mit}\mu_{it} m_{it} + s_{Lit}\mu_{it} l_{it} + \left(\gamma - s_{Mit}\mu_{it} - s_{Lit}\mu_{it}\right) k_{it} \\ &\quad -\frac{1}{2}\alpha_{MM} m_{it}^2 - \frac{1}{2}\alpha_{LL} l_{it}^2 - \frac{1}{2}\alpha_{KK} k_{it}^2 - \alpha_{MK} k_{it} m_{it} - \alpha_{ML} m_{it} l_{it} - \alpha_{LK} l_{it} k_{it} + a_{it}. \end{aligned}$$

In this setting equation (9) holds and so by adding λ_{it} on both sides and dividing by $1/\mu_{it}$ one gets:

$$\begin{aligned} LHS_{it}\frac{\partial q_{it}}{\partial m_{it}} &= (q_{it} + \lambda_{it}) - \mu_{it}\left[s_{Lit}(l_{it} - k_{it}) + s_{Mit}(m_{it} - k_{it})\right] \qquad \text{(C-1)} \\ &= \gamma k_{it} + \alpha_{MM} m_{it}^2 + \alpha_{LL} l_{it}^2 + \alpha_{KK} k_{it}^2 + \alpha_{MK} k_{it} m_{it} + \alpha_{ML} m_{it} l_{it} + \alpha_{LK} l_{it} k_{it} + (a_{it} + \lambda_{it}), \end{aligned}$$

where LHS_{it} is the same as in equation (10), i.e., a function of observables and $\partial q_{it}/\partial m_{it}$=$\alpha_M + \alpha_{MM} m_{it} + \alpha_{MK} k_{it} + \alpha_{ML} l_{it}$, i.e., a function of some parameters as well as m_{it}, l_{it} and k_{it}.

By substituting equation (11) to equation (13) into equation (C-1) (while replacing the old α_M with $\partial q_{it}/\partial m_{it}$) and dividing both sides by γ we get:

$$
\begin{aligned}
LHS_{it}\frac{\partial q_{it}}{\partial m_{it}}\frac{1}{\gamma} &= e_{it}-\phi_a e_{it-1}+\partial q_{it}/\partial m_{it}\frac{1}{\gamma}\phi_a LHS_{it-1} \\
&+ \partial q_{it}/\partial m_{it}\frac{1}{\gamma}\left(\phi_\lambda-\phi_a\right)\left(\frac{r_{it-1}}{s_{Mit-1}}-\frac{1}{\partial q_{it}/\partial m_{it}}q_{it-1}\right)+u_{it}
\end{aligned}
\tag{C-2}
$$

where $e_{it} = k_{it} + \frac{\alpha_{MM}}{\gamma}m_{it}^2 + \frac{\alpha_{LL}}{\gamma}l_{it}^2 + \frac{\alpha_{KK}}{\gamma}k_{it}^2 + \frac{\alpha_{MK}}{\gamma}k_{it}m_{it} + \frac{\alpha_{ML}}{\gamma}m_{it}l_{it} + \frac{\alpha_{LK}}{\gamma}l_{it}k_{it}$ and $u_{it}=\frac{1}{\gamma}\left(\nu_{ait}+\nu_{\lambda it}\right)$.

Estimation of the various parameters in equation (C-2) can be carried by non-linear GMM, as in De Loecker and Warzynski (2012) for example, by considering that u_{it} is a function of some data as well as of the parameters, and builds on moment conditions such as $E\left[k_{it}u_{it}\right] = E\left[k_{it-1}u_{it}\right] = E\left[m_{it-1}u_{it}\right] = E\left[l_{it-1}u_{it}\right] = 0$ as well as $E\left[m_{it-1}^2u_{it}\right] = E\left[m_{it-1}k_{it-1}u_{it}\right] = E\left[m_{it-1}l_{it-2}u_{it}\right] = E\left[k_{it-2}l_{it-2}u_{it}\right] = 0$, and so on and so forth. Considering moments up to $t-2$ $(t-1)$ there are 30 (13) such moments conditions that can be exploited. As in the Cobb-Douglas case, it is perhaps best not to exploit parameters' constraints (this mean for example estimating $\frac{\alpha_{MM}}{\gamma}$ rather than trying to separately identify α_{MM} and γ from the revenue equation) and extract some reduced form parameters to be used in a second stage regression based on the quantity equation.

For the quantity equation we have:

$$
q_{it} = \frac{\partial q_{it}}{\partial m_{it}}\frac{\chi_{it}}{s_{Mit}}+\gamma k_{it}-\frac{1}{2}\alpha_{MM}m_{it}^2-\frac{1}{2}\alpha_{LL}l_{it}^2-\frac{1}{2}\alpha_{KK}k_{it}^2-\alpha_{MK}k_{it}m_{it}-\alpha_{ML}m_{it}l_{it}-\alpha_{LK}l_{it}k_{it}+a_{it},
\tag{C-3}
$$

where $\chi_{it} = s_{Lit}\left(l_{it}-k_{it}\right)+s_{Mit}\left(m_{it}-k_{it}\right)$. All of the parameters in equation (C-4) have been identified up to the scaling γ in the previous stage and we can write it as:

$$
q_{it} = \gamma z_{it}+a_{it},
\tag{C-4}
$$

where:

$$
z_{it} = \frac{\partial q_{it}}{\partial m_{it}}\frac{1}{\gamma}\frac{\chi_{it}}{s_{Mit}}+k_{it}-\frac{1}{2}\frac{\alpha_{MM}}{\gamma}m_{it}^2-\frac{1}{2}\frac{\alpha_{LL}}{\gamma}l_{it}^2-\frac{1}{2}\frac{\alpha_{KK}}{\gamma}k_{it}^2-\frac{\alpha_{MK}}{\gamma}k_{it}m_{it}-\frac{\alpha_{ML}}{\gamma}m_{it}l_{it}-\frac{\alpha_{LK}}{\gamma}l_{it}k_{it}+a_{it}.
$$

As in the Cobb-Douglas case we can further substitute for a_{it} using equation (13) (while replacing the old α_M with $\partial q_{it}/\partial m_{it}$) and use the moment condition $E\left[k_{it}\nu_{ait}\right]=0$ on a simple linear model with a single regressor to identify γ and ultimately productivity, demand and markup shocks.

D More general processes for a and λ

Our model can be easily extended to non-linear Markov processes for a and λ as well as to the presence of time-invariant unobserved heterogeneity and measurement error in capital. Consider, for example, the first case and in particular:

$$\begin{aligned} a_{it} &= \phi_{1a} a_{it-1} + \phi_{2a} a_{it-1}^2 + \nu_{ait} \\ \lambda_{it} &= \phi_{1\lambda} \lambda_{it-1} + \phi_{2\lambda} \lambda_{it-1}^2 + \nu_{\lambda it}. \end{aligned} \tag{D-1}$$

By substituting equations (12), (13) and (D-1) into (10) we obtain:

$$\begin{aligned} LHS_{it} &= \frac{\gamma}{\alpha_M} k_{it} + \phi_{1a} LHS_{it-1} - \phi_{1a} \frac{\gamma}{\alpha_M} k_{it-1} \\ &+ (\phi_{1\lambda} - \phi_{1a}) \left(\frac{r_{it-1}}{s_{Mit-1}} - \frac{1}{\alpha_M} q_{it-1} \right) \\ &+ (\phi_{2\lambda} - \phi_{2a}) \left(\alpha_M (\frac{r_{it-1}}{s_{Mit-1}})^2 + \frac{1}{\alpha_M} (q_{it-1})^2 - 2(\frac{r_{it-1}}{s_{Mit-1}} q_{it-1}) \right) \\ &+ \phi_{2a} \left(\alpha_M (LHS_{it-1})^2 + \frac{\gamma^2}{\alpha_M} (k_{it-1})^2 - 2\gamma LHS_{it-1} k_{it-1} - 2\alpha_M (LHS_{it-1} \frac{r_{it-1}}{s_{Mit-1}}) \right) \\ &+ \phi_{2a} \left(+2\gamma (k_{it-1} \frac{r_{it-1}}{s_{Mit-1}}) + 2(LHS_{it-1} q_{it-1}) - 2\frac{\gamma}{\alpha_M} (k_{it-1} q_{it-1}) \right) + \frac{1}{\alpha_M} (\nu_{ait} + \nu_{\lambda it}). \end{aligned} \tag{D-2}$$

Equation (D-2) can be used to estimate $\frac{\gamma}{\alpha_M} \equiv \beta$, ϕ_{1a} and ϕ_{2a}[30] by a suitable linear regression with a change in variables and some reduced-form parameters. In turn these estimates could be employed in the corresponding expression of (18):

[30] As in the baseline case, β and ϕ_{1a} can be directly obtained from, respectively, the coefficients of k_{it} and LHS_{it-1} with no need to exploit reduced-form parameters constraints. As for ϕ_{2a}, this can be obtained as 1/2 times the coefficient of the interaction between LHS_{it-1} and q_{it-1}.

$$
\begin{aligned}
q_{it} &= \frac{\gamma}{\hat{\beta}}\frac{s_{Lit}}{s_{Mit}}\left(l_{it}-k_{it}\right)+\frac{\gamma}{\hat{\beta}}\left(m_{it}-k_{it}\right)+\gamma k_{it} \\
&+ \hat{\phi}_{1a}\frac{\gamma}{\hat{\beta}}LHS_{it-1}-\hat{\phi}_{1a}\gamma k_{it-1}-\hat{\phi}_{1a}\left(r_{it-1}\frac{\gamma}{\hat{\beta}s_{Mit-1}}-q_{it-1}\right) \\
&+ \hat{\phi}_{2a}(\frac{\gamma}{\hat{\beta}})^2(LHS_{it-1})^2+\hat{\phi}_{2a}\gamma^2(k_{it-1})^2+\hat{\phi}_{2a}(\frac{\gamma}{\hat{\beta}})^2(\frac{r_{it-1}}{s_{Mit-1}})^2+\hat{\phi}_{2a}(q_{it-1})^2 \\
&- 2\hat{\phi}_{2a}\frac{\gamma^2}{\hat{\beta}}LHS_{it-1}k_{it-1}-2\hat{\phi}_{2a}(\frac{\gamma}{\hat{\beta}})^2LHS_{it-1}\frac{r_{it-1}}{s_{Mit-1}}+2\hat{\phi}_{2a}\frac{\gamma}{\hat{\beta}}LHS_{it-1}q_{it-1} \\
&+ 2\hat{\phi}_{2a}\frac{\gamma^2}{\hat{\beta}}k_{it-1}\frac{r_{it-1}}{s_{Mit-1}}-2\hat{\phi}_{2a}\gamma k_{it-1}q_{it-1}+\nu_{ait}.
\end{aligned}
\tag{D-3}
$$

from which the γ parameter can be obtained by a suitable linear regression where the dependent variable is $q_{it}-\hat{\phi}_{1a}q_{it-1}-\hat{\phi}_{2a}(q_{it-1})^2$ and the right-hand side variables are grouped into two sets: one in which the only unknown coefficient is γ and the other where the only unknown coefficient is γ^2. As in the baseline case, instrumenting is needed.

The case of time-invariant unobserved heterogeneity is easier to handle. In this scenario we have:

$$
\begin{aligned}
a_{it} &= \phi_a a_{it-1}+u_{ai}+\nu_{ait} \\
\lambda_{it} &= \phi_\lambda \lambda_{it-1}+u_{\lambda i}+\nu_{\lambda it}.
\end{aligned}
\tag{D-4}
$$

By substituting equations (12), (13) and (D-4) into equation (10) we obtain:

$$
\begin{aligned}
LHS_{it} &= \frac{\gamma}{\alpha_M}k_{it}+\phi_a LHS_{it-1}-\phi_a\frac{\gamma}{\alpha_M}k_{it-1} \\
&+ (\phi_\lambda-\phi_a)\left(\frac{r_{it-1}}{s_{Mit-1}}-\frac{1}{\alpha_M}q_{it-1}\right)+\frac{1}{\alpha_M}\left(u_{ai}+u_{\lambda i}\right)+\frac{1}{\alpha_M}\left(\nu_{ait}+\nu_{\lambda it}\right).
\end{aligned}
\tag{D-5}
$$

Equation (D-5) can be transformed into a linear regression model similar to equation (15) with the only difference being that, the simultaneous presence of an unobservable time-invariant component correlated with regressors ($\frac{1}{\alpha_M}\left(u_{ai}+u_{\lambda i}\right)$) and the lag of the dependent variable (LHS_{it-1}), calls for the use of, for example, a dynamic panel data estimator rather than simple OLS. Similar arguments apply to quantity equation (19).

Last but not least the presence of standard measurement error in, for example, capital is relatively straightforward to accommodate in our framework. Such measurement error would imply that the error term in equation (14) is correlated with k_{it} and k_{it-1} and so simple OLS cannot be used any more. Yet, very much like in Wooldridge (2009), the simple solution to this problem is to use appropriate instruments for k_{it} and k_{it-1} like, for example, suitable lags of capital and inputs. At the same time, the moment condition $E\left\{\nu_{ait}k_{it}\right\}=0$ used to identify γ in equation (18) would be violated by the presence of measurement error but can, for example, be replaced by the following alternative:

$$E\left\{\nu_{ait}k_{it-2}\right\}=0.$$

E Multi-product firms

There are several issues related to multi-product firms. We focus on the issue of the assignment of inputs to outputs. Produced quantities and generated revenues may be observable for individual products in some databases. However, in many instances information on inputs used for a specific product is not available for multi-product firms. We propose an extension of our baseline model and procedure to solve the problem of assigning inputs to outputs for multi-product firms. In doing so we assume, as in De Loecker et al. (2016), there is a limited role for economies (or diseconomies) of scope on the cost side. However, contrary to De Loecker et al. (2016), we do not impose multi-product firms to be characterized by a common productivity across the different products they produce. We also allow for firm-product-time specific markups but impose demand shocks to be common across products within a firm. This corresponds to a setting where firms can be distinguished into those consistently selling high quality products and those consistently selling low quality products. Yet firms are allowed to be more or less efficient in the production of a specific product and charge different markups. The framework we propose is consistent with a monopolistically competitive market structure where firms ignore cannibalisation effects. The model could be potentially enriched to cope with such cannibalisation effects by building, for example, on the demand and market structure developed in Hottman et al. (2016).

As usual we denote a firm by i and time by t. A firm i produces in t one or more products indexed by p. The number of products is denoted by I_{it}. We assume demand shocks are firm-time specific (λ_{it}) while we allow markups (μ_{ipt}) and productivity shocks (a_{ipt}) to be firm-product-time specific. The production function for product p produced by firm i is given by:

$$Q_{ipt}=C_{pt}A_{ipt}L_{ipt}^{\alpha_{Lp}}M_{ipt}^{\alpha_{Mp}}K_{ipt}^{\gamma_p-\alpha_{Mp}-\alpha_{Lp}}, \tag{E-1}$$

where C_{pt} is an innocuous product-time constat we disregard in what follows. This means we allow for technology (α_{Lp}, α_{Mp}, γ_p) to differ across the different products p produced by a multi-product firm. At the same time productivity is allowed to vary across products within a firm and information coming from single-product firms can be used to infer the technology of multi-product firms, i.e., we rule out physical synergies in production but allow for some of the economies (diseconomies) of scope discussed in De Loecker et al. (2016). Furthermore, we assume firms maximize profits for each product independently which is consistent with monopolistic competition. At each point in time a firm chooses (for each product p) the amount of labour L_{ipt} and materials M_{ipt} in order to minimize short-term costs and takes capital K_{ipt}, as well as productivity a_{ipt} and demand λ_{it} shocks as given. We make us of Generalised CES preferences meaning that demand for a given product p is given by:

$$Q_{ipt} = P_{ipt}^{-\eta_{ipt}} \Lambda_{it}^{\eta_{ipt}-1} \kappa_{pt}^{-\eta_{ipt}}$$

while profit maximization requires:

$$P_{ipt} = \mu_{ipt} \frac{\partial C_{ipt}}{\partial Q_{ipt}}, \tag{E-2}$$

where marginal cost is equal to[31]

$$\frac{\partial C_{ipt}}{\partial Q_{ipt}} = A_{ipt}^{-\frac{1}{\alpha_{Lp}+\alpha_{Mp}}} Q_{ipt}^{\frac{1-\alpha_{Lp}-\alpha_{Mp}}{\alpha_{Lp}+\alpha_{Mp}}} K_{ipt}^{\frac{\gamma_p-\alpha_{Lp}-\alpha_{Mp}}{\alpha_{Lp}+\alpha_{Mp}}} \tag{E-3}$$

and the markup is simply a function of the elasticity of demand: $\mu_{ipt} = \frac{\eta_{ipt}}{\eta_{ipt}-1}$. Markups can be obtained from:

$$\mu_{ipt} = \frac{\alpha_{Mp}}{s_{Mipt}} \tag{E-4}$$

where s_{Mipt} is the expenditure share of materials for product p at time t in firm revenue for product p at time t. At the same time we can write log revenue for product p at time t, up to an innocuous constant, as:

$$r_{ipt} = q_{ipt} + p_{ipt} = \frac{1}{\mu_{ipt}} \left(q_{ipt} + \lambda_{it}\right). \tag{E-5}$$

Finally, we assume that both a_{ipt} and λ_{it} evolve over time as linear stochastic Markov processes:

$$\begin{aligned} a_{ipt} &= \phi_{ap}\ a_{ipt-1} + \nu_{aipt} \\ \lambda_{it} &= \phi_{\lambda}\lambda_{it-1} + \nu_{\lambda it} \end{aligned}$$

where ν_{aipt} and $\nu_{\lambda it}$ can be correlated with each other.

[31] We omit the innocuous product-time constant $\left(\frac{W_{Lpt}}{\alpha_{Lp}}\right)^{\frac{\alpha_{Lp}}{\alpha_{Lp}+\alpha_{Mp}}} \left(\frac{W_{Mpt}}{\alpha_{Mp}}\right)^{\frac{\alpha_{Mp}}{\alpha_{Lp}+\alpha_{Mp}}}$

As far as single-product firms are concerned the assumptions above are such that the parameters of the production function (α_{Mp} and γ_p), as well as single-product firms productivity, demand and markups, can be obtained using the baseline MULAMA procedure. Estimations need to be carried on single-product firms separately for each product p. Turning to multi-product firms we impose, as in De Loecker et al. (2016), that the same technology parameters coming from single-product producers extend to each of their products. Yet, in order to quantity multi-product firms productivity, markups and demand, we still need to solve the issue of how to assign inputs to outputs and we do so by building on the above assumptions. As far as materials are concerned we need to assign the observable total firm material expenditure M_{it} across the I_{it} products produced by firm i at time t, i.e., we need to assign values to M_{ipt} such that $\sum_{p=1}^{I_{it}} M_{ipt} = M_{it}$. We can use this condition along with (E-4) and (E-5) to operate this assignment. Substituting (E-4) into (E-5) and adding $\sum_{p=1}^{I_{it}} M_{ipt} = M_{it}$ provides a system of $I_{it} + 1$ equations in $I_{it} + 1$ unknowns; the I_{it} inputs expenditures M_{ipt} plus λ_{it}. Indeed, at this stage we have data on r_{ipt}, q_{ipt}, α_{Mp} and M_{it}. By recovering inputs expenditures M_{ipt} we can subsequently compute materials expenditure shares in revenues s_{Mipt} and so use (E-4) to recover our firm-product-time specific markups μ_{ipt}. Since labour is a variable input a condition analogous to (E-4) holds for this input and so we can use the computed markups μ_{ipt} and information on α_{Lp} to derive labour expenditure shares in revenues s_{Lipt} and ultimately recover L_{ipt}.[32]

The above procedure allows so far to obtain markups and demand shocks, as well as information on labour and materials use, for each of the products of a multi-product firm. However, in order to recover productivity a_{ipt} we still need values for capital K_{ipt}. To do this one can proceed as follows. By substituting (E-3) into (E-2), while using observed prices and the computed values of μ_{ipt}, L_{ipt}, M_{ipt} and production function coefficients, one gets a system of I_{it} equations. This can be further complemented by (E-1) as well as $\sum_{p=1}^{I_{it}} K_{ipt} = K_{it}$ and observed quantities to obtain a system of $2 \times I_{it} + 1$ equations in $2 \times I_{it}$ unknowns: the a_{ipt} and K_{ipt}. The system is over-identified and can thus be solved by imposing $\sum_{p=1}^{I_{it}} K_{ipt} = K_{it}$ holds and minimizing the squared differences between actual prices and quantities and those delivered by the system.

[32] As a matter of fact in the MULAMA model we do not need to impose α_{Lp} to be the same across firms. From every single product firm equation (E-4) applied to labour delivers, using the computed markups and the observed labour expenditure share in revenue, a different α_{Lp}. One can compute the mean value of these coefficients across firms producing product p. Doing this for the I_{it} products of a given multi-product firm one can then re-scale all of the labour coefficient in order to get values for L_{ipt} that add up to the observed total labour expenditure of that firm: $\sum_{p=1}^{I_{it}} L_{ipt} = L_{it}$.

F Aggregation

We now provide an example in which it makes sense to aggregate quantities produced of different products within a firm while using average log prices (across firms within a product) as weights. In terms of our data a product has to be thought of as an 8-digit Prodcom code produced by a firm belonging to a given industry, i.e., at the 3 digit-unit of measurement level.

Suppose that firm i produces many products indexed by p and that the log production function of product p by firm i can be simplified as q_{ip}=$q_i + s_p^q$ where s_p^q is an Hicks-neutral shifter specific to the product and constant across firms. Further assume that the log demand shock corresponding to product p produced by firm i is λ_{ip}=$\lambda_i + s_p^\lambda$ where s_p^λ is specific to the product and constant across firms. Now impose markups $\mu_{ip} = \mu_i$. We thus get:

$$r_{ip} = p_{ip} + q_{ip} = \frac{1}{\mu_i}\left(q_{ip} + \lambda_{ip}\right) = \frac{1}{\mu_i}\left(s_p + q_i + \lambda_i\right) = \frac{1}{\mu_i}\left(\tilde{q}_{ip} + \lambda_i\right),$$

where s_p=$s_p^q + s_p^\lambda$ and $\tilde{q}_{ip} = s_p + q_i$. This equation shows that, within our assumptions, everything works as if the firm was producing identical products, i.e., having the same productivity, demand and markup shocks as well as technology constraint, in different quantities $\tilde{q}_{ip}$. The problem is that q_{ip} is directly observable in our data while $\tilde{q}_{ip}$ is not. Yet, from the above equation we get:

$$p_{ip} = r_{ip} - q_{ip} = \frac{1}{\mu_i}\left(\tilde{q}_{ip} + \lambda_i\right) - q_i - s_p^q = \frac{1}{\mu_i}\left(q_i + \lambda_i\right) - q_i + \frac{1}{\mu_i}s_p - s_p^q.$$

We finally posit $\mathbb{E}\left[p_{ip}|i \in I^p\right] = a + bs_p - s_p^q$ where I^p is the set of firms i producing product p, a=$\mathbb{E}\left[\frac{1}{\mu_i}\left(q_i + \lambda_i\right) - q_i | i \in I^p\right]$ and b=$\mathbb{E}\left[\frac{1}{\mu_i} | i \in I^p\right]$. This property would be automatically satisfied if all firms were producing all products within an industry. On a broader basis, this amounts assuming that the distributions of productivity, markups and demand shocks corresponding to firms selling a given product are similar across the 8-digit products belonging to a given industry. We thus allow for the distributions of firm-level productivity, markups and demand shocks to be different across industries.

The above assumption implies that, by summing the average (across firms within a product) log price p_{ip} observed in the data to the output of product p by firm i (q_{ip}) we get a monotonous transformation of $\tilde{q}_{ip} = s_p + q_i$:

$$\mathbb{E}\left[p_{ip}|i \in I^p\right] + q_{ip} = a + bs_p + q_i$$

that we can use to aggregate outputs of the different products of firm i.