Gains from parameters: Flexibility in the productivity distribution and international trade

Ruben Dewitte*, Michel Dumont†, Glenn Rayp‡, Peter Willemé§

Abstract

Since the early 2000s, the Pareto distribution has been dominating the prevailing class of heterogeneous firms models and its resulting social welfare and policy analysis. However, the assumption of Pareto-distributed heterogeneity is perceived statistically misspecified and theoretically oversimplifying. This paper introduces Finite Mixture Models (FMM) as a superior alternative to the Pareto distribution. FMMs build on the existence of discrete subpopulations in the data. As such, they allow for a very flexible fit to the heterogeneity distribution and can be supported by a general specification of firm dynamics. We build a statistical framework that demonstrates their excellent empirical performance on the population of Portuguese firms. Moreover, FMMs can easily be implemented into the class of heterogeneous firms models à la Melitz (2003). A policy exercise calculates the Gains From Trade between rivaling distributions and demonstrates the importance of shifting away from Pareto.

*Ghent University and SHERPPA

†Federal Planning Bureau

‡Ghent University and SHERPPA

§Federal Planning Bureau

1 Introduction

Since the early 2000s, the Pareto distribution has been dominating the prevailing class of heterogeneous firms models (Melitz, 2003). Even though the Melitz (2003)-model is not restricted to this distributional choice, its convenience lead to a widespread reliance on the Pareto distribution for social welfare and economic policy analysis.

However, the assumption of Pareto-distributed heterogeneity is perceived statistically misspecified and theoretically oversimplifying. The empirical literature on firm size distributions provides idiosyncratic evidence in favour of alternatives such as the Lognormal (Bas et al., 2015; Fernandes et al., 2015; Head et al., 2014), the Weibull (Bee and Schiavo, 2017), the Double Pareto (Arkolakis, 2016), the Lognormal-pareto (Nigai, 2017), ... respective to Pareto. Also, the firm dynamics that underpin the stationary Pareto(-tailed) distribution rely on rigidly structured mechanisms (Córdoba, 2008; Gabaix, 2009; Reed, 2001) subject to discussion. Moreover, the Gains From Trade (GFT) literature critiques the Pareto distribution for its far-reaching simplification of reality, masking important channels through which trade can operate (Arkolakis et al., 2012; Bas et al., 2015; Fernandes et al., 2015).[1]

This paper introduces Finite Mixture Models (FMM) as a superior alternative to the Pareto distribution. We provide a theoretical motivation for FMMs and build a statistical framework to showcase its excellent empirical performance on the population of Portuguese firms. A policy application demonstrates the importance of correctly approximating the heterogeneity distribution in the class of heterogeneous firms models à la Melitz (2003). It also underlines the straightforward implementation of FMM into the existing theoretical models.

FMMs allow for a very flexible fit to the heterogeneity distribution and can be supported by a general specification of firm dynamics. A FMM is essentially a weighted sum of an à priori unknown number of individual densities. As such, it is a semi-parametric approximation that allows for discrete subpopulations to define the overall distribution. Moreover, these discrete subpopulations can be defined by independent firm-level stochastic processes of entry, exit and growth. This allows for a more general description of firm dynamics than the rigidly structured mechanisms underlying the Pareto(-tailed) distributions.

Our dataset allows for a representative evaluation of the *complete* distributions of firm size. Research up to date has been confined to data with a truncated or unrepresentative

[1]Under a standard Pareto-Melitz (2003)-model, the elasticity of trade with respect to variable trade costs is constant across trading partners (Bas et al., 2015). Moreover, this elasticity combined with the domestic trade share are sufficient statistics to calculate aggregate GFT (Arkolakis et al., 2012), neutralizing the firm-level heterogeneity dimension.

left tail of the distribution. We have access to Portuguese firm-level sales of the universe of firms in manufacturing and services over the years 2004-2007. This allows us to evaluate the performance of parametric distributions for the complete as well as specifically the *left*- and right- tail of the firm size distribution.

We build a statistical framework that refutes the false perception of Pareto-distributed firm sales in favour of a general class of distributions that mimic Pareto(-tail) behaviour. The graphical representation of Pareto's Cumulative Distribution Function (CDF) features as a straight line on a log-log scale. For long, the apparent matching straight-line behaviour of truncated firm sales was used as evidence in favour of Pareto. However, just as every bend line appears straight if you zoom in close enough, so too does firm sales only appear to be Pareto when truncated sufficiently (Perline, 2005). The combination of our complete dataset and a clear statistical framework allows us to strongly reject the Pareto-assumption in favour of the alternative, better-fitting FMMs.

A policy exercise calculates the Gains From Trade between rivaling distributions and demonstrates the importance of shifting away from Pareto. We rely on a standard heterogeneous firms models à la Melitz (2003) and showcase how FMMs can easily be implemented in this class of models. Specifically, we introduce a multitude of new distributions to the current mix being considered for GFT, revealing the critical importance of the distributional assumption when calculating GFT.

The message this paper carries has consequences for several strands of the micro-founded economic literature. First, our results advance most significantly the discussion on the choice of the heterogeneity distribution in the GFT literature. Recent studies compared the performance of Pareto in terms of fit and welfare gains with distributions such as the bounded (from above) Pareto distribution (Feenstra, 2014; Helpman et al., 2008; Melitz and Redding, 2015), the Lognormal distribution (Bas et al., 2015; Fernandes et al., 2015; Freund and Pierola, 2015; Head et al., 2014; Yang, 2017), the Weibull distribution (Bee and Schiavo, 2017) or a mix of distributions tailored to both the bulk and the tail(s) of the empirical distribution (Nigai, 2017; Sager and Timoshenko, 2017). Most closely related to our work is the work of Fernandes et al. (2015), who relies on firm-level export data to consider a finite mixture of Lognormal distributions using export destinations as mixing parameter. Our work is more general as it takes no specific stance on distribution type nor mixing parameter, while having access to untruncated firm sales data. As such, we are also the first able to investigate the influence of the left tail of the firms sales distribution. Overall, we introduce a multitude of new distributions to the current mix, implement clear statistical distribution tests and provide the first overview of the playing field in GFT relying on the population of firm sales.
Second, it puts into question recent firm-level stochastic processes assumed in the micro-founded growth literature (Arkolakis, 2016). Processes assumed up till now result in a

structured *infinite* mixture of Lognormal distributions with mixing parameter firm age, parametrized as Double Pareto or Double-Pareto Lognormal distributions (Reed, 2001; Reed and Jorgensen, 2004). Our results provide evidence in favour of *finite* mixtures, without taking a stance both on the distribution type and mixing parameter.

Third, it affects the granularity literature. The mapping of large firm dynamics into aggregate fluctuations usually relies on the Pareto-distributed firm sales, at least in the right tail (Carvalho and Grassi, 2015; di Giovanni and Levchenko, 2012; Gabaix, 2011), an assumption refuted by our empirical results.

Finally, our paper is also related to those literatures that investigate the Pareto behaviour for a diverse set of subjects (see Clauset et al. (2009) for an overview). More specifically, our work is closely related to the discussion on the distributions of city sizes being Pareto (Gabaix and Ibragimov, 2011), Lognormal (Eeckhout, 2004, 2009), a combination of Lognormal and Pareto (Giesen et al., 2010; Ioannides and Skouras, 2013; Luckstead and Devadoss, 2017) or FMMs (Kwong and Nadarajah, 2019). We also greatly benefited from the discussion in the actuarial science literature on the fitting abilities of FMMs (Miljkovic and Grün, 2016) versus composite models (Bakar et al., 2015).

The paper is organized as follows. In the ensuing section we examine the graphical and theoretical motivation in favour of both Pareto and FMM. Section 3 contains the methodology, discussing the estimation methods for the considered distributions and showcasing the methods that allows to differentiate between them. Section 4 discusses the data. We provide our empirical results in section 5 and discuss policy implications of these results in section 6. Section 7 concludes.

2 To Pareto or not to Pareto

In this section, we dissect the flawed motivation for the Pareto distribution and introduce FMM as superior alternatives. We rely on graphical tools to link the perception of Pareto behaviour to the use of truncated data, or tail data only. In the same vein, we scrutinize the reasoning behind the unnecessarily strong focus on the right-tail of firm sales. The alternative FMMs allow for a very flexible fit to the complete sales distribution and can be supported by a more general specification of firm dynamics.

We deduce from graphical representations that firm sales only appear to follow Pareto in the tails and can just as well be approximated by a large class of distributions appropriately denoted *'False Pareto'*. Figure 1 displays the empirical survival function (upper tail) of Portuguese productivity on a log-log scale. If our complete data would be Pareto, it would follow a straight line. This does not seem true for the complete data due to the existence of a hump in the middle (see also the Probability Density Function in Appendix Figure A1). Nevertheless, both the left (lower left panel) and right tail (lower right panel) showcase linearity of the CDF and survival function respectively on a log-log scale, in line with (Inverse) Pareto behaviour in the (left) right tail of the distribution. However, just as every bend line looks straight when you zoom in close enough, so too does firms sales only appear to be straight when truncated sufficiently.[2] This apparent straight line behaviour of the tails can just as well be approximated by a surprisingly large class of distributions including, but not restricted to, (finite mixtures of) the Exponential, Lognormal, Gamma and Weibull distribution.[3] We denote this class of distributions that mimic Pareto in the tails as False Pareto.

While the focus on Pareto is often motivated from its good fit to the right tail of firms sales, the remainder of the distribution also contains vital information on firm behaviour. The historical preference for the right tail of firms sales can be rationalized by the relative weight of individual firms, the empirical availability of data on larger firms and its accommodation of the extensive margin of several firm-level activities such as exporting, FDI, innovation ... (Arkolakis et al., 2012; Chaney, 2008). Recent work, however, emphasizes the importance of matching the bulk of the distribution given its overall relative weight (the bulk accounts for up to 72% percent of firm sales in our data) (Fernandes et al., 2015; Head et al., 2014; Nigai, 2017). Lastly, the left tail incorporates the entry and exit of firms. While the entry and exit behaviour already has been investigated (see, f.i. Hopenhayn (1992)), there is to our knowledge no work that analyses the importance of a good fit for the left tail of the firm size distribution. This paper builds a framework which allows to focus on the specific areas of the distribution, but treats both the tails and the bulk of the distribution a priori as equally important.

2.1 Finite Mixture Models

Finite Mixture models allows, contrary to Pareto, for a very flexible fit to the complete (both bulk and tails) firm size distribution. The FMM $g(\cdot)$ is essentially a weighted sum of K individual densities $m_k(\cdot)$:

$$g(\omega|\boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k m_k(\omega|\boldsymbol{\theta}_k), \qquad \pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1$$

where K represents the number of components or discrete subpopulations, π_k is the prior probability of component k, $\boldsymbol{\theta_k}$ the component-specific parameter vector of density $m(\cdot)$ and $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{K-1}, \theta_1, \ldots, \theta_k)$ is the vector of all model parameters. These

[2] Note how the data of Axtell (2001), and more recently Chaney (2018), are not truncated but unintentionally favour the Pareto distribution due to data binning (Virkar and Clauset, 2014).

[3] This class of distributions is defined within the Gumbel domain of attraction (see (Perline, 2005)).

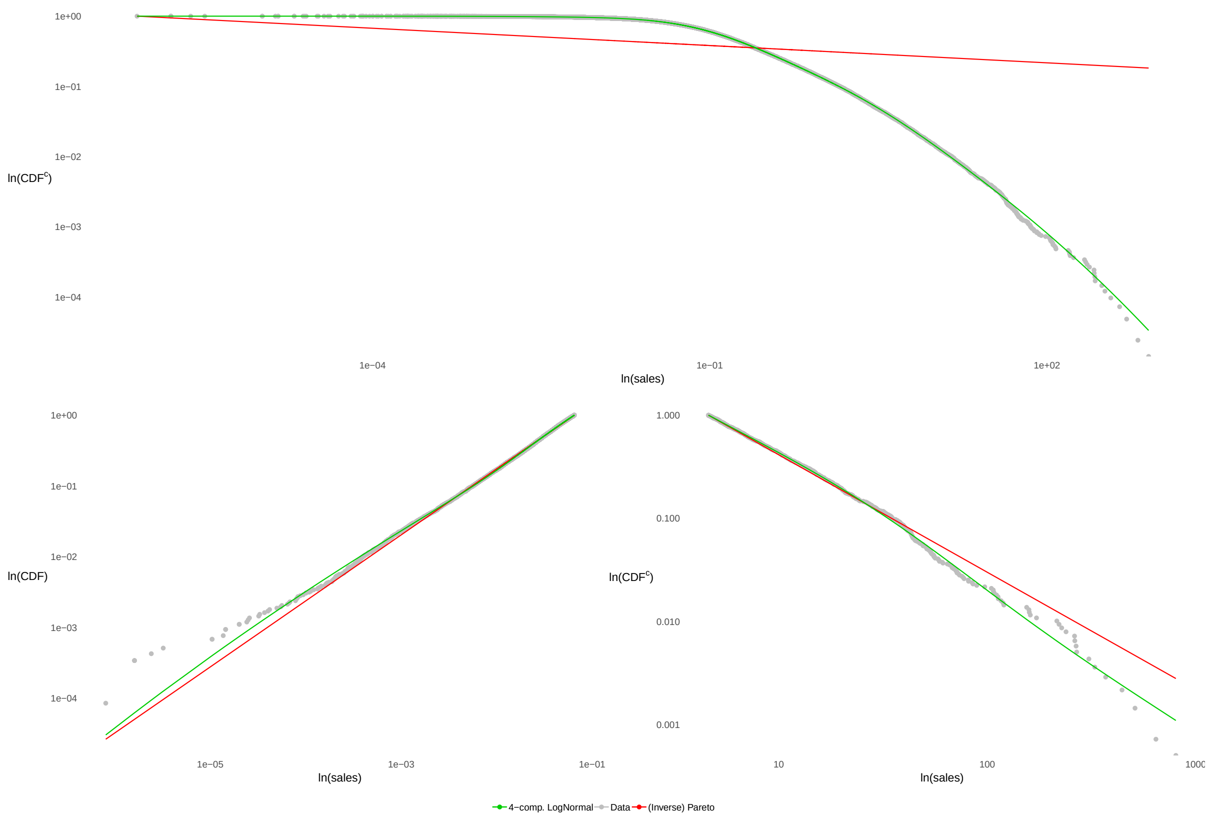

Figure 1: Empirical survival function of Portuguese firm sales in 2006 (upper panel) on a log-log scale with fitted (Inverse) Pareto and Finite Mixture of Lognormal distributions reveal the bad fit of a Pareto distribution. Focusing on the left tail (lower left panel) right tail (lower right panel) provides a seemingly better-fitting Pareto distribution, but the FMM still looks superior. **Notes:** (Truncated) Distributions are fitted using maximum likelihood methods (cf. infra) to the complete and truncated dataset independently. Tail truncation points are determined by the best-fitting (Inverse) Pareto distributions

models are also referred to as latent-class models given that the number of components, and thus also the mixing parameter itself, don't have to be specified à priori but are determined by the data. As such, a finite mixture model resembles a semi-parametric approach.[4] Moreover, they fall within the class of False Pareto distributions (depending on the chosen individual densities) that mimic Pareto tails and can easily be integrated in the class of heterogeneous firms models à la Melitz (2003).[5]

Not only is a finite mixture model empirically appealing, it also generalizes the existing generative processes of Pareto-tailed distributions. We provide an example using the Lognormal distribution. Assume productivity $\omega \in \Omega$ follows a geometric Brownian motion over time t

$$d\omega_t = \mu(\omega_t, t)dt + \sigma(\omega_t, t)d\omega_t, \tag{1}$$

with a systematic component, drift μ and a stochastic component, the Brownian motion dw. It is well-known in the micro-founded growth literature that this process implies:[6]

1. ... a lognormal productivity distribution at fixed time T with fixed initial state Ω_0:

$$\Omega \sim LN\left(\Omega_0 + \left(\mu - \frac{\sigma^2}{2}\right)T, \sigma^2 T\right) \tag{2}$$

2. ... a Pareto distribution if the growth process is reflected at ω_{min}:

$$\Omega \sim Pareto\,(k, \omega_{min}) \tag{3}$$

3. ... a Double Pareto distribution if the growth process is augmented with a Yule process of birth such that $T \sim Exp\,(\lambda)$:

$$\Omega(T) \sim DoublePareto\,(\alpha, \beta, \omega_0)$$

4. ... a Double-Pareto Lognormal distribution if the growth process is augmented with a Yule process of birth such that $T \sim Exp\,(\lambda)$ and a lognormally distributed initial

[4] A semi-parametric approach is to be favoured over a nonparametric approach in the case of heavy-tailed distributions such as firm sales. This because the heavy tails renders nonparametric procedures less efficient (Clauset et al., 2009).

[5] Finite mixtures of distributions in the Gumbel domain of attraction (Perline, 2005) ensure a unique equilibrium in the Melitz (2003)-model. This means that the zero-profit cutoff monotonically decreases with productivity. A sufficient condition is that $\frac{g(\omega)\omega}{1-G(\omega)}$ is increasing to infinity on $(0, \infty)$.

[6] See o.a. Arkolakis (2016); Eeckhout (2004); Gabaix (2009); Reed and Jorgensen (2004).

state $\Omega_0 \sim LN\left(\mu_0, \sigma_0^2\right)$:

$$\Omega(T, \omega_0) \sim DoubleParetoLN\left(\alpha, \beta, \mu_0, \sigma_0\right).$$

Both the Double-Pareto and Double-Pareto lognormal distributions are structured infinite mixtures of (Lognormal) distributions with mixing parameter firm age T. However, firm age is not the only possible mixing parameter we can think of, to the contrary. It has already been argued firm size distributions could be influenced by the fact that firms belong to different industries (Luttmer, 2007), firms innovate or not (Costantini and Melitz, 2008), firm export or not (f.i. learning-by-exporting (De Loecker, 2013)), ... Overall, there are "many sources of heterogeneity that support the idea of discrete subpopulations likely to differ in important characteristics ..." (Perline (2005),p.80).

This is exactly what Finite Mixture Models are intended for. They allow for discrete subpopulations without having to specify those à priori. This is a large advantage compared to the rigid structuring that is implied by infinite mixture models such as the Double-Pareto (Lognormal) distributions. Such models carry the danger of misspecification and/or oversimplification, as they are based on a single, prespecified mixing parameter.

3 Methodology

From the graphical presentation in the previous section, we learned that a clear statistical framework is needed to differentiate between actual Pareto and False Pareto behaviour, both for the bulk and the tail of the data. We first establish our methodology to fit the distributions both to the tails and the complete firm sales. We then determine the statistical tests to differentiate between the fitted distributions.

3.1 Distribution fitting

We rely on augmented maximum likelihood[7] procedures to fit all our distributions to the data: (Inverse) Pareto, FMM and Pareto-tailed Lognormals.[8] In the case of FMM, ML is wrapped in an Expectation-Maximization (EM) algorithm, allowing us to determine both the distribution parameters and component weights at once. We also augment the procedures for both (Inverse) Pareto and FMM in order to be able to fit truncated data. This will allow us to focus on tail performance and to generalize our framework to unrepresentative/truncated data.

3.1.1 (Inverse) Pareto

Complete data The probability density functions of the (Inverse) Pareto distributions are specified as

$$ParetoInv(\omega;\omega_{max},k_1) = \frac{k_1}{\omega_{max}^{k_1}}\omega^{k_1-1}, \qquad Pareto(\omega;\omega_{min},k_2) = \frac{k_2}{\omega_{min}^{-k_2}}\omega^{-k_2-1}.$$

for $\omega \leq \omega_{max}$ and $\omega_{min} \leq \omega$ respectively. The maximim likelihood estimator for the shape parameters $k_{1,2}$ can be easily obtained as

$$k_1 = \left[\frac{1}{F}\sum_{f=1}^{F} ln\frac{\omega_{max}}{\omega_f}\right]^{-1}, \qquad k_2 = \left[\frac{1}{F}\sum_{f=1}^{F} ln\frac{\omega_f}{\omega_{min}}\right]^{-1},$$

while the scale parameters estimates would be $\hat{\omega}_{min} = \min(\omega), \hat{\omega}_{max} = \max(\omega))$, as the likelihood function is monotonically increasing (decreasing) in ω_{min} (ω_{max}).

Truncated data If we want to fit the (Inverse) Pareto distribution to one of the tails, we need to adapt the MLE-procedure accordingly. Due to the fact that the truncation

[7] Popular fitting techniques in the literature rely on the minimization of squared errors between a log-linearization of the theoretical and empirical PDF's/CDF's (Axtell, 2001; Bas et al., 2015; Bee and Schiavo, 2017; di Giovanni and Levchenko, 2013; Freund and Pierola, 2015; Head et al., 2014; Nigai, 2017). Such methods, however, might not be apt to fit distribution functions. For instance, reported parameters in the literature are, to our knowledge, not obtained from a regression procedure restricted to estimate a properly normalized distribution function. Parameters obtained from an estimation procedure must result in a probability density function that integrates to 1 over the range from the lower bound up to the upper bound (due to its normalization properties) (Clauset et al., 2009). While it is possible to incorporate such constraints in the regression analysis, it has never been reported to our knowledge. Moreover, it is unclear to what extent the standard errors obtained from these methods are valid (Clauset et al., 2009). Maximum likelihood methods do not suffer from such problems.

[8] See Appendix Tables A1,A2 for an overview of the distribution specifications.

points are parameters, truncation has no influence on the MLE-estimator for the shape parameter k. However, we need an estimation method for the truncation points.

There is no consensus on the practices to determine the minimum (maximum) of the (Inverse) Pareto-distribution. Some rely on visual techniques, looking for a 'kink' in the distribution above which the relationship between log rank and log size is approximately linear (Bas et al., 2015; di Giovanni and Levchenko, 2013). Some use export sales, and assume as such a truncation parameter equal to the minimum of sales, f.i. Freund and Pierola (2015). Some determine their minimum to ensure a Pareto parameter large enough to deliver finite moments when calibrating their theoretical models (Bee and Schiavo, 2017; Head et al., 2014). Others estimate the minimum, assuming a mixed Log-normal/Pareto distribution (Bakar and Nadarajah, 2013; Malevergne et al., 2011; Nigai, 2017). Such methods are however, either subject to possibly large measurement errors and inconsistencies or restrictive in their need to assume a distributional relation both for the bulk and the tail of the distribution.

Obtaining an accurate estimate for the lower bound is, however, vital to the accuracy of the estimated shape parameter $\hat{k}$. Choosing a minimum too low results in a biased shape parameter, as we will be fitting a power-law to non-power-law data. Choosing a value too high, on the other hand, increases the statistical error and bias from finite size effects on the shape parameter, as we throw legitimate data points away. Moreover, it is widely documented that the minimum and shape parameter of the (Truncated) Pareto distribution exhibit a positive correlation Bee and Schiavo (2017); di Giovanni and Levchenko (2013); Eeckhout (2004); Freund and Pierola (2015); Head et al. (2014).

In order to obtain an accurate estimate for the lower bound, we rely on a formal decision rule developed by Clauset et al. (2009). For the ordered productivity set $\{\omega_f; f = 1, \ldots, F\}$, we evaluate every ω_f as a potential ω_{min} (ω_{max}), estimating the ML estimate of the power-law exponent $k_{1,2}$. We then use the Kolmogorov-Smirnov statistic to select the optimum ω_{min} (ω_{max}). It is defined as the cut-off which minimizes the quantity

$$\begin{aligned} T_{KS,\hat{\omega}_{max}} &= \sup_{\omega \leq \hat{\omega}_{max}} |S(\omega) - ParetoInv(\omega; \hat{k}_1, \hat{\omega}_{max})| \\ T_{KS,\hat{\omega}_{min}} &= \sup_{\omega \geq \hat{\omega}_{min}} |S(\omega) - Pareto(\omega; \hat{k}_2, \hat{\omega}_{min})| \end{aligned}$$

where $S(\omega)$ is the cumulative distribution function (CDF) of the observed values for $\omega_f \geq \omega_{min}$ ($\omega_f \leq \omega_{max}$).

3.1.2 FMM

Complete data The maximum likelihood estimation of a FMM (see eq. 1) is not straightforward. This because the number of components is unobserved. The log-likelihood function can be written as

$$logL(\omega|\mathbf{\Psi}) = \sum_{f=1}^{F}\sum_{k=1}^{K} z_{fk}\left[log(\pi_k) + log(m_k(\omega|\theta_k))\right] \tag{4}$$

where z_{fk} is the unobserved component that equals one if ω_f originated from subpopulation k and zero otherwise. Two steps need to be taken iteratively in order to be able to maximize this equation. The E-step of the s-th iteration consists of determining the conditional expectation of eq. 4 given the observed data and the current parameter estimates from iteration $s-1$:

$$\begin{aligned} Q(\mathbf{\Psi}|\mathbf{\Psi}^{(s-1)}) &= E\left[logL(\omega|\mathbf{\Psi})|\omega, \mathbf{\Psi}^{(s-1)}\right] \\ &= \sum_{f=1}^{F}\sum_{k=1}^{K} \pi_{fk}^{(s)}\left[log(\pi_k) + log(m_k(\omega|\boldsymbol{\theta}_k))\right] \end{aligned} \tag{5}$$

where the missing data z_{fk} is replaced by the posterior probability that ω_f belongs to the kth mixture:

$$\pi_{fk}^{(s)} = E\left[z_{fk}|\omega_f, \mathbf{\Psi}^{(s-1)}\right] = \frac{\pi_k^{(s-1)} m_k(\omega_f|\boldsymbol{\theta}_k^{(s-1)})}{\sum_{k=1}^{K}\pi_k^{(s-1)} m_k(\omega_f|\boldsymbol{\theta}_k^{(s-1)})}.$$

The M-step then, consists of maximizing the Q-function over the parameters Ψ:

$$\mathbf{\Psi}^{(s)} = \max_{\mathbf{\Psi}} Q(\mathbf{\Psi}|\mathbf{\Psi}^{(s-1)}) \tag{6}$$

Each iteration updates the E- and M-step until the algorithm converges (See Miljkovic and Grün (2016) for a more elaborate overview). This algorithm can be applied to many distributions. We restrict ourselves to a range of distributions that are commonly used in the economics literature and which ensure a unique equilibrium in the Melitz (2003)-model: the Lognormal, Weibull, Fréchet, Gamma, Burr and Exponential distribution.

Truncated data The EM-algorithm can easily be adapted to fitting data only to the tail(s) of the data. For this we need to specify the conditional densities

$$\begin{aligned} g(\omega|\boldsymbol{\Psi}, \omega^l \leq \omega \leq \omega^u) &= \frac{\sum_{k=1}^{K} \pi_k m_k(\omega|\boldsymbol{\theta_k})}{G(\omega^u|\boldsymbol{\Psi}) - G(\omega^l|\boldsymbol{\Psi})} \\ &= \sum_{k=1}^{K} \pi_k \frac{M_k(\omega^u|\boldsymbol{\theta_k}) - M_k(\omega^l|\boldsymbol{\theta_k})}{G(\omega^u|\boldsymbol{\Psi}) - G(\omega^l|\boldsymbol{\Psi})} \frac{m_k(\omega|\boldsymbol{\theta_k})}{M_k(\omega^u|\boldsymbol{\theta_k}) - M_k(\omega^l|\boldsymbol{\theta_k})} \\ &= \sum_{k=1}^{K} \eta_k m_k(\omega|\boldsymbol{\theta_k}, \omega^l \leq \omega \leq \omega^u) \end{aligned} \tag{7}$$

with $\eta_k > 0$, $\sum_{k=1}^{K} \eta_k = 1$. The Q-function becomes

$$\begin{aligned} Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(s-1)}) &= E\left[logL(\omega|\boldsymbol{\Psi})|\omega, \boldsymbol{\Psi}^{(s-1)}\right] \\ &= \sum_{f=1}^{F} \sum_{k=1}^{K} \pi_{fk}^{(s)} \left[log(\eta_k) + log(m_k(\omega|\boldsymbol{\theta}_k, \omega^l \leq \omega \leq \omega^u))\right] \end{aligned} \tag{8}$$

where the the posterior probability that ω_f comes from the kth mixture is not affected by the truncation:

$$\pi_{fk}^{(s)} = \frac{\eta_k^{(s-1)} m_k(\omega_f|\boldsymbol{\theta}_k^{(s-1)}, \omega^l \leq \omega \leq \omega^u))}{\sum_{k=1}^{K} \eta_k^{(s-1)} m_k(\omega_f|\boldsymbol{\theta}_k^{(s-1)}), \omega^l \leq \omega \leq \omega^u)} = \frac{\pi_k^{(s-1)} m_k(\omega_f|\boldsymbol{\theta}_k^{(s-1)})}{\sum_{k=1}^{K} \pi_k^{(s-1)} m_k(\omega_f|\boldsymbol{\theta}_k^{(s-1)})}.$$

The M-step then again consists of maximizing the Q-function over the parameters Ψ. Iterating over the E- and M-step until the algorithm converges provides us with distributions fitted to the truncated data.

3.1.3 Pareto-tailed distributions

When we evaluate the fit of Pareto-tailed distributions to the complete dataset and not only the tails, we need to specify a parametric form for the bulk of the distribution. The most common choice is the Lognormal distribution, as it is the only distribution we know of that already has parametric specifications with Pareto tails on both ends and has been used extensively in the academic literature. We consider both the Double-Pareto Lognormal distribution (Reed, 2002), which is the product of a Double-Pareto with a Lognormal distribution, and the Pareto-tailed lognormal distribution, which is a component distribution of Pareto tails and the Lognormal (Ioannides and Skouras, 2013; Luckstead and Devadoss, 2017). Reducing the parameter spaces of these distributions allows us to consider distributions with only a left- or right-Pareto tail, resulting in the Left- and Right-Pareto Lognormal and Left- and Right-Pareto-tailed Lognormal distributions respectively. We estimate these distributions using standard ML methods.

3.2 Distribution selection

We rely on several criteria to differentiate between the distributions. First, we consider whether the proposed parametric distribution is a sufficiently good fit to the data. Then, we differentiate between distributions using information criteria.

Goodness of fit We use the Kolmogorov-Smirnov (KS) test to evaluate whether the proposed parametric distribution is a good fit for the empirical distribution. The KS-statistic quantifies the largest distance between the empirical distribution $S(\omega)$ and the parametric distribution $G(\omega; \cdot)$:

$$T_{KS} = \sup |S(\omega) - G(\omega; \cdot)|$$

As we estimate our distributions on the empirical data itself, the Kolmogorov distribution cannot be used to define critical values for this test statistic. Therefore, we calculate the p-value using a Monte Carlo-analysis (Clauset et al., 2009). We draw multiple, MC times, sample data under the null hypothesis that our data is generated by the proposed parametric distribution and calculate how many times we get a KS-test statistic at least as big as the original test statistic $\frac{\#(T^*_{KS} \geq T_{KS})}{MC}$. When we rely on the KS-test for truncated distributions, we sample data both from the empirical data and from the parametric distribution in line with the null hypothesis.

Information Criteria We differentiate between the distributions based on the negative log-likelihood, the Akaike or Bayesian Information Criteria.

The Negative Log-Likelihood is $NLL = -\sum_{f=1}^{F} lng(\omega_f; \cdot)$. Differentiation between two distributions is based on the likelihood ratio

$$LR = \sum_{f=1}^{F} ln\frac{g_1(\omega_f; \cdot)}{g_2(\omega_f, \cdot)} \tag{9}$$

with $g_{1,2}$ the probability densities of the rivaling distributions. If these distributions are *nested*, the test statistic amounts to minus two times this ratio, which follows a chi-squared distribution with 1 degree of freedom Wilks (1938).If these distributions are *non-nested*, the test statistic will be the sample average of this ratio, standardized by a consistent estimate of its standard deviation. The null hypothesis states that both classes of distributions are equally far (in the Kullback-Leibler divergence/relative entropy sense) from the true distribution. If this is true, our test statistic will follow (asymptotically) a Gaussian distribution with mean zero. If the null is false, and $g_1(\cdot)$ is closer to the truth, the test statistic diverges to $+\infty$ with probability one. If $g_2(\cdot)$ fits the data better, it diverges to $-\infty$.

The Aikaike Information criteria is defined as $AIC = 2np - 2ln(L)$ with np the number of parameters and $ln(L)$ the log likelihood. The Bayesian Information criteria is defined as $BIC = npln(F) - 2ln(L)$. Differentiation between distributions relies on relative distance of the BIC's: $\Delta BIC = BIC_1 - BIC_2$. The value of ΔBIC can be implies strong evidence in favor of distribution 1 if $B > 10$, moderate evidence if $6 < B \leq 10$ and weak evidence if $2 < B \leq 6$ (Kass and Raftery, 1995).

4 Data

Our dataset is based on the Sistema de Contas Integradas das Empresas (SCIE, Enterprise Integrated Accounts System) of Portugal.[9] This database covers the universe of firms in manufacturing and services over the years 2004–2007, containing detailed information on firm's balance sheets and income statements. As Cabral and Mata (2003) already demonstrated, Portugal is not special as to what concerns the firm sales distribution in comparison with other countries such as France, the U.S. . . . In a later stadium of the paper, we plan to validate our Portuguese dataset relying on Orbis data, similar to Nigai (2017).

We compare the number of enterprises, total employment and turnover present in our dataset with the OECD structural SBDS database for the year 2006 in Appendix Table A3.[10] This table demonstrates the full coverage of our dataset, containing 100% percent of the SBDS data for all size classes and industries. The large coverage of our dataset implies a coverage of both individual entrepreneurs and other corporations. For this paper, we discard individual entrepreneurs and focus on the manufacturing sector.

[9]The SCIE has already been used by a.o. (Bastos et al., 2018; Carreira and Teixeira, 2016; Dias et al., 2016; Fernandes and Ferreira, 2017; Fonseca et al., 2018).

[10]Representativity is guaranteed over the complete sample period, see Figures A2. The divergence in 2004 is probably due a censoring problem where non-active firms are present in the dataset but not in official figures. These non-active firms disappear, however, when we discard the individual entrepreneurs from the dataset.

5 Results

We fit the considered distributions to Portuguese manufacturing sales in the year 2006. We focus initially on FMM of Lognormal distributions, as it allows for a straightforward comparison with the Pareto-tailed Lognormal distributions. We limit our discussion to up to 4 component FMM, as this proves sufficient for our main message. We show that our results can be extended to other distributions. Our results are also externally validated using city size distributions.

Our empirical results strongly reject the Pareto distribution in favour of FMMs. Table 1 displays the distributions fits to Portuguese firms sales in 2006 and results from the selection tests, both for the complete and tails data. The distribution are ordered according to their Negative Log-likelihood ranking R_{NLL}, with Log-likelihood rankings with parameter correction $R_{AIC,BIC}$ stated in the adjacent columns. Starting with the complete data, we see that a Pareto distribution is a really bad fit, with a Log-Likelihood, a KS-test statistic and cumulative KS-deviance up to 10 times bigger than a simple Lognormal distribution. These results matches the graphical image from Figure 1. Going further, we observe that Pareto-tailed Lognormals provide a better fit to the data than the Lognormal distribution. However, if we allow for the Lognormal distribution to differ between discrete subpopulations, it performs significantly better in all goodness-of-fit measures and information criteria, even when correcting for the number of parameters.

The Pareto distribution performs also worse than FMMs in the tails of the distribution. We fitted the (Inverse) Pareto to the (left) right tail of the distribution using the methods described in the methodology section (Section 3). We then recovered the best-fitting truncation parameter for the Pareto distribution and fitted truncated FMMs of Lognormals to both tails of the distribution. This test puts the Pareto distribution twice in the advantage. First, it is free from a parametric specification for the bulk of the distribution and second, the truncation parameter is chosen where it fits best for Pareto. Nevertheless, we still find a better fit for FMMs of Lognormals to the tails of the distribution. Likelihood ratio tests indicate that from 3-component Lognormals onwards this fit is significantly better than Pareto. These results provide, given the bias towards Pareto in these tests, strong evidence in favour False Pareto and FMM.

Figure 2 provides more insight into the numerical results of Table 1. It plots both the absolute KS-deviance (left) and cumulative absolute KS-deviance (right) between the parametric and empirical distribution. The figure shows the large errors from the Lognormal distribution. We also observe how augmenting the Lognormal distribution with a right tail, as in Nigai (2017), improves the fit only marginally. While it does allow for a slightly better fit in the right tail of the distribution, this comes at the cost of a worse fit to the left-tail of the distribution and an almost as-bad fit to the bulk of the distribution. The best-fitting Pareto-tailed Lognormal, the Double-Pareto Lognormal does a better job at fitting the distribution. However, it clearly lags behind in comparison with the best-fitting FMM which only displays marginal errors both in the bulk and the tails of the data.

Table 1: Distribution fits to the Portuguese firms sales distribution in 2006 demonstrate the superior fit of FMM to the complete dataset as well as to the tails.

			Goodness of fit		Information Criteria			
Distribution	Comp.	Par.	T_{KS}	$\sum_{f=1}^{F} KS_f$	Loglike[1]	R_{NLL}	R_{AIC}	R^2_{BIC}
Complete data								
Lognormal	4	11	0.0044	49.44	-3728.47	1	1	2^{+++}
Lognormal	3	8	0.0044	56.83	-3733.37	2	2	1^{+++}
Lognormal	2	5	0.0086	133.17	-3843.45	3	3	3^{+}
Double-Pareto Lognormal	1	4	0.0108	170.39	-3853.37	4	4	4
Pareto-tailed Lognormal	1	4	0.0318	635.16	-4174.49	5	5	5
Left-Pareto-tailed Lognormal	1	3	0.0414	801.71	-4471.81	6	6	6
Right-Pareto-tailed Lognormal	1	3	0.0324	640.37	-4654.69	7	7	7
Left-Pareto Lognormal	1	3	0.0434	886.51	-4698.91	8	8	8
Right-Pareto Lognormal	1	3	0.0350	732.97	-4744.90	9	9	9
Lognormal	1	2	0.0393	872.65	-4847.17	10	10	10
Pareto	1	1	0.4914	9455.27	-67831.05	11	11	11
Left tail								
Lognormal	4	11	0.0070	27.68	32005.55**	1	2	4
Lognormal	3	8	0.0080	32.55	32003.44**	2	1	3
Lognormal	2	5	0.0157	77.93	31996.02	3	4	2
Inverse Pareto	1	-	0.0121	68.92	31994.79	4	-	-
Lognormal	1	2	0.9928	5762.88	31994.64	5	3	1
Right tail								
Lognormal	3	8	0.6205	363.45	-3977.68***	1	1	1
Lognormal	4	11	0.0102	4.10	-4531.72*	2	4	4
Lognormal	1	2	0.0148	6.11	-4532.18**	3	2	2
Lognormal	2	5	0.0148	6.12	-4532.22**	4	3	3
Pareto	1	-	0.0251	16.13	-4537.17	5	-	-

Notes: The FMM are fitted to the tails with truncation parameters determined by the best-fitting (Inverse) Pareto distribution. The distributions indicated in grey are the best-fitting (paretly) Pareto distributions to the Log-likelihood.
[1] ***,**,* indicates significance at the 0.01,0.05,0.1 level respectively for the p-value of the Log-likelihood ratio test **in the tails** between FMM and the (Inverse) Pareto distribution.
[2] +++ indicates that the difference between BIC of the distribution and the best-scoring Pareto-Lognormal combination (grey) **for the complete dataset** is greater than 10:$\Delta BIC > 10$, ++ indicates that $6 < \Delta\ BIC \leq 10$ and + indicates that $2 < \Delta\ BIC \leq 6$.

5.1 Extension to other distributions

The superior fit of FMMs is not limited to the Lognormal distribution, and can be expanded to f.i. the Burr and/or Weibull distribution. Table A4 displays the test results of fits to the complete data expanding to FMMs of distributions often used in the economic literature such as the Exponential, Gamma, Weibull, Burr and Fréchet distribution. Overall, most of these distributions are not able to match the performance of the Lognormal. The

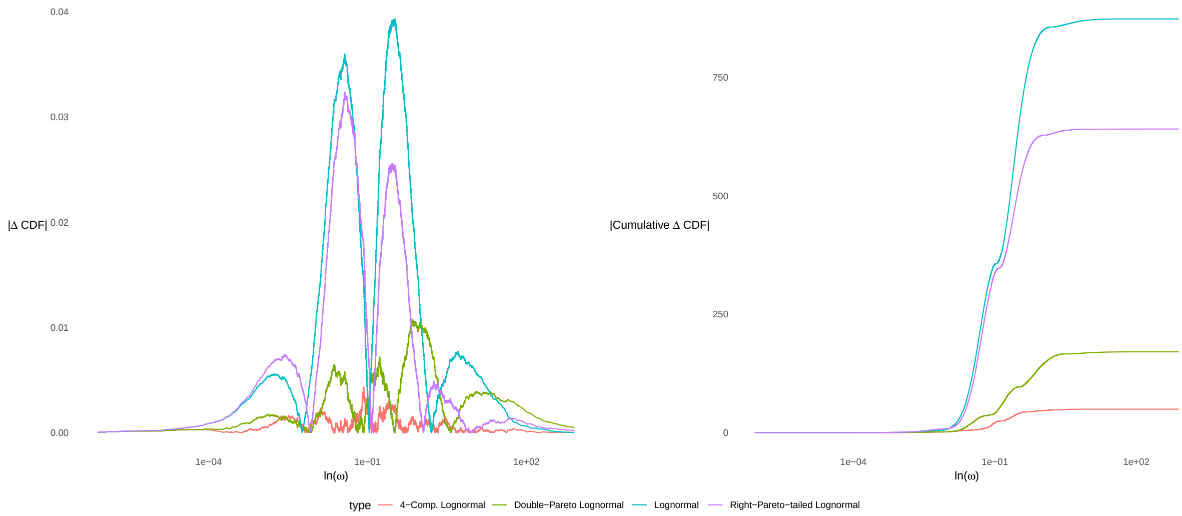

Figure 2: Absolute deviations (left) and cumulative absolute deviations (right) of 1-, and 4-component Lognormal, Double-Pareto Lognormal and Right-Pareto-tailed Lognormal distributions from the empirical distribution (left) reveal the superior performance of multi-component FMMs over the complete range of the data – Portugal 2006.

Burr distribution, on the other hand, provides an even better fit to the data, while also FMM of the Weibull and exponential distribution provide a better fit than Pareto-tailed Lognormals.

5.2 External validation: city size distributions

We validate our results externally fitting the same distributions the the U.S. Census 200 city size distribution data. This dataset has been subject to an extensive debate in the city size literature, including the famous discussion between Eeckhout (2009) and Levy (2009). Table A5 showcases the test results, demonstrating that the city size distribution is neither Lognormal, nor Pareto or Pareto-tailed Lognormal. It is best approximated by a FMM of Burr distributions. These results are in line with Kwong and Nadarajah (2019).

6 Policy application: Gains From Trade[11]

We investigate the consequences of the findings from the previous section for quantitative policy recommendations. Specifically, we perform a GFT exercise similar to Melitz and Redding (2015) and Bee and Schiavo (2017). In the meantime, we showcase the practical convenience of combining FMM with heterogeneous firms models à la Melitz (2003).

[11]See Appendix B for a full workout of the model.

We consider trade between two symmetric countries i and j, and choose labor in one country as the numeraire, so that $W_i = W_j = 1$. We calibrate the structural parameters in order to match stylized facts in our dataset, similar to (Bee and Schiavo, 2017; Head et al., 2014; Melitz and Redding, 2015). The variable trade costs τ is calibrated to match the average fraction of exports in firms sales in manufacturing[12]. We choose fixed entry costs f^e to deliver a entry rate of 0.5 (Head et al., 2014) and fixed exporting costs f_{ij} to match the export participation rate. Domestic fixed costs are set to one $f_{ii} = 1$, as the quantity of relevance to calculate GFT is the ratio between domestic and exporting costs f_{ij}/f_{ii}. This calibration setup ensures equal export participation at the initial situation, disregarding the distribution (Bee and Schiavo, 2017). All distributional parameters are recovered from our empirical analysis. We adapt the parameters obtained from firm size distribution to match the productivity distribution using the power transformation $Cr^{\frac{1}{1-\sigma}} \sim \omega$.[13] We set the elasticity of substitution σ equal to four. The average fraction of exports in firm sales and the export participation rates are obtained from our dataset.

We compute the GFT by evaluating the change in welfare from its calibrated state U due to an exogenous increase in variable/fixed trade costs τ/f to τ'/f':

$$\text{GFT} = \frac{U - U'}{U'} \text{ x } 100. \tag{10}$$

Results of the calibration exercise in Table 2 show the procentual losses/gains from moving a symmetric economy from a calibrated state to autarky or free trade via variable trade costs for considered distributions. The distributions are ordered according to their Negative Log-likelihood. As such, we can interpret the four-component Burr distribution as providing the closest values to the truth.[14]. We see that jumping to a less-able distribution involves large errors in the calculations of the GFT.

Note how the Pareto distribution is not included in the calibration exercise. This is due to the requirement that for the Pareto distribution, the shape parameter is larger than the elasticity of substitution minus one $k > \sigma - 1$. The parameter estimates for k being close to one are too small to be considered for the model. This issue is common in the literature, and is usually resolved by fitting the Pareto distribution only to a small part of

[12] $\frac{\tau^{1-\sigma_s}}{1+\tau^{1-\sigma_s}} = \frac{\sum_{j \neq i} X_{ijs}}{X_{is}}$, assuming a CES demand system and symmetry across countries.

[13] The change in variables needed for this transformation per distribution are provided in Table A1. The constant is unknown, but not of relevance for GFT-calculations. FMM of distribution types are equally closed to power transformations as their underlying distributions.

[14] Nigai (2017) provides nonparametric baseline values. We also tested such nonparametric estimates, and found them to, to a small extent, underestimate the GFT. This is not surprising, as we mentioned the difficulty to fit a heavy tail without parametric assumptions (cf. infra). A nonparametric density estimate will decrease to zero very fast after the last observation, while a parametric estimate has to follow the imposed shape.

the dataset (Bas et al., 2015; Bee and Schiavo, 2017; di Giovanni and Levchenko, 2013). Given our setup of investigating gains from trade for the complete distribution, however, we refrain from such methods.

Table 2: Procentual losses/gains from moving a symmetric economy from a calibrated state[1] to autarky or free trade via variable trade costs

Distribution	Comp.	Autarky	Free Trade
Burr	4	-8.314	13.692
Lognormal	4	-5.413	13.916
Weibull	4	-5.580	14.106
Double-Pareto Lognormal	1	-21.562	16.034
Left-Pareto Lognormal	1	-12.198	12.929
Right-Pareto Lognormal	1	-15.359	13.706
Lognormal	1	-4.069	4.415

Notes: [1]Fixed entry costs are calibrated to deliver a entry rate of 0.5 and fixed exporting costs to match the export participation rate of Portugal in 2006. This calibration setup ensures equal export participation at the initial situation, disregarding the distribution.

7 Conclusion

This paper challenges the dominance of Pareto in heterogeneous firms models. Relying on the population of Portuguese firms, we build a clear statistical framework that refutes the Pareto assumption in favour of Finite Mixture Models. FMMs are theoretically favourable, building on the existence of discrete subpopulations in the data and generalize as such existing theories on firm dynamics. Moreover, they can be easily implemented into the class of heterogeneous firms models à la Melitz (2003). A policy exercise calculates the Gains From Trade between rivaling distributions and demonstrates the importance of shifting away from Pareto.

While we find strong evidence in favour of FMMs, we take no stance on distribution type nor on the mixing parameter that defines the discrete subpopulations. It is clear that the two are closely interconnected, and therefore not easily identifiable. Future research will be necessary to take a closer look at the sources of distinctive distributions within the population of firms.

The idea of FMMs opens many new venues for research. For instance, the estimation of firm heterogeneity, productivity, usually relies a first-order Markov process that is identical for the complete population. Concurrently, however, it is recognized that firms can differ in their productivity evolution depending on o.a. firm age Olley and Pakes (1996), exporting behaviour De Loecker (2013), innovation behavior Costantini and Melitz (2008) .... Introducing Finite Mixture Modeling into the estimation procedures would allow, semi-parametrically, to control for such discrete subpopulations without the risk of model misspecification. Simultaneously, it would provide the opportunity to research the sources of discrete subpopulations within sales from a deeper, more detailed level.

References

Arkolakis, C. (2016). A Unified Theory of Firm Selection and Growth. *The Quarterly Journal of Economics 131*(1), 89.

Arkolakis, C., A. Costinot, and A. Rodríguez-Clare (2012). New Trade Models, Same Old Gains? *American Economic Review 102*(1), 94–130.

Axtell, R. L. (2001). Zipf Distribution of U.S. Firm Sizes. *Science 293*(5536), 1818–1820.

Bakar, S. A., N. Hamzah, M. Maghsoudi, and S. Nadarajah (2015). Modeling loss data using composite models. *Insurance: Mathematics and Economics 61*, 146 – 154.

Bakar, S. A. A. and S. Nadarajah (2013). CompLognormal: An R Package for Composite Lognormal Distributions. *The R Journal 5*(2), 97–103.

Bas, M., T. Mayer, and M. Thoenig (2015). From micro to macro: demand, supply, and heterogeneity in the trade elasticity. mimeo.

Bastos, P., J. Silva, and E. Verhoogen (2018). Export Destinations and Input Prices. *American Economic Review 108*(2), 353–392.

Bee, M. and S. Schiavo (2017). Powerless: gains from trade when firm productivity is not pareto distributed. *Review of World Economics*.

Cabral, L. M. B. and J. Mata (2003, September). On the evolution of the firm size distribution: Facts and theory. *American Economic Review 93*(4), 1075–1090.

Carreira, C. and P. Teixeira (2016). Entry and exit in severe recessions: lessons from the 2008–2013 Portuguese economic crisis. *Small Business Economics 46*(4), 591–617.

Carvalho, V. and B. Grassi (2015, April). Large firm dynamics and the business cycle. Economics Working Papers 1481, Department of Economics and Business, Universitat Pompeu Fabra.

Chaney, T. (2008, September). Distorted gravity: The intensive and extensive margins of international trade. *American Economic Review 98*(4), 1707–21.

Chaney, T. (2018). The gravity equation in international trade: An explanation. *Journal of Political Economy 126*(1), 150–177.

Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009). Power-Law Distributions in Empirical Data. *SIAM Review 51*(4), 661–703.

Costantini, J. and M. Melitz (2008). The dynamics of firm-level adjustment to trade liberalization. *The organization of firms in a global economy 4*, 107–141.

Córdoba, J. C. (2008). A generalized gibrat's law. *International Economic Review 49*(4), 1463–1468.

De Loecker, J. (2013, August). Detecting learning by exporting. *American Economic Journal: Microeconomics 5*(3), 1–21.

di Giovanni, J. and A. A. Levchenko (2012). Country size, international trade, and aggregate fluctuations in granular economies. *Journal of Political Economy 120*(6), 1083–1132.

di Giovanni, J. and A. A. Levchenko (2013). Firm entry, trade, and welfare in zipf's world. *Journal of International Economics 89*(2), 283–296.

Dias, D. A., C. R. Marques, and C. Richmond (2016). Misallocation and productivity in the lead up to the Eurozone crisis. *Journal of Macroeconomics 49*, 46–70.

Eeckhout, J. (2004). Gibrat's Law for (All) Cities. *American Economic Review 94*(5), 1429–1451.

Eeckhout, J. (2009). Gibrat's Law for (All) Cities: Reply. *The American Economic Review 99*(4), 1676–1683.

Feenstra, R. C. (2014). Restoring the Product Variety and Pro-competitive Gains from Trade with Heterogeneous Firms and Bounded Productivity. Working Paper 19833, National Bureau of Economic Research.

Fernandes, A. M., P. J. Klenow, S. Meleshchuk, M. D. Pierola, and A. Rodríguez-Clare (2015). The Intensive Margin in Trade: Moving Beyond Pareto. mimeo.

Fernandes, A. P. and P. Ferreira (2017). Financing constraints and fixed-term employment: Evidence from the 2008-9 financial crisis. *European Economic Review 92*, 215–238.

Fonseca, T., F. Lima, and S. C. Pereira (2018). Understanding productivity dynamics: A task taxonomy approach. *Research Policy 47*(1), 289–304.

Freund, C. and M. D. Pierola (2015). Export Superstars. *The Review of Economics and Statistics 97*(5), 1023–1032.

Gabaix, X. (2009). Power laws in economics and finance. *Annual Review of Economics 1*(1), 255–294.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica 79*(3), 733–772.

Gabaix, X. and R. Ibragimov (2011). Rank - 1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents. *Journal of Business & Economic Statistics 29*(1), 24–39.

Giesen, K., A. Zimmermann, and J. Suedekum (2010). The size distribution across all cities – double pareto lognormal strikes. *Journal of Urban Economics 68*(2), 129 – 137.

Head, K., T. Mayer, and M. Thoenig (2014). Welfare and Trade without Pareto. *American Economic Review 104*(5), 310–16.

Helpman, E., M. Melitz, and Y. Rubinstein (2008). Estimating Trade Flows: Trading Partners and Trading Volumes. *The Quarterly Journal of Economics 123*(2), 441.

Hopenhayn, H. A. (1992). Entry, exit, and firm dynamics in long run equilibrium. *Econometrica 60*(5), 1127–1150.

Hsieh, C.-T. and R. Ossa (2016). A global view of productivity growth in China. *Journal of International Economics 102*, 209–224.

Ioannides, Y. and S. Skouras (2013). Us city size distribution: Robustly pareto, but only in the tail. *Journal of Urban Economics 73*(1), 18 – 29.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Kwong, H. S. and S. Nadarajah (2019). A note on "pareto tails and lognormal body of us cities size distribution". *Physica A: Statistical Mechanics and its Applications 513*, 55 – 62.

Levy, M. (2009). Gibrat's Law for (All) Cities: Comment. *The American Economic Review 99*(4), 1672–1675.

Luckstead, J. and S. Devadoss (2017). Pareto tails and lognormal body of us cities size distribution. *Physica A: Statistical Mechanics and its Applications 465*, 573 – 578.

Luttmer, E. G. (2007). Selection, growth, and the size distribution of firms. *The Quarterly Journal of Economics 122*(3), 1103–1144.

Malevergne, Y., V. Pisarenko, and D. Sornette (2011, Mar). Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. *Phys. Rev. E 83*, 036111.

Melitz, M. J. (2003). The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity. *Econometrica 71*(6), 1695–1725.

Melitz, M. J. and S. J. Redding (2015). New Trade Models, New Welfare Implications. *American Economic Review 105*(3), 1105–46.

Miljkovic, T. and B. Grün (2016). Modeling loss data using mixtures of distributions. *Insurance: Mathematics and Economics 70*, 387 – 396.

Nigai, S. (2017). A tale of two tails: Productivity distribution and the gains from trade. *Journal of International Economics 104*, 44–62.

Olley, S. and A. Pakes (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica 64*(6), 1263–1297.

Perline, R. (2005). Strong, weak and false inverse power laws. *Statistical Science 20*(1), 68–88.

Reed, W. J. (2001). The pareto, zipf and other power laws. *Economics Letters 74*(1), 15 – 19.

Reed, W. J. (2002). On the rank-size distribution for human settlements. *Journal of Regional Science 42*(1), 1–17.

Reed, W. J. and M. Jorgensen (2004). The double pareto-lognormal distribution—a new parametric model for size distributions. *Communications in Statistics - Theory and Methods 33*(8), 1733–1753.

Sager, E. and O. A. Timoshenko (2017). The EMG distribution and Trade Elasticities. mimeo, Bureau of Labor Statistics.

Virkar, Y. and A. Clauset (2014). Power-law distributions in binned empirical data. *The Annals of Applied Statistics*, 89–119.

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics 9*(1), 60–62.

Yang, M.-J. (2017). Global Superstars under Log-Normally Distorted Gravity. mimeo.

A Appendix

A.1 Figures

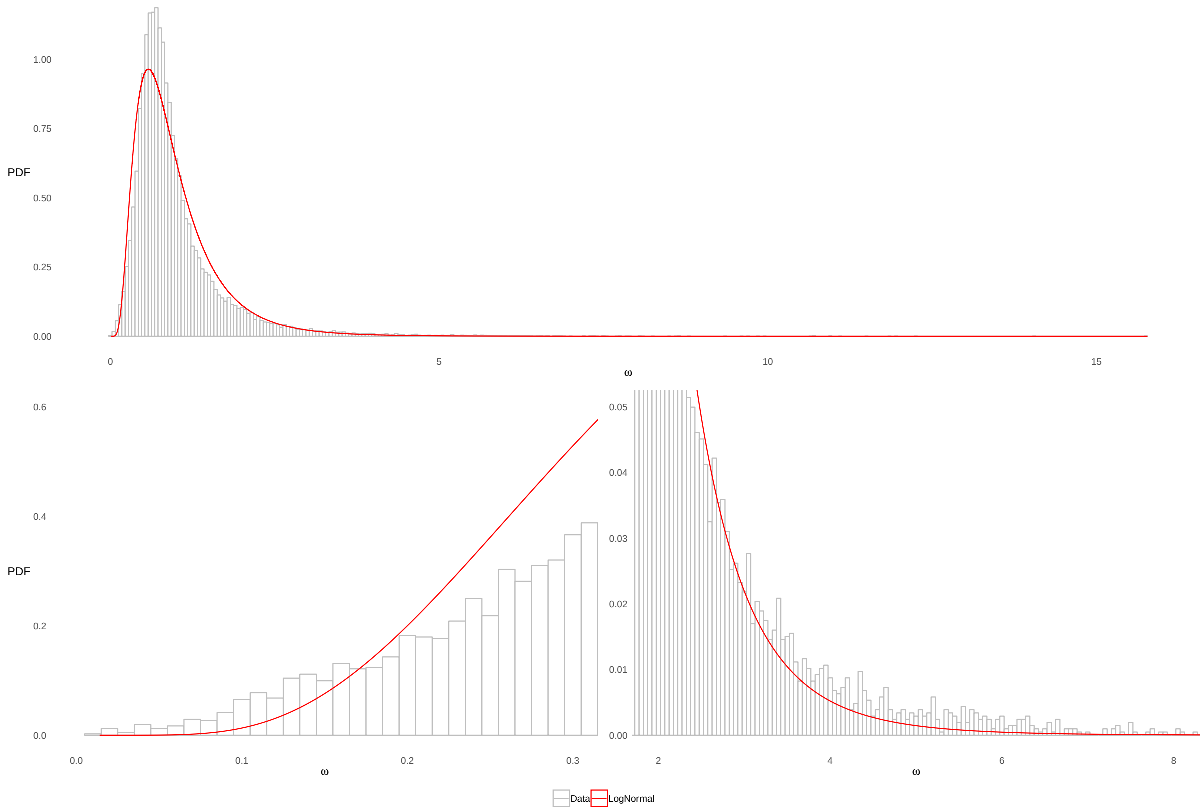

Figure A1: Empirical and Log-normal probability density functions of Portuguese productivity in 2006 (upper panel) with focus on the left (lower left panel) and right tail (lower right panel). **Notes:** Productivity, ω, is measured as domestic sales tot the power of $1/(\sigma - 1)$ with $\sigma = 4$. The Log-normal distribution is fitted using Maximum-Likelihood.

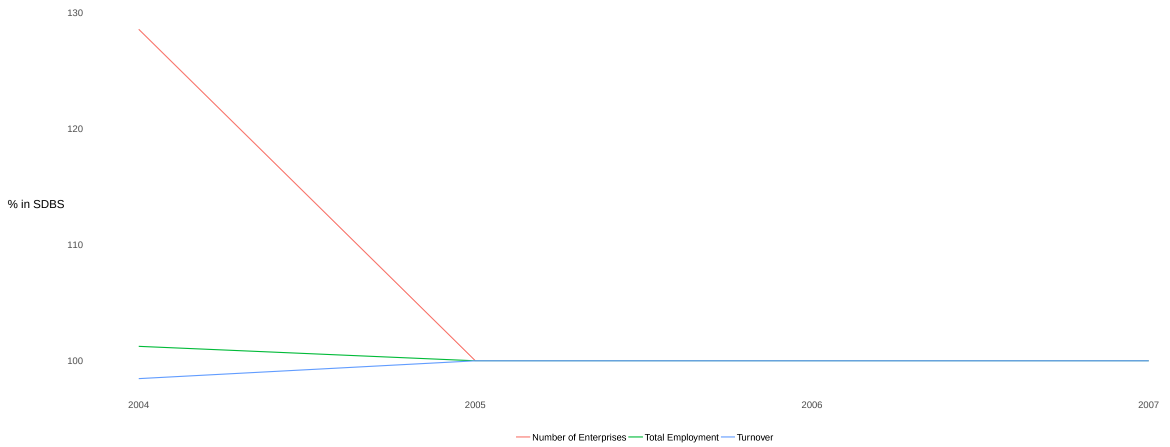

Figure A2: Evolution of the ratio of the number of enterprises, total employment and turnover of the SCIE sample of firms compare to the OECD structural SBDS database over time showcases a divergence in the year 2004.

A.2 Tables

Table A1: General overview of the distributions considered

Distribution	Abbreviation	Support	Parameters	Power transformation $a\omega^b$
(Truncated) Pareto	(T)P	$[\omega_{min}, \omega_{max}]$	$k, \omega_{min}, \omega_{max}$	$kb, \omega_{min}/a, \omega_{max}/a$
Lognormal	LN	$[0, \infty[$	μ, Var	$1/b(\mu - lna)$, Var/b
Weibull	W	$[0, \infty[$	k, s	$bk, s/a$
Burr	B	$[0, \infty[$	k, c, s	k, bc, s/a
Fréchet	F	$[0, \infty]$	k, s	kb, s/a
Gamma	G	$[0, \infty[$	k, s	-
Exponential	Exp	$[0, \infty[$	k, s	-
Double-Pareto Lognormal	DPLN	$[0, \infty[$	k_1, μ, Var, k_2	$\frac{k_1}{b}, b\mu + log(a), Var, \frac{k_2}{b}$
Left-Pareto Lognormal	LPLN	$[0, \infty[$	k_1, μ, Var	$\frac{k_1}{b}, b\mu + log(a), Var$
Right-Pareto Lognormal	RPLN	$[0, \infty[$	μ, Var, k_2	$b\mu + log(a), Var, \frac{k_2}{b}$
Pareto-tailed Lognormal	PTLN	$[0, \infty[$	μ, Var, η, τ	
Pareto-left-tail Lognormal	PLTLN	$[0, \infty[$	μ, Var, η	
Pareto-right-tail Lognormal	PRTLN	$[0, \infty[$	μ, Var, τ	

Table A2: Overview of the probability and cumulative density functions of distributions considered.

Distribution	**PDF**	**CDF**
(T)P	$\frac{k}{\omega_{min}^{-k}-\omega_{max}^{-k}}\omega^{-k-1}$	$1-\frac{\omega^{-k}-\omega_{max}^{-k}}{\omega_{min}^{-k}-\omega_{max}^{-k}}$
LN	$\frac{1}{Var\sqrt{2\pi}}e^{-(ln\omega-\mu)^2/2Var^2}$	$\Phi\left(\frac{ln\omega-\mu}{Var}\right)$
W	$\frac{k}{s}\left(\frac{\omega}{s}\right)^{k-1}e^{-\left(\frac{\omega}{s}\right)^{k}}$	$1-e^{-\left(\frac{\omega}{s}\right)^{k}}$
B	$\frac{\frac{kc}{s}\left(\frac{\omega}{s}\right)^{c-1}}{\left(1+\left(\frac{\omega}{s}\right)^{c}\right)^{k+1}}$	$1-\frac{1}{\left(1+\left(\frac{\omega}{s}\right)^{c}\right)^{k}}$
F	$\frac{k}{s}\left(\frac{\omega}{s}\right)^{-1-k}e^{-\left(\frac{\omega}{s}\right)^{-k}}$	$e^{-\left(\frac{\omega}{s}\right)^{-k}}$
G	$\frac{1}{s^k\Gamma(k)}\omega^{k-1}e^{-\frac{\omega}{s}}$	$\frac{1}{\Gamma(k)}\gamma(k,\frac{\omega}{\theta})$
Exp	$\frac{e^{-\frac{\omega}{s}}}{s}$	$1-e^{-\frac{\omega}{s}}$

Table A2: Overview of the probability and cumulative density functions of distributions considered.

Distribution	**PDF**	**CDF**
DPLN[1]	$\frac{k_2k_1}{k_2+k_1}\left[\omega^{-k_2-1}e^{k_2\mu+\frac{k_2^2Var^2}{2}}\Phi\left(\frac{ln\omega-\mu-k_2Var^2}{Var}\right)+\omega^{k_1-1}e^{-k_1\mu+\frac{k_1^2Var^2}{2}}\Phi^c\left(\frac{ln\omega-\mu+k_1Var^2}{Var}\right)\right]$	$\Phi\left(\frac{ln\omega-\mu}{Var}\right)-\frac{1}{k_2+k_1}\left[k_1\omega^{-k_2}e^{k_2\mu+\frac{k_2^2Var^2}{2}}\Phi\left(\frac{ln\omega-\mu-k_2Var^2}{Var}\right)-k_2\omega^{k_1}e^{-k_1\mu+\frac{k_1^2Var^2}{2}}\Phi^c\left(\frac{ln\omega-\mu+k_1Var^2}{Var}\right)\right]$
LPLN[1]	$k_1\omega^{k_1-1}e^{-k_1\mu+\frac{k_1^2Var^2}{2}}\Phi^c\left(\frac{ln\omega-\mu+k_1Var^2}{Var}\right)$	$\Phi\left(\frac{ln\omega-\mu}{Var}\right)-\omega^{k_1}e^{-k_1\mu+\frac{k_1^2Var^2}{2}}\Phi^c\left(\frac{ln\omega-\mu+k_1Var^2}{Var}\right)$
RPLN[1]	$k_2\omega^{-k_2-1}e^{k_2\mu+\frac{k_2^2Var^2}{2}}\Phi\left(\frac{ln\omega-\mu-k_2Var^2}{Var}\right)$	$\Phi\left(\frac{ln\omega-\mu}{Var}\right)-\omega^{-k_2}e^{k_2\mu+\frac{k_2^2Var^2}{2}}\Phi\left(\frac{ln\omega-\mu-k_2Var^2}{Var}\right)$
PTLN[1,2]	$\begin{cases} cb\omega^{\beta-1} & \text{if } \omega<\eta \\ \frac{b}{(2\pi)^{1/2}\omega Var}e^{-\frac{(log(\omega)-\mu)^2}{2Var^2}}, & \text{if } \eta\le\omega\le\tau \\ ab\omega^{-\zeta-1}, & \text{if } \tau<\omega \end{cases}$	$\begin{cases} \frac{cb\omega^{\beta}}{\beta} & \text{if } \omega<\eta \\ \ldots+\frac{b}{2}\left(2\Phi\left(\frac{log(q)-\mu}{Var}\right)-1\right)-\frac{b}{2}\left(2\Phi\left(\frac{log(\eta)-\mu}{Var}\right)-1\right), & \text{if } \eta\le\omega\le\tau \\ \ldots-ab\frac{\omega^{-\zeta}-\tau^{-\zeta}}{\zeta} & \text{if } \tau<\omega \end{cases}$
PLTLN[1,2]	$\begin{cases} cb\omega^{\beta-1} & \text{if } \omega<\eta \\ \frac{b}{(2\pi)^{1/2}\omega Var}e^{-\frac{(log(\omega)-\mu)^2}{2Var^2}}, & \text{if } \eta\le\omega \end{cases}$	$\begin{cases} \frac{cb\omega^{\beta}}{\beta} & \text{if } \omega<\eta \\ \ldots+\frac{b}{2}\left(2\Phi\left(\frac{log(q)-\mu}{Var}\right)-1\right)-\frac{b}{2}\left(2\Phi\left(\frac{log(\eta)-\mu}{Var}\right)-1\right), & \text{if } \eta\le\omega \end{cases}$
PRTLN[1,2]	$\begin{cases} \frac{b}{(2\pi)^{1/2}\omega Var}e^{-\frac{(log(\omega)-\mu)^2}{2Var^2}}, & \text{if } \omega\le\tau \\ ab\omega^{-\zeta-1}, & \text{if } \tau<\omega \end{cases}$	$\begin{cases} \frac{b}{2}\left(2\Phi\left(\frac{log(q)-\mu}{Var}\right)-1\right)-\frac{b}{2}\left(2\Phi\left(\frac{log(\eta)-\mu}{Var}\right)-1\right), & \text{if } \omega\le\tau \\ \ldots-ab\frac{\omega^{-\zeta}-\tau^{-\zeta}}{\zeta} & \text{if } \tau<\omega \end{cases}$

Notes: [1] Φ and Φ^c stand for the standard normal and complementary standard normal cdfs. [2] $\beta=-\frac{log(\eta)-\mu}{Var^2}$, $\zeta=\frac{log(\tau)-\mu}{Var^2}$, $a=\frac{\tau^{\zeta+1}}{(2\pi)^{1/2}\tau Var}e^{-\frac{(log(\tau)-\mu)^2}{2Var^2}}$, $c=\frac{\eta^{1-\beta}}{(2\pi)^{1/2}\eta Var}e^{-\frac{(log(\eta)-\mu)^2}{2Var^2}}$, $b=\frac{\eta}{\beta(2\pi)^{1/2}\omega Var}e^{-\frac{(log(\omega)-\mu)^2}{2Var^2}}+\Phi\left(\frac{\tau-\mu}{Var}\right)-\Phi\left(\frac{\eta-\mu}{Var}\right)$

Table A3: SCIE vs OECD structural SDBS database

	Number of Enterprises						Total Employment						Turnover					
Industry	1-9	10-19	20-49	50-249	> 250	Total	1-9	10-19	20-49	50-249	> 250	Total	1-9	10-19	20-49	50-249	> 250	Total
15	100	100	100	100	100	100	100	100	100			100	100	100	100			100
17	100	100	100	100	100	100				100	100	100				100	100	100
18	100	100	100	100	100	100				100	100	100				100	100	100
19	100	100	100	100	100	100		100	100	100				100	100	100		
20	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
21	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
22	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
24	100	100	100	100	100	100				100	100					100	100	
25	100	100	100	100	100	100				100	100	100				100	100	100
26	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
27	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
28	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
29	100	100	100	100	100	100	100			100	100	100	100			100	100	100
30	100	100	100	100	100	100												
31	100	100	100	100	100	100			100	100	100	100			100	100	100	100
32	100	100	100	100	100	100		100	100	100	100			100	100	100	100	
33	100	100	100	100	100	100	100	100	100				100	100	100			
34	100	100	100	100	100	100			100	100	100				100	100	100	
35	100	100	100	100	100	100			100	100	100				100	100	100	
36	100	100	100	100	100	100		100		100	100	100		100		100	100	100
37	100	100	100	100		100				100		100				100		100

Each cell corresponds to the ratio of our dataset compared to the data from the OECD structural SDBS database for the year 2006. Size classes are based on total employment.

Table A4: Distribution fits to the Portuguese firms sales distribution in 2006 demonstrate how the superiority of FMM is dependent on distribution type.

			Goodness of fit		Information Criteria			
Distribution	Comp.	Par.	T_{KS}	$\sum_{f=1}^{F} KS_f$	Loglike	R_{NLL}	R_{AIC}	R^1_{BIC}
Burr	4	15	0.0022	23.12	-3708.79	1	2	5^{+++}
Burr	3	11	0.0027	26.73	-3711.72	2	1	2^{+++}
Lognormal	4	11	0.0044	49.44	-3728.47	3	4	4^{+++}
Burr	2	7	0.0038	60.58	-3729.07	4	3	1^{+++}
Lognormal	3	8	0.0044	56.83	-3733.37	5	5	3^{+++}
Weibull	4	11	0.0057	79.85	-3746.92	6	6	6^{+++}
Weibull	3	8	0.0084	150.35	-3823.35	7	7	7^{+++}
Exponential	4	7	0.0118	155.98	-3834.93	8	8	9^{++}
Lognormal	2	5	0.0086	133.17	-3843.45	9	9	8^{+}
Double-Pareto Lognormal	1	4	0.0108	170.39	-3853.37	10	10	10
Burr	1	3	0.0116	251.28	-3889.44	11	11	11
Exponential	3	5	0.0141	254.05	-4151.30	12	12	12
Pareto-tailed Lognormal	1	4	0.0318	635.16	-4174.49	13	13	13
Left-Pareto-tailed Lognormal	1	3	0.0414	801.71	-4471.81	14	14	14
Weibull	2	5	0.0217	429.86	-4505.41	15	15	15
Right-Pareto-tailed Lognormal	1	3	0.0324	640.37	-4654.69	16	16	16
Left-Pareto Lognormal	1	3	0.0434	886.51	-4698.91	17	17	17
Right-Pareto Lognormal	1	3	0.0350	732.97	-4744.90	18	18	18
Lognormal	1	2	0.0393	872.65	-4847.17	19	19	19
Gamma	4	11	0.0559	802.53	-5008.75	20	20	20
Gamma	3	8	0.0567	817.61	-5070.34	21	21	21
Fréchet	4	11	0.0792	1783.03	-6470.70	22	24	24
Fréchet	3	8	0.0792	1783.19	-6470.77	23	23	23
Fréchet	2	5	0.0792	1783.17	-6470.77	24	22	22
Gamma	2	5	0.0724	1186.15	-6529.06	25	25	25
Exponential	2	3	0.0558	1163.66	-6907.25	26	26	26
Weibull	1	2	0.0988	2416.23	-9873.50	27	27	27
Fréchet	1	2	0.1003	2616.61	-11504.42	28	28	28
Gamma	1	2	0.1812	4179.77	-19646.73	29	29	29
Pareto	1	1	0.4914	9455.27	-67831.05	30	30	30
Exponential	1	1	0.4261	10227.39	-Inf	31	31	31

Notes: The distribution indicated in grey is the best-fitting combination of Pareto and Lognormal according to BIC. [1] $^{+++}$ indicates that the difference between BIC of the distribution and the best-scoring Pareto-Lognormal combination (grey) is greater than 10:$\Delta BIC > 10$, $^{++}$ indicates that $6 < \Delta\ BIC \leq 10$ and $^{+}$ indicates that $2 < \Delta\ BIC \leq 6$.

Table A5: Distribution fits to the U.S. Census 2000 city size distribution confirm the supremacy of FMM.

			Goodness of fit		Information Criteria			
Distribution	Comp.	Par.	T_{KS}	$\sum_{f=1}^{F} KS_f$	Loglike	R_{NLL}	R_{AIC}	R^1_{BIC}
Burr	4	15	0.0036	17.76	-6009.36	1	2	5^{+++}
Burr	3	11	0.0041	24.03	-6012.93	2	1	2^{+++}
Lognormal	4	11	0.0066	50.48	-6018.37	3	4	4^{+++}
Lognormal	3	8	0.0069	51.15	-6019.41	4	3	1^{+++}
Lognormal	2	5	0.0079	56.80	-6046.38	5	5	3^{+++}
Burr	2	7	0.0095	89.02	-6065.49	6	6	7
Right-Pareto Lognormal	1	3	0.0132	166.64	-6084.57	7	7	6
Double-Pareto Lognormal	1	4	0.0132	166.65	-6084.58	8	8	8
Pareto-tailed Lognormal	1	4	0.0175	252.85	-6134.77	9	9	9
Right-Pareto-tailed Lognormal	1	3	0.0167	231.73	-6141.82	10	10	10
Lognormal	1	2	0.0189	268.16	-6152.28	11	11	11
Left-Pareto-tailed Lognormal	1	3	0.0189	268.19	-6152.30	12	12	12
Left-Pareto Lognormal	1	3	0.0189	268.21	-6152.33	13	13	13
Weibull	4	11	0.0136	130.93	-6189.17	14	14	14
Fréchet	4	11	0.0145	144.51	-6222.94	15	15	15
Fréchet	3	8	0.0172	195.27	-6281.96	16	16	16
Exponential	4	7	0.0186	136.46	-6297.88	17	17	17
Burr	1	3	0.0218	295.48	-6370.10	18	18	18
Weibull	3	8	0.0174	185.45	-6393.30	19	19	19
Fréchet	2	5	0.0256	320.89	-6538.43	20	20	20
Exponential	3	5	0.0276	276.10	-6632.86	21	21	21
Weibull	2	5	0.0303	322.74	-6919.85	22	22	22
Gamma	4	11	0.0635	442.06	-7384.14	23	23	24
Fréchet	1	2	0.0459	641.06	-7403.89	24	24	23
Gamma	3	8	0.0645	454.42	-7432.55	25	25	25
Gamma	2	5	0.0850	722.70	-8485.61	26	26	27
Exponential	2	3	0.0716	840.90	-8488.20	27	27	26
Weibull	1	2	0.0838	1130.31	-9029.84	28	28	28
Gamma	1	2	0.1686	2071.06	-13893.39	29	29	29
Pareto	1	1	0.4176	5059.16	-31612.23	30	30	30
Exponential	1	1	0.3770	5574.95	-Inf	31	31	31

Notes: The distribution indicated in grey is the best-fitting combination of Pareto and Lognormal according to BIC. [1] $^{+++}$ indicates that the difference between BIC of the distribution and the best-scoring Pareto-Lognormal combination (grey) is greater than 10:$\Delta BIC > 10$, $^{++}$ indicates that $6 < \Delta\ BIC \leq 10$ and $^{+}$ indicates that $2 < \Delta\ BIC \leq 6$

B Heterogeneous firms model

We present a multi-country, multi-sector generalization of the workhorse heterogeneous firm model of Melitz and Redding (2015), in accordance with Hsieh and Ossa (2016). This allows us to investigate the importance of variations in the parameters of the productivity distribution.

Setup Imagine a world with N countries, each country $i \in N$ populated by L_i identical households. Each household supplies inelastically one unit of labor, earning wage W_i. They make their consumptions choices according to a nested Cobb-Douglas-CES utility function:

$$U_i = \prod_{s=1}^{S} \left(\sum_{i=1}^{N} \int_0^{M_{ijs}} q_{ijs} \left(\nu_{is}\right)^{\frac{\sigma_s - 1}{\sigma_s}} d\nu_{is} \right)^{\frac{\sigma_s}{\sigma_s - 1}\alpha_{js}} , \quad (11)$$

where S denotes the number of industries, M_{ijs} is the number of firms from country i serving market j in industry s, q_{ijs} is the quantity of an industry s variety from country i consumed in country j, α_{is} is the fraction of country j income spend on industry s varieties and $\sigma_s > 1$ the elasticity of substitution between industry s varieties.

Sectoral equilibrium Profit maximization of the firm results in an equilibrium price as a constant markup over marginal costs $p_{ijs} = \frac{\sigma_s}{\sigma_s - 1} \frac{\tau_{ijs} W_i}{\omega}$, resulting in the revenue of a firm from i in j with productivity ω:

$$r_{ijs}(\omega) = \left(\frac{\sigma_s}{\sigma_s - 1} \frac{\tau_{ijs} W_i}{\omega} \right)^{1-\sigma_s} P_{js}^{\sigma_s - 1} \alpha_s W_j L_j, \quad (12)$$

with P_{js} the CES price index in country j and sector s. Only firms that are productive enough can survive $\omega \geq \omega^*_{iis}$ and/or export $\omega \geq \omega^*_{ijs}$ defined under the zero-profit condition:

$$\sigma_s W_i f_{ijs} = \left(\frac{\sigma_s}{\sigma_s - 1} \frac{\tau_{ijs} W_i}{\omega^*_{ijs}} \right)^{1-\sigma_s} P_{js}^{\sigma_s - 1} \alpha_s W_j L_j$$

Moreover, free entry implies that in equilibrium, the expected value of entry must be equal to the sunk cost of entry:

$$\sum_{j=1}^{N} f_{ijs} \int_{\omega^*_{ijs}}^{\infty} \left[\left(\frac{\omega}{\omega^*_{ijs}} \right)^{\sigma_s - 1} - 1 \right] p(\omega) d\omega = \sum_{j=1}^{N} f_{ijs} J(\omega^*_{ijs}) = f^e_{is}$$

$S * N$ free-entry equations and $S * N$ cut-off conditions determine the $2(S * N)$ cut-off conditions in function of wages, fixed costs and distributional parameters. These cut-offs, then, allow us to determine all firm performance measures (relative to wages).

General equilibrium Imposing an equilibrium on the labour market (total demand = total supply) in each country allows us to obtain an expression for the mass of entering firms that only depends on labor supply

$$L_{is} = M^e_{is} \sum_{j=1}^{N} \sigma_s f_{ijs} \omega_{ijs}^{1-\sigma_s} \int_{\omega^*_{ijs}}^{\infty} \omega^{\sigma_s - 1} f_{is}(\omega) d\omega = M^e_{is} \sum_{j=1}^{N} \sigma_s f_{ijs} \left(\omega^*_{ijs} \right)^{1-\sigma_s} I(\omega^*_{ijs}) \tag{13}$$

where the industry labour shares are obtained from cost-minimization

$$L_{is} = \alpha_s L_i.$$

An equilibrum on the goods market is obtained from imposing trade balance

$$W_i L_i = \sum_{s=1}^{S} \sum_{j=1}^{N} \frac{X_{jis}}{X_{js}} \alpha_s W_j L_j, \tag{14}$$

where X_{jis} denotes total sales by firms from j in i for sector s

$$X_{ijs} = M^e_{is} \sigma_s W_i f_{ijs} \left(\omega^*_{ijs} \right)^{1-\sigma_s} I(\omega^*_{ijs}).$$

B.1 Parametrization

In order to solve the model, we need to parametrize two statistics:

$$I(\omega^*_{ijs}) = \int_{\omega^*_{ijs}}^{\infty} (\omega)^{\sigma_s - 1} f_{is}(\omega) d\omega$$

and

$$
\begin{aligned}
J(\omega_{ijs}^*) &= \int_{\omega_{ijs}^*}^{\infty} \left[\left(\frac{\omega}{\omega_{ijs}^*} \right)^{\sigma_s - 1} - 1 \right] f_{is}(\omega) d\omega \\
&= \left(\omega_{ijs}^*\right)^{1-\sigma_s} I(\omega_{ijs}^*) - \int_{\omega_{ijs}^*}^{\infty} f_{is}(\omega) d\omega \\
&= \left(\omega_{ijs}^*\right)^{1-\sigma_s} I(\omega_{ijs}^*) - \int_{0}^{\infty} f_{is}(\omega) d\omega + \int_{0}^{\omega_{ijs}^*} f_{is}(\omega) d\omega \\
&= \left(\omega_{ijs}^*\right)^{1-\sigma_s} I(\omega_{ijs}^*) - 1 + F_{is}\left(\omega_{ijs}^*\right)
\end{aligned}
$$

For most of the considered distributions, we approach the $I(\omega_{ijs}^*)$ numerically relying on the Matlab command 'integral'. We provide parameterizations for the (finite mixture) of the Lognormal distributions.

B.1.1 LogNormal

The parametrization of the K-Mixture of Lognormals equals

$$1 - F_{is}\left(\omega_{ijs}^*\right) = 1 - \sum_{k=1}^{K} \Phi\left(\frac{ln(\omega_{ijs^*}) - \mu_k}{Var_k}\right)$$

and

$$I(\omega_{ijs}^*) = \sum_{k=1}^{K} e^{\frac{1}{2}[Var_k(\sigma_s - 1)]^2 + \mu_k(\sigma - 1)} \left(1 - \Phi\left(\frac{ln(\omega_{ijs^*}) - \mu_k}{Var_k} - Var_k\left(\sigma_s - 1\right)\right)\right).$$

It is easy to see how these parametrizations reduce to the standard Lognormal parametrizations with $K = 1$:

$$1 - F_{is}\left(\omega_{ijs}^*\right) = 1 - \Phi\left(\frac{ln(\omega_{ijs^*}) - \mu}{Var}\right)$$

$$I(\omega_{ijs}^*) = e^{\frac{1}{2}[Var(\sigma_s - 1)]^2 + \mu(\sigma - 1)} \left(1 - \Phi\left(\frac{ln(\omega_{ijs^*}) - \mu}{Var} - Var\left(\sigma_s - 1\right)\right)\right).$$

This parametrization clearly showcases the easy integration of FMM into Melitz (2003)-type models.