Unobserved Heterogeneity in the Productivity Distribution and Gains From Trade.

Ruben Dewitte, Michel Dumont, Glenn Rayp, Peter Willemé*

26th May 2020

Abstract

Finding a good parametric approximation to the productivity distribution is a problem of general interest. This paper argues that heterogeneity in productivity is best captured by Finite Mixture Models (FMMs). FMMs build on the existence of unobserved subpopulations in the data. As such, they are generally consistent with models of firm dynamics differing between groups of firms and allow for a very flexible distribution fit. We find FMMs to increase this fit by more than 70% compared to currently considered distributions. A Gains From Trade exercise reveals that only FMMs approximate the 'true gains' reasonably well.

Keywords: Finite Mixture Model, firm size distribution, productivity distribution, Gains From Trade

JEL Codes: L11, F11, F12

*Dewitte: Ghent University, Tweekerkenstraat 2, 9000 Ghent, Belgium, rubenl.dewitte@ugent.be, Dumont: Federal Planning Bureau and Ghent University, Rue Belliard - Belliardstraat 14-18 1040 Brussels, dm@plan.be , Rayp: Ghent University, Tweekerkenstraat 2, 9000 Ghent, Belgium, glenn.rayp@ugent.be, Willemé: Federal Planning Bureau and Ghent University, Rue Belliard - Belliardstraat 14-18 1040 Brussels, pw@plan.be. We thank Pete Klenow, Luca David Opromolla, Gonzague Vannoorenberghe, Gerdie Everaert, Bruno Merlevede, Marijn Verschelde and participants at the 2018 ETSG conference in Warsaw, the 2019 ETSG conference in Bern, the DEGIT XIV conference in Odense and the 2019 workshop on "Firm Heterogeneity in Technical Change" in Ghent for their helpful comments and suggestions.

1 Introduction

Finding a good parametric approximation to the productivity distribution is a problem of general interest. Its importance can be appraised by its large influence on various research fields. First, the mechanisms driving firm-level dynamics in aggregate growth models are determined by the parametric approximation of the productivity distribution (see for instance Luttmer (2007); Arkolakis (2016)). Second, the propagation of firm-level volatility to the aggregate level mainly relies on a Pareto specification for the right tail of the productivity distribution (Gabaix, 2011; di Giovanni and Levchenko, 2012; Carvalho and Grassi, 2019). In the international trade literature, it is recognized that different choices for the productivity distribution significantly affect Gains From Trade (GFT) estimates (Head et al., 2014; Nigai, 2017; Bee and Schiavo, 2018) and alters the channels through which trade affects welfare (Arkolakis et al., 2012; Bas et al., 2017; Melitz and Redding, 2015; Fernandes et al., 2018).

To date, there is no consensus on what this parametric approximation should be. Some authors argue a single distributional form such as Pareto (Axtell, 2001), Lognormal (Head et al., 2014) or Weibull (Bee and Schiavo, 2018) suffices to define the productivity distribution. Others build on the idea that a single distribution can not adequately capture the heterogeneity in productivity. This results in combinations of distributions such as the Double-Pareto (Arkolakis, 2016), Double-Pareto Lognormal (Sager and Timoshenko, 2019) or Lognormal-Pareto (Nigai, 2017). Nevertheless, Dewitte (2020) demonstrates that none of these currently considered distributions are able to provide a sufficiently good fit to the data.

This paper argues that heterogeneity in the productivity distribution can be captured most adequately by Finite Mixture Models (FMMs). A FMM is essentially a weighted sum of an a priori unknown number of individual densities. As such, it is a semi-parametric approximation that allows for discrete subpopulations to define the overall distribution. The flexible, semi-parametric nature of FMMs renders them favorable both from a theoretical and empirical point of view.

From a theoretical point of view, the generative process of a FMM corresponds to a simple combination of the generative processes of the underlying individual densities. A FMM can therefore easily generalize, and is generally consistent with, existing models of firm dynamics. First, FMMs allows to combine a specified generative process of firm dynamics across groups of firms to capture additional, unspecified heterogeneity. Luttmer (2007), for instance, generalizes his single-sector model with a finite mixture specification to a multi-sector model. This in order to capture additional heterogeneity across industries and obtain a satisfactory fit to the data. Second, a finite mixture specification is generally consistent with the mechanisms considered to differentiate firm dynamics between groups of firms. The differences in growth rates between financially constrained and unconstrained firms by Cabral and Mata (2003), for instance, can be respecified into a finite mixture specification. FMMs provide an empirical tool that can account for dynamics to differ between groups of firms without having, but not excluding the possibility, to specify the mechanisms that drive these differences a priori. These mechanisms can be left ‘unobserved’.

We illustrate the excellent empirical performance of FMMs using the domestic sales[1] of the *population* of active Portuguese firms in 2006. Our contributions to the literature are threefold. First, we have access to a representative dataset on the sales distribution. This allows us to evaluate the performance of parametric distributions on the complete productivity distribution as well as to focus on both the *left* and right tail. Moreover, it insulates us from erroneous conclusions due to truncated or unrepresentative data in the left tail of the distribution (Perline, 2005). Second, we introduce a multitude of new, economically relevant distributions to the productivity distribution literature. Fitting and comparing up to 52 different distributions helps to reveal features of the data that are of importance when deciding on a specific parametric distribution. Third, our analysis relies on a clear statistical framework to distinguish between distributional fits. Based on the Bayesian Information Criterion (BIC), the currently favored Double-Pareto Lognormal (Sager and Timoshenko, 2019) and Lognormal-Pareto (Nigai, 2017) come in ranked sixteenth and thirty-first out of 52 distributions respectively, while FMMs top the charts. Moreover, a Kolmogorov-Smirnov test reveals that only FMMs provide a distribution fit that is not rejected by the data. FMMs reduce the maximum deviation from the empirical Cumulative Distribution Function (the Kolmogorov-Smirnov test statistic) by more than 70% compared to the Double-Pareto Lognormal distribution and by more than 90% compared to the Lognormal-Pareto distribution. This performance is not surprising, as we show that the Double-Pareto Lognormal and Lognormal-Pareto distribution can be interpreted as constraints of the more general mixture specification.

A Gains From Trade application demonstrates the importance of correctly approximating the productivity distribution in heterogeneous firms models à la Melitz (2003), and underlines the straightforward implementation of FMMs into such models. We contribute to the literature providing quantitative expressions necessary to calibrate a heterogeneous firms model for all distributions considered. Our calibration exercise reveals that when reducing variable trade costs by two thirds, FMMs are able to track the 'true GFT' (obtained from the empirical distribution) closely, while a single Lognormal distribution underestimates these GFT by ±11% and a Lognormal-Pareto distribution overestimates them by ±13%.

The paper is organized as follows. In the following section we start by linking the large literature on the parametric approximation of size distributions, spanning the fields of efficiency analysis, physics, regional and actuarial science, to the productivity distribution literature. From this overview, it becomes apparent that the literature on productivity distributions lacks a clear statistical framework that differentiates between a sufficiently large number of alternative distributions over a representative data range. We therefore establish a methodology that uniformly fits a large number of distributions both to complete and truncated datasets, and present evaluation methods to differentiate between these distribution in section 3. Our database on firm sales is discussed in section 4. We provide our empirical results in section 5 and discuss the implications of these results for GFT calculations in section 6. Section 7 concludes.

[1] We rely on the distributional relation between productivity and positive domestic sales, under specific model assumptions (Nigai, 2017; Dewitte, 2020), to evaluate parametric approximations of the productivity distribution.

2 Literature Review

This section provides an overview of the literature related to firm size/productivity distributions. We discuss why the Pareto distribution can only match the tail of size distributions while single hump-shaped distributions such as the Lognormal or the Weibull distribution can not accurately match both the tail and the bulk of the distribution. Size distributions are therefore best approximated by a combination of distributions, of which we consider three types: mixture, piecewise composite and multiplicative distributions. We argue that finite mixtures are preferable both from an empirical and theoretical point of view because of their flexible, semi-parametric nature.

2.1 Single distributions

The *Pareto distribution* has been dominating heterogeneous firms models (Melitz, 2003). Even though the Melitz (2003)-model is not restricted to this distributional choice, its empirical performance (see for instance Axtell (2001); Gabaix (2009); Levy (2009); di Giovanni et al. (2011)) and convenience led to a widespread reliance on the Pareto distribution for social welfare and economic policy analysis.[2] The fit of a Pareto distribution is usually evaluated using its Cumulative Distribution Function (CDF), which follows a straight line on a log-log scale with the shape parameter (k) as slope:

$$G_P(x; x_{min}, k) = 1 - \left(\frac{x_{min}}{x}\right)^k, \qquad x \geq x_{min}. \tag{1}$$

Figure 1 compares a fitted Pareto survival function (CDF^c = 1−CDF) with the empirical survival function of Portuguese firm-level sales in 2006 on a log-log scale for the complete dataset (upper panel). It is immediately clear that the Pareto distribution is not a good fit to the complete distribution due to the existence of a hump in the middle.[3]

The popularity of the Pareto distribution, however, rests on its ability to provide a close fit to lower-truncated[4] data with predominantly large observations.[5] Just as every curved line looks straight when one zooms in close enough, so too does the distribution of firm sales appear to be straight when truncated sufficiently. Both the left (lower left panel) and right tail (lower right panel) exhibit linearity of the CDF and survival functions respectively on a log-log scale, in line with

[2] See Arkolakis et al. (2012) for an overview of work relying on the Melitz-Pareto combination.

[3] See also the Probability Density Function (PDF) in Appendix Figure 2.

[4] An *upper-truncated* version of the Pareto distribution has also been used to explain the existence of zero trade flows across country pairs (Helpman et al., 2008; Feenstra, 2018) and to demonstrate the relevance of heterogeneous firms models (Melitz and Redding, 2014). A discussion on the economic relevance of, and an extension of the analysis to, upper-truncated distributions falls outside the scope of this paper. The methodology set out in this paper allows to truncate any kind of distribution both from above and/or below (see section 3).

[5] Note that the influential paper of Axtell (2001) does not rely on truncated data but unintentionally favors the Pareto distribution due to data binning (Virkar and Clauset, 2014) and methodological choices (Clauset et al., 2009; Bottazzi et al., 2015) characteristic of that time.

Pareto behavior in the tails of the distribution.[6] The apparent straight line behavior of the tails can therefore 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 distributions.[7] Proof of which is the performance of the Lognormal distribution in the lower panels of Figure 1.[8]

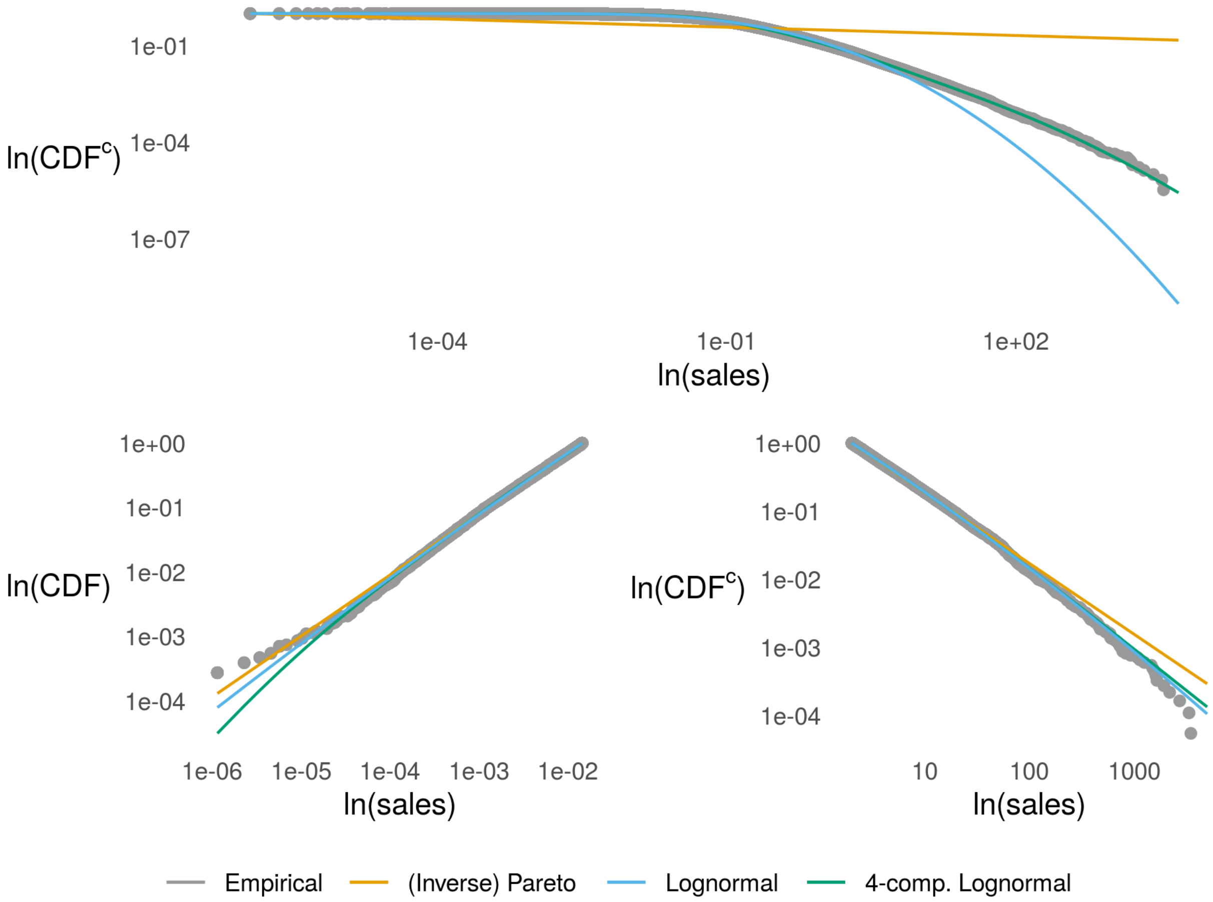

Figure 1: Empirical survival function of Portuguese domestic sales in 2006 (upper panel) on a log-log scale with fitted (Inverse) Pareto and (4-component mixture of) the Lognormal distributions. The lower left and right panels focus on distributions fitted solely to the left and right tail respectively. **Notes:** (Truncated) Distributions are fitted using maximum likelihood methods (cf. infra) to the complete and truncated datasets independently. Tail truncation points are determined by the best-fitting (Inverse) Pareto distributions according to the Kolmogorov-Smirnov statistic.

These alternative *hump-shaped distributions* are claimed to provide a better fit to complete size

[6] The Inverse Pareto distribution is specified as

$$G_{IP}(x; x_{max}, k) = 1 - \left(\frac{x_{max}}{x}\right)^{-k}, \qquad x \leq x_{max}.$$

[7] Perline (2005) defines this class of distributions within the Gumbel domain of attraction.

[8] Even though Pareto and Lognormal distributions exhibit qualitatively different behavior in their upper tails, their apparent quantitative similar behavior in the upper tail for Lognormals with large variance is well-documented (Malevergne et al., 2011).

distributions (see Bee and Schiavo (2018) for the Weibull and Eeckhout (2004, 2009); Head et al. (2014); Fernandes et al. (2018) for the Lognormal distribution). In the firm size literature, this claim is usually supported by comparing their performance with a limited number of alternative distributions, mostly Pareto, using the low-powered R-squared.[9] Even though homogeneous hump-shaped distributions such as the Lognormal can adequately fit the tail or the bulk of the empirical distribution, they cannot do both simultaneously. This is easily observable from the upper panel of Figure 1 where the single Lognormal distribution, when fitted to the complete size distribution, does not fit the right tail of the complete productivity distribution while matching the bulk rather satisfactorily.

2.2 Combined distributions

As single distributions are not capable of accurately matching both the bulk and the tail(s) of the productivity distribution, recent research focuses on combinations of distributions. We consider three types of combinations: mixture, piecewise composite and product distributions. To our knowledge, mixture distributions have not been fitted to the productivity distribution. Nevertheless, current applications of both the piecewise composite and product distributions can be interpreted as constraints of the more general mixture specification.

2.2.1 Mixture distributions

Finite Mixture Models (FMMs) are essentially a weighted sum of I individual densities $m_i(\cdot)$:

$$g(x|\boldsymbol{\Psi}) = \sum_{i=1}^{I} \pi_i m_i(x|\boldsymbol{\theta}_i), \qquad \pi_i \geq 0, \quad \sum_{i=1}^{I} \pi_i = 1 \tag{2}$$

where I represents the number of components or discrete subpopulations, π_i is the probability of belonging to component i, $\boldsymbol{\theta_i}$ the component-specific parameter vector of density $m_i(\cdot)$ and $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{I-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_I)$ is the vector of all model parameters (McLachlan and Peel, 2000). They are also referred to as Latent Class Models (LCM) provided that the number of components, and thus also the mixing parameter itself, does not have to be specified a priori but is determined by the data. As such, a finite mixture model provides a semi-parametric approach ideal to fully capture the heterogeneity of size distributions.[10]

The aptitude of Finite Mixture models has already been explored in the context of efficiency analysis (see for instance Beard et al. (1997); Orea and Kumbhakar (2004); El-Gamal and Inanoglu (2005); Greene (2005)), city sizes (Kwong and Nadarajah, 2019) and actuarial losses (Miljkovic and Grün, 2016). It has, to our knowledge, not been applied to productivity distributions before.

[9] See Clauset et al. (2009) for an explanation as to why the R-squared has low power in a distributional context.

[10] A semi-parametric approach is to be favored over a nonparametric approach in the case of heavy-tailed distributions such as firm size. This is because the heavy tails renders nonparametric procedures less efficient (Clauset et al., 2009; Dewitte, 2020).

The generative process of a FMM corresponds to a simple combination of the generative processes of the underlying individual densities and can therefore easily generalize, and is generally consistent with, existing models of firm dynamics.[11] First, FMMs allows to combine a specified generative process of firm dynamics across groups of firms to capture additional, unspecified heterogeneity. Luttmer (2007), for instance, generalizes his single-sector model with a finite mixture specification to a multi-sector model. This allows to capture additional heterogeneity across industries and obtain a satisfactory fit to the data. Similarly, Rossi-Hansberg and Wright (2007) argue the need to account for cross-sectoral differences in their initial single-sector model specification to achieve an accurate description of the cross-sectional size distribution of US firms.

Second, a finite mixture specification is generally consistent with the mechanisms that differentiate firm dynamics between groups of firms. Firm dynamics are argued to differ between groups of firms depending on whether or not they are financially constrained (Cooley and Quadrini, 2001; Cabral and Mata, 2003; Desai et al., 2003; Albuquerque and Hopenhayn, 2004; Clementi and Hopenhayn, 2006; Angelini and Generale, 2008), innovate (Costantini and Melitz, 2008; Atkeson and Burstein, 2010), add or drop products (Klette and Kortum, 2004; Lentz and Mortensen, 2008), add or drop management layers (Caliendo and Rossi-Hansberg, 2012; Caliendo et al., 2020), incur specific market penetration costs (Arkolakis, 2016), et cetera. As (Rossi-Hansberg and Wright, 2007, p. 1641) paraphrase Jovanovic (1982): “many of the mechanisms in the literature undoubtedly contributed toward an explanation of establishment dynamics”. To date, however, it remains unclear which mechanism, or mechanisms, dominate. There are “many sources of heterogeneity that support the idea of discrete subpopulations likely to differ in important characteristics” (Perline, 2005, p.80). Finite Mixture Models provide an empirical tool that can account for dynamics to differ between groups of firms as determined by the data. As such, they can account for most, or even a combination, of the proposed mechanisms without having to specify these mechanisms a priori. The mechanisms can be left ‘unobserved’.

2.2.2 Piecewise composite distributions

Piecewise composite distributions have a probability density specified as:

$$g(x|\boldsymbol{\theta}) = \begin{cases} \alpha_1 m_1^*(x|\boldsymbol{\theta}_1) & \text{if} \quad c_0 < x \leq c_1 \\ \alpha_2 m_2^*(x|\boldsymbol{\theta}_2) & \text{if} \quad c_1 < x \leq c_2 \\ \vdots & \qquad \vdots \\ \alpha_I m_I^*(x|\boldsymbol{\theta}_I) & \text{if} \quad c_{I-1} < x \leq c_I \end{cases} \tag{3}$$

where $\forall i \in I : m_i^*(x|\boldsymbol{\theta}_i) = \frac{m_i(x|\boldsymbol{\theta}_i)}{\int_{c_{i-1}}^{c_i} m_i(x|\boldsymbol{\theta}_i)dx}$ is the probability density function (PDF) of $m_i(x|\boldsymbol{\theta}_i)$

[11] Note that while this paper conceptualizes the generality of FMMs from a generative perspective, it is not able to provide evidence in favor of any specific generative process. See the methodology section (section 4), Appendix B and the conclusion (Section 7) for a more elaborate evaluation of current limitations regarding this paper’s discussion of (the generative processes of) FMMs.

truncated at the cutoffs c_{i-1}, c_i. For this distribution to be well-behaved, additional differentiability and continuity conditions are imposed that determine the value of both component cutoffs (c_i) and probabilities (α_i) (Bakar et al., 2015), so that the vector of all model parameters reduces to the combination of the component-specific parameter vectors: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_I)$.

While these composite distributions can be formed from many individual parametric distributions, applications mostly focus on Lognormal distributions with Pareto tails. The 'Inverse Pareto-Lognormal-Pareto' distribution has been applied in the city size literature (Ioannides and Skouras, 2013; Luckstead and Devadoss, 2017), while the 'Lognormal-Pareto' version was applied by Nigai (2017) to the Melitz (2003) model for GFT calculations. Dewitte (2020) generalizes the implementation of the piecewise composite distributions to allow for any underlying density in three-, and two- piecewise composite distributions, mainly focusing on Pareto-tailed piecewise composites.

From the distribution specification in equation 3, it can be observed that piecewise composite distributions can be interpreted as mixtures of truncated densities with component probabilities restricted to ensure continuity and differentiability (Scollnik, 2007).[12] This contrasts with the general mixture specification (eq. 2), where component probabilities can be interpreted as the probability that an individual observation belongs to a certain group of observations. Moreover, the generative process of piecewise distributions is rather ambiguous. It is for instance not clear yet which firm dynamics could explain the existence of hard cutoffs that separate the Lognormal from the Pareto distribution.

2.2.3 Product distributions

Alternatively, distributions can be combined into a product distribution: a probability distribution constructed as the distribution of the product of random variables having two other known distributions. The product distribution mainly used in the literature, the Double-Pareto Lognormal distribution, results from the product of a Lognormal with a (Double-)Pareto distributed random variable (Reed and Jorgensen, 2004). This distribution is found to approximate city size distributions well (Reed, 2002; Giesen et al., 2010), while Sager and Timoshenko (2019) applied the distribution to Brazilian export data.

A generative process for this Double-Pareto Lognormal distribution exists (Reed and Hughes, 2002; Reed, 2002; Reed and Jorgensen, 2004) and is applicable to heterogeneous firms models (Arkolakis, 2016). Interestingly, the Double-Pareto Lognormal distribution can be seen as a structured infinite mixture of Lognormal distributions (Reed, 2002, p.13).[13] The Double-Pareto Lognormal distribution can therefore be absorbed by the more flexible mixture distributions as specified

[12] This becomes even more clear when we rewrite the specification of the piecewise composite distribution (eq. 3) as the weighted sum of truncated densities: $g(x|\boldsymbol{\theta}) = \alpha_1 \mathbb{I}(c_0 < x \leq c_1) m_1^*(x|\boldsymbol{\theta}_1) + \alpha_2 \mathbb{I}(c_1 < x \leq c_2) m_2^*(x|\boldsymbol{\theta}_2) + \ldots + \alpha_I \mathbb{I}(c_{I-1} < x \leq c_I) m_I^*(x|\boldsymbol{\theta}_I)$.

[13] In the context of firm size, this could mean that each age (= time since entry in the market) group of firms is distributed Lognormally at a certain point in time. The reason the overall firm size distribution is not Lognormal is that these groups of firms have not all been evolving for the same length of time. The overall distribution of size will be a mixture of Lognormal distributions (across age groups) with time since entry as mixing parameter. When this mixing parameter is exponentially distributed, firm size will be Double-Pareto Lognormally distributed.

in equation 2. Whereas the Double-Pareto Lognormal may suffer from misspecification and/or oversimplification by imposing a structure on the mixture distribution, a FMM allows the data to determine the mixture structure needed to capture the heterogeneity that is present in the data.

3 Methodology

The literature review reveals the myriad of empirical evidence in favor of qualitatively very different distributions fits to productivity. This points at the lack of a clear statistical framework that differentiates between a sufficiently large number of distributions over a representative data range. In this section, we establish a methodology that uniformly fits the large, but relevant, range of single and combined distributions to both complete and truncated data. We then present statistical tests to differentiate between the fitted distributions.

3.1 Distribution fitting

We rely on Maximum Likelihood (ML)[14] over all firms $b \in B$ to fit all considered distributions to the data. We consider the (Inverse) Pareto, hump-shaped distributions (Lognormal, Weibull, Fréchet, Gamma, Exponential and Burr) and combinations of these distributions in the form of mixtures, piecewise composite or product distributions. We limit piecewise composite and product distributions to available Pareto-tailed extensions of the considered hump-shaped distributions.[15] In the case of FMMs, ML is wrapped in an Expectation-Maximization (EM) algorithm to estimate the component probabilities. The estimation methods allow to fit the distributions to both complete and truncated data. This will not only allow us to single out and focus on tail performance, but also to generalize the proposed distributional fits to unrepresentative and/or truncated data.

3.1.1 (Inverse) Pareto

Complete data The ML estimator for the shape parameter k over all firms $b \in B$ can easily be obtained as

[14] The choice for Maximum Likelihood contrasts with the productivity distribution literature, where popular fitting techniques rely on the minimization of squared errors between a log-linearization of the theoretical and empirical PDFs/CDFs (Axtell, 2001; di Giovanni and Levchenko, 2013; Head et al., 2014; Freund and Pierola, 2015; Bas et al., 2017; Nigai, 2017; Bee and Schiavo, 2018). 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 which extent the standard errors obtained from these methods are valid (Clauset et al., 2009; Bottazzi et al., 2015). Maximum likelihood methods do not suffer from such problems.

[15] See Appendix Tables 1, 2 and 3 for an overview of the specifications for all distributions considered. Considered distributions are chosen based on their occurrence in the economic literature.

$$k_{IP} = \left[\frac{1}{B} \sum_{b=1}^{B} ln \frac{x_{max}}{x_b} \right]^{-1}, \qquad k_P = \left[\frac{1}{B} \sum_{b=1}^{B} ln \frac{x_b}{x_{min}} \right]^{-1}. \tag{4}$$

The ML estimator of the scale parameters equals the maximum and minimum observation: $\hat{x}_{min} = \min(x)$, $\hat{x}_{max} = \max(x)$, as the likelihood function is monotonically increasing (decreasing) in x_{min} (x_{max}).

Truncated data The (Inverse) Pareto distribution is a special distribution, being truncated from (above) below by definition.[16] This means that the (upper) lower truncation point lies within the parameter space of the distribution, and distribution fits can be optimized accordingly. The ML estimator as specified above merely assumes the exogenously applied truncation points as scale parameter.

Obtaining an accurate estimate for the (upper) lower bound is, however, vital to the accuracy of the estimated shape parameter $\hat{k}$. Choosing a (maximum) minimum too (high) low results in a biased shape parameter, as one will be fitting a power-law to non-power-law data. Choosing a value too (low) high, on the other hand, increases the statistical error and bias from finite size effects on the shape parameter, as one discards legitimate data points. Moreover, it is widely documented that the minimum and shape parameter of the Pareto distribution exhibit a positive correlation (Eeckhout, 2004; di Giovanni and Levchenko, 2013; Head et al., 2014; Freund and Pierola, 2015; Bee and Schiavo, 2018).

Many practices therefore co-exist to determine the (upper) lower truncation point, without consensus on the best practice to determine this scale parameter of the (Inverse) Pareto-distribution. In the case of the 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 (di Giovanni and Levchenko, 2013; Bas et al., 2017). Some use export sales, and assume as such a truncation parameter equal to the minimum of sales, e.g. Freund and Pierola (2015). Others determine their minimum to ensure a Pareto parameter large enough to deliver finite moments when calibrating their theoretical models (Head et al., 2014; Bee and Schiavo, 2018). Still others estimate the minimum, assuming a mixed Lognormal-Pareto distribution (Malevergne et al., 2011; Bakar and Nadarajah, 2013; Nigai, 2017). Such methods are either subject to possibly large measurement errors and inconsistencies or restrictive in their need to assume a distributional relation between the bulk and the tail of the distribution.

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 $\{x_b; b = 1, \ldots, B\}$, we evaluate every x_b as a potential (x_{max}) x_{min}, estimating the ML estimate of the power-law exponent k. We

[16] Fully truncated (both from below and above) Pareto distributions can be deduced from a truncated probability density function (see eq. 6) and have been used in the economic literature (Helpman et al., 2008; Melitz and Redding, 2014; Feenstra, 2018).

then use the Kolmogorov-Smirnov statistic to select the optimum (x_{max}) x_{min}. It is defined as the cutoff which minimizes the maximum absolute deviation of the empirical from the theoretical CDF:

$$\begin{aligned} T_{KS,\hat{x}_{max}} &= \sup_{x \leq \hat{x}_{max}} \left| \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(x_b \leq \hat{x}_{max}) - G_{IP}(x; \hat{k}, \hat{x}_{max}) \right| \\ T_{KS,\hat{x}_{min}} &= \sup_{x \geq \hat{x}_{min}} \left| \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(x_b \geq \hat{x}_{min}) - G_{P}(x; \hat{k}, \hat{x}_{min}) \right|, \end{aligned} \tag{5}$$

where $\mathbb{I}_A$ is the indicator of event A.

3.1.2 Hump-shaped, piecewise composite and product distributions

Complete data The maximum likelihood of the considered *hump-shaped distributions* (Lognormal, Weibull, Fréchet, Gamma, Exponential and Burr) is straightforward and estimation methods are widely available. We also consider *piecewise composite distributions* as Pareto-tailed extensions of these hump-shaped distributions. The ML estimator of these distributions has no closed form and needs to be approached numerically, see Dewitte (2020). Pareto-tailed extensions in the form of *product distributions*, on the other hand, are less generally available. We consider the Double-Pareto Lognormal distribution (Reed and Jorgensen, 2004). This distribution is the result of multiplying a Double Pareto, used by among others Arkolakis (2016), with a Lognormal distribution. Reducing the parameter space of the Double Pareto allows us to consider the Left- and Right-Pareto Lognormal distribution respectively. Also in this case, the ML estimator has no closed form solution and needs to be approached numerically (Reed and Jorgensen, 2004).

Truncated data Consisting of individual truncated densities, the estimation of piecewise composite distributions on truncated data is by its definition straightforward. Maximum likelihood methods for the remaining hump-shaped and product distributions can easily be adapted by truncating the distribution to be restricted within the domain of the data. The resulting truncated probability density function ($g^*(x)$) is then specified within the (exogenously determined) boundaries $x \in \left[c^l, c^u\right]$:

$$g^*(x) = \frac{g(x)}{G(c^u) - G(c^l)}. \tag{6}$$

3.1.3 FMM

Complete data Direct maximum likelihood estimation of a FMM (see eq. 2) is not straightforward, since the number of components I is a priori unknown. The log-likelihood function can be written as

$$logL(x|\boldsymbol{\Psi}) = \sum_{b=1}^{B}\sum_{i=1}^{I} z_{bi}\left[log(\pi_i) + log(m_i(x_b|\boldsymbol{\theta}_i))\right], \tag{7}$$

where z_{bi} is an unobserved component indicator equal to one if the observation x_b originates from subpopulation i and zero otherwise. Two steps need to be taken iteratively in order to be able to maximize this equation. The Expectation (E)-step of the s-th iteration consists of determining the conditional expectation of eq. 7 given the observed data and the current parameter estimates from iteration $s-1$:

$$\begin{aligned} Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(s-1)}) &= E\left[logL(x|\boldsymbol{\Psi})|x, \boldsymbol{\Psi}^{(s-1)}\right] \\ &= \sum_{b=1}^{B}\sum_{i=1}^{I} \pi_{bi}^{(s)}\left[log(\pi_i) + log(m_i(x_b|\boldsymbol{\theta}_i))\right], \end{aligned} \tag{8}$$

where the missing data z_{ni} is replaced by the posterior probability that x_b belongs to the ith mixture:

$$\pi_{bi}^{(s)} = E\left[z_{bi}|x_b, \boldsymbol{\Psi}^{(s-1)}\right] = \frac{\pi_i^{(s-1)} m_i(x_b|\boldsymbol{\theta}_i^{(s-1)})}{\sum_{i=1}^{I}\pi_i^{(s-1)} m_i(x_b|\boldsymbol{\theta}_i^{(s-1)})}. \tag{9}$$

The Maximization (M)-step then, consists of maximizing the Q-function over the parameter vector $\boldsymbol{\Psi}$:

$$\boldsymbol{\Psi}^{(s)} = \max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(s-1)}). \tag{10}$$

Each iteration updates the E- and M-step until the algorithm converges (See Miljkovic and Grün (2016) and McLachlan and Peel (2000) for a more elaborate overview).

The validity of the proposed estimation technique does not depend on its ability to identify the unobserved component indicator z_{bi}. FMMs can be utilized in two ways. First, they can be used as a semi-parametric, flexible approximation of the overall distribution. Second, they are model-based clustering methods when a certain distribution is imposed (Fop et al., 2018; Grün, 2018). While both applications rely on the idea that discrete subpopulations define the overall distribution, the semi-parametric approximation does not claim to correctly identify these subpopulations (z_{bi}). This paper relies on FMMs as a semi-parametric approximation of the productivity distribution. See Appendix B for a more elaborate discussion on the difference between both applications and their relevance for the current analysis.

Truncated data The EM-algorithm can be adapted to fitting data only to truncated data within the (exogenously determined) boundaries $x \in \left[c^l, c^u\right]$. We specify the conditional densities

$$\begin{aligned} g(x|\boldsymbol{\Psi}, c^l \leq x \leq c^u) &= \frac{\sum_{i=1}^{I} \pi_i m_i(x|\boldsymbol{\theta_i})}{G(c^u|\boldsymbol{\Psi}) - G(c^l|\boldsymbol{\Psi})} \\ &= \sum_{i=1}^{I} \pi_i \frac{M_i(c^u|\boldsymbol{\theta_i}) - M_i(c^l|\boldsymbol{\theta_i})}{G(c^u|\boldsymbol{\Psi}) - G(c^l|\boldsymbol{\Psi})} \frac{m_i(x|\boldsymbol{\theta_i})}{M_i(c^u|\boldsymbol{\theta_i}) - M_i(c^l|\boldsymbol{\theta_i})} \\ &= \sum_{i=1}^{I} \eta_i m_i(x|\boldsymbol{\theta_i}, c^l \leq x \leq c^u), \end{aligned} \tag{11}$$

with $\eta_i > 0$, $\sum_{i=1}^{I} \eta_i = 1$ and M_i the component-specific Cumulative Distribution Function. The Q-function becomes

$$\begin{aligned} Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(s-1)}) &= E\left[logL(x|\boldsymbol{\Psi})|x, \boldsymbol{\Psi}^{(s-1)}\right] \\ &= \sum_{b=1}^{B} \sum_{i=1}^{I} \pi_{bi}^{(s)} \left[log(\eta_i) + log(m_i(x_b|\boldsymbol{\theta}_i, c^l \leq x_b \leq c^u))\right], \end{aligned} \tag{12}$$

where the posterior probability that x_b comes from the ith mixture is not affected by the truncation:

$$\pi_{bi}^{(s)} = \frac{\eta_i^{(s-1)} m_i(x_b|\boldsymbol{\theta}_i^{(s-1)}, c^l \leq x_b \leq c^u))}{\sum_{i=1}^{I} \eta_i^{(s-1)} m_i(x_b|\boldsymbol{\theta}_i^{(s-1)}), c^l \leq x_b \leq c^u)} = \frac{\pi_i^{(s-1)} m_i(x_b|\boldsymbol{\theta}_i^{(s-1)})}{\sum_{i=1}^{I} \pi_i^{(s-1)} m_i(x_b|\boldsymbol{\theta}_i^{(s-1)})}. \tag{13}$$

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

3.2 Distribution evaluation

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

Goodness of fit We follow Dewitte (2020) in evaluating the parametric distributions by summarizing the distance between the empirical and parametric rth moment of the distribution by the 1- and ∞-norm:

$$S^r = \sum_{y} \Delta^r(y), \qquad T^r = \sup_{y} \Delta^r(y), \tag{14}$$

where $\Delta^r(y)$ is the normalized absolute deviation:

$$\Delta^r(y) = \frac{\left|\frac{1}{B}\sum_{b=1}^{B}\mathbb{I}(x_b \geq y)x_b^r - \int_y^\infty x^r g(x|\mathbf{\Psi})dx\right|}{\frac{1}{B}\sum_{b=1}^{B} x_b^r}. \tag{15}$$

$\mathbb{I}(A)$ is the indicator of event A and $\mu_y^r = \int_y^\infty x^r g(x|\mathbf{\Psi})dx$ is the y-bounded, rth-moment of the parametric distribution, with r taking positive values. Evaluated at the 0th-moment of the distribution, the test statistic T^0 corresponds with the Kolmogorov-Smirnov (KS) test statistic, quantifying the largest distance between the empirical and parametric CDF. This is the sole specification of the statistic specified on which we can rely to provide statistically underpinned claims regarding the accuracy of the distributional assumption with respect to its empirical counterpart. Nevertheless, Dewitte (2020) argues that evaluating these test statistics at higher moments of the distribution ($r > 0$) can be informative on the distributional fit, especially relating to their use in heterogeneous firms models (see also section 6).[17] Whereas the ∞-norm contains only information on the largest distance, the 1-norm provides information on the distance between both distributions over the complete distributional space, weighting all distances equally. The normalization factor allows us to interpret the distances on a scale of zero to one for all moments, similar to the interpretation of the standard KS test statistic.

As we rely on estimated parameters, asymptotic distributions are not available for the test statistics. We therefore rely on a parametric bootstrap:

1. Assume B i.i.d. random variables with distribution $G(\cdot|\mathbf{\Psi})$;
2. Estimate the parameters $\mathbf{\Psi}$ of the distribution using MLE and calculate the rth moment implied by the parametric distribution: $\hat{\mu}^r$;
3. $H_0 : \mu^r = \hat{\mu}^r$ with test statistic $t \in \{S^r, T^r\}$;
4. Draw N bootstrap samples of size B from $G(\cdot|\hat{\mathbf{\Psi}})$;
5. For each sample of the parametric distribution, calculate the bootstrapped test statistics $t^* \in \left\{(S^{\tilde{r}})^*, (T^{\tilde{r}})^*\right\}$;[18]
6. The p-value is then defined as

$$\hat{p} = \frac{1}{N+1}\left[\sum_{n=1}^{N}\mathbb{I}\left(t_n^* \geq t\right) + 1\right]. \tag{16}$$

The bootstrap exercise should therefore be interpreted as 'the likelihood of observing a deviation between the moments of the empirical and parametric distribution as large as t under the null

[17] We have no knowledge of statistical tests that evaluate distributional fits based on bounded higher moments of the distribution.

[18] Note that we do not re-fit the parametric distribution to the bootstrap sample. The vastness of the dataset at our availability in the empirical section results both in a large computational burden but also a very precise estimation of the distribution parameters. The influence of not refitting the parametric distribution to the bootstrap sample is therefore negligent.

hypothesis', allowing us to evaluate whether the distributional assumption provides a good fit to the evaluated moments of the distribution.

Information Criteria We differentiate between distributions based on the log-likelihood, the Aikaike or Bayesian Information Criteria. When possible, we can differentiate between two distributions based on the ratio of their likelihoods:

$$LR = \sum_{b=1}^{B} ln \frac{g_1(x_b; \cdot)}{g_2(x_b, \cdot)} \tag{17}$$

with $g_{1,2}$ the probability densities of the respective distributions. If these distributions are non-nested (Vuong, 1989), the test statistic amounts to 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 and Leibler (1951) 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 $-\infty$ (Vuong, 1989).

The Aikaike Information criterion penalizes the log-likelihood information for the number of parameters (to avoid overfitting) and is defined as $AIC = 2np - 2ln(L)$ with np the number of parameters and $ln(L)$ the log-likelihood. Similarly, the Bayesian Information criterion corrects for the number of parameters as $BIC = npln(B) - 2ln(L)$. Differentiation between distributions relies then on te relative distance of the BICs: $\Delta BIC = BIC_1 - BIC_2$. The value of ΔBIC 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). AIC and BIC statistics are considered adequate when choosing the number of components for a suitable FMM (McLachlan and Peel, 2000).

4 Data

We use firm-level data from Portugal to evaluate the empirical performance of FMMs compared to "traditional" distributions such as the Log-normal or Pareto distribution. The main source of information is Sistema de Contas Integradas das Empresas (SCIE, Enterprise Integrated Accounts System) in the year 2006, a dataset covering the universe of active Portuguese firms that has been used already by, among others: (Carreira and Teixeira, 2016; Dias et al., 2016; Fernandes and Ferreira, 2017; Bastos et al., 2018; Fonseca et al., 2018).[19] It contains data both on firm-level sales and number of employees. Moreover, each firm has a unique identification number that allows us to link this dataset with a dataset on international trade.

The firm size distribution of Portugal was earlier the object of study by Cabral and Mata (2003), who relied on a longitudinal matched employer-employee dataset covering all business units with

[19] A comparison between SCIE and the OECD SBDS database proves the full coverage of firms in our dataset for the Portuguese economy (see Table 6).

at least one wage earner in the Portuguese economy (Quadros de Pessoal). They provide evidence that the firm size distribution of Portugal is not very different from other countries such as France, the United States, Germany, Japan and the United Kingdom.

We rely on the distributional relation between productivity and positive domestic sales, under specific model assumptions (Nigai, 2017; Dewitte, 2020), to evaluate parametric approximations of the productivity distribution. Relying on domestic rather than total sales corrects for the impact of international trade on the firm size distribution (di Giovanni et al., 2011). We reduce our dataset discarding self-employed companies[20], resulting in a dataset covering the positive domestic sales of 299,935 Portuguese firms in 2006.

5 Results

We fit the distributions to Portuguese domestic sales in the year 2006. We initially focus on fitting the Pareto, Lognormal, combinations of Pareto and Lognormal and up to 5-component mixtures of Lognormals to the complete data. This proves to be sufficient for our main message. We show that our results hold when focusing on the tails of the data, can be extended to other economically relevant distributions, are robust to sample selection and outliers and can be externally validated on city size data.

5.1 Complete data

Single distributions can not sufficiently capture the heterogeneity of the productivity distribution. Table 1 displays the selected distribution fits, ordered according to their log-likelihood. One immediately observe that single parametric distributions provide the worst fits. This demonstrates the need, as the evolution of the literature indicates (Nigai, 2017; Sager and Timoshenko, 2019), to combine distributions in order to adequately capture heterogeneity in productivity. The Pareto distribution, for instance, provides a really bad fit to the distribution with a Goodness of fit statistic of up to 267 times bigger than the best fitting mixture of Lognormals.[21].

Finite mixture models greatly improve the distributional fit, without over-fitting the data. According to the log-likelihood, distributions with a larger number of parameters provide a better fit to the data, even when parameter correction ($R_{AIC,BIC}$) is applied. The BIC values indicate that the 4-component Lognormal provides the best fit to the data. This demonstrates that the performance of FMMs is not the result of over-fitting, but of FMMs being able to capture heterogeneity of which other distributional forms are not capable. The currently favored Double-Pareto Lognormal (Sager and Timoshenko, 2019) and Lognormal-Pareto (Nigai, 2017) distribution are ranked fourth and eighth respectively. The structure imposed on a general mixture specification in order to attain these specific piecewise composite or product distributions (see section 2.2) is therefore not

[20] Disregarding individual companies renders our dataset more comparable with earlier datasets used to evaluate productivity distributions such as the ORBIS database used by Nigai (2017).

[21] The higher the Goodness of fit statistic, the larger the deviation between the empirical and parametric distribution (see eq. 15)

Table 1: Selected distribution fits to Portuguese domestic sales in 2006.

		Goodness of fit		Information Criteria		
Distribution	Parms.	T_a^0	S_b^0	Loglike	R_{AIC}	R_{BIC}
5-comp. Lognormal	14	0.18	0.11	12,776	1	2^{+++}
		(0.10;0.25)	(0.08;0.32)			
4-comp. Lognormal	11	0.19	0.11	12,770	2	1
		(0.09;0.25)	(0.08;0.32)			
3-comp. Lognormal	8	0.29	0.34	12,723	3	3^{+++}
		(0.10;0.24)**	(0.09;0.32)**			
Double-Pareto Lognormal	4	0.66	0.80	12,429	4	4^{+++}
		(0.09;0.25)***	(0.08;0.33)***			
2-comp. Lognormal	5	0.53	0.71	12,401	5	5^{+++}
		(0.10;0.24)***	(0.09;0.32)***			
Inv. Pareto-Lognormal-Pareto	4	0.81	1.01	12,231	6	6^{+++}
		(0.09;0.26)***	(0.08;0.34)***			
Inv. Pareto-Lognormal	3	3.02	4.26	9,198	7	7^{+++}
		(0.09;0.24)***	(0.08;0.31)***			
Lognormal-Pareto	3	2.56	3.78	8,721	8	8^{+++}
		(0.09;0.25)***	(0.08;0.32)***			
Left-Pareto Lognormal	3	3.23	4.91	8,059	9	9^{+++}
		(0.10;0.25)***	(0.09;0.32)***			
Right-Pareto Lognormal	3	2.82	4.38	8,028	10	10^{+++}
		(0.09;0.25)***	(0.08;0.32)***			
Lognormal	2	2.93	5.03	7,372	11	11^{+++}
		(0.10;0.25)***	(0.08;0.33)***			
Pareto	2	48.34	68.18	-436,227	12	12^{+++}
		(0.09;0.25)***	(0.08;0.33)***			

Notes: All distributions fitted using Maximum Likelihood.
Values between parentheses report the 5th and 95th quantile of the parametric bootstrapped test statistic with 999 replications. ***, **, * indicate significance of this test at 1%, 5% and 10% respectively.
$^{+++}$, $^{++}$, $^{+}$ indicates the difference between this distribution's BIC and the first-ranked distribution in terms of BIC (ΔBIC) providing strong evidence in favour of the first-ranked distribution ($\Delta BIC > 10$), moderate evidence ($6 < \Delta BIC \leq 10$) and weak evidence ($2 < \Delta BIC \leq 6$) respectively.
$_a$ Values multiplied by 100 for expositional purpose, $_b$ Values divided by 1,000 for expositional purpose.

warranted.

Finite mixture models are the sole parametric specifications that are not rejected by the data. Focusing on goodness-of-fit criteria around the 0th moment, we observe that these follow the log-likelihood ranking closely. The 4- and 5-component Lognormal distributions reduce the largest deviation from the empirical CDF (T^0) by more than 70% ($\frac{0.66-0.19}{0.66} \times 100$) compared to the Double-Pareto Lognormal distribution and by more than 90% compared to the Lognormal-Pareto distribution. This pattern is consistent over the complete range of the data, as is apparent from the cumulative error of the CDF fit (S^0). Moreover, none of the currently favored parametric distributions provide a good fit to the data. Only for the 4- and 5-component Lognormal distributions the null hypothesis that the data originates from the proposed parametric distribution can not be rejected.

Figure 2[22] provides a visual insight into the numerical results of Table 1. It plots the normalized absolute deviation between the empirical and parametric CDF. The figure shows the large errors related to the Lognormal distribution. Augmenting the Lognormal distribution with a Pareto right-tail as in Nigai (2017) improves the fit only marginally. While it does provide 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 equally bad fit to the bulk of the distribution as the Lognormal 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 4-component Lognormal, which only displays marginal errors both in the bulk and the tails of the data. This tail performance becomes even more apparent when considering the Quantile-Quantile plot in Figure 3.

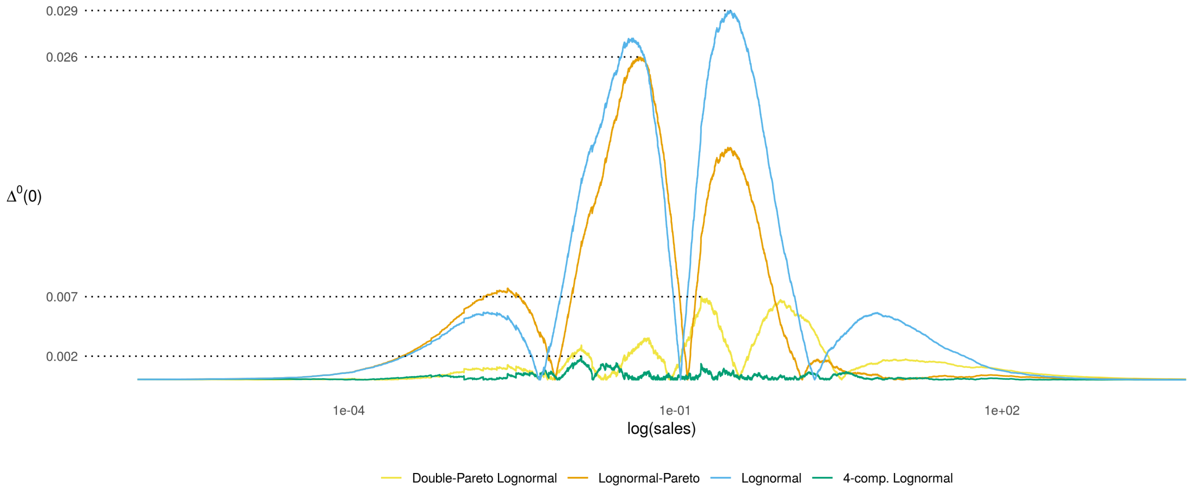

Figure 2: Normalized Absolute Deviation between the empirical and Double-Pareto Lognormal, Lognormal-Pareto, Lognormal and 4-component Lognormal CDFs over the complete range of domestic sales in Portugal, 2006.

[22] This representation of the results is essentially a visually more interpretable version of the Probability-Probability plot (see Appendix Figure 3).

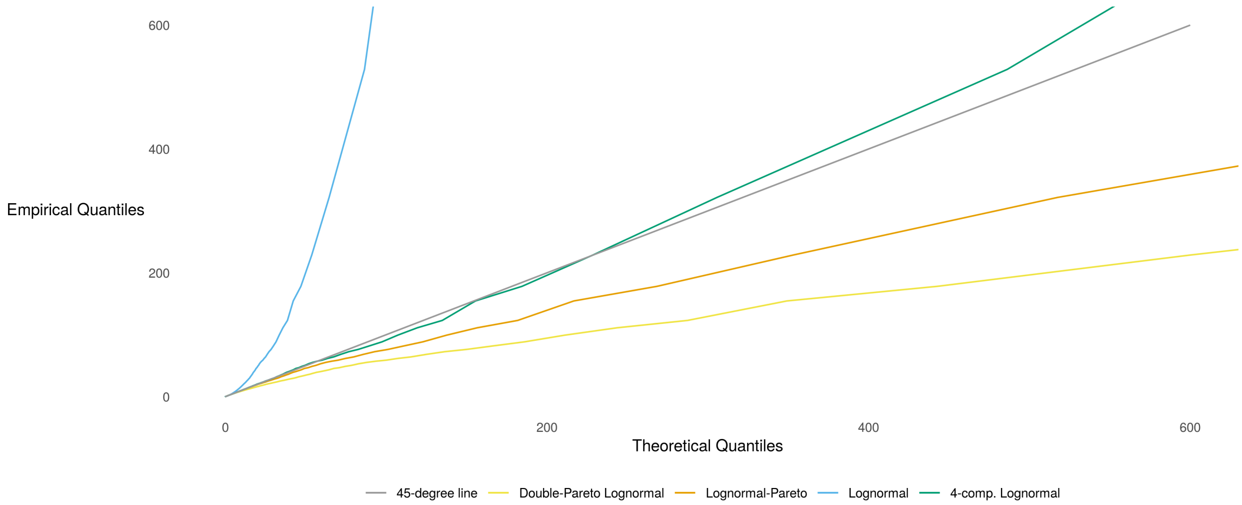

Figure 3: Quantile-Quantile plot for the Double-Pareto Lognormal, Lognormal-Pareto, Lognormal and 4-component Lognormal over approximately 99.99% of domestic sales in Portugal, 2006.
Note: Quantiles are capped at 600 for expositional purposes, leaving out approximately the upper 0.01% of the data.

5.2 Truncated data

Allowing for heterogeneity in distributions clearly provides a better fit when fitting the complete distribution, but what about when we fit the tails only? This is mostly interesting from the Pareto point of view, which is often claimed to be a good fit to the right tail of the productivity distribution.[23]

Table 2 displays the results of fitting the (Inverse) Pareto to the (left) right tail of the distribution using the methods described in section 3. We recovered the best-fitting truncation point for the (Inverse) Pareto distribution, assigning 8.53% and 6.07% of the data to the left and right tail respectively. We reduced our dataset according to these truncation parameters and fitted truncated mixtures of Lognormals to both tails of the distribution for comparison. This approach puts the Pareto distribution twice in the advantage. First, it is free from a parametric specification for the bulk of the distribution. Second, the truncation parameter is chosen in function of the best-fitting (Inverse) Pareto distribution. As a result, the (Inverse) Pareto, as well as (mixtures of) the Lognormal, provide a good fit to the tails according to the Kolmogorov-Smirnov test.

Nevertheless, despite the advantage for the (Inverse) Pareto distribution, it seems that (mixtures of) the Lognormal distribution provide a significantly better fit to the tails of the data. (Mixtures of) the Lognormal distribution have a higher log-likelihood and lower deviation from the empirical CDF than the (Inverse) Pareto distribution. This results in the likelihood ratio test significantly rejecting Pareto in favor of (mixtures of) the Lognormal distribution, which is in line with earlier results reported in related literature (Clauset et al., 2009). When correcting for the number of parameters, the BIC reveals that the single Lognormal distribution is sufficient to fit the tail only.

[23] Note that this argument carries the normative value that obtaining a good fit for larger firms is absolute, regardless of the implications for the fit to smaller firms.

A mixture of Lognormals insufficiently improves the fit in order to justify the corresponding increase in number of parameters.

Table 2: Selected distribution fits to the tails of Portuguese domestic sales in 2006.

		Goodness of fit		Information Criteria		
Distribution	Parms.	T^0_a	S^0_b	Loglike	R_{AIC}	R_{BIC}
		Left tail (N=25,588, 8.53% of the data)				
5-comp. Trunc. Lognormal	14	0.63	0.04	108,196.19***	5	6^{+++}
		(0.32;0.85)	(0.02;0.10)			
4-comp. Trunc. Lognormal	11	0.61	0.04	108,195.05***	4	5^{+++}
		(0.33;0.85)	(0.02;0.09)			
3-comp. Trunc. Lognormal	8	0.58	0.04	108,194.44***	1	4^{+++}
		(0.33;0.86)	(0.02;0.10)			
2-comp. Trunc. Lognormal	5	0.77	0.06	108,189.93***	3	3^{+++}
		(0.32;0.84)*	(0.02;0.09)			
Trunc. Lognormal	2	1.02	0.10	108,186.99***	2	1
		(0.32;0.85)**	(0.02;0.09)**			
Inv. Pareto	2	0.80	0.10	108,183.90	6	2^{++}
		(0.33;0.84)*	(0.02;0.09)**			
		Right tail (N=18,217, 6.07% of the data)				
5-comp. Trunc. Lognormal	14	0.62	0.03	-47,896.59***	5	6^{+++}
		(0.39;1.00)	(0.02;0.07)			
Trunc. Lognormal	2	0.70	0.04	-47,897.86***	1	1
		(0.38;0.97)	(0.02;0.08)			
2-comp. Trunc. Lognormal	5	0.71	0.04	-47,897.99***	2	3^{+++}
		(0.38;1.01)	(0.02;0.08)			
3-comp. Trunc. Lognormal	8	0.68	0.04	-47,898.60***	3	4^{+++}
		(0.38;0.99)	(0.02;0.08)			
4-comp. Trunc. Lognormal	11	0.68	0.04	-47,898.62***	4	5^{+++}
		(0.39;1.00)	(0.02;0.08)			
Pareto	2	0.86	0.08	-47,910.44	6	2^{+++}
		(0.38;0.99)	(0.02;0.08)*			

Notes: All distributions fitted using Maximum Likelihood.
Values between parentheses report the 5th and 95th quantile of the parametric bootstrapped test statistic with 999 replications. ***, **, * indicate significance of this test at 1%, 5% and 10% respectively.
Similarly, ***, **, * indicate significance at 1%, 5% and 10% respectively for the likelihood ratio test between (Inverse) Pareto and (mixtures of) the Lognormal distribution.
$^{+++}$, $^{++}$, $^{+}$ indicates the difference between this distribution's BIC and the first-ranked distribution in terms of BIC (ΔBIC) providing strong evidence in favour of the first-ranked distribution ($\Delta BIC > 10$), moderate evidence ($6 < \Delta BIC \leq 10$) and weak evidence ($2 < \Delta BIC \leq 6$) respectively.
$_a$ Values multiplied by 100 for expositional purpose, $_b$ Values divided by 1,000 for expositional purpose.

5.3 Extension to other distributions

The superior performance of FMMs is not limited to the Lognormal distribution. Appendix Table 7 displays the 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. Most of these mixtures are not able to match the performance of the Lognormal. Only the Burr distribution provides an equivalent fit to the PDF and CDF.[24] Compared to Pareto-tailed combinations of distributions, we find that also mixtures of Weibull and Gamma are able to provide an improved distribution fit. Overall, the currently favored Double-Pareto Lognormal (Sager and Timoshenko, 2019) and Lognormal-Pareto (Nigai, 2017) distribution are ranked sixteenth and thirty-first respectively according to BIC, out of 52 considered distributions.

The consistent excellent performance of the Lognormal distribution can be motivated from two perspectives. From the perspective of overall fit, a mixture of (log-) normal distributions with sufficient components is assumed to be able to approach all distributions (McLachlan and Peel, 2000). From a generative perspective for individual components, the Lognormal distribution is the realization of applying the Central Limit Theorem (CLT) in the log domain: firm heterogeneity will approximately be Lognormal if it is the multiplicative product of many independent random variables. This corresponds with extensions of heterogeneous firms models à la Melitz (2003) that consider multi-dimensional firm heterogeneity, taking into consideration the product dimension (Bernard et al., 2009) or uncertainty in demand and/or supply (see for instance De Loecker (2011); Bas et al. (2017); Sager and Timoshenko (2019); Gandhi et al. (0)).

5.4 Robustness

We scrutinize the robustness of our results with a number of additional analyses. First, we examine whether our results are not caused by sample selection. We therefore restrict our dataset to the manufacturing sector only (see Appendix Table 8) and find the performance of FMMs to improve relative to Pareto-tailed distributions. Second, we inspect whether our results are not due to outliers in the tails of the distribution. We discard the first and last 1,000 observations of the dataset. Results in Appendix Table 9 again confirm the superiority of FMMs.

We validate our approach externally fitting the considered distributions to the U.S. Census 2000 city size distribution data. This dataset has been subject to an extensive debate in the city size literature, including the discussion between Eeckhout (2004, 2009) and Levy (2009).[25] Appendix Table 10 provides the test results, demonstrating that the city size distribution is neither Lognormal, Pareto nor Pareto-tailed Lognormal. It is best approximated by a 2-component Lognormal distribution (according to the BIC). These results provide an overview of the city size literature up till now and are in line with the findings of Kwong and Nadarajah (2019).

6 Gains From Trade implications

As stated in the introduction, the Gains From Trade literature is an area where finding a good parametric approximation of the productivity distributions is of critical importance. In this section, we integrate the distribution fits from the previous section into a heterogeneous firms framework à la

[24] The Burr distribution fails to match higher moments of the data, however. See also section 6.

[25] The dataset is available at https://www.aeaweb.org/aer/data/sept09/20071478_data.zip.

Melitz (2003). This allows us to perform a GFT exercise along the lines of (Melitz and Redding, 2015; Bee and Schiavo, 2018) and investigate the importance of providing a good fit to the productivity distribution.

Our setup is a two-country symmetric heterogeneous firms model with a finite number of firms.[26] The parameterization of our model is standard (Head et al., 2014; Melitz and Redding, 2015; Bee and Schiavo, 2018). We work with two symmetric countries i and j and choose labor in one country as the numeraire, so that $W^i = W^j = 1$. We choose fixed entry costs $f^e = 0.545$ and set fixed costs equal to one ($f^{ii} = f^{ij} = 1$). The elasticity of substitution is set to four.

Finally, we need to capture the heterogeneity distribution. Assuming a parametric distribution and under the assumption of an *infinite* number of firms, we can calculate the necessary analytical expressions using the distributional parameters from our empirical analysis to capture heterogeneity. Following Nigai (2017), we can also capture heterogeneity directly from the empirical, *finite*, data. To allow comparison between GFT obtained assuming a parametric distribution and GFT obtained from the finite data, we perform a parametric bootstrap. This parametric bootstrap generates a range of finite sample estimates under the hypothesis that the observed data is generated by a certain parametric distribution, which can be compared with the observed finite data (Dewitte, 2020).

We calculate the changes in welfare due to a trade shock (Gains From Trade), which can be written as log changes in real per-capita income due to an exogenous increase in variable trade costs τ_{ij} to τ'_{ij}. This can be further decomposed into the channels through which trade affects welfare: trade costs (τ^{ij}), the number of firms (M^i), the probality of successful entry into the domestic market ($m^0_{\omega^{ii*}}$), the average productivity of firms exporting from i to j ($m^{\sigma-1}_{\omega^{ij*}}$)[27] and the bilateral trade share (λ^{ij}):

$$\begin{aligned} 100 \times ln\frac{(\mathbb{W}^i)'}{\mathbb{W}^i} &= 100 \times -ln\frac{(P^i)'}{P^i} \qquad (18) \\ &= 100 \times -\left[ln\frac{(\tau^{ij})'}{(\tau^{ij})} - \frac{1}{\sigma-1}\left(ln\frac{(M^i)'}{M^i} - ln\frac{(m^0_{\omega^{ii*}})'}{m^0_{\omega^{ii*}}} + ln\frac{(m^{\sigma-1}_{\omega^{ij*}})'}{m^{\sigma-1}_{\omega^{ij*}}} - ln\frac{(\lambda^{ij})'}{\lambda^{ij}} \right)\right]. \end{aligned}$$

Our exercise reduces the variable trade costs from $\tau^{ij} = 3$ to $(\tau^{ij})' = 1$. The obtained GFT are displayed in Figure 4. This figure presents the parametric bootstrapped distribution of GFT by means of box-plots delineating the 5th, 25th, 50th, 75th and 95th quantile. Empirical GFT are indicated by the vertical blue line. Green circles are the average parametric finite sample GFT and the parametric plug-in population estimates of GFT are shown by yellow diamonds.

We observe that heavy-(Pareto-) tailed distributions significantly overestimate GFT, while relatively light-tailed distributions underestimate GFT. Mixture models are the only distributions that provide an approximation of GFT that is not rejected by the data. The distributions in Figure

[26] See Appendix C for a full workout of the model.

[27] We define average productivity here as average productivity unconditional on successful entry, in contrast to the definition conditional on successful entry in (Melitz, 2003, p.1702).

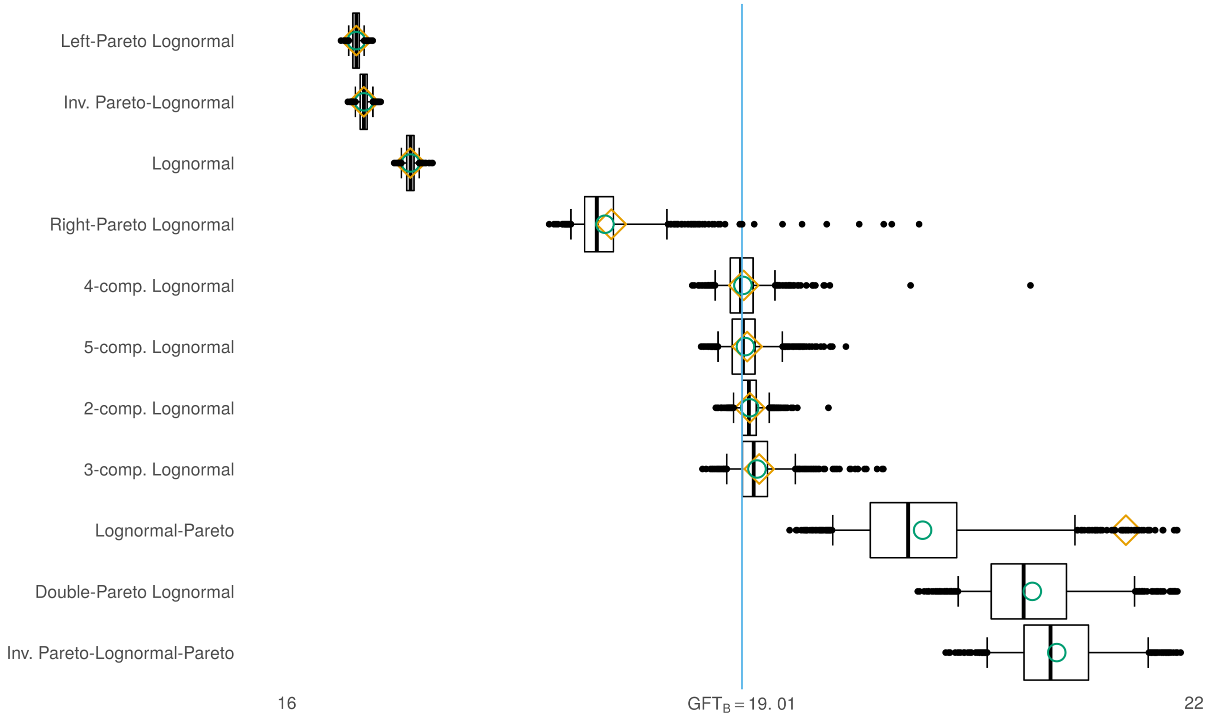

Figure 4: Gains from a reduction in variable trade costs $\tau^{ij} = 3$ to $(\tau^{ij})' = 1$.

Notes: Box-plots display the 5th, 25th, 50th, 75th and 95th quantile of the asymptotic distribution of parametric finite sample GFT obtained from a bootstrap with 999 replications. Yellow diamonds represent the parametric plug-in (population) estimates of GFT. Green circles are the average parametric bootstrapped finite sample GFT and the empirical sample GFT are indicated by the vertical blue line. All sample values obtained from a sample of 299,935 firms.

4 are ordered according to their distance from the empirical GFT. As such, we can interpret the 4-component Lognormal distribution as providing the closest fit to the GFT obtained from the empirical distribution. Where the empirical values imply an increase in real income per capita of 19.01% when reducing variable trade costs from 3 to 1, the 4-component Lognormal distribution closely predicts this to be 19.02%, as can be deduced from the parametric plug-in population estimates (yellow diamonds). Moreover, the close fit results in a very good approximation of the empirical GFT, as can be deduced from the parametric bootstrapped finite sample GFT being at least as small as the empirical GFT in more than 5% of the cases (the box-plot overlaps with the vertical blue line). This contrasts with the simple Lognormal distribution underestimating the empirical GFT by about 11% with 16.8% predicted GFT, and with the Lognormal-Pareto distribution overestimating the empirical GFT by approximately 13%, with 21.55% predicted GFT.

Deviations from GFT calculations can be mainly attributed to errors in capturing the evolution of average productivity of exporting firms and bilateral trade shares. Table 3 reports the weighted components of welfare gains (see eq. 18) for all considered distributional forms, allowing us to evaluate the channels trough which the differences in GFT between distributions come about. We observe that the deviation of the parametric results compared to the empirical distribution are relatively small for the changes in number of firms and in the probability of successful entry into the domestic market. The largest differences can be found for the changes in average productivity of exporting firms and in the trade shares. Heavy-tailed distributions largely underestimate the positive effect of the increase in average productivity of exporting firms and the negative effect of the increase in the bilateral trade shares compared to the empirical distribution, while the reverse is true for lighter-tailed distributions.

Table 3: Decomposition of procentual welfare gains from a reduction in variable trade costs $\tau^{ij} = 3 \rightarrow (\tau^{ij})' = 1$.

Distribution	Parms.	$ln\frac{(\mathbb{W}^i)'}{\mathbb{W}^i}$	$-\ln\frac{(\tau^{ij})'}{(\tau^{ij})}$	$\frac{1}{\sigma-1}ln\frac{(M^i)'}{M^i}$	$\frac{1}{\sigma-1}ln\frac{(m^0_{\omega ii*})'}{m^0_{\omega ii*}}$	$\frac{1}{\sigma-1}ln\frac{(m^{\sigma-1}_{\omega ij*})'}{m^{\sigma-1}_{\omega ij*}}$	$-\frac{1}{\sigma-1}ln\frac{(\lambda^{ij})'}{\lambda^{ij}}$
Pareto	2	-	1.10	-	-	-	-
		(-0.00;0.00)***	(1.10;1.10)	(-0.22;-0.22)***	(-0.00;0.00)***	(0.00;0.00)***	(-0.88;-0.88)***
Left-Pareto Lognormal	3	0.16	1.10	-0.17	0.15	0.60	-1.51
		(0.16;0.17)***	(1.10;1.10)	(-0.17;-0.17)***	(0.15;0.15)***	(0.58;0.62)***	(-1.53;-1.49)***
Inv. Pareto-Lognormal	3	0.17	1.10	-0.17	0.15	0.58	-1.49
		(0.16;0.17)***	(1.10;1.10)	(-0.17;-0.17)***	(0.15;0.15)***	(0.56;0.60)***	(-1.51;-1.47)***
Lognormal	2	0.17	1.10	-0.17	0.15	0.53	-1.44
		(0.17;0.17)***	(1.10;1.10)	(-0.17;-0.17)***	(0.15;0.15)***	(0.51;0.55)***	(-1.46;-1.42)***
Right-Pareto Lognormal	3	0.18	1.10	-0.18	0.17	0.28	-1.19
		(0.18;0.19)**	(1.10;1.10)	(-0.19;-0.18)	(0.17;0.18)	(0.23;0.33)**	(-1.24;-1.13)**
Empirical	0	0.19	1.10	-0.18	0.18	0.20	-1.10
4-comp. Lognormal	11	0.19	1.10	-0.18	0.18	0.20	-1.10
		(0.19;0.19)	(1.10;1.10)	(-0.19;-0.18)	(0.17;0.18)	(0.18;0.22)	(-1.13;-1.08)
5-comp. Lognormal	14	0.19	1.10	-0.19	0.18	0.20	-1.10
		(0.19;0.19)	(1.10;1.10)	(-0.19;-0.18)	(0.17;0.19)	(0.17;0.22)	(-1.12;-1.07)
2-comp. Lognormal	5	0.19	1.10	-0.17	0.17	0.23	-1.13
		(0.19;0.19)	(1.10;1.10)	(-0.18;-0.17)***	(0.16;0.17)***	(0.22;0.25)***	(-1.15;-1.12)***
3-comp. Lognormal	8	0.19	1.10	-0.18	0.18	0.19	-1.09
		(0.19;0.19)	(1.10;1.10)	(-0.19;-0.18)	(0.17;0.18)	(0.16;0.22)	(-1.12;-1.06)
Lognormal-Pareto	3	0.22	1.10	-0.22	0.22	0.02	-0.90
		(0.20;0.21)***	(1.10;1.10)	(-0.22;-0.20)***	(0.20;0.22)***	(0.04;0.14)***	(-1.04;-0.93)***
Double-Pareto Lognormal	4	-	1.10	-	-	-	-
		(0.20;0.22)***	(1.10;1.10)	(-0.20;-0.19)***	(0.19;0.20)***	(0.02;0.09)***	(-0.98;-0.90)***
Inv. Pareto-Lognormal-Pareto	4	-	1.10	-	-	-	-
		(0.21;0.22)***	(1.10;1.10)	(-0.20;-0.18)*	(0.18;0.20)***	(0.01;0.08)***	(-0.97;-0.89)***

Notes: Values between parentheses report the 5th and 95th quantile of the parametric bootstrapped statistics with 999 replications. ***, **, * indicate the rejection of a signifcant overlap of the parametric bootstrapped statistic with the empirical statistic at 1%, 5% and 10% respectively.

Our results confirm the findings of Dewitte (2020) that a good fit to truncated average sales proves to be a critical predictor of the performance of GFT calculations. A ranking of the distributions according to GFT performance does not closely follow the ranking of the fit to the 0th moment of the distribution (the CDF). The Double-Pareto Lognormal, for instance, provides a closer fit to the empirical CDF than the Right-Pareto Lognormal, but provides worse GFT approximations. This can be attributed to the relatively heavy tail of the Double-Pareto Lognormal, resulting in a large error when calculating higher moments of the distribution. As such, a ranking of distributions based on the fit to average sales proves to be a better indicator of GFT performance, as can be deduced from the statistics T^1 in Appendix Table 7.

These findings are not the result of a specific parametrization of the model. Figure 4 displays the percentage errors in parametric GFT calculations relative to the empirical benchmark for different parametrization scenarios. Our findings are robust for different values of the elasticity of substitution (left upper panel) and fixed entry costs (left bottom panel), as well as for different starting values for the iceberg trade costs (right upper panel) and for a reduction in fixed rather than variable trade costs (right bottom panel).

7 Conclusion

This paper provides evidence that heterogeneity in the productivity distribution can be captured most adequately by Finite Mixture Models. A clear statistical framework differentiates between the fit of 52 distributions to domestic sales of the population of active Portuguese firms in 2006. The flexible, semi-parametric nature of FMMs results in a substantial empirical performance improvement compared to currently favored distributions in the firm size literature. Moreover, FMMs are the only distributions providing an approximation of Gains From Trade that is not rejected by the data.

Even though our results provide strong evidence in favor of FMMs, we take no stance on distribution type nor on the mixing parameter (or mechanism) that defines the underlying discrete subpopulations. It is clear that the two are closely interconnected, and therefore not easily identifiable. Further research is necessary to be able to define which mechanisms result in multiple individual densities defining the overall productivity distribution.

The idea of FMMs also opens many new venues for ongoing research. For instance, the estimation of productivity usually relies on a first-order Markov process that is identical for the complete population. Concurrently, however, it is recognized that productivity dynamics are endogenous to exporting (De Loecker, 2013), importing (Kasahara and Rodrigue, 2008), innovation (Aw et al., 2011), management practices (Bloom and Reenen, 2011; Caliendo et al., 2020), et cetera. Introducing Finite Mixture Modeling into the estimation procedures would allow, semi-parametrically, to control for such discrete subpopulations without the risk of model misspecification. This would be in line with the exploration of Finite Mixture Models in a stochastic frontier context (see for instance Beard et al. (1997); Orea and Kumbhakar (2004); El-Gamal and Inanoglu (2005); Greene (2005)).

Moreover, the potential identification of these subpopulations provides the opportunity to discriminate between the many different mechanisms (see for instance Cabral and Mata (2003); Klette and Kortum (2004); Rossi-Hansberg and Wright (2007); Atkeson and Burstein (2010); Caliendo et al. (2020)) that drive the existence of such subpopulations.

References

Albuquerque, R. and H. A. Hopenhayn (2004). Optimal lending contracts and firm dynamics. *The Review of Economic Studies 71*(2), 285–315.

Angelini, P. and A. Generale (2008). On the evolution of firm size distributions. *The American Economic Review 98*(1), 426–438.

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.

Atkeson, A. and A. Burstein (2010). Innovation, firm dynamics, and international trade. *Journal of Political Economy 118*(3), 433–484.

Aw, B. Y., M. J. Roberts, and D. Y. Xu (2011, June). R&D Investment, Exporting, and Productivity Dynamics. *American Economic Review 101*(4), 1312–44.

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 (2017). From micro to macro: Demand, supply, and heterogeneity in the trade elasticity. *Journal of International Economics 108*, 1 – 19.

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

Beard, T. R., S. B. Caudill, and D. M. Gropper (1997). The diffusion of production processes in the us banking industry: A finite mixture approach. *Journal of Banking & Finance 21*(5), 721–740.

Bee, M. and S. Schiavo (2018). Powerless: gains from trade when firm productivity is not Pareto distributed. *Review of World Economics 154*(1), 15–45.

Bernard, A. B., S. J. Redding, and P. K. Schott (2009). Products and productivity. *The Scandinavian Journal of Economics 111*(4), 681–709.

Bloom, N. and J. V. Reenen (2011). Chapter 19 - human resource management and productivity. Volume 4 of *Handbook of Labor Economics*, pp. 1697–1767. Elsevier.

Bottazzi, G., D. Pirino, and F. Tamagni (2015). Zipf law and the firm size distribution: a critical discussion of popular estimators. *Journal of evolutionary economics 25*(3), 585–610.

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.

Caliendo, L., G. Mion, L. D. Opromolla, and E. Rossi-Hansberg (2020). Productivity and organization in portuguese firms. *Journal of Political Economy forthcoming.*

Caliendo, L. and E. Rossi-Hansberg (2012). The impact of trade on organization and productivity. *The Quarterly Journal of Economics 127*(3), 1393–1467.

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. M. and B. Grassi (2019, April). Large firm dynamics and the business cycle. *American Economic Review 109*(4), 1375–1425.

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

Clementi, G. L. and H. A. Hopenhayn (2006, 02). A Theory of Financing Constraints and Firm Dynamics*. *The Quarterly Journal of Economics 121*(1), 229–265.

Cooley, T. F. and V. Quadrini (2001, December). Financial markets and firm dynamics. *American Economic Review 91*(5), 1286–1310.

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.

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

Desai, M., P. Gompers, and J. Lerner (2003). Institutions, capital constraints and entrepreneurial firm dynamics: Evidence from europe. Working Paper 10165, National Bureau of Economic Research.

Dewitte, R. (2020). From Heavy-tailed Micro to Macro: On the characterization of ffirm-level heterogeneity and its aggregation properties. mimeo, Ghent University.

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.

di Giovanni, J., A. A. Levchenko, and R. Rancière (2011). Power laws in firm size and openness to trade: Measurement and implications. *Journal of International Economics 85*(1), 42–52.

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. *American Economic Review 99*(4), 1676–1683.

El-Gamal, M. A. and H. Inanoglu (2005). Inefficiency and heterogeneity in turkish banking: 1990–2000. *Journal of Applied Econometrics 20*(5), 641–664.

Feenstra, R. C. (2018). Restoring the product variety and pro-competitive gains from trade with heterogeneous firms and bounded productivity. *Journal of International Economics 110*, 16–27.

Fernandes, A. M., P. J. Klenow, S. Meleshchuk, M. D. Pierola, and A. Rodríguez-Clare (2018). The Intensive Margin in Trade: Moving Beyond Pareto. Policy Research working paper WPS 8625, World Bank Group.

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.

Fop, M., T. B. Murphy, et al. (2018). Variable selection methods for model-based clustering. *Statistics Surveys 12*, 18–65.

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.

Gandhi, A., S. Navarro, and D. A. Rivers (0). On the identification of gross output production functions. *Journal of Political Economy forthcoming.*

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.

Greene, W. (2005). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics 126*(2), 269–303.

Grün, B. (2018). Model-based clustering. arXiv preprint 1807.01987, arXiv.

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.

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.

Jovanovic, B. (1982). Selection and the evolution of industry. *Econometrica 50*(3), 649–670.

Kasahara, H. and J. Rodrigue (2008). Does the use of imported intermediates increase productivity? plant-level evidence. *Journal of Development Economics 87*(1), 106–118.

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

Klette, T. and S. Kortum (2004). Innovating firms and aggregate innovation. *Journal of Political Economy 112*(5), 986–1018.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

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.

Lentz, R. and D. T. Mortensen (2008). An empirical model of growth through product innovation. *Econometrica 76*(6), 1317–1373.

Levy, M. (2009). Gibrat's Law for (All) Cities: Comment. *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.

McLachlan, G. J. and D. Peel (2000). *Finite mixture models.* New York: Wiley Series in Probability and Statistics.

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 (2014). Chapter 1 - Heterogeneous Firms and Trade. In E. H. Gita Gopinath and K. Rogoff (Eds.), *Handbook of International Economics*, Volume 4 of *Handbook of International Economics*, pp. 1–54. Elsevier.

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.

Orea, L. and S. C. Kumbhakar (2004). Efficiency measurement using a latent class stochastic frontier model. *Empirical Economics 29*(1), 169–183.

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

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

Reed, W. J. and B. D. Hughes (2002). From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. *Physical Review E 66*(6), 067103.

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

Rossi-Hansberg, E. and M. L. J. Wright (2007, December). Establishment size dynamics in the aggregate economy. *American Economic Review 97*(5), 1639–1666.

Sager, E. and O. A. Timoshenko (2019). The double emg distribution and trade elasticities. *Canadian Journal of Economics/Revue canadienne d'économique 52*(4), 1523–1557.

Scollnik, D. P. M. (2007). On composite lognormal-pareto models. *Scandinavian Actuarial Journal 2007*(1), 20–33.

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

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica 57*(2), 307–333.