SALES AND MARKUP DISPERSION:
THEORY AND EMPIRICS

Monika Mrázová†
University of Geneva,
and CEPR

J. Peter Neary‡
University of Oxford,
CEPR and CESifo

Mathieu Parenti§
ULB and CEPR

September 7, 2017

*Earlier versions were circulated under the title “Technology, Demand, and the Size Distribution of Firms”. We are particularly grateful to Jan De Loecker and Julien Martin for assisting us with the data, and to Abi Adams, Costas Arkolakis, Jonathan Dingel, Peter Egger, Basile Grassi, Joe Hirschberg, Oleg Itsikhoki, Luca Macedoni, Rosa Matzkin, Marc Melitz, Steve Redding, Kevin Roberts, Ina Simonovska, Stefan Sperlich, Gonzague Vannoorenberghe, Frank Verboven, Frank Windmeijer, Tianhao Wu, and participants at various conferences and seminars for helpful comments. Monika Mrázová thanks the Fondation de Famille Sandoz for funding under the “Sandoz Family Foundation - Monique de Meuron” Programme for Academic Promotion. Peter Neary thanks the European Research Council for funding under the European Union’s Seventh Framework Programme (FP7/2007-2013), ERC grant agreement no. 295669.

†Geneva School of Economics and Management (GSEM), University of Geneva, Bd. du Pont d’Arve 40, 1211 Geneva 4, Switzerland; e-mail: monika.mrazova@unige.ch.

‡Department of Economics, University of Oxford, Manor Road, Oxford OX1 3UQ, UK; e-mail: peter.neary@economics.ox.ac.uk.

§ECARES, Université Libre de Bruxelles, Av. F.D. Roosevelt, 42, 1050 Brussels, Belgium; mparenti@ulb.ac.be.
Abstract

We derive exact conditions relating the distributions of firm productivity, sales, output, and markups to the form of demand. In particular, for a large family (including Pareto, lognormal, and Fréchet), the distributions of productivity and sales are the same if and only if demand is “CREMR” (Constant Revenue Elasticity of Marginal Revenue). Based on this demand function, we uncover a new class of distributions that are well-suited to capture the dispersion of markups. Empirically, we show that the choice between Pareto and lognormal productivity distributions matters less in explaining sales and markups than the choice between CREMR and other demands.

Keywords: CREMR Demands; Heterogeneous Firms; Kullback-Leibler Divergence; Lognormal Distribution; Pareto Distribution; Sales and Markup Distributions.

JEL Classification: F15, F23, F12
1 Introduction

The hypothesis of a representative agent has provided a useful starting point in many fields of economics. However, sooner or later, intellectual curiosity and the exigencies of matching the empirical evidence make it essential to recognize that agents are heterogeneous. In many cases, this involves constructing models with three components. First is a distribution of agent characteristics, usually assumed exogenous; second is a model of individual agent behavior; and third, implied by the first two, is a predicted distribution of outcomes. Such a pattern can be seen in income distribution theory, in the theory of optimal income taxation, in macroeconomics, and in urban economics.\footnote{For examples, see Stiglitz (1969), Mirrlees (1971), Krusell and Smith (1998), and Behrens, Duranton, and Robert-Nicoud (2014), respectively.} In the field of international trade it has rapidly become the dominant paradigm, since the increasing availability of firm-level export data from the mid-1990s onwards undermined the credibility of representative-firm models, and stimulated new theoretical developments. A key contribution was Melitz (2003), who built on Hopenhayn (1992) to derive an equilibrium model of monopolistic competition with heterogeneous firms. In this setting, the model structure combines assumptions about the distribution of firm productivity and about the form of demand that firms face to derive predictions about the distribution of firm sales. Such models have provided a fertile laboratory for studying a wide range of problems relating to the process of globalization.

Although the pioneering work of Melitz (2003) avoided making specific distributional assumptions, trade models have typically been parameterized in a canonical way, that combines a Pareto distribution of firm productivity on the supply side with CES preferences on the demand side. As shown by Helpman, Melitz, and Yeaple (2004) and Chaney (2008), this combination of assumptions predicts a Pareto distribution of firm sales. This parametrization can be justified on at least two grounds. First is its theoretical tractability, which makes it relatively easy to extend the model to incorporate various real-world features of the global economy, such as outsourcing, multi-product firms, and global value chains.\footnote{See Antràs and Helpman (2004), Bernard, Redding, and Schott (2011), and Antràs and Chor (2013), respectively.} Second is the
empirical regularity that the distribution of firm sales is plausibly close to Pareto, at least in the upper tail. (See, for example, Axtell (2001) and Gabaix (2009).)

However, at least two difficulties arise when this canonical model is confronted with data. First is that not all studies find a Pareto distribution of firm sales, especially if smaller firms are included. Bee and Schiavo (2014) and Head, Mayer, and Thoenig (2014) consider the implications of a lognormal distribution, while Combes, Duranton, Gobillon, Puga, and Roux (2012), and Nigai (2017) explore mixtures and piecewise combinations of Pareto and lognormal, respectively. Analytically, this literature yields a second result: Head, Mayer, and Thoenig (2014) show that lognormal productivities plus CES demands yield a lognormal distribution of firm sales. The parallel between this result and the Helpman-Melitz-Yeaple-Chaney result for the Pareto-CES case is suggestive, but to date the literature gives no guidance on what may happen with different combinations of assumptions.

A second problem with the canonical framework is that a CES demand function has strong counterfactual implications. In particular, it implies that markups are constant across space and time: in the cross section, all firms should have the same markup in all markets; respectively.
while, in the time series, exogenous shocks such as globalization cannot affect markups and so competition effects will never be observed. Trade economists have been uneasy with these stark predictions for some time, and a number of contributions has explored the implications of relaxing the CES assumption, though to date without considering their implications for sales and markup distributions.\footnote{The implications of preferences other than CES have been considered by Melitz and Ottaviano (2008), Zhelobodko, Kokovin, Parenti, and Thisse (2012), Arkolakis, Costinot, Donaldson, and Rodríguez-Clare (2012), Mrázová and Neary (2013), Bertolletti and Epifani (2014), Fabinger and Weyl (2012), Simonovska (2015), and Parenti, Ushchev, and Thisse (2017), among others.} However, only recently has it been possible to confront the predictions of CES-based models with data, following the development of techniques for measuring markups that do not impose assumptions about market structure or the functional form of demand. Figure 1(a) from De Loecker, Goldberg, Khandelwal, and Pavcník (2016) shows that the distribution of markups from a sample of Indian firms is very far from being concentrated at a single value.\footnote{We discuss these data in more detail in Section 6.2 below.} A possible explanation is that such markup heterogeneity arises from aggregation across sectors with different elasticities of substitution. However, Figure 1(b) from Lamorgese, Linarello, and Warzynski (2014) shows that even greater heterogeneity is observed when the data are disaggregated by sector. Taken together, this evidence suggests that markup distributions are far from the Dirac form implied by CES preferences, but to date there is no model of industry equilibrium that would generate such patterns.

In this paper, we provide a general characterization of the problem of explaining the distribution of firm size and firm markups, given particular assumptions about the structure of demand and the distribution of firm productivities. We present two different kinds of results. On the one hand, we present exact conditions under which specific assumptions about the distribution of firm productivity are consistent with a particular form of the distribution of sales revenue, output, or markups. On the other hand, we use the Kullback-Leibler Divergence to quantify the information loss when a predicted distribution fails to match the actual one. We show that applying this tool in the context of models of heterogeneous firms
leads to new insights about the relationship between fundamentals and the size distribution of firms, and also provides a quantitative framework for gauging how well a given set of assumptions explain a given data set.

It hardly needs emphasizing that the assumptions made about productivity distributions and demand structure have crucial implications for a wide range of questions. We mention just three. First is the interpretation of the trade elasticity. The elasticity of trade with respect to trade costs is a constant when demands are CES as shown by Chaney (2008), and this allows a parsimonious expression for the gains from trade as shown by Arkolakis, Costinot, and Rodríguez-Clare (2012); see also Melitz and Redding (2015). Similar results hold with non-CES demands if the distribution of firm productivities is Pareto, as shown by Arkolakis, Costinot, Donaldson, and Rodríguez-Clare (2012). However, when the distribution of firm productivities is lognormal, the trade elasticity is variable and does not take a simple analytic form even with CES demands, as shown by Head, Mayer, and Thoenig (2014); see also Bas, Mayer, and Thoenig (2017). A second reason why these assumptions matter is the granular origins of aggregate fluctuations. Gabaix (2011) showed that relaxing the continuum assumption implies that the largest firms can have an impact on aggregate fluctuations, and di Giovanni and Levchenko (2012) showed that similar effects arise in open economies. These arguments rely on the assumption that the upper tail of the distribution of firm size is a power law, so understanding the mechanisms that may generate this is a key research question. Finally, the interaction of distributional and demand assumptions matters for quantifying the misallocation of resources. The pioneering study of Hsieh and Klenow (2007) showed that close to half the difference in efficiency between China and India on the one hand and the U.S. on the other could be attributed to an inefficient allocation of labor. However, this was under the maintained hypothesis that demands were CES, which, as Dixit and Stiglitz (1977) showed, implies that goods markets are constrained efficient. With non-CES demands, inefficiency may be partly a reflection of goods-market rather than factor-market distortions, with very different implications for welfare-enhancing
policies. (See, for example, Epifani and Gancia (2011) and Dhingra and Morrow (2011).)

In all these cases, the assumptions made about the productivity distribution and demand structure matter for key economic issues, yet the existing literature gives little guidance on the implications of relaxing the standard assumptions, nor how best to proceed when the assumptions of the canonical model do not hold. Our paper aims to throw light on both these questions.

The first part of the paper presents exact characterizations of the links between the distributions of firm attributes, technology and preferences. We begin in Section 2 with two general propositions which characterize the form that very general distributions of firm characteristics and general models of firm behavior must take if they are to be mutually consistent. Sections 3 and 4 then apply these results to distributions of sales and markups respectively. Among the results we derive is a characterization of the demand functions which are necessary and sufficient for productivity and sales to exhibit the same distribution from a wide family which includes Pareto, lognormal and Fréchet. We show that this property is implied by a new family of demands, a generalization of the CES, which we call “CREMR” for “Constant Revenue Elasticity of Marginal Revenue.” The CREMR class has many desirable properties; in particular, it allows for variable markups in a parsimonious way. However, it is very different from most of the non-CES demand systems used in applied economics. We also derive the distributions of markups that are implied by CREMR and other demand functions.

The second part of the paper addresses the question of how to proceed when the conditions for exact consistency between distributions, preferences and technology do not hold. Section 5 presents the Kullback-Leibler Divergence (KLD), which has a natural application to evaluating how “close” a predicted distribution comes to an actual one, and shows how this criterion allows us to quantify the cost of using the “wrong” assumptions about demand or technology to calibrate a hypothetical distribution of firm sales or markups. Section 6

---

5 We use “sales” throughout to refer to sales revenue.

6 “CREMR” rhymes with “dreamer”.

7
illustrates the results of applying the KLD to actual data sets. Finally, Section 7 concludes, while the Appendix contains proofs and more technical details.

2 Characterizing Links Between Distributions

The two central results of the paper link the distributions of two firm characteristics to a general specification of the relationship between them: we make no assumptions about whether either characteristic is exogenous or endogenous, nor about the details of the technological and demand constraints faced by firms which generate the relationship. All we assume is a hypothetical dataset of a continuum of firms, which reports for each firm $i$ its characteristics $x(i)$ and $y(i)$, both of which are monotonically increasing functions of $i$.\(^7\) Formally:

Assumption 1. $\{i, x(i), y(i)\} \in [\Omega \times (\underline{x}, \overline{x}) \times (\underline{y}, \overline{y})]$, where $\Omega$ is the set of firms, with both $x(i)$ and $y(i)$ monotonically increasing functions of $i$.

Examples of $x(i)$ and $y(i)$ include firm productivity, sales and output in most models of heterogeneous firms.

The first result restates a standard result in mathematical statistics in our context; it is closely related to Lemma 1 of Matzkin (2003).

Proposition 1. Given Assumption 1, any two of the following imply the third:

(1) $x$ is distributed with CDF $G(x)$, where $g(x) \equiv G'(x) > 0$;

(2) $y$ is distributed with CDF $F(y)$, where $f(y) \equiv F'(y) > 0$;

(3) Firm behavior, given technology and demand, is such that: $x = v(y)$, $v'(y) > 0$;

\(^7\)The assumption that they are increasing functions is without loss of generality. For example, if $x(i)$ is increasing and $y(i)$ is decreasing, Proposition 1 can easily be reformulated using the survival function of $y$. Monotonicity here is a property of theoretical models. In our empirical applications we do not need to assume that any measured firm characteristics are monotonic in $i$. We follow standard models of firm heterogeneity under monopolistic competition by considering a continuum of firms whose characteristics are realizations of a random variable. Because we work with a continuum, the c.d.f. of this random variable is the actual distribution of these realizations. Henceforward, we use lower-case variables to describe both a random variable and its realization.
where the functions are related as follows:

(i) (1) and (3) imply (2) with \( F(y) = G[v(y)] \) and \( f(y) = g[v(y)]v'(y) \); similarly, (2) and (3) imply (1) with \( G(x) = F[v^{-1}(x)] \) and \( g(x) = f[v^{-1}(x)]\frac{dv^{-1}(x)}{dx} \).

(ii) (1) and (2) imply (3) with \( v(y) = G^{-1}[F(y)] \).

Part (i) of the proposition is a standard result on transformations of variables. Part (ii) is less standard, and requires Assumption 1: characteristics \( x(i) \) and \( y(i) \) must refer to the same firm and must be monotonically increasing in \( i \). The proof is in Appendix A. The importance of the result is that it allows us to characterize fully the conditions under which assumptions about distributions and about the functional forms that link them are mutually consistent. Part (ii) in particular provides an easy way of determining which specifications of firm behavior are consistent with particular assumptions about the distributions of firm characteristics. All that is required is to derive the form of \( v(y) \) implied by any pair of distributional assumptions.

Our next result shows how Proposition 1 is significantly strengthened when the distributions of the two firm characteristics share a common parametric structure, which is given by the following:

**Definition 1.** A sub-family of probability distributions is a member of the “Generalized Power Function” [“GPF”] family if there exists a continuously differentiable function \( H(\cdot) \) such that the cumulative distribution function of every member of the sub-family can be written as:

\[
G(x; \theta) = H \left( \theta_0 + \frac{\theta_1 x^{\theta_2}}{\theta_2} \right) \tag{1}
\]

where each member of the sub-family corresponds to a particular value of the vector \( \theta \equiv \{\theta_0, \theta_1, \theta_2\} \).

The function \( H(\cdot) \) is completely general, other than exhibiting the minimal requirements of a probability distribution: \( G(x_{\min}; \theta) = 0 \) and \( G(x_{\max}; \theta) = 1 \), where \( x_{\min} \) and \( x_{\max} \) are the

---

\(^8\)This implies that the Spearman rank correlation between \( x \) and \( y \) is one.
bounds of the support of $G$; and, to be consistent with a strictly positive density function, $G_x > 0$, $H(\cdot)$ must satisfy the restriction: $\theta_1 H' > 0$. The great convenience of the GPF family given by (1) is that it nests many of the most widely-used distributions in applied economics, including Pareto, truncated Pareto, lognormal, uniform, Fréchet, Gumbel, and Weibull. (See Appendix B for details of members of the GPF family.)

Combining Proposition 1 and Definition 1 gives the following:

**Proposition 2.** Given Assumption 1, any two of the following imply the third:

1. The distribution of $x$ is a member of the GPF family: $G(x; \theta) = H \left( \theta_0 + \frac{\theta_1}{\theta_2} x^{\theta_2} \right)$, $G_x > 0$;
2. The distribution of $h(y)$ has the same form as that of $x$ but with different parameters: $F(y; \theta') = G \left[ h(y); \theta' \right] = H \left( \theta_0 + \frac{\theta'_1}{\theta'_2} h(y)^{\theta'_2} \right)$, $F_y > 0$;
3. $x$ is a power function of $h(y)$: $x = x_0 h(y)^E$;

where the parameters are related as follows:

(i) (1) and (3) imply (2) with $\theta'_1 = E \theta_1 x_0^{\theta_2}$ and $\theta'_2 = E \theta_2$; similarly, (2) and (3) imply (1) with $\theta_1 = E^{-1} \theta'_1 x_0^{-E^{-1} \theta'_2}$ and $\theta_2 = E^{-1} \theta'_2$.

(ii) (1) and (2) imply (3) with $x_0 = \left( \frac{\theta_2 \theta'_1}{\theta_1 \theta'_2} \right)^{\frac{1}{E}}$ and $E = \frac{\theta'_2}{\theta_2}$.

Comparing the distributions of $x$ and $y$ in (1) and (2), they are the same member of the GPF family, except that the $\theta_1$ and $\theta_2$ parameters are different, and that $y$ is subject to an arbitrary monotonic transformation $h(y)$. The $h(\cdot)$ function is completely general, and the elements of the parameter vector $\theta$ can take on any values, except in two respects: $h$ must be monotonically increasing from the strict monotonicity restriction on $F$: $h' > 0$ since $F_y = G_x h' > 0$; and $\theta_0$ must be the same in both distributions, so both $G(x; \theta)$ and $F(y; \theta')$ are two-parameter distributions.

\(^9\)The proof is in Appendix C.
Each choice of the $h(\cdot)$ function generates in turn a further family, such that the transformation $h(y)$ follows a distribution from the GPF family.\textsuperscript{10} Proposition 2 shows that these families are intimately linked via a simple power function that expresses one of the two firm characteristics as a transformation of the other. In much of the paper we will concentrate on two special forms for the $h(\cdot)$ function. The identity transformation, $h(y) = y$, implies from Proposition 2 a property we call “self-reflection”, since the distributions of $x$ and $y$ are the same. This case proves particularly useful when we consider distributions of firm sales and output in Section 3. The other case we consider in detail is the odds transformation, $h(y) = \frac{y}{1-y}$, where $0 \leq y \leq 1$. This implies a property we call “odds-reflection”, since the distribution of $y$ is an odds transformation of that of $x$. This case proves particularly useful when we consider distributions of firm markups in Section 4.

In the next two sections we give some examples of links between distributions and models of firm behavior implied by Propositions 1 and 2, with detailed derivations in Appendix F.

3 Backing Out Demands

The first set of applications of Proposition 2 apply part (ii) of the proposition: we ask what demand functions are consistent with assumed distributions of two different firm attributes. Moreover, following the existing literature, we ask when will we observe self-reflection, in the sense that the distributions of the two attributes are the same (though with different parameters of course). Figure 2 summarizes schematically the results of this section, which specify the demand functions that are necessary and sufficient for self-reflection between the distributions of any two of firm output $x$, sales revenue $r$, and productivity $\varphi$.

\textsuperscript{10}Assuming that a transformation of a variable follows a standard distribution is a well-known method of generating new functional forms for distributions. See Johnson (1949), who attributes it to Edgeworth, and Jones (2015).
3.1 Self-Reflection of Productivity and Sales

We begin in this sub-section by focusing on the two central attributes of productivity and sales revenue. We know from Helpman, Melitz, and Yeaple (2004) and Chaney (2008) that CES demands are sufficient to bridge the gap between two Pareto distributions; and we also know from Head, Mayer, and Thoenig (2014) that a lognormal distribution of productivity coupled with CES demands implies a lognormal distribution of sales. We want to establish the necessary conditions for these links, which in turn will tell us whether there are other demand systems that ensure an exact correspondence between the form of the productivity and sales distributions.

The answer to these questions is immediate from Proposition 2: if both productivity $\varphi$ and sales $r$ follow the same distribution, which can be any member of the GPF family, including the Pareto and the lognormal, then they must be related by a power function:

$$\varphi = \varphi_0 r^E$$  \hspace{1cm} (2)

To infer the implications of this for demand, we use two properties of a monopolistically competitive equilibrium. First, firms equate marginal cost to marginal revenue, so $\varphi = c^{-1} = \left(\frac{\partial r}{\partial x}\right)^{-1}$. Second, all firms face the same residual demand function, so firm sales

\[\text{Our approach does not require that the marginal costs be exogenous. They could be chosen endogenously by firms either by optimizing subject to a variable cost function, as in Zhelobodko, Kokovin, Parenti, and Thisse (2012), or as the outcome of investment in R&D.}\]
conditional on output are independent of productivity $\varphi$: $r(x) = xp(x)$ and $\frac{\partial r}{\partial x} = r'(x)$. Combining these with (2) gives a simple differential equation in sales revenue:

$$[r'(x)]^{-1} = \varphi_0 r(x)^E$$

(3)

Integrating this we find that a necessary and sufficient condition for self-reflection of productivity and sales is that the inverse demand function take the following form:

$$p(x) = \frac{\beta}{x}(x - \gamma)^{\frac{\sigma - 1}{\sigma}}$$

$1 < \sigma < \infty$, $x > \gamma\sigma$, $\beta > 0$ (4)

We are not aware of any previous discussion of the family of inverse demand functions in (4), which expresses expenditure $r(x) = xp(x)$ as a power function of consumption relative to a benchmark $\gamma$. We detail its properties in Appendix D. Its key property, from (3), is that the elasticity of marginal revenue with respect to total revenue is constant: $E = \frac{1}{\sigma-1}$. Hence we call it the “CREMR” family, for “Constant Revenue Elasticity of Marginal Revenue.” It includes CES demands as a special case: when $\gamma$ equals zero, (4) reduces to $p(x) = \beta x^{-\frac{1}{\sigma}}$, and the elasticity of demand is constant, equal to $\sigma$. More generally, the elasticity of demand varies with consumption, $\varepsilon(x) = -\frac{p(x)}{xp(x)} = \frac{x-\gamma}{x-\gamma\sigma}$, though it approaches $\sigma$ for large firms.

To give some intuition for the result that CREMR demands link GPF productivity and sales, consider the Pareto case. A Pareto distribution of productivities $\varphi$ implies that the elasticity of the density of the productivity distribution is constant: if $G(\varphi)$ is Pareto, so $G(\varphi) = 1 - (\frac{\varphi}{\bar{\varphi}})^{-k}$, with density function $g(\varphi) = G'(\varphi)$, then the elasticity of density is $\frac{\varphi g'(\varphi)}{g(\varphi)} = -(k+1)$. Similarly, a Pareto distribution of sales, $r = px$, implies that the elasticity of the density of the sales distribution is constant: if $F(r) = 1 - (\frac{r}{\bar{r}})^{-n}$, with density function $f(r) = F'(r)$, then the elasticity of density is $\frac{rf'(r)}{f(r)} = -(n+1)$. These two log-linear relationships are only consistent with each other if demands also imply a log-linear relationship between firm productivity and firm sales. In a Melitz-type model, productivity is the inverse of marginal cost, which equals marginal revenue. Hence Pareto productivities and
Pareto sales are only consistent with each other if there is a log-linear relationship between marginal and total revenue, which is the eponymous defining feature of CREMR demands. To see this slightly more formally, suppose that the distribution of productivity is Pareto with shape parameter $k$. Then for any two levels of productivity, $c_1^{-1}$ and $c_2^{-1}$, the ratio of their survival functions (one minus their cumulative probabilities) is $\left(\frac{c_2}{c_1}\right)^k$. Since firms are profit-maximizers, this is also the ratio of the survival functions of marginal revenues, $\left[\frac{r'(x_2)}{r'(x_1)}\right]^k$. But if the elasticity of marginal revenue to sales revenue is constant and equal to $\frac{1}{\sigma-1}$, this in turn equals $\left(\frac{c_2}{c_1}\right)^{\frac{k}{\sigma-1}}$. Since this is true for any arbitrary level of sales, it implies that sales are distributed as a Pareto with scale parameter $n = \frac{k}{\sigma-1}$. This result was derived for the case of Pareto productivities and CES demands by Chaney (2008). (See also Helpman, Melitz, and Yeaple (2004).) The formal proof, a corollary of Proposition 2, shows that it generalizes from CES to CREMR, and that GPF productivities and CREMR demands are necessary as well as sufficient for this outcome.

Figure 3 shows three representative inverse demand curves from the CREMR family, along with their corresponding marginal revenue curves. The CES case in panel (a) combines the familiar advantage of analytic tractability with the equally familiar disadvantage of imposing strong and counter-factual properties. In particular, the proportional markup $\xi$ must be the same for all firms in all markets. By contrast, members of the CREMR family with non-zero values of $\gamma$ avoid this restriction. Moreover, we show in Appendix D
that the sign of $\gamma$ determines whether a CREMR demand function is more or less convex than a CES demand function. The case of a positive $\gamma$ as in panel (b) corresponds to demands that are “subconvex”: less convex at each point than a CES demand function with the same elasticity. In this case the elasticity of demand falls with output, which implies that larger firms have higher markups and that globalization has a pro-competitive effect. These properties are reversed when $\gamma$ is negative as in panel (c). Now the demands are “superconvex” – more convex than a CES demand function with the same elasticity – and larger firms have smaller markups. CREMR demands thus allow for a much wider range of comparative statics responses than the CES itself. Finally, CREMR demands can be rationalized by an additively separable utility function where the sub-utility is a hypergeometric function. (For details, see Appendix E.) Since this is an analytic function, CREMR demands can be used as a foundation for quantitative analysis of normative issues.

How do CREMR demands compare with other better-known demand systems? Inspecting the demand functions themselves is not so informative, as they depend on three different parameters. Instead, we use the approach of Mrázová and Neary (2013), who show that any well-behaved demand function can be represented by its “demand manifold”, a smooth curve relating its elasticity $\varepsilon(x) \equiv -\frac{p'(x)}{xp''(x)}$ to its convexity $\rho(x) \equiv -\frac{xp''(x)}{p'(x)}$. We show in Appendix
D that the CREMR demand manifold can be written in closed form as follows:

\[ \bar{\rho}(\varepsilon) = 2 - \frac{1}{\sigma - 1} \frac{(\varepsilon - 1)^2}{\varepsilon} \]  

(5)

Whereas the demand function (4) depends on three parameters, the corresponding demand manifold only depends on \( \sigma \). Panel (a) of Figure 4 illustrates some manifolds from this family for different values of \( \sigma \), while panel (b) shows the manifolds of some of the most commonly-used demand functions in applied economics: linear, CARA, Translog and Stone-Geary (or Linear Expenditure System).\(^{12}\) It is clear that CREMR manifolds, and hence CREMR demand functions, behave very differently from the others. The arrows in Figure 4 denote the direction of movement as sales increase. In the empirically relevant subconvex region, where demands are less convex than the CES, CREMR demands are more concave at low levels of output (i.e., at high demand elasticities) than any of the others, and their elasticity of demand falls more slowly with convexity as sales rise.

### 3.2 CREMR and GPF Distributions: Some Special Cases

While the result of the previous sub-section holds for any distributions from the GPF family, it is useful to consider in more detail the Pareto and lognormal cases. Starting with the Pareto, since it is a member of the GPF family of distributions, it follows immediately as a corollary of Proposition 2 that CREMR demands are necessary and sufficient for self-reflection in this case. We state the result formally for completeness, and because it makes explicit the links that must hold between the parameters of the two Pareto distributions and

\(^{12}\)These manifolds are derived in Mrázová and Neary (2013). We confine attention to the admissible region, \( \{ \varepsilon > 1, \rho < 2 \} \), defined as the region where firms’ first- and second-order conditions are satisfied. The curve labeled “CES” is the locus \( \varepsilon = \frac{1}{\rho - 1} \), each point on which corresponds to a particular CES demand function; this is also equation (5) with \( \varepsilon = \sigma \). To the right of the CES locus is the superconvex region (where demand is more convex than the CES), while to the left is the subconvex region. The curve labeled “SM” is the locus \( \varepsilon = 3 - \rho \); to the right is the “supermodular” region (where selection effects in models of heterogeneous firms have the conventional sign, e.g., more efficient firms serve foreign markets by foreign direct investment rather than exports); while to the left is the submodular region. See Mrázová and Neary (2011) for further discussion.
the demand function. (In what follows we use \( r \sim \mathcal{P}(r, n) \) to indicate that \( r \) follows a Pareto distribution with threshold parameter \( r \) and shape parameter \( n \), so \( F(r) = 1 - \left( \frac{r}{r} \right)^{-n} \).

**Corollary 1. Pareto Productivity and Sales Revenue:** Any two of the following statements imply the third: 1. Firm productivity \( \varphi \sim \mathcal{P}(\varphi, k) \); 2. Firm sales revenue \( r \sim \mathcal{P}(r, n) \); 3. The demand function belongs to the CREMR family in (4); where the parameters are related as follows:

\[
\sigma = \frac{k + n}{n} \Leftrightarrow n = \frac{k}{\sigma - 1} \quad \text{and} \quad \beta = \left( \frac{k + n \varphi^k}{k} \right)^{\frac{k}{k + n}} \Leftrightarrow r = \beta^\sigma \left( \frac{\sigma - 1}{\varphi} \right)^{\sigma - 1} \tag{6}
\]

Note that the demand parameter \( \gamma \) does not appear in (6), so these expressions hold for all members of the CREMR family, including the CES. This confirms that Corollary 1 extends a result of Chaney (2008), as noted earlier.

Although it has become customary to assume that actual firm size distributions can be approximated by the Pareto, at least for larger firms, there are other candidate explanations for the pattern of firm sales. Head, Mayer, and Thoenig (2014) and Bee and Schiavo (2014) argue that firm size distribution is better approximated by a lognormal distribution than a Pareto. We have already noted that the lognormal distribution is a special case of the GPF family in Proposition 2. It follows immediately from the proposition that the CREMR relationship \( \varphi = \varphi_0 r^k \) is necessary and sufficient for self-reflection in the lognormal case. However, unlike in the Pareto case, this does not imply that all CREMR demand functions are consistent with lognormal productivity and sales. The reason is that, except in the CES case (when the CREMR parameter \( \gamma \) is zero), the value of sales revenue for the smallest firm is strictly positive.\(^{13}\) Strictly speaking, this is inconsistent with the lognormal distribution, whose lower bound is zero. We can summarize this result as follows. (We use \( r \sim \mathcal{LN}(\mu, s) \) to indicate that \( r \) follows a lognormal distribution with location parameter \( \mu \) and scale

\(^{13}\)Since \( p'(x) = -\frac{\beta}{\sigma^2} (x - \gamma)^{-\frac{1}{2}} (x - \gamma \sigma) \), the output of the smallest firm when \( \gamma \) is strictly positive is \( \gamma \sigma \), while its sales revenue is \( r(x) = \beta [\gamma (\sigma - 1)]^{-\frac{\sigma - 1}{\sigma - 2}} > 0 \). When demands are strictly superconvex, so \( \gamma \) is strictly negative, sales revenue is discontinuous at \( x = 0 \): \( \lim_{x \to 0^+} r(x) = \beta (\gamma)^{\frac{\sigma - 1}{\sigma - 2}} > 0 \), but \( r(0) = 0 \).
parameter $s$, equal to the mean and standard deviation of the natural logarithm of $r$. Hence $F(r) = \Phi \left( \frac{\log r - \mu}{s} \right)$, where $\Phi$ is the cumulative distribution function of the standard normal distribution.

**Corollary 2. Lognormal Productivity and Sales Revenue:** Any two of the following statements imply the third: 1. Firm productivity follows a $\mathcal{LN}(\mu, s)$ distribution; 2. Firm sales follow a $\mathcal{LN}(\mu', s')$ distribution; 3. The demand function is CES: $p(x) = \beta x^{-\frac{1}{\sigma}}$; where the parameters are related as follows:

$$\sigma = \frac{s + s'}{s} \quad \Leftrightarrow \quad s' = (\sigma - 1)s \quad \text{and} \quad \beta = \frac{s + s'}{s} \exp \left( \frac{s}{s'} \mu' - \mu \right) \quad \Leftrightarrow \quad \mu' = (\sigma - 1) \left[ \mu + \log \left( \frac{\beta}{\sigma} \right) \right]$$

Hence, unlike the Pareto case, the only demand function that is exactly compatible with lognormal productivity and sales is the CES. Relaxing the assumption of Pareto productivity in favor of lognormal productivity comes at the expense of ruling out pro-competitive effects. However, in practical applications, where there is a finite interval between the output of the smallest firm and zero, we may not wish to rule out combining lognormal productivity with members of the CREMR family other than the CES.

**3.3 Self-Reflection of Productivity and Output**

The distribution of sales revenue is not the only outcome predicted by models of heterogeneous firms. We can also ask what are the conditions under which output follows the same distribution as productivity. Proposition 2 implies that a necessary and sufficient condition for this form of self-reflection is that the elasticity of productivity with respect to output be constant. This turns out to be related to a different demand family:

$$p(x) = \frac{1}{x} \left( \alpha + \beta x^{\frac{s-1}{s}} \right)$$

18
The demand function in (8) plays the same role with respect to the characteristic of interest, in this case firm output, as the CREMR family does with respect to firm sales. It is necessary and sufficient for a constant elasticity of marginal revenue with respect to output, equal to \( \frac{1}{\sigma} \). Hence we call it “CEMR” for “Constant (Output) Elasticity of Marginal Revenue.”\(^{14}\)

Unlike CREMR, there are some precedents for this class. It has the same functional form, except with prices and quantities reversed, as the direct PIGL (“Price-Independent Generalized Linearity”) class of Muellbauer (1975).\(^{15}\) In particular, the limiting case where \( \sigma \) approaches one is the inverse translog demand function of Christensen, Jorgenson, and Lau (1975). However, except for the CES (the special case when \( \alpha = 0 \)), CEMR demands bear little resemblance to commonly-used demand functions.\(^{16}\)

When the common distribution of productivity and output is a Pareto, we can immediately state a further corollary of Proposition 2:

**Corollary 3. Pareto Productivity and Output:** Any two of the following statements imply the third: 1. Firm productivity \( \phi \sim P(\phi, k) \); 2. Firm output \( x \sim P(x, m) \); 3. The demand function belongs to the CEMR family (8); where the parameters are related as follows:

\[
\sigma = \frac{k}{m} \Leftrightarrow m = \frac{k}{\sigma} \quad \text{and} \quad \beta = \frac{k}{k - m} \frac{x^m}{\phi} \Leftrightarrow x = \left( \beta \frac{\sigma - 1}{\sigma} \phi \right)^{\sigma} \tag{9}
\]

However, when both productivity and output follow a lognormal distribution, we encounter a similar though less extreme restriction on the range of admissible CEMR demand functions to that in the CREMR case of Corollary 2. Now the requirement that output be zero for the smallest firm is only possible if both the parameters \( \alpha \) and \( \beta \) in the CEMR demand function (8) are positive. As shown by Mrázová and Neary (2013), this corresponds

\(^{14}\)“CEMR” rhymes with “seemer.”

\(^{15}\)For this reason, Mrázová and Neary (2013) called it the “inverse PIGL” class of demand functions.

\(^{16}\)As shown by Mrázová and Neary (2013), the CEMR demand manifold implies a linear relationship between the convexity and elasticity of demand, passing through the Cobb-Douglas point \((\varepsilon, \rho) = (1, 2)\): \( \rho = 2 - \frac{\varepsilon - 1}{\sigma} \). The manifold for the inverse translog special case \((\sigma \to 1)\) coincides with the SM locus in Figure 4(b). For high elasticities (corresponding to small firms when demand is subconvex), CEMR demands are qualitatively similar to CREMR, except that they are somewhat more elastic: the CEMR manifold can be written as \( \varepsilon = (2 - \rho)\sigma + 1 \), while for high \( \varepsilon \) the CREMR manifold becomes \( \varepsilon = (2 - \rho)(\sigma - 1) + 1 \).
to the case where CEMR demands are superconvex. By contrast, if either $\alpha$ or $\beta$ is strictly negative, then demands are strictly subconvex: more plausible in terms of its implications for the distribution of markups, but not compatible with a lognormal distribution of output. Summarizing:

**Corollary 4. Lognormal Productivity and Output:** Any two of the following statements imply the third: 1. Firm productivity follows a $\mathcal{LN}(\mu, s)$ distribution; 2. Firm output follows a $\mathcal{LN}(\mu', s')$ distribution; 3. The demand function belongs to the superconvex subclass of the CEMR family (8) with $\alpha \geq 0$, $\beta \geq 0$, and $\alpha \beta > 0$; where the parameters are related as follows:

$$\sigma = \frac{s'}{s} \Leftrightarrow s' = \sigma s \quad \text{and} \quad \beta = \frac{s'}{s'-s} \exp \left( \frac{s}{s'} \mu' - \mu \right) \Leftrightarrow \mu' = \sigma \left[ \mu + \log \left( \frac{\beta \sigma - 1}{\sigma} \right) \right]$$

(10)

3.4 Self-Reflection of Output and Sales

A final self-reflection corollary of Proposition 2 relates to the case where output and sales follow the same distribution. This requires that the elasticity of one with respect to the other is constant, which implies that the demand function must be a CES.\(^{17}\) Formally:

**Corollary 5. Pareto Output and Sales Revenue:** Any two of the following statements imply the third: 1. The distribution of firm output $x$ is a member of the generalized power function family; 2. The distribution of firm sales revenue $r$ is the same member of the generalized power function family; 3. The demand function is CES: $p(x) = \beta x^{-\frac{1}{\sigma}}$, where $\beta = x_0^{-\frac{1}{\sigma}}$ and $\sigma = \frac{E}{E-1}$.

In the Pareto case, the sufficiency part of this result is familiar from the large literature on the Melitz model with CES demands: it is implicit in Chaney (2008) for example. The necessity part, taken together with earlier results, shows that it is not possible for all three

\(^{17}\)Suppose that $x = x_0 r(x)^E$. Recalling that $r(x) = xp(x)$, it follows immediately that the demand function must take the CES form.
firm attributes, productivity, sales and revenue, to have the same distribution from the
generalized power family class under any demand system other than the CES. Corollary 5
follows immediately from previous results when productivities themselves have a generalized
power function distribution, since the only demand function which is a member of both the
CEMR and CREMR families is the CES itself. However, it is much more general than that,
since it does not require any assumption about the underlying distribution of productivities.
It is an example of a corollary to Proposition 2 which relates two endogenous firm outcomes
rather than an exogenous and an endogenous one. Taken together, the results of this section
show that exactly matching a Pareto or lognormal distribution of firm sales or output, when
productivity is assumed to have the same distribution, places strong restrictions on the
admissible demand function. The elasticity of marginal revenue with respect to the firm
outcome of interest must be constant, and the implied demand function must be consistent
with the range of the distribution assumed. However, that leaves open the question of how
great an error would be made by using a demand function which does not allow for an exact
fit. We address this question in Section 5. First, we turn to consider the implications of
different demand functions for the distributions of sales and markups.

4 Inferring Sales and Markup Distributions

The previous section used part (ii) of Proposition 2 to back out the demands implied by
assumed distributions of two firm characteristics. In this section we show how part (i) can be
used to derive the distributions of firm characteristics given the distribution of productivity
and the form of the demand function. Section 4.1 considers the distributions of markups
implied by CREMR demands, while Section 4.2 presents the distributions of both sales and
markups implied by a number of widely-used demand functions.
4.1 CREMR Markup Distributions

We begin with the case of CREMR demands, since they imply a particularly simple form for the markup distribution. In order to be able to invoke Proposition 2, we need to express productivity as a function of the markup.

The first step is to express output as a function of the markup. In general, with the markup \( m \) defined as \( \frac{p}{c} \), we can write the markup as a function of output by invoking a standard expression in terms of the elasticity of demand: \( m(x) = \frac{\varepsilon(x)}{\varepsilon(x)-1} \). Specializing to the case of CREMR demands, recall from Section 3.1 that the elasticity of demand for CREMR demand functions is \( \varepsilon(x) = \frac{x - \gamma}{x - \gamma \sigma} \). Hence, we can write the CREMR markup as a function of output: \( m(x) = \frac{x - \gamma}{x - \gamma \sigma} \). We concentrate on the case of subconvex demands (i.e., \( \gamma > 0 \)), which implies that larger firms have higher markups: \( m(x) \in [m, \frac{x}{\sigma-1}] \) as \( x \in [x, \infty] \). Define the relative markup as the markup relative to its maximum value, \( \frac{\sigma}{\sigma-1} \), which is the value that obtains under CES preferences with the same value of \( \sigma \): \( \bar{m} \equiv \frac{m}{\bar{m}} = \frac{\sigma-1}{\sigma} m \in [\bar{m}, 1] \). Hence it follows that: \( \bar{m}(x) = \frac{x - \gamma}{x} \). Inverting this allows us to express output as a function of the relative markup: \( x(\bar{m}) = \gamma \frac{1}{1-\bar{m}} \).

The next step is to express productivity \( \varphi \) as a function of output. This follows from profit-maximization, which implies that marginal cost \( \varphi^{-1} \) equals marginal revenue, given by equation (22) in Appendix D: \( \varphi(x) = \frac{1}{\beta} \frac{\sigma}{\sigma-1} (x - \gamma)^{\frac{1}{\sigma}} \). Finally, combining \( \varphi(x) \) and \( x(\bar{m}) \), gives the desired relationship between productivity and the markup:

\[
\varphi(\bar{m}) = \varphi_0 \left( \frac{\bar{m}}{1 - \bar{m}} \right)^{\frac{1}{\sigma}} \quad \varphi_0 \equiv \frac{1}{\beta} \frac{\sigma}{\sigma-1} \gamma^{\frac{1}{\sigma}} \quad (11)
\]

Clearly this satisfies Proposition 2’s conditions for “Odds Reflection”. Hence, if productivity follows any distribution in the GPF class, and if the demand function belongs to the subconvex CREMR family, equation (4) with \( \gamma > 0 \), then Proposition 2 implies that markups follow the corresponding “GPF-odds” distribution.

Once again, we focus on three particularly interesting cases:
1. Pareto: If demands are subconvex CREMR and productivity $\varphi$ is distributed as a Pareto, so $G(\varphi) = 1 - \frac{\varphi}{k} \varphi^{-k}$, then the relative markup must follow a “Pareto-Odds” distribution:

$$F(\tilde{m}) = 1 - \left( \frac{\tilde{m}}{1 - \tilde{m}} \right)^{n'} \left( \frac{\tilde{m}}{1 - \tilde{m}} \right)^{-n'} \quad \tilde{m} \in \{\tilde{m}, 1\} \quad \tilde{m} \equiv \frac{m}{\bar{m}}, \tilde{m} \equiv \frac{m}{\bar{m}}. \quad (12)$$

where $n' \equiv \frac{k}{s}$ and $\tilde{m} \equiv \frac{\varphi_0}{\varphi_0 + \varphi_0}$. This distribution appears to be new, and may prove useful in future applications. However, it implies that the distribution of markups is U-shaped, which is less in line with the available evidence than the next case we consider, although the minimum value of the U may lie to the left of the relevant $[0, 1]$ interval.

2. Lognormal: If demands are subconvex CREMR and productivity follows a lognormal distribution, so $G(\varphi) = \Phi \left[ \frac{1}{s} \{ \log \varphi - \mu \} \right]$, then the relative markup must follow a “Lognormal-Odds” distribution:

$$F(\tilde{m}) = \Phi \left[ \frac{1}{s'} \left\{ \log \frac{\tilde{m}}{1 - \tilde{m}} - \mu' \right\} \right] \quad (13)$$

where: $s' \equiv \sigma s$ and $\mu' \equiv \sigma (\mu - \log \varphi_0)$. This distribution has been studied in the statistics literature where it is known as the “Logit-Normal”, though we are not aware of a theoretical rationale for its occurrence as here.\(^{18}\) Figure 5 illustrates some mem-

\(^{18}\)See Johnson (1949) and Mead (1965).
bers of this family of distributions. Comparing these with the empirical results from De Loecker, Goldberg, Khandelwal, and Pavcnik (2016) and Lamorgese, Linarello, and Warzynski (2014) illustrated in Figure 1, which also exhibit inverted-U-shaped profiles, suggests that the lognormal-odds distribution provides a good fit for the empirical markup distribution. Of course, a more precise way of measuring goodness of fit of distributions would be preferable; we will turn to this in the next section.

3. Fréchet: Finally, if productivity follows a Fréchet distribution and demands are CREMR, then the relative markup must follow a “Fréchet-Odds” distribution. Once again, this distribution appears to be new. It provides an exact characterization of the distribution of profit margins for a firm that sells in many foreign markets, where the distribution of productivity draws across markets follows a Fréchet distribution, as in Tintelnot (2017).

4.2 Other Sales and Markup Distributions

<table>
<thead>
<tr>
<th>Demand Function</th>
<th>( p(x) ) or ( x(p) )</th>
<th>( \varphi(r) ) or ( \varphi(\tilde{r}) )</th>
<th>( \varphi(m) ) or ( \varphi(\tilde{m}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREMR</td>
<td>( \frac{\beta}{\alpha} \left( x - \gamma \right) \left( \frac{\gamma}{\alpha} \right)^{\frac{\alpha - 1}{\gamma}} )</td>
<td>( \varphi_0 \left( \frac{\alpha}{\alpha - 1} \right)^{\frac{1}{\alpha - 1}} )</td>
<td>( \varphi_0 \left( \frac{\beta}{\beta - 1} \right)^{\frac{1}{\beta - 1}} )</td>
</tr>
<tr>
<td>Linear</td>
<td>( \alpha - \beta x )</td>
<td>( \frac{1}{\alpha} \left( \frac{1}{1 - \tau} \right)^{\frac{1}{2}} )</td>
<td>( \frac{2m - 1}{\alpha} )</td>
</tr>
<tr>
<td>LES</td>
<td>( \frac{\delta}{x + \gamma} )</td>
<td>( \gamma \delta \left( \frac{1}{1 - \tau} \right)^{2} )</td>
<td>( \frac{\gamma^2 m^2}{\delta} )</td>
</tr>
<tr>
<td>Translog/AI</td>
<td>( \frac{1}{p} (\gamma - \eta \log p) )</td>
<td>( \varphi_0(r + \eta) \exp \left( \frac{r}{\eta} \right) )</td>
<td>( m \exp \left( m - \frac{\eta}{\eta + \gamma} \right) )</td>
</tr>
</tbody>
</table>

Table 1: Productivity as a Function of Sales and Markups for Selected Demand Functions

Proposition 2 can be used to derive the distributions of sales and markups implied by any demand function. In particular, closed-form expressions for productivity as a function of sales or markups can be derived for some of the most widely-used demand functions in applied
economics. Table 1 gives results for linear, Stone-Geary or linear expenditure system (LES), and translog demands, along with the CREMR results already derived.\textsuperscript{19} Combining these with different assumptions about the distribution of productivity, and invoking Proposition 2, it is clear that a wide variety of sales and markup distributions are implied.\textsuperscript{20} For example, the relationships between productivity and sales implied by linear and LES demands have the same form, so the sales distributions implied by these two very different demand systems are observationally equivalent. The same is not true of their implied markup distributions, however: in the LES case, productivity is a simple power function of markups, so the LES implies self-reflection of the productivity and markup distributions if either is a member of the GPF class.\textsuperscript{21}

It is clearly desirable to compare the distributions implied by these different demand functions with each other and with a given empirical distribution. In the remainder of the paper we turn to this task.

5 Comparing Predicted and Actual Distributions

5.1 From Theory to Calibration

So far we have characterized the exact distributions of firm size and firm markups implied by particular assumptions about the primitives of the model: the structure of demand and the distribution of firm productivities. Results of this kind provide an essential benchmark, but they are not so helpful from a quantitative perspective: they do not tell us by how much a theoretically-implied distribution departs from a given distribution, whether hypothetical

\textsuperscript{19}From a firm’s perspective, the translog is observationally equivalent to the almost ideal (AI) model of Deaton and Muellbauer (1980).

\textsuperscript{20}For parameter restrictions and other details, such as the form of $\varphi_0$ (which differs in each case), see Appendix G. Note that in some cases it is desirable to express the results in terms of sales relative to the maximum level, $\hat{r} \equiv \hat{r}$, just as with CREMR demands the markup distribution is most easily expressed in terms of the relative markup $\hat{m}$.

\textsuperscript{21}For example, a lognormal distribution of productivity and LES demand imply a lognormal distribution of markups, so providing microfoundations for an assumption made by Epifani and Gancia (2011).
or observed. In the remainder of the paper, we turn to explore the quantitative implications of our approach when applied to actual data sets. In particular, we quantify the differences between the actual distributions in the data and a variety of distributions implied by different theoretical models. To measure the “goodness of fit” of different models, we use the Kullback-Leibler divergence (denoted “KLD” hereafter), introduced by Kullback and Leibler (1951). We also present results for the QQ estimator as a robustness check.\(^{22}\) The next sub-section sketches the theoretical properties of the KLD, while Section 5.3 shows how we operationalize it. To fix ideas, we focus on explaining the distribution of firm sales. Adapting the framework to explain the distribution of output, markups, or any other firm outcome, is straightforward.

### 5.2 The Kullback-Leibler Divergence

The KLD measures the “information loss” or “relative entropy” when one distribution, \(F\), is used to approximate another, \(\tilde{F}\):

\[
D_{KL}(F||\tilde{F}(\cdot; \theta)) = \int_{\mathbb{R}} \log \left( \frac{f(r)}{\tilde{f}(r; \theta)} \right) f(r) dr
\]

In our context, the observed distribution \(F(r)\) is the actual distribution of firm sales. As for \(\tilde{F}(r; \theta)\), it is the distribution of firm sales implied by the underlying distribution of firm productivities, \(G(\varphi)\), combined with a model of firm behavior, \(\varphi(r; \theta)\), parameterized by \(\theta\).

The KLD has a number of desirable features, the first two of which are well-known. First, it has an axiomatic foundation in information theory: we give further details in Appendix H.1. Second, it has an elegant statistical interpretation: it equals the expected value of the inverse log-likelihood ratio, so choosing the parameter vector \(\theta\) to minimize the KLD

\(^{22}\)Other criteria could be used, though none is as satisfactory as the KLD. A first- or second-order stochastic dominance criterion is not informative about the dissimilarity between the two firm size distributions if their cumulative distributions intersect more than once. The Kolmogorov-Smirnov test privileges the maximum deviation between the two cumulative distributions, and ignores information about the distributions at other points. As for matching moments, this does not guarantee a close fit unless many moments are used. Moreover, there is a specific problem with matching moments for the Pareto distribution. The \(t^{th}\) moment exists if and only if the dispersion parameter \(k\) exceeds \(t\); however, empirically, raw data often exhibit values of \(k\) that are less than one, so even the mean does not exist.
is asymptotically equivalent to maximizing the likelihood. Third, and new in this paper, is that it links directly with Proposition 2: the KLD in our context can be decomposed to show how it relates to the Revenue Elasticity of Marginal Revenue (REMR) $E$:

$$
\mathcal{D}_{KL}(F||\tilde{F}(\cdot; \theta)) = \log f(r) - \log \left[ g(\varphi(r)) \frac{d\varphi}{dr} \right]_r^{\varphi(r)} + \int_{\underline{r}}^{r} \frac{1 - F(r)}{r} \left[ \frac{rf'(r)}{f(r)} + 1 \right] - \left[ \frac{\varphi g'(\varphi(r))}{g(\varphi(r))} + 1 \right] E(r) - \frac{rE'(r)}{E(r)} dr
$$

(15)

Recall that Proposition 2 derived necessary and sufficient conditions for an exact match between the distributions of two firm characteristics when both distributions belong to the same member of the generalized power function family: the elasticity of one characteristic with respect to the other should be constant, and its value should be consistent with the parameters of the two distributions. Equation (15) goes further and quantifies the information loss when the assumptions of Proposition 2 do not hold. In particular, it identifies three distinct sources of information loss in matching a fitted distribution $\tilde{F}(r)$ to an actual distribution of firm sizes $F(r)$. First is a failure to match the lower end-point of the distribution, $\underline{r}$. Second is a mismatch at each point in the range between the actual elasticity of density of the firm size distribution, $\frac{rf'(r)}{f(r)}$, and that predicted by the assumptions about the productivity distribution and the REMR, $\frac{\varphi g'(\varphi(r))}{g(\varphi(r))} E(r)$. And third is a failure to allow for variations in the REMR $E$; i.e. a failure to allow for deviations from part (iii) of Proposition 2. Each of these three components can be positive or negative, but their sum must be non-negative. Appendices H.2 and H.3 give details and applications.

5.3 Operationalizing the KLD

To compare the fit of predicted and actual distributions we use the discrete counterpart of the continuous KLD introduced in Section 5.2. We choose the parameter vector $\theta$ to
minimize the KLD between the empirical c.d.f. \( F \) defined over the support \([r, \bar{r}]\), and the theoretical c.d.f. \( \tilde{F}(.; \theta) \). Considering the histogram corresponding to \( F \), defined over \( n_b \) bins with width equal to \( b \), the KLD becomes:

\[
\mathcal{D}_{KL}(F \mid\mid \tilde{F}(.; \theta)) = \sum_{i=1}^{n_b} \left( F(r + i \times b) - F(r + (i - 1) \times b) \right) \log \left( \frac{F(r + (i + 1) \times b) - F(r + i \times b)}{F(r + (i - 1) \times b) - F(r + i \times b)} \right)
\]

(16)

We report below \( \mathcal{D}_{KL}(F \mid\mid \tilde{F}(.; \hat{\theta})) \), where \( \hat{\theta} \) is the parameter vector that minimizes \( \mathcal{D}_{KL}(F \mid\mid \tilde{F}(.; \theta)) \).

In the next section, we set the number of bins equal to 1,000. Fortunately, the ranking of different models is not very sensitive to the number of bins considered. When the number of bins increases without bound, our estimator is asymptotically equivalent to the maximum likelihood estimator under the additional constraint that the empirical support of the distribution is included in the one predicted by the theory.\(^{23}\) As for the units of measurement for the KLD, information scientists typically present values in “bits” (log to base 2) or “nats” (log to base \( e \)). Such units have little intuitive appeal in economics. Instead, we present the values of the KLD normalized by the value implied by a uniform distribution of sales.\(^{24}\) This is an uninformative prior in the spirit of the “dartboard” approach to benchmarking the geographic concentration of manufacturing industry of Ellison and Glaeser (1997), or the “balls and bins” approach to benchmarking the world trade matrix of Armenter and Koren (2014). The value of the KLD is unbounded, but a specification that gave a value greater than that implied by a uniform distribution could not be considered a satisfactory explanation of the data.

\(^{23}\)When the distribution is lognormal this difference is immaterial as the support consists of \( \mathbb{R}^+ \). This is no longer the case when the distribution is Pareto.

\(^{24}\)See (37) in Appendix H.1 for the explicit expression.
6 Empirical Applications

To illustrate how the KLD can be used to compare the goodness of fit of different assumptions about demand and the distribution of productivity, we end with two empirical applications. The first, in Section 6.1, uses data on French exports to Germany in 2005, drawn from the same source as that used by Head, Mayer, and Thoenig (2014). The second, in Section 6.2, uses firm-level data on Indian sales and markups, as used by De Loecker, Goldberg, Khandelwal, and Pavcnik (2016). Section 6.3 explores how robust are the results with Indian data to dropping smaller observations. Finally, as a further robustness check, Section 6.4 confirms that an alternative criterion for choosing between distributions, the QQ estimator, give qualitatively similar results to the KLD.

6.1 French Exports to Germany

(a) A First Look: Obviously Pareto? (b) A Second Look: Obviously Lognormal?

Figure 6: Alternative Perspectives on the Data

The data consists of the universe of French exports to Germany in 2005. Figure 6 shows that the data exhibit some typical features of such data sets. When we plot a histogram with the log frequency on the vertical axis and actual sales on the horizontal, as in Panel

\footnote{It contains 161,191 firm-product observations on export sales by 27,550 firms: 5.85 products per firm. We are very grateful to Julien Martin for performing the analysis for us on French Customs data.}
(a), the long tail is clearly in evidence, and it seems plausible that the data are generated by a Pareto distribution. However, the first bin contains over half the firms, which is brought out more clearly when we plot the actual frequency on the vertical axis and log sales on the horizontal, as in Panel (b). Now the data seem self-evidently lognormal. Yet a third perspective comes from the vertical lines in Panel (b). The line labeled (1) is at median sales, with 50% of firms to the left, but these account for only 0.1% of sales; the line labeled (2) is at 76.7% of firms, but these account for only 1.0% of sales; finally, the line labeled (3) is at 99.6% of firms, which account for only 50% of sales. Thus, we might reasonably conclude that the data are Pareto where it matters, with the top firms dominating.

![Figure 7: KLD-Minimizing Predicted Distributions: Pareto (green) and Lognormal (red)](image)

These subjective considerations provide a poor basis for discriminating between rival views of the best underlying distribution, and justify our turning to use the KLD as a more objective indicator of how well different assumptions fit the data. Figure 7 compares the best-fit Pareto (in green) and lognormal (in red). From Section 3, each of these amounts to assuming that demand is CREMR, and that the underlying productivity distribution is either Pareto or lognormal. (Note that the distribution and demand parameters are not separately identified.) In each case, we choose parameter values for the specification in question that minimize the KLD: recall that this is asymptotically equivalent to a maximum likelihood
estimation of those parameters, conditional on the specification. Panel (a) illustrates the results for all firms, while panel (b) avoids the distorted perspective caused by including the largest firms, by omitting the top 89 firms, which account for 0.05% of observations but 32% of sales. Inspecting the fitted distributions, it is evident that the lognormal matches the smaller firms better, and conversely for the Pareto. The values of the minimized KLD show that the lognormal provides a better overall fit than the Pareto: 0.0001 as opposed to 0.0012. (Recall that the data are normalized by the value of the KLD for a uniform distribution, which for this data set is 6.8082.)

<table>
<thead>
<tr>
<th></th>
<th>CREMR/CES</th>
<th>Translog/AI</th>
<th>Linear and LES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pareto</td>
<td>0.0012</td>
<td>0.3819</td>
<td>0.4711</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.0001</td>
<td>0.7315</td>
<td>0.8314</td>
</tr>
</tbody>
</table>

Table 2: KLD for French Exports Compared with Predictions from Selected Demand Functions and Productivity Distributions

Table 2 gives the values of the KLD for the Pareto and lognormal cases shown in Figure 7, and also for the distributions implied by either translog or linear demand functions combined with either Pareto or lognormal productivities. These distributions are calculated by combining the relevant productivity distribution with the relationships between productivity and sales implied by translog and linear demands from Table 1. (Recall from that table that the linear and LES specifications are observationally equivalent.) Each entry in the table is the value of the KLD that measures the information loss when the combination of assumptions indicated by the row and column is used to explain the observed distribution of sales.

To assess whether the values are significantly different from one another, we use a bootstrapping approach. We construct one thousand samples of the same size as the data (i.e., 161,191 observations), by sampling with replacement from the original data. For each sample, we then compute the KLD value for each of the six models. Table 3 gives the results. Each entry in the table is the proportion of samples in which the combination in the relevant
column gives a higher value of the KLD than that in the relevant row. All the values are equal to or very close to 100%, which confirms that the results in Table 2 are robust.

<table>
<thead>
<tr>
<th>CREMR + LN</th>
<th>CREMR + P</th>
<th>TLog + P</th>
<th>Lin + P</th>
<th>TLog + LN</th>
<th>Lin + LN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREMR + LN</td>
<td>–</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>CREMR + P</td>
<td>100%</td>
<td>–</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>TLog + P</td>
<td>100%</td>
<td>100%</td>
<td>–</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Lin + P</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>–</td>
<td>0%</td>
</tr>
<tr>
<td>TLog + LN</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>–</td>
</tr>
<tr>
<td>Lin + LN</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99.7%</td>
</tr>
</tbody>
</table>

Table 3: Bootstrapped Robustness of the KLD Ranking: French Sales
(See text for explanation)

Turning to the results in Table 2, recall that panel (a) of Figure 7 showed that the lognormal matches the smaller firms better, and conversely for the Pareto. Table 2 provides a quantitative confirmation of this. With a preponderance of the bins corresponding to smaller firms, it is not surprising that the lognormal does better as measured by the KLD: as shown in the second column, it yields a value of 0.0017, considerably lower than the value of 0.0090 for the Pareto. However, the difference between distributions turns out to be much less significant than that between different specifications of demand. The KLD values for the translog/AI and linear/LES specifications are much higher than for the CREMR case, as shown in the third and fourth columns of Table 2, with the Pareto now preferred to the lognormal. The overwhelming conclusion from these results is that, if we want to fit the distribution of sales in this data set, then the choice between Pareto and lognormal distributions is less important than the choice between CREMR and other demands.

6.2 Indian Sales and Markups

The second data set we use has 2,457 firm-product observations on both sales and markups in Indian manufacturing for the year 2001. (See De Loecker, Goldberg, Khandelwal, and Pavcnik (2016) for a detailed description of the data, which come from the Prowess data set collected by the Centre for Monitoring the Indian Economy (CMIE).) While the sales
Table 4: KLD for Indian Sales and Markups Compared with Predictions from Selected Demand Functions and Productivity Distributions

<table>
<thead>
<tr>
<th></th>
<th>CREMR</th>
<th>Translog/AI</th>
<th>LES</th>
<th>Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Sales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pareto</td>
<td>0.2253</td>
<td>0.1028</td>
<td>0.1837</td>
<td>0.1837</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.0140</td>
<td>0.5825</td>
<td>0.7266</td>
<td>0.7266</td>
</tr>
<tr>
<td>B. Markups</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pareto</td>
<td>0.1851</td>
<td>0.2205</td>
<td>0.2191</td>
<td>0.2512</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.1863</td>
<td>0.2228</td>
<td>0.2083</td>
<td>0.2075</td>
</tr>
</tbody>
</table>

data are directly observed, the markup data are estimated, using the so-called “production approach”. This approach relies solely on cost-minimization: markups are calculated by computing the gap between the output elasticity with respect to variable inputs and the share of those inputs in total revenue. It is particularly well-suited to our purposes, since it does not impose any restrictions on consumer demand and is consistent with a variety of market structures including monopolistic competition. Since the empirical markup distribution has been obtained without making any assumption about functional forms, we can therefore compare objectively the performance of different productivity distributions combined with different demand systems based on the distributions of sales and markups that they imply. The empirical markup distribution was shown in Figure 1(a) above. Observations with negative markups (about 20% of the total) are not included in the sample, as they are inconsistent with steady-state equilibrium behavior by firms. The remaining observations are demeaned by product-year and firm-year fixed effects, so the sample mean equals one by construction.

The KLD results are given in Table 4 and illustrated in Figure 8. As with the French data, these results are normalized by the KLD for a uniform distribution, and bootstrapping confirms that the differences between them are highly robust. (Appendix I gives details.) The KLD values for sales are broadly in line with those from the French data. The one major difference is that, conditional on a Pareto distribution of productivities, CREMR
demands give the worst fit to sales, with translog demands performing best, and linear-LES intermediate between the others. However, the differences between the KLD values for these specifications are much less than those conditional on lognormal productivities. Here the ranking is the same as with the French data: CREMR does best, with translog performing much less well and Linear-LES worst of all.

Of most interest are the results for markups. Here CREMR demands clearly do best, irrespective of the assumed distribution, with translog and LES performing at the same level, and linear doing equally well under Pareto assumptions but less well in the lognormal case. (Recall from Table 1 that linear and LES demands are not separately identified for sales, but they are for markups.) These results for markups reinforce the finding from the French data that the choice between Pareto and lognormal distributions is less important than the choice between CREMR and other demands. For sales a similar pattern applies conditional on lognormal productivity, whereas in the Pareto case the choice of demand is less important, with CREMR doing worst of all.
6.3 Robustness to Truncation

As we have seen in the two preceding sub-sections, the results with French and Indian sales data are very similar, except for the case of CREMR demands combined with Pareto productivity: this gives a good fit with French data but performs less well with Indian data. One possible explanation for this is that the French data relate to exports, whereas the Indian data are for total domestic production. Presumptively, smaller firms have been selected out of the French data, so we might expect the Pareto assumption to be more appropriate. To throw light on this issue, we explore the robustness of the Indian results to left-truncating the data: specifically, we repeat a number of the comparisons between different specifications for the Indian sales distribution dropping one observation at a time.

Figure 9: CREMR vs. CREMR: KLD for Indian Sales

Table 9 compares the KLD for the Pareto and lognormal, conditional on CREMR demands, starting on the left-hand side with all observations (so the values are the same as in Figure 8) and successively dropping up to 809 observations.\textsuperscript{26} Although the curves are not

\textsuperscript{26}Each KLD value is normalized by the value of the KLD for a uniform distribution corresponding to the number of observations dropped. Alternative approaches would make very little difference however, as the KLD value for the uniform varies very little, from 3.9403 with no observations dropped to 3.5598 with 809 observations dropped.
precisely monotonic, the broad picture is clear: conditional on CREMR demands, Pareto does better and lognormal does worse as more and more observations are dropped. The Pareto specification dominates when we drop 663 or more observations: these account for 27% of all firm-product observations, but only 1.2% of total sales.

Figure 10: CREMR vs. The Rest, Given Pareto: KLD for Indian Sales

Figure 10 shows that a similar pattern emerges when we compare the performance of different demand functions in explaining the sales distribution, conditional on a Pareto distribution for productivity. In this case, the CREMR specification overtakes the linear one when we drop 11 or more observations, which account for 0.44% of all firm-product observations, and only 0.0002% of sales. While it overtakes the translog when we drop 118 or more observations, which account for 4.80% of observations, and 0.03% of sales.

These findings confirm that both CREMR and Pareto fit the sales data relatively better when the smallest observations are dropped. They also make precise the pattern observed in Figure 6 and in many other datasets: the right tail of the sales distribution, where the Pareto assumption outperforms the lognormal, begins at exactly 663 observations.

36
6.4 Robustness: The QQ Estimator

A different kind of robustness check is to consider an alternative criterion to the KLD for comparing predicted and actual distributions. Here we consider the QQ estimator, developed by Kratz and Resnick (1996), and previously applied by Head, Mayer, and Thoenig (2014) and Nigai (2017). Unlike the KLD, this estimator does not have a maximum likelihood interpretation. However, it is more intuitive, since the QQ distance measure is simply the sum of the squared deviations of the quantiles of the predicted distribution from those of the actual distribution:

$$QQ(\tilde{F} \parallel F(\cdot; \theta)) = \sum_{i=1}^{n} (\log \tilde{q}_i - \log q_i(\theta))^2$$  \hspace{1cm} (17)

where $\tilde{q}_i = \tilde{F}^{-1}(i/n)$ is the $i$'th quantile observed in the data, while $q_i(\theta) = F^{-1}(i/n; \theta)$ is the $i$'th quantile predicted by the theory. The QQ estimator $\hat{\theta}$ is defined as the parameter vector that minimizes the sum of squares $QQ(\tilde{F} \parallel F(\cdot; \theta))$ in (17).

To implement the QQ estimator we need analytic expressions for the quantiles under each of the eight combinations of assumptions about demand and the distribution of productivity we consider. These are given in Appendix J. We set the number of quantiles $n$ equal to 100. The resulting values of the QQ estimator for Indian sales and markups are given in Table 5, and they are illustrated in Figure 11. As with the KLD values in Section 6.2, we scale these

<table>
<thead>
<tr>
<th></th>
<th>CREMR</th>
<th>Translog</th>
<th>LES</th>
<th>Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Sales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pareto</td>
<td>58.939</td>
<td>12.693</td>
<td>24.484</td>
<td>24.484</td>
</tr>
<tr>
<td>Lognormal</td>
<td>3.078</td>
<td>116.918</td>
<td>133.274</td>
<td>133.274</td>
</tr>
<tr>
<td>B. Markups</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pareto</td>
<td>0.113</td>
<td>0.978</td>
<td>1.133</td>
<td>3.606</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.110</td>
<td>0.990</td>
<td>0.340</td>
<td>0.325</td>
</tr>
</tbody>
</table>

Table 5: QQ Estimator for Indian Sales and Markups

To implement the QQ estimator we need analytic expressions for the quantiles under each of the eight combinations of assumptions about demand and the distribution of productivity we consider. These are given in Appendix J. We set the number of quantiles $n$ equal to 100. The resulting values of the QQ estimator for Indian sales and markups are given in Table 5, and they are illustrated in Figure 11. As with the KLD values in Section 6.2, we scale these
Figure 11: QQ Estimator for Indian Sales and Markups

by the uniform benchmark, which is 2.420 for sales and 90.133 for markups.

Comparing Table 5 with Table 4, and Figure 11 with Figure 8, it is evident that the results based on the QQ estimator are qualitatively very similar to those for the KLD. In particular, the Pareto assumption gives a better fit for sales than for markups, except in the CREMR case; while the lognormal assumption tends to give a better fit for markups than for sales. Comparing different demand functions, CREMR demands give a better fit to the markup distribution than any other demands, irrespective of which productivity distribution is assumed. As for sales, the results differ between the Pareto and lognormal cases. Conditional on lognormal, CREMR again performs much better, whereas, conditional on Pareto, it performs least well, with the translog doing best. The only qualitative difference between the results using the two criteria is that with the QQ estimator the translog does somewhat better than the LES in fitting the markup distribution. Overall, we can conclude that the rankings given earlier are not unduly sensitive to our choice of criterion for comparing actual and predicted distributions.
7 Conclusion

This paper has addressed the question of how to explain the distributions of firm size and firm markups using models of heterogeneous firms. We provide a general necessary and sufficient condition for consistency between arbitrary assumptions about the distributions of two firm characteristics and an arbitrary model of firm behavior which relates those two characteristics at the level of an individual firm. In the specific context of Melitz-type models of heterogeneous firms competing in monopolistic competition, we showed that our condition implies a new demand function that generalizes the CES. The CREMR or “Constant Revenue Elasticity of Marginal Revenue” family of demands is necessary and sufficient for a Pareto or lognormal distribution of firm productivities to be consistent with a similar distribution of firm sales.

In addition to exact results of this kind, we showed how the Kullback-Leibler divergence can be used to compare a predicted with an observed distribution of firm size. The value of the Kullback-Leibler divergence can be expressed in terms of the difference between the elasticities of density of the two distributions, which in turn can be related to errors in estimating the level and the rate of change of the elasticity of revenue with respect to marginal revenue. Simulations show that the information cost of using the “wrong” parameter to calibrate an observed distribution can be highly asymmetric. Finally, two empirical applications of our approach, to a sample of French exports to Germany and to a dataset of sales and markups for Indian firms, suggest that the choice between Pareto and lognormal distributions is less important than the choice between CREMR and other demands.

While we have concentrated on explaining the distributions of firm sales and markups given assumptions about the distribution of firm productivity, it is clear that our approach has many other potential applications. Linking observed heterogeneity of outcomes to underlying heterogeneity of agents’ characteristics via an assumed model of agent behavior is a common research strategy in many fields of economics. Both our exact results and our approach to measuring the information cost of incorrect assumptions about behavior should
prove useful in many other contexts.
Appendices

A Proof of Proposition 1

To show that (1) and (3) imply (2), let \( \tilde{F}(y) \) denote the distribution of \( y \) implied by (1) and (3). Since \( v \) is strictly increasing from (3), we have \( y = v^{-1}(x) \). Therefore the CDF of \( x \) is \( \tilde{F}[v^{-1}(x)] \). By Assumption 1, it has to coincide with \( G \) so:

\[
\tilde{F}[v^{-1}(x)] = G(x) \quad \forall x \in (x, \bar{x})
\] (18)

Therefore, \( \tilde{F}(y) = G[v(y)] \), which is the function assumed in (2), as was to be proved. A similar proof shows that (2) and (3) imply (1).

Next, we wish to prove that (1) and (2) imply (3). We start by picking an arbitrary firm \( i \) with characteristics \( x(i) \) and \( y(i) \). Because \( x(i) \) and \( y(i) \) are strictly increasing in \( i \), the fraction of firms with characteristics below \( x(i) \) and, respectively, \( y(i) \), are equal:

\[
G[x(i)] = F[y(i)] \quad \forall i \in \Omega
\] (19)

Inverting gives \( x(i) = G^{-1}[F(y(i))] \). Since this holds for any firm \( i \in \Omega \), it follows that \( x = v(y) = G^{-1}[F(y)] \), as required.

B Generalized Power Function Distributions

Table 6 shows that many well-known distributions are members of the Generalized Power Function family, \( G(x; \theta) = H\left(\theta_0 + \frac{\theta_1}{\theta_2} x^{\theta_2}\right) \), introduced in Definition 1. Hence Proposition 2 can immediately be applied to deduce a constant-elasticity relationship between any two firm characteristics which share any of the distributions in the table, provided the two distributions have compatible supports, and the same value of the parameter \( \theta_0 \).
Table 6: Some Members of the Generalized Power Function Family of Distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>PDF</th>
<th>Support</th>
<th>CDF</th>
<th>( \theta_0 )</th>
<th>( \theta_1 )</th>
<th>( \theta_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pareto</td>
<td>( 1 - \left( \frac{z}{\theta} \right)^{-k} )</td>
<td>([x, \infty))</td>
<td>( z )</td>
<td>1</td>
<td>( k )</td>
<td>(-k)</td>
</tr>
<tr>
<td>Truncated Pareto</td>
<td>( \frac{1-x^{-k}}{1-x^{-k}} )</td>
<td>([x, \bar{x}])</td>
<td>( z )</td>
<td>( \frac{1}{1-x^{-k}} )</td>
<td>( \frac{kx^k}{1-x^{-k}} )</td>
<td>(-k)</td>
</tr>
<tr>
<td>Lognormal</td>
<td>( \Phi \left( \frac{\log x - \mu}{s} \right) )</td>
<td>([0, \infty))</td>
<td>( \Phi \left[ \log (z) \right] )</td>
<td>0</td>
<td>( \frac{1}{s} )</td>
<td>( \frac{1}{s} )</td>
</tr>
<tr>
<td>Uniform</td>
<td>( \frac{x-x}{x-\bar{x}} )</td>
<td>([x, \bar{x}])</td>
<td>( z )</td>
<td>( \frac{x-x}{x-\bar{x}} )</td>
<td>( \frac{1}{x-\bar{x}} )</td>
<td>1</td>
</tr>
<tr>
<td>Fréchet</td>
<td>( \exp \left[ - \left( \frac{x-\mu}{s} \right)^{-\alpha} \right] )</td>
<td>([\mu, \infty))</td>
<td>( \exp \left[ -z^{-\alpha} \right] )</td>
<td>(-\frac{\mu}{s})</td>
<td>( \frac{1}{s} )</td>
<td>1</td>
</tr>
<tr>
<td>Gumbel</td>
<td>( \exp \left[ - \exp \left{ - \left( \frac{x-\mu}{s} \right) \right} \right] )</td>
<td>( (-\infty, \infty))</td>
<td>( \exp \left[ - \exp \left{ -z \right} \right] )</td>
<td>(-\frac{\mu}{s})</td>
<td>( \frac{1}{s} )</td>
<td>1</td>
</tr>
<tr>
<td>Reversed Weibull</td>
<td>( \exp \left[ - \left( \frac{\mu-x}{s} \right)^{-\alpha} \right] )</td>
<td>( (-\infty, \mu))</td>
<td>( \exp \left[ -z^{\alpha} \right] )</td>
<td>( \frac{\mu}{s} )</td>
<td>(-\frac{1}{s} )</td>
<td>1</td>
</tr>
</tbody>
</table>

A simple example of a distribution which is not a member of the GPF family is the exponential: \( G(x; \theta) = 1 - \exp(-\lambda x) \). This one-parameter distribution does not have the flexibility to match either the sufficiency or the necessity part of Proposition 2. If \( x \) is distributed as an exponential and \( x = x_0 y^E \), then \( y \) is distributed as a Weibull: \( F(y; \theta') = 1 - \exp(-\lambda x_0 y^E) \). Whereas if both \( x \) and \( y \) are distributed as exponentials, then \( x = x_0 y \), i.e., \( E = 1 \). For similar reasons, the one-parameter version of the Fréchet (used in Eaton and Kortum (2002)) is not a member of the GPF family, though as Table 6 shows, both its two-parameter version (used in many applications of the Eaton-Kortum model) and the three-parameter “Translated Fréchet” (with one of the parameters set equal to \( \theta_0 \)) can be written as members of the family.

### C Proof of Proposition 2

To show that (1) and (3) imply (2), assume \( G(x; \theta) = H \left( \theta_0 + \frac{\theta_1}{\theta_2} x^\theta_2 \right) \), \( G_x > 0 \), and \( x = x_0 h(y)^E \). Then the implied distribution of \( y \) is:

\[
F(y; \theta) = H \left[ \theta_0 + \frac{\theta_1}{\theta_2} \left\{ x_0 h(y)^E \right\}^{\theta_2} \right] = H \left[ \theta_0 + \frac{\theta_1'}{\theta_2'} h(y)^{\theta_2'} \right]
\]
where: $\theta'_2 = E\theta_2$ and $\frac{\theta'_1}{\theta'_2} = \frac{\theta_1}{\theta_2}x^2$ so $\theta'_1 = \frac{\theta_1}{\theta_2}\theta'_2x^2 = E\theta_1x^2$. Thus (1) and (3) imply (2). A similar proof shows that (2) and (3) imply (1).

Next, to show that (1) and (2) imply (3), assume $G(x; \theta) = H\left(\theta_0 + \frac{\theta_1}{\theta_2}x^2\right)$, $G_x > 0$, and $F(y; \theta') = H\left(\theta_0 + \frac{\theta_1}{\theta_2}h(y)^{\theta_2}\right)$, $F_y > 0$. From part (ii) of Proposition 1, $x = G^{-1}[F(y; \theta') ; \theta]$. Inverting $G(x; \theta)$ gives $\theta_0 + \theta'_1x^2 = \frac{1}{\theta_1}\left\{H^{-1}(G(x; \theta)) - \theta_0\right\}$. Now substitute $F(y; \theta')$ for $G(x; \theta)$:

$$x = \left[\frac{\theta_2}{\theta_1}\left\{H^{-1}(G(x; \theta)) - \theta_0\right\}\right]^{\frac{1}{\theta_2}} = x_0h(y)^E$$

where: $E = \frac{\theta'_2}{\theta_2}$ and $x_0 = \left(\frac{\theta_2}{\theta_1}\right)^{\frac{1}{\theta_2}} = \left(\frac{1}{\theta_1}\right)^{\frac{1}{\theta_2}}$. Thus (1) and (2) imply (3).

### D Properties of CREMR Demand Functions

First, we wish to show that the CREMR property $\varphi = (r')^{-1} = \varphi_0^E$ is necessary and sufficient for the CREMR demands given in (4). To prove sufficiency, note that, from (4), total and marginal revenue are:

$$r(x) \equiv xp(x) = \beta(x - \gamma)\frac{x^{-1}}{\sigma} \quad r'(x) = p(x) + xp'(x) = \beta\frac{\sigma - 1}{\sigma}(x - \gamma)^{-\frac{1}{\sigma}}$$

Combining these gives:

$$r'(x) = \beta\frac{\sigma - 1}{\sigma} \frac{\sigma - 1}{\sigma} r(x)^{-\frac{1}{\sigma - 1}}$$

Hence, the revenue elasticity of marginal revenue is indeed constant, equal to $\frac{1}{\sigma - 1}$. For later use it is also useful to express these equations in terms of proportional changes (where a circumflex denotes a logarithmic derivative, so $\hat{r} \equiv \frac{dr}{r}$, $r > 0)$:

$$\begin{align*}
\hat{r} &= \frac{\sigma - 1}{\sigma} \frac{x}{x - \gamma} \hat{x} \\
\hat{r}' &= -\frac{1}{\sigma} \frac{x}{x - \gamma} \hat{x}
\end{align*}$$

$$\Rightarrow \hat{r}' = -\frac{1}{\sigma - 1} \hat{r}$$
To prove necessity, invert equation (3) to obtain \( r'(x) = \varphi_0^{-1} r(x)^{-E} \). This is a standard first-order differential equation in \( r(x) \) with constant coefficients. Its solution is:

\[
r(x) = \left[ (E + 1) (\varphi_0^{-1} x - \kappa) \right]^{\frac{1}{E+1}}
\]  

(25)

where \( \kappa \) is a constant of integration. Collecting terms, recalling that \( r(x) = xp(x) \), gives the CREMR demand system (4), where the coefficients are: \( \sigma = \frac{E+1}{E} \), \( \beta = (E+1)^{\frac{1}{E+1}} \varphi_0^{-\frac{1}{E+1}} \), and \( \gamma = \varphi_0 \kappa \). Note that it is the constant \( \kappa \) which makes CREMR more general than CES. Since the CREMR property \( \varphi = (r')^{-1} = \varphi_0 r^E \) is both necessary and sufficient for the demands given in (4), we call the latter CREMR demands.

Next, we wish to derive the demand manifold for CREMR demand functions. Mrázová and Neary (2013) show that, for a firm with constant marginal cost facing an arbitrary demand function, the elasticities of total and marginal revenue with respect to output can be expressed in terms of the elasticity and convexity of demand. Combining their results leads to an expression for the revenue elasticity of marginal revenue which holds for any demand function:

\[
\hat{r} = \frac{\varepsilon - 1}{\varepsilon} \hat{x} \quad \hat{r}' = -\frac{2 - \rho}{\varepsilon - 1} \varepsilon \hat{r}
\]  

(26)

Equating this to (24) leads to the CREMR demand manifold in the text, equation (5). Note that requiring marginal revenue to be positive (\( \varepsilon > 1 \)) and decreasing (\( \rho < 2 \)) implies that \( \sigma > 1 \), just as in the familiar CES case.

To establish conditions for demand to be superconvex, we solve for the points of intersection between the demand manifold and the CES locus, the boundary between the sub- and superconvex regions. From Mrázová and Neary (2013), the expression for the CES locus is: \( \rho = \frac{\varepsilon + 1}{\varepsilon} \). Eliminating \( \rho \) using the CREMR demand manifold (5) and factorizing gives:

\[
\rho - \frac{\varepsilon + 1}{\varepsilon} = -\frac{(\varepsilon - \sigma)(\varepsilon - 1)}{(\sigma - 1)\varepsilon} = 0
\]  

(27)
Given $1 < \sigma \leq \infty$, this expression is zero, and so every CREMR manifold intersects the CES locus, at two points. One is at $\{\varepsilon, \rho\} = \{1, 2\}$, implying that all CREMR demand manifolds must pass through the Cobb-Douglas point. The other is at $\{\varepsilon, \rho\} = \{\sigma, 1 + \frac{1}{\sigma}\}$. Hence every CREMR demand manifold lies strictly within the superconvex region (where $\rho > \frac{(1+\sigma)}{\varepsilon}$) for $\sigma > \varepsilon > 1$, and strictly within the subconvex region for $\varepsilon > \sigma$. The condition for superconvexity, $\varepsilon \leq \sigma$, can be reexpressed in terms of $\gamma$ by using the fact that the elasticity of demand is $\varepsilon = \frac{x - \gamma}{x - \gamma x}$ $\sigma$. Substituting and recalling that $\sigma$ must be strictly greater than one, we find that CREMR demands are superconvex if and only if $\gamma \leq 0$. As with many other demand manifolds considered in Mrázová and Neary (2013), this implies that, for a given value of $\sigma$, the demand manifold has two branches, one in the superconvex region corresponding to negative values of $\gamma$, and the other in the subconvex region corresponding to positive values of $\gamma$. Along each branch, the equilibrium point converges towards the CES locus as output rises without bound, as shown by the arrows in Figure 4.

Similarly, to establish conditions for profits to be supermodular, we solve for the points of intersection between the demand manifold and the SM locus, the boundary between the sub- and supermodular regions. From Mrázová and Neary (2013), the expression for the SM locus is: $\rho = 3 - \varepsilon$. Eliminating $\rho$ using the CREMR demand manifold and factorizing gives:

$$\rho + \varepsilon - 3 = \frac{[(\sigma - 2)\varepsilon + 1](\varepsilon - 1)}{(\sigma - 1)\varepsilon} = 0$$

Once again, this expression is zero at two points: the Cobb-Douglas point $\{\varepsilon, \rho\} = \{1, 2\}$, and the point $\{\varepsilon, \rho\} = \{\frac{1}{2 - \sigma}, \frac{5 - 3\sigma}{2 - \sigma}\}$. The latter is in the admissible region only for $\sigma < 2$. Hence for $\sigma \geq 2$, the CREMR demand manifold is always in the supermodular region.

### E CREMR Preferences

We seek a specification of preferences that rationalizes CREMR demands. One way of doing this is to assume additively separable preferences, $U = \int_{i \in \Omega} u(x(i)) \, di$, which implies that
\( p(i) = \lambda^{-1} u'(x(i)) \), where \( \lambda \) is the marginal utility of income. Integrating the CREMR demand function (4), we can solve for the sub-utility \( u\{x(i)\} \), which is a hypergeometric function:

\[
u\{x(i)\} = \beta \frac{\sigma}{1 - \sigma} x(i) \frac{\sigma - 1}{\sigma} 2F_1\left(-\frac{\sigma - 1}{\sigma}, -\frac{\sigma - 1}{\sigma}; \frac{\gamma}{\sigma}; \frac{x(i)}{x(i)}\right) + \kappa
\]  

(29)

Here \( 2F_1(a, b; c; z) \), \(|z| < 1\), is the Gaussian hypergeometric function:

\[
2F_1(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n z^n}{(c)_n n!}
\]

(30)

and \((q)_n\) is the (rising) Pochhammer symbol:

\[
(q)_n = \begin{cases} 
1 & n = 0 \\
q(q + 1)...(q + n - 1) & n > 0
\end{cases}
\]

(31)

When \( \gamma \) is zero, (29) reduces to the CES utility function, \( u\{x(i)\} = \beta \frac{\sigma}{1 - \sigma} x(i) \frac{\sigma - 1}{\sigma} + \kappa \); when \( \gamma \) is positive, so demands are subconvex, utility is positive; and conversely when \( \gamma \) is negative.

Setting \( \kappa \), the constant of integration in (29), equal to zero implies that \( u(0) = 0 \). In this case, the utility function always exhibits a taste for diversity. To see this, note that \( u(x) \) must be increasing (since otherwise \( p(x) \) would not be positive) and concave (since otherwise \( p(x) \) would not be decreasing in \( x \)). Any concave and differentiable function \( u(x) \) is bounded above by its Taylor approximation: \( u(x_0) \leq u(x) + (x_0 - x) u'(x) \). Setting \( x_0 = 0 \) and using the fact that \( u(0) = 0 \) implies that \( 0 \leq u(x) - xu'(x) \). Hence the elasticity of utility \( \xi(x) \equiv \frac{xu'(x)}{u(x)} \) is always less than one. This in turn implies a taste for diversity in the sense that fixing total consumption \( X \equiv nx \), where \( x \) is the same for all varieties and \( n \) is the measure of varieties, implies that \( U = nu(x) = nu\left(\frac{X}{n}\right) \). Logarithmically differentiating with respect to \( n \) yields: \( \dot{U} = (1 - \xi)\dot{n} \).

46
F Proofs of Corollaries 1, 2, 3, and 4

Corollaries 1 and 2 (Productivity and Sales with Pareto or Lognormal):

Proposition 2 holds for any distribution in the generalized power function class. The particular solutions for the constant terms in equations (6) and (7) are derived by substituting the parameters of the Pareto and lognormal distributions into the relevant expressions in Proposition 2. Finally, as discussed in the text, all members of the CREMR class with non-zero $\gamma$ (i.e., non-zero $\kappa$) are, strictly speaking, inconsistent with a lognormal distribution, since they imply that the smallest firm has strictly positive sales revenue.

Corollaries 3 and 4 (Productivity and Output with Pareto or Lognormal):

In these cases, Proposition 2 implies that productivity must be a simple power function of output: $\varphi = \varphi_0 x^E$. Replacing $\varphi$ by $r'(x)^{-1}$ as before yields a new differential equation in $r(x)$, with solution:

$$r(x) = \varphi_0^{-1} \frac{x^{1-E}}{1-E} + \kappa$$

where $\kappa$ is once again a constant of integration. This is the CEMR demand system (8), where $\sigma = \frac{1}{E}$ and $\beta = \frac{1}{\varphi_0(1-E)}$. The final step, as in the case of Corollaries 1 and 2, is to solve for the constant terms when the distributions are either Pareto or lognormal.

G Derivations Underlying Table 1

As in Mrázová and Neary (2013), we give the demand functions from a “firm’s-eye view”; many of the parameters taken as given by the firm are endogenous in industry and general equilibrium. For each demand function, we follow a similar approach to that used with CREMR demands in Sections 3.1 and 4.1: we use the first-order condition to solve for productivity as a function of either output or price; the definition of sales revenue to solve for output or price as a function of sales; and the relationship between markups and elasticities to solve for either output or price as a function of the markup. Combining yields $\varphi(r)$ and
\( \varphi(m) \) as required.

**Linear:** \( p(x) = \alpha - \beta x, \alpha > 0, \beta > 0 \). Sales revenue is quadratic in output, \( r(x) = \alpha x - \beta x^2 \), but only the root corresponding to positive marginal revenue, \( r'(x) = \alpha - 2\beta x > 0 \), is admissible. Since maximum output is \( \bar{x} = \frac{\alpha}{2\beta} \), maximum sales revenue is \( \bar{r} = \frac{\alpha^2}{4\beta} \), and we work with sales relative to their maximum: \( \tilde{r} \equiv \frac{r}{\bar{r}} \). Hence output as a function of relative sales is: \( x(\tilde{r}) = \frac{\alpha}{2\beta} \left[ 1 - (1 - \tilde{r})^{\frac{1}{2}} \right] \). Equating marginal revenue to marginal cost gives \( \varphi(x) = \frac{1}{\alpha - 2\beta x} \).

Finally, the elasticity of demand is \( \varepsilon(x) = \frac{\alpha - \beta x}{\beta x} \), so the markup as a function of output is \( m(x) = \frac{\alpha - \beta x}{\alpha - 2\beta x} \). We do not work with the relative markup in this case, since \( m(x) \to \infty \) as \( x \to \bar{x} \). Inverting \( m(x) \) gives \( x(m) = \frac{\alpha}{\beta} \frac{m - 1}{2m - 1} \).

**LES:** \( p(x) = \frac{\delta}{x + \gamma}, \gamma > 0, \delta > 0 \). We use the inverse demand function rather than the more familiar direct one: \( x(p) = \frac{\delta}{p} - \gamma \). Note that, in monopolistic competition, the second-order condition requires that \( \gamma \) be positive, so its usual interpretation as (minus) a subsistence level of consumption is not admissible. Sales revenue is \( r(x) = \frac{\delta x}{x + \gamma} \), attaining its maximum at \( \bar{r} = \delta \), so we work with relative sales: \( \tilde{r} \equiv \frac{r}{\bar{r}} = \frac{x}{x + \gamma} \). Inverting gives: \( x(\tilde{r}) = \gamma \frac{1}{1 - \tilde{r}} \). The first-order condition yields: \( \varphi(x) = \frac{(x + \gamma)^2}{\gamma \delta} \). Finally, the elasticity of demand is \( \varepsilon(x) = \frac{x + \gamma}{\gamma} \), so the markup as a function of output is \( m(x) = \frac{x + \gamma}{\gamma} \); inverting gives \( x(m) = \gamma(m - 1) \).

**Translog:** \( x(p) = \frac{1}{p} (\gamma - \eta \log p), \gamma > 0, \eta > 0 \). From the direct demand function, sales revenue as a function of price is \( r(p) = \gamma - \eta \log p \), which when inverted gives \( p(r) = \exp \left( \frac{\gamma - r}{\eta} \right) \). From the first-order condition, \( \varphi(p) = \frac{x'(p)}{r'(p)} = \frac{\eta + \gamma - \eta \log p}{\eta p} \). Combining this with \( p(r) \) gives the expression for \( \varphi(r) \) in Table 1, with: \( \varphi_0 = \frac{1}{\exp \left( \frac{\gamma}{\eta} \right)} \). Finally, the elasticity of demand is \( \varepsilon(p) = \frac{\eta + \gamma - \eta \log p}{\gamma - \eta \log p} \), so the markup as a function of price is \( m(p) = \frac{\eta + \gamma - \eta \log p}{\eta} \); inverting gives \( p(m) = \exp \left( \frac{\eta + \gamma}{\eta} - m \right) \).
H The Kullback-Leibler Divergence

H.1 Information-Theoretic Foundations of the KLD

The starting point of information theory is an axiomatic basis for a quantitative measure of the information content of a single draw from a known distribution \( F(r) \).\(^{27}\) It is natural that a measure of information should be additive, non-negative, and inversely related to the probability of the draw. The only function satisfying these requirements is minus the log of the probability: \( I(r) = -\log(f(r)) \).\(^{28}\) This in turn leads to the concept of the Shannon entropy of \( F(r) \), which is the expected value of information from a single draw:\(^{29}\)

\[
S_F \equiv E[I(r)] = - \int_{\underline{r}}^{\bar{r}} \log(f(r)) f(r) dr
\] (33)

(See Shannon (1948).) Intuitively, Shannon entropy can be thought of as a measure of the unpredictability or uncertainty about an individual draw implied by the known distribution \( F(r) \). In general it ranges from zero to infinity. It equals zero when \( F(r) \) is a Dirac distribution with all its mass concentrated at a single point: in this case, knowing the distribution tells us everything about individual draws, so an extra draw conveys no new information. By contrast, Shannon entropy can be arbitrarily large when \( F(r) \) is a uniform distribution:

\[
F(r) = \frac{r - \underline{r}}{\bar{r} - \underline{r}}, \ r \in [\underline{r}, \bar{r}] \quad \Rightarrow \quad S_F = S_{\text{Uniform}} = \log(\bar{r} - \underline{r})
\] (34)

\(^{27}\)See Cover and Thomas (2012) for an introduction to information theory. Previous applications of Shannon entropy to economics include the work on inequality by Theil (1967), and the theory of rational inattention developed by Sims (2003), and applied to international trade by Dasgupta and Mondria (2014). Applications of the KLD to economics include Vuong (1989), Cameron and Windmeijer (1997) and Ullah (2002) in econometrics, Adams (2013) in empirical demand analysis, and Galle, Rodríguez-Clare, and Yi (2017) in international trade.

\(^{28}\)In information theory it is customary to take all logarithms to base 2, so information is measured in bits. For some theoretical results it is more convenient to use natural logarithms, though most results hold irrespective of the logarithmic base used.

\(^{29}\)Shannon entropy was first introduced for discrete distributions. The application to continuous distributions is also called “differential entropy.”
In this case, knowing the distribution conveys no information whatsoever about individual draws, so, as the upper bound \( \tau \) becomes arbitrarily large, the same happens to Shannon entropy.

While Shannon entropy measures the expected information gain conveyed by a draw from a single distribution, the KLD measures the information loss when one distribution is used to approximate another one, typically the one observed in the data. Formally, if \( F(r) \) is the observed c.d.f. of firms’ sales, and \( \tilde{F}(r) \) is the distribution used to approximate \( F(r) \), then the KLD is defined as follows:

\[
D_{KL}(F||\tilde{F}) = \int_{\tau} \log \left( \frac{f(r)}{\tilde{f}(r)} \right) f(r)dr
\]

To get some intuition for the KLD, it is helpful to rewrite it as follows:

\[
D_{KL}(F||\tilde{F}) = -\int_{\tau} \log \left( \tilde{f}(r) \right) f(r)dr - S_F
\]

The first term on the right-hand side of (36) measures the cross-entropy between \( F(r) \) and \( \tilde{F}(r) \). Intuitively, this is a measure of the unpredictability of an individual draw from the benchmark distribution \( F(r) \) implied by the tested distribution \( \tilde{F}(r) \). Equation (36) thus shows that the KLD equals the difference between the cross-entropy and Shannon entropy. Heuristically, it can be interpreted as the “excess” unpredictability of \( F(r) \) implied by \( \tilde{F}(r) \) relative to the unpredictability of \( F(r) \) implied by itself; or as the informativeness of a draw from \( \tilde{F}(r) \) relative to one from \( F(r) \). The KLD also has a statistical interpretation: it equals the expected value of the log likelihood ratio, so choosing the parameters of a distribution to minimize KLD is equivalent to maximizing the likelihood of the sample. By Gibbs’ inequality, the KLD is always non-negative, \( D_{KL}(F||\tilde{F}) \geq 0 \), and attains its lower bound of zero if and only if \( F(r) = \tilde{F}(r) \) almost everywhere, when the distribution \( \tilde{F}(r) \) is completely informative about \( F(r) \). As for its upper bound, the KLD value is unbounded unlike Shannon entropy. However, as discussed in the text, we take its value when \( \tilde{F} \) is uniform as a benchmark for
a “reasonable” fit. This is given by:

$$\mathcal{D}_{KL}(F||F_{Uniform}) = \log(\bar{r} - \underline{r}) - S_F = S_{Uniform} - S_F$$  \hfill (37)$$

where the second equality follows from (35).

A number of qualifications need to be kept in mind when we use the KLD as a measure of the “closeness” of two distributions. First, the KLD is not symmetric with respect to both distributions: $\mathcal{D}_{KL}(\tilde{F}||F) \neq \mathcal{D}_{KL}(F||\tilde{F})$. Formally, the KLD is a pre-metric, not a metric, and it does not satisfy the triangle inequality. In our application, this does not pose a problem, since it is natural to take the actual firm size distribution as a benchmark, whether it comes from theory or from empirical observation. The role of the KLD is then to quantify how well different candidate methods of calculating a distribution $\tilde{F}(r)$ approximate the “true” distribution $F(r)$: it measures the divergence of $\tilde{F}(r)$ from $F(r)$, not the distance between them.

Second, for the KLD to be well defined, the tested distribution $\tilde{F}(r)$ must have a strictly positive density, $\tilde{f}(r) > 0$, at every point in $[\underline{r}, \bar{r}]$.\(^{30}\) In principle, this can pose problems when we wish to compare a distribution implied by a demand function (such as the linear) that implies a saturation consumption level with an unbounded distribution such as the Pareto or lognormal. This is not a problem in practical applications, however, since we can always calibrate demand to fit the upper limit of the observed values of $F(r)$. Even in theoretical contexts, it is an advantage rather than a disadvantage in our context, since it leads us to consider right-truncated distributions. This is a particularly desirable direction to explore in the light of Feenstra (2014), who shows that, without truncation, a Pareto distribution does not allow us to distinguish between the product-variety and pro-competitive gains from trade.

Third, the KLD, like Shannon entropy, attaches the same weight to all observations. In a

\(^{30}\)The converse is not needed, since by convention $\lim_{f(r) \to 0} f(r) \log (f(r)) = 0.$
heterogeneous-firms context, we may be more interested in explaining the behavior of large firms, which account for a disproportionate share of total production and exports. One way of implementing this would be to calculate a “weighted KLD”, where higher weights are attached to larger firms. A more direct approach is to see how the KLD behaves as we drop more observations on smaller firms: we pursue this in Section 6.3.

H.2 Decomposing the KLD

Because our main focus is on comparing an observed distribution with one predicted by a model, it is helpful to relate the KLD to the elasticities of density of the two underlying distributions. To do this we use integration by parts. First, rewrite the definition of Shannon entropy in (33) as \( \int u dv \), where \( u \equiv \log f(r) \), so \( du = \frac{f'(r)}{f(r)} dr \), and \( dv \equiv f(r) dr \), so \( v = F(r) + C \), where \( C \) is an arbitrary constant of integration. Integrate by parts:

\[
S_F = - (1 + C) \log f(\bar{r}) + C \log f(\bar{r}) + \int_{\bar{r}}^{\bar{r}} \frac{F(r) + C r f'(r)}{r} \frac{dr}{f(r)}
\]

Setting \( C \) equal to \(-1\) gives:

\[
S_F = - \log f(\bar{r}) - \int_{\bar{r}}^{\bar{r}} \frac{1 - F(r)}{r} \frac{r f'(r)}{f(r)} dr
\]

This shows that Shannon entropy can be decomposed into two terms. The first is the information content of the lower limit of the distribution, i.e., in our application, the information content of the marginal firms. The second equals the integral of the elasticity of the density, \( \frac{r f'(r)}{f(r)} \), times the relative survival function, \( \frac{1 - F(r)}{r} \). The latter is declining in sales, so, when written in this way, Shannon entropy attaches more weight to the elasticities of density of

\[31\text{For a discrete version of such a measure, called a “quantitative-qualitative measure of relative information,” see Taneja and Tuteja (1984) and Kvåleseth (1991). A more satisfactory alternative is the generalization of KLD known as the Rényi divergence of order } \alpha, \alpha \geq 0 \text{ (see Rényi (1959)): } D_{\alpha}(F||\tilde{F}) \equiv \frac{1}{\alpha-1} \log \left( \int_{\bar{r}}^{\bar{r}} \frac{f(r)^{\alpha}}{f(r)^{\alpha-1}} dr \right). \text{ The KLD is the limiting case of this as } \alpha \to 1: D_1(F||\tilde{F}) = D_{KL}(F||\tilde{F}). \text{ For values of } \alpha \text{ between zero and one, the Rényi divergence weights all possible draws more equally than the KLD, regardless of their probability.} \]
larger firms.\footnote{The rate at which the relative survival function declines is one plus the proportional hazard rate: $d \log \left[ \frac{1 - F(r)}{r} \right] = - \left( 1 + \frac{rf'}{1 - F} \right) d \log r$.} If instead we set $C$ in (38) equal to zero rather than one, we get an alternative decomposition expressed in terms of the upper bound:

$$S_F = - \log f(\tau) + \int_\tau^\infty \frac{F(r) r f'(r)}{f(r)} dr$$  \hspace{1cm} (40)

Now the first term is the information content of the upper limit of the distribution. However, this is less useful than (39) for our purposes, since, for many distributions, including the Pareto and the lognormal, $\log f(b) = -\infty$.

Repeating this process for the KLD gives in a similar fashion two alternative decompositions, one expressed in terms of the lower bounds of the distribution:

$$D_{KL}(F||\tilde{F}) = \log f(\underline{r}) - \log \tilde{f}(\underline{r}) + \int_{\underline{r}}^\tau \frac{1 - F(r)}{r} \left[ rf'(r) - r\tilde{f}'(r) \right] dr$$  \hspace{1cm} (41)

and the other in terms of the upper bounds:

$$D_{KL}(F||\tilde{F}) = \log f(\bar{r}) - \log \tilde{f}(\bar{r}) - \int_{\underline{r}}^\tau \frac{F(r)}{r} \left[ rf'(r) - r\tilde{f}'(r) \right] dr$$  \hspace{1cm} (42)

Again we concentrate on the first of these, which, as before, can be decomposed into two terms. The first is the difference between the information contents of the lower limits of the two distributions. The second equals the integral of the difference between their elasticities of density, times the relative survival function, $\frac{1 - F(r)}{r}$. Recalling that the latter is declining in sales shows that the KLD attaches less weight to underestimates of the elasticity of density of larger firms.

The decomposition of the KLD in (41) proves particularly insightful when the predicted size distribution is derived from an underlying model of firm behavior. As in Section 3, this comes from a distribution of firm productivity $g(\varphi)$ and a model that links productivity to sales via a function $\varphi(r)$. From the standard result on densities of transformed variables
(part (i) of Proposition 1), we can relate the density of the derived distribution of sales to the density of the underlying distribution of firm productivity: \( \tilde{f}(r) = g(\varphi(r)) \frac{d\varphi}{dr} \). Totally differentiating this gives an expression in terms of elasticities:

\[
\frac{r\tilde{f}'(r)}{\tilde{f}(r)} = \frac{\varphi(r)g'(\varphi(r))}{g(\varphi(r))} \frac{r\varphi'(r)}{\varphi(r)} + \frac{r\varphi''(r)}{\varphi'(r)}
\] (43)

We can relate the second term to the elasticity of marginal revenue with respect to total revenue, \( E(r) \equiv \frac{r\varphi'(r)}{\varphi(r)} \):

\[
\frac{r\varphi''(r)}{\varphi'(r)} = E(r) - 1 + \frac{rE'(r)}{E(r)}
\] (44)

(See Lemma 5 in Mrázová and Neary (2013).) Substituting into (43), the density elasticity of the derived sales distribution \( \tilde{F}(r) \) can be written in terms of underlying elasticities as follows:

\[
\frac{r\tilde{f}'(r)}{\tilde{f}(r)} = \left[ \frac{\varphi(r)g'(\varphi(r))}{g(\varphi(r))} + 1 \right] E(r) - 1 + \frac{rE'(r)}{E(r)}
\] (45)

Substituting this into (41) gives the full decomposition of the KLD in equation (15) in the text. When \( G(\varphi) \) is Pareto, so \( G(\varphi) = 1 - (\frac{\varphi}{\tilde{n}})^{-k} \), the elasticity of density is \( \frac{\varphi g'(\varphi)}{g(\varphi)} = -(1+k) \). Hence (45) simplifies to the following:

\[
\frac{r\tilde{f}'(r)}{\tilde{f}(r)} = -[kE(r) + 1] + \frac{rE'(r)}{E(r)}
\] (46)

H.3 Quantifying the Information Cost of Incorrect Assumptions

To illustrate the application of the KLD decomposition in equation (15), we show its implications in the benchmark case where both productivity and sales have a Pareto distribution, and demands are of the CREMR type. This eliminates the third source of information loss in (15), since \( E' = 0 \). However, this does not mean that a perfect calibration is guaranteed, as we shall see.

When \( \tilde{F} \) and \( F \) are both Pareto with parameters \( \tilde{n} \) and \( n \), the KLD can be calculated...
from equation (41):
\[ D_{KL}(F||\tilde{F}) = \log \frac{n}{\tilde{n}} + \frac{\tilde{n}}{n} - 1 \] (47)

To relate this to primitive parameters, recall from Section 3 that with CREMR demands the elasticity of marginal revenue with respect to total revenue, \( E \), equals \( \frac{1}{\sigma - 1} \), and so, with a Pareto distribution of productivity, the shape parameter for the derived distribution of sales is \( \tilde{n} = Ek = \frac{k}{\sigma - 1} \). Substituting into (47) gives the KLD decomposition, equation (15), in the Pareto-CREMR case:
\[ D_{KL}(F||\tilde{F}) = \log \frac{n}{k} + \log(\sigma - 1) + \frac{k}{n \sigma - 1} - 1 \] (48)

The first derivative of this with respect to \( \sigma \) is:
\[ \frac{dD_{KL}}{d\sigma} = (\sigma - 1)^{-1} \left( 1 - \frac{n}{\tilde{n}} \right) = (\sigma - 1)^{-2} \left( \sigma - \frac{k+n}{n} \right). \]
This is positive, and so \( D_{KL} \) is increasing, if and only if \( \sigma \geq \frac{k+n}{n} \). The second derivative is:
\[ \frac{d^2D_{KL}}{d\sigma^2} = -\left( \sigma - 1 \right)^{-2} \left( 1 - 2\frac{n}{\tilde{n}} \right) = -\left( \sigma - 1 \right)^{-3} \left( \sigma - \frac{2k+n}{n} \right). \] This is negative, and so \( D_{KL} \) is concave, if and only if \( \sigma \geq \frac{2k+n}{n} \). Equation (48) is illustrated in Figure 12 as a function of \( \sigma \), drawn for values of \( k = 1 \) and \( n = 2 \).

Figure 12 shows clearly that the information cost of using the “wrong” estimate of \( \sigma \) is highly asymmetric. For given values of \( k \) and \( n \), the true value of \( \sigma \) equals \( \frac{k+n}{n} \). (Recall equation (6).) Given the assumed values of \( k \) and \( n \), this equals 1.5, which is the value of \( \sigma \) at
Figure 13: Decomposition of KLD in the Pareto-CREMR Case

which the KLD is minimized. For other values, it is much more sensitive to underestimates than to overestimates of the true value of $\sigma$. Why this is so is shown from two different perspectives in Figure 13. Panel (a) shows the two components of the KLD from (48), while panel (b) shows how a higher assumed value of $\sigma$ affects the location of the predicted distribution relative to that of the true distribution corresponding to $\sigma = 1.5$. Clearly, underestimating $\sigma$ means overestimating the mass of the smallest firms and underestimating the mass of the larger firms. From (48), the cost of the former is increasing in the log of $\sigma - 1$, whereas the cost of the latter is falling in the reciprocal of $\sigma - 1$. For values of $\sigma$ below 1.5, the second effect dominates: because the Pareto has an infinite tail, it is more important to fit the larger firms than the smaller ones. This is clear from panel (b), while the numerical values of the components of the KLD in panel (a) show explicitly how the gains and losses in information that come from an increase in $\sigma$ are traded off against each other.

A further implication of equation (48) is that, with Pareto productivity and CREMR demands, the KLD depends on only one of the three parameters in the CREMR demand function. Figure 12 applies equally well to the CES case (where the CREMR parameter $\gamma$ is zero) as it does to any other member of the CREMR class. This suggests a further role
for the CREMR family in calibrations. To calibrate the size distribution of firms, the only

demand parameter that is needed is \( \sigma \). Hence the values of the other parameters \( \beta \) and \( \gamma \) can

be chosen to match other features of the data: \( \gamma \) to match the size distribution of markups

across firms, and \( \beta \) to match the level of demand.

Figure 14: Components of KLD in the Truncated-Pareto-CREMR Case

It is also illuminating to see how the KLD extends to the case of CREMR demands

combined with a sales distribution that is a right-truncated Pareto, \( F(r) = \frac{1 - r^n r^{-n}}{1 - \bar{r}^{-n}} \), where

\( r \in [\underline{r}, \bar{r}] \). In the untruncated Pareto case, the KLD was independent of the lower bound of the

Pareto \( r \). This is no longer true, though it depends only on the ratio of the lower and upper

bounds: \( \lambda \equiv \frac{\bar{r}}{\underline{r}} \in [0, 1] \). (This reduces to zero in the untruncated case.) Straightforward

calculations give the extension of equation (47) to the truncated case:

\[
\mathcal{D}_{KL}(F||\tilde{F}) = \log \frac{n}{\tilde{n}} + \log \frac{1 - \lambda^{\tilde{n}}}{1 - \lambda^{n}} + \frac{\tilde{n} - n}{n} + (\tilde{n} - n) \frac{\lambda^n}{1 - \lambda^n} \log \lambda
\] (49)

Replacing some, though not all, occurrences of \( \tilde{n} \) by \( \frac{k - 1}{\sigma - 1} \) allows us to write this in terms of

primitives in a manner which parallels equation (48):

\[
\mathcal{D}_{KL}(F||\tilde{F}) = \log \frac{n}{k} + \log (\sigma - 1) + \log \frac{1 - \lambda^{\tilde{n}}}{1 - \lambda^{n}} + \frac{k}{n \sigma - 1} - 1 + (\tilde{n} - n) \frac{\lambda^n}{1 - \lambda^n} \log \lambda
\] (50)
This is illustrated in Figure 14 as a function of $\sigma$. The dashed loci repeat the KLD and its components for the untruncated case from Figure 13(a). The solid loci give the KLD and its components for the truncated case, assuming the same values of $k$ and $n$ as before and a value of $\lambda = 0.1$. It is clear that right-truncation makes no quantitative difference to the cost of underestimating $\sigma$. The main effect is to reduce the cost of mismeasuring the mass of the smallest firms when $\sigma$ is overestimated.

I Bootstrapped Comparisons on Indian Data

Tables 7 and 8 repeat for Indian sales and markup data respectively the bootstrapping comparisons presented for French exports data in Table 3. It is clear that the comparisons between different values of the KLD for Indian data shown in Table 4 and Figure 8 are just as robust as those for the French data shown in Table 2.

<table>
<thead>
<tr>
<th>CREMR + LN</th>
<th>CREMR + P</th>
<th>TLog + P</th>
<th>Lin + P</th>
<th>TLog + LN</th>
<th>Lin + LN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREMR + LN</td>
<td></td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>CREMR + P</td>
<td>100%</td>
<td></td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>TLog + P</td>
<td>100%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Lin + P</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>TLog + LN</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Lin + LN</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 7: Bootstrapped Robustness of the KLD Ranking: Indian Sales
(See text for explanation)

<table>
<thead>
<tr>
<th>CREMR + P</th>
<th>CREMR + LN</th>
<th>Lin + LN</th>
<th>LES + LN</th>
<th>LES + P</th>
<th>TLog + P</th>
<th>TLog + LN</th>
<th>Lin + P</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREMR + P</td>
<td></td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>CREMR + LN</td>
<td>98%</td>
<td></td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Lin + LN</td>
<td>100%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>LES + LN</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>LES + P</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>16%</td>
<td>6%</td>
<td>0%</td>
</tr>
<tr>
<td>TLog + P</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>84%</td>
<td>–</td>
<td>0%</td>
</tr>
<tr>
<td>TLog + LN</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>94%</td>
<td>100%</td>
<td>–</td>
</tr>
<tr>
<td>Lin + P</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 8: Bootstrapped Robustness of the KLD Ranking: Indian Markups
(See text for explanation)
Tables 9 and 10 give the expressions for the quantiles of the sales and markup distributions respectively that are implied by our assumptions about demand and the distribution of firm productivities. These expressions are used to calculate the entries in Figure 11.

<table>
<thead>
<tr>
<th>Demand</th>
<th>Pareto $\mathcal{P}(\xi, k)$</th>
<th>lognormal $\mathcal{LN}(\mu, s)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREMR</td>
<td>$\frac{1}{\bar{m}} (1 - y)^{-\frac{s-1}{k}}$</td>
<td>$\exp(\mu + s \cdot \Phi^{-1}[y])$</td>
</tr>
<tr>
<td>Translog</td>
<td>$\eta \cdot \left( W \left[ e \cdot (1 - y)^{-\frac{1}{k}} \right] - 1 \right)$</td>
<td>$\eta \cdot \left( W \left[ \exp\left( \frac{\gamma}{\eta} + 1 + s \cdot \Phi^{-1}[y] + \mu \right) \right] - 1 \right)$</td>
</tr>
<tr>
<td>Linear/LES</td>
<td>$\bar{r} \left( 1 - (1 - y)^{\frac{1}{k}} \right)$</td>
<td>$\bar{r} - \frac{\exp(-2(\mu+s \cdot \Phi^{-1}[y]))}{43}$</td>
</tr>
</tbody>
</table>

Table 9: Quantiles for Sales

- $\Phi[z]$: c.d.f. of a standard normal
- $W$: The Lambert function

<table>
<thead>
<tr>
<th>Demand</th>
<th>Pareto $\mathcal{P}(\xi, k)$</th>
<th>lognormal $\mathcal{LN}(\mu, s)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREMR</td>
<td>$\frac{\bar{m} \cdot (1 - y)^{-\frac{1}{k}}}{\bar{m} - 1 + (1 - y)^{-\frac{1}{k}}}$</td>
<td>$\bar{m} \left[ 1 + \exp\left( -\mu - s \cdot \zeta \left[ y; \frac{1}{s} \cdot \left( \ln\left( \frac{1}{\bar{m} - 1} \right) - \mu \right) \right] \right]^{-1}$</td>
</tr>
<tr>
<td>Translog</td>
<td>$W \left[ e \cdot (1 - y)^{-\frac{1}{k}} \right]$</td>
<td>$W \left[ \exp\left[ s \cdot \zeta \left[ y; -\frac{1}{s} \cdot \left( \frac{\gamma}{\eta} + \mu \right) \right] + \frac{\gamma}{\eta} + \mu + 1 \right] \right]$</td>
</tr>
<tr>
<td>Linear</td>
<td>$\frac{1}{2} + \frac{1}{2} \cdot (1 - y)^{-\frac{1}{k}}$</td>
<td>$\frac{1}{2} + \frac{\eta}{2} \exp\left[ \mu + s \cdot \zeta \left[ y; -\frac{1}{s} \cdot \left( \ln(\alpha) + \mu \right) \right] \right]$</td>
</tr>
<tr>
<td>LES</td>
<td>$(1 - y)^{-\frac{1}{k}}$</td>
<td>$\sqrt{\frac{2}{\gamma}} \cdot \exp\left( \frac{\mu}{2} + \frac{s}{2} \cdot \zeta \left[ y; \frac{1}{s} \cdot \left( \ln\left( \frac{2}{3} \right) - \mu \right) \right] \right)$</td>
</tr>
</tbody>
</table>

Table 10: Quantiles for Markups

- $\zeta[y; z] \equiv \Phi^{-1}\left(1 - \Phi[z]\right) \cdot y + \Phi[z]$
References


