On the Measure of Distortions

Hugo A. Hopenhayn

May 19, 2012

Abstract

The paper considers formally the mapping from distortions to the allocations of resources across firms to aggregate productivity. TFP gaps are characterized as the integral of a strictly concave function with respect to an employment-weighted measure of distortions. Size related distortions are shown to correspond to a mean preserving spread of this measure, explaining the stronger effects on TFP found in the literature. An empirical lower bound on distortions based on size distribution of firms is derived and analyzed. The effect of curvature on the impact and measurement of distortions is also considered.
1 Introduction

Among the factors explaining the disparity of aggregate productivity across countries, the misallocation of resources across firms has been receiving much attention in recent papers [1, 2, 3, 4, 5, 8, 12]. The basic idea is that institutions and policies might prevent the equalization of the marginal value of inputs across firms, thus resulting in aggregate productivity losses. The benchmark models used in many of these papers [7, 6, 9, 10, 11] share a similar structure. The main objective of this paper is to provide a precise characterization of the link between these inter-firm distortions and aggregate productivity in this class of models, that is summarized in a measure of distortions.

The basic setting considered here has a set of firms producing a homogenous product using labor as the only inputs. Firms produce output with a homogenous production function that exhibits decreasing returns and with an idiosyncratic productivity. The optimal allocation of labor across firms was considered by [10]. It leads to an endogenous size distribution of firms after equating marginal product of labor and a simple expression for the aggregate production function. Aggregate productivity in this undistorted economy is a geometric mean of firm level productivities.

Barriers to the reallocation of labor resulting from firing costs were first considered in [6] as a source of misallocation. The literature that followed (see in particular [12, 5, 3]) abstracts from policies and considers the quantitative effect of hypothetical barriers preventing the reallocation of labor as firm specific wedges. The main results suggested by this literature are: 1) Large distortions can lead to large effects on productivity; 2) More concentrated distortions have larger effects; 3) distortions that result in a reallocation of labor from more establishments with higher TFP to those with lower TFP are more detrimental to productivity than those that inefficiently reallocate labor within size classes. Such is also the case of size dependent policies considered in [5].

The analysis in these papers is purely quantitative and there is no theoretical analysis establishing these results. This paper attempts to fill the gap, providing a transparent characterization of the mapping between distortions and aggregate productivity. Consider distortions that move employment from a set of firms to another set, starting at the efficient allocation. First note that small changes have a second order effect, regardless of the original size (proportional to TFP) of the firms involved since at the efficient allocation marginal productivities are equalized. Since first order effects are zero, the effects of the reallocation on aggregate TFP must depend on infra-
marginal considerations. The first observation is that, fixing the number of workers reallocated if the reallocation involves small firms it will affect a larger proportion of their employment, digging deeper in the infra-marginal distortion. This is more damaging to aggregate TFP than a reallocation involving the same number of workers and firms, but the involved firms are larger (this would include reallocating those workers from large to small firms.) Second, for a given amount of total employment reallocated, the negative effect on productivity depends on the fraction of original efficient employment being reallocated (depth) and not on the specific source and destination. For example taking 10% of employment from one firm with 1000 employees is equivalent to taking 10% employment from 10 firms with 100 employees each.

These two observations lead to the following measure of distortions. Let \( n_i \) be the employment of a firm under the efficient allocation and \( \theta_i n_i \) its distorted employment. The measure of distortions \( N(\theta) \) counts the total fraction of aggregate original employment -regardless of source- that was was distorted by \( \theta \). This measure is sufficient to derive the effects on aggregate productivity. Moreover, the ratio of TFP in the distorted economy to the undistorted level has a simple representation: \( \int \theta^\alpha dN(\theta) \), where \( \alpha \) is the degree of decreasing returns faced by firms. As \( 0 < \alpha < 1 \), it follows immediately that mean preserving spread of this measure leads to lower productivity. The notion of a mean preserving spread can be interpreted as putting more employment mass at "larger distortions" and "concentrating" more the distortions. This also explains the quantitative results found in [12].

As an application of our results, we provide an answer to the following question: given two size distributions of firms \( F \) and \( G \), where \( F \) corresponds to an undistorted economy and \( G \) to a distorted one, what are the minimum set of distortions (in terms of their effect on TFP) that rationalize \( G \)? The characterization, confirming the method used in [1], reduces the problem to one of assortative matching with the simple solution of setting \( \theta(n) \) so that \( F(n) = G(\theta(n)) \) for all \( n \). I then apply this method to obtain lower bounds on distortions for India, China and Mexico taking the US as a benchmark. The effects on productivity according to this lower bound are meager: about 3.5% for India and Mexico and 0.5% for China. I also show that the size distributions generated by the distortions considered in [12] can be also rationalized with distortions that imply much smaller TFP decrease (e.g. 7% instead of 49\%). This negative result suggests that without explicitly measuring distortions as in [8] or deriving them from observed policies, there is not much hope of establishing large effects if they are to be consistent with
measured size distributions.

Another important factor in determining the effect of distortions is the assumed curvature in the firm level production function (or demand), for which there is no general consensus. For instance, while [12] and many of the papers that follow take a value $\alpha = 0.85$, the analysis in [8] use an implied value $\alpha = 1/2$. We first establish that for given measure of distortions, the effects on productivity are zero at the extremes $\alpha = 0$ and $\alpha = 1$, so that the impact of $\alpha$ in the calculations is non-monotonic. We then consider the question of curvature in the context of the calculations carried in [8]. There are more subtleties to the analysis as given data on firms inputs and output, the implied productivities and optimal input choices depend on $\alpha$. Thus the measure of distortions also varies with $\alpha$. Surprisingly, the effect of $\alpha$ is perfectly determined going monotonically from no productivity losses when $\alpha = 0$ to maximum losses for $\alpha = 1$. The proof is remarkable as it reduces the TFP ratio to a certainty equivalent and then uses standard analysis of risk aversion.

The paper is organized as follows. Section 2 sets up the benchmark model. Section 3 discusses the distorted economy. Section 4 develops the measure of distortions, derives the mains Propositions and discusses its implications. Section 5 derives the lower bound of distortions using the size distribution of firms and provides calculations of those bounds for a set of countries. Section 6 considers the role of curvature and Section 7 concludes.

2 Baseline model

This section describes a simple baseline model that will be used throughout the paper. The model is a simplified version of [6] that builds on [7] but without entry and exit. and closely related to [10] and [9]. As in [10] we consider here a static version is a collection of firms $i = 1, \ldots, M$, with production functions

$$y_i = z_i n_i^\alpha,$$

where $z_i$ is an idiosyncratic productivity shock for firm/establishment $i = 1, \ldots, n$. Production displays decreasing returns ($\alpha < 1$) in the only input labor and total endowment in the economy $N$ is supplied inelastically. Firms behave competitively taking prices as given. This economy has a unique competitive equilibrium $\{n_i, w\}$, where $n_i$ is the profit maximizing input choice for firm $i$ and labor market clears. The competitive equilibrium is also the solution to the planners problem:
\[
\max_{n_i} \sum_i z_i n_i^\alpha \\
\text{subject to: } \sum n_i \leq N.
\]

The first order conditions for this problem imply that
\[
\ln n_i = a_0 + \frac{1}{1 - \alpha} \ln z_i \tag{2.1}
\]
where \(a_0\) is a constant that depends on \(\alpha\), \(N\) and the vector of firm level productivities. Substituting in the production function,
\[
\ln y_i = \ln z_i + \alpha \left( a_0 + \frac{1}{1 - \alpha} \ln z_i \right) \tag{2.2}
\]
\[
= \alpha a_0 + \frac{1}{1 - \alpha} \ln z_i \tag{2.3}
\]
\[
= \ln n_i - a_0 (1 - \alpha) \tag{2.4}
\]
is also proportional to \(z_i\), implying that at the efficient allocation \(y_i/n_i = y/n = a_0^{\alpha-1} \equiv a\) for all \(i\). \(^1\) Finally, using the aggregate resource constraint to substitute for \(a\), it follows that
\[
y = \left( \sum_i z_i^{1-\alpha} \right)^{1-\alpha} N^\alpha.
\]

This is an aggregate production function of the same class as the underlying firm-level production function, with TFP parameter given by \(\left( \sum_i z_i^{1-\alpha} \right)^{1-\alpha}\). This technology exhibits decreasing returns in the aggregate, as firms here are treated as a fixed factor. This can be more clearly seen, dividing the first term by \(M^{1-\alpha}\)
\[
y = \left( E z_i^{1-\alpha} \right)^{1-\alpha} M^{1-\alpha} N^\alpha. \tag{2.5}
\]

This aggregate production function has constant returns to scale in firms and other inputs (in our example, labor), where aggregate TFP is a geometric mean of firm level productivity.

\(^1\)This would obviously not hold when with a fixed cost in terms of overhead labor, as used in [3].
3 The distorted economy

This section analyzes the consequences of deviations from the optimal allocation of resources across productive units. Figure 3.1 provides a useful picture of the type of distortions that might occur:

![Figure 3.1: Wedges in marginal product](image)

The solid line shows an optimal allocation, where \( \ln n_i \) is a linear function of \( \ln z_i \). The dots represent actual employment.

1. \( n_i \) not equal for all firms with the same \( z_i \), termed uncorrelated distortions;

2. average \( \ln n_i (z) \neq a + \frac{1}{1-\alpha} z \), termed correlated distortions, in the case of Figure 3.1 it is a distortion that results in reallocation of labor from more to less productive firms.

Both of these distortions result in losses of productivity as marginal product (or the marginal value of labor) is not equated across productive units. As an accounting device, it is useful to model these distortions as firm-specific implicit taxes/subsidies that create a wedge between its revenues and output:

\[
\begin{align*}
r_i &= (1 - \tau_i) y_i = (1 - \tau_i) z_i n_i^\alpha \\
&= \alpha (z_i (1 - \tau_i))^{\frac{1}{1-\alpha}},
\end{align*}
\]

where \( \alpha \) is a constant that depends only on the equilibrium wage. Equilibrium in this economy will be identical in terms of allocations to the equilibrium of an undistorted economy where the distribution of firm productivities
is changed to $z_i (1 - \tau_i)$. Total revenues are given by

$$r = N^\alpha M^{1-\alpha} \left( E[z_i (1 - \tau_i)]^{\frac{\alpha}{1-\alpha}} \right)^{1-\alpha} \tag{3.1}$$

and total output

$$y = \int y_i di = \int r_i (1 - \tau_i)^{-1} di$$

$$\frac{y}{r} = \frac{\int r_i (1 - \tau_i)^{-1} di}{\int r_i} = \frac{E(1 - \tau_i)^{-1} (z_i (1 - \tau_i))^{\frac{1}{1-\alpha}}}{E(z_i (1 - \tau_i))^{\frac{1}{1-\alpha}}}.$$ \tag{3.2}

Using equations (3.1) and (3.2), it follows that

$$y = N^\alpha M^{1-\alpha} \frac{E(1 - \tau_i)^{-1} (z_i (1 - \tau_i))^{\frac{1}{1-\alpha}}}{E(z_i (1 - \tau_i))^{\frac{1}{1-\alpha}}}.$$ \tag{3.3}

### 3.1 Distortions and aggregate productivity: some examples

It has been suggested [12] that correlated distortions that implicitly tax high productivity firms and subsidize low productivity ones are more damaging to aggregate productivity than those that are uncorrelated to size. This section shows that under a homogenous production function, such presumption is not necessarily true.

I will first consider three examples that illustrate some of the key insights.

**Example 1.** There are two types of firms with productivities $z_1 = 1$ and $z_2 = 2$. Suppose there are 16 firms of each type, a total labor endowment $N = 2000$ and $\alpha = \frac{1}{2}$. It is easily verified that the optimal allocation requires $n_1 = 25$ and $n_2 = 100$ and total output $y = 400$.

*Uncorrelated distortions for low productivity firms.* Now suppose that 12 of the type 1 firms are excluded from production while 4 of them get 100 workers. This gives a feasible set of distortions as it is easily verified that total employment doesn’t change and total output $y = 360$.

*Uncorrelated distortions for high productivity firms.* Assume instead that 3 type 2 firms are excluded from production while one of them gets 400 workers. This does not change aggregate employment and also gives total output $y = 360$.

*Correlated distortions.* Now assume 12 firms of type 1 are excluded from production while 1 firm of type 2 gets 400 workers. Again, this does not change aggregate employment and also total output $y = 360$. 
What do all these examples have in common? In all cases employment in some firms is dropped to zero while for other firms it is multiplied by 4, relative to the efficient allocation. Moreover, in all cases the original amount of employment that is affected by each of these distortions is exactly the same. In the first case, the employment dropped to zero is that of 12 type 1 firms giving a total of $12 \times 25 = 300$. In turn, 4 of these firms had employment quadrupled so the total employment affected by this distortion is $4 \times 25 = 100$. In the second case, the three firms of type 2 excluded from production represent an original total employment of 300 while the one firm whose employment quadrupled had 100 workers. It is easily verified that the same is true for the last case.

The examples suggest that what matters for total productivity is not what type of firms are hit by each distortion but the total original employment affected by them. In the remaining of this chapter we prove this conjecture, providing a general characterization of distortions.

4 The Measure of Distortions and Aggregate Productivity

For the undistorted economy, employment of each firm $n(z) = az^{\frac{1}{1-\alpha}}$, where $a$ is a constant that depends on the total labor endowment $N$ and the distribution of productivities. I will define a distortion as a ratio $\theta$ from actual employment to the undistorted one: $n = \theta n(z)$, where $\theta \geq 0$. (It is easy to see that this is equivalent to a wedge $(1 - \tau) = \theta^{1-\alpha}$.) Distortions reallocate resources across firms, so they generate the same level of aggregate employment. This motivates the following definition:

Definition 1. A feasible distortion is a conditional probability distribution $P(\theta|z)$ such that

$$N/M = \int \theta n(e) dP(\theta|e) dG(e),$$

where $N$ is employment allocated to production and $M$ is the number of firms, so $N/M$ is the average size of a firm. Since there is a one to one mapping between $z$ and $n(z)$, it is more convenient to summarize these distortions, after the corresponding change of variables with a joint measure $\mu(\theta, n_u)$ with mass $M$ and such that:

$$N = \int n_u d\mu(\theta, n_u) = \int \theta n_u d\mu(\theta, n_u)$$
where $n_u$ is undistorted employment corresponding to $n(z)$ for some $z$.

For every $\theta$ let

$$N(\hat{\theta}) = \int_{\theta \leq \hat{\theta}} n_u d\mu.$$ 

It is easy to see that this defines a measure on $\theta$ with the property that $N = \int dN(\theta)$ and by feasibility $N = \int \theta dN((\theta))$. Here $N(\hat{\theta})$ corresponds to the total original (undistorted) employment affected by a distortion $\theta \leq \hat{\theta}$. Notice that it is silent about the productivity of the firms underlying these distortions.\(^2\)

Consider now total output in the distorted economy:

$$y_d = M \int y(\theta, z) dP(\theta|z) dG(z).$$ 

Using

$$y(\theta, z) = z(\theta n(z))^{\alpha} = \theta^\alpha y(z) = \theta^\alpha a n(z)$$

where $a$ is average labor productivity in the undistorted economy, it follows that:

$$y_d = a \int \theta^\alpha n(z) d\mu(\theta, z)$$

$$= a \int \theta^\alpha dN(\theta)$$

Since $a$ is average labor productivity in the undistorted economy, it follows that $y_u = aN$ so the ratio of TFP in the distorted economy to the undistorted one is:

$$\frac{TFP_d}{TFP_u} = \frac{\int \theta^\alpha dN(\theta)}{N} \tag{4.1}$$

which simply corresponds to integrating $\theta^\alpha$ with the normalized measure of distortions, i.e. by the corresponding employment weights.

A similar result can be obtained considering the employment weighted distribution of wedges, since there is a one-one mapping between wedges and $\theta$ given by $\theta = (1 - \tau)^{\frac{1}{1-\alpha}}$.

\(^2\)Note that in the the three examples above this measure is given by:

$$N(\theta) = \begin{cases} 300 & \text{for } 0 \leq \theta < 1 \\ 1900 & \text{for } 1 \leq \theta < 4 \\ 2000 & \text{for } 4 \leq \theta. \end{cases}$$
4.1 Ordering distortions

Not all measure of distortions correspond to feasible distortions, for they need to be consistent with total employment.

Definition 2. A feasible measure of distortions is a measure $N(\theta)$ that integrates to $N$ and such that

$$\int \theta dN(\theta) = N.$$ 

It follows immediately that a means preserving spread of a feasible measure of distortions is also a feasible measure of distortions. Together with equation (4.1) this suggests a very natural order on measures of distortions given by second order stochastic dominance. Indeed, as the function under integration is concave, a mean preserving spread of a measure of distortions gives another measure of distortions with lower associated TFP. The intuition behind this result is that the infra-marginal effect of distortions is stronger the more deep and concentrated they are. The following corollary is a direct consequence of this observation.

Corollary 1. The effect of uncorrelated distortions on TFP increases with the total employment of the group involved. In particular, holding fixed the number of firms affected, uncorrelated distortions to large firms are more detrimental to TFP than uncorrelated distortions to small firms.

4.2 Generalization to more inputs

The above analysis generalizes easily to more inputs with a Cobb-Douglass specification. For exposition, we consider here the case of two inputs, $n$ and $k$ and production function $y_i = z_i n_i^\alpha k_i^\beta$. Letting $n(z)$ and $k(z)$ denote the optimal allocation and $\theta_L$ and $\theta_k$ the corresponding distortions, total output is:

$$y = M \int \theta_L^\alpha \theta_k^\beta z n(z)^\alpha k(z)^\beta dP(\theta_L, \theta_k | z) dG(z)$$

where $K/M = \int k(z) dP(\theta_L, \theta_k) dG(z)$ and $N/M = \int \theta_L d(\theta_L, \theta_k) dG(z)$.

Using linearity between $k(z)$ and $n(z)$ it follows that total output

$$y = a M \int \theta_L^\alpha \theta_k^\beta z z^{\frac{\alpha+\beta}{1-\beta-\alpha}} dPdG$$

$$= a_0 \int \theta_L^\alpha \theta_k^\beta dN(\theta_L, \theta_K)$$
for some constants $a$ and $a_0$ that are independent of the $\theta's$, and consequently

$$\frac{TFFP}{TFFP_{eff}} = \frac{1}{N} \int \theta_L^\alpha \theta_K^\beta dN (\theta_L, \theta_K).$$

This is again an employment weighted measure integrating $\theta_L^\alpha \theta_K^\beta$, that is in turn homogeneous aggregator of distortions. In the particular case where $\theta_L = \theta_K = \theta$ this aggregator is $\theta^{\alpha+\beta}$ that setting $\alpha = \alpha + \beta$ is the same as the one obtained before. A caveat to this extension is that we are treating total capital as given. This in general is not the case, but as we will see is justified in the analysis of the next section.

4.3 Restuccia and Rogerson: explaining the results.

In a recent paper, [12] examine the potential effects of distortions in an economy similar to the one described here. The specification is similar to the one above with the addition of capital with production function $y_i = z_i n_i^\alpha k_i^\beta$. They consider taxes on output, so profits are of the form $(1 - \tau_i) y_i - w_l i - r k_i$, where $\tau_i$ denotes a sales tax. In their numerical exercises, $\tau_i$ takes two values $\tau_1 > 0 > \tau_2$ applied to two subsets of firms. The subsidy $\tau_2$ is chosen so that total capital remains constant. Note that the effect of this sales tax is the same as that of a tax/subsidy $(1 - \tau_i)^{-1}$ on labor and capital and equivalent to setting $\theta_L = a_L (1 - \tau_i)^{\frac{1}{1-\alpha+\beta}}$ and $\theta_K = a_k (1 - \tau_i)^{\frac{1}{1-\alpha+\beta}}$ where $a_L$ and $a_k$ are constants that guarantees that resource constraints are satisfied (implying a feasible measure of distortions).

The following table is taken from the simulations reported by Restuccia and Rogerson. The first two columns consider the case of uncorrelated distortions, where a fraction of establishments is taxed and the counterpart subsidized at a rate so that total capital stock is unchanged. The second pair of columns consider the case where the $x\%$ most productive establishments are taxed while the counterpart is subsidized, again at a rate that maintains the total capital stock unchanged.

There are two distinguishing features of these distorted economies: 1) the larger the share of establishments taxed, the larger the negative effect on TFP and 2) Correlated distortions seem to have a larger effect than uncorrelated ones. Restuccia and Rogerson interpret this last result as follows: "A key difference is that in this case the distortion is not to the size distribution of establishments of a given productivity, but rather to the distribution of resources across establishments of varying productivity."

To examine the first feature, take an uncorrelated distortion that sets a tax $\tau_i > 0$ to a fraction $\alpha$ of establishments while subsidizing the rest
Table 1: Uncorrelated and Correlated Distortions

<table>
<thead>
<tr>
<th>% Estab. taxed</th>
<th>Uncorrelated</th>
<th>Correlated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\tau_t$</td>
<td>$\tau_s$</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>90%</td>
<td>0.84</td>
<td>0.66</td>
</tr>
<tr>
<td>50%</td>
<td>0.96</td>
<td>0.80</td>
</tr>
<tr>
<td>10%</td>
<td>0.99</td>
<td>0.92</td>
</tr>
</tbody>
</table>

with $\tau_s < 0$. Consider the corresponding distortions $\theta_t < 1$ and $\theta_s > 1$. To preserve total employment (and use of capital) it must be that

$$\theta_s - 1 = \frac{\alpha}{1 - \alpha} (1 - \theta_t).$$

(4.2)

The corresponding measure of distortions is: $\{(\alpha N, \theta_t), ((1 - \alpha) N, \theta_s)\}$. An increase in $\alpha$ can be interpreted as a mean preserving spread of the original measure. To see this, take $\alpha' > \alpha$ and new measure $\{(\alpha' N, \theta_t), ((1 - \alpha') N, \theta_s')\}$. It follows immediately that $\theta_s' > \theta_s$ and that

$$\frac{(1 - \alpha') \theta_s' + (\alpha' - \alpha) \theta_t}{1 - \alpha} = \theta_s.$$ 

So the new measure can be constructed by taking $(\alpha' - \alpha) N$ from the original mass at $\theta_s$ and assigning it to $\theta_t$ and mass $(1 - \alpha') N$ and assigning it to $\theta_s'$, which as shown is mean preserving.

Consider now the second feature. Correlated distortions do in fact have larger effects in this setting, but not for the reasons claimed above. Indeed, our examples in Section 3.1 suggest that this need not be true. Even though correlated distortions move resources from establishments with higher TFP to those with lower ones, at the efficient allocation marginal productivities are equated and so the nature of marginal distortions does not matter. The reason why correlated distortions are more detrimental to productivity in the Restuccia and Rogerson simulations is that they hit a larger fraction of the population and result in a more dispersed measure of distortions. The analysis follows very similar lines to comparative statics with respect to $\alpha$ considered above. To preserve equality of total employment, the following

---

3More precisely, it is the product of this measure since the same distortions apply to labor and capital.
must hold:

\[ (\theta_s - 1) = \frac{N_t}{N_s} (1 - \theta_t), \]

where \( N_t \) corresponds to the total employment (at the efficient allocation) of establishments taxed and \( N_s \) of those subsidized. With correlated distortions where larger establishments are taxed and holding constant \( \alpha \), \( N_t/N_s \) will be higher. Following the same logic as above, the corresponding measure of distortions is a mean preserving spread of the distribution of uncorrelated distortions with the same \( \alpha \) where \( N_t/N_s \) would be lower.

5 Distortions and the size distribution of firms

One feature of underdeveloped economies is a different size distribution, characterized by a "missing middle" and in particular the large fraction of employment in small firms. How much do these differences in size distribution explain the TFP gap? What is the role played by distortions? In a very interesting paper, [1] provide a clever answer to this question which we formally analyze below. This analysis provides a potentially very valuable tool, as information on size distribution of firms is usually available for a large set of countries, much more so than access to individual firm data.

There are obvious identification problems with this approach. Even if the economies compared had the same distribution of firm level productivities, the mapping between distortions and size distribution is not invertible. As an example, the same size distribution for an undistorted economy can be obtained by another one where only the least productive firms produce with a distribution of wedges that generates this size distribution. However, as shown below and following the procedure suggested in [1], it is very straightforward to find a lower bound on distortions under some identifying assumptions.

Let \( Q (\theta, n) \) be a joint measure on \((\theta, n)\). The interpretation is as follows: if the economy were undistorted, the size distribution of firms would be the marginal on \( n \):

\[ F (n) = \int_{z \leq n} dQ (\theta, z) \]  

while the actual size distribution

\[ G (n) = \int_{\theta z \leq n} dQ (\theta, z) \].

This joint distribution has an associated measure of distortions \( N (\theta) = M \int nQ (\theta, dn) \). The lower bound on distortions is obtained as the solution
to the following program:

$$\max_{Q(\theta,n)} \int \theta^n Q(d\theta, dn)$$

subject to (5.1) and (5.2), where $F$ is a given distribution of efficient establishment size and $G$ the actual size distribution.

The following Proposition characterizes the solution to this problem, but it is very intuitive. For continuous size distributions the solution is for each $n$ a point mass at $\theta(n)$ so that $F(n) = G(\theta(n)n)$. Letting $h(n) = \theta(n)n$ this function provides an assortative match between efficient and actual firm size by matching the percentiles of the corresponding distributions.

**Proposition 1.** Suppose $F$ and $G$ are continuous distributions. The solution to (5.3) is given the joint measure $Q(\theta,n)$ that puts all mass on the graph of $\theta(n)$ where $F(n) = G(\theta(n)n)$ and $dQ(\theta(n), n) = dF(n)$.

**Proof.** We show the solution to the above problem can be cast as an optimal matching problem. Any pair $(\theta,n)$ can be also represented by $(m,n)$ where $m = \theta n$. Hence for any joint measure $Q(d\theta, dn)$ there exists a corresponding measure $P(dm, dn)$ with first marginal $G$ and second marginal $F$. Rewrite (5.3) as:

$$\max_{P(dm, dn)} \int \left( \frac{m}{n} \right)^{\alpha} nP(dm, dn)$$

subject to:

$$G(dm) = P(dm, N)$$
$$F(dn) = P(M, dn)$$

where $N$ and $M$ are the support of $F$ and $G$, respectively. This is an assignment problem with match-return function $u(m,n) = m^\alpha n^{1-\alpha}$. Since this function is supermodular, the solution is perfectly assortative matching, so that $m(n)$ satisfies $F(n) = G(m(n)) = G(\theta(n)n)$ where $\theta(n) = m(n)/n$.

The above procedure would require to know, in addition to the actual size distribution of firms, the hypothetical size distribution for that economy in absence of distortions. This can be done (with somewhat strong assumptions) by benchmarking that economy with an undistorted one under the following identifying assumptions:

**Assumption 1.** (a) The benchmark economy is undistorted. (b) The underlying distribution of productivities for both economies is the same.
Take for instance the size distributions of India and US as the benchmark economy as shown in the left side of Figure 5.1. The average size of a US firm is approximately 270 while in India it is only 50 and that is apparent from the strong stochastic dominance observed in this figure. This fact needs to be considered when calculating the hypothetical size distribution India would have if the economy were undistorted. More generally, letting $\bar{n}_d$ denote the average size of the distorted economy and $\bar{n}_u$ that of the undistorted one, a firm that is of size $n$ in the undistorted economy (e.g. US) would have been of size $\gamma^{-1}n$ in the distorted one (e.g. India) in the absence of distortions, where $\gamma = \bar{n}_u/\bar{n}_d$. This adjustment is done in the right panel of Figure 5.1. Once this adjustment is made, it shows that India’s size distribution is compressed relative to that of the US.

![Figure 5.1: Size distribution of Firms: India and US](image)

Let $F$ denote the size distribution of the benchmark economy and $G$ the that of the distorted one. Define the hypothetical size distribution of the distorted economy in absence of distortions $\tilde{F}$ by $\tilde{F}(n) = F(\gamma n)$. Our upper bound is constructed setting $G(m(n)) = \tilde{F}(n)$ and computing:

$$\frac{TFP_d}{TFP_u} = \frac{1}{\bar{n}_d} \int m(n)\alpha n^{1-\alpha}d\tilde{F}(n),$$

$$= \frac{1}{\bar{n}_d} \gamma \int m(n)\alpha n^{1-\alpha}dF(\gamma n)$$

$$= \frac{\gamma^{\alpha-1}}{\bar{n}_d} \int m\left(\frac{z}{\gamma}\right)^\alpha z^{1-\alpha}dF(z)$$

$$= \frac{\gamma^\alpha}{\bar{n}_u} \int \tilde{m}(z)^\alpha z^{1-\alpha}dF(z)$$

where $\bar{n}_u = \int ndF(n)$ and $\bar{n}_d = \int ndG(n)$ are average firm employment in each of the two economies and $\tilde{m}(z) = m\left(\frac{z}{\gamma}\right)$. Using the above definition
of $m(n)$, it follows that:

$$G(\tilde{m}(z)) = G\left(m\left(\frac{z}{\gamma}\right)\right) = \tilde{F}\left(\frac{z}{\gamma}\right) = F(z).$$

So the procedure consists in defining $\tilde{m}(z)$ by matching percentiles of the distributions $G$ and $F$, then computing the average distortion and multiplying by the factor $\gamma^\alpha$ to account for the differences in average firm size.

The identifying assumptions are strong, but I believe they can be weakened. In particular, I conjecture that (b) can be weakened to assuming that the distribution of productivities of the distorted economy is dominated by that of the distorted one. Differences in productivities would arise from both, difference in underlying productivities together with distortions. A firm could be small in the distorted economy either because its employment is less than optimal or because its productivity is lower. For a given level of employment, the firm’s output would be higher in the latter case. This suggests that to explain a lower size distribution in the distorted economy, it is less damaging to the TFP of that economy if this comes from distortions rather than lower distribution of firm productivities. If this is true, the assumption that both firms have the same distribution of firm productivities provides an upper bound for $TFP_d/TFP_u$.

### 5.1 Example: Pareto distributions

Consider two economies a benchmark economy $b$ and a distorted economy $d$ that have Pareto size distributions with parameters $(x_b, b, x_d, d)$, respectively, and mean employments $\bar{n}_b = \frac{bx_b}{b-1}$ and $\bar{n}_d = \frac{dx_d}{d-1}$ satisfying the above assumptions, so $1-F(n) = \left(\frac{n}{x_b}\right)^{-b}$ and $1-G(n) = \left(\frac{n}{x_d}\right)^{-d}$ and $\gamma = \frac{bx_b}{dx_d} \frac{d-1}{b-1}$. Define $\tilde{m}(n)$ by $F(n) = G(\tilde{m}(n))$:

$$\tilde{m}(n) = x_d \left(\frac{n}{x_b}\right)^{b/d}. $$
The bound on relative TFP is given by:

\[
\frac{TFP_d}{TFP_e} = \frac{\gamma^\alpha}{\tilde{n}_b} \int_{x_b} \tilde{n}(n)^\alpha n^{1-\alpha} dF(n) \\
= \frac{b \gamma^\alpha x_d^\alpha}{\tilde{n}_b x_b^{\alpha / d-b}} \int_{x_b} n^{\frac{b}{\alpha} (1-b)} n^{1-\alpha} n^{-(1+b)} dn \\
= \frac{b \gamma^\alpha x_d^\alpha}{\tilde{n}_b x_b^{\alpha / d-b}} \int_{x_b} n^{\frac{b}{\alpha} (1-b)} dn \\
= \frac{b \gamma^\alpha x_d^\alpha}{\tilde{n}_b x_b^{\alpha / d-b}} \left[ \int_{x_b} n^{\frac{b}{\alpha} (1-b)} dn \right]_x \infty \\
= \frac{\gamma^\alpha x_d^\alpha x_b^{-\alpha} (1-b)}{1 + \frac{b}{\alpha} - b - \alpha} \\
= \frac{\left( \frac{b}{\alpha} - 1 \right) x_d^\alpha x_b^{-\alpha} (1-b)}{1 + \frac{b}{\alpha} - b - \alpha}
\]

which is independent of the scaling parameters \(x_b\) and \(x_d\). The following table gives the TFP ratios for some values of \(b\) and \(d\) (not sure why, but the computed values are the same when \(b\) and \(d\) are inverted, but I cannot see symmetry in the above formula.)

<table>
<thead>
<tr>
<th>b</th>
<th>d</th>
<th>TFP ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>1.2</td>
<td>0.96</td>
</tr>
<tr>
<td>1.5</td>
<td>1.5</td>
<td>0.82</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.72</td>
</tr>
<tr>
<td>1.5</td>
<td>2</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.94</td>
</tr>
</tbody>
</table>

5.2 Application: Bounds for some countries

This section computes the TFP bound for three economies: India, China and Mexico. The benchmark size distribution is the one corresponding to the US. Figure 5.2 provides on the left panel the size distribution of firms for the four countries. The right panel gives the distributions adjusted to the average firm size of the US (the average sizes are as follows: Mexico=15,
India=50, US=272, China=558.) It is worth noting that the normalized
distribution for China is close to the one for the US, while those of Mexico
and India are similar to each other but more compressed than the US.

Figure 5.2: Size Distributions: India, China, Mexico and US

Figure 5.3 plots the corresponding measures of distortion. Recall that an
undistorted economy corresponds to a point mass measure at one. China’s
measure is very close to undistorted, while India and Mexico’s measures, be-
ing very similar to each other, are a clear mean preserving spread of China’s.
The corresponding TFP ratios are as given in Table 3, giving the above
calculation for different values of $\alpha$. As was apparent from before, China
appears as almost undistorted while the TFP losses for India and Mexico
are relatively small, especially when taking $\alpha = 0.85$ which is one of the
standard value used in the literature. The value of $\alpha = 0.5$ is consistent with
the markup values used by [8].

Table 3: TFP ratios

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 1/2$</th>
<th>$\alpha = 2/3$</th>
<th>$\alpha = 0.85$</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>0.991</td>
<td>0.992</td>
<td>0.995</td>
</tr>
<tr>
<td>India</td>
<td>0.928</td>
<td>0.937</td>
<td>0.964</td>
</tr>
<tr>
<td>Mexico</td>
<td>0.931</td>
<td>0.939</td>
<td>0.966</td>
</tr>
<tr>
<td>R&amp;R</td>
<td>0.851</td>
<td>0.873</td>
<td>0.929</td>
</tr>
<tr>
<td>USA mean</td>
<td>0.655</td>
<td>0.693</td>
<td>0.817</td>
</tr>
</tbody>
</table>

18
These bounds on distortion losses are very small when compared to what has been suggested in the literature. To put our results in perspective, I now consider the most extreme hypothetical case considered in [12], where the top 90% firms are taxed away 40% of their output giving a TFP ratio of 0.51. Using $\alpha = 0.85$ as done in their paper, the implied measure of distortions is $\{(0.033, 90\%), (470, 10\%)\}$ as depicted in the left panel of Figure 5.4. These distortions give rise to a substantial spread in the size distribution and a distance to the US undistorted distribution that is much larger than than of India, as shown in Figure 5.5. Using this size distribution, we can now calculate our lower bound of distortion. The corresponding measure is shown in the right panel of Figure 5.4. It is considerably more dispersed than the one obtained for India but orders of magnitude less disperse than the actual measure of distortions as shown in the left panel of the figure. As a consequence, this lower bound on the measure of distortions is considerably less damaging to TFP. The fourth row in Table 3 gives the corresponding TFP ratios. Compare the bound for $\alpha = 0.85$ that gives a 7% decrease in TFP to the one obtained in [12] which is almost 50%!
In contrast to [12] and [5], our method disciplines the calculations of policy distortions with the size distribution of different countries. But at the same time, ours is a lower bound which is only attained when distortions do not lead to rank reversals in firm size. These rank reversals inevitably occur both for the correlated and uncorrelated distortions considered in [12]. To illustrate this point, consider the extreme case of correlated distortions analyzed above: a firm with 2 employees in the undistorted economy will have approximately 1,000 in the distorted one and a firm with an original employment of 9,000 employees ends up with less than 300! The full extent of rank reversals for [12] is shown in Figure 5.6. A useful comparison is to consider the largest distortions that can be obtained without rank reversals, which is to set all firms’ employment identical to the average size. The corresponding TFP ratios are given in the last row of Table 3: in case of $\alpha = 0.85$ this falls short of a 20% decrease in TFP.

Our bounds on distortions are also very small when compared to [8], derived from establishment level data for China and India, that find a TFP ratios in the order of 45%. Table 4 provides some statistics of the dispersion
in $\theta'$s found in [8] and in our calculation. All measures of dispersion are orders of magnitude higher in their data.

Table 4: Dispersion of $\ln \theta'$s

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>India (94)</th>
<th>China (98)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H-K Bound</td>
<td>H-K Bound</td>
</tr>
<tr>
<td>SD</td>
<td>4.47</td>
<td>4.93</td>
</tr>
<tr>
<td>75-25</td>
<td>5.4</td>
<td>6.27</td>
</tr>
<tr>
<td>90-10</td>
<td>10.7</td>
<td>12.4</td>
</tr>
</tbody>
</table>

6 On the impact of curvature

This section considers the impact of curvature - the degree of decreasing returns at the firm level - on the analysis of distortions. Recall the representation of TFP derived in Section 4

$$\frac{TFP_d}{TFP_e} = \int \frac{\theta^\alpha dN(\theta)}{N}.$$  

By Jensen’s inequality, for any $0 < \alpha < 1$ this is strictly less than one. For $\alpha = 1$ it equals one for it must integrate to employment. But also when $\alpha = 0$ this ratio is equal to one. Hence, the relationship between the TFP gap an curvature is not monotonic, for a fixed measure of distortion.

The caveat to this analysis is that we consider fixed the measure of distortions, while this might also be affected by $\alpha$. This follows from the fact that the optimal distribution of employment across firms is a function of $\alpha$: as $\alpha$ increases, employment becomes more concentrated in large firms and whether this results in a smaller or larger TFP gap depends on the distribution of distortions. If distortions are obtained as the result of firm level output wedges $(1 - \tau_i)$ as in [12], using equation (3.3) it is straightforward to see that the TFP gap disappears as $\alpha \rightarrow 0$. On the other extreme, the results are ambiguous and might depend on the nature of distortions: with uncorrelated distortions, output will still be concentrated in one firm with highest productivity but if they are positively correlated, it will not.

When distortions are uncovered from the data as in [8] there is an additional reasons why curvature will matter in the calculations: both, the distribution of TFP and implicit distortions (i.e. wedges) depend on $\alpha$. Interestingly enough, a sharp result emerges in this case as detailed below.
A stylized version of the procedure followed by [8] is as follows. The data consists of establishment levels of inputs and outputs:

\[(n_1, y_1, n_2, y_2, ..., n_M, y_M)\]

where \(M\) is the number of establishments. Using this data and a production function of the form \(y_i = z_i n_i^\alpha\), we can solve for the vector of productivities \((z_1, z_2, ..., z_M)\) and compute the counterfactual efficient level of output.

As shown in Section 2, aggregate TFP in the undistorted economy is:

\[TFP_e = \left( \sum_{i} z_i^{1-\alpha} \right)^{1-\alpha} \]

Substituting \(z_i = y_i/n_i^\alpha\) gives:

\[TFP_e = \left( \sum_{i} \left( \frac{y_i}{n_i^{\alpha}} \right)^{1-\alpha} \right)^{1-\alpha}\]

and dividing by actual TFP in this economy \(y/n^\alpha\) gives:

\[\frac{TFP_e}{TFP} = \left( \sum_{i} \left( \frac{y_i}{n_i^{\alpha}} \right)^{1-\alpha} \right)^{1-\alpha} = \left( \sum_{i} \frac{n_i}{n} LPR_i^{1-\alpha} \right)^{1-\alpha}\]  \(6.1\)

where \(LPR_i = \frac{y_i}{n_i^{\alpha}}\) stands for labor productivity ratio. From equation (6.1) it follows immediately that:

\[\left( \frac{TFP_e}{TFP} \right)^{1-\alpha} = \sum \frac{n_i}{n} LPR_i^{1-\alpha}\]  \(6.2\)

Equation (6.2) expresses the TFP ratio as the certainty equivalent of the lottery \(\{(LPR_1, \frac{n_i}{n}), (LPR_2, \frac{n_i}{n}), ..., (LPR_M, \frac{n_i}{n})\}\) under utility function \(u(x) = x^{1-\alpha}\). Note precisely because these preferences are risk loving they imply a TFP coefficient ratio greater than one. An increase in \(\alpha\) implies more risk loving and hence higher \(TFP_e/TFP\), so the TFP gap increases with \(\alpha\). At the extreme, when \(\alpha = 0\) utility is linear and there is no TFP gap. In the other extreme, when \(\alpha = 1\) and assuming firm \(M\) has the highest productivity the \(TFP_e/TFP = (\sum (z_i/z_M)(n_i/n))^{-1}\). This proves the following Proposition:
Proposition 2. When firm level tfp and wedges are obtained from the data as in [8], the ratio $\frac{TFP}{TFP_e}$ decreases with $\alpha$ and it is equal to one (i.e. no gap) at $\alpha = 0$.

As an example, suppose the economy consists of two firms and $n_1/n = n_2/n = 1/2$. The following table gives the TFP ratios for different levels of curvature and degree of distortions, as measured by the relative average output of the two firms.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>0</th>
<th>0.2</th>
<th>0.5</th>
<th>0.8</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>relative $y_i/n_i$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>1.09</td>
<td>1.28</td>
<td>1.57</td>
<td>1.74</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>1.05</td>
<td>1.17</td>
<td>1.39</td>
<td>1.55</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>1.02</td>
<td>1.08</td>
<td>1.22</td>
<td>1.35</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>1.01</td>
<td>1.02</td>
<td>1.07</td>
<td>1.16</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

It can be seen that TFP is very sensitive to the degree of curvature and as stated in the Proposition increases with $\alpha$.

7 Final remarks

To be included.
References


