

Robust estimation of production frontiers

Stima robusta di funzioni frontiera

Sergio Destefanis, Giuseppe Storti ^{1 2}

Dipartimento di Scienze Economiche e Statistiche

Università di Salerno

storti@unisa.it, destefanis@unisa.it

Riassunto: Si è più volte osservato in letteratura come gli approcci non parametrici alla stima della frontiera di produzione siano scarsamente robusti rispetto alla presenza di valori eccezionali. Il presente lavoro propone una classe di stimatori robusti della frontiera basati sulla teoria dei valori estremi. Il metodo proposto è di natura parametrica e consente di definire un test asintotico utile al fine di verificare la presenza di outliers nell'insieme di produzione osservato. Vengono infine presentati i risultati di un'applicazione del metodo descritto ad un campione di imprese estratto dalle indagini del Mediocredito Centrale sulle imprese manifatturiere italiane relative al periodo (1989-1997).

Keywords: production frontier, nonparametric estimation, Expected Maximal Output frontier, extreme value theory.

1. Introduction and statement of the problem

In the last two decades non-parametric methods for the estimation of production frontiers, such as Data Envelopment Analysis (DEA) and Free Disposal Hull (FDH), have been widely applied in the economic literature. These methods have the advantage of being based on a relatively small set of easily interpretable economic assumptions with no need for the formulation of restrictive hypotheses on the functional shape of the relationship between input and output variables on the frontier of the production set. One of the main drawbacks associated to these estimators is related to their robustness since, as pointed out by recent literature (Cazals et al., 2002), they are very sensitive to the presence of outliers. An alternative nonparametric approach which, by construction, is more robust to the presence of extreme observations, is that based on the concept of Expected Maximal Output (or, equivalently, Expected Minimum Input) function proposed by Cazals et al. (2002). However, in a nonparametric setting, the derivation of a test for detecting potential outliers is in general not straightforward. Furthermore, for many applications, the availability of a parametric model often allows for a more thorough economic interpretation of the phenomenon to be modeled. To conjugate the advantages of both approaches, Florens and Simar (2001) suggest a two stage procedure in which first, the frontier is estimated nonparametrically and, then, the best parametric approximation (in the least squares sense) of the estimated nonparametric frontier is found.

In this paper we propose a parametric approach to the estimation of the production frontier which is robust to the presence of extreme values and outliers and, at the same time,

¹Indirizzo per corrispondenza: Sergio Destefanis, DISES, Università di Salerno, Via Ponte Don Melillo, 84084 Fisciano (SA)

²Il lavoro è stato svolto con finanziamenti MIUR. Un ringraziamento speciale a Stuart Coles per aver fornito le routines in S-PLUS utilizzate per l'analisi empirica.

still relies on relatively weak distributional assumptions. The structure of the paper is as follows. Section 2 illustrates the proposed estimator while in Section 3 the results of an application to a microeconomic dataset are presented.

2. Robust estimation of the production frontier

Consider a reference set including n producers characterized by a technology with one output and $q \geq 1$ inputs. For each producer, let Y_i be the observed output and X_i be the associated input vector ($q \times 1$), for $i = 1, \dots, n$. Cazals et al. (2002) define the expected maximal output frontier of order- m as:

$$\phi_m(x) = E(\max(Y^1, \dots, Y^m) | X \leq x) \quad (1)$$

where $(Y^1, \dots, Y^m)$ are m independent identically distributed random variables generated from the distribution of Y given $X \leq x$. In practice the expectation in (1) can be approximated by a simple Monte Carlo procedure. First a random sample of size m is drawn with replacement from the empirical distribution of Y given $X \leq x$. Then, denoting this sample by $(y_b^1, \dots, y_b^m)$, the sample maximum will be given by $\tilde{\phi}_{b,m}(x) = \max_{i=1, \dots, m}(y_b^i)$ with $m < n(x)$, where $n(x)$ is the number of producers using input $X_i \leq x$, $i = 1, \dots, n(x)$. Finally, the above steps are repeated for $b = 1, \dots, B$ and the Expected Maximal Output Frontier of order m ($EMOF_m$) is estimated as:

$$\hat{\phi}_{m,n}(x) = B^{-1} \sum_{b=1}^B \tilde{\phi}_{b,m}(x) \quad (2)$$

where the quality of the approximation depends on the value of B . The sampling properties of this estimator are discussed by Cazals et al. (2002).

The $EMOF_m$ can be regarded as a nonparametric approach since it does not require the specification of a parametric model for the frontier. Differently, the method which is here proposed is based on a parametric fit even if it relies on weak distributional assumptions and does not imply any hypothesis on the nature of the functional shape of the frontier. More precisely, given a sample of i.i.d. random variables $(Y_1, \dots, Y_m)$ it follows that, under suitable assumptions, by the Extremal Types theorem (Coles, 2001) the sample maximum $M_m = \max(Y_1, \dots, Y_m)$ asymptotically converges in distribution to a Generalized Extreme Value (GEV) distribution :

$$M_m \xrightarrow{d} Z \quad (3)$$

as $m \rightarrow \infty$, where $Z \sim GEV(\mu, \sigma, \xi)$ with probability distribution function given by:

$$G(z) = \exp \left\{ - \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \right\} \quad (4)$$

for $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. The expected maximum is given by

$$E(Z) = \mu + \frac{\sigma}{\xi} [\Gamma(1 - \xi) - 1] \quad \text{for } \xi < 1 \quad (5)$$

while the quantile of order $(1 - p)$ of the GEV distribution is:

$$z_p \begin{cases} = \mu - \frac{\sigma}{\xi} \left\{ 1 - [-\log(1 - p)]^{-\xi} \right\} & \text{for } \xi \neq 0 \\ = \mu - \sigma \log [-\log(1 - p)] & \text{for } \xi = 0 \end{cases} \quad (6)$$

Hence, for a given value of the input vector x , a GEV distribution with parameters $\mu(x)$ (location), $\sigma(x)$ (scale) and $\xi(x)$ (shape) can then be fitted by maximum likelihood to the series of sample maxima $\tilde{\phi}_{1,m}(x), \dots, \tilde{\phi}_{B,m}(x)$ (Coles, 2001). The expected value of this distribution can be derived by equation (5) and can be regarded as the parametric counterpart of the $EMOF_m$. An alternative, and more general, approach would be to define the frontier as $z_p(x)$ i.e. as the percentile of order $1 - p$ of the underlying GEV distribution. From an economic standpoint, $z_p(x)$ can be directly interpreted as the output level which will be exceeded with probability p by the maximum output observed among m randomly chosen producers using less or equal input than x . Henceforth, we will refer to this frontier as the *Expected Return Level* of order (m, p) ($ERL_{m,p}$). Also note that, in practical applications, for assessment purposes, the reference standard could be made more or less severe by tuning the parameter p . Estimates $\hat{z}_{m,p}(x)$ and $\hat{E}_m(x)$ of $z_p(x)$ as well as of the expected maximum in (5) can be obtained by a simple plug-in of the ML estimates of $(\mu(x), \sigma(x), \xi(x))$ in equations (6) and (5), respectively.

Furthermore, this approach also offers a convenient framework for testing for the presence of outliers. This could be done referring to the symmetric $(1 - 2\alpha)\%$ confidence interval based on the profile likelihood of $z_p(x)$. If, for a given producer (x_0, y_0) the observed output y_0 exceeds the upper boundary of the confidence interval for $z_p(x_0)$, the producer will be then classified as a potential outlier at the level α . In this way, only producers characterized by extremely high output levels can be detected while, for the purpose of detecting producers characterized by extremely low production levels, a slight modification to the above described framework should be considered. More precisely, we could refer to the empirical distribution of Y given $X \geq x_0$, draw B random samples from this distribution and then, as suggested in Coles (2001), fit a GEV distribution to the sample maxima of the negated data. Then, again, if the observed output y_0 , multiplied by -1 , exceeds the upper boundary of the estimated confidence interval for $z_p(x_0)$, the producer will be classified as a potential outlier characterized by an extremely low output level.

3. The empirical analysis

In this section the proposed approach is applied to a panel of Italian manufacturing firms extracted from the three surveys performed by Mediocredito Centrale during the period 1989-1997. The sample includes 4322 firms operating in the Machinery sector. For each firm we observe one output, given by value added, and three inputs given by gross fixed assets at book value and labour disaggregated in the number of ‘blue’ and ‘white’ collars, respectively. Value added and fixed assets were deflated using annual price indices based on Istat national accounting. There are well-known problems with using assets at book value as a measure of capital stock. Yet our data afford no alternative and the literature always reports reasonable parameter estimates for this variable. More details on the data can be found in Destefanis (2000). In our analysis the production frontier has been estimated by using the $EMOF_m$ as well as the approach based on the $ERL_{m,p}$ for two different values of m (30,50) and three different values of p (0.10,0.05,0.01) (Table 1). Furthermore, the parametric estimator of the expected maximum $\hat{E}_m(x)$ has been considered. The number of replicates used for the estimation was $B = 200$.

The last column in Table 1 reports the number of times in which the algorithm for the numerical maximization of the GEV likelihood converged. Failures were mainly concen-

	$EMOF_m$	$\hat{E}_m$	$\hat{z}_{m,0.10}$	$\hat{z}_{m,0.05}$	$\hat{z}_{m,0.01}$	nobs
m=30	2180	2181	1155	814	273	3407
m=50	1828	1830	916	624	178	3442
m=100 (par.)	657	657	355	276	181	4322

Table 1: Number of observations exceeding the frontier for different estimators and different values of m .

trated at the beginning of the sample (ordered by increasing values of the input). In this sense we can say that a substantial drawback of the $\hat{E}_m$ and $ERL_{m,p}$ estimators is represented by end effects. In order to overcome this problem it is possible to replace non-parametric resampling from the empirical CDF by a parametric resampling procedure. This alternative approach has been here applied (Table 1) with a window of length $m=100$ and assuming an exponential distribution for the output i.e. $(Y|X \leq x) \sim Exp(\theta(x))$ with density:

$$f(y|X \leq x) = \theta(x)e^{-\theta(x)y}$$

For each producer (x, y) , the rate of the relevant exponential distribution was estimated as the reciprocal of the sample mean of the output values Y_i observed for producers with input level $X_i \leq x$.

The results in Table 1 show that the estimators considered do not envelope all the data points and are in this sense more robust to the presence of extreme values than conventional estimators such as DEA or FDH. The value of m acts as a smoothing parameter since the number of observations exceeding the frontier reduces as m increases. For more details on the impact of the value of m on the estimated frontier, the reader may refer to Cazals et al. (2002). Similarly, for fixed m , the estimated frontier output level $\hat{z}_{m,p}$ is a monotonic function of p : as p increases the frontier tends to get closer to the bulk of the observed data. It is also interesting to note how the results for the $EMOF_m$ and for the ‘parametric’ estimate of the expected maximum are very close, in the non-parametric case as well as when a parametric resampling procedure is adopted. The estimator based on parametric resampling looks like a promising alternative to its non-parametric counterpart since it allows to drastically reduce the impact of end effects. A more thorough investigation of the economic implications and of the statistical properties of the approaches based on the $\hat{E}_m$ and $ERL_{m,p}$ estimators is left for future research.

References

- Cazals, C., Florens, J. and Simar, L. (2002) Nonparametric frontier estimation: a robust approach, *Journal of Econometrics*, 106, 1-25.
- Coles S. G. (2001) *An Introduction to Statistical Modelling of Extreme Values*, Springer.
- Destefanis S. (2000) Differenziali territoriali di produttività ed efficienza e sviluppo dualistico, *Lavoro e Relazioni Industriali*, 2, 103-143.
- Florens, J. and Simar, L. (2001) Parametric approximations of nonparametric frontiers, Technical Report, IDEI-GREMAQ.