Nonparametric Frontier estimation: A Multivariate Conditional Quantile Approach

Abdelaati Daouia[1] and Léopold Simar[2]

[1] GREMAQ, Université de Toulouse I
21 allée de Brienne
31000 TOULOUSE, France
(e-mail: daouia@cict.fr)

[2] Institut de Statistique, Université de Louvain-la-Neuve
20 voie du roman pays
1348, Louvain-la-Neuve, Belgique
(e-mail: simar@stat.ucl.ca.be)

Abstract. This paper proposes a probabilistic framework for efficiency and productivity analysis in a complete multivariate setup (multiple inputs and multiple outputs). Properties of the Farrell's efficiency scores are derived in terms of the characteristics of the probability distribution of the data generating process. This allows to introduce a notion of α-quantile efficiency scores related to a non-standard conditional α-quantile frontier and nonparametric robust estimators are provided. The asymptotic behavior of the estimator is provided with numerical illustration.

Keywords: Frontier estimation, Robust nonparametric estimators, Conditional quantiles.

1 Introduction and Basic Concepts

Foundations of the economic theory on productivity and efficiency analysis date back to the works of [Koopmans, 1951] and [Debreu, 1951] on activity analysis. We consider a production technology where the activity of the production units is characterized by a set of inputs $x \in I\!R_+^p$ used to produce a set of outputs $y \in I\!R_+^q$. The production set is the set of technically feasible combinations of (x, y):

$$\Psi = \{(x, y) \in I\!R_+^{p+q} \mid x \text{ can produce } y\}. \tag{1}$$

Assumptions are usually done on this set, such as free disposability of inputs and outputs, meaning that if $(x, y) \in \Psi$, then $(x', y') \in \Psi$, as soon as $x' \geq x$ and $y' \leq y$.

The Farrell-Debreu efficiency scores for a given production scenario $(x, y) \in \Psi$, are defined as:

$$\text{Input oriented} \quad : \quad \theta(x, y) = \inf\{\theta \mid (\theta x, y) \in \Psi\} \tag{2}$$

$$\text{Output oriented} : \quad \lambda(x, y) = \sup\{\lambda \mid (x, \lambda y) \in \Psi\} \tag{3}$$

In practice Ψ is unknown and so has to be estimated from a random sample of production units $\mathcal{X} = \{(X_i, Y_i) \,|\, i = 1, \ldots, n\}$, where we assume that $\text{Prob}((X_i, Y_i) \in \Psi) = 1$ (called deterministic frontier models). So the problem is related to the problem of estimating the support of the random variable (X, Y) where Ψ is supposed to be compact. The most popular nonparametric estimators are based on the envelopment ideas (see *e.g.* [Simar and Wilson, 2000], for a recent survey).

The Free Disposal Hull (FDH) estimator ([Deprins *et al.*, 1984]) is provided by the free disposal hull of the sample points $\mathcal{X}$:

$$\widehat{\Psi}_{FDH} = \left\{(x, y) \in I\!R_+^{p+q} \,|\, y \leq Y_i, \; x \geq X_i, \quad i = 1, \ldots, n\right\}. \tag{4}$$

The FDH efficiency scores are obtained by plugging $\widehat{\Psi}_{FDH}$ in equations (2) and (3) in place of the unknown Ψ. The asymptotic properties of the resulting estimators are provided by [Park *et al.*, 2000]. In summary, the error of estimation converges at a rate $n^{1/(p+q)}$ to a limiting Weibull distribution.

The FDH estimators envelop all the data points and so are very sensitive to outliers and/or to extreme values. [Cazals *et al.*, 2002] have introduced the concept of partial frontiers (order-m frontiers) with a nonparametric estimator which does not envelop all the data points. The value of m may be considered as a trimming parameter and as $m \to \infty$ the partial order-m frontier converges to the full-frontier. It is shown that by selecting the value of m as an appropriate function of n, the non-parametric estimator of the order-m efficiency scores provides a robust estimator of the corresponding efficiency scores sharing the same asymptotic properties as the FDH estimators but being less sensitive to outliers and/or extreme values.

Recently [Aragon *et al.*, 2002] have proposed an alternative to order-m partial frontiers by introducing quantile based partial frontiers. The idea is to replace this concept of "discrete" order-m partial frontier by a "continuous" order-α partial frontier where $\alpha \in [0, 1]$ corresponds to the level of an appropriate non-standard conditional quantile frontier. Unlike the order-m partial frontiers, due to the absence of natural ordering of Euclidean spaces for dimension greater than one, the α-quantile approach is limited to one-dimensional input for the input oriented frontier and to one-dimensional output for the output oriented frontier.

In this paper, we overcome this difficulty and we propose an extension to the full multivariate case, introducing the concept of α-quantile efficiency scores and the corresponding α-quantile frontier set.

2 Probabilistic Formulation and Nonparametric Estimation

[Daraio and Simar, 2002] propose a probabilistic formulation of efficiency concepts. The Data Generating Process (DGP) of (X, Y) is completely characterized by

$$H_{XY}(x,y) = \text{Prob}(X \leq x, Y \geq y). \tag{5}$$

The support of $H_{XY}(\cdot,\cdot)$ is Ψ and $H_{XY}(x,y)$ can be interpreted as the probability for a unit operating at the level (x,y) to be dominated. This joint probability can be decomposed as follows:

$$\begin{aligned} H_{XY}(x,y) &= \text{Prob}(X \leq x \,|\, Y \geq y)\, \text{Prob}(Y \geq y) = F_{X|Y}(x|y)\, S_Y(y) && (6)\\ &= \text{Prob}(Y \geq y \,|\, X \leq x)\, \text{Prob}(X \leq x) = S_{Y|X}(y|x)\, F_X(x), && (7)\end{aligned}$$

where we suppose the conditional probabilities exit (*i.e.*, when needed, $F_X(x) > 0$ or $S_Y(y) > 0$).

An input oriented efficiency score $\tilde{\theta}(x,y)$ for $(x,y) \in \Psi$ is defined for all y with $S_Y(y) > 0$ as

$$\widetilde{\theta}(x,y) = \inf\{\theta \,|\, F_{X|Y}(\theta x|y) > 0\} = \inf\{\theta \,|\, H_{XY}(\theta x, y) > 0\}. \tag{8}$$

For the output oriented case, for all x such that $F_X(x) > 0$, we define the output efficiency score as

$$\widetilde{\lambda}(x,y) = \sup\{\lambda \,|\, S_{Y|X}(\lambda y|x) > 0\} = \sup\{\lambda \,|\, H_{XY}(x, \lambda y) > 0\}. \tag{9}$$

This input (resp. output) efficiency score can be interpreted as the proportionate reduction (resp. increase) of inputs (resp. outputs) a unit working at the level (x,y) should perform to be dominated with probability zero.

If Ψ is free disposal (a minimal assumption), it can be shown that: $\widetilde{\theta}(x,y) \equiv \theta(x,y)$ and $\widetilde{\lambda}(x,y) \equiv \lambda(x,y)$.

Natural nonparametric estimators of $\theta(x,y)$ and of $\lambda(x,y)$ are obtained by plugging the empirical distribution $\widehat{H}_{XY,n}$ in place of H_{XY} in the definition of the efficiency scores, where

$$\widehat{H}_{XY,n}(x,y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x, Y_i \geq y), \tag{10}$$

As pointed out in [Daraio and Simar, 2002], these estimators are the FDH estimators of the Farrell-Debreu efficiency scores.

3 Conditional Quantile Based Efficiency Scores

[Aragon *et al.*, 2002] have introduced the conditional quantile frontier function for a production (output) function when the output is unidimensional and for a cost (input) function when the input is one dimensional. We extend the ideas to a full multivariate setup. Since a natural ordering of Euclidean spaces of dimension greater than one does not exist, we overcome the difficulty by defining α-quantile efficiency scores as follows.

Definition 1 *For all y such that $S_Y(y) > 0$ and for $\alpha \in]0,1]$, the α-quantile input efficiency score for the unit $(x,y) \in \Psi$ is defined as*

$$\theta_\alpha(x,y) = \inf\{\theta \mid F_{X|Y}(\theta x|y) > 1-\alpha\} \tag{11}$$

For all x such that $F_X(x) > 0$ and for $\alpha \in]0,1]$, the α-quantile output efficiency score for the unit $(x,y) \in \Psi$ is defined as

$$\lambda_\alpha(x,y) = \sup\{\lambda \mid S_{Y|X}(\lambda y|x) > 1-\alpha\} \tag{12}$$

For instance, in the input case, $\theta_\alpha(x,y)$ is the proportionate reduction (if < 1) or increase (if > 1) of inputs, a unit working at the level (x,y) should perform to be dominated by firms producing more than the output level y with probability $1-\alpha$. If $\theta_\alpha(x,y) = 1$, we will say that the unit is input efficient at the level $\alpha \times 100\%$. Clearly when $\alpha = 1$, this is, under free disposability of Ψ, the Farrell-Debreu input efficiency score. In a certain sense, we can say that $\theta_\alpha(x,y)$ is the input efficiency of (x,y) at the level $\alpha \times 100\%$. The same is true in the output direction. We define Ψ^* as being the interior of Ψ.

Proposition 1 *Assume that $F_{X|Y}$ is continuous and monotone increasing in x and that $S_{Y|X}$ is continuous and monotone decreasing in y. Then, for all $(x,y) \in \Psi^*$, there exist α and β in $]0,1]$ such that*

$$\theta_\alpha(x,y) = 1, \qquad \textit{where } \alpha = 1 - F_{X|Y}(x|y) \tag{13}$$

$$\lambda_\beta(x,y) = 1, \qquad \textit{where } \beta = 1 - S_{Y|X}(y|x). \tag{14}$$

Proposition 1 shows that any point (x,y) in the interior of Ψ, belongs to an appropriate α-quantile efficient frontier in both directions (input and output). When $\alpha \to 1$, the α-quantile efficient scores converge monotonically to the Farrell-Debreu efficiency scores:

Proposition 2 *For all y such that $S_Y(y) > 0$, we have $lim_{\alpha\to 1} \searrow \theta_\alpha(x,y) = \theta(x,y)$ and for all x such that $F_X(x) > 0$, $lim_{\alpha\to 1} \nearrow \lambda_\alpha(x,y) = \lambda(x,y)$.*

The α-quantile input efficiency score $\theta_\alpha(x,y)$ is clearly monotone nonincreasing with x but it is in general not monotone in y, unless we add an assumption on $F_{X|Y}$:

Proposition 3 *Assume that $F_{X|Y}(\cdot|y)$ is continuous for any y. Then, the two following properties are equivalent.*

$$F_{X|Y}(x|y) \textit{ is monotone nonincreasing with } y \tag{15}$$

$$\theta_\alpha(x,y) \textit{ is monotone nondecreasing with } y \textit{ for all } \alpha. \tag{16}$$

Points (x,y) here are such that $F_{X|Y}(x|y) < 1$.

Proposition 4 *The two following properties are equivalent.*

$$S_{Y|X}(y|x) \textit{ is monotone nondecreasing with } x \tag{17}$$

$$\lambda_\alpha(x,y) \textit{ is monotone nondecreasing with } x \textit{ for all } \alpha. \tag{18}$$

Points (x,y) here are such that $S_{Y|X}(y|x) < 1$.

4 Nonparametric Estimator

A natural nonparametric estimator of the α-quantile efficiency scores is obtained by plugging the empirical $\widehat{H}_{XY,n}(x,y)$ in the above formulas

$$\widehat{\theta}_{\alpha,n}(x,y) = \inf\{\theta \mid \widehat{F}_{X|Y,n}(\theta x|y) > 1-\alpha\}, \tag{19}$$

$$\widehat{\lambda}_{\alpha,n}(x,y) = \sup\{\lambda \mid \widehat{S}_{Y|X,n}(\lambda y|x) > 1-\alpha\}, \tag{20}$$

These nonparametric estimators can be computed very easily. When $\alpha \to 1$, the estimators converge monotonically to the FDH efficiency scores $\widehat{\theta}_n(x,y)$ and $\widehat{\lambda}_n(x,y)$, respectively:

Proposition 5 *For all y such that $\widehat{H}_{XY,n}(\infty,y) > 0$, we have $lim_{\alpha\to 1} \searrow \widehat{\theta}_{\alpha,n}(x,y) = \widehat{\theta}_n(x,y)$ and for all x such that $\widehat{H}_{XY,n}(x,0) > 0$, $lim_{\alpha\to 1} \nearrow \widehat{\lambda}_{\alpha,n}(x,y) = \widehat{\lambda}_n(x,y)$.*

The asymptotic behavior of our estimator is given by the following theorems (only presented for the output direction: we have the same results for the input oriented case).

Theorem 1 *Let $(x,y) \in \Psi$ be such that $F_X(x) > 0$ and let $0 < \alpha < 1$. Assume that $\lambda \mapsto S_{Y|X}(\lambda y|x)$ is decreasing in a neighborhood of $\lambda_\alpha(x,y)$. Then, for every $\varepsilon > 0$,*

$$Prob(|\widehat{\lambda}_{\alpha,n}(x,y) - \lambda_\alpha(x,y)| > \varepsilon) \leq 2e^{-2n\delta^2_{\varepsilon,x,y}}, \quad \textit{for all } n \geq 1,$$

where

$$\delta_{\varepsilon,x,y} = \frac{F_X(x)}{(2-\alpha)} \min\big\{(1-\alpha) - S_{Y|X}((\lambda_\alpha(x,y)+\varepsilon)y|x) \\ ; S_{Y|X}((\lambda_\alpha(x,y)-\varepsilon)y|x) - (1-\alpha)\big\}.$$

Theorem 2 *Let $0 < \alpha < 1$ be a fixed order and let $(x,y) \in \Psi$ be a fixed unit such that $F_X(x) > 0$. Assume that $G(\lambda) = S_{Y|X}(\lambda y|x)$ is differentiable at $\lambda_\alpha(x,y)$ with negative derivative $G'(\lambda_\alpha(x,y)) =< \bigtriangledown S_{Y|X}(\lambda_\alpha(x,y)y|x), y >$. Then,*

$$\sqrt{n}\left(\widehat{\lambda}_{\alpha,n}(x,y) - \lambda_\alpha(x,y)\right) \xrightarrow{\mathcal{L}} N\left(0, \sigma^2_\alpha(x,y)\right) \quad \textit{as} \quad n \to \infty,$$

where

$$\sigma^2_\alpha(x,y) = \frac{\alpha(1-\alpha)}{[G'(\lambda_\alpha(x,y))]^2 F_X(x)}.$$

A more robust estimator of the Farrell-Debreu efficiency scores $\lambda(x,y)$ than the standard FDH estimator $\widehat{\lambda}_n(x,y)$, which however shares similar asymptotic properties with this latter one, can be derived as follows.

Lemma 1 *Assume that the support of Y is bounded. Then, for any $(x, y) \in \Psi$,*

$$n^{1/(p+q)}\left(\widehat{\lambda}_n(x,y) - \widehat{\lambda}_{\alpha(n),n}(x,y)\right) \xrightarrow{a.s.} 0 \quad as \quad n \to \infty,$$

where the order $\alpha(n) > 0$ is such that

$$n^{(p+q+1)/(p+q)}\left(1 - \alpha(n)\right) \longrightarrow 0 \quad as \quad n \to \infty.$$

Making use of this lemma and the following decomposition

$$\begin{aligned} n^{1/(p+q)}(\lambda(x,y) - \widehat{\lambda}_{\alpha(n),n}(x,y)) &= n^{1/(p+q)}(\lambda(x,y) - \widehat{\lambda}_n(x,y)) \\ &\quad + n^{1/(p+q)}(\widehat{\lambda}_n(x,y) - \widehat{\lambda}_{\alpha(n),n}(x,y)) \end{aligned}$$

we get immediately from Corollary 3.2 of [Park *et al.*, 2000] the following result:

Theorem 3 *Under Assumptions AI-AIII of [Park* et al.*, 2000], we have for any (x, y) interior to Ψ,*

$$n^{1/(p+q)}\left(\lambda(x,y) - \widehat{\lambda}_{\alpha(n),n}(x,y)\right) \xrightarrow{\mathcal{L}} Weibull(\mu_{NW,0}^{p+q}, p+q) \quad as \quad n \to \infty,$$

where $\mu_{NW,0}$ is a constant (see [Park et al.*, 2000]) .*

The latter results show that with an appropriate choice of α, we obtain a nonparametric estimator of the Farrell-Debreu efficiency score $\lambda(x, y)$ sharing the same properties than the FDH estimator, but since it does not envelop all the data points, it will be more robust to extreme and/or outlying observations.

5 Numerical Illustrations

We illustrate the α-quantile efficiency scores and their estimation by using some of simulated data set used in [Daraio and Simar, 2002] with multi-input ($p = 2$) and multi-output ($q = 2$) and Z is favorable to output production. The results are displayed in Figure 1. We see that all the ratios allow to detect the favorable effect of Z on the production process. The α-quantile measures being less sensitive to extreme values, give a better picture.

In order to appreciate the robustness to outliers, and compare the performance of the order-m and of the α-quantile measures, we introduce in the same data set 5 outliers by projecting, in the Y coordinates 5 points in a radial expansion by a factor 1/0.6. The results of this data set with $n = 105$ points are shown in Figure 2. It is clear that the full frontier approach is unable to detect the favorable effect of Z, at least for values larger than the mean of Z (2.5), the order-m does better but again fails for large values of Z. On the contrary, the order-α quantile frontier are much more robust to the 5 outliers and we obtain similar results as in Figure 1, where no outliers where introduced.

References

[Aragon *et al.*, 2002]Y. Aragon, A. Daouia, and C. Thomas-Agnan. Nonparametric frontier estimation: A conditional quantile-based approach. *to appear, Econometric Theory*, 2002.

[Cazals *et al.*, 2002]C. Cazals, J.P. Florens, and L. Simar. Nonparametric frontier estimation: A conditional quantile-based approach. *Journal of Econometrics*, 106:1–25, 2002.

[Daraio and Simar, 2002]C. Daraio and L. Simar. Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *to appear, Journal of Productivity Analysis*, 2002.

[Debreu, 1951]G. Debreu. The coefficient of resource utilization. *Econometrica*, 19(3):273–292, 1951.

[Deprins *et al.*, 1984]D. Deprins, L. Simar, and H. Tulkens. Measuring labor inefficiency in post offices. In M. Marchand, P. Pestieau, and H. Tulkens, editors, *The Performance of Public Enterprises: Concepts and measurements*, pages 243–267. North-Holland, Amsterdam, 1984.

[Koopmans, 1951]T.C. Koopmans. An analysis of production as an efficient combination of activities. In T.C. Koopmans, editor, *Activity Analysis of Production and Allocation*. Cowles Commission for Research in Economics, Monograph 13, John-Wiley and Sons, Inc., New York, 1951.

[Park *et al.*, 2000]B. Park, L. Simar, and Ch. Weiner. The fdh estimator for productivity efficiency scores : Asymptotic properties. *Econometric Theory*, 16:855–877, 2000.

[Simar and Wilson, 2000]L. Simar and P.W. Wilson. The fdh estimator for productivity efficiency scores : Asymptotic properties. *Journal of Productivity Analysis*, 13:49–78, 2000.

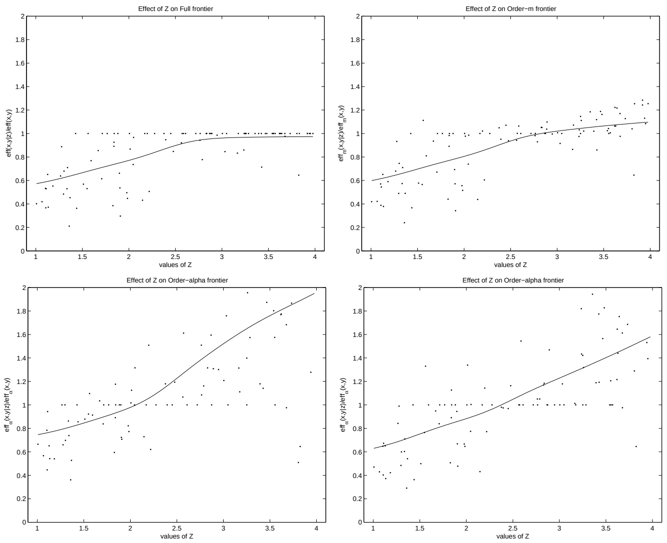

Fig. 1. *Simulated example, $n = 100$: "positive" effect of Z on production efficiency (output oriented framework). Scatterplot and smoothed regression of the ratios $\hat{\lambda}_n(x, y \mid z)/\hat{\lambda}_n(x, y)$ on Z (top left), of $\hat{\lambda}_{m,n}(x, y \mid z)/\hat{\lambda}_{m,n}(x, y)$ on Z (top right, with $m = 25$) and of $\hat{\lambda}_{\alpha,n}(x, y \mid z)/\hat{\lambda}_{\alpha,n}(x, y)$ on Z (bottom panel, left $\alpha = 0.80$ and right $\alpha = 0.90$). Here k-NN=17.*

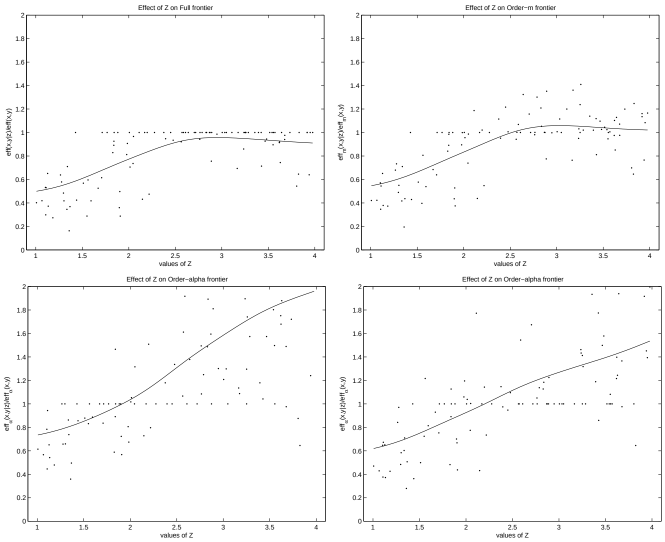

Fig. 2. *Simulated example, $n = 105$ including 5 outliers: "positive" effect of Z on production efficiency (output oriented framework). Scatterplot and smoothed regression of the ratios $\hat{\lambda}_n(x, y \mid z)/\hat{\lambda}_n(x, y)$ on Z (top left), of $\hat{\lambda}_{m,n}(x, y \mid z)/\hat{\lambda}_{m,n}(x, y)$ on Z (top right, with $m = 25$) and of $\hat{\lambda}_{\alpha,n}(x, y \mid z)/\hat{\lambda}_{\alpha,n}(x, y)$on Z (bottom panel, left $\alpha = 0.80$ and right $\alpha = 0.90$). Here k-NN=20.*