T E C H N I C A L

R E P O R T

0318

INTRODUCING ENVIRONMENTAL VARIABLES IN NONPARAMETRIC FRONTIER MODELS: A PROBABILISTIC APPROACH

C. DARAIO and L. SIMAR

I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Introducing Environmental Variables in Nonparametric Frontier Models: a Probabilistic Approach

Cinzia Daraio*
Laboratory of Economics and Management
Scuola Superiore S. Anna, Italy
cinzia@sssup.it

Léopold Simar†
Institut de Statistique
Université Catholique de Louvain, Belgium
simar@stat.ucl.ac.be

May 12, 2003

Abstract

This paper proposes a general formulation of a nonparametric frontier model introducing external environmental factors that might influence the production process but are neither inputs nor outputs under the control of the producer. A representation is proposed in terms of a probabilistic model which defines the data generating process. Our approach extends the basic ideas from Cazals, Florens and Simar (2002) to the full multivariate case. We introduce the concepts of conditional efficiency measure and of conditional efficiency measure of order-m. Afterwards we suggest a practical way for computing the nonparametric estimators. Finally, a simple methodology to investigate the influence of these external factors on the production process is proposed. Numerical illustrations through some simulated examples and through a real data set on Mutual Funds show the usefulness of the approach.

Keywords: production function, frontier, nonparametric estimation, environmental factors, robust estimation.

JEL Classification: C13, C14, D20.

*Research support from the "Progetto Giovani Ricercatori 2002 (EGIOV03CD)", Scuola Superiore S. Anna, is gratefully acknowledged.

†Research support from "Projet d'Actions de Recherche Concertées" (No. 98/03–217) and from the "Interuniversity Attraction Pole", Phase V (No. P5/24) from the Belgian Government are also acknowledged.

1 Introduction

Most of the economic theory on efficiency analysis dates back to Koopmans (1951) and Debreu (1951) on activity analysis. We might 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 $y \in I\!R^q_+$. In this framework the production set is the set of technically feasible combinations of (x, y). It is defined as

$$\Psi = \{(x, y) \in I\!R^{p+q}_+ \mid x \text{ can produce } y\}. \tag{1.1}$$

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[1] $x' \geq x$ and $y' \leq y$. Often convexity of Ψ is also assumed, and so on (see *e.g.* Shephard, 1970, for a modern formulation of the problem).

As far as efficiency is of concern, the boundaries of Ψ are of interest. For instance, if we are looking in the input direction, the Farrell-Debreu measure of input-oriented efficiency score for a unit operating at the level (x, y) is usually defined as:

$$\theta(x, y) = \inf\{\theta \mid (\theta x, y) \in \Psi\}. \tag{1.2}$$

If (x, y) is inside Ψ, $\theta(x, y) \leq 1$ is the proportionate reduction of inputs a unit working at the level (x, y) should perform to achieve efficiency. The corresponding radial efficient frontier in the input space, for units producing a level y of outputs, is defined by points with efficiency scores equal to 1. This frontier is then described as the set $(x^\partial(y), y) \in \Psi$, where $x^\partial(y) = \theta(x, y)x$ is the radial projection of (x, y) on the frontier, in the input direction (orthogonal to the vector y).

If we are looking in the output direction, the Farrell-Debreu measure of output-oriented efficiency score for a unit operating at the level (x, y) is similarly defined as:

$$\lambda(x, y) = \sup\{\lambda \mid (x, \lambda y) \in \Psi\}. \tag{1.3}$$

Here $\lambda(x, y) \geq 1$ represent the proportionate increase of outputs the unit operating at level (x, y) should attain to be considered as being efficient. The efficient frontier corresponds to those points where $\lambda(x, y) = 1$.

In empirical studies, the set Ψ is unknown and so are the efficiency scores. The econometric problem is therefore to estimate these quantities from a random sample of production units $\mathcal{X} = \{(X_i, Y_i) | i = 1, \ldots, n\}$. Since the pioneering work of Farrell (1957), the literature has developed a lot of different approaches to achieve this goal.

[1]From here and below inequalities between vectors have to be understood element by element.

The nonparametric models are particularly appealing since they don't rely on restrictive hypothesis on the Data Generating Process (DGP). The most popular approaches are based on envelopment estimators in the spirit of Farrell approach. Deprins, Simar and Tulkens (1984) have proposed the Free Disposal Hull (FDH) of the set of the observations to estimate Ψ:

$$\widehat{\Psi}_{FDH} = \left\{(x,y) \in I\!\!R_+^{p+q} | y \leq Y_i,\ x \geq X_i, \quad i = 1, \ldots, n\right\}. \tag{1.4}$$

The convex hull of $\widehat{\Psi}_{FDH}$ provides the Data Envelopment Analysis (DEA) estimator of Ψ, popularized as linear programming estimator by Charnes, Cooper and Rhodes (1978):

$$\begin{aligned}\widehat{\Psi}_{DEA} \quad = \quad & \{(x,y) \in I\!\!R_+^{p+q} | y \leq \sum_{i=1}^{n} \gamma_i Y_i \,;\, x \geq \sum_{i=1}^{n} \gamma_i X_i \text{ for } (\gamma_1, \ldots, \gamma_n) \\ & \text{such that } \sum_{i=1}^{n} \gamma_i = 1 \,;\, \gamma_i \geq 0, i = 1, \ldots, n\},\end{aligned} \tag{1.5}$$

it is the smallest free disposal convex set covering all the data. The corresponding estimators of the efficiency scores are then obtained by plugging $\widehat{\Psi}$ in the equations (1.2) and (1.3) above in place of the unknown Ψ.

Today, statistical inference based on DEA/FDH type of estimators is available either by using asymptotic results (Kneip, Park and Simar, 1998 and Park, Simar and Weiner, 2000) or by using the bootstrap, see Simar and Wilson (2000) for a recent survey of the available results. In summary, if the true attainable set is free disposal, then $\widehat{\Psi}_{FDH}$ is a consistent estimator of Ψ, but $\widehat{\Psi}_{DEA}$ is not. If Ψ is free disposal and convex then both estimators are consistent, but the DEA estimator takes advantage of the convexity assumption and achieves a slightly faster rate of convergence.

During the last decades, the literature on efficiency estimation has been extended to explore the reasons of different level of efficiencies across production units. The idea was to relate efficiency measures to some external or environmental factors which might influence the production process but that are not under the control of the producers.

The evaluation of the influence of environmental factors on the efficiency of producers is indeed a relevant issue related to the explanations of efficiency, the identification of economic conditions that create inefficiency, and finally to the improvement of managerial performance.

When categorical factors are considered (like the form of ownership,...), we are in the presence of different groups of producers; in this situation, testing issues for comparing group efficiency scores can be proposed using appropriate bootstrap algorithms (in the spirit of Simar and Wilson, 2002). When these external factors $z \in I\!\!R^r$ are continuous mainly two approaches have been proposed in literature but both are flawed by restrictive prior assumptions on the DGP and/or on the role of these external factors on the production process.

The first family of models is based on a *one-stage* approach (see *e.g.* Banker and Morey, 1986; Fare, Grosskopf, Lovell and Pasurka, 1989; Fare, Grosskopf and Lovell, 1994, p. 223-226), where these factors z are considered as free disposal inputs and /or outputs which contribute to define the attainable set $\Psi \subset I\!R_+^p \times I\!R_+^q \times I\!R^r$, but which are not active in the optimization process defining the efficiency scores. For instance, the analog of (1.2), would be:

$$\theta(x, y|z) = \inf\{\theta \mid (\theta x, y, z) \in \Psi\}, \tag{1.6}$$

and the estimator of Ψ is defined as above by adding the variables z in defining the FDH and /or the DEA enveloping set, with a variable z being considered as an input if it is conducive (favorable, advantageous, beneficial) to efficiency and as an output if it is detrimental (damaging, unfavorable) to efficiency. The drawback of this approach is twofold: first we have to know *a priori* what is the role of z on the production process, and second we assume the free disposability (and eventually convexity, if DEA is used) of the corresponding attainable extended set Ψ.

The second family of models is based on a *two-stage* approach. Here the estimated efficiency scores are regressed, in an appropriated limited dependent variable parametric regression model (like truncated normal regression models) on the environmental factors z. Some models in this family propose also three-stage and four-stage analysis as extension of the two-stage approach (for more details see Fried, Schmidt, and Yaisawarng 1999; Fried, Lovell, Schmidt, and Yaisawarng 2002). As pointed out by Simar and Wilson (2003), most of these models are flawed by the fact that usual inference on the obtained estimates of the regression coefficient is not available. Simar and Wilson (2003) give a list of references where this approach has been used and propose a bootstrap algorithm to obtain more accurate inference. However, also this bootstrap-based approach, even when corrected, has two inconveniences. First, it relies on a separability condition between the input $\times$ output space and the space of values for z: the extended attainable set is the cartesian product $\Psi \times I\!R^r$ and so the value of z does not influence the position of the frontier of the attainable set. Second, the regression in the second stage relies on some parametric assumptions (like linear model and truncated normal error term).

In this paper, we propose a more general full nonparametric approach which overcomes most of the drawbacks mentioned above. It relies on a probabilistic definition of the frontier and of the efficiency which is equivalent to the definition proposed above but allows an easy introduction of environmental factors. The basic ideas where proposed in Cazals, Florens and Simar (2002) (from hereafter CFS). Here, we extend to a more general multivariate setup and we propose a practical methodology to evaluate the estimators. We will define conditional efficient frontier and also conditional order-m frontier and their corresponding nonparametric estimators. In particular, order-m frontier estimators are known as being more robust to outliers and/or extreme values than the full frontier estimates. We also suggest an easy procedure for evaluating the impact of these environmental factors on the production process.

The paper is organized as follows. The next section introduces the multivariate probabilistic model for defining the DGP of a production process. This section includes also the definition of the full frontier and of the order-m frontier. Section 3 shows how this framework can easily be adapted to the introduction of environmental factors. Section 4 addresses some practical computational issues and Section 5 illustrates the methodology by using some simulated data sets and a real data set on mutual funds. Section 6 concludes.

2 Production Frontiers: a probabilistic formulation

The production process is here described by the joint probability measure of (X, Y) on $\mathbb{R}_+^p \times \mathbb{R}_+^q$. The support of (X, Y) is the attainable set Ψ. In terms of the joint probability measure of (X, Y), the Farrell-Debreu input efficiency defined in (1.2) can also be characterized, under free disposability, as:

$$\theta(x, y) = \inf\{\theta \mid F_X(\theta x \mid y) > 0\}, \tag{2.1}$$

where $F_X(x \mid y) = \text{Prob}(X \leq x \mid Y \geq y)$.

A nonparametric estimator of $\theta(x, y)$ can be provided by plugging the empirical version of $F_X(x \mid y)$ in (2.1) given by

$$\widehat{F}_{X,n}(x \mid y) = \frac{\sum_{i=1}^n \mathbb{I}(X_i \leq x, Y_i \geq y)}{\sum_{i=1}^n \mathbb{I}(Y_i \geq y)}, \tag{2.2}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Then, the estimator of the input efficiency score for a given point (x, y) is the solution of

$$\hat{\theta}_n(x, y) = \inf\{\theta \mid \widehat{F}_{X,n}(\theta x \mid y) > 0\}. \tag{2.3}$$

Now, as pointed in CFS, this coincides to the FDH estimator of $\theta(x, y)$ given by

$$\hat{\theta}_n(x, y) = \inf\{\theta \mid (\theta\, x, y) \in \widehat{\Psi}_{FDH}\} = \min_{i \mid Y_i \geq y} \left\{ \max_{j=1,\ldots,p} \left(\frac{X_i^j}{x^j} \right) \right\}, \tag{2.4}$$

where a^j denotes the jth component of a vector a.

We know that under the free disposal assumption, this is a consistent estimator of $\theta(x,y)$ but with a poor rate of convergence $n^{1/(p+q)}$: this is the curse of dimensionality shared by most nonparametric estimators (see Park, Simar and Weiner, 2000 for more properties of $\hat{\theta}_n(x,y)$).

The FDH estimator $\widehat{\Psi}_{FDH}$ is very sensitive to extreme points, since as an estimator of the full-frontier, it envelops all the cloud of points $\mathcal{X}$. Therefore, CFS propose to estimate an order-m frontier, which corresponds to another definition of the benchmark against which units will be compared. The idea can be summarized as follows (we extend somewhat the presentation of CFS, introducing here the concept of order-m efficiency).

For a given level of outputs y in the interior of the support of Y, consider now m i.i.d. random variables $X_i, i = 1, \ldots, m$ generated by the conditional p-variate distribution function $F_X(x \mid y)$ and define the set:

$$\Psi_m(y) = \{(x, y') \in I\!R_+^{p+q} \mid x \geq X_i, y' \geq y, i = 1, \ldots, m\}. \tag{2.5}$$

Then, for any x, we may define

$$\tilde{\theta}_m(x,y) = \inf\{\theta \mid (\theta x, y) \in \Psi_m(y)\}. \tag{2.6}$$

Note that $\tilde{\theta}_m(x,y)$ may be computed by the following formula:

$$\tilde{\theta}_m(x,y) = \min_{i=1,\ldots,m} \left\{ \max_{j=1,\ldots,p} \left(\frac{X_i^j}{x^j} \right) \right\}. \tag{2.7}$$

$\tilde{\theta}_m(x,y)$ is a random variable since the X_i are random variables generated by $F_X(x \mid y)$.

Now, adapting the Definition 5.1 in CFS for the expected order-m frontier, we can define the expected order-m input efficiency measure, or in shorter, the order-m input efficiency measure as follows:

Definition 2.1 *For any $x \in I\!R_+^p$, the (expected) order-m input efficiency measure denoted by $\theta_m(x,y)$ is defined for all y in the interior of the support of Y as:*

$$\theta_m(x,y) = E(\tilde{\theta}_m(x,y) \mid Y \geq y), \tag{2.8}$$

where we assume the existence of the expectation.

So, in place of looking for the lower boundary of the support of $F_X(x \mid y)$, as was typically the case for the full-frontier and for the efficiency score $\theta(x,y)$, the order-m efficiency score can be viewed as the expectation of the minimal input efficiency score of the unit (x,y), when compared to m units randomly drawn from the population of units producing more outputs than the level y. This is certainly a less extreme benchmark for the unit (x,y) than the "absolute" minimal achievable level of inputs: it is compared to a set of m peers producing more than its level y and we take as benchmark, the expectation of the minimal achievable input in place of the absolute minimal achievable input.

Note that the order-m efficiency score is not bounded by 1: a value of $\theta_m(x,y)$ greater than one indicates that the unit operating at the level (x,y) is more efficient than the average of m peers randomly drawn from the population of units producing more output than y. Then for any $x \in \mathbb{R}^p_+$, the expected minimum level of inputs of order-m is defined as $x^\partial_m(y) = \theta_m(x,y)\, x$ which can be compared with the full-frontier $x^\partial(y) = \theta(x,y)\, x$. From Theorem 5.1 and Theorem 5.2 of CFS one immediately obtains:

Theorem 2.1 *For any $x \in \mathbb{R}^p_+$ and for all y in the interior of the support of Y, if $\theta_m(x,y)$ exists, we have:*

$$\begin{aligned}\theta_m(x,y) &= \int_0^\infty (1 - F_X(ux \mid y))^m du &(2.9)\\ &= \theta(x,y) + \int_{\theta(x,y)}^\infty (1 - F_X(ux \mid y))^m du, &(2.10)\\ \lim_{m\to\infty} \theta_m(x,y) &= \theta(x,y). &(2.11)\end{aligned}$$

A nonparametric estimator of $\theta_m(x,y)$ is straightforward: we replace the true $F_X(\cdot \mid y)$ by its empirical version, $\widehat{F}_{X,n}(\cdot \mid y)$. We have

$$\begin{aligned}\hat{\theta}_{m,n}(x,y) &= \widehat{E}(\tilde{\theta}_m(x,y) \mid Y \geq y)\\ &= \int_0^\infty (1 - \widehat{F}_{X,n}(ux \mid y))^m du, &(2.12)\\ &= \hat{\theta}_n(x,y) + \int_{\hat{\theta}_n(x,y)}^\infty (1 - \widehat{F}_{X,n}(ux \mid y))^m du &(2.13)\end{aligned}$$

This leads to an estimator of the frontier, which for finite m, does not envelop all the observed data points and so, is less sensitive to extreme points and /or to outliers. As shown by (2.13), as m increases and for fixed n, $\hat{\theta}_{m,n}(x,y) \to \hat{\theta}_n(x,y)$. Simar (2003) proposes a semi-automatic procedure to flag potential outliers by investigating the convergence of $\hat{\theta}_{m,n}(x,y)$ to $\hat{\theta}_n(x,y)$ as m increases: if $\hat{\theta}_{m,n}(x,y)$ is still larger than 1 even for large values of m, then (x,y) could be an extreme points of the cloud $\mathcal{X}$.

CFS analyze the asymptotic properties of the proposed nonparametric estimators. In particular, they show the $\sqrt{n}$-consistency of $\hat{\theta}_{m,n}(x,y)$ to $\theta_m(x,y)$ for m fixed, as $n \to \infty$. Note that we avoid the curse of dimensionality for the nonparametric estimator of the order-m efficiency.

We now briefly sketch the main differences for the output oriented case. The Farrell-Debreu output efficiency score can be characterized as

$$\lambda(x,y) = \sup\{\lambda \mid S_Y(\lambda y \mid x) > 0\}, \tag{2.14}$$

where $S_Y(y \mid x) = \text{Prob}(Y \geq y \mid X \leq x)$. A nonparametric estimator of $\lambda(x, y)$ is provided by the empirical version of $S_Y(y \mid x)$:

$$\widehat{S}_{Y,n}(y \mid x) = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i \leq x, Y_i \geq y)}{\sum_{i=1}^{n} \mathbb{1}(X_i \leq x)}. \tag{2.15}$$

Then, the estimator of the output efficiency score for a given point (x, y) is the solution of

$$\hat{\lambda}_n(x, y) = \sup\{\lambda \mid \widehat{S}_{Y,n}(\lambda y \mid x) > 0\}, \tag{2.16}$$

which coincides to the FDH estimator:

$$\hat{\lambda}_n(x, y) = \sup\{\lambda \mid (x, \lambda y) \in \widehat{\Psi}_{FDH}\} = \max_{i \mid X_i \leq x} \left\{ \min_{j=1,\dots,p} \left(\frac{Y_i^j}{y^j} \right) \right\}, \tag{2.17}$$

For a given level of inputs x in the interior of the support of X, consider m i.i.d. random variables $Y_i, i = 1, \dots, m$ generated by the conditional q-variate distribution function $F_Y(y \mid x) = \text{Prob}(Y \leq y \mid X \leq x)$ and define the set:

$$\Psi_m(x) = \{(x', y) \in \mathbb{R}_+^{p+q} \mid x' \leq x, Y_i \leq y, i = 1, \dots, m\}. \tag{2.18}$$

Then, for any y, we may define

$$\begin{aligned} \tilde{\lambda}_m(x, y) &= \sup\{\lambda \mid (x, \lambda y) \in \Psi_m(x)\} && (2.19) \\ &= \max_{i=1,\dots,m} \left\{ \min_{j=1,\dots,p} \left(\frac{Y_i^j}{y^j} \right) \right\}. && (2.20) \end{aligned}$$

The order-m output efficiency measure is defined as follows.

Definition 2.2 *For any $y \in \mathbb{R}_+^q$, the (expected) order-m output efficiency measure denoted by $\lambda_m(x, y)$ is defined for all x in the interior of the support of X as:*

$$\lambda_m(x, y) = E(\tilde{\lambda}_m(x, y) \mid X \leq x), \tag{2.21}$$

where we assume the existence of the expectation.

As above, we obtain

Theorem 2.2 *For any $y \in \mathbb{R}_+^q$ and for all x in the interior of the support of X, if $\lambda_m(x, y)$ exists, we have:*

$$\begin{aligned} \lambda_m(x, y) &= \int_0^\infty [1 - (1 - S_Y(uy \mid x))^m] \, du && (2.22) \\ &= \lambda(x, y) - \int_0^{\lambda(x,y)} (1 - S_Y(uy \mid x))^m du, && (2.23) \\ \lim_{m \to \infty} \lambda_m(x, y) &= \lambda(x, y). && (2.24) \end{aligned}$$

A nonparametric estimator of $\lambda_m(x, y)$ is given by:

$$\begin{aligned} \hat{\lambda}_m(x, y) &= \int_0^\infty \left[1 - (1 - \widehat{S}_{Y,n}(uy \mid x))^m\right] du && (2.25) \\ &= \hat{\lambda}_n(x, y) - \int_0^{\hat{\lambda}_n(x,y)} (1 - \widehat{S}_{Y,n}(uy \mid x))^m du. && (2.26) \end{aligned}$$

3 Introducing Environmental Factors

The analysis of the preceding section can easily be extended to the case where additional information is provided by other variables $Z \in I\!R^r$, exogenous to the production process itself, but which may explain part of it. The basic idea for introducing this additional information in the model is to condition the production process to a given value of $Z = z$. CFS propose the idea for order-m frontiers and for the univariate case (one input for the input oriented case or one output for the output oriented case). We propose below a more general presentation inspired from Section 5.1 and 5.2 of CFS, allowing to handle the multi-input (multi-output) and the full frontier cases. To save place, we describe the basic ideas in the input oriented framework. Practical computational issues are addressed in Section 4.

The joint distribution on (X, Y) conditional on $Z = z$ defines the production process if $Z = z$. In particular the efficiency measure defined above in (2.1) has to be adapted to the condition $Z = z$ as follows:

$$\theta(x, y \mid z) = \inf\{\theta \mid F_X(\theta x \mid y, z) > 0\}, \tag{3.1}$$

where $F_X(x \mid y, z) = \text{Prob}(X \leq x \mid Y \geq y, Z = z)$.

A nonparametric estimator of the conditional full-frontier efficiency $\theta(x, y|z)$ is given by plugging a nonparametric estimator of $F_X(x|y, z)$. This requires some smoothing techniques in z. At this purpose we use a kernel estimator of $F_X(x|y, z)$ defined as:

$$\widehat{F}_{X,n}(x \mid y, z) = \frac{\sum_{i=1}^n I\!I(x_i \leq x, y_i \geq y) K((z - z_i)/h_n)}{\sum_{i=1}^n I\!I(y_i \geq y) K((z - z_i)/h_n}, \tag{3.2}$$

where $K(\cdot)$ is the kernel and h_n is the bandwidth of appropriate size (we discuss practical bandwidth selection issues in the next section). Hence, we obtain the "conditional FDH efficiency measure" as follows:

$$\hat{\theta}_n(x, y \mid z) = \inf\{\theta \mid \widehat{F}_{X,n}(\theta x \mid y, z) > 0\}. \tag{3.3}$$

Note that the asymptotic properties of $\hat{\theta}_n(x, y \mid z)$ have not yet been derived in the literature, but we might expect that the rate of convergence of the usual FDH estimator will deteriorate with the dimension of Z, due to the smoothing in getting $\widehat{F}_{X,n}(x \mid y, z)$.

The conditional order-m input efficiency measure is introduced accordingly. For a given level of outputs y in the interior of the support of Y, consider the m i.i.d. random variables $X_i, i = 1, \ldots, m$ generated by the conditional p-variate distribution function $F_X(x \mid y, z)$ and define the set:

$$\Psi_m^z(y) = \{(x, y') \in I\!R_+^{p+q} \mid x \geq X_i, y' \geq y\}. \tag{3.4}$$

Note that this set depends on the value of z since the X_i are generated through $F_X(x \mid y, z)$. Then, for any x, we may define

$$\tilde{\theta}^z_m(x, y) = \inf\{\theta \mid (\theta x, y) \in \Psi^z_m(y)\}. \tag{3.5}$$

Note that $\tilde{\theta}^z_m(x, y)$ may be computed by the following formula:

$$\tilde{\theta}^z_m(x, y) = \min_{i=1,\ldots,m} \left\{ \max_{j=1,\ldots,p} \left(\frac{X_i^j}{x^j} \right) \right\}. \tag{3.6}$$

Now we can define the conditional order-m input efficiency measure by following the idea of Definition 2.1.

Definition 3.1 *For any $x \in I\!R^p_+$, the conditional order-m input efficiency measure given that $Z = z$, denoted by $\theta_m(x, y|z)$ is defined for all y in the interior of the support of Y as:*

$$\theta_m(x, y|z) = E(\tilde{\theta}^z_m(x, y) \mid Y \geq y, Z = z), \tag{3.7}$$

where we assume the existence of the expectation.

Therefore, for any $x \in I\!R^p_+$, the expected minimum level of inputs of order m, given that $Z = z$, is defined as $x^{\partial}_m(y|z) = \theta_m(x, y|z)\, x$. As above we have immediately the following theorem.

Theorem 3.1 *For any $x \in I\!R^p_+$ and for all y in the interior of the support of Y, if $\theta_m(x, y|z)$ exists, we have:*

$$\theta_m(x, y|z) = \int_0^\infty (1 - F_X(ux \mid y, z))^m du, \tag{3.8}$$

$$\lim_{m \to \infty} \theta_m(x, y|z) = \theta(x, y|z). \tag{3.9}$$

A nonparametric estimator of $\theta_m(x, y|z)$ is provided by plugging the nonparametric estimator of $F_X(x|y, z)$ proposed above in (3.2). As showed in CFS, the resulting estimator of the order-m efficiency measure achieves the rate of convergence $\sqrt{nh^r_n}$, where $r = \dim(Z)$, so here, due to the smoothing in Z, we cannot avoid the curse of dimensionality in the dimension of Z.

Formally, the estimator is obtained as follows

$$\hat{\theta}_{m,n}(x, y|z) = \widehat{E}(\tilde{\theta}^z_m(x, y) \mid y, z) = \int_0^\infty (1 - \widehat{F}_{X,n}(ux \mid y, z))^m du \tag{3.10}$$

where $\tilde{\theta}^z_m(x, y)$ is defined above in (3.6), and the m random variables X_i are generated according to the estimated $\widehat{F}_{X,n}(x \mid y, z)$. For a given kernel and a given bandwidth, the univariate integral in (3.10) can be evaluated for any point (x, y) and for any level of the environmental factors $Z = z$, by using an appropriate numerical method. Note that here again, for a fixed value of n we have $\lim_{m\to\infty} \hat{\theta}_{m,n}(x, y|z) = \hat{\theta}_n(x, y \mid z)$.

The derivations of the formulae for the definition and the estimation of the conditional output efficiency scores (full-frontier and order-m) are obtained in a similar way by replacing in Section 2, $S_Y(\lambda y \mid x)$ by $S_Y(\lambda y \mid x, z)$ and $\widehat{S}_{Y,n}(\lambda y \mid x)$ by $\widehat{S}_{Y,n}(\lambda y \mid x, z)$.

4 Practical computations

Again, the presentation here is limited to the input oriented case to save place.

FDH and conditional FDH efficiency estimates

For any given point (x, y), the FDH estimator $\hat{\theta}_n(x, y)$ is very easy and fast to compute. The operational formula comes from (2.4):

$$\hat{\theta}_n(x, y) = \inf\{\theta \mid \widehat{F}_{X,n}(\theta x \mid y) > 0\} = \min_{i|Y_i \geq y} \left\{ \max_{j=1,\ldots,p} \left(\frac{X_i^j}{x^j} \right) \right\}. \tag{4.1}$$

It is easy to show that for any (symmetric) kernel with compact support ($K(u) = 0$ if $|u| > 1$, as for the uniform, triangle, epanechnikov or quartic kernels), the conditional FDH efficiency estimator is given by:

$$\hat{\theta}_n(x, y|z) = \inf\{\theta \mid \widehat{F}_{X,n}(\theta x \mid y, z) > 0\} = \min_{\{i|Y_i \geq y, |Z_i - z| \leq h\}} \left\{ \max_{j=1,\ldots,p} \left(\frac{X_i^j}{x^j} \right) \right\}, \tag{4.2}$$

where h is the chosen bandwidth. It is interesting to note that our plug-in estimates $\widehat{F}_{X,n}(\theta x \mid y, z) > 0$ is such that for kernels with unbounded support, like the gaussian kernel, $\hat{\theta}_n(x, y|z) \equiv \hat{\theta}_n(x, y)$: the estimate of the full-frontier efficiency is unable to detect any influence of the environmental factors. Therefore, in this framework of conditional boundary estimation, kernels with compact support have to be used.

Order-m and conditional order-m efficiencies

For the order-m efficiency $\hat{\theta}_{m,n}(x, y)$ and $\hat{\theta}_{m,n}(x, y|z)$, the univariate integrals (2.13) and (3.10) could be evaluated by numerical methods[2], even when $p \geq 1$. The algorithms are very fast: the computation of such integrals for one point, is of the order of a hundredth of a second on a "old" Pentium III, 450 Mghz machine. However numerical integration can be avoided by an easy Monte-Carlo algorithm (proposed in CFS for the order-m frontier), that

[2]All the Matlab codes allowing to compute the input and/or output oriented efficiency measures described in this paper (unconditional, conditional to $Z = z \in I\!R$ and of any order m) are freely available on request at simar@stat.ucl.ac.be. When numerical integration is required, the build-in Matlab "quad" procedure (based on adaptive Simpson quadrature) is used.

we describe below, as fast for small values of m such as $m = 10$, but much slower when m increases:

[1] For a given y, draw a sample of size m with replacement among those X_i such that $Y_i \geq y$ and denote this sample by $(X_{1,b}, \ldots, X_{m,b})$;

[2] Compute $\tilde{\theta}^b_m(x, y) = \min_{i=1,\ldots,m} \left\{ \max_{j=1,\ldots,p} \left(\frac{X^j_{i,b}}{x^j} \right) \right\}$.

[3] Redo [1]-[2] for $b = 1, \ldots, B$, where B is large.

[4] Finally, $\hat{\theta}_{m,n}(x, y) \approx \frac{1}{B} \sum_{b=1}^{B} \tilde{\theta}^b_m(x, y)$.

The quality of the approximation can be tuned by increasing B but in most applications, say $B = 200$, seems to be a reasonable choice (see Simar, 2003, for a code written in Matlab).

This Monte-Carlo algorithm can be adapted as follows for the conditional order-m efficiency score. Suppose that h is the chosen bandwidth for a particular kernel $K(\cdot)$:

[1] For a given y, draw a sample of size m with replacement, and with a probability $K((z - z_i)/h)/ \sum_{j=1}^{n} K((z - z_j)/h)$, among those X_i such that $Y_i \geq y$. Denote this sample by $(X_{1,b}, \ldots, X_{m,b})$;

[2] compute $\tilde{\theta}^{b,z}_m(x, y) = \min_{i=1,\ldots,m} \left\{ \max_{j=1,\ldots,p} \left(\frac{X^j_{i,b}}{x^j} \right) \right\}$.

[3] Redo [1]-[2] for $b = 1, \ldots, B$, where B is large.

[4] Finally, $\hat{\theta}_{m,n}(x, y \mid z) \approx \frac{1}{B} \sum_{b=1}^{B} \tilde{\theta}^{b,z}_m(x, y)$.

Bandwidth selection: a simple data-driven method

It is well known that the choice of the bandwidth is important in nonparametric smoothing. We propose in this paper a very simple and easy to compute rule based on a k-Nearest Neighbor (k-NN) method.

The idea is that the smoothing in computing our Z-conditional efficiency estimators (3.3) and (3.10), comes from the smoothing in the estimation of the conditional distribution function $\widehat{F}_{X,n}(x \mid y, z)$ (see equation (3.2)). This is due to the continuity of the variable Z. Hence, we suggest in a first step to select a bandwidth h which optimizes in a certain sense the estimation of the density of Z. We propose to use the likelihood cross validation criterion (see Silverman, 1986 for details), using a k-NN method: this allows to obtain bandwidths which are localized, insuring we have always the same number of observations Z_i in the local neighbor of the point of interest z when estimating the density of Z.

So, for a grid of values of k, we evaluate the leave-one-out kernel density estimate of Z, $\hat{f}_k^{(-i)}(Z_i)$ for $i = 1, \dots, n$ and find the value of k which maximizes the score function:

$$CV(k) = n^{-1} \sum_{i=1}^{n} \log \left(\hat{f}_k^{(-i)}(Z_i) \right),$$

where

$$\hat{f}_k^{(-i)}(Z_i) = \frac{1}{(n-1)h_{Z_i}} \sum_{j=1, j \neq i}^{n} K\left(\frac{Z_j - Z_i}{h_{Z_i}} \right),$$

and h_{Z_i} is the local bandwidth chosen such that there exist k points Z_j verifying $|Z_j - Z_i| \leq h_{Z_i}$.

Afterwards, in a second step, in order to compute $\widehat{F}_{X,n}(x \mid y, z)$, we have to take into account for the dimensionality of x and y, and the sparsity of points in larger dimensional spaces. Consequently, we expand the local bandwidth h_{Z_i} by a factor $1+n^{-1/(p+q)}$, increasing with $(p+q)$ but decreasing with n.

Stressing the influence of Z on the production process

The comparison of $\hat{\theta}_n(x, y \mid z)$ with $\hat{\theta}_n(x, y)$ is certainly of interest for analyzing the global influence of Z on the production process. When Z is univariate, a scatter plot of the ratios[3] $\hat{\theta}_n(x, y \mid z)/\hat{\theta}_n(x, y)$ against Z and its smoothed nonparametric regression line would be helpful to describe the influence of Z on efficiency. If this regression is increasing, it indicates that Z is detrimental (unfavorable) to efficiency and when this regression is decreasing, it specifies a Z factor conducive (favorable) to efficiency.

We recall indeed that here we are in an input oriented framework. In the first case (unfavorable Z) the environmental variable acts like an "extra" *undesired* output to be produced asking for the use of more inputs in production activity, hence Z has a "negative" effect on the production process. In this case $\hat{\theta}_n(x, y \mid z)$, the efficiency computed taking Z into account, will be much larger than the unconditional efficiency $\hat{\theta}_n(x, y)$ for large values of Z then for small value of Z. Consequently, the ratios $\hat{\theta}_n(x, y \mid z)/\hat{\theta}_n(x, y)$ will increase, on average, with Z.

In the second case (favorable Z), the environmental variable plays a role of a "substitutive" input in the production process, giving the opportunity to "save" inputs in the activity of production; in this case, Z has a "positive" effect on the production process. It follows that the conditional efficiency $\hat{\theta}_n(x, y \mid z)$ will be much larger than $\hat{\theta}_n(x, y)$ for small values of Z (less substitutive inputs) than for large values of Z. Therefore, the ratios $\hat{\theta}_n(x, y \mid z)/\hat{\theta}_n(x, y)$ will, on average, decrease when Z increases.

[3]We can do the same with the differences $\hat{\theta}_n(x, y \mid z) - \hat{\theta}_n(x, y)$, but since efficiency scores are proportions, ratios seem very natural.

Since we know that full-frontier estimates, and the derived estimated efficiency scores, are very sensitive to outliers and extreme values, we do also the same analysis for the more robust order-m efficiency scores. Thus we present also the nonparametric smoothed regression of the ratios $\hat{\theta}_{m,n}(x, y \mid z)/\hat{\theta}_{m,n}(x, y)$ on Z. This could be done for some values of m, knowing that when m increases, this converges to the preceding case (full-frontier). As pointed in CFS, m can also be viewed as a trimming parameter and several values of m could be used to provide a sensitivity analysis. This allows to detect potential outliers, *i.e.*, points such that their order-m efficiency scores are still larger than 1, even when m increases, see Simar (2003).

Mutatis mutandis, the same could be done in the output oriented case, with similar conclusions to detect the influence of Z on efficiency. In this case, the influence of Z goes in the opposite direction: an increasing regression corresponds to favorable environmental factor and a decreasing regression indicates an unfavorable factor. In an output oriented framework, a favorable Z means that the environmental variable operates as a sort of "extra" input *freely available*: for this reason the environment is "favorable" to the production process. Consequently, the value of $\hat{\lambda}_n(x, y \mid z)$ will be much smaller (greater efficiency) than $\hat{\lambda}_n(x, y)$ for small values of Z than for large values of Z: the ratios $\hat{\lambda}_n(x, y \mid z)/\hat{\lambda}_n(x, y)$ will increase with Z, on average.

In the case of unfavorable Z, the environmental variable works as a "compulsory" or *unavoidable* output to be produced to face the negative environmental condition. Z in a certain sense penalizes the production of the outputs of interest. In this situation, $\hat{\lambda}_n(x, y|z)$ will be much smaller than $\hat{\lambda}_n(x, y)$ for large values of Z. As a result, the regression line of $\hat{\lambda}_n(x, y \mid z)/\hat{\lambda}_n(x, y)$ over Z will be decreasing.

Of course, we do not propose any inference here, but only an easy and useful descriptive diagnostic tool.

5 Empirical illustrations

5.1 Classroom simulated data sets

We begin with some very simple simulations where all the units produce the same quantity of output by using a single input X. Now suppose that Z is unfavorable to the production process (suppose each unit has to produce 1 liter of ice from water at 20 degrees centigrade: X is the required energy and Z is the environmental temperature: if Z is large, the process, even efficient will require more input). We simulated a sample size $n = 100$ from $Z \sim$ Uniform$(1, 10)$ and compare 3 different scenarios for generating X.

- Case 1: $X = Z^{3/2} + \varepsilon$, here Z is unfavorable to X for all values of its range and ε is the random true inefficiency $\varepsilon \sim \text{Exp}(3)$, *i.e.* an exponential r.v. with mean 3.

- Case 2: $X = 5^{3/2} + \varepsilon$, here Z is independent of X and ε is as above.

- Case 3: $X = 5^{3/2}\, \mathbb{1}(Z \le 5) + Z^{3/2}\, \mathbb{1}(Z > 5) + \varepsilon$, *i.e.*, the unfavorable effect of Z on X starts only after the value of Z larger than 5, with the same inefficiency term ε.

We computed the FDH, conditional to Z-FDH, the order-m and conditional Z-order-m input efficiency scores of all the 100 units. For the illustration, we have chosen here $m = 25$. For larger values of m the results converge very quickly to the full-frontier results. We present the results for a triangle kernel (we obtain very similar results with other kernels with compact support).

The 3 following pictures (Figures 1 to 3) illustrate how the nonparametric regression of the ratios between the conditional and the unconditional efficiency measures on Z is able to capture the real effect of Z on the production process. We recover exactly what we expected through the 3 different simulation scenarios. So it seems that our estimation procedure works pretty well.

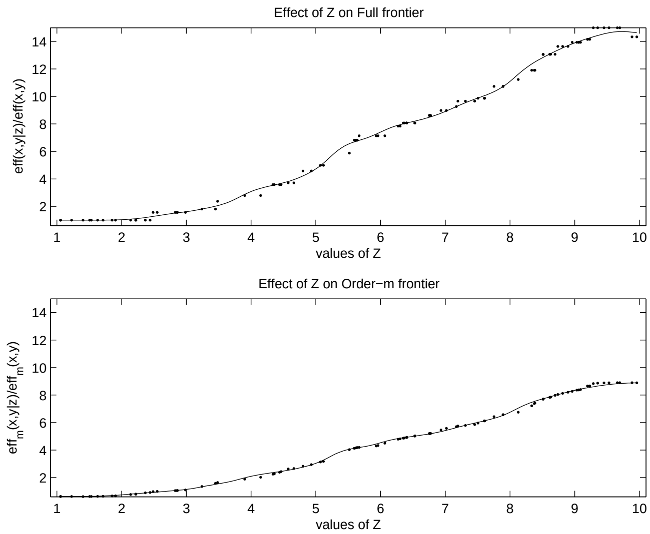

Figure 1: *Classroom example, case 1, "unfavorable" effect of Z on production efficiency (input oriented framework). Scatterplot and smoothed regression of $\hat{\theta}_n(x, y \mid z)/\hat{\theta}_n(x, y)$ on Z (top) and of $\hat{\theta}_{m,n}(x, y \mid z)/\hat{\theta}_{m,n}(x, y)$ on Z, with $m = 25$ (bottom). Here k-NN=19.*

Table 1 gives the average values of the 4 different input efficiency measures for the 3 cases. Again we obtain the expected results under the 3 different scenarios. For instance, in case 2, the true mean efficiency score is about $5^{3/2}/(5^{3/2} + 3) \approx 0.79$. Note that the full detailed

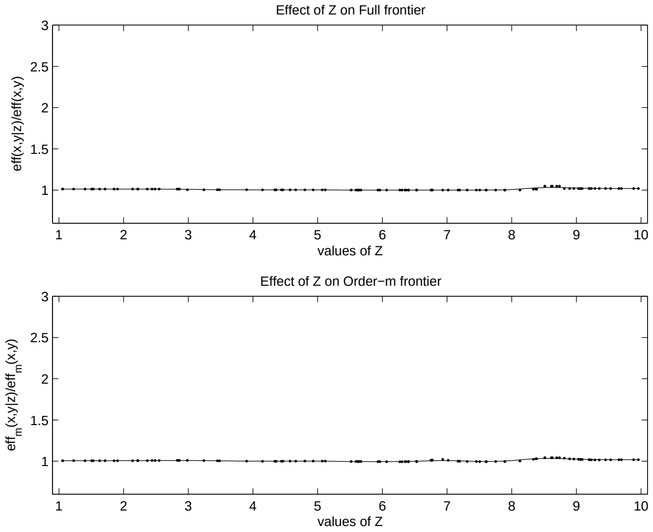

Figure 2: *Classroom example, case 2, no effect of Z on production efficiency (input oriented framework). Scatterplot and smoothed regression of* $\hat{\theta}_n(x, y \mid z)/\hat{\theta}_n(x, y)$ *on* Z *(top) and of* $\hat{\theta}_{m,n}(x, y \mid z)/\hat{\theta}_{m,n}(x, y)$ *on* Z*, with* $m = 25$ *(bottom). Here k-NN=19.*

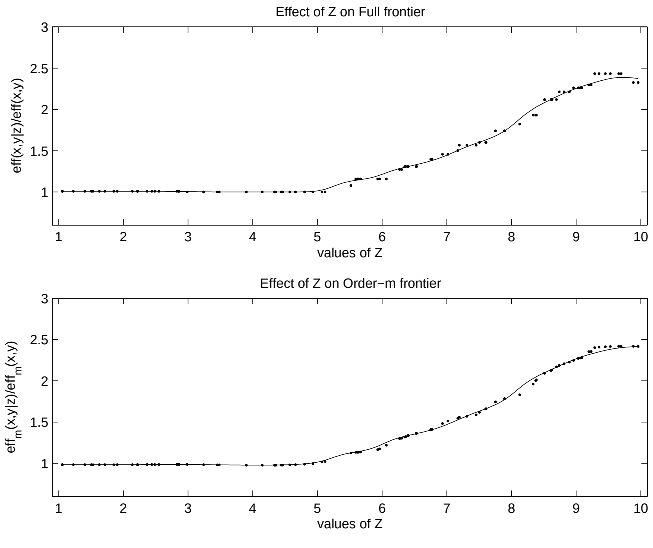

Figure 3: *Classroom example, case 3, "unfavorable" effect of Z on production efficiency, only after* $Z > 5$ *(input oriented framework). Scatterplot and smoothed regression of* $\hat{\theta}_n(x, y \mid z)/\hat{\theta}_n(x, y)$ *on* Z *(top) and of* $\hat{\theta}_{m,n}(x, y \mid z)/\hat{\theta}_{m,n}(x, y)$ *on* Z*, with* $m = 25$ *(bottom). Here k-NN=19.*

table of results for the 100 units (not reproduced here to save place) provides two interesting information: for each unit (X_i, Y_i), the number of dominating units N, *i.e.*, the number of points $j \neq i$ such that $X_j \leq X_i$ and $Y_j \geq Y_i$. The same is done for the Z-conditional measure where N_z is the number of points dominating (X_i, Y_i), with in addition $|Z_j - Z_i| \leq h_{Z_i}$. The summary Table 1 gives the average values of N and N_z over the $n = 100$ observations.

Case	N	$\hat{\theta}_n(x,y)$	$\hat{\theta}_{n,m}(x,y)$	h	N_z	$\hat{\theta}_n(x,y\vert z)$	$\hat{\theta}_{n,m}(x,y\vert z)$
1	49.5	0.1616	0.2774	0.9384	9.4	0.6990	0.7482
2	49.5	0.8302	0.8356	0.9384	9.7	0.8381	0.8424
3	49.5	0.6209	0.6367	0.9384	9.5	0.8249	0.8462

Table 1: *Average efficiency scores over the 100 observations, for the classroom example. N is the average number of observations dominating (x, y) and N_z the average number of dominating points given $Z = z$. h is the average of the selected local bandwidths (with k-NN=19).*

5.2 Multivariate simulated data sets

In this simulated example, we simulate a multi-input ($p = 2$) and multi-output ($q = 2$) data set. We follow the ideas proposed by Park, Simar and Weiner (2000) and by Simar (2003) to simulate the data set and then we introduce some dependency to an environmental factor Z.

In this set-up, the function describing the efficient frontier is given by:

$$y^{(2)} = 1.0845(x^{(1)})^{0.3}(x^{(2)})^{0.4} - y^{(1)}$$

where $y^{(j)}$, $(x^{(j)})$, denotes the jth component of y, (of x), for $j = 1, 2$. We draw $X_i^{(j)}$ independent uniforms on $(1, 2)$ and $\tilde{Y}_i^{(j)}$ independent uniform on $(0.2, 5)$. Then the generated random rays in the output space are characterized by the slopes $S_i = \tilde{Y}_i^{(2)}/\tilde{Y}_i^{(1)}$. Finally, the generated random points on the frontier are defined by:

$$\begin{aligned} Y_{i,eff}^{(1)} &= \frac{1.0845(X_i^{(1)})^{0.3}(X_i^{(2)})^{0.4}}{S_i + 1} \\ Y_{i,eff}^{(2)} &= 1.0845(X_i^{(1)})^{0.3}(X_i^{(2)})^{0.4} - Y_{i,eff}^{(1)}. \end{aligned}$$

We chose, as above, the efficiencies generated by $\exp(-U_i)$ where U_i are drawn from an exponential with mean $\mu = 1/3$. Finally, in a standard setup (without environmental factors), we define $Y_i = Y_{i,eff} * \exp(-U_i)$.

Now we introduce the dependency on Z in the latter expression as follows: $Z \sim \text{Uniform}(1, 4)$

- Case 1, Z is favorable to output production but differently for $Y^{(1)}$ than for $Y^{(2)}$. We define $V = Z$ and set

$$Y_i^{(1)} = V^2 * Y_{i,eff}^{(1)} * \exp(-U_i)$$
$$Y_i^{(2)} = V^{1/2} * Y_{i,eff}^{(2)} * \exp(-U_i).$$

- Case 2, Z is independent of Y. We define $V = 2.5$, the mean of Z and use the same latter expressions to generate Y.

We computed the FDH, conditional to Z-FDH, the order-m and conditional Z-order-m output efficiency scores of all the units. We have chosen again $m = 25$, for larger values of m, say $m \geq 100$, the results are very similar to the full-frontier (FDH) results. We present the results for a triangle kernel: here again the results are very stable with respect to other choice of the kernel with compact support. Figure 4 and 5 indicate very clearly the differences between the two scenarios even with a small sample size of $n = 100$ (remember that we are in a space of dimension 5). For a larger sample size ($n = 500$) the effect of Z on the efficiency appears still more clearly as in Figure 6. Here also, the difference between the full frontier and the order-25 frontier is more visible.

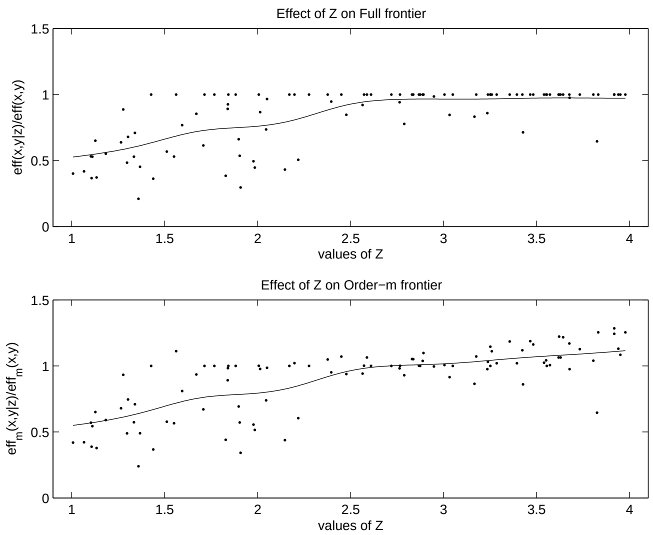

Figure 4: *Multivariate example, case 1, $n = 100$: "positive" effect of Z on production efficiency (output oriented framework). Scatterplot and smoothed regression of $\hat{\lambda}_n(x, y \mid z)/\hat{\lambda}_n(x, y)$ on Z (top) and of $\hat{\lambda}_{m,n}(x, y \mid z)/\hat{\lambda}_{m,n}(x, y)$ on Z, with $m = 25$ (bottom). Here k-NN=17.*

Table 2 presents again a summary of the results, as in the preceding classroom example. We see here, by looking at the average values of N and N_z, that in this 4 (or 5) dimensional

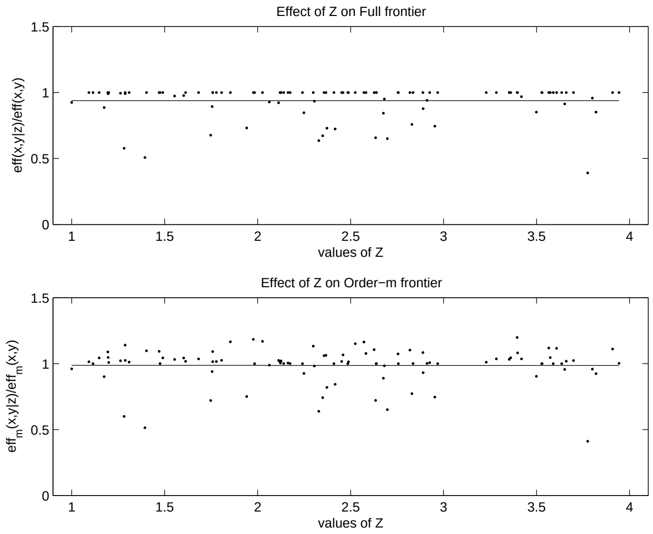

Figure 5: *Multivariate example, case 2, $n = 100$: no effect of Z on production efficiency (output oriented framework). Scatterplot and smoothed regression of $\hat{\lambda}_n(x, y \mid z)/\hat{\lambda}_n(x, y)$ on Z (top) and of $\hat{\lambda}_{m,n}(x, y \mid z)/\hat{\lambda}_{m,n}(x, y)$ on Z, with $m = 25$ (bottom). Here k-NN=19.*

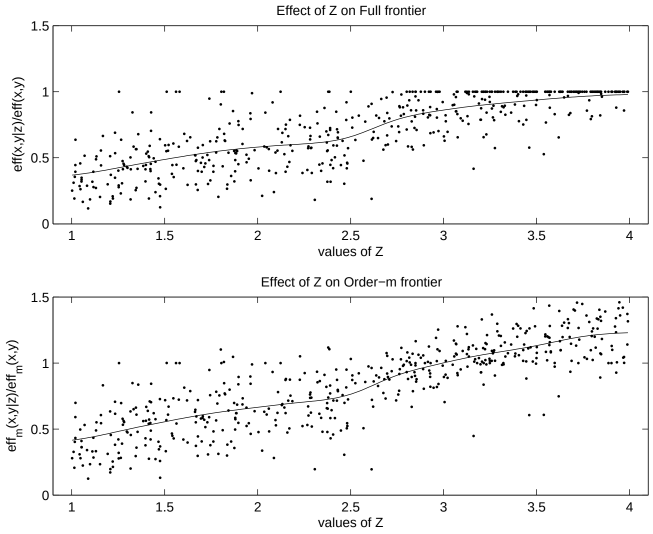

Figure 6: *Multivariate example, case 1, $n = 500$: "positive" effect of Z on production efficiency (output oriented framework). Scatterplot and smoothed regression of $\hat{\lambda}_n(x, y \mid z)/\hat{\lambda}_n(x, y)$ on Z (top) and of $\hat{\lambda}_{m,n}(x, y \mid z)/\hat{\lambda}_{m,n}(x, y)$ on Z, with $m = 25$ (bottom). Here k-NN=40.*

space, it is better to rely on more observations than $n = 100$ to get more sensible results, at least for the full-frontier estimates (this is the curse of dimensionality of the FDH and Z-FDH estimators). However, even when $n = 100$, the Figures 4 and 5 allows to detect the effect of Z on the production process.

Case	N	$\hat{\theta}_n(x, y)$	$\hat{\theta}_{n,m}(x, y)$	h	N_z	$\hat{\theta}_n(x, y\|z)$	$\hat{\theta}_{n,m}(x, y\|z)$
case 1, $n = 100$	4.8	1.5819	1.4788	0.3864	0.7	1.1679	1.1679
case 2, $n = 100$	2.9	1.1988	1.1393	0.3236	0.6	1.0965	1.0965
case 1, $n = 500$	26.9	2.0453	1.7350	0.1467	1.6	1.1760	1.1760

Table 2: *Multivariate example, Average efficiency scores: N is the average number of observations dominating (x, y) and N_z the average number of dominating points given $Z = z$. h is the average of the selected local bandwidths.*

5.3 Mutual funds data

We illustrate our methodology analyzing US Mutual Funds data. We use a cross-section data set, collected by the reputed Morningstar, which consists of the US Mutual Funds universe updated at 05-31-2002. Among this universe we select the Aggressive-Growth (AG) category of Mutual Funds. These are funds that seek rapid growth of capital and that may invest in emerging market growth companies.

From a first data set of 247 observations, we end up, for this illustration, with a sample of 129 mutual funds, after dropping 103 observations for missing values and 15 observations detected by the Simar's (2003) procedure as being outliers.

The selection of variables has been done by taking the same variables chosen in earlier studies (Murthy, Choi and Desai, 1997; Sengupta, 2000) that used a (deterministic) nonparametric approach.

Following these previous studies, we apply an input oriented framework in order to evaluate the performance of mutual funds in terms of their risk (as expressed by standard deviation of return) and transaction costs (including expense ratio, loads and turnover) management. Murthy, Choi and Desai (1997) used as inputs: risk (standard deviation, or volatility of the return), expense ratio (the percentage of fund assets paid for operating expenses, management fees, administrative fees, and all other asset-based costs), loads (percentage for the front-end and back-end sales charges of each fund) and turnover ratio (a measure of the fund's trading activity). The 3 latter inputs are considered as a measure of the transaction costs.

The traditional output in this framework is the total return of funds (the annual return at the 05-31-2002, expressed in percentage terms). Most returns where negative in this period, hence we shift them to get all positive returns by adding 100. This does not change our input oriented analysis. Sengupta (2000) uses market risks of mutual funds (the percentage of fund's movements that can be explained by movements in its benchmark index) as an input in his analysis, underlying that the effect of market risks is conducive for mutual funds performance. In our illustration we use this variable (market risks) as environmental variable, to investigate its effect on our data, *i.e.* if it is detrimental or favorable to the performance of mutual funds in the period under consideration.

In our illustration we decided to eliminate one of the inputs previously considered, the loads, for the following reasons: the curse of dimensionality (6 variables, with only 129 observations); loads in the data set is typically a discrete variable with not many different values (round to the percentage), with a majority of funds having loads equal to zero; and finally, the correlation of this variable with any of the 5 others (X, Y and Z) is smaller than 0.07, which might indicate an orthogonal aspect of the activity. So, we end up with 3 inputs, 1 output, 1 environmental factor and 129 observations.

Table 3 displays some summary statistics of the chosen variables.

Variable	mean	std	min	max	iqr
Y= return	81.8329	9.8416	40.1200	103.7600	13.4825
$X^{(1)}$= volatility	34.9777	8.8845	14.7300	81.0500	9.8875
$X^{(2)}$= turnover	155.1938	99.1631	15.0000	642.0000	129.7500
$X^{(3)}$= exp. ratio	1.6815	1.2859	0.4800	14.7000	0.8400
Z= market risk	0.4721	0.1571	0.0584	1.0000	0.1362

Table 3: *Summary statistics for the $n = 129$ Aggressive-Growth US Mutual Funds. Average, standard deviation, minimum, maximum and interquartile range.*

Table 4 presents the results coming from our Matlab code. In order to save place, we present only 15 funds chosen at random from the full table (presented at length in the Appendix). We have chosen a triangle kernel for the smoothing and the likelihood cross-validation procedure provided $k = 21$ as the optimal choice for the k-NN method. We select again the value of $m = 25$, although we did the computations for several values of m. If $10 \leq m \leq 50$ we obtain very similar results and when m is larger than, say, 100 we obtain very similar results as for the full-frontier efficiency scores.

Looking at the last row of Table 4, we see that the global effect of the market risk factor Z on the full efficiency measures is an increase from 0.6083 to 0.8825. For the order-m frontier we have a similar mean effect going from 0.8149 to 0.9215. The effect is more important for the full FDH frontier, as expected, since these measures are more sensitive to extreme points.

Fund	N	$\hat{\theta}_n(x,y)$	$\hat{\theta}_{n,m}(x,y)$	h	N_z	$\hat{\theta}_n(x,y\|z)$	$\hat{\theta}_{n,m}(x,y\|z)$
1	20	0.4200	0.6923	0.2589	4	0.8056	0.9355
2	7	0.4956	0.6113	0.1449	1	0.8722	0.9997
3	0	1.0000	1.0005	0.2208	0	1.0000	1.0000
6	3	0.5138	0.8777	0.0524	0	1.0000	1.0001
15	9	0.4396	0.6925	0.0226	0	1.0000	1.0013
33	0	1.0000	1.0576	0.0238	0	1.0000	1.0000
36	7	0.4860	0.7144	0.0155	1	0.8855	0.8961
37	15	0.4910	0.6179	0.0149	4	0.7457	0.7578
40	3	0.8323	0.8449	0.0193	0	1.0000	1.0000
74	8	0.4831	0.8772	0.1762	2	0.4831	0.5835
111	1	0.9182	0.9300	0.0188	0	1.0000	1.0000
112	5	0.7976	0.8571	0.0220	1	0.9587	0.9593
124	19	0.4790	0.7182	0.1710	5	0.8487	0.8998
127	47	0.3098	0.5472	0.0160	6	0.5707	0.6008
129	5	0.4453	0.7846	0.0296	2	0.9062	0.9312
mean	9.2	0.6083	0.8149	0.0823	1.8	0.8825	0.9215

Table 4: *Results from 15 selected funds from the Aggressive-Growth US Mutual Funds. N is the number of observations dominating (x, y) and N_z the number of dominating points given $Z = z$. h is the selected local bandwidth (k-NN=21). Last row is the average over all the 129 observations*

We propose here some descriptive comments on the figures of Table 4: a few funds have a huge increase of their efficiencies when Z-conditional measures are considered (funds like #1,#2,#3,…, even some like fund #6 becomes efficient). Some other funds have a very poor performance, even if we take the environmental factor into account: these are funds like #37, #74, #127, …. In practical applications, these funds should deserve more attention.

To have a global idea of the effect of the risk factor Z on mutual funds performance, we regress nonparametrically the ratios between the conditional efficiency measures and the unconditional efficiency measures on Z: we obtain the picture displayed in Figure 7.

Looking at this picture, we can see a global positive effect of the market risk factor Z on the performance of mutual funds. When looking at the full conditional efficiency measures (top panel of Figure 7), this effect seems to be more important when $Z \geq 0.5$. Note that for low values of Z, the regression line in this case is attracted to low values of isolated points on the left of the picture. This global effect is confirmed on the bottom panel of the same picture, where the effect seems to start around $Z = 0.2$. These pictures confirm that with our data market risk acts as a “ substitutive input” in the mutual funds management process.

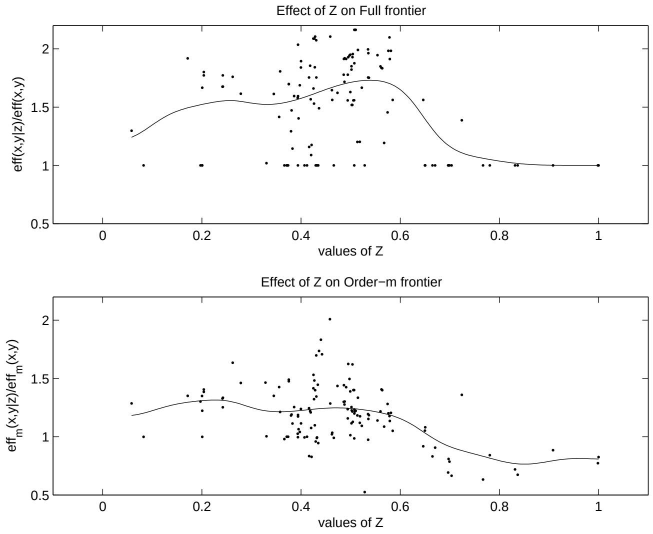

Figure 7: Aggressive-Growth US Mutual Funds. Scatterplot and smoothed regression of the ratios $\hat{\theta}_n(x, y \mid z)/\hat{\theta}_n(x, y)$ on Z (top) and of $\hat{\theta}_{m,n}(x, y \mid z)/\hat{\theta}_{m,n}(x, y)$ on Z, with $m = 25$ (bottom). Here k-NN=21.

6 Conclusion

In this paper, developing ideas proposed by Cazals, Florens and Simar (2002), we provide a full probabilistic formulation of a nonparametric frontier model and of a nonparametric frontier model of order-m. This formulation allows the introduction in both models (full frontier and order-m frontier) of environmental factors which may influence the production process but that are neither inputs nor outputs under the control of the producer.

The presentation allows general multi-input/ multi-output situations and provides a practical way for evaluating the nonparametric estimators. A data-driven procedure for choosing the bandwidth, based on a k-nearest neighbor method is suggested. Furthermore, we propose a useful graphical tool for highlighting the eventual influence of Z on the production process. Our method will tell us if the environmental factor is conducive or detrimental to the production activity.

The approach is illustrated by some simulated data set and with a real data set on US mutual funds, where the risk of the market shows a positive influence on the performance (management process) of mutual funds.

Some interesting theoretical issues are still open in this framework. For instance, what are the statistical properties of the conditional full frontier efficiency estimator? Or, how could we select optimal bandwidth in the estimation procedure? We propose a very simple sensible technique, based on likelihood cross-validation for the density of Z, but other criterion could be investigated.

Appendix

Full Table of Results for the Mutual Funds Example

Units	N	$\hat{\theta}_n(x,y)$	$\hat{\theta}_{n,m}(x,y)$	h	N_z	$\hat{\theta}_n(x,y\vert z)$	$\hat{\theta}_{n,m}(x,y\vert z)$
1	20	0.4200	0.6923	0.2589	4	0.8056	0.9355
2	7	0.4956	0.6113	0.1449	1	0.8722	0.9997
3	0	1.0000	1.0005	0.2208	0	1.0000	1.0000
4	4	0.6138	0.6314	0.2214	4	0.6138	0.8526
5	3	0.6138	0.6266	0.2256	3	0.6138	0.8162
6	3	0.5138	0.8777	0.0524	0	1.0000	1.0001
7	4	0.5407	0.8224	0.0604	0	1.0000	1.0010
8	6	0.5446	0.7112	0.0631	0	1.0000	1.0003
9	6	0.5456	0.7145	0.0651	0	1.0000	1.0001
10	5	0.5147	0.6686	0.0146	0	1.0000	1.0000
11	4	0.5129	0.7198	0.0145	0	1.0000	1.0000
12	7	0.5185	0.6170	0.0141	1	1.0000	1.0000
13	6	0.5185	0.6155	0.0157	0	1.0000	1.0000
14	7	0.4750	0.7152	0.0162	0	1.0000	1.0007
15	9	0.4396	0.6925	0.0226	0	1.0000	1.0013
16	17	0.4560	0.6231	0.2166	6	0.8080	0.8755
17	2	0.6000	0.8299	0.2207	0	1.0000	1.0148
18	6	0.5552	0.7223	0.2166	1	1.0000	1.0004
19	18	0.5095	0.6737	0.0173	3	0.7938	0.8330
20	34	0.3497	0.5334	0.0513	10	0.6318	0.6468
21	5	0.5278	0.8996	0.0322	1	1.0000	1.0023
22	29	0.3735	0.6700	0.0296	4	0.6304	0.6976
23	2	0.6333	1.0038	0.0300	0	1.0000	1.0296
24	6	0.7931	0.8088	0.0838	2	0.8080	0.8118
25	13	0.4317	0.6622	0.0157	4	0.8102	0.8167
26	3	0.5093	0.8681	0.0393	0	1.0000	1.0002
27	7	0.5107	0.7876	0.0399	2	0.8952	0.9332
28	5	0.4648	0.8013	0.0385	1	0.9275	0.9572
29	0	1.0000	1.0427	0.0171	0	1.0000	1.0003
30	0	1.0000	1.0101	0.0199	0	1.0000	1.0000
31	0	1.0000	1.0056	0.0205	0	1.0000	1.0000
32	2	0.6537	0.7565	0.0178	1	1.0000	1.0000
33	0	1.0000	1.0576	0.0238	0	1.0000	1.0000
34	4	0.4500	0.8576	0.0224	0	1.0000	1.0085
35	6	0.4855	0.7255	0.0200	1	0.9667	0.9682
36	7	0.4860	0.7144	0.0155	1	0.8855	0.8961
37	15	0.4910	0.6179	0.0149	4	0.7457	0.7578
38	15	0.4904	0.6218	0.0145	4	0.7447	0.7566
39	8	0.4896	0.7161	0.0256	1	0.8406	0.9143
40	3	0.8323	0.8449	0.0193	0	1.0000	1.0000
41	3	0.6418	0.7141	0.0148	0	1.0000	1.0000
42	4	0.6416	0.7139	0.0155	1	1.0000	1.0000
43	1	1.0000	1.0140	0.0155	0	1.0000	1.0000
44	6	0.4790	0.6746	0.0169	0	1.0000	1.0002
45	13	0.4825	0.5918	0.0167	1	0.8010	0.9064
46	13	0.4823	0.5894	0.0184	0	1.0000	1.0007
47	3	0.5700	0.7466	0.0189	0	1.0000	1.0042
48	5	0.5429	0.8070	0.0318	1	0.9987	0.9998
49	19	0.4042	0.5872	0.0298	3	0.6444	0.6890
50	20	0.4041	0.5907	0.0365	4	0.6444	0.7412
51	2	0.7125	0.9431	0.0302	0	1.0000	1.0041
52	5	0.6196	0.9513	0.0856	1	0.9674	0.9999
53	0	1.0000	1.1030	0.1429	0	1.0000	1.0000
54	0	1.0000	1.0095	0.0429	0	1.0000	1.0000
55	0	1.0000	1.5839	0.2497	0	1.0000	1.0026
56	0	1.0000	1.3897	0.3282	0	1.0000	1.0000
57	0	1.0000	1.4835	0.3352	0	1.0000	1.0000
58	0	1.0000	1.0000	0.0359	0	1.0000	1.0000
59	0	1.0000	1.0000	0.0375	0	1.0000	1.0000
60	1	1.0000	1.0000	0.0376	1	1.0000	1.0000
61	5	0.5212	0.7692	0.0250	0	1.0000	1.0022
62	10	0.5226	0.6932	0.0266	1	1.0000	1.0000
63	7	0.5226	0.7019	0.0213	0	1.0000	1.0003
64	1	0.5700	1.0591	0.0387	0	1.0000	1.0322
65	21	0.3913	0.6888	0.0855	2	0.9710	1.0083

Units	N	$\hat{\theta}_n(x,y)$	$\hat{\theta}_{n,m}(x,y)$	h	N_z	$\hat{\theta}_n(x,y\|z)$	$\hat{\theta}_{n,m}(x,y\|z)$
66	1	0.8507	1.2603	0.0186	0	1.0000	1.0426
67	1	0.8636	1.2598	0.0216	0	1.0000	1.0513
68	1	0.6404	0.9775	0.0415	0	1.0000	1.0116
69	5	0.5098	0.7110	0.0678	1	0.8227	0.9608
70	5	0.6000	0.9246	0.0265	0	1.0000	1.0114
71	1	0.8385	0.9198	0.0701	0	1.0000	1.0000
72	2	0.6872	0.7803	0.0790	0	1.0000	1.0000
73	2	0.8327	0.9902	0.2674	2	0.8327	0.8336
74	8	0.4831	0.8772	0.1762	2	0.4831	0.5835
75	23	0.4458	0.6961	0.1735	8	0.4458	0.5637
76	22	0.4458	0.6948	0.1758	7	0.4458	0.5462
77	2	0.6129	0.9795	0.1743	1	0.6129	0.6787
78	3	0.6163	0.6963	0.0390	0	1.0000	1.0000
79	6	0.6193	0.6857	0.1254	0	1.0000	1.0020
80	0	1.0000	1.0009	0.3742	0	1.0000	1.0000
81	24	0.4191	0.6900	0.0413	7	0.6896	0.7046
82	54	0.3301	0.5468	0.0393	14	0.6948	0.7027
83	12	0.6793	0.7154	0.0347	4	0.7769	0.7967
84	6	0.6793	0.7502	0.0354	1	0.8783	0.8869
85	19	0.3335	0.4509	0.0390	3	0.5659	0.6721
86	19	0.3333	0.4484	0.0390	3	0.5659	0.6619
87	1	0.6793	0.8402	0.0337	0	1.0000	1.0005
88	0	1.0000	1.0061	0.0268	0	1.0000	1.0000
89	0	1.0000	1.0001	0.0237	0	1.0000	1.0000
90	22	0.4711	0.6990	0.0211	8	0.8264	0.8704
91	61	0.3777	0.5585	0.0198	10	0.5918	0.6749
92	61	0.3777	0.5585	0.0164	8	0.7885	0.7905
93	6	0.5279	0.8069	0.0213	1	0.9792	0.9845
94	1	0.6404	1.0990	0.1199	0	1.0000	1.0091
95	8	0.4172	0.8174	0.1249	2	0.4172	0.8835
96	6	0.4166	0.8385	0.1240	2	0.4166	0.8828
97	4	0.5182	0.8847	0.0848	1	0.9909	1.0042
98	15	0.3884	0.6541	0.0863	5	0.7701	0.7883
99	14	0.3885	0.6526	0.0810	4	0.7703	0.7842
100	12	0.3873	0.6994	0.0840	4	0.8125	0.8235
101	6	0.6991	0.7352	0.4055	2	0.9075	0.9456
102	2	0.7207	0.7356	0.2030	0	1.0000	1.0000
103	5	0.6706	0.8301	0.5469	3	0.6706	0.6853
104	3	0.6674	0.9093	0.5449	2	0.6674	0.7029
105	2	0.6136	1.0064	0.0150	0	1.0000	1.0196
106	2	0.5403	0.8998	0.0154	0	1.0000	1.0040
107	2	0.5115	0.8889	0.0133	0	1.0000	1.0028
108	1	0.5625	0.8641	0.0170	0	1.0000	1.0000
109	2	0.5625	0.7694	0.0272	0	1.0000	1.0000
110	5	0.5649	0.6556	0.0539	2	0.8000	0.9354
111	1	0.9182	0.9300	0.0188	0	1.0000	1.0000
112	5	0.7976	0.8571	0.0220	1	0.9587	0.9593
113	14	0.4464	0.4979	0.0386	0	1.0000	1.0001
114	1	0.8189	0.8228	0.0295	1	0.8189	0.8197
115	10	0.4457	0.5857	0.0271	0	1.0000	1.0003
116	16	0.4486	0.5459	0.0266	0	1.0000	1.0003
117	16	0.4482	0.5421	0.0259	1	0.6680	0.9412
118	0	1.0000	2.3707	0.0332	0	1.0000	1.2449
119	8	0.4128	0.7950	0.0176	2	0.8929	0.9684
120	6	0.4125	0.8183	0.0158	1	0.8922	0.9812
121	10	0.4142	0.7648	0.0163	3	0.8959	0.9350
122	0	1.0000	1.1304	0.4276	0	1.0000	1.0000
123	0	1.0000	1.0217	0.0437	0	1.0000	1.0010
124	19	0.4790	0.7182	0.1710	5	0.8487	0.8998
125	65	0.3461	0.5564	0.1705	15	0.5800	0.7427
126	64	0.3463	0.5557	0.1716	14	0.5800	0.7378
127	47	0.3098	0.5472	0.0160	6	0.5707	0.6008
128	7	0.5848	0.8348	0.1361	2	0.5848	0.6938
129	5	0.4453	0.7846	0.0296	2	0.9062	0.9312

References

[1] Banker, R.D. and R.C. Morey (1986), Efficiency analysis for exogeneously fixed inputs and outputs, *Operations Research*, 34(4), 513–521.

[2] Cazals, C., J.P. Florens and L. Simar (2002), "Nonparametric frontier estimation: a robust approach", *Journal of Econometrics*, 106, 1-25.

[3] Charnes, A., Cooper, W.W. and E. Rhodes (1978), Measuring the inefficiency of decision making units. *European Journal of Operational Research*, 2, 429–444.

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

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

[6] Fare, R., S. Grosskopf, C.A. K. Lovell and C. Pasurka (1989), "Multilateral Productivity Comparisons when some Outputs are Undesirable: a Nonparametric Approach", *Review of Economics and Statistics* 71:1 (February), 90-98.

[7] Fare, R., S. Grosskopf and C.A. K. Lovell (1994), *Production Frontiers*, Cambridge University Press.

[8] Farrell, M.J. (1957), The measurement of productive efficiency. *Journal of the Royal Statistical Society, Series A*, 120, 253–281.

[9] Fried, H.O., S.S. Schmidt and S. Yaisawarng (1999), Incorporating the operating environment into a nonparametric measure of technical efficiency, *Journal of Productivity Analysis*, 12, 249–267.

[10] Fried, H.O., C.A.K. Lovell, S.S. Schmidt and S. Yaisawarng (2002), Accounting for environmental effects and statistical noise in Data Envelopment Analysis, *Journal of Productivity Analysis*, 17, 157–174.

[11] Kneip, A., B.U. Park, and L. Simar (1998), A note on the convergence of nonparametric DEA estimators for production efficiency scores, *Econometric Theory*, 14, 783–793.

[12] Koopmans, T.C. (1951), An Analysis of Production as an Efficient Combination of Activities, in *Activity Analysis of Production and Allocation*, ed. by T.C. Koopmans, Cowles Commission for Research in Economics, Monograph 13. New York: John-Wiley and Sons, Inc.

[13] Murthi, B., Choi, Y. and Desai, P. (1997), "Efficiency of Mutual Funds and Portfolio Performance Measurement: a Nonparametric Measurement", *European Journal of Operational Research*, 98, 408-418.

[14] Park, B. Simar, L. and Ch. Weiner (2000), The FDH Estimator for Productivity Efficiency Scores : Asymptotic Properties, *Econometric Theory*, Vol 16, 855-877.

[15] Sengupta, J. K. (2000), *Dynamic and Stochastic Efficiency Analysis, Economics of Data Envelopment Analysis*, World Scientific, Singapore.

[16] Shephard, R.W. (1970), *Theory of Cost and Production Function.* Princeton: Princeton University Press.

[17] Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[18] Simar, L. (2003), Detecting Outliers in Frontiers Models: a Simple Approach , Discussion paper #0146, Institut de Statistique, UCL, Louvain-la-Neuve, Belgium. Forthcoming in *Journal of Productivity Analysis.*

[19] Simar, L., and P.W. Wilson (2000), Statistical inference in nonparametric frontier models: The state of the art, *Journal of Productivity Analysis* 13, 49–78.

[20] Simar L. and P. Wilson (2002), Nonparametric Test of Return to Scale, *European Journal of Operational Research*, 139, 115–132.

[21] Simar L. and P. Wilson (2003), Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes, Discussion paper #0307, Institut de Statistique, UCL, Louvain-la-Neuve, Belgium.