INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

1050

HOW TO MEASURE THE IMPACT OF ENVIRONMENTAL FACTORS IN A NONPARAMETRIC PRODUCTION MODEL?

BADIN, L., DARAIO, C. and L. SIMAR

How to Measure the Impact of Environmental Factors in a Nonparametric Production Model?

Luiza Bădin Cinzia Daraio Léopold Simar*

December 17, 2010

Abstract:

The measurement of technical efficiency of decision making units is useful for making comparisons and informing managers and policy makers on existing differentials and potential improvements across a sample of analyzed units. The step further is to relate the obtained efficiency estimates to some external or environmental variables which may influence the production process and hence, affect the performance evaluation and explain the efficiency differentials. Conditional efficiency measures (Daraio and Simar, 2005; 2007a), including conditional FDH, conditional DEA, conditional order$-m$ and conditional order$-\alpha$, have been recently introduced and became rapidly a useful tool to investigate the impact of external-environmental factors on the performance of Decision Making Units in a nonparametric framework. In this paper, we clarify what can be learned by analyzing these conditional efficiency scores, showing that the impact of these factors on the production process can have different facets: impact on the attainable set in the input $\times$ output space, and/or impact on the distribution of the inefficiency scores. The approach proposes statistical inference on the level of the impact, using up-to-dated bootstrap algorithms for which we prove the consistency. The procedure is illustrated through simulated samples and with a real data set in the Banking industry.

Keywords: conditional efficiency measures, robust frontiers, nonparametric frontiers, bootstrap, subsampling, Banking industry

JEL Classification: C14, C40, C60, D20

*__Bădin__: Department of Mathematics, Bucharest Academy of Economic Studies and *Gh. Mihoc-C. Iacob* Institute of Mathematical Statistics and Applied Mathematics, Bucharest, Romania; email luiza.badin@csie.ase.ro. **Daraio**: Dipartimento di Scienze Aziendali, Università di Bologna, Bologna, Italy; email cinzia.daraio@unibo.it. **Simar**: Institut de Statistique, Université Catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium; email leopold.simar@uclouvain.be. Financial support from the "Inter-university Attraction Pole", Phase VI (No. P6/03) of the Belgian Government (Belgian Science Policy) and from the INRA-GREMAQ, Toulouse, France are gratefully acknowledged.

1 Introduction and Basic Notations

In productivity analysis, one is interested in the evaluation of the performances of firms to identify inefficient units where improvements could help to increase their profitability or to reduce their costs. Most of the efficiency analysis literature focused on the estimation of the production frontier, which provides the benchmark against which the economic producers are evaluated. Nevertheless, a very important component, that recent studies are more concerned with, is the explanation of efficiency differentials by including in the analysis exogenous variables or environmental factors, that cannot be controlled by the producer, but may influence the production process. From a managerial point of view, it is important to identify the "particularities" of the production process or the economic conditions that might be responsible for inefficiency as well as to detect and analyze possible influential factors that can determine changes in productivity patterns. The meaning and the economic role played by external-environmental variables is strictly linked to the economic field firms are operating in. The choice of the environmental variables has to be done on a case-by-case basis, having a good knowledge of the production process characteristics and by taking into account the economic field of application.

In this paper, we will formalize a nonparametric production model where the role of these environmental factors is explicitly introduced in a non-restrictive way. Then we will explain how in these models, we can measure and infer about the impact of these factors on the production process. By doing so, we will clarify the usefulness and limitations of some previous tools developed in the literature and suggest practical algorithms to implement them.

We will first introduce the notations and the basic assumptions on the Data Generating Process (DGP) characterizing the production process in the presence of environmental factors. Let $X \in \mathbb{R}_+^p$ denote the vector of inputs and let $Y \in \mathbb{R}_+^q$ denote the vector of outputs. We consider a vector of environmental factors $Z \in \mathcal{Z} \subset \mathbb{R}^r$ that may influence the process and the productivity patterns. Firms transform quantities of inputs into outputs, but the environmental variables may affect this process. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be the probability space on which the random variables are defined, we denote by $\mathcal{P}$ the support of the joint distribution of (X, Y, Z) and we denote a particular DGP by $P \in \mathbb{P}$.

A large part of the literature on this topic has been focused on so-called 2-stage analysis, where typically, some first stage estimates of the efficiency of the firms are regressed in a second stage on these additional factors to investigate their effect on efficiency. Simar and Wilson (2007) clarified that these two stages approaches are restricted to models where these factors do not influence the shape of the production set (this is the "separability" condition detailed in the following). Banker and Natarajan (2008) suggest another model where a two-stage approach is valid but the model heavily depends on quite restrictive and unrealistic

assumptions on the production process, as described and commented in details in Simar and Wilson (2010b). If the 2-stage approach is validated (by the appropriate test, see Daraio et al. 2010), one can indeed in a first stage estimate the efficiency scores of the units relative to the boundary of the unconditional attainable set in the inputs $\times$ outputs space and then regress, in a second stage, the obtained efficiencies on the environmental factors. We know that even if an appropriate model is used (Logit, Truncated Normal, Nonparametric truncated regression,...), the inference on the impact of Z on the efficiency measures has to be carefully conducted, using adapted bootstrap techniques (see Simar and Wilson, 2007 and 2010b for details).

The impact and influence of Z on the production process may be multiple and can be quite different from one application to the other. The effect of Z on the production may either affect the range of achievable values for the couples (X, Y), including the shape of the boundaries of the attainable set, or it may only affect the distribution of the inefficiencies inside a set with boundaries not depending on Z (only the probability of being more or less far from the efficient frontier may depend on Z) or it can affect both. Finally, the environmental factors Z may also be completely independent of (X, Y).

Cazals et al. (2002) and Daraio and Simar (2005) provide a quite general and unrestricted framework to investigate the joint behavior of (X, Y, Z) from a productivity point of view. They consider a probability model that generates the variables (X, Y, Z) where the conditional distribution of (X, Y) given a particular value of Z will be of particular interest. This conditional process can be described by

$$H(x, y|z) = \text{Prob}(X \leq x, Y \geq y|Z = z), \tag{1.1}$$

or any equivalent variation of it (the joint conditional density function or the joint conditional cumulative distribution function, ...). The function $H(x, y|z)$ is simply the probability for a unit operating at level (x, y) to be dominated by firms facing the same environmental conditions z. Given that $Z = z$, the range of possible combinations of inputs $\times$ outputs, Ψ^z, is the support of $H(x, y|z)$:

$$\Psi^z = \{(x, y)|\, Z = z,\ x \text{ can produce } y\}, \tag{1.2}$$

If $H(x, y)$ denotes the unconditional probability of being dominated, we have

$$H(x, y) = \int_{\mathcal{Z}} H(x, y|z)\, f_Z(z)\, dz, \tag{1.3}$$

having support Ψ, the marginal (unconditional) attainable set defined as

$$\Psi = \{(x, y)|\, x \text{ can produce } y\} = \bigcup_{z \in \mathcal{Z}} \Psi^z. \tag{1.4}$$

Remember that the joint support of the variables (X, Y, Z) is denoted by $\mathcal{P}$. It is clear that, by construction, for all $z \in \mathcal{Z}$, $\Psi^z \subseteq \Psi$.

The "separability" condition, described in Simar and Wilson (2007) states that the support of (X, Y) is not dependent of Z, equivalently

$$\text{"Separability" condition:} \quad \Psi^z = \Psi, \quad \text{for all } z \in \mathcal{Z}. \tag{1.5}$$

In this latter case, the support of (X, Y, Z) can be written as $\mathcal{P} = \Psi \times \mathcal{Z}$, where $\times$ represents the cartesian product. As clearly illustrated by Figures 1 and 2 in Simar and Wilson (2010), it is important to understand the implications of condition (1.5). If the condition is verified, the only potential remaining impact of the environmental factors on the production process may be on the distribution of the efficiencies. This justifies the use of 2-stage approaches as illustrated in Simar and Wilson (2007). If the condition (1.5) is not verified, the measure of the distance of a unit (x, y) to the boundary of Ψ, even if it can be well defined and estimated (see details below), has little economic interest, because it ignores the heterogeneity introduced by Z on the attainable sets of values for (X, Y).

Whether or not Ψ^z is independent of z is an empirical issue and Daraio et al. (2010) provide a statistical procedure to test this hypothesis. The test is a "global" test of separability since it tests the null hypothesis $\Psi^z = \Psi$, $\forall z \in \mathcal{Z}$ against its complement: $\exists z \in \mathcal{Z}$ such that $\Psi^z \neq \Psi$.

As described e.g. in Daraio and Simar (2007a), the two measures $H(x, y|z)$ and $H(x, y)$ allow to define conditional and marginal efficiency scores that can be estimated by nonparametric methods. The comparison of the conditional and marginal efficiency scores can be used to investigate the impact of Z on the production process. One of the objectives of this paper is to clarify what can be learned from the analysis of these conditional efficiency scores and focusing on the particular role of efficiency scores relative to partial order frontiers (order-m frontiers from Cazals et al., 2002 and order-α quantile type frontiers from Daouia and Simar, 2006). In this paper, we suggest also a procedure allowing to make local inference on the impact of Z on the process (as opposed to the global test of separability developed in Daraio et al. 2010). Confidence intervals for the local impact of Z will be obtained by adapting the subsampling ideas from Simar and Wilson (2010a)

The paper is organized as follows. Section 2 revisits the concept of conditional efficiency scores and explains what can effectively be learned by comparing conditional and unconditional efficiencies. Section 3, provides nonparametric estimates for the local effect of Z on the production including a consistent bootstrap algorithm to produce confidence intervals for the measures of the impact. We illustrate the procedure with a real data set in the banking sector in Section 4.2. Section 5 summarizes the main findings and concludes the paper.

2 Effect of Z on Efficiency Measures

2.1 Farrell Efficiency scores

The literature on efficiency analysis propose several ways for measuring the distance of a firm operating at the level (x_0, y_0) to the efficient boundary of the attainable set. In the lines of the pioneering work of Debreu (1950), Farrell (1957) and Shephard (1970), radial distances became very popular in the efficiency literature. They can be input or output oriented (maximal radial contraction of the inputs or maximal radial expansion of the outputs to reach the efficient boundary). Recently, Färe et al. (1985) introduced hyperbolic radial distances that avoid some of the ambiguity in choosing output or input orientation. In this case, input and output levels are adjusted simultaneously. These radial measures can be defined as follows:

$$\begin{aligned}\theta(x_0, y_0) &= \inf\{\theta > 0 | (\theta x_0, y_0) \in \Psi\} \\ \lambda(x_0, y_0) &= \sup\{\lambda > 0 | (x_0, \lambda y_0) \in \Psi\} \\ \gamma(x_0, y_0) &= \sup\{\gamma > 0 | (\gamma^{-1} x_0, \gamma y_0) \in \Psi\}.\end{aligned}$$

In what follows, we will focus the presentation for the output orientation and it is easy to adapt the presentation for the input oriented and for the hyperbolic cases. From Cazals et al. (2002) and Daraio and Simar (2005), we know that under the assumption of free disposability of the inputs and of the outputs, these measures can be characterized by some appropriate probability function determined by $H(x, y)$. We have, for the marginal Farrell output measure of efficiency,

$$\lambda(x_0, y_0) = \sup\{\lambda > 0 | S_{Y|X}(\lambda y_0 | X \leq x_0) > 0\}, \tag{2.1}$$

where $S_{Y|X}(y_0|X \leq x_0) = \text{Prob}(Y \geq y_0 | X \leq x_0) = \dfrac{H(x_0, y_0)}{H(x_0, 0)}$ is the (nonstandard) conditional survival function of Y, nonstandard because the condition is $X \leq x_0$ and not $X = x_0$.

If the firm is facing environmental factors $Z = z_0$, then Daraio and Simar (2005) define the conditional Farrell output measure of efficiency as

$$\begin{aligned}\lambda(x_0, y_0 | z_0) &= \sup\{\lambda > 0 | (x_0, \lambda y_0) \in \Psi^{z_0}\} \\ &= \sup\{\lambda > 0 | S_{Y|X,Z}(\lambda y_0 | X \leq x_0, Z = z_0) > 0\},\end{aligned} \tag{2.2}$$

where $S_{Y|X,Z}(y_0|X \leq x_0, Z = z_0) = \text{Prob}(Y \geq y_0 | X \leq x_0, Z = z_0) = \dfrac{H(x_0, y_0|z_0)}{H(x_0, 0|z_0)}$ is the conditional survival function of Y, here we condition on $X \leq x_0$ and $Z = z_0$. Since for all $z_0 \in \mathcal{Z}$, $\Psi^{z_0} \subseteq \Psi$, we have for all $(x_0, y_0, z_0) \in \mathcal{P}$ the relations $1 \leq \lambda(x_0, y_0|z_0) \leq \lambda(x_0, y_0)$.

Daraio et al. (2010) uses these two measures to conduct a global test of separability. In their approach, using unconditional and conditional efficiency measures, they propose to

estimate (by using FDH or DEA techniques) a kind of mean integrated square difference between $\mathcal{P}$ and $\Psi \times \mathcal{Z}$. This provide a test statistic whose sampling distribution is approximated by the bootstrap. We propose below, as a complementary analysis, to investigate the local impact of Z on the process.

It has been shown in details in Daraio and Simar (2005, 2007a) that the ratios of conditional to unconditional measures may be informative to investigate the impact of Z on the production process. The ratios are defined as follows, for all $(x, y, z) \in \mathcal{P}$,

$$R(x, y|z) = \frac{\lambda(x, y|z)}{\lambda(x, y)}. \tag{2.3}$$

If we consider a generic random observation $(X, Y) \in \Psi^z$ of a firm facing environmental factors $Z = z$, we can define the random variable $R(X, Y|Z = z)$ having the following properties: for all $z \in \mathcal{Z}$, $R(X, Y|Z = z) \overset{a.s.}{\leq} 1$, but if the separability condition (1.5) holds then for all z, $R(X, Y|Z = z) \overset{a.s.}{=} 1$. A population parameter of particular interest will be the conditional average of these ratio. For any DGP $P \in \mathbb{P}$, we define the mean and variance of $R(X, Y|Z = z)$:

$$\begin{aligned} \tau^z(P) &= \mathbb{E}(R(X, Y|Z = z)) \\ \sigma^{2,z}(P) &= \mathbb{V}(R(X, Y|Z = z)). \end{aligned} \tag{2.4}$$

Clearly, for all $P \in \mathbb{P}$, $\tau^z(P) \leq 1$ but if $\Psi^z = \Psi$, $\tau^z(P) = 1$ and if $\Psi^z \neq \Psi$, then $\tau^z(P) < 1$. So, $\tau^z(P)$ will be our basic quantity of interest that allows to make a local analysis on the impact of Z on the production set when $Z = z$. We will provide nonparametric estimate of $\tau^z(P)$ and their analysis as a function of z will help to understand how the impact of Z on the attainable set may vary with z. We will also provide bootstrap confidence intervals for $\tau^z(P)$, for all $z \in \mathcal{Z}$. By looking to the confidence interval, we will be able to check if locally, Z has a significant effect on the boundary of the attainable set.

2.2 Partial order Frontiers

Partial frontiers, and the resulting partial efficiency scores, have been proposed to provide robust measures of efficiencies, robust to extreme data points or outliers (a survey and a detailed analysis of these approaches can be found in Daraio and Simar, 2007a). In our setup here, this remains true when we will use partial frontiers of extreme orders, as explained below. However, when using partial frontiers of lower order, we will see that we obtain useful complementary information on the impact of Z on the distribution of the inefficiencies inside the attainable set. To save space, we limit the presentation to the output oriented case and to the order-α quantile frontiers. The extension to other orientations (input and hyperbolic) is immediate. The case of the partial output order-m frontier is summarized in Appendix A.2.

Order-α quantile frontiers

Extending previous work of Aragon et al. (2000) for the univariate case, Daouia and Simar (2006) define for any $\alpha \in (0, 1]$ the order-α output efficiency score as

$$\lambda_\alpha(x_0, y_0) = \sup\{\lambda > 0 | S_{Y|X}(\lambda y_0 | X \leq x_0) > 1 - \alpha\}. \tag{2.5}$$

We see that if $\alpha \to 1$, $\lambda_\alpha(x_0, y_0) \to \lambda(x_0, y_0)$. If $\lambda_\alpha(x_0, y_0) = 1$, the point (x_0, y_0) belongs to the order-α quantile frontier, meaning that only $(1 - \alpha) \times 100\%$ of the firms using less resources than x_0, dominate the unit (x_0, y_0). A value $\lambda_\alpha(x_0, y_0) < 1$ indicates a firm producing more than the level determined by the order-α frontier at x_0.

By conditioning on $Z = z_0$, Daouia and Simar (2006) define similarly the conditional order-α output efficiency score of (x_0, y_0) as

$$\lambda_\alpha(x_0, y_0 | z_0) = \sup\{\lambda > 0 | S_{Y|X,Z}(\lambda y_0 | X \leq x_0, Z = z_0) > 1 - \alpha\}. \tag{2.6}$$

Again the ratios of conditional to unconditional scores will be of interest. We define

$$R_\alpha(x, y|z) = \frac{\lambda_\alpha(x, y|z)}{\lambda_\alpha(x, y)}, \tag{2.7}$$

and when considering a generic observation $(X, Y) \in \Psi^z$ of a firm facing environmental factors $Z = z$, we obtain the random variable $R_\alpha(X, Y|Z = z)$. For any DGP $P \in \mathbb{P}$, we can thus define the conditional average of this ratio:

$$\tau^z_\alpha(P) = \mathbb{E}(R_\alpha(X, Y|Z = z)), \tag{2.8}$$

where again, if $\alpha \to 1$, $\tau^z_\alpha(P) \to \tau^z(P)$.

In the Appendix A.2, we define the order-m efficiency scores, the ratios $R_m(x, y|z) = \lambda_m(x, y|z)/\lambda_m(x, y)$ and their expectation $\tau^z_m(P) = \mathbb{E}(R_m(X, Y|Z = z))$. In this case, when $m \to \infty$, $\tau^z_m(P) \to \tau^z(P)$.

2.3 What do we learn by the analysis of $\tau^z(P)$, $\tau^z_\alpha(P)$ and $\tau^z_m(P)$?

It has been described in details in Daraio and Simar (2005, 2007a) how useful is the analysis of the regression line of $\tau^z(P)$ over z. For instance, in the output orientation, an increasing regression corresponds to a favorable effect of Z (higher values of Z allow to reach higher outputs, Z is acting as a free available input) and the opposite for a decreasing regression (Z is acting as an undesirable output). A nonparametric estimator of the regression line will be introduced below and an algorithm for providing pointwise confidence intervals will also be described.

We will now clarify what the expected ratio $\tau^z(P)$ really measures and what the partial ratios can add in the analysis.

First, it should be noticed that the conditional "full" parameter $\tau^z(P)$ only brings information on potential differences between the boundaries of Ψ and Ψ^z and is not sensitive to changes in the distribution of inefficiencies. It is obvious that the measure $R(x, y|Z = z) \leq 1$ for a fixed point (x, y) only depends on the relative position of the boundaries of Ψ and Ψ^z (in the radial direction given by y). This is true for all $(x, y, z) \in \mathcal{P}$, so it is true for $R(X, Y|Z = z)$ and for its expectation $\tau^z(P)$. This is illustrated below, in Figure 1, for the particular case of a univariate output. Here $\lambda(x_0, y_0) = \varphi(x_0)/y_0$ and $\lambda(x_0, y_0|z_0) = \varphi_{z_0}(x_0)/y_0$ so that $R(x_0, y_0|z_0) = \varphi_{z_0}(x_0)/\varphi(x_0)$, with a similar expression for $R_\alpha(x_0, y_0|z_0)$. Different distributions of the inefficiencies (conditional and unconditional) but having same support, result in ratios $R(x_0, y_0|z_0) = 1$, as illustrated in the left panels of Figure 1: we see indeed in panel II and III that $\varphi_{z_0}(x_0) \equiv \varphi(x_0)$.

Second, the information carried by the conditional "partial" parameter $\tau^z_\alpha(P)$ is multiple. Suppose that $\Psi^z = \Psi$ and so $\tau^z(P) = 1$ (the support of (X, Y) is not changed) then, if the distribution of inefficiencies is affected by Z, the quantiles of $S_{Y|X,Z}$ will be different from those of $S_{Y|X}$. Therefore for all $(x, y) \in \Psi^z$, the ratio $R_\alpha(x, y|z)$ will be affected and so will be their average. Note that in this case ($\Psi^z = \Psi$), the changes can go in two directions for the partial parameter: if the distribution of the inefficiency is more spread in the direction of less efficient behavior (as in panel II), we observe $\varphi_{\alpha,z_0}(x_0) < \varphi_{z_0}(x_0)$ giving $R_\alpha(x_0, y_0|z_0) < 1$ and so the expectation $\tau^{z_0}_\alpha(P)$ may be less than 1. On the contrary, if z_0 provides a favorable environment to efficient behavior of the firms, the distribution of Y will be more concentrated near the efficient boundary when $Z = z_0$ (as in panel III), we have $\varphi_{\alpha,z_0}(x_0) > \varphi_{z_0}(x_0)$ giving $R_\alpha(x_0, y_0|z_0) > 1$ and we might have on the average $\tau^{z_0}_\alpha(P) > 1$. That is the reason why the global test of "separability" of Daraio et al. (2010) uses statistics only based on the full measures of efficiency and not on the partial efficiency scores.

Third, if there is a shift on the frontier $\Psi^z \neq \Psi$ with $\tau^z(P) < 1$, it is much more difficult to interpret the ratios $R_\alpha(x, y|z)$. It is clear that a shift of the boundary will be transferred to the partial frontier, at least for large values of α , but this effect can either be increased or compensated by a simultaneous change of the distribution of the inefficiencies. So, in the case of a shift of the boundary (see the right panels of Figure 1), we could observe $R_\alpha(x_0, y_0|z_0)$ less, equal or greater than 1. We illustrate 3 cases in Figure 1. We see that in panel IV, the shift of $\varphi_{\alpha,z_0}(x_0)$ with respect to $\varphi_\alpha(x_0)$ is the same as the shift of $\varphi_{z_0}(x_0)$ with respect to $\varphi(x_0)$, giving here $R_\alpha(x_0, y_0|z_0) < R(x_0, y_0|z_0) < 1$. In panel V, we have more spread toward inefficiencies when conditioning on z_0, the shift of the quantile of the conditional distribution is much more important so $R_\alpha(x_0, y_0|z_0) \ll R(x_0, y_0|z_0) < 1$. But we could observe, as in panel VI, a different behavior when given z_0 it is more probable to reach the frontier $\varphi_{z_0}(x_0)$

implying that we could obtain for some quantiles $R_\alpha(x_0, y_0|z_0) > R(x_0, y_0|z_0)$. So even if $R(x_0, y_0|z_0) < 1$ we could have in extreme cases $R_\alpha(x_0, y_0|z_0) \geq 1$ (in panel VI, we illustrate the case where $R_\alpha(x_0, y_0|z_0) > 1$).

So, to summarize the second and third points above, if $\Psi^z = \Psi$, $\tau^z_\alpha(P)$ is useful to shed light on the local impact of Z on the shape of the distribution of the inefficiencies. But it does not allow to detect, when considered alone, a local shift of the boundary of the support of (X, Y). Unless $\alpha \to 1$, because in this case, the partial frontier can serve as a robust estimator of the full frontier (see in the next section).

In any cases, these partial measures bring useful complementary information of the relative position of the quantiles of $S_{Y|X,Z}$ with respect to those of $S_{Y|X}$. It will therefore be useful to provide the regression lines $\tau^z(P), \tau^z_{\alpha_1}(P), \ldots, \tau^z_{\alpha_k}(P)$ on z, for a grid of selected values for α like, $0.99, 0.95, 0.90; \ldots, 0.50$. The latter case $\alpha = 0.50$ is providing for instance, a picture on the impact of z on the median of the inefficiency distribution as a function of z.

The same is true for the order-m partial parameters $\tau^z_m(P)$ where the particular case $m = 1$ would provide a picture of the effect of z on the average frontier. Here, the choice of large values of m would provide the same information as the full frontier parameter (see in the next section).

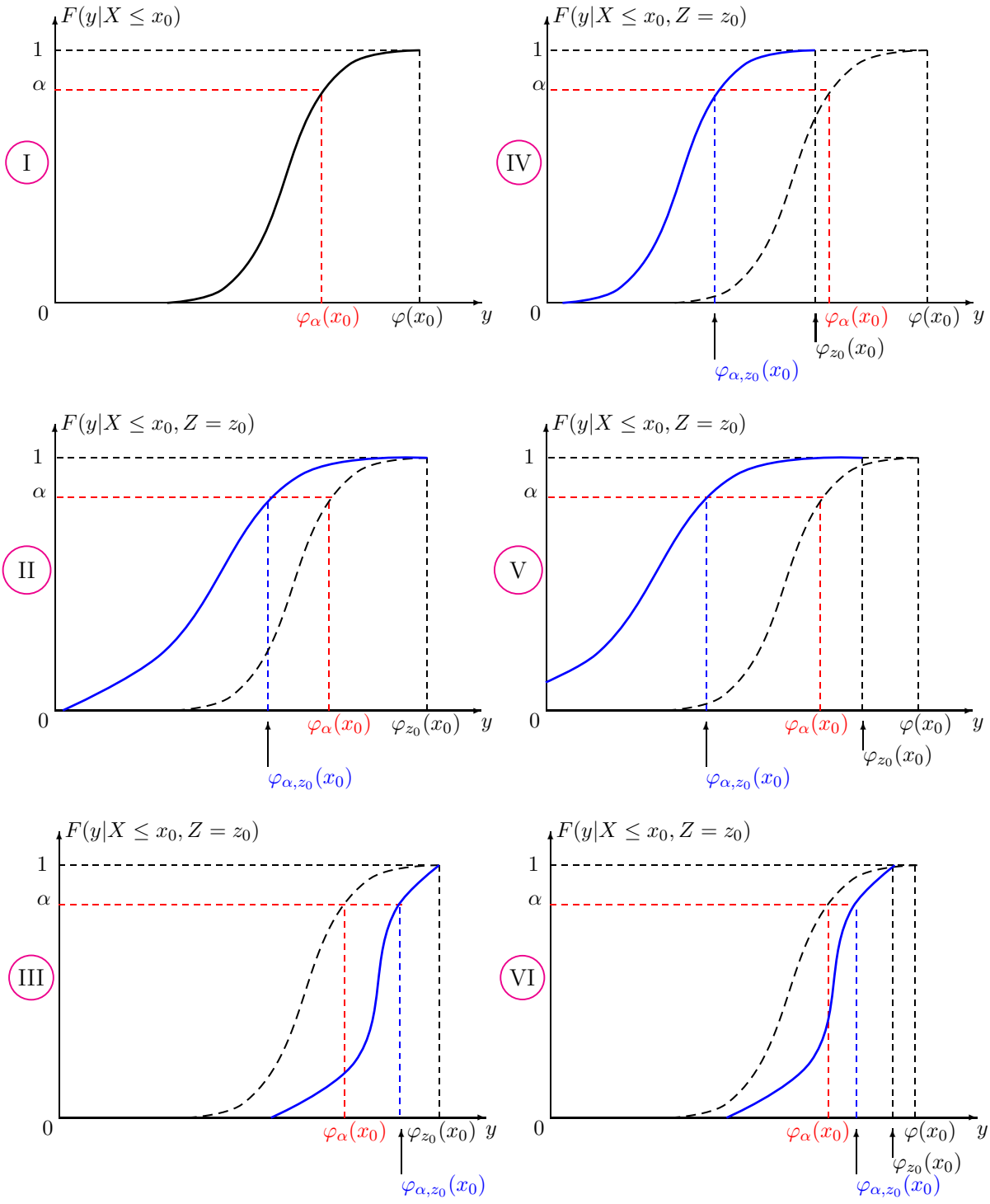

Figure 1: *Various scenarios for* $F(y|X \le x_0)$ *and* $F(y|X \le x_0, Z = z_0)$ *. In the left panels the "separability" condition is verified, while on the right panels, this condition is violated.*

3 Nonparametric Estimator

3.1 Efficiency Estimators

Nonparametric estimators of the conditional and unconditional efficiency scores are very easy to obtain. We summarized the notations and properties here to what is needed for the rest of the paper (details can be found in Daraio and Simar, 2007a, or Simar and Wilson, 2008). We will denote $\mathcal{S}_n = \{(X_i, Y_i, Z_i)|\ i = 1, \ldots, n\}$ the sample of n iid observations on (X, Y, Z) generated in $\mathcal{P}$ according the DGP $P \in \mathbb{P}$. If we plug nonparametric estimators of $S_{Y|X}$ and $S_{Y|X,Z}$ in all the formulae above, we obtain very natural nonparametric estimators of the efficiencies. For the $S_{Y|X}$ we can use the empirical probabilities

$$\widehat{S}_{Y|X}(y_0|X \leq x_0) = \frac{1/n \sum_{i=1}^n \mathbb{1}(X_i \leq x_0, Y_i \geq y_0)}{1/n \sum_{i=1}^n \mathbb{1}(X_i \leq x_0)}, \tag{3.1}$$

where $\mathbb{1}(\cdot)$ is the indicator function. This provides the popular FDH estimator of $\lambda(x_0, y_0)$

$$\widehat{\lambda}(x_0, y_0) = \max_{\{i|X_i \leq x_0\}} \left\{ \min_{j=1,\ldots,q} \frac{Y_i^j}{y_0^j} \right\} \tag{3.2}$$

whose statistical properties are well known (see e.g. Simar and Wilson, 2008). To summarize, under mild regularity conditions:

$$n^{1/(p+q)} \left(\lambda(x_0, y_0) - \widehat{\lambda}(x_0, y_0) \right) \xrightarrow{\mathcal{L}} \text{Weibull}(\mu_0^{p+q}, p+q), \tag{3.3}$$

where μ_0 is a constant depending on the DGP $P \in \mathbb{P}$ that is described in Park et al. (2000). For the conditional (conditional to $Z = z_0$) some smoothing techniques are required. We have the estimator

$$\widehat{S}_{Y|X,Z}(y_0|X \leq x_0, Z = z_0) = \frac{1/n \sum_{i=1}^n \mathbb{1}(X_i \leq x_0, Y_i \geq y_0) K((z_0 - Z_i)/b)}{1/n \sum_{i=1}^n \mathbb{1}(X_i \leq x_0) K((z_0 - Z_i)/b)}, \tag{3.4}$$

where for simplicity, we wrote the expression for a univariate Z. Here $K(\cdot)$ is a kernel with compact support and $b > 0$ is the bandwidth. For the general multivariate case, see Daraio and Simar (2007a). In the general multivariate setup, an optimal bandwidth selection procedure has been suggested in Bădin et al. (2010), it is based on a least-squares cross validation technique. This leads to the conditional efficiency estimator

$$\widehat{\lambda}(x_0, y_0|z_0) = \max_{\{i|X_i \leq x_0, ||Z_i - z_0|| \leq b\}} \left\{ \min_{j=1,\ldots,q} \frac{Y_i^j}{y_0^j} \right\} \tag{3.5}$$

So, it appears that the estimation of the conditional efficiency score is a kind of "restricted" FDH program (restricted to data points having $||Z_i - z_0|| \leq b$). The statistical properties

of the estimators of the conditional measures have been determined in Jeong et al (2010). To summarize and roughly speaking, these estimators keep similar properties as the FDH estimator but with an "effective" sample size depending on the bandwidth: n is replaced by nb^r, where r is the dimension of Z. In practice since the optimal bandwidth has a size $n^{-1/(r+4)}$ (see Bădin et al., 2010 for details), this gives a rate of convergence for the conditional measures estimators of $n^{4/((r+4)(p+q))}$ in place of the better rate $n^{1/(p+q)}$ achieved by the FDH estimators. It is important to report these rates in order to derive below a consistent bootstrap algorithm.

The nonparametric partial frontier efficiency estimates are obtained in a similar way, by plugging the estimators $\widehat{S}_{Y|X}$ and $\widehat{S}_{Y|X,Z}$ in the expressions defining the partial efficiency scores: algorithms have been proposed in Cazals et al. (2002), Daraio and Simar (2005, 2007a) for the order-m case and in Daouia and Simar (2006) and Daraio and Simar (2007a) for the order-α quantile case. Their statistical properties have been also established. Under mild regularity conditions, we have for instance

$$\sqrt{n}\left(\lambda_\alpha(x_0,y_0) - \widehat{\lambda}_\alpha(x_0,y_0)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0,\sigma^2(\alpha,x_0)\big), \tag{3.6}$$

where an expression for $\sigma^2(\alpha,x_0)$ is given in Daouia and Simar (2006). A similar result holds for the order-m case (see Cazals et al. 2002).

For the estimators of the conditional partial measures, we have similar results where the rate of convergence $\sqrt{n}$ deteriorates to $\sqrt{nb^r} = n^{2/(r+4)}$ when the optimal bandwidth of Bădin et al. (2010) described above is used.

Robust Estimators of the Full Frontier

As explained above the partial frontiers may have their own interest providing less extreme surfaces to benchmark individual units and allowing to investigate the impact of Z on the distribution of the efficiencies. In particular for $m = 1$, the order-m frontier is not looking to an optimal behavior but rather to an average behavior of firms (the same is true for the order-α frontier with $\alpha = 0.50$).

But as pointed and illustrated in Daraio and Simar (2007a) it may happen that outliers or extreme data points can hide the real effect of the environmental factors. So, in this case, it is particularly useful to build robust estimators of the full frontier. This can be achieved by using partial order frontier with extreme orders.

Indeed, if we let $\alpha = \alpha(n) \to 1$ (or $m = m(n) \to \infty$) when $n \to \infty$ fast enough (see Cazals et al., 2002 and Daouia and Simar, 2006, for details), the respective partial frontier estimators will converge to the full frontier sharing the same properties as the FDH estimator (with the same limiting Weibull distribution). But for finite n (as we use in practice), $\alpha(n)$ will be less than 1 (and $m(n)$ will be less than infinity) and so the corresponding estimate

of the full frontier will not envelop all the data points being more robust and resistant to outliers and extreme values than the standard envelopment estimators like FDH or DEA.

Simar (2003) has suggested some data driven techniques to select reasonable values of α and m by analyzing the proportion of data points remaining outside the corresponding partial frontiers over a grid of values of the orders. This allows to detect potential outliers. Daouia and Gijbels (2009) propose a theoretical comparisons of both partial frontiers in terms of their robustness properties and give a rule with more theoretical background for selecting appropriate levels of α or m. In Daouia and Gijbels (2010), a semi-automatic practical rule is given to select the appropriate order of the partial frontier for obtaining robust estimators of the full frontier (and the corresponding efficiency scores) in the presence of outliers.

3.2 Estimation of the Regression

To save place we only present the full frontier case, where we want to estimate $\tau^z(P) = \mathbb{E}(R(X,Y|Z=z)$ by using basic tools from the nonparametric econometrics literature (see e.g. Pagan and Ullah, 1999). We will simplify the presentation to univariate continuous Z, but this can be done for any dimension r of Z.[1]

We do not have iid observations of $R(X_i,Y_i|Z=z)$, neither iid observations $R(X_i,Y_i|Z_i) = \lambda(X_i,Y_i|Z_i)/\lambda(X_i,Y_i)$ because the lambda's are unknown. What we only have is the set of the n estimators (obtained from the sample $\mathcal{S}_n$):

$$\widehat{R}(X_i,Y_i|Z_i) = \frac{\widehat{\lambda}(X_i,Y_i|Z_i)}{\widehat{\lambda}(X_i,Y_i)},$$

so that we have a sample of n pairs $\big(Z_i, \widehat{R}(X_i,Y_i|Z_i)\big)$, $i=1,\ldots,n$ from which we will estimate $\tau^z(P)$. Most of the nonparametric estimates of the regression function (including Nadaraya-Watson, local linear, etc...) can be written as

$$\hat{\tau}_n^z = \sum_{i=1}^n W_n(Z_i,z,h_z)\widehat{R}(X_i,Y_i|Z_i), \tag{3.7}$$

with the weights $W_n(Z_i,z,h_z) \geq 0$ summing up to one. This is a local average of the $\widehat{R}(X_i,Y_i|Z_i)$, the localization being tuned by the bandwidth h_z. The Nadaraya-Watson kernel weights are given by

$$W_n(Z_i,z,h_z) = \frac{K\big((Z_i-z)/h_z\big)}{\sum_{i=1}^n K\big((Z_i-z)/h_z\big)}.$$

For local linear estimator we have rather (see Fan and Gijbels, 1996)

$$W_n(Z_i,z,h_z) = \frac{w_n(Z_i,z,h_z)}{\sum_{i=1}^n w_n(Z_i,z,h_z)}, \text{ where } w_n(Z_i,z,h_z) = K\Big(\frac{Z_i-z}{h_z}\Big)\left[S_{2,n} - (Z_i-z)S_{1,n}\right],$$

[1] For more details on how to handle discrete variables in this framework, see Bădin and Daraio (2010).

where $S_{j,n} = \sum_1^n K\big((Z_i - z)/h_z\big)(Z_i - z)^j$, $j = 1, 2$.

As usual in nonparametric regression, bandwidth h_z with appropriate size (i.e. $h_z = c\, n^{-1/(r+4)}$) can be obtained by least-squares crossvalidation criterion (see e.g. Li and Racine, 2007 for details).

In order to provide pointwise confidence intervals for $\tau^z(P)$ we need to derive the statistical properties of $\hat{\tau}_n^z$. Standard theory cannot be applied because we do not observe independent pairs $(Z_i, R(X_i, Y_i|Z_i))$ but rather the pairs $(Z_i, \widehat{R}(X_i, Y_i|Z_i))$, where, as noticed above, $\widehat{R}(X_i, Y_i|Z_i)$ are estimates of $R(X_i, Y_i|Z_i)$. Hopefully, we can follow the same argument as in Simar and Wilson (2010) and Daraio et al. (2010) to obtain the asymptotic distribution of our regression estimate. This will be sufficient to prove the consistency of the bootstrap we propose in the next section.

Under regularity conditions, we obtain (a sketch of the proof is proposed in Appendix A.1) the following result, as $n \to \infty$, $h_z \to 0$ with $nh_z^r \to \infty$

$$\sqrt{nh_z^r}\Big(\hat{\tau}_n^z - \tau^z(P) - n^{-\kappa}\mu_{Q_P}^z - h_z^2(B^z + n^{-\kappa}C^z)\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, V^z\big). \tag{3.8}$$

where B^z, C^z, μ_Q^z, V^z are bounded constants described in the Appendix and $\kappa = 4/((r + 4)(p + q))$ is determined by the rate of convergence of the conditional efficiency estimator that is used.[2]

We see that we have the usual bias term B^z coming from the nonparametric regression and a second bias term, coming from the estimation of the $R(X_i, Y_i|Z_i)$ by $\widehat{R}(X_i, Y_i|Z_i)$ that disappears when n increases at a rate $n^{-\kappa}$. As shown below this latter bias term can be neglected in our bootstrap approach.

Balancing the bias coming from h_z and the variance term, it is well known in the nonparametric literature (see e.g. Pagan and Ullah, 1999) that the optimal size of the bandwidth h_z for the regression is $h_z = cn^{-1/(r+4)}$ which is achieved by using least-squares crossvalidation for selecting h_z. With this choice we have

$$n^{2/(r+4)}\Big(\hat{\tau}_n^z - \tau^z(P) - \frac{\mu_Q^z}{n^\kappa}\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(cB^z, V^z\big), \tag{3.9}$$

where in the asymptotic normal, the bias term $n^{-\kappa}C^z$ can be neglected since it is an $o(1)$.

3.3 Confidence Intervals for the Regression

For building confidence intervals for $\tau^z(P)$ by using the bootstrap, we cannot use the standard algorithms as in Härdle and Bowman (1988) or Härdle and Marron (1991), because

[2]To save place we do not explicit the results for the partial frontier measures, but we would obtain in this case similar results as (3.8) with $\kappa = 2/(r + 4)$, due to the better rate of convergence of the efficiency estimators. This does not change the nonparametric rate $\sqrt{nh_z^r}$ we obtain in (3.8) for $\hat{\tau}_n^z$. So confidence intervals for $\tau_\alpha^z(P)$ could be obtained by following the same algorithm as the one described in the next section.

the $R(X_i, Y_i|Z_i)$ are not directly observed and the available pairs $(Z_i, \widehat{R}(X_i, Y_i|Z_i))$ are not independent. In addition bootstrapping on the pairs $(Z_i, \widehat{R}(X_i, Y_i|Z_i))$ would neglect all the noise introduced by estimating $R(X_i, Y_i|Z_i)$ by $\widehat{R}(X_i, Y_i|Z_i)$.[3]

The original independent data are the (X_i, Y_i, Z_i), $i = 1, \ldots, n$. So we will use, as in Simar and Wilson (2010), the m out of n bootstrap on the triple (X_i, Y_i, Z_i) to approximate the sampling distribution of $\left(\hat{\tau}_n^z - \tau^z(P)\right)$. Its consistency is established by Theorem 2.1 in Politis et al (2002) but we give below the main ideas.

We will consider a bootstrap sample of m observations drawn without replacement from the sample $\mathcal{S}_n = \{(X_i, Y_i, Z_i)|\ i = 1, \ldots, n\}$. Since the original sample was an iid random sample of size n generated by the DGP $P \in \mathbb{P}$, this subsample, denoted by $\mathcal{S}_m$, can be considered as a random iid sample of size m drawn from the same P. We will consider $m = m(n) \to \infty$ as $n \to \infty$ with $m/n \to 0$. So by (3.9), we have

$$\left(\hat{\tau}_n^z - \tau^z(P)\right) \sim \frac{\mu_Q^z}{n^\kappa} + \mathcal{AN}\left(\frac{cB^z}{n^{2/(r+4)}}, \frac{V^z}{n^{4/(r+4)}}\right) \tag{3.10}$$

$$\left(\hat{\tau}_m^z - \tau^z(P)\right) \sim \frac{\mu_Q^z}{m^\kappa} + \mathcal{AN}\left(\frac{cB^z}{m^{2/(r+4)}}, \frac{V^z}{m^{4/(r+4)}}\right), \tag{3.11}$$

where $\hat{\tau}_m^z$ is the same estimator as $\hat{\tau}_n^z$ but computed with the sample $\mathcal{S}_m$. If the distribution appearing in (3.10) was completely known, it would be easy to find its quantiles $q_{n;\alpha}$, where $\text{Prob}\left(\hat{\tau}_n^z - \tau^z(P)\right) \leq q_{n;\alpha}\big) = \alpha$. Then a $(1-\alpha) \times 100\%$ confidence interval would be given by

$$\tau^z(P) \in \left[\hat{\tau}_n^z - q_{n;1-\alpha/2}, \hat{\tau}_n^z - q_{n;\alpha/2}\right]. \tag{3.12}$$

The quantiles $q_{n;\alpha}$ are unknown but they can be approximated from the normal approximation (3.10) by

$$q_{n;\alpha} \approx \frac{\mu_Q^z}{n^\kappa} + \frac{cB^z}{n^{2/(r+4)}} + \frac{\sqrt{V^z}}{n^{2/(r+4)}}\, z_\alpha,$$

where z_α is the quantile of the standard normal. Note that since μ_Q, B^z and V^z are unknown, this is not very helpful. But we have the same relation for the quantiles $q_{m;\alpha}$ of $\hat{\tau}_m^z - \tau^z(P)$:

$$q_{m;\alpha} \approx \frac{\mu_Q^z}{m^\kappa} + \frac{cB^z}{m^{2/(r+4)}} + \frac{\sqrt{V^z}}{m^{2/(r+4)}}\, z_\alpha.$$

So we see that for $m, n \to \infty$ with $m/n \to 0$ we have

$$q_{n;\alpha} - (m/n)^{2/(r+4)} q_{m;\alpha} = \mu_Q^z \left[\frac{1}{n^\kappa} - \frac{(m/n)^{2/(r+4)}}{m^\kappa}\right] = o(1), \tag{3.13}$$

[3]It should be noticed that we are not interested in the individual random variables $R(X_i, Y_i|Z_i))$, but rather in the expectation $\tau^z(P)$, given that $Z = z$, and to analyze this as a function of z. Individual confidence interval for a particular fixed point of interest for $R(x_0, y_0|z_0)$ could be obtained by standard bootstrap techniques as described in Kneip et al. (2008, 2010) or in Simar and Wilson (2010).

that indicates that the quantiles of $(m/n)^{2/(r+4)}\left(\hat{\tau}_m^z - \tau^z(P)\right)$ can be used to approximate those of $\left(\hat{\tau}_n^z - \tau^z(P)\right)$, and that the bias term introduced by μ_Q^z can be neglected (as confirmed empirically in Simar and Wilson, 2010a, by intensive Monte-Carlo experiments in similar setups).

Of course the quantiles of $\left(\hat{\tau}_m^z - \tau^z(P)\right)$ are also unknown but they can be approximated by the Monte-Carlo part of the bootstrap algorithm. First we see that in $(m/n)^{2/(r+4)}\left(\hat{\tau}_m^z - \tau^z(P)\right)$ we can replace $\tau^z(P)$ by $\hat{\tau}_n^z$ since by doing so we add an error $o_p(1)$ of smaller order (the order is $(m/n)^{2/(r+4)} \times n^{-(\kappa \wedge 2/(r+4))}$).

So, in practice the unknown quantiles $q_{n;\alpha}$ will be approximated by $(m/n)^{2/(r+4)}q^*_{m;\alpha}$ where $q^*_{m;\alpha}$ are the bootstrap approximations of $q_{m;\alpha}$. For a given m, we construct the N_m subsets $\mathcal{S}^*_{m,b}$, $b = 1, \ldots, N_m$, of size m drawn without replacement from $\mathcal{S}_n$.[4] The sampling distribution of $\left(\hat{\tau}_m^z - \tau^z(P)\right)$ is then approximated by

$$\widehat{G}_{m,n}(w) = \frac{1}{N_m}\sum_{b=1}^{N_m} \mathbb{1}\left(\hat{\tau}^{*,z}_{m,b} - \tau_n^z \leq w\right), \tag{3.14}$$

where $\hat{\tau}^{*,z}_{m,b}$ is the version of $\hat{\tau}_m^z$ applied to the sample $\mathcal{S}^*_{m,b}$. The quantiles of $\widehat{G}_{m,n}(w)$ are given by

$$q^*_{m;\alpha} = \inf\{w | \widehat{G}_{m,n}(w) \leq \alpha\}. \tag{3.15}$$

The bootstrap $(1-\alpha)\times 100\%$ confidence interval for $\tau^z(P)$ is thus given by

$$\tau^z(P) \in \left[\hat{\tau}_n^z - (m/n)^{2/(r+4)}q^*_{m;1-\alpha/2}, \hat{\tau}_n^z - (m/n)^{2/(r+4)}q^*_{m;\alpha/2}\right]. \tag{3.16}$$

A formal proof of the consistency of this m out of n bootstrap is given in Theorem 2.1 in Politis et al.(2002). The only remaining question is how to select m in practice. We follow the data driven method described in Simar and Wilson (2010a).

3.4 The bootstrap algorithm

The bootstrap algorithm can thus be described as follows.

[1] First we compute from the sample $\mathcal{S}_n = \{(X_i, Y_i, Z_i)|\ i = 1, \ldots, n\}$ the n efficiency scores $\widehat{\lambda}(X_i, Y_i)$ and their conditional version $\widehat{\lambda}(X_i, Y_i|Z_i)$. By doing so, for each data point we compute the optimal bandwidth for the conditional survival function at Z_i (we do this by using the Bădin et al. (2010) approach). We thus have n optimal bandwidths $h_{n,i}$ each attached to the ith observation. We compute the n ratios $\widehat{R}(X_i, Y_i|Z_i)$.

[4]The number of subsets N_m can be a huge number: $N_m = \binom{n}{m}$. In practice, of course, we do not compute all these subsets, but we would just take a random selection of B such subsamples, where B should not be too small.

[2] We select a fixed grid of values for Z, say $\{z_1, \ldots, z_k\}$ where the regression will be evaluated. We compute the nonparametric regression by one of the methods described in (3.7): this provides $\hat{\tau}_n^{z_j}$ for $j = 1, \ldots, k$. Here the bandwidth h_n^z is selected by least-squares crossvalidation.

[3] For a given value of $m < n$, we will repeat the next steps [3.1] to [3.3] B times, for $b = 1, \ldots, B$, where B is large enough (say, $B = 2000$).

[3.1] Draw a random sample $\mathcal{S}_{m,b}^* = \{(X_i^{*,b}, Y_i^{*,b}, Z_i^{*,b}) |\ i = 1, \ldots, m\}$ without replacement from $\mathcal{S}_n$. By doing so, we keep also the value of the bandwidth $h_{n,i}^{*,b}$ computed at step [1] attached to the corresponding selected data $(X_i^{*,b}, Y_i^{*,b}, Z_i^{*,b})$.

[3.2] We compute the m ratios $\widehat{R}^{*,b}(X_i^{*,b}, Y_i^{*,b} | Z_i^{*,b})$, $i = 1, \ldots, m$ by the same techniques as in [1]. Note that here we have to rescale the corresponding bandwidths $h_{n,i}^{*,b}$ at the appropriate size. So we will use the bandwidths $h_{m,i}^{*,b} = (n/m)^{1/(r+4)} h_{n,i}^{*,b}$ for computing the conditional scores in the bootstrap sample $\mathcal{S}_{m,b}^*$.

[3.3] By the same nonparametric method as in [2], we estimate the regressions $\hat{\tau}_m^{*,b,z_j}$ at the fixed points z_j, for $j = 1, \ldots, k$. For doing so we use the same bandwidth computed in [2] but rescaled to the appropriate size.[5] So we will use here $h_m^z = (n/m)^{1/(r+4)} h_n^z$. We obtain $\hat{\tau}_m^{*,b,z_j}$ for $j = 1, \ldots, k$.

[4] For each $j = 1, \ldots, k$, compute $(q_{m;\alpha/2}^{*,z_j}, q_{m;1-\alpha/2}^{*,z_j})$, the $\alpha/2$ and $1 - \alpha/2$ quantiles of the B bootstrapped values of $\hat{\tau}_m^{*,b,z_j} - \hat{\tau}_n^{z_j}$. This provides the k confidence intervals of $\tau^{z_j}(P)$ at each fixed z_j:

$$\tau^{z_j}(P) \in \left[\hat{\tau}_n^{z_j} - (m/n)^{2/(r+4)} q_{m;1-\alpha/2}^{*,z_j}, \hat{\tau}_n^{z_j} - (m/n)^{2/(r+4)} q_{m;\alpha/2}^{*,z_j} \right]. \tag{3.17}$$

The selection of m is done as follows. We redo the steps [3] to [4] over a grid of L values of m, say, $m_1 < m_2 < \ldots < m_L$ and we obtain for each m_ℓ, the k resulting confidence intervals (3.17).[6] Then we compute the volatility of the quantity of interest seen as a function of m. Here the two bounds of the confidence intervals (3.17) are of the quantities of interest, Politis et al. (2002) suggest in this case to take $c^{z_j}(m) = (1/2)[\text{low}_m^{z_j} + \text{up}_m^{z_j}]$, where the notation is implicit. The volatility is measured by the "moving" standard deviation of 3 adjacent values of $c^{z_j}(m)$ centered at the current value of m_ℓ, $\ell = 2, \ldots, L-1$. As explained in Politis et al. (2002), a reasonable value for m^{z_j} should correspond to the value that minimizes this

[5] Here we could recompute the bandwidth h_m^z by crossvalidation, but at a computational cost. By doing what is suggested in [3.3], we achieved the desired theoretical order of the bandwidth.

[6] The choice of this grid is really open and depends on the computational burden: we should cover a wide spectrum of values for m. Simar and Wilson (2010a) and Daraio et al. (2010) suggest, for instance, to choose the 49 subsamples sizes $m \in \{[n/50], 2[n/50], \ldots, 49[n/50]\}$, where $[a]$ denotes the integer parts of a.

volatility. Intensive Monte-Carlo experiments in Simar and Wilson (2010a) and Daraio et al. (2010), in similar setups of nonparametric frontier estimation, indicate that this procedure provides very good results in terms of coverage, size of tests, power of tests, etc.

A simpler alternative is to select a common value of m for the different values of z_j. We could for instance select the m equal to the average of all the m^z. We could also use the same approach as above, but here, the volatility would be measured on an average value $c(m) = (1/k)\sum_j c^{z_j}(m)$. This approach could provide more stable behavior of $c(m)$ as a function of m. In the simulation examples shown below, it appears that we have very little differences by using either approach. The optimal values for m^{z_j} were rather stable across the selected grid for z. In all the results shown below, we used the more general approach where we select an optimal m different for each z_j.

4 Numerical Illustrations

4.1 Simulated Examples

To illustrate how the procedure can work in practice, we first introduced some simulated examples, because there we know what we expect to find. We will use, as simulated scenario, an example inspired from Simar and Wilson (2010b) where we see clearly the 2 different ways an environmental factor can influence the production process. We analyze the three following different DGPs:

$$Y = g(X)e^{-U} \tag{4.1}$$

$$Y^* = g(X)e^{-U(1+|Z-2|/2)} \tag{4.2}$$

$$Y^{**} = g(X)e^{-(1+|Z-2|/2)}e^{-U}, \tag{4.3}$$

where $g(X) = [1-(X-1)^2]^{1/2}$ with $X \sim U(0,1)$ and $Z \sim U(0,4)$. Finally $U \geq 0$ with $U \sim \mathcal{N}^+(0,\sigma_U^2)$ and we choose for the illustration $\sigma_U^2 = 0.10$.

In the first DGP1 (4.1), Z has no effect on the production process (Z is independent of (X,Y)). In the DGP2 (4.2), we have the "separability" condition $\Psi^z \equiv \Psi$, $\forall z$ but Z influences the distribution of the inefficiencies (higher probability of being inefficient when $|Z-2|$ increases). In the last DGP3 (4.3), the effect of Z is only on the boundary of the attainable (X,Y), violating the "separability" condition (the shift is multiplicative and more important when $|Z-2|$ increases). A summary of the results for the case $n = 100$ is displayed in Figure 2; the following comments will be useful to understand what we learn and what we do not learn by looking to these pictures.

In DGP1 (the 2 top panels), the Z is independent of (X,Y), we see indeed the flat behavior of our regressions. With 100 observations, the random fluctuations of the estimates

of the efficiency scores (full or α-quantile frontiers) are responsible on the right panels for some deviations from the flat lines. There is also an edge effect due to the nonparametric regression, we have always less precision of the regressing estimate near the edge of the cloud of points (this will remain true for all the pictures below).

For the first 2 DGP's, the "separability" condition is verified and so the true (unobserved) $R(Z) = 1$ with probability 1. This explains why the confidence intervals for DGP1 and DGP2 are so narrow in both left panels: the randomness comes only from the fact that we use estimators of $\widehat{R}(Z)$ in place of $R(Z)$. For DGP2, in the left panel, we see at both ends of the picture some small spurious effect on $\hat{\tau}_n^z$, where it should be flat. This is due to the fact that when Z approaches 0 or 4, the inefficiency distribution gives more probability to inefficiency. It is well known, that the precision of the FDH estimator deteriorates when the probability of observing firms near the boundary decreases (which is the case when $|Z-2|$ increases). The greater statistical noise induces the small spurious effect at both ends (we will see that this effect disappears when the sample size increases). However we see clearly on the right panel of DGP2 that the effect of Z on the inefficiency distribution is really present. The inverse-U-shaped effect reflects the heteroscedasticity in the distribution of inefficiencies. The lower level quantile frontiers are more sensitive to the change of Z. The effect is more important near the center: when $Z = 2$, there is no "bad" effect on the efficiency distribution, so, using conditional measures (to $Z = 2$), a firm is benchmarked, at the α level, against very efficient firms leading to higher values of $\hat{\lambda}_\alpha(x, y|z)$ with respect to the marginal measure $\hat{\lambda}_\alpha(x, y)$ where we do not take into account for the favorable environment and $\hat{R}_\alpha(z)$ will be larger than 1. This is not possible to detect when $\alpha \to 1$, because the boundary has not moved.

For DGP3, we see here clearly that the separability condition is violated. We observe that the confidence intervals are wider because $R(Z)$ has now its own randomness, increasing statistical imprecision. It seems clear that the interpretation of the $\hat{\tau}_{\alpha,n}^z$ for small values of α, without looking to what happens when $\alpha \to 1$ is very difficult. Compare the upper-most dash dot red lines ($\alpha = 0.50$) for DGP2 and DGP3: without a clear information on the separability issue, we cannot identify the source of the impact of Z.

When n is larger, we would confirm all the above facts with more evidence. Figure 3 illustrates this for $n = 200$. For instance, in DGP2, the quantile frontiers are better estimated and the effect of Z on the distribution of inefficiencies appears more clearly. But the full frontier case (left panel) displays a quite flat shape, as it should. In the DGP3, the left panel indicates clearly the role of Z on the boundary (almost no effect when Z is around 2). In the right panel, the regression curves, corresponding to different levels of the quantiles, confirm the inverse-U-shape effect of Z on the frontier level.[7] The curves are not parallel because

[7]In order to understand the impact of Z on the production processes in DGP2 and DGP3, the reader

the effect in DGP3 is multiplicative and not additive.

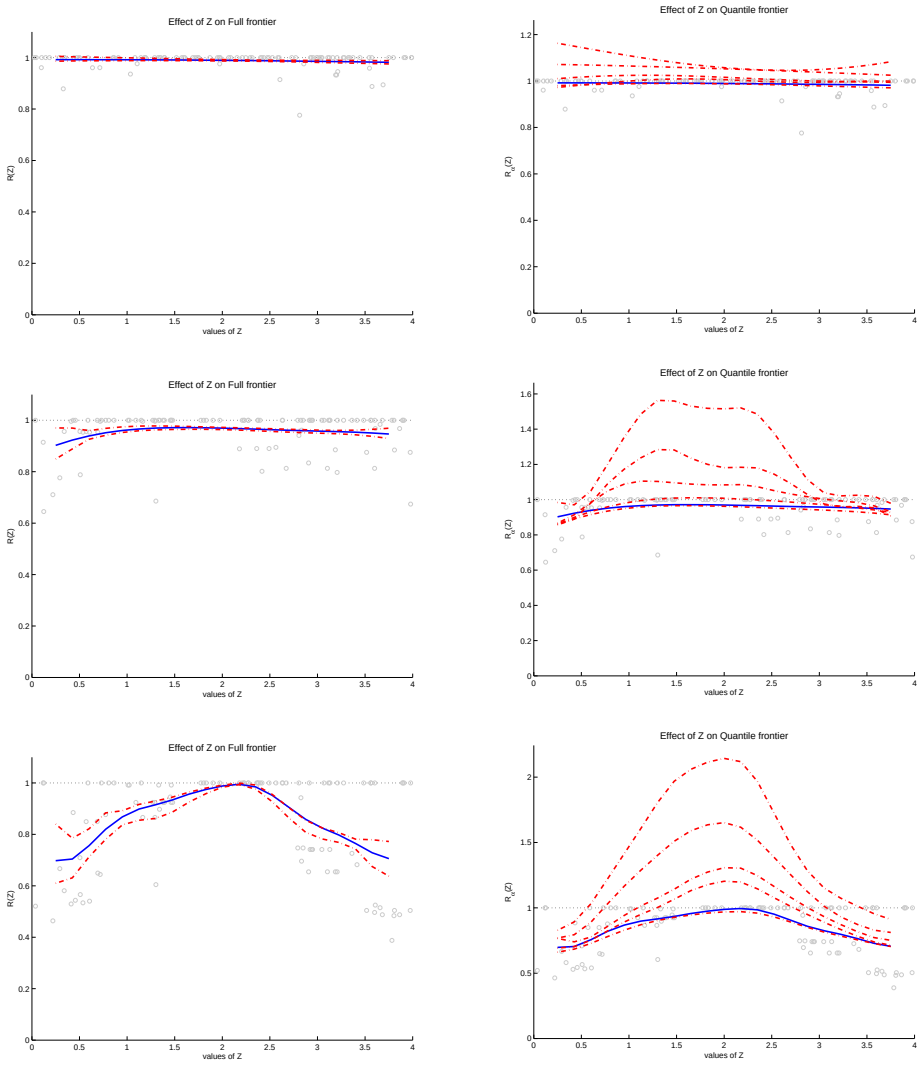

Figure 2: *Regressions of the ratios $R(Z)$ on Z. From top to bottom: DGP1, DGP2 and DGP3. Left panels, full frontier $\hat{\tau}_n^z$ with 95% confidence intervals for $\tau^z(P)$. Right panels, $\hat{\tau}_{\alpha,n}^z$ for $\alpha = (0.5, 0.75, 0.90, 0.95, 0.99, 1.00)$, the last one (full frontier case) in solid line. Here $n = 100$ and the circles are the estimated data points $(Z_i, \widehat{R}(Z_i))$.*

can verify that if the factors $(1 + |Z - 2|/2)$ would be replaced by $(1 - |Z - 2|/2)$ all the inverse-U-shaped curves would be replaced by U-shaped curves.

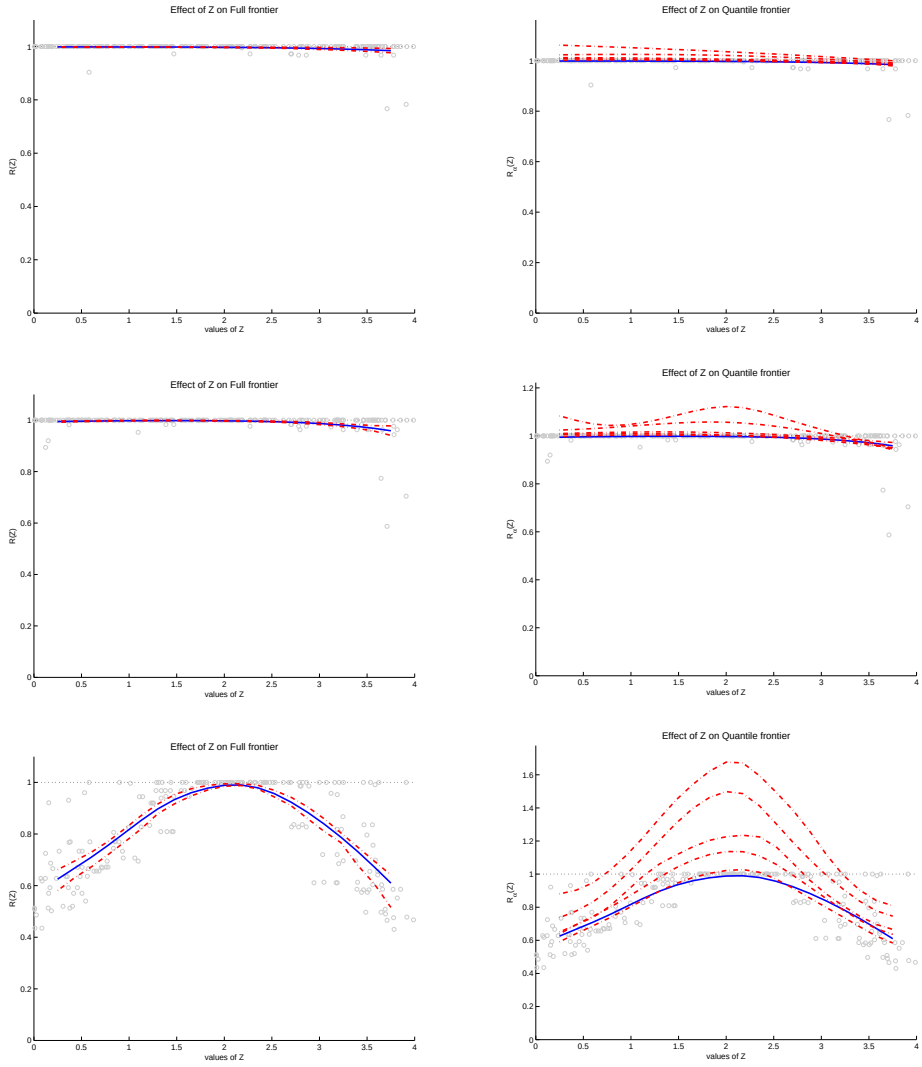

Figure 3: *Regressions of the ratios $R(Z)$ on Z. From top to bottom: DGP1, DGP2 and DGP3. Left panels, full frontier $\hat{\tau}_n^z$ with 95% confidence intervals for $\tau^z(P)$. Right panels, $\hat{\tau}_{\alpha,n}^z$ for $\alpha = (0.5, 0.75, 0.90, 0.95, 0.99, 1.00)$, the last one (full frontier case) in solid line. Here $n = 200$ and the circles are the estimated data points $(Z_i, \widehat{R}(Z_i))$.*

4.2 Efficiency in the Banking Sector

Simar and Wilson (2007) includes an empirical example based on Aly et al. (1990) using data on 6.955 US Commercial Banks observed at the end of the 4th quarter, 2002.[8] They run a truncated regression on the trCoutput oriented DEA estimates of efficiency in a second

[8]We would like to thank Paul W. Wilson who provided us this data set.

stage (as suggested in Aly et al., 1990). Daraio et al. (2010) used the same data set to test the "separability" condition which was rejected at any reasonable level, indicating that any two-stage procedure is meaningless for this dataset. This was a global test; we will rather here proceed to a local analysis.

The original data set contains 3 inputs (purchased funds, core deposits and labor) and 4 outputs (consumer loans, business loans, real estate loans, and securities held) for banks. Aly et al.1990 considered 2 continuous environmental factors, the size of the banks Z_1, and a measure of the diversity of the services proposed by the banks Z_2 (see Aly. et al., 1990, for details) and one binary variable indicating if the banks belong or not to a Metropolitan Statistical Area. We will use, as in Simar and Wilson (2007), a measure of the size of the banks by the log of the total assets, rather than the total deposit as in Aly et al. For simplifying the presentation, we will illustrate our procedure with a subsample of 322 Banks (also used in Simar and Wilson, 2007).

Some prior exploratory data analysis indicates that the 3 inputs are highly correlated among themselves and the same is true for the 4 outputs. So, due the dimensionality of the problem (3 inputs, 4 outputs, and 3 environmental factors) with the limited sample used here (322 units), we first reduce the dimension in the input $\times$ output space by using the methodology suggested in Daraio and Simar (2007a).

Since the radial measures are scale invariant, we divide each inputs and outputs by their mean (to be "unit" free) and replace the 3 scaled inputs by their best (non-centered) linear combination (we use here a kind of non-centered PCA, as explained in details in Daraio and Simar, 2007a), and we check that we did not loose much information by doing so, and that the resulting univariate input factor is highly correlated with the 3 original inputs. We follow the same procedure with the 4 outputs. The results are

$$\begin{aligned} IF &= 0.5707X_1 + 0.5731X_2 + 0.5881X_3, \\ OF &= 0.4851Y_1 + 0.4875Y_2 + 0.5095Y_3 + 0.5172Y_4, \end{aligned}$$

indicating that both the input and the output factor are a kind of average of the scaled inputs and outputs respectively (the weights are equal). We obtain the following correlations $\hat{\rho}_{IF,X_j} = (0.972,\ 0.971,\ 0.996)$ for $j = 1, 2, 3$ and IF explains 96% of total inertia of the original data (X_1, X_2, X_3). We obtain similar results when reducing the dimension in the output space: $\hat{\rho}_{OF,Y_j} = (0.924,\ 0.938,\ 0.975,\ 0.990)$ for $j = 1, \ldots, 4$, and OF explains 92% of total inertia of the original data $(Y_1, \ldots, Y_4)$. Hence we can conclude that we do not loose much information by this dimension reduction and the factors IF and OF are good representatives of the input and output activities of the Banks.

Remember that with the full data set and with all the original variables, Daraio et al. (2010) rejected the null hypothesis of global separability. We will here illustrate in our

simplified version of the examples what we can learn by the methodology we proposed in this paper. Figure 4 below illustrates the marginal analysis of the impact of the SIZE variable on the production process. The question of the separability condition for $Z_1 = \text{SIZE}$ does not appear clearly on the left panel of the figure. We observe a slight positive effect on the boundary of the attainable set. Larger size banks allow to reduce the inputs, for a given level of the outputs more than smaller banks (the last decrease on the right is spurious and is due to the isolated big units having their FDH scores equal to 1; automatically their conditional FDH is 1 and by construction, so are the ratios corresponding to these units). However, in the same time, the pointwise confidence intervals at each grid point cover in most of the case the value 1. Since for several values of z, the expected ratio is really (significantly) bigger than one, we can confirm that globally a test of separability would reject the null hypothesis (as was found for a global test in Daraio et al., 2010, in a similar setup, but with a larger data set).

The right panel of Figure 4 gives a more clear picture. Looking to the α-quantiles regressions here (from bottom to top going from $\alpha = 0.5$ till 0.99), we learn a lot. Even the regression corresponding to the 99%-quantile frontier is quite different from the regression with the full ratios. The general shape of the quantiles curves is more stable than the limit (full frontier) one. So, it seems that the full frontier analysis is perturbed by some extreme data points, and that the effect of Z_1 may be masked by some outlying points. The analysis of the regressions with more robust quantiles frontiers show a very regular positive slope, confirming again the positive effect of the SIZE on the process, as extensively explained in Daraio and Simar (2005, 2007a). Since they are parallel, it seems that the effect is additive with respect to the different quantiles of the distribution of the inefficiencies, it is only the boundary that is shifted (remember panel IV of Figure 1 above).

The above analysis illustrates how useful it is to look simultaneously to the full frontier results but also to the α-quantile results with a grid of values for α.

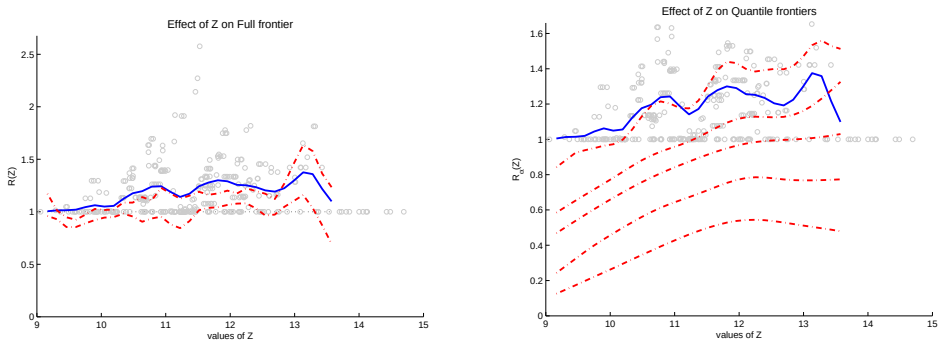

Figure 4: *Marginal effect of* Z_1 = SIZE *on the production. Left panel, full frontier* $\hat{\tau}_n^z$ *with 95% confidence intervals for* $\tau^z(P)$. *Right panels,* $\hat{\tau}_{\alpha,n}^z$ *for* $\alpha = (0.5, 0.75, 0.90, 0.95, 0.99, 1.00)$, *the last one (full frontier case) in solid line. Here* $n = 322$ *and the circles are the estimated data points* $(Z_i, \widehat{R}(Z_i))$.

We do the same univariate exercise to investigate the marginal effect of the variable Z_2 (DIVERSE: a measure of the diversity of the products of the Banks). Figure 5 displays the results. We do not see a clear effect of Z_2 on the boundary, except for small values of z (less diverse Banks should be able to reduce more their inputs given the level of their outputs). This is confirmed by the right panel but the distribution of the inefficiencies seems to be homoscedastic relative to Z_2 (the different quantile results are similar and rather stable with respect to z_2). Note that the downward curvature at the extreme left part of the picture (for the full frontier case and for the 99%-quantile case) are due to the edge effect mentioned before.

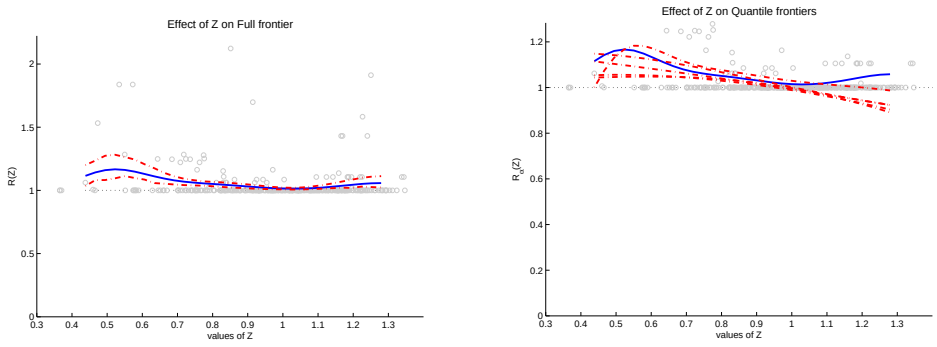

Figure 5: *Marginal effect of* Z_2 = DIVERSE *on the production. Left panel, full frontier* $\hat{\tau}_n^z$ *with 95% confidence intervals for* $\tau^z(P)$. *Right panels,* $\hat{\tau}_{\alpha,n}^z$ *for* $\alpha = (0.5, 0.75, 0.90, 0.95, 0.99, 1.00)$, *the last one (full frontier case) in solid line. Here* $n = 322$ *and the circles are the estimated data points* $(Z_i, \widehat{R}(Z_i))$.

We can now illustrate how the effect of Z can be displayed for bivariate cases. Here we estimate from the start the conditional survival function by conditioning on both $Z_1 =$ SIZE and $Z_2 =$ DIVERSE. We first do the exercise for the full sample of 322 units. To save place we only show the results for the order-α quantile frontier, to be less sensitive to extreme points (as noticed above). Figure 6 shows on the left panel the estimate of the regression of $\widehat{R}(Z_i)$ on Z_i evaluated on a grid of values for Z_1 and Z_2. We see the increasing slope relative to Z_1 and the flat average effect of Z_2. We do not observe a clear interaction effect between Z_1 and Z_2, although the effect of Z_1 seems bigger for middle values of Z_2. The right panel indicates the resulting marginal effects obtained from the preceding surface regression, evaluated at the observations (X_i, Y_i, Z_i) and then viewed marginally as a function of each component Z_1 and Z_2 separately. The marginal effects confirm what has been seen in the "pure" marginal analysis above.

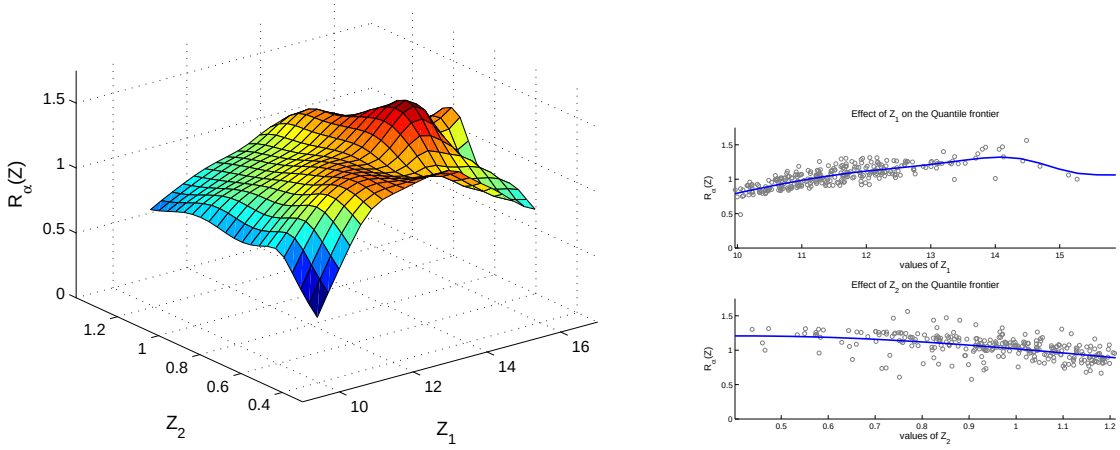

Figure 6: *Joint effect of* (Z_1, Z_2) *on the production process. We use here* $\hat{\tau}^z_{\alpha,n}$ *for* $\alpha = 0.95$. *Here* $n = 322$ *and the circles are the estimated data points* $(Z_i, \widehat{R}_\alpha(Z_i))$.

Of course for this bivariate analyis we can also compute at selected grid points (z_1, z_2) confidence intervals for $\tau^z(P)$ and $\tau^z_\alpha(P)$. We illustrate this in Table 1 where we select for z_ℓ the 3 quartiles of Z_ℓ, $\ell = 1, 2$ giving the 9 selected pairs for (z_1, z_2). The resulting confidence intervals shown in the table confirm the analysis done above. We note also (as in Figure 4) that the estimates of the expected ratios with the full frontier are sometimes outside the 95% confidence intervals, showing that the point estimates are in some cases biased (the basic bootstrap techniques for confidence intervals, automatically correct for the bias). This is less apparent for partial frontier, as expected, partial efficiency scores having an asymptotic normal distribution and not a Weibull type one.

Finally, we want to investigate if the effects of Z_1 and Z_2 are similar across the two groups of banks $MSA = 1$ (174 banks are in a Metropolitan Statistical Area) and $MSA = 0$ (148 units are not belonging to a MSA). Since the sample size is large in both groups, we can do two separate analysis. The results are displayed in Figure 7.

It seems that the marginal effects have approximately the same shape and same size in the two groups, but we detect a slight difference of the interaction between the two variables. In the group MSA, we have the highest ratios for large Z_1 and small Z_2 whereas for the not MSA banks the size effect Z_1 is larger for higher values of Z_2. This slight interaction was hidden in the global picture with all the observations in Figure 6.

z_1	z_2	$\hat{\tau}_n^z$	low	up	$\hat{\tau}_{\alpha,n}^z$	low	up
10.6699	0.8514	1.4368	1.2663	1.4254	1.0175	0.9844	1.0312
10.6699	0.9998	1.3748	1.2069	1.3499	0.9731	0.9533	0.9901
10.6699	1.1391	1.2080	1.0448	1.1754	0.8809	0.8569	0.8884
11.3696	0.8514	1.4569	1.2524	1.4324	1.1098	1.0801	1.1234
11.3696	0.9998	1.3551	1.1839	1.3417	1.0338	1.0124	1.0437
11.3696	1.1391	1.2042	1.0131	1.1274	0.9707	0.9398	0.9812
12.1351	0.8514	1.3016	1.0528	1.1561	1.1874	1.1269	1.2250
12.1351	0.9998	1.2391	1.0395	1.1210	1.1023	1.0517	1.1232
12.1351	1.1391	1.1602	0.9700	1.0729	1.0419	0.9939	1.0600

Table 1: *Point estimates and 95% confidence intervals for* $\tau^z(P)$ *and* $\tau_\alpha^z(P)$*, with* $\alpha = 0.95$ *at selected grid points* (z_1, z_2)*.*

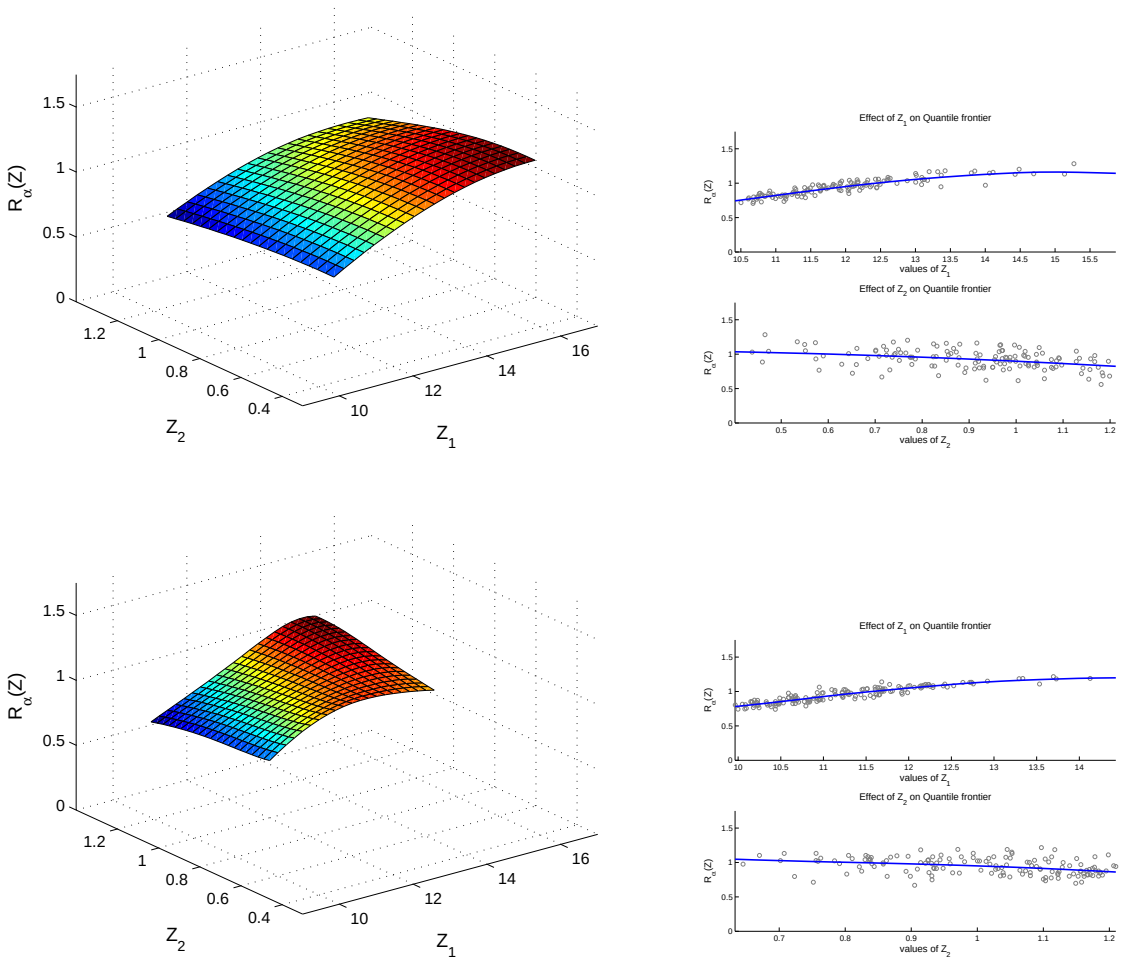

Figure 7: *Joint effect of* (Z_1, Z_2) *on the production process for the two groups. Top panels for 174 Banks in MSA and bottom panels for 148 Banks not in a MSA. We use here* $\hat{\tau}_{\alpha,n}^z$ *for* $\alpha = 0.95$*.*

5 Conclusions

This paper has formalized in a nonparametric model of production the role of environmental variables by introducing these external factors in a non-restrictive way.
The paper clarifies what can be learned by analyzing the conditional efficiency measures and proposes a general approach to measure and infer about the impact of these factors on the production process.
By using conditional efficiency measures we can indeed measure the impact of external factors on the attainable set in the input-output space, and/or we can investigate the impact of the external factors on the distribution of inefficiency scores.
The paper proposes a statistical approach to make inference on the level of the impact by using up-to dated bootstrap algorithms for which we prove the consistency. In the paper we have provided practical information to implement the bootstrap and have shown its general and wide usefulness for empirical applications by illustrating its functioning by means of several simulated examples and a real dataset on US commercial banks.

A Appendix

A.1 Asymptotic Properties of $\hat{\tau}_n^z$

The argument is the one developed in the Appendix of Simar and Wilson (2010b) where OLS is used on DEA efficiency scores. We adapt it to the usual standard setup of nonparametric regression (see e.g. Pagan and Ullah, 1999). We must first obtain the stochastic properties of the random variables $\widehat{R}(X_i, Y_i|Z_i)$ which play the role of the dependent variable in the regression. Here we summarize the arguments used in Section 4.2 of Daraio et al. (2010). To simplify the notation we will define the n unobserved iid ratios $R_i = R(X_i, Y_i|Z_i)$ and the n available estimators $\widehat{R}_i = \widehat{R}(X_i, Y_i|Z_i)$.

Under mild regularity conditions, we have for a generic observation (X_i, Y_i), as $n \to \infty$

$$n^{\kappa}\big(\widehat{R}_i - R_i\big) \xrightarrow{\mathcal{L}} Q_P^{Z_i}(\cdot), \tag{A.1}$$

where $Q_P^{Z_i}(\cdot)$ is a nondegenerate distribution (i.e. it is not a Dirac distribution with mass 1 at one single value) with finite mean $\mu_Q^{Z_i}$ and variance $\sigma_Q^{2,Z_i} > 0$. The rate of convergence is governed by the worst rate of convergence of the estimators used to estimate the λ's. In our case here, it is the rate of the conditional measure, so $\kappa = 4/((r+4)(p+q))$ (see Section 3).

Since $\widehat{R}_i = R_i + n^{-\kappa}\zeta_i$, where ζ_i is implicitly defined by the equation, ζ_i must have limiting distribution $Q_P^{Z_i}(\cdot)$. So, we have for large n

$$\mathbb{E}(\widehat{R}_i) = \mathbb{E}(\widehat{R}_i) + \mu_Q^{Z_i}/n^{\kappa} \tag{A.2}$$

$$\mathbb{V}(\widehat{R}_i) = \mathbb{V}(\widehat{R}_i) + O(n^{-\kappa}), \tag{A.3}$$

where the second term in (A.3) accounts for the variance of $n^{-\kappa}\zeta_i$ and the covariance between R_i and $n^{-\kappa}\zeta_i$ is indeed $O(n^{-\kappa})$.

Of course there is some dependence between the $\widehat{R}_i$ and so between the $n^{-\kappa}\zeta_i$. Following Daraio et al. (2010), and based on the local nature of the envelopment estimators (for the FDH case, asymptotically, the estimator at (X_i, Y_i) depends only on at most one other data point), all the covariances between $n^{-\kappa}\zeta_i$ and $n^{-\kappa}\zeta_j$ for $j \neq i$ are asymptotically equal to zeros, except for at most 2 values of $j \neq i$. It is also clear that when nonzero, this covariance is bounded by the product of the two standard deviations, i.e. $O(n^{2\kappa})$.

Now, as defined in (3.7), the nonparametric regression estimator is given by

$$\begin{aligned}
\hat{\tau}_n^z &= \sum_{i=1}^{n} W_n(Z_i, z, h_z)\widehat{R}_i \\
&= \sum_{i=1}^{n} W_n(Z_i, z, h_z)R_i + n^{-\kappa}\sum_{i=1}^{n} W_n(Z_i, z, h_z)\zeta_i \\
&= \hat{\tau}^z(P) + \zeta^z. \qquad \text{(A.4)}
\end{aligned}$$

The first term of the last equation $\hat{\tau}^z(P)$ is the standard nonparametric regression estimator of $\tau^z(P)$ one would obtain by observing the iid true values R_i; the second term ζ^z accounts for the bias and the dependence introduced by replacing R_i by its estimates $\widehat{R}_i$.

Now we analyze the asymptotic behavior of ζ^z. Again, to simplify the notation we particularize to the case of the Nadaraya-Watson estimator and for one dimensional z ($r = 1$) (the same idea can be used for the local linear estimator case, but at a cost of notational complexity, see Fan and Gijbels, 1996). We can now summarize the main steps. First we can write

$$\zeta^z = \frac{1}{\hat{f}_Z(z)}\,\frac{1}{nh_z}\sum_{i=1}^{n} K_i\frac{\zeta_i}{n^{\kappa}}$$

where $\hat{f}_Z(z)\xrightarrow{p} f_Z(z)$ and $K_i = K((Z_i - z)/h_z)$. Due to the first two moments of $n^{-\kappa}\zeta_j$ summarized above (in particular the bounded number of nonzero covariances) it is easy to show (see details e.g. in Lemma 3.1 of Pagan and Ullah, 1999) that

$$\mathbb{E}\left(\frac{1}{nh_z}\sum_{i=1}^{n} K_i\frac{\zeta_i}{n^{\kappa}}\right) = f_Z(z)\frac{\mu^z_{Q_P} + h_z^2\, C^z}{n^{\kappa}} + o(n^{-\kappa}h_z^2) \qquad \text{(A.5)}$$

$$\mathbb{V}\left(\frac{1}{nh_z}\sum_{i=1}^{n} K_i\frac{\zeta_i}{n^{\kappa}}\right) \leq \frac{f_Z(z)D^z}{nh_z\, n^{2\kappa}} + O\big(n^{-(1+2\kappa)}\big) \qquad \text{(A.6)}$$

where C^z and D^z are bounded constants. Due to this, one can see that

$$\sqrt{nh_z}\left(\zeta^z - \frac{\mu^z_{Q_P} + h_z^2\, C^z}{n^{\kappa}}\right)\xrightarrow{p} 0.$$

The multivariate extension is given by

$$\sqrt{nh_z^r}\left(\zeta^z - \frac{\mu^z_{Q_P} + h_z^2\, C^z}{n^\kappa}\right) \xrightarrow{p} 0. \tag{A.7}$$

Therefore, since from (A.4)

$$\hat{\tau}^z_n - \frac{\mu^z_{Q_P} + h_z^2\, C^z}{n^\kappa} = \hat{\tau}^z(P) + \zeta^z - \frac{\mu^z_{Q_P} + h_z^2\, C^z}{n^\kappa},$$

we obtain as $n \to \infty$

$$\sqrt{nh_z^r}\Big(\hat{\tau}^z(P) - \tau^z(P) - h_z^2 B^z\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, V^z\big), \tag{A.8}$$

where the bias B^z and the variance V^z are bounded constants that can be found in any textbook on nonparametric econometrics. They depend on the particular estimator used (Nadaraya-Watson, Local linear,etc.) and on characteristics of the Kernel and of the DGP; see e.g. Pagan and Ullah (1999) or Li and Racine (2007) for details and comments.

Now, from (A.7) we obtain as $n \to \infty$, $h_z \to 0$ with $nh_z^r \to \infty$:

$$\sqrt{nh_z^r}\Big(\hat{\tau}^z_n - \tau^z(P) - n^{-\kappa}\mu^z_{Q_P} - h_z^2(B^z + n^{-\kappa}C^z)\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, V^z\big). \tag{A.9}$$

A.2 Order-m partial frontiers

An alternative partial frontier has been introduced by Cazals et al. (2002): the order-m frontiers. Roughly speaking, in the output orientation case, the idea is to take as benchmark for evaluating firms, the expectation of the best practice among m peers drawn at random in the population of firms using less resources than x_0. Specifically, consider m i.i.d. random variables Y_i, $i = 1, \ldots, m$ generated according the survival $S_{Y|X}(y|X \leq x_0)$ and we define the random set $\Psi_m(x_0) = \{(x', y) \in \mathbb{R}_+^{p+q} | x' \leq x_0, y \leq Y_i, i = 1, \ldots, m\}$. Then, we can define

$$\begin{aligned}\tilde{\lambda}_m(x_0, y_0) &= \sup\{\lambda > 0 | (x_0, \lambda y) \in \Psi_m(x_0)\} \\ &= \max_{i=1,\ldots,m}\left\{\min_{j=1,\ldots,q} \frac{Y_i^j}{y_0^j}\right\}.\end{aligned}$$

This is the maximal output radial expansion ($\leq$ of ≥ 1) for (x_0, y_0) to reach the FDH of the random set of firms (x_0, Y_i), $i = 1, \ldots, m$. Finally, the order-m output efficiency score is given by the conditional expectation of $\tilde{\lambda}_m(x_0, y_0)$:

$$\lambda_m(x_0, y_0) = \mathbb{E}\big(\tilde{\lambda}_m(x_0, y_0) | X \leq x_0\big). \tag{A.10}$$

It is easy to see that if $m \to \infty$, $\lambda_m(x_0, y_0) \to \lambda(x_0, y_0)$. See Daraio and Simar (2007a) for details. Since the benchmark is against an average of the best among m peers, the

corresponding frontier (the set of points (x, y) where $\lambda_m(x, y) = 1$) is less extreme. For instance if $m = 1$, the m-frontier represent an average production frontier among producers using less resources than the current value x_0. It has been shown in Cazals et al. (2002) that if $\lambda_m(x_0, y_0)$ exists, it can be computed by the following univariate integral

$$\lambda_m(x_0, y_0) = \int_0^\infty \left[1 - \left(1 - S_{Y|X}(uy_0|X \leq x_0)\right)^m\right] du. \tag{A.11}$$

When facing environmental conditions $Z = z_0$, we can define the conditional order -m measures by conditioning every random event to $Z = z_0$. As described in Daraio and Simar (2007a), thus leads to the expression

$$\lambda_m(x_0, y_0|z_0) = \int_0^\infty \left[1 - \left(1 - S_{Y|X,Z}(uy_0|X \leq x_0, Z = z_0)\right)^m\right] du, \tag{A.12}$$

leading, in our purpose, to the ratios

$$R_m(x, y|z) = \frac{\lambda_m(x, y|z)}{\lambda_m(x, y)}. \tag{A.13}$$

The parameter of interest will be here

$$\tau_m^z(P) = \mathbb{E}(R_m(X, Y|Z = z)), \tag{A.14}$$

where again, if $m \to \infty$, $\tau_m^z(P) \to \tau^z(P)$.

What has been said above about $\tau_\alpha^z(P)$, remains valid for $\tau_m^z(P)$: the parameter will mainly capture the local effect of Z on the distribution of the inefficiencies when the boundary is not changing ($\Psi^z = \Psi$). But it does not allow, when considered alone, to capture a shift of the boundary. Unless m increases to infinity and we search a robust estimator of the full frontier (see Section 3).

References

[1] Banker, R.D. and R. Natarajan (2008),Evaluating Contextual Variables Affecting Productivity Using Data Envelopment Analysis, *Operations Research*, 56(1), 48–58.

[2] Bădin, L., Daraio, C. and L. Simar (2010), Optimal Bandwidth Selection for Conditional Efficiency Measures: a Data-driven Approach, *European Journal of Operational Research*, 201, 2, 633–640.

[3] Bădin, L., Daraio, C. (2010), Explaining Efficiency with Nonparametric Frontier Models. Recent developments in statistical inference, in *Exploring research frontiers in contemporary statistics and econometrics*, ed. by I. Van Keilegom and P.W. Wilson, Springer, Berlin (in press).

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

[5] Daouia, A. and I. Gijbels (2009), Robustness and inference in nonparametric partial-frontier modeling, manuscript.

[6] Daouia, A. and I. Gijbels (2010), Estimating frontier cost models using extremiles, in *Exploring research frontiers in contemporary statistics and econometrics*, ed. by I. Van Keilegom and P.W. Wilson, Springer, Berlin (in press).

[7] Daouia, A. and L. Simar (2007), Nonparametric Efficiency Analysis: A Multivariate Conditional Quantile Approach, *Journal of Econometrics*, 140, 375–400.

[8] Daraio, C. and L. Simar (2005), Introducing Environmental Variables in Nonparametric Frontier Models: a Probabilistic Approach, *Journal of Productivity Analysis*, 24, 93–121.

[9] Daraio, C. and L. Simar (2006), A robust nonparametric approach to evaluate and explain the performance of mutual funds, *European Journal of Operational Research*, Vol 175 (1), 516–542.

[10] Daraio, C. and L. Simar (2007a), *Advanced Robust and Nonparametric Methods in Efficiency Analysis. Methodology and applications*, Springer, New York.

[11] Daraio, C. and L. Simar (2007b), Conditional nonparametric Frontier models for convex and non convex technologies: A unifying approach, *Journal of Productivity Analysis*, 28, 13–32.

[12] Daraio, C., Simar, L. and P. Wilson (2010), Testing whether two-stage estimation is meaningful in nonparametric models of production, Discussion Paper #1030, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

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

[14] Deprins, D., Simar, L. and H. Tulkens (1984), Measuring labor-efficiency in post offices, in Marchand, M., Pestieau, P. and Tulkens, H. (eds.) *The Performance of public enterprises - Concepts and Measurement*, Amsterdam, North-Holland, 243–267.

[15] Fan J. and I. Gijbels (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall.

[16] Farrell, M.J. (1957), The measurement of the Productive Efficiency, *Journal of the Royal Statistical Society*, Series A, CXX, Part 3, 253–290.

[17] Färe, R., Grosskopf, S. and C.A.K. Lovell (1985), *The Measurement of Efficiency of Production*. Boston, Kluwer-Nijhoff Publishing.

[18] Jeong, S.O. , B. U. Park and L. Simar (2010), Nonparametric conditional efficiency measures: asymptotic properties. *Annals of Operations Research*, 173, 105–122.

[19] Kneip, A, L. Simar and P.W. Wilson (2008), Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models, *Econometric Theory*, 24, 1663–1697.

[20] Kneip, A., Simar, L. and P.W. Wilson (2010), A Computational Efficient, Consistent Bootstrap for Inference with Non-parametric DEA Estimators. Discussion paper 0903, Institut de Statistique, UCL, in press *Computational Economics*.

[21] Li, Q. and J. Racine (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

[22] Pagan, A. and A. Ullah (1999), *Nonparametric Econometrics*, Cambridge University Press.

[23] Park, B., Simar, L. and C. Weiner (2000), The FDH estimator for productivity efficiency scores: asymptotic properties, *Econometric Theory* 16, 855-877.

[24] Racine, J. and Q. Li (2004), Nonparametric estimation of regression functions with both categorical and continuous data, *Journal of Econometrics*, 119, 99–130.

[25] Shephard, R.W. (1970). *Theory of Cost and Production Function*. Princeton University Press, Princeton, New-Jersey.

[26] Simar, L. and P.W. Wilson (2007), Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes, *Journal of Econometrics*, Vol 136, 1, 31–64.

[27] Simar, L. and P.W. Wilson (2008), Statistical Inference in Nonparametric Frontier Models: recent Developments and Perspectives, in *The Measurement of Productive Efficiency*, 2nd Edition, Harold Fried, C.A.Knox Lovell and Shelton Schmidt, (Eds), Oxford University Press.

[28] Simar, L. and P.W. Wilson (2010a), Inference by the m out of n bootstrap in Nonparametric Frontier Models, in press, *Journal of Productivity Analysis*.

[29] Simar, L. and P.W. Wilson (2010b), Two-Stage DEA: *Caveat Emptor*, Discussion Paper #1041, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.