Laboratory of Economics and Management
Sant'Anna School of Advanced Studies

Piazza Martiri della Libertà, 33 - 56127 PISA (Italy)
Tel. +39-050-883-343 Fax +39-050-883-344
Email: lem@sssup.it Web Page: http://www.lem.sssup.it/

LEM
Working Paper Series

Conditional Nonparametric Frontier Models for Convex and Non Convex Technologies: a Unifying Approach

Cinzia DARAIO*
Leopold SIMAR[†]

* IIT-CNR, and Sant'Anna School of Advanced Studies, Pisa, Italy
[†] Université Catholique de Louvain, Belgium

2005/12 **May 2005**

ISSN (online) 2284-0400

Conditional Nonparametric Frontier Models for Convex and Non Convex Technologies: a Unifying Approach*

Cinzia Daraio†
IIT-CNR and Scuola Superiore S. Anna, Italy
cinzia@sssup.it , cinzia.daraio@iit.cnr.it

Léopold Simar‡
Université Catholique de Louvain, Belgium
simar@stat.ucl.ac.be

5 May, 2005

Abstract

The explanation of productivity differentials is very important to identify the economic conditions that create inefficiency and to improve managerial performance. In literature two main approaches have been developed: one-stage approaches and two-stage approaches. Daraio and Simar (2003) propose a full nonparametric methodology based on *conditional* FDH and *conditional* order-m frontiers without any convexity assumption on the technology. On the one hand, convexity has always been assumed in mainstream production theory and general equilibrium. On the other hand, in many empirical applications, the convexity assumption can be reasonable and sometimes natural. Leading by these considerations, in this paper we propose a unifying approach to introduce external-environmental variables in nonparametric frontier models for convex and non convex technologies. Developing further the work done in Daraio and Simar (2003) we introduce a conditional DEA estimator, *i.e.*, an estimator of production frontier of DEA type conditioned to some external-environmental variables which are neither inputs nor outputs under the control of the producer. A robust version of this conditional estimator is also proposed. These various measures of efficiency provide also indicators of convexity. Illustrations through simulated and real data (mutual funds) examples are reported.

*This paper has been prepared within the AQuaMethPSR Project under the PRIME Network of Excellence supported by the European Commission, 6th Framework Programme. Previous versions of this paper were presented at the 4^{th} International DEA Symposium, Birmingham, 4-7 September 2004 and at the First Italian Congress in Econometrics and Empirical Economics, 24-25 January 2005. We would like to thank conferences participants for helpful comments. The usual disclaimers apply.

†Work partially supported by the Italian Registry of ccTLD .it.

‡Research support from the "Interuniversity Attraction Pole", Phase V (No. P5/24) from the Belgian Government (Belgian Science Policy) is acknowledged.

Keywords: convexity, external-environmental factors, production frontier, nonparametric estimation, robust estimation.
JEL Classification: C13, C14, D20.

1 Introduction

Efficiency and productivity literature primarily focused on the measurement of decision making units (DMUs)' performance.

In recent decades there has been a growing interest for the logical step ahead: the explanation of DMUs productivity differentials. As a matter of fact, the impact of external-environmental factors on the efficiency of producers is 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. These factors are neither inputs nor outputs under the control of the producer, but can affect the performance of the production process. In literature, two main approaches have been developed.

In the "one-stage" approach the environmental variables are directly included in the linear programming formulation along with the inputs and outputs. In the "two-stage" approach the technical efficiency, computed in a standard way, is used as dependent variable in a second-stage regression. Some authors propose also three-stage and four-stage analysis as extension of the two-stage approach[1].

The main disadvantage of the one-stage approach is that it requires the classification of environmental factor as an input or an output *prior* to the analysis. The main shortcoming of the two-stage approach, as pointed out in Simar and Wilson (2003), is that the efficiency estimates are serially correlated in a complicated way and that the first stage efficiency scores are biased. Hence, they propose a procedure based on bootstrap techniques to permit a more accurate inference in the second-stage. Note that all these two stage approaches have an additional drawback: they rely on a separability condition between the input-output space and the space of environmental variables. In addition, in all the studies published so far, a restrictive parametric model is used for the second-stage regression.

Daraio and Simar (2003), hereafter DS, propose a full nonparametric approach which overcomes most of the drawbacks mentioned above. They define *conditional* (to external-environmental factors) frontiers and *conditional* order-m frontiers together with their related efficiency scores and the corresponding nonparametric estimators. In particular, order-m frontier estimators (Cazals, Florens and Simar, 2002, hereafter CFS) are known as being more robust to outliers and extreme values than the full frontier estimates.

[1]See Daraio and Simar (2003), and the references cited there.

In this paper we provide a unifying approach to introduce external environmental variables in nonparametric models of production frontiers. Completing the work done in DS we introduce a conditional Data Envelopment Analysis (DEA) estimator, *i.e.*, a DEA estimator of production frontiers conditioned to some external-environmental variables that are neither inputs nor outputs under the control of the producer. In order to control for the influence of extremes or outliers we introduce also a robust version of our conditional DEA estimator, based on the concept of order$-m$ frontiers. The motivation for this paper is threefold.

Firstly, convexity has always been an usual assumption on the production set structure, very often used by economists and practitioners. DEA, in fact, is the most popular nonparametric estimator in empirical applications[2], and its convexity assumption on the production set is widely used in mainstream theories of production and general equilibrium (see e.g. Mas-Colell, Whinston and Green, 1995). Several recent studies focus on the convexity assumption in frontier models (*e.g.*, Bogetoft, 1995; Bogetoft, Tama, and Tind, 2000; Briec, Kerstens and Vanden Eeckaut, 2004; Podinovski, 2004).

Secondly, in some fields of application, allowing for the convexity of the production possibility set is *natural* given the characteristics of the underlying technology. Consider, for instance, the industry of mutual funds. A mutual fund is managed by an economic operator which selects a set of bonds/stocks according to an investment objective or a mix of investment goals, focusing on the return, or the risk of the portfolio or on a balance among these two. Owing these features of their management process, it seems quite normal to allow for the feasibility of some portfolios that are linear combinations of actually observed funds. In this framework, the assumption that "the mean of any two combinations that can be produced can itself be produced (Farrell, 1959, p. 377)" seems quite natural.

Hence, in this paper we aim at enriching the toolbox of applied researchers in productivity analysis offering a complete range of conditional measures of efficiency, *i.e.*, measures of performance which take into account the operating environment (or other external factors) in which firms operate in, without imposing their positive or negative impact, but letting the data themselves to tell if and how they affect the performance.

Therefore, the *conditional* DEA estimator, as well as its robust version, is useful to explain efficiency differentials when the convexity hypothesis is reasonable for the technology analyzed.

Thirdly, we lay down the ground for the development of a statistical test of convexity, which could offer a *rigorous* way to choose among a set of efficiency measures (convex and not convex) those ones appropriate to explain efficiency differentials in the empirical context analyzed.

[2]See Cooper, Seiford and Tone (1999) for about 15,000 references of DEA applications.

The paper is structured as follows. In Section 2 we describe the frontier estimation setting and we propose, extending DS, a unifying formalization of the production process based on a probabilistic approach, where the FDH and DEA estimators can be naturally introduced. Section 3 presents the concept of order-m frontiers, as based on CFS and DS ideas, and analyzes how convexity can be introduced in these partial frontiers. This leads to define efficiency scores of order-m with respect to convex technologies. Nonparametric estimators are then described and some of their properties are investigated. Section 4 shows how the probabilistic formulation allows to introduce conditional efficiency measures and, extending DS, defines a *conditional* DEA efficiency score and its robust (order-m) version. In Section 5 we propose a series of indicators of the type of those proposed in Briec, Kerstens and Vanden Eeckaut (2004), extending its application to robust order-m efficiency measures and to conditional and robust measures of performance. Section 6 illustrates the different concepts trough some simulated data sets as well as real data on mutual funds. Section 7 concludes, outlining future development to address. In the Appendix we address some issues about the bandwidth selection procedure necessary for estimating most of the conditional measures.

2 Formalizing the Production Process

2.1 The activity analysis framework

In an activity analysis framework (Koopmans, 1951; Debreu, 1951) the activity of production units (or a production technology) 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:

$$\Psi = \{(x, y) \in I\!R_+^{p+q} \mid x \text{ can produce } y\}. \tag{2.1}$$

Usually, the free disposability of inputs and outputs is assumed, meaning that if $(x, y) \in \Psi$, then $(x', y') \in \Psi$, as soon as[3] $x' \geq x$ and $y' \leq y$.

The boundaries of Ψ becomes of interest when we want to estimate efficiency. If we are looking in the input direction, the Farrell measure of input-oriented efficiency score[4] for a unit operating at the level (x, y) is defined as:

$$\theta(x, y) = \inf\{\theta \mid (\theta x, y) \in \Psi\}. \tag{2.2}$$

[3]From here and below inequalities between vectors are element-wise.

[4]Here and below we consider only the input oriented framework to save place. The extension to the output oriented framework is straightforward (see DS and Daraio, 2003 for details).

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 can then be described as the set $(x^{\partial}(y), y) \in \Psi$, where $x^{\partial}(y) = \theta(x, y)x$ is the radial projection of $(x, y) \in \Psi$ on the frontier, in the input direction (orthogonal to the vector y).

In empirical applications, the set Ψ is unknown as well as 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.

Envelopment estimators (Data Envelopment Analysis (DEA): Charnes, Cooper and Rhodes, 1978/ Free Disposal Hull (FDH): Deprins, Simar and Tulkens, 1984) within the nonparametric approach are particularly appealing since they do not rely on restrictive hypothesis on the Data Generating Process (DGP).

In this framework, an observed production unit, (x_i, y_i), defines an individual production possibilities set $\psi(x_i, y_i)$, which under the free disposability of inputs and outputs, can be written as:

$$\psi(x_i, y_i) = \{(x, y) \in I\!R_+^{p+q} |\ x \geq x_i, y \leq y_i\} \tag{2.3}$$

The union of these individual production possibilities sets provides the FDH estimator of the whole production set Ψ:

$$\begin{aligned} \widehat{\Psi}_{FDH} &= \bigcup_{i=1}^{n} \psi(x_i, y_i) \qquad (2.4) \\ &= \{(x, y) \in I\!R_+^{p+q} |\ x \geq x_i, y \leq y_i,\ i = 1, ..., n\}. \end{aligned}$$

The DEA estimator[5] of the frontier of Ψ, $\widehat{\Psi}_{DEA}$, is obtained by the convex hull of $\widehat{\Psi}_{FDH}$:

$$\widehat{\Psi}_{DEA} = \mathcal{CH}\Big(\bigcup_{i=1}^{n} \psi(x_i, y_i)\Big) \tag{2.5}$$

$$= \Big\{(x, y) \in I\!R_+^{p+q} \mid y \leq \sum_{i=1}^{n} \gamma_i\, y_i\ ;\ x \geq \sum_{i=1}^{n} \gamma_i\, x_i, \tag{2.6}$$

$$\text{for } (\gamma_1, ..., \gamma_n) \quad \text{s.t.} \sum_{i=1}^{n} \gamma_i = 1; \gamma_i \geq 0, i = 1, ..., n\Big\}.$$

where $\mathcal{CH}$ stands for 'the convex hull of'. It is the smallest free disposal convex set covering all the data points.

The corresponding FDH and DEA estimators of efficiency scores are obtained by plugging $\widehat{\Psi}_{FDH}$ and $\widehat{\Psi}_{DEA}$, respectively, in equation (2.2) above.

[5]Note that here we consider only the Variable Returns to Scale case; the extension to different returns to scale situations is straightforward.

2.2 A probabilistic formulation of the production process

DS, generalizing results obtained in CFS, propose a probabilistic formulation of the production process in which it is easy to introduce external-environmental factors. The production process can indeed be described by the joint probability measure of (X, Y) on $\mathbb{R}_+^p \times \mathbb{R}_+^q$. This joint probability measure is completely characterized by the knowledge of the probability function $H_{XY}(\cdot, \cdot)$ defined as

$$H_{XY}(x, y) = \text{Prob}(X \leq x, Y \geq y). \tag{2.7}$$

The support of $H_{XY}(\cdot, \cdot)$ is Ψ and $H_{XY}(x, y)$ can be interpreted as the probability for a unit operating at the level (x, y) to be dominated. Note that this function is a non-standard distribution function, having a cumulative distribution form for X and a survival form for Y. In the input orientation chosen here, it is useful to decompose this joint probability as follows:

$$H_{XY}(x, y) \;=\; \text{Prob}(X \leq x \,|\, Y \geq y)\, \text{Prob}(Y \geq y) = F_{X|Y}(x|y)\, S_Y(y), \tag{2.8}$$

where we suppose the conditional probabilities exist (*i.e.*, $S_Y(y) > 0$). The conditional distribution $F_{X|Y}$ is non-standard due to the event describing the condition (*i.e.*,$Y \geq y$ instead of $Y = y$). We can now define the efficiency scores in terms of the support of these probabilities. The input oriented efficiency score $\theta(x, y)$ for $(x, y) \in \Psi$ is defined for all y with $S_Y(y) > 0$ as

$$\theta(x, y) = \inf\{\theta \,|\, F_{X|Y}(\theta x|y) > 0\} = \inf\{\theta \,|\, H_{XY}(\theta x, y) > 0\}. \tag{2.9}$$

The idea here is that the support of the conditional distribution $F_{X|Y}(\cdot \mid y)$ can be viewed as the attainable set of input values X for a unit working at the output level y. Under the free disposability assumption, the lower boundary of this support (in a radial sense) provides the Farrell-efficient frontier, or the input benchmarked value (see CFS and DS for details).

A nonparametric estimator is then easily obtained by replacing the unknown $F_{X|Y}(x \mid y)$ by its empirical version:

$$\widehat{F}_{X|Y,n}(x \mid y) = \frac{\sum_{i=1}^n \mathbb{1}(X_i \leq x, Y_i \geq y)}{\sum_{i=1}^n \mathbb{1}(Y_i \geq y)}, \tag{2.10}$$

where $\mathbb{1}(\cdot)$ is the indicator function.

As shown by CFS, the resulting estimator of the input efficiency score for a given point (x, y) coincides with the FDH estimator of $\theta(x, y)$:

$$\widehat{\theta}_{FDH}(x, y) \;=\; \inf\{\theta \mid (\theta\, x, y) \in \widehat{\Psi}_{FDH}\} \tag{2.11}$$

$$\;=\; \inf\{\theta \mid \widehat{F}_{X|Y,n}(\theta x \mid y) > 0\}. \tag{2.12}$$

We know that under the free disposal assumption, this is a consistent estimator of $\theta(x,y)$ with a rate of convergence of $n^{1/(p+q)}$ (see Park, Simar and Weiner, 2000).

Slightly faster is the rate of convergence of the DEA estimator (which relies on the additional convexity assumption of Ψ) that is of $n^{2/(p+q+1)}$ (see Kneip, Park and Simar, 1998). It is usually obtained by solving the linear program involved by:

$$\widehat{\theta}_{DEA}(x,y) = \inf\{\theta \mid (\theta\, x, y) \in \widehat{\Psi}_{DEA}\}. \tag{2.13}$$

where $\widehat{\Psi}_{DEA}$ was defined in (2.6).

For the extensions below, it is useful to notice that in the probabilistic formulation developed here, the DEA estimator of the efficiency score could also be obtained by convexifying the FDH input efficient boundary obtained by solving (2.12) for each data point (x_i, y_i). Namely:

$$\begin{aligned} \widehat{\theta}_{DEA}(x,y) \;\; = \;\; & \inf\{\theta \mid y \leq \sum_{i=1}^{n} \gamma_i\, y_i \;;\; \theta x \geq \sum_{i=1}^{n} \gamma_i\, \hat{x}_i^{\partial,FDH}, \\ & \qquad \text{for } (\gamma_1,...,\gamma_n) \quad \text{s.t.} \sum_{i=1}^{n} \gamma_i = 1; \gamma_i \geq 0, i = 1,...,n\Big\}. \end{aligned} \tag{2.14}$$

where $\hat{x}_i^{\partial,FDH} = \widehat{\theta}_{FDH}(x_i,y_i)\, x_i$ is the FDH-input efficient level computed by using (2.12) at the observed point (x_i, y_i), *i.e.*, the lower boundary, on the ray (input mix) x_i, of the support of $\widehat{F}_{X|Y,n}(\cdot \mid y_i)$.

3 Order-m frontiers and efficiency scores

The FDH estimator $\widehat{\Psi}_{FDH}$, as well as its convex version $\widehat{\Psi}_{DEA}$, are very sensitive to extremes and outliers, since they envelop all the data points of the observed set $\mathcal{X}$. To be more robust to extreme values CFS propose to estimate an order-m frontier, which corresponds to another benchmark frontier against which units will be compared.

3.1 General formulation

As pointed above the support of $F_{X|Y}(\cdot \mid y)$ defines the attainable set of input values X for a unit working at the output level y. Now instead of looking at the lower boundary of this support, we prefer to define as a benchmark value, the average of the minimal value of inputs for m units randomly drawn according $F_{X|Y}(\cdot \mid y)$, *i.e.*, units producing at least the output level y. This defines the input order-m frontier.

Formally, for a given level of output y, we consider m i.i.d. random variables $X_1,...,X_m$ generated by the conditional $p-$variate distribution function $F_{X|Y}(\cdot \mid y)$ and obtain the

random production set of order$-m$ for units producing more than y:

$$\widetilde{\Psi}_m(y) = \{(x, y') \in I\!R_+^{p+q} |\ x \geq X_i, y' \geq y, i = 1, ..., m\}. \tag{3.1}$$

Then, the order-m input efficiency score is defined as:

$$\theta_m(x, y) = E_{X|Y}(\tilde{\theta}_m(x, y) \mid Y \geq y), \tag{3.2}$$

where $\tilde{\theta}_m(x, y) = \inf\{\theta \mid (\theta x, y) \in \widetilde{\Psi}_m(y)\}$ and $E_{X|Y}$ is the expectation relative to the distribution $F_{X|Y}(\cdot \mid y)$.

Hence, the order-m efficiency score is 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 at least the output 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 or the same level than its level y and we take as benchmark, the expectation of the minimal achievable inputs in place of the absolute minimal achievable inputs.

The order-m frontier can be described by the set $(x_m^{\partial}(y), y) \in \Psi$, where $x_m^{\partial}(y) = \theta_m(x, y)\, x$ is the radial projection of $(x, y) \in \Psi$ on the order-m frontier, in the input direction (orthogonal to the vector y). We can also define the resulting attainable set of order-m by:

$$\Psi_m = \{(x, y) \in \Psi \mid x \geq x_m^{\partial}(y)\}. \tag{3.3}$$

Note that since $\theta_m(x, y)$ may be $\geq$ or ≤ 1, some $(x, y) \in \Psi$, may be outside the order-m set Ψ_m. As $m \to \infty$, of course, $\Psi_m \to \Psi$ and $\theta_m(x, y) \to \theta(x, y)$.

A nonparametric estimator $\widehat{\theta}_m(x, y)$ of order-m efficiency scores $\theta_m(x, y)$ (and of the corresponding frontier) is obtained by plugging the empirical version of $F_{X|Y}(\cdot \mid y)$ in the formulae above. The computations involves the computation of the following one-dimensional integral,

$$\begin{aligned} \widehat{\theta}_m(x, y) &= \widehat{E}_{X|Y}(\tilde{\theta}_m(x, y) \mid Y \geq y) && (3.4) \\ &= \int_0^{\infty} (1 - \widehat{F}_{X|Y}(ux \mid y))^m du, \\ &= \widehat{\theta}_{FDH}(x, y) + \int_{\widehat{\theta}_{FDH}(x,y)}^{\infty} (1 - \widehat{F}_{X|Y}(ux \mid y))^m du. && (3.5) \end{aligned}$$

Note that a simple Monte-Carlo procedure, as described in DS and CFS, may approximate the empirical expectation in (3.4) and so avoiding numerical integration (for large values of m, the integral is much faster to compute).

One of the main advantage of this estimator is that it does not suffer from the so called 'curse of dimensionality' characterizing most nonparametric estimators and implying for

great values of $(p+q)$ the need of large data sets in order to reduce statistical imprecision (length of confidence intervals, bias of the estimators, ...). We achieve here, for a fixed value of m, the standard root-n convergence rate of $\widehat{\theta}_m(x,y)$ to $\theta_m(x,y)$ and a Normal limiting distribution.

Another main advantage of this estimator is that it also provides a much more robust estimator to outliers or extreme values than the full frontier estimator since by construction, it does not envelop all the data points. We noticed above that when $m \to \infty$, the order-m frontier converges to the full frontier. The same is true for the estimator: $\widehat{\theta}_m(x,y) \to \widehat{\theta}_{FDH}(x,y)$ when $m \to \infty$. Therefore, choosing appropriately $m(n) \to \infty$ as a function of n, we can use $\widehat{\theta}_{m(n)}(x,y)$ as an estimator of the full frontier efficient level $\theta(x,y)$: this is a way of defining a robust estimator of the full frontier, since for any finite m, the corresponding frontier will not envelop all the data points. CFS show indeed that this robust estimator of $\theta(x,y)$ shares the asymptotic properties of the FDH estimator, in particular, $\widehat{\theta}_{m(n)}(x,y) \to \theta(x,y)$ when $n \to \infty$.

In practice for finite samples, several values of m are chosen and a particular value of m can be specified by looking at the percentage of points in the sample which stands outside $\widehat{\Psi}_m$. This percentage could be interpreted as the robustness level of the estimator (we could choose such a percentage as, say, 5% or 10%,...). These percentages have been used in Simar (2003) to warn or detect potential outliers in the data set.

3.2 Introducing convexity

In this section, we discuss issues concerning the convexity of the attainable production set of order-m, Ψ_m, as defined in (3.3). To the best of our knowledge, no general results have been published so far on the shape of Ψ_m. CFS give some monotonicity properties of the frontier, as a function of y in the case where $p = 1$ (see Theorem 2.4 in CFS: $F_{X|Y}(x \mid y)$ has to be monotone non-increasing with y to obtain a monotone frontier). Florens and Simar (2005) give some bivariate examples ($p = q = 1$) where the order-m frontier can be analytically computed and where Ψ and Ψ_m are both convex. As a matter of fact, there is basically no reason why Ψ_m should be convex, even if Ψ is convex, unless some very peculiar structure is imposed on $H_{X,Y}(x,y)$. This is due to the expectation defining the efficient level of order-m in (3.2) and then on its dependence on y.

However, we have seen that order-m frontiers are particularly useful to provide robust and consistent estimators of the full frontier when $m(n) \to \infty$ with n at the appropriate rate. Hence, if the true attainable set Ψ is convex, it is useful to impose some convexity assumptions on order-m attainable sets and their estimators, in order to provide a robust estimation of the full frontier. This can be done at two levels: either locally (for a given

value of y), or globally.

• Local convexity

We can indeed for a given level of output y and for a given value of m, introduce the random convex production set of order-m for units producing more than y, denoted by $\widetilde{\Psi}_m^C(y)$, as the convex hull of $\widetilde{\Psi}_m(y)$ (defined in equation (3.1)):

$$\begin{aligned}\widetilde{\Psi}_m^C(y) &= \mathcal{CH}\Big(\widetilde{\Psi}_m(y)\Big) &(3.6)\\ &= \{(x,y') \in I\!R_+^{p+q} |\ x \geq \sum_{i=1}^{m} \gamma_i X_i,\ \text{for } (\gamma_1,\ldots,\gamma_n) &(3.7)\\ &\qquad \text{such that } \sum_{i=1}^{m} \gamma_i = 1\,;\ \gamma_i \geq 0,\ y' \geq y,\ i = 1,...,m\},\end{aligned}$$

where the X_i are generated by $F_{X|Y}(\cdot|y)$, as above. Then for the order-m efficiency score, we define a locally-convex order-m input efficiency measure as:

$$\theta_m^{LC}(x,y) = E_{X|Y}(\tilde{\theta}_m^{LC}(x,y) \mid Y \geq y), \tag{3.8}$$

where $\tilde{\theta}_m^{LC}(x,y) = \inf\{\theta \mid (\theta x, y) \in \widetilde{\Psi}_m^C(y)\}$. The resulting order-$m$ frontier, is described by the set $(x_m^{\partial,LC}(y), y) \in \Psi$, where $x_m^{\partial,LC}(y) = \theta_m^{LC}(x,y)\, x$ is the radial projection of $(x,y) \in \Psi$ on the corresponding order-m frontier, in the input direction (orthogonal to the vector y).

Note that when $p = 1$ we have $\theta_m^{LC}(x,y) \equiv \theta_m(x,y)$ so that the "local-convex" order-m frontier is identical to the basic order-m frontier $(x_m^{\partial}(y), y) \in \Psi$ described in the preceding section.

It should be noticed that the local convex constraint for a given y, in (3.7), does not provide a global convex attainable set of order-m. Denoting this set by Ψ_m^{LC}, it is defined through:

$$\Psi_m^{LC} = \{(x',y) \in I\!R_+^{p+q} \mid x' \geq x_m^{\partial,LC}(y) \text{ for } (x,y) \in \Psi\}. \tag{3.9}$$

Nothing indeed ensures that Ψ_m^{LC} is convex. We will discuss later how to estimate these quantities.

• Global convexity

A natural way to obtain a convex set of order-m is to convexify Ψ_m globally and not only locally. As a matter of fact, we can define the convex attainable set of order-m, Ψ_m^C, as the convex closure of Ψ_m:

$$\Psi_m^C = \mathcal{CH}\Big(\Psi_m\Big). \tag{3.10}$$

If the set Ψ_m is convex, then of course $\Psi_m^C \equiv \Psi_m$. A corresponding order-m efficiency score, with this convex reference set, is then defined by:

$$\theta_m^C(x,y) = \inf\{\theta \mid (\theta\, x, y) \in \Psi_m^C\}. \tag{3.11}$$

This order-m efficiency score has the property of being defined with respect to a convex attainable set of order-m. As seen below, it has the advantage of being easy to estimate and it will provide a robust version of the DEA estimator.

• **Estimation of** $\theta_m^{LC}(x,y)$

The idea is, as above, to plug-in the empirical version of $F_{X|Y}(\cdot|y)$ in the expressions (3.6) to (3.8). A nonparametric estimator of $\theta_m^{LC}(x,y)$ is then obtained by using the empirical version of the expectation in (3.8):

$$\widehat{\theta}_m^{LC}(x,y) = \widehat{E}_{X|Y}(\tilde{\theta}_m^{LC}(x,y) \mid Y \geq y). \tag{3.12}$$

This can be approximated by a simple Monte-Carlo procedure, similar to the Monte-Carlo procedure described in DS and CFS:

[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] Solve the following linear program

$$\begin{aligned} \tilde{\theta}_m^{LC,b}(x,y) \;\; = \;\; \inf\{\theta \mid \theta x \geq \sum_{i=1}^m \gamma_i\, X_{i,b}, \quad \text{for } (\gamma_1, ..., \gamma_m) \\ \text{s.t.} \sum_{i=1}^m \gamma_i = 1; \gamma_i \geq 0, i = 1, ..., m\Big\}. \end{aligned} \tag{3.13}$$

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

[4] Finally, $\widehat{\theta}_m^{LC}(x,y) \approx \frac{1}{B}\sum_{b=1}^B \tilde{\theta}_m^{LC,b}(x,y)$.

The quality of the approximation can be tuned by increasing B, but at a computational cost since at each step, we have to run the linear program (3.13).

• **Estimation of** $\theta_m^C(x,y)$

An estimator for the order-m efficient score relative to a global convex attainable set of order-m is even easier to obtain. In analogy with (2.14), we only have to project all the points on the estimated order$-m$ frontier and then run a DEA program, as follows:

$$\begin{aligned} \widehat{\theta}_m^C(x,y) \;\; = \;\; \inf\{\theta \mid y \leq \sum_{i=1}^n \gamma_i\, y_i \;;\; \theta x \geq \sum_{i=1}^n \gamma_i\, \hat{x}_{m,i}^{\partial}, \\ \text{for } (\gamma_1, ..., \gamma_n) \quad \text{s.t.} \sum_{i=1}^n \gamma_i = 1; \gamma_i \geq 0, i = 1, ..., n\Big\}. \end{aligned} \tag{3.14}$$

where $\hat{x}^{\partial}_{m,i} = \hat{\theta}_m(x_i, y_i)\, x_i$ is the estimated order-m input efficient level for the i^{th} observation.

• Properties

The statistical properties of these "convex" order-m estimators have still to be investigated, but under the appropriate convexity assumptions on Ψ_m, we conjecture that they share the same properties as the original order-m estimators. However, it is easy to analyze the behavior of these convex order-m measures when $m \to \infty$.

– By construction and under the convexity of Ψ, for all $(x, y) \in \Psi$ and $m \geq 1$, we have:

$$\theta(x, y) \leq \theta^C_m(x, y) \leq \theta^{LC}_m(x, y) \leq \theta_m(x, y), \tag{3.15}$$

and so, when $m \to \infty$, all the order-m efficiency scores converge to $\theta(x, y)$. Also, in practice, we expect that when Ψ_m is really convex, $\theta^C_m(x, y)$ will be very similar to $\theta^{LC}_m(x, y)$.

– For the estimators, we have the following similar relations. For all $(x, y) \in \Psi$, $m \geq 1$ and n, we have:

$$\widehat{\theta}_{DEA}(x, y) \leq \widehat{\theta}^C_m(x, y) \leq \widehat{\theta}^{LC}_m(x, y) \leq \widehat{\theta}_m(x, y). \tag{3.16}$$

Clearly, when $m \to \infty$, $\widehat{\theta}^C_m(x, y) \to \widehat{\theta}_{DEA}(x, y)$: compare (3.14) with (2.14) and note that when $m \to \infty$, $\hat{x}^{\partial}_{m,i} \to \hat{x}^{\partial,FDH}_i$.

For non-convex technologies, we have seen that $\widehat{\theta}_m(x, y)$ is a more robust estimator of the Farrell efficiency score $\theta(x, y)$ than the FDH estimator $\widehat{\theta}_{FDH}(x, y)$ (see the discussion above, end of Section 3.1). The same is true for convex technologies. Let $m(n)$ be a function of n going to infinity when $n \to \infty$, $\widehat{\theta}^C_{m(n)}(x, y)$ is a more robust estimator of the Farrell efficiency score $\theta(x, y)$ than the traditional DEA estimator $\widehat{\theta}_{DEA}(x, y)$, because for finite m the corresponding estimated frontier will not envelop all the data points. In practice, for convex technologies, the choice of m is done as for non-convex ones, by tuning the desired level of robustness.

As far as order-m efficiency scores themselves have to be estimated, $\widehat{\theta}_m(x, y)$ converges at the $\sqrt{n}$-rate to $\theta_m(x, y)$, Ψ being convex or non-convex. However, the two sets of relations (3.15) and (3.16) indicate that under the convexity assumption of Ψ_m, $\widehat{\theta}^C_m(x, y)$ is a more appropriate estimator of $\theta_m(x, y)$.

4 Conditional measures of efficiency

As shown in DS, the probabilistic formulation of the production process allows to introduce external-environmental factors. We denote by $Z \in I\!R^r$ these factors. The idea is that the joint distribution of (X, Y) conditional on $Z = z$ defines the production process if $Z = z$. By analogy with (2.7), the support of $H_{X,Y|Z}(x, y|z) = \text{Prob}(X \leq x, Y \geq y \,|\, Z = z)$ defines Ψ^z, the attainable production set when $Z = z$. For an input conditional measure of efficiency, the natural decomposition of this joint distribution is given by:

$$H_{X,Y|Z}(x, y|z) = F_{X|Y,Z}(x \mid y, z) S_{Y|Z}(y|z), \tag{4.1}$$

for all y such that $S_{Y|Z}(y|z) = \text{Prob}(Y \geq y \,|\, Z = z) > 0$ and where $F_{X|Y,Z}(x \mid y, z) = \text{Prob}(X \leq x \mid Y \geq y, Z = z)$. So, for all y such that $S_{Y|Z}(y|z) > 0$, Ψ^z can also be defined by the support of $F_{X|Y,Z}(\cdot \mid y, z) = \text{Prob}(X \leq x \mid Y \geq y, Z = z)$. Then, as above in (2.9), the lower boundary of the latter will define the lower boundary achievable for a unit producing an output level y with an environment described by the value z. Formally we have:

$$\theta(x, y \mid z) = \inf\{\theta \mid F_{X|Y,Z}(\theta x \mid y, z) > 0\}. \tag{4.2}$$

Note again that the conditioning on Y is the event $Y \geq y$ (because Y is an output) and the conditioning on Z is defined, as in a regression framework, by $Z = z$. Note also that Ψ^z can be described as:

$$\Psi^z = \{(x', y) \in I\!R_+^{p+q} \mid x' \geq x^{\partial,z}(y) \text{ for } (x, y) \in \Psi\}, \tag{4.3}$$

where $x^{\partial,z}(y)$ is the efficient level of input, conditional on $Z = z$, for an output level y: $x^{\partial,z}(y) = \theta(x, y \mid z)\, x$, where $(x, y) \in \Psi$. Clearly, $\Psi^z \subseteq \Psi$.

4.1 Conditional FDH

• **Definition of $\widehat{\theta}_{FDH}(x, y \mid z)$**

A natural nonparametric estimator is obtained by plugging a nonparametric estimator of $F_{X|Y,Z}(\cdot \mid y, z)$ in the expression above (4.2). Due to the equality in the conditioning on Z this requires some smoothing techniques. At this purpose we use a kernel estimator defined as:

$$\widehat{F}_{X|Y,Z,n}(x \mid y, z) = \frac{\sum_{i=1}^n 1\!\!1(x_i \leq x, y_i \geq y) K((z - z_i)/h)}{\sum_{i=1}^n 1\!\!1(y_i \geq y) K((z - z_i)/h)}, \tag{4.4}$$

where $K(\cdot)$ is the kernel and h is the bandwidth of appropriate size[6]. Hence, we obtain the "conditional FDH efficiency measure" as follows:

$$\widehat{\theta}_{FDH}(x, y \mid z) = \inf\{\theta \mid \widehat{F}_{X|Y,Z,n}(\theta x \mid y, z) > 0\}. \tag{4.5}$$

[6]Issues about the practical choice of the bandwidth are discussed in the Appendix.

As pointed in DS, for any (symmetric) kernel with compact support[7] (*i.e.*, $K(u) = 0$ if $|u| > 1$, as for the uniform, triangle, epanechnikov or quartic kernels), the conditional FDH efficiency estimator is given by:

$$\widehat{\theta}_{FDH}(x, y|z) = \inf\{\theta \mid \widehat{F}_{X|Y,Z,n}(\theta x \mid y, z) > 0\} = \min_{\{i|Y_i \geq y, |Z_i - z| \leq h\}} \left\{ \max_{j=1,\ldots,p} \left(\frac{X_i^j}{x^j} \right) \right\}. \quad (4.6)$$

Therefore, it does not depend on the chosen kernel but only on the selected bandwidth. This will be different for the conditional order-m measures defined below.

• Conditional FDH attainable set

The conditional attainable set Ψ^z is estimated by:

$$\widehat{\Psi}^z_{FDH} = \{(x', y) \in I\!R^{p+q}_+ \mid x' \geq \hat{x}^{\partial, FDH, z}(y) \text{ for } (x, y) \in \widehat{\Psi}_{FDH}\} \quad (4.7)$$

where $\hat{x}^{\partial, FDH, z}(y)$ is the estimated conditional efficient level of inputs:

$$\hat{x}^{\partial, FDH, z}(y) = \widehat{\theta}_{FDH}(x, y|z)\, x \;\text{ for } (x, y) \in \widehat{\Psi}_{FDH}.$$

Note that the conditional FDH attainable set can also be defined as follows. A production unit characterized by the observation (x_i, y_i, z_i) defines an individual attainable set $\psi(x_i, y_i \mid z_i)$, which under free disposability of inputs and outputs can be written as in (2.3):

$$\psi(x_i, y_i \mid z_i) = \{(x, y) \in I\!R^{p+q}_+ \mid x \geq x_i, y \leq y_i\}. \quad (4.8)$$

Indeed, for this value of $Z = z_i$, all the points in $\psi(x_i, y_i \mid z_i)$ are, under free disposability, attainable. Now for any given value of $Z = z$, the 'global' attainable set will be obtained by the union of all the attainable sets $\psi(x_i, y_i \mid z_i)$ for z_i being in a h-neighborhood of z:

$$\begin{aligned} \widehat{\Psi}^z_{FDH} &= \bigcup_{\{i|z-h \leq z_i \leq z+h\}} \psi(x_i, y_i \mid z_i) \quad (4.9) \\ &= \{(x, y) \in I\!R^{p+q}_+ \mid x \geq x_i, y \leq y_i, \text{ for } i \text{ s.t. } z - h \leq z_i \leq z + h\}. \end{aligned}$$

The conditional FDH efficiency score can thus be equivalently defined by:

$$\widehat{\theta}_{FDH}(x, y|z) = \inf\{\theta \mid (\theta x, y) \in \widehat{\Psi}^z_{FDH}\}, \quad (4.10)$$

• Properties

Note that the union of all the conditional attainable sets over all the observed values $z_i \in I\!R^r, i = 1, \ldots, n$ will recover the full FDH production set. In symbols:

$$\bigcup_{i=1,\ldots,n} \widehat{\Psi}^{z_i}_{FDH} \equiv \widehat{\Psi}_{FDH}. \quad (4.11)$$

[7] DS pointed out that for kernels with unbounded support, like the gaussian kernel, it is easy to show that $\widehat{\theta}_{FDH}(x, y|z) \equiv \widehat{\theta}_{FDH}(x, y)$: 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.

Indeed we have:

$$\begin{aligned}\bigcup_{i=1,\ldots,n} \widehat{\Psi}^{z_i}_{FDH} &= \bigcup_{i=1,\ldots,n}\left(\bigcup_{\{j|z_i-h\le z_j\le z_i+h\}} \psi(x_j,y_j \mid z_j)\right)\\ &= \bigcup_{i=1,\ldots,n}\left(\bigcup_{\{j|z_i-h\le z_j\le z_i+h\}} \Big\{(x,y)\in I\!R^{p+q}_+ \mid x\ge x_j, y\le y_j\Big\}\right)\\ &= \bigcup_{i=1,\ldots,n}\Big\{(x,y)\in I\!R^{p+q}_+ \mid x\ge x_i, y\le y_i\Big\} \equiv \widehat{\Psi}_{FDH}.\end{aligned}$$

4.2 Conditional DEA

• Conditional DEA attainable set

Now, by analogy with (2.6), if we suppose that the true conditional attainable set Ψ^z is convex, we can introduce an additional convexity constraints on our estimator. This defines the conditional DEA attainable set:

$$\widehat{\Psi}^z_{DEA} = \mathcal{CH}\left(\widehat{\Psi}^z_{FDH}\right) = \mathcal{CH}\Big(\bigcup_{\{i|z-h\le z_i\le z+h\}} \psi(x_i,y_i \mid z_i)\Big) \tag{4.12}$$

$$= \left\{(x,y)\in I\!R^{p+q}_+ \mid y\le \sum_{\{i|z-h\le z_i\le z+h\}} \gamma_i y_i \,;\, x\ge \sum_{\{i|z-h\le z_i\le z+h\}} \gamma_i x_i \right.$$
$$\left. \text{for nonnegative } \gamma\text{'s such that } \sum_{\{i|z-h\le z_i\le z+h\}} \gamma_i = 1\right\}. \tag{4.13}$$

Note that this provides a local convex attainable set, local in the sense of conditional on the external factors $Z=z$. This is true for all values of $z\in I\!R^r$.

As a matter of fact, $\widehat{\Psi}^z_{FDH}$ is an estimator of the attainable set conditional on $Z=z$, relying only on free disposability and $\widehat{\Psi}^z_{DEA}$ is an estimator relying on the additional assumption of convexity.

• Definition of $\widehat{\theta}_{DEA}(x,y \mid z)$

A conditional DEA-efficiency score may be defined by:

$$\widehat{\theta}_{DEA}(x,y|z) = \inf\{\theta \mid (\theta x, y)\in \widehat{\Psi}^z_{DEA}\}. \tag{4.14}$$

It can be computed by solving the linear program:

$$\widehat{\theta}_{DEA}(x,y|z) = \inf\{\theta \mid y\le \sum_{\{i|z-h\le z_i\le z+h\}} \gamma_i\, y_i \,;\, \theta x\ge \sum_{\{i|z-h\le z_i\le z+h\}} \gamma_i\, x_i,$$
$$\text{for nonnegative } \gamma\text{'s} \quad \text{s.t.} \sum_{\{i|z-h\le z_i\le z+h\}} \gamma_i = 1\Big\}. \tag{4.15}$$

Of course, in the latter linear program, x_i could be replaced by its projection on the FDH efficient frontier, *i.e.*, by $\hat{x}_i^{\partial,FDH} = \widehat{\theta}_{FDH}(x_i, y_i)\, x_i$.

• Properties

Note that here, the union of all these sets over all the observed values $z_i \in I\!R^r, i = 1, \ldots, n$ will recover partly the full DEA production set. This is because the union of convex hull of sets is a subset of the convex hull of the union of the sets. So we have:

$$\bigcup_{i=1,\ldots,n} \widehat{\Psi}_{DEA}^{z_i} = \bigcup_{i=1,\ldots,n} \mathcal{CH}\left(\widehat{\Psi}_{FDH}^{z_i}\right) \subseteq \mathcal{CH}\left(\bigcup_{i=1,\ldots,n} \widehat{\Psi}_{FDH}^{z_i}\right) = \mathcal{CH}\left(\widehat{\Psi}_{FDH}\right) = \widehat{\Psi}_{DEA}. \quad (4.16)$$

But of course, the convex hull of the union will coincide with the DEA set:

$$\mathcal{CH}\left(\bigcup_{i=1,\ldots,n} \widehat{\Psi}_{DEA}^{z_i}\right) \equiv \widehat{\Psi}_{DEA} \quad (4.17)$$

4.3 Conditional order-m measures

• General approach

The conditional order-m input efficiency measure is defined in DS, where only free disposability is assumed. For a given level of outputs y in the interior of the support of Y, we consider the m i.i.d. random variables $X_i, i = 1, \ldots, m$ generated by the conditional p-variate distribution function $F_{X|Y,Z}(x \mid y, z)$ and we define the conditional random set:

$$\widetilde{\Psi}_m^z(y) = \{(x, y') \in I\!R_+^{p+q} \mid x \geq X_i, y' \geq y,\ i = 1, \ldots, m\}. \quad (4.18)$$

Note that this set depends on the value of z since the X_i are generated through the conditional distribution function. 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 then defined as:

$$\theta_m(x, y|z) = E_{X|Y,Z}(\tilde{\theta}_m^z(x, y) \mid Y \geq y, Z = z), \quad (4.19)$$

where $\tilde{\theta}_m^z(x, y) = \inf\{\theta \mid (\theta x, y) \in \widetilde{\Psi}_m^z(y)\}$ and the expectation is relative to the distribution $F_{X|Y,Z}(\cdot \mid y, z)$. It is shown by DS (Theorem 3.1) that $\theta_m(x, y|z)$ converges to $\theta(x, y|z)$ when $m \to \infty$.

• Definition of $\widehat{\theta}_m(x, y|z)$

A nonparametric estimator of $\theta_m(x, y|z)$ is provided by plugging the nonparametric estimator of $F_{X|Y,Z}(x|y, z)$ proposed in (4.4), which depends on the kernel and on the chosen bandwidth. Formally, the estimator can be obtained by:

$$\begin{aligned} \widehat{\theta}_m(x, y|z) &= \widehat{E}_{X|Y,Z}(\tilde{\theta}_m^z(x, y) \mid y, z) && (4.20) \\ &= \int_0^\infty (1 - \widehat{F}_{X,n}(ux \mid y, z))^m du. && (4.21) \end{aligned}$$

This involve the computation of a one-dimensional numerical integral. Note that DS propose also a Monte-Carlo algorithm to approximate the empirical expectation in (4.20), but for large m solving the integral is much faster.

Since $\widehat{\theta}_m(x,y|z) \to \widehat{\theta}_{FDH}(x,y|z)$ when $m \to \infty$, the order-m conditional efficiency score can again be viewed as a robust estimator of the conditional efficiency score $\theta(x,y|z)$ when chosing $m = m(n) \to \infty$ with $n \to \infty$. For finite m, the corresponding attainable set will not envelop all the data points and so is more robust to extremes or outlying data points.

• Introducing convexity

Following the same idea as in Section 3.2 we can provide a robust estimator of the conditional DEA efficiency score by convexifying the conditional attainable set obtained by the estimates $\widehat{\theta}_m(x,y|z)$. We first define, as above, $\widetilde{\Psi}_m^{C,z}(y)$ as being the convex hull of $\widetilde{\Psi}_m^z(y)$:

$$\begin{aligned} \widetilde{\Psi}_m^{C,z}(y) &= \mathcal{CH}\Big(\widetilde{\Psi}_m^z(y)\Big) && (4.22)\\ &= \{(x,y') \in I\!R_+^{p+q} |\ x \geq \sum_{i=1}^m \gamma_i X_i, \text{ for } (\gamma_1,\ldots,\gamma_n) && (4.23)\\ &\qquad \text{such that } \sum_{i=1}^m \gamma_i = 1\,;\, \gamma_i \geq 0,\ y' \geq y,\ i = 1,...,m\}, \end{aligned}$$

this random convex set depends on z through the random generation of the X_i, $i = 1,\ldots,m$. The corresponding conditional efficiency score of order-m is then defined by:

$$\theta_m^C(x,y|z) = E_{X|Y,Z}(\tilde{\theta}_m^{C,z}(x,y) \mid Y \geq y, Z = z), \tag{4.24}$$

where $\tilde{\theta}_m^{C,z}(x,y) = \inf\{\theta \mid (\theta x, y) \in \widetilde{\Psi}_m^{C,z}(y)\}$ and the expectation is relative to the distribution $F_{X|Y,Z}(\cdot \mid y,z)$. Of course, when $p = 1$, $\theta_m^C(x,y|z) \equiv \theta_m(x,y|z)$, because $\widetilde{\Psi}_m^z(y)$ is trivially convex.

Clearly, when $m \to \infty$, and if Ψ^z is convex, $\theta_m^C(x,y|z)$ converges to $\theta(x,y|z)$: this can be seen when realizing that for all m, under convexity of Ψ^z, $\theta(x,y|z) \leq \theta_m^C(x,y|z) \leq \theta_m(x,y|z)$ and that $\theta_m(x,y|z)$ converges to $\theta(x,y|z)$ as $m \to \infty$.

• Definition of $\widehat{\theta}_m^C(x,y \mid z)$

The conditional efficiency scores of order-m, relative to a convex conditional attainable set, can be estimated by replacing the unknown $F_{X|Y,Z}(\cdot|y,z)$ needed in computing (4.24) by its nonparametric estimator proposed in (4.4).

$$\widehat{\theta}_m^C(x,y \mid z) = \widehat{E}_{X|Y,Z}(\tilde{\theta}_m^{C,z}(x,y) \mid Y \geq y, Z = z). \tag{4.25}$$

In practice, this can be computed by the following Monte-Carlo algorithm (adapted from DS for convex sets). Suppose that h is the chosen bandwidth for a particular kernel $K(\cdot)$ with bounded support:

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

[2] Solve the following linear program

$$\tilde{\theta}_m^{C,z,b}(x,y) = \inf\{\theta \mid \theta x \geq \sum_{i=1}^{m} \gamma_i \, X_{i,b}, \quad \text{for } (\gamma_1, ..., \gamma_m) \text{ s.t.} \sum_{i=1}^{m} \gamma_i = 1; \gamma_i \geq 0, i = 1, ..., m\Big\}. \tag{4.26}$$

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

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

As usual, the quality of the Monte-Carlo approximation can be tuned by the choice of B.

• **Properties**

Here, when $m \to \infty$, $\widehat{\theta}_m^C(x, y \mid z)$ will converge to the conditional DEA efficiency score $\widehat{\theta}_{DEA}(x, y \mid z)$, so again, this version of order-m estimator relative to convex conditional attainable sets can be viewed as a robust version of the conditional DEA estimator.

5 Indicators for convexity

Extending ideas from Briec, Kerstens and Vanden Eeckaut (2004) and the references reported there, we can built indicators of convexity with some simple ratios of the several measures of efficiency introduced above. Table 1 summarizes the measures of interest.

Table 1: *Efficiency estimators: a summary table with references*

	Unconditional		Conditional	
	Non-Convex Tech.	**Convex Tech.**	**Non-Convex Tech.**	**Convex Tech.**
Full frontiers	$\widehat{\theta}_{FDH}(x,y)$	$\widehat{\theta}_{DEA}(x,y)$	$\widehat{\theta}_{FDH}(x,y\|z)$	$\widehat{\theta}_{DEA}(x,y\|z)$
	Deprins, Simar and Tulkens (1984)	*Charnes, Cooper and Rodhes(1978)*	*Daraio and Simar(2003)*	*This paper*
Robust frontiers	$\widehat{\theta}_{m}(x,y)$	**Local Conv.** $\widehat{\theta}^{LC}_{m}(x,y)$	$\widehat{\theta}_{m}(x,y\|z)$	$\widehat{\theta}^{c}_{m}(x,y\|z)$
	Cazals, Florens and Simar(2002)	*This paper*	*Cazals, Florens, and Simar (2002), Daraio and Simar (2003)*	*This paper*
		Global Conv. $\widehat{\theta}^{C}_{m}(x,y)$		
		This paper		

When using the indicators of convexity, we prefer here to avoid the words "goodness of fit tests for convexity", as used by Briec et al (2004), because, formally they do not provide a "test" in a statistical sense, but rather indicators or descriptive statistics.

These ratios provide indeed useful indications about the convexity assumption by comparing the convex and non convex version of the various efficiency scores. Along these lines we can built the following indicators of convexity for each DMU:

- Indicator of Convexity for the full frontier efficiency score estimates:

$$IC_i = \frac{\hat{\theta}_{DEA}(x_i, y_i)}{\hat{\theta}_{FDH}(x_i, y_i)}$$

- Indicator of Convexity for the *conditional* full frontier efficiency score estimates:

$$ICZ_i = \frac{\hat{\theta}_{DEA}(x_i, y_i \mid z_i)}{\hat{\theta}_{FDH}(x_i, y_i \mid z_i)}$$

- Indicator of Convexity for the order-m frontier efficiency score estimates:

$$IC_{m,i} = \frac{\hat{\theta}_m^C(x_i, y_i)}{\hat{\theta}_m(x_i, y_i)}$$

- Indicator of Convexity for the *conditional* order-m frontier efficiency score estimates.

$$ICZ_{m,i} = \frac{\hat{\theta}_m^C(x_i, y_i \mid z_i)}{\hat{\theta}_m(x_i, y_i \mid z_i)},$$

where of course the latter indicator is trivially equal to 1 when $p = 1$.

Table 2 summarizes the different indicators.

Table 2: *Indicators for convexity*

	UNCONDITIONAL	CONDITIONAL
FULL FRONTIERS	$IC_i = \frac{\widehat{\theta}_{DEA}(x,y)}{\widehat{\theta}_{FDH}(x,y)}$ *Briec, Kerstens and Vanden Eeckaut (2004)*	$ICZ_i = \frac{\widehat{\theta}_{DEA}(x,y\|z)}{\widehat{\theta}_{FDH}(x,y\|z)}$ *This paper*
ROBUST FRONTIERS	$IC_{m,i} = \frac{\widehat{\theta}_m^C(x,y)}{\widehat{\theta}_m(x,y)}$ *This paper*	$ICZ_{m,i} = \frac{\widehat{\theta}_m^C(x,y\|c)}{\widehat{\theta}_m(x,y\|c)}$ *This paper*

By construction, all these ratios are less or equal to one (in the input oriented framework adopted here) and under the convexity assumption, they should not be far from one at least for large sample sizes. A statistical test could be developed according these lines, by building some appropriate test statistics (like average of the indicators over the sample units). Then we would reject the null hypothesis of convexity if the test statistic is too small. Bootstrap techniques are the only way to perform these tests in a rigorous way by evaluating the appropriate p-values. The implementation of the bootstrap should follow the lines of Simar and Wilson (2001, 2002). This will not be pursued here and is left for future work.

6 Empirical illustrations

We illustrate our methodology using some simulated data and a real data set on US mutual funds, belonging to the Aggressive Growth category.

6.1 Simulated datasets

We simulated a simple Cobb-Douglas technology with 3 different scenarios for the external - environmental variable Z. We simulated a sample of size $n = 100$ from $Z \sim \text{Uniform}(1, 2)$ and compare three different scenarios for generating X. As above, we adopt an input orientation.

Simulated example 1 $X = Y^2\ Z^{-2}\ \varepsilon$, where $Y \sim \text{Uniform}(1, 2)$, ε is the random true inefficiency given by $\varepsilon = e^{0.4u}$, and $u \sim N^+(0, 1)$. Here Z is *favorable* for the production process: it is, in a certain sense, a *substitute* of the input X;

Simulated example 2 $X = Y^2\ Z^2\ \varepsilon$, where Y and ε are as above. Here we have a scenario similar to example 1 except that the effect of Z is *unfavorable*: if the value of Z augments, also X augments;

Simulated example 3 $X = Y^2\ \varepsilon$, where Y and ε are as above. In this case Z is independent of X and hence *neutral* for the production process.

6.1.1 Simulated example 1

Figure 1 illustrates how the nonparametric regression of the ratios between the conditional and unconditional efficiency measures on Z is able to capture the favorable effect of Z on the production process[8]. Although we are only working with estimated values, it also shows that our method for detecting the effect of Z is not affected by the convexity assumption, which was expected since the true sets are convex.

Figure 2 provides the same plot for the robust (order-m) version of the efficiency scores, where m was chosen to be equal to 25. As expected (there are no outliers here), the message of these plots is the same as for their full frontier correspondents.

Table 3 offers some descriptive statistics of the different input efficiency measures used here. To investigate the usefulness of the descriptive indicators of convexity, the table provides also some information on the distribution of these indicators in the sample (by giving

[8] As explained in DS, in an input oriented framework, an *increasing* smoothing nonparametric regression line describes a *negative* effect of the external factor Z on the production process. Whilst a *decreasing* nonparametric regression line highlights a *positive* effect of Z on the production process; and finally, a *straight* line denotes a *neutral* effect of Z on the production process. For more details on this topic see DS.

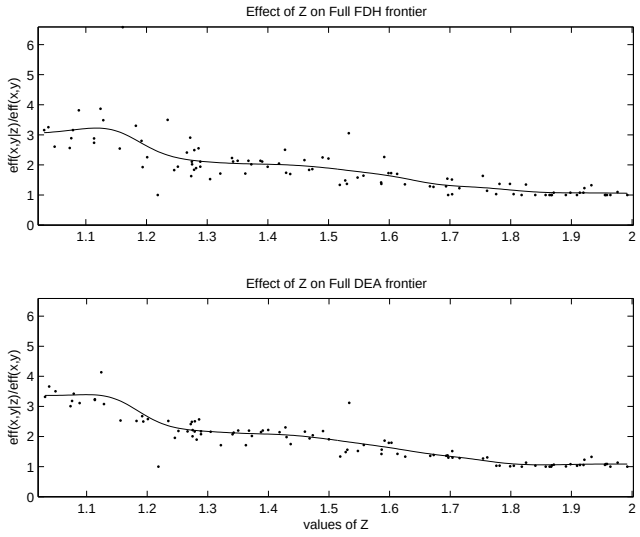

Figure 1: *Simulated example 1, positive (favorable) effect of Z on production efficiency (input oriented framework). Scatter plot and smoothed regression of $\hat{\theta}_{FDH,n}(x, y \mid z)/\hat{\theta}_{FDH,n}(x, y)$ on Z (top panel) and of $\hat{\theta}_{DEA,n}(x, y \mid z)/\hat{\theta}_{DEA,n}(x, y)$ on Z (bottom panel).*

the number of observations for which the indicators are $\geq 0.99, \ldots, 0.65$). The table deserves some comments:

- Since the true sets, Ψ and Ψ^z, are convex, the estimators are not too different using convex and non-convex approaches.
- We know that the true sets are convex, so the indicators IC and ICZ for the full frontier should be near one. The two distributions have indeed most of their mass above, say 0.90. Nevertheless, these indicators would be more useful for "testing" the convexity assumption within a formal inferential procedure (using the bootstrap as mentioned above).
- We do not know if the order-m attainable sets are convex, but comparing the distribution of the indicators IC_m with the distribution of IC, it seems that the distribution of robust indicators (IC_m) is less concentrate near 1 (*i.e.*, IC_m has a smaller proportion of values larger than 0.95: 34 observations against 42 for IC). Here again, a formal test should indicate if these differences are significant.
- Since $p = 1$, $\hat{\theta}^C_{m,n}(.|z) \equiv \hat{\theta}_{m,n}(.|z)$, hence the last column ICZ_m is identically equal to 1 for all the 100 observations.

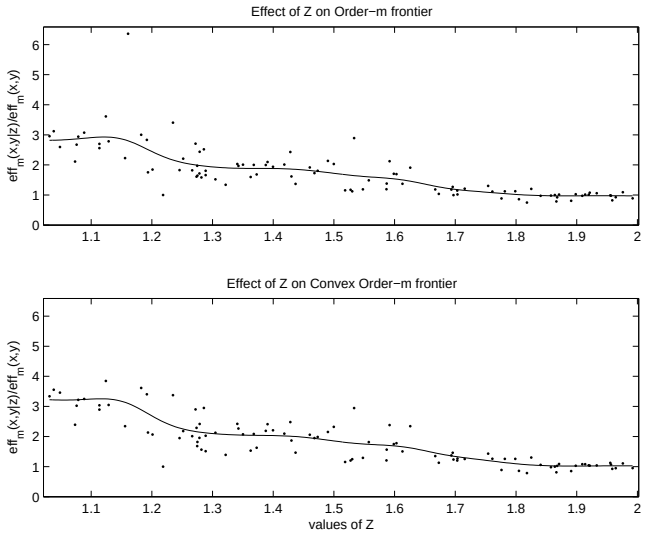

Figure 2: *Simulated example 1, positive effect of Z on production efficiency (input oriented framework). Scatter plot and smoothed regression of $\hat{\theta}_{m,n}(x,y \mid z)/\hat{\theta}_{m,n}(x,y)$ on Z (top panel), and of $\hat{\theta}^{C}_{m,n}(x,y \mid z)/\hat{\theta}^{C}_{m,n}(x,y)$ on Z (bottom panel).*

	$\hat{\theta}_{DEA,n}(x,y)$	$\hat{\theta}_{DEA,n}(x,y \mid z)$	$\hat{\theta}^{C}_{m,n}(x,y)$	$\hat{\theta}^{C}_{m,n}(x,y \mid z)$
Average	0.531	0.859	0.596	0.927
St. Dev.	0.241	0.178	0.274	0.134
Minimum	0.136	0.385	0.140	0.440
# Eff. Obs	3	32	10	63
	$\hat{\theta}_{FDH,n}(x,y)$	$\hat{\theta}_{FDH,n}(x,y \mid z)$	$\hat{\theta}_{m,n}(x,y)$	$\hat{\theta}_{m,n}(x,y \mid z)$
Average	0.579	0.917	0.647	0.927
St. Dev.	0.247	0.147	0.282	0.134
Minimum	0.152	0.437	0.157	0.440
# Eff. Obs	12	63	14	63
Indicators	IC	ICZ	IC_m	ICZ_m
$\# \geq 0.99$	26	50	11	100
$\# \geq 0.95$	42	61	34	100
$\# \geq 0.90$	59	74	59	100
$\# \geq 0.85$	79	81	90	100
$\# \geq 0.80$	90	89	97	100
$\# \geq 0.75$	97	92	98	100
$\# \geq 0.70$	99	97	99	100
$\# \geq 0.65$	100	98	100	100

Table 3: *Descriptive statistics of efficiency scores estimated over 100 observations, for the simulated example 1. Average, Standard Deviation, Minimum value, number of efficient observations, and distribution of the indicators of convexity.*

6.1.2 Simulated example 2

Figures 3 and 4 show that the nonparametric regression of the ratios between the conditional and unconditional efficiency measures on Z allows to capture, in this case, the unfavorable effect of Z on the production process (increasing nonparametric regression of the efficiency ratios on Z). Here again, as expected, the method for detecting the effect of Z is not affected by the convexity assumption.

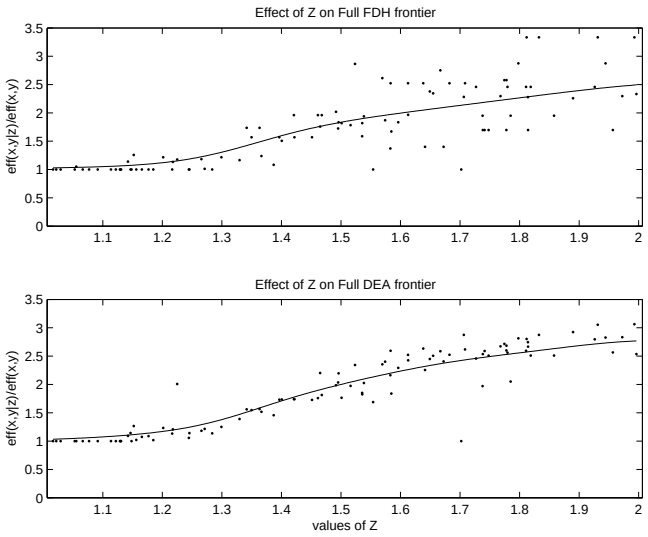

Figure 3: *Simulated example 2, negative (unfavorable) effect of Z on production efficiency (input oriented framework). Scatter plot and smoothed regression of $\hat{\theta}_{FDH,n}(x, y \mid z)/\hat{\theta}_{FDH,n}(x, y)$ on Z (top panel) and of $\hat{\theta}_{DEA,n}(x, y \mid z)/\hat{\theta}_{DEA,n}(x, y)$ on Z (bottom panel).*

Table 4 reports some descriptive statistics on the results of simulated example 2. The qualitative comments made above for Table 3 roughly apply in this case too: the orders of magnitude of the figures appearing in the tables are very similar. Note that here the difference between the distributions of IC and of ICm seems not significant (the number of observations larger than 0.95 is 15 for IC and 17 for ICm).

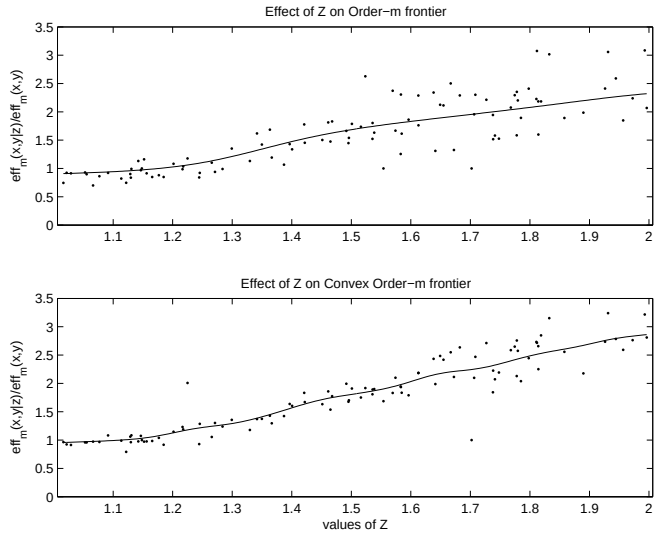

Figure 4: *Simulated example 2, negative effect of Z on production efficiency (input oriented framework). Scatter plot and smoothed regression of $\hat{\theta}_{m,n}(x, y \mid z)/\hat{\theta}_{m,n}(x, y)$ on Z (top panel), and of $\hat{\theta}^{C}_{m,n}(x, y \mid z)/\hat{\theta}^{C}_{m,n}(x, y)$ on Z (bottom panel).*

	$\hat{\theta}_{DEA,n}(x,y)$	$\hat{\theta}_{DEA,n}(x,y \mid z)$	$\hat{\theta}^{C}_{m,n}(x,y)$	$\hat{\theta}^{C}_{m,n}(x,y \mid z)$
Average	0.475	0.783	0.557	0.892
St. Dev.	0.221	0.169	0.267	0.155
Minimum	0.145	0.365	0.154	0.402
# Eff. Obs	5	8	10	42
	$\hat{\theta}_{FDH,n}(x,y)$	$\hat{\theta}_{FDH,n}(x,y \mid z)$	$\hat{\theta}_{m,n}(x,y)$	$\hat{\theta}_{m,n}(x,y \mid z)$
Average	0.581	0.872	0.653	0.892
St. Dev.	0.254	0.162	0.298	0.155
Minimum	0.201	0.393	0.213	0.402
# Eff. Obs	13	41	20	42
Indicators	IC	ICZ	IC_m	ICZ_m
$\# \geq 0.99$	6	19	5	100
$\# \geq 0.95$	15	36	17	100
$\# \geq 0.90$	28	49	41	100
$\# \geq 0.85$	35	63	54	100
$\# \geq 0.80$	61	87	76	100
$\# \geq 0.75$	81	97	82	100
$\# \geq 0.70$	85	99	94	100
$\# \geq 0.65$	89	100	98	100

Table 4: *Descriptive statistics of efficiency scores estimated over 100 observations, for the simulated example 2. Average, Standard Deviation, Minimum value, number of efficient observations, and distribution of the indicators of convexity.*

6.1.3 Simulated example 3

Figures 5 and 6 below illustrate that the nonparametric regression of the ratios between the conditional and unconditional efficiency measures on Z allows again to capture the real (neutral) effect of Z on the production process (straight nonparametric regression of the efficiency ratios on Z), with or without the convexity assumption.

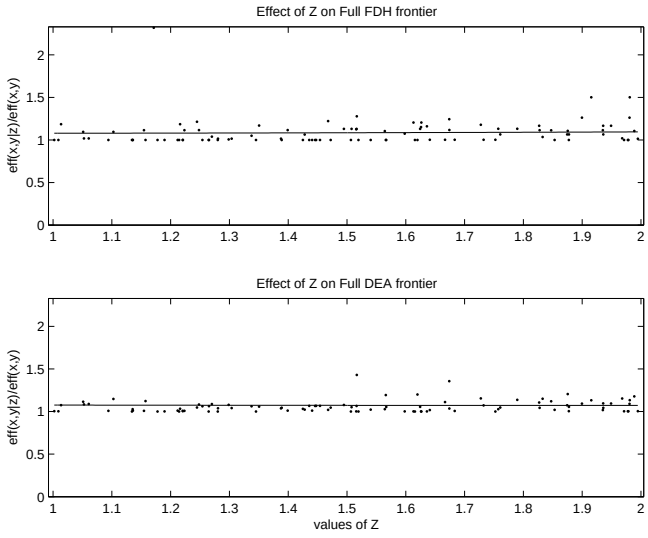

Figure 5: *Simulated example 3, no effect of Z on production efficiency (input oriented framework). Scatter plot and smoothed regression of* $\hat{\theta}_{FDH,n}(x, y \mid z)/\hat{\theta}_{FDH,n}(x, y)$ *on Z (top panel) and of* $\hat{\theta}_{DEA,n}(x, y \mid z)/\hat{\theta}_{DEA,n}(x, y)$ *on Z (bottom panel).*

The statistics on the efficiency scores and the indicators of convexity are given in Table 5. They mainly confirm the comments given for the preceding scenarios.

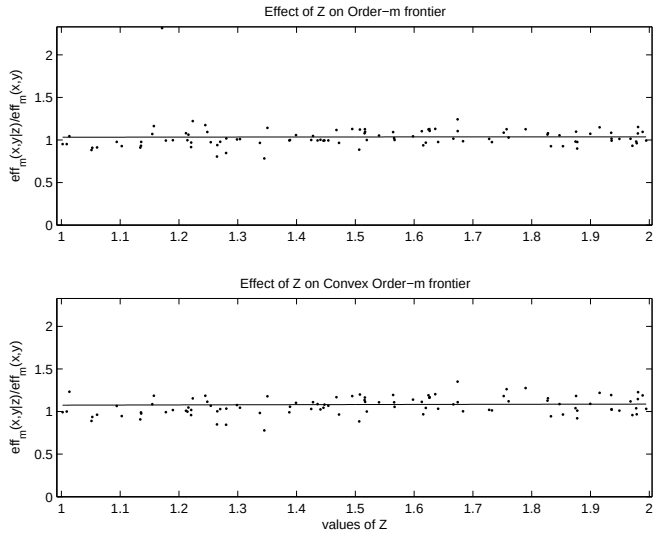

Figure 6: *Simulated example 3, no effect of Z on production efficiency (input oriented framework). Scatter plot and smoothed regression of* $\hat{\theta}_{m,n}(x, y \mid z)/\hat{\theta}_{m,n}(x, y)$ *on Z (top panel), and of* $\hat{\theta}^{C}_{m,n}(x, y \mid z)/\hat{\theta}^{C}_{m,n}(x, y)$ *on Z (bottom panel).*

	$\hat{\theta}_{DEA,n}(x,y)$	$\hat{\theta}_{DEA,n}(x,y \mid z)$	$\hat{\theta}^{C}_{m,n}(x,y)$	$\hat{\theta}^{C}_{m,n}(x,y \mid z)$
Average	0.761	0.810	0.836	0.889
St. Dev.	0.170	0.174	0.188	0.162
Minimum	0.306	0.344	0.353	0.378
# Eff. Obs	5	15	19	42
	$\hat{\theta}_{FDH,n}(x,y)$	$\hat{\theta}_{FDH,n}(x,y \mid z)$	$\hat{\theta}_{m,n}(x,y)$	$\hat{\theta}_{m,n}(x,y \mid z)$
Average	0.811	0.872	0.872	0.889
St. Dev.	0.176	0.170	0.190	0.162
Minimum	0.344	0.344	0.380	0.378
# Eff. Obs	17	41	29	42
Indicators	IC	ICZ	IC_m	ICZ_m
$\# \geq 0.99$	13	31	14	100
$\# \geq 0.95$	43	50	58	100
$\# \geq 0.90$	78	68	98	100
$\# \geq 0.85$	98	84	100	100
$\# \geq 0.80$	100	94	100	100
$\# \geq 0.75$	100	97	100	100
$\# \geq 0.70$	100	100	100	100
$\# \geq 0.65$	100	100	100	100

Table 5: *Descriptive statistics of efficiency scores estimated over 100 observations, for the simulated example 3. Average, Standard Deviation, Minimum value, number of efficient observations, and distribution of the indicators of convexity.*

6.2 Real data set

We illustrate our methodology analyzing Aggressive-Growth US Mutual Funds data. Several studies have applied efficiency analysis methods to evaluate the performance of mutual funds (see *e.g.* Murthi, Choi, and Desai,1997, and the references reported in Daraio and Simar, 2004a). We apply an input oriented framework in order to evaluate how mutual funds perform in terms of their risk (as expressed by standard deviation of return) and transaction costs (including expense ratio and turnover) management (so that we have $p = 3$ inputs). The traditional output in this framework is the total return of funds. Sengupta (2000) uses market risks as an input in his work, assuming that it has a favorable (positive) effect on the performance of the funds. In our illustration we use market risks as environmental variable (Z), 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. We used 3 inputs (risk, expense ratio and turnover), 1 output (return), 1 environmental factor (market risks) and 129 observations. For a detailed description and analysis of these data as well as a comparison with other US mutual funds category by objectives, see Daraio and Simar (2004a).

Figure 7 provides the nonparametric regression of the ratios between the conditional and unconditional efficiency measures on Z (market risks) for the US Aggressive Growth mutual funds.

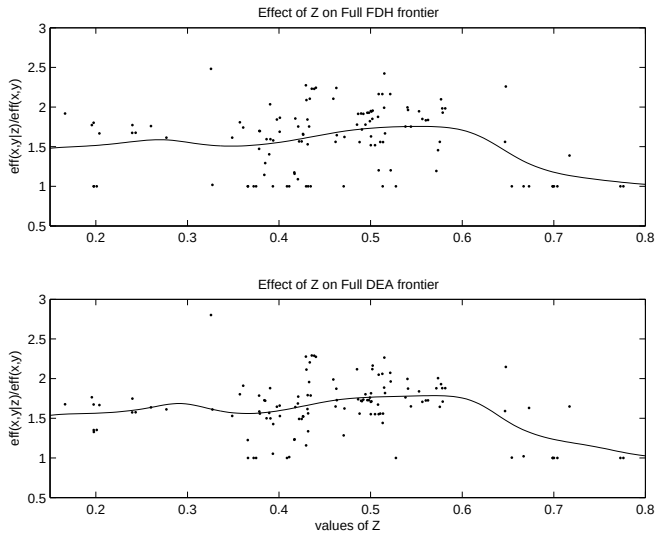

Figure 7: *US Aggressive Growth Mutual Funds data (input oriented framework). Scatter plot and smoothed regression of* $\hat{\theta}_{FDH,n}(x, y \mid z)/\hat{\theta}_{FDH,n}(x, y)$ *on* Z *(top panel) and of* $\hat{\theta}_{DEA,n}(x, y \mid z)/\hat{\theta}_{DEA,n}(x, y)$ *on* Z *(bottom panel).*

Globally these plots indicate that for a large part of the range of Z ($Z \leq 0.6$), a neu-

tral effect of the market risk is observed and that the positive effect (globally assumed in Sengupta's approach) appears only for larger values of Z. This illustrates how our tools can be useful in an exploratory phase to detect the effect of environmental variables on the production process, without any *a priori* assumption.

Figure 8 shows the picture for the robust versions of the frontiers. For order$-m$ efficiency measures we choose a value of $m = 75$, which corresponds to a level of robustness at 10%. The plots lead roughly to the same conclusions on the effect of Z on the production process for the non-convex case (top panel) but for the robust order-m convex frontier estimators, the effect of Z is less clear to interpret: here some favorable effect is also detected for smaller values of Z. Since non-convex estimators are always consistent (even under convexity) but convex estimators are only consistent under convexity, this difference for the robust efficiency estimators should warn for potential non-convexities in the production process. This will be confirmed in the analysis of the indicators of convexity below.

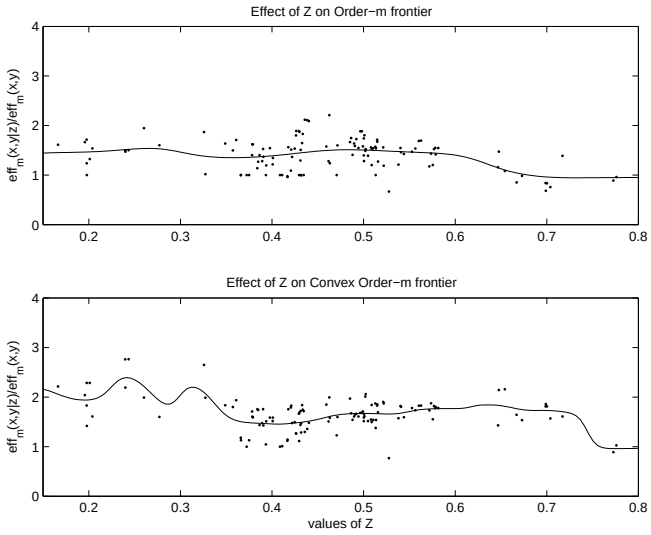

Figure 8: *US Aggressive Growth Mutual Funds data (input oriented framework). Scatter plot and smoothed regression of* $\hat{\theta}_{m,n}(x, y \mid z)/\hat{\theta}_{m,n}(x, y)$ *on* Z *(top panel), and of* $\hat{\theta}^{C}_{m,n}(x, y \mid z)/\hat{\theta}^{C}_{m,n}(x, y)$ *on* Z *(bottom panel).*

In this mutual funds example, an empirical investigation on convexity is indeed of great importance. This analysis could be useful to reveal the strategic behavior of mutual funds managers concerning the *substitutability* among the management dimensions: risks, turnover and transaction costs.

From a performance point of view, the knowledge of the *strategic behavior* adopted by funds in managing risks, turnover and transaction costs as *substitute* resources (disclosed by the verification of the convexity hypothesis) or as *non substitute* inputs (disclosed by

the refusing of the convexity hypothesis) could shed lights on the type of strategic goals pursued: mixed strategy (substitution) the latter, pure strategy (specialization) the former. In particular, a simple check might be done on the analyzed funds to see how the funds that apply a mixed strategy (*i.e.* use their inputs as substitutes, *i.e.* verify convexity) have performed compared with the funds that have specialized their management along some non substitutive combinations of inputs (as here we applied an input oriented framework).

To investigate convexity with this data set, we provide in Table 6, as for the simulated examples, some descriptive statistics of the different input efficiency measures and some information on the distribution of the indicators of convexity. This table deserves some comments:

- The efficiency scores for convex and non-convex technologies have the same order of magnitude when we look at their average, although their ratio is substantially lower than 1 in all the cases. The full distribution of the convexity indicators brings more information:
  - The distribution of IC is not very concentrated near 1 (only 23 % of observations - 30 over 129 - have values higher than 0.99). The robust version of the indicator, IC_m, is even less concentrated near 1: around 50 % of observations - 66 over 129 - have values larger than 0.85. Hence, in this exploratory phase, the assumption of convexity of the attainable set seems to be not confirmed.
  - The analysis of the distributions of ICZ and $ICZm$ (very similar), might indicate more convexity when looking at the attainable production sets, conditionally to the level of the market risks Z, since the distributions are more concentrated near 1 (more than 100 observations over 129 have the indicators ICZ and $ICZm$ greater than 0.90).
- All these comments are based on descriptive considerations. As a matter of fact, the observed differences may or may not be significant: this indicates the need for formal testing procedures (evaluation of p-values,...) particularly in the analysis of real data sets.

	$\hat{\theta}_{DEA,n}(x,y)$	$\hat{\theta}_{DEA,n}(x,y \mid z)$	$\hat{\theta}^{C}_{m,n}(x,y)$	$\hat{\theta}^{C}_{m,n}(x,y \mid z)$
Average	0.549	0.844	0.582	0.902
St. Dev.	0.170	0.164	0.178	0.135
Minimum	0.305	0.417	0.340	0.489
# Eff. Obs	6	21	6	39
	$\hat{\theta}_{FDH,n}(x,y)$	$\hat{\theta}_{FDH,n}(x,y \mid z)$	$\hat{\theta}_{m,n}(x,y)$	$\hat{\theta}_{m,n}(x,y \mid z)$
Average	0.608	0.888	0.687	0.904
St. Dev.	0.207	0.159	0.197	0.144
Minimum	0.310	0.417	0.362	0.453
# Eff. Obs	20	69	22	69
Indicators	IC	ICZ	IC_m	ICZ_m
$\# \geq 0.99$	30	56	12	82
$\# \geq 0.95$	84	69	27	96
$\# \geq 0.90$	93	101	44	107
$\# \geq 0.85$	106	123	66	121
$\# \geq 0.80$	110	127	89	123
$\# \geq 0.75$	114	129	110	124
$\# \geq 0.70$	116	129	120	124
$\# \geq 0.65$	121	129	123	126

Table 6: *Descriptive statistics of efficiency scores estimated over the 129 observations, for the US Aggressive Growth Mutual Funds data. Average, Standard Deviation, Minimum value, number of efficient observations, and distribution of the indicators of convexity.*

7 Conclusions

Motivated by the consideration that there exist empirical applications in which convexity could be reasonable we propose in this paper a conditional DEA estimator and a robust version of it based on the concept of order-m frontier. We describe also how these measures can be estimated and we address the problem of their practical computation. These newly introduced measures complete the exploratory tools available for gauging the performance of DMUs when extra information on operating environment are available.

We report also some indicators of convexity for several conditional and unconditional, full frontier and robust efficiency measures, extending previous indicators proposed in the literature. Finally, we illustrate all these concepts trough the analysis of some empirical examples: simulated and real data sets.

The analysis of the distributions of convexity estimators in the mutual funds example

shows that convexity is not clearly established and non-substitutability among the management dimensions (risks, turnover and transaction costs) might be at place in US Aggressive Growth funds. This illustration suggests that the convexity issue should be carefully taken into account in applied works. As a matter of fact, even when convexity could be reasonable from a theoretical point of view, its validity should be empirically checked and verified. Moreover, non-convex estimators are always consistent (even under convexity), whilst convex estimators are consistent only under the convexity assumption.

The indicators of convexity presented here, even if useful for descriptive and exploratory purpose, are not able to give a definitive answer about the convexity assumption of the corresponding attainable sets. In fact, the conclusions are drawn in terms of estimated technologies instead of true technologies. A statistical test procedure is requested to make inference with respect to the true technology. In other words, for a particular observation or for the global technology, without a formal testing procedure, it is impossible to determine if the values of the various indicators of convexity less than one are due to non convexity or due to sampling variation. Bootstrap techniques are the only way to perform these tests in a rigorous way. The implementation of the bootstrap should follow the lines of Simar and Wilson (2001, 2002).

Rigorous statistical procedures for testing convexity both in the traditional inputs-outputs representation of the production process and in the enlarged inputs-outputs-external factors framework are left for future development of this work.

References

[1] Bogetoft, P. (1996): "DEA on Relaxed Convexity Assumptions", *Management Science* 42, 457- 465.

[2] Bogetoft, P., J.M. Tama, and J. Tind (2000): "Convex Input and Output Projections of Nonconvex Production Possibility Sets", *Management Science* 46(6), 858-869.

[3] Briec, W., Kerstens, K. and P. Vanden Eeckaut (2004), "Non-convex technologies and cost functions: definitions, duality and nonparametric tests of convexity", *Journal of Economics*, Vol. 81 (2004), No. 2, pp. 155-192.

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

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

[6] Cooper, W.W., Seiford L.M., and Tone K. (1999), *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*,Kluwer Academic Publishers, Boston.

[7] Daraio, C. (2003), *Comparative Efficiency and Productivity Analysis based on Nonparametric and Robust Nonparametric Methods. Methodology and Applications*, Ph.D. Dissertation, Sant'Anna School of Advanced Studies, Pisa.

[8] Daraio C. and Simar, L. (2003), "Introducing environmental variables in nonparametric frontier estimation: a probabilistic approach", Discussion Paper no. 0313, Institut de Statistique, UCL, Belgium, and LEM WP no. 2003/17, *forthcoming* in *The Journal of Productivity Analysis*.

[9] Daraio C. and Simar, L. (2004a), "A robust nonparametric approach to evaluate and explain the performance of mutual funds", Discussion Paper no. 0412, Institut de Statistique, UCL, Belgium.

[10] Daraio C. and Simar, L. (2004b), "Introducing External Factors in Nonparametric and Robust Frontier Models: further investigations", Invited paper at the CORS-INFORMS Joint International Meeting, May 16-19 2004, Banff Alberta (Canada).

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

[12] 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.

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

[14] Farrell, M.J. (1959), "Convexity assumption in theory of competitive markets", *Journal of Political Economy*, 67, 377-391.

[15] Florens, J.P. and L. Simar, (2005), Parametric Approximations of Nonparametric Frontier, *Journal of Econometrics*, 124, 91–116.

[16] 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.

[17] 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.

[18] Mas-Colell, A., Whinston M.D., and Green, J.R. (1995), *Microeconomic Theory*, Oxford University Press.

[19] 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.

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

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

[22] Podinovski, V. V. (2004), "Selective convexity in DEA models", *forthcoming* in *European Journal of Operational Research*.

[23] Sheather S.J., and Jones M.C. (1991), "A relyable data-based bandwidth selection method for kernel density estimation", *Journal of the Royal Statistical Society*, Series B, 53:3, pp. 683-690.

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

[25] Simar, L. (2003), Detecting Outliers in Frontiers Models: a Simple Approach, *Journal of Productivity Analysis*, 20, 391–424.

[26] 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.

[27] Simar L. and P. Wilson (2001), Testing Restrictions in Nonparametric Efficiency Models, *Communications in Statistics, simulation and computation*, 30 (1), 159–184.

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

[29] 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.

Appendix: Bandwidth selection

DS propose a simple data-driven procedure for choosing the bandwidth, based on a k-nearest neighbor method, based on likelihood cross-validation for the density of Z.

In a *first step*, a bandwidth h which optimizes the estimation of the density of Z is selected, based on the likelihood cross validation criterion, using a k-NN (Nearest Neighborhood) method (see e.g. Silverman, 1986). 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.

Hence, 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 $i = 1, \ldots, n$ 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 $|Z_j - Z_i| \leq h_{Z_i}$.

In a *second step*, taking into account for the *dimensionality* of x and y, and the sparsity of points in larger dimensional spaces, the local bandwidths h_{Z_i} are expanded by a factor $1 + n^{-1/(p+q)}$, increasing with $(p+q)$ but decreasing with n. For more details, see Daraio (2003).

We notice that the calculations of efficiency scores and the evaluation of the influence of external factors is not too sensitive to the choice of the procedure for bandwidth selection. As a matter of fact, we obtained very similar results by applying the global bandwidth obtained with the Sheather and Jones (1991) method for kernel density estimation of Z. See Daraio and Simar (2004b), where a comparison of these bandwidth selection methods is proposed.