INSTITUT DE STATISTIQUE

BIOSTATISTIQUE ET

SCIENCES ACTUARIELLES

(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

2013/47

Statistical Approaches for Nonparametric Frontier Models: A Guided Tour

SIMAR, L. and P. WILSON

Statistical Approaches for Nonparametric Frontier Models: A Guided Tour

LÉOPOLD SIMAR   PAUL W. WILSON*

November 2013

Abstract

A rich theory of production and analysis of productive efficiency has developed since pioneering work by Koopmans (1951) and Debreu (1951). Farrell (1957) is the earliest published empirical study, and appeared in a statistical journal (*JRSS*), even though Farrell provided no statistical theory. The literature in econometrics, management sciences, operations research and mathematical statistics has since been enriched by hundreds of papers trying to develop or implement new tools for analyzing productivity and efficiency of firms. Both parametric and nonparametric approaches have been proposed. The mathematical challenge is to derive estimators of production, cost, revenue, or profit frontiers which represent, in the case of production frontiers, the optimal loci of combinations of inputs (like labor, energy, capital, etc.) and outputs (the products or services produced by the firms). Optimality is defined in terms of various economic considerations. Then the efficiency of a particular unit is measured by its distance to the estimated frontier. The statistical problem can be viewed as the problem of estimating the support of a multivariate random variable, subject to some shape constraints, in multiple dimensions. These techniques are applied in thousands of papers in the economic and business literature. This "Guided Tour" reviews the development of various nonparametric approaches since the early work of Farrell. Remaining challenges and open issues in this challenging arena are also described.

Keywords: Productivity, Efficiency, Data Envelopment Analysis (DEA), Free Disposal Hull (FDH), Nonparametric Frontiers, Boundary Estimation, Extreme Value Theory, Bootstrap
AMS2000 Subject Classification: 62-02, 62G05, 62G20

*Simar: Institut de Statistique, Biostatistique, et Sciences Actuarielles, Université Catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium; email leopold.simar@uclouvain.be.

Wilson: Department of Economics and School of Computing, 228 Sirrine Hall, Clemson University, Clemson, South Carolina 29634–1309, USA; email pww@clemson.edu.

Research support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) and from U.S. National Science Foundation (grant no. SMA-1243436) is gratefully acknowledged. Any remaining errors are solely our responsibility.

Contents

1 Introduction

Production theory and efficiency analysis examine how firms or production units in a particular sector of activity transform their inputs (e.g., labor, energy, capital, etc.) into quantities of outputs, i.e., the goods or services that are produced by the firms. The analysis is not limited to business firms such as manufacturing concerns, electricity plants, banks, for-profit hospitals, etc.; it is used to examine schools, universities, provision of public goods and services, non-profit organizations including hospitals and credit unions, etc. The efficient production frontier is defined in the relevant input-output space as the locus of the maximal attainable level of outputs corresponding to given levels of inputs. Alternatively, if prices of inputs are available, one can consider a cost frontier defined by the minimal cost of producing various levels of outputs. Intermediate cases are also possible; for example, one might consider maximization of quantities of a subset of outputs while minimizing quantities of some (perhaps all) inputs, holding other quantities fixed. In all cases, the problem amounts to estimating a boundary surface in the relevant space (e.g., the input-output space in the case of a production frontier, or the output-cost space in terms of a cost frontier) under shape constraints induced by some economic assumptions (e.g., monotonicity, concavity, etc.). The technical efficiency of a particular production plan (characterized geometrically by a point in the input-output space) is then determined by an appropriate measure of distance between this point and the optimal frontier. The background of the economic theory behind these analysis is due to Koopmans (1951) and Debreu (1951); see Shephard (1970) for a comprehensive presentation of the underlying economic theory.

In empirical studies, the attainable set in the input-output space is unobserved, and hence the efficiency of a given firm is also unknown. These quantities must be estimated from a sample of observed combinations of input and output quantities obtained from existing production units operating in the activity sector being studied. Many different approaches have been investigated in the literature, including statistical models of varying degrees of sophistication and ranging from fully parametric to fully nonparametric approaches. This literature has developed in a variety of academic fields, including economics, management and management science, operations research, econometrics, and statistics; in each case field, papers ranging from "very theoretical" to "very applied" can be found.

This "Guided Tour" focuses on statistical results obtained in the nonparametric branch of the literature, while stressing the inherent difficulty of the problem and solutions that have been developed. The tour begins in Section 2 by defining the basics of an economic model for production theory. The most popular nonparametric estimators, based on envelopment techniques, are then presented in Section 3. Statistical properties of the estimators and practical aspects of inference-making (mostly by bootstrap methods) are discussed in Section 4. Section 5 presents various extensions that have been proposed in the literature to address some of the inherent drawbacks of envelopment estimators (e.g., sensitivity to extreme data points and outliers). Section 6 shows how environmental factors that may influence the production process can be included in the analysis, allowing for heterogeneity. Finally, Section 7 briefly describes additional, interesting issues and challenges that remain open questions, including (i) how to use nonparametric methods to improve some parametric estimators, (ii) introducing noise in the observational process; (iii) testing issues; and (iv) nonparametric frontier models for panel data. The existing, first solutions to these problems are described, but more work is needed on these issues.

2 The Economic and The Statistical Paradigms

2.1 Production theory and efficiency scores

Following Koopmans (1951) and Debreu (1951), the production process can be described as follows. Let $x \in \mathbb{R}_+^p$ denote a vector of p input quantities and let $y \in \mathbb{R}_+^q$, denote a vector of q output quantities. The production set

$$\Psi = \left\{(x, y) \in \mathbb{R}_+^{p+q} \mid x \text{ can produce } y\right\} \tag{2.1}$$

describes the set of physically attainable points (x, y). For efficiency evaluation, the efficient boundary of Ψ, i.e., the technology, is of interest. This is defined by

$$\Psi^\partial = \left\{(x, y) \in \Psi \mid (\gamma^{-1}x, \gamma y) \notin \Psi \text{ for any } \gamma > 1\right\}. \tag{2.2}$$

Some minimal economic assumptions on Ψ are typically made. Most studies assume that inputs and outputs are freely (or strongly) disposable, i.e.,

$$\forall\ (x, y) \in \Psi,\ \text{ and any } (x', y') \text{ such that } x' \geq x \text{ and } y' \leq y,\ (x', y') \in \Psi. \tag{2.3}$$

This hypothesis assumes that it is always possible (even if not economically sound) to waste resources and implies monotonicity of the technology. Another mild assumption is that all production requires use of some positive input quantities (this is called the "no free lunch" assumption):

$$(x, y) \notin \Psi \text{ if } x = 0 \text{ and } y \geq 0,\ y \neq 0. \tag{2.4}$$

In addition, it is often assumed that the attainable set Ψ is convex so that if $(x_1, y_1),\ (x_2, y_2) \in \Psi$, then for all $\alpha \in [0, 1]$,

$$(x, y) = \alpha(x_1, y_1) + (1 - \alpha)(x_2, y_2) \in \Psi. \tag{2.5}$$

Other economic assumptions on Ψ (e.g., returns to scale) are sometimes made, but at this stage only those introduced above are needed.

The technical efficiency of a given production plan (x, y) can now be measured along the lines of Debreu (1951) and Farrell (1957). The *input* measure of technical efficiency $\theta(x, y)$ is given by the minimal radial contraction of the inputs to project the point (x, y) onto the efficient frontier Ψ^∂:

$$\theta(x, y) = \inf \{\theta \mid (\theta x, y) \in \Psi\}, \tag{2.6}$$

where for all points $(x, y) \in \Psi$, $\theta(x, y) \leq 1$. A value of 1 indicates an efficient point lying on the boundary of Ψ. Similarly, the *output*-oriented efficiency score $\lambda(x, y)$ is the maximal radial expansion of the outputs that projects the points (x, y) onto the efficient frontier, i.e.,

$$\lambda(x, y) = \sup \{\lambda \mid (x, \lambda y) \in \Psi\}. \tag{2.7}$$

Here, for all points $(x, y) \in \Psi$, $\lambda(x, y) \geq 1$. A value of 1 indicates an efficient point lying on the boundary of Ψ.

Other distance measures have been proposed in the economic literature, including hyperbolic distance

$$\gamma(x, y) = \sup \left\{\gamma > 0 \mid (\gamma^{-1} x,\ \gamma y) \in \Psi\right\} \tag{2.8}$$

due to Färe et al. (1985) (see also Färe and Grosskopf, 2004), where input and output quantities are adjusted simultaneously to reach the boundary along an hyperbolic path. Note $\gamma(x, y) = 1$ iff (x, y) belongs to the efficient boundary Ψ^∂. More recently, a lot of interest has been devoted to directional distances (Chambers et al., 1998; Färe and Grosskopf, 2000) due to their great flexibility. Here the projection of (x, y) onto the frontier is along a path in a given direction $d = (-d_x, d_y)$, where $d_x \in \mathbb{R}^p_+$ and $d_y \in \mathbb{R}^q_+$. The flexibility is due to the fact that some values of the direction vector can be set to zero. Distance from a point (x, y) to Ψ^∂ is then measured by

$$\delta(x, y \mid d_x, d_y) = \sup \{\delta \mid (x - \delta d_x, y + \delta d_y) \in \Psi\}, \tag{2.9}$$

where for all points $(x, y) \in \Psi$, $\delta(x, y) \geq 0$. A value of 0 indicates an efficient point lying on the boundary of Ψ. Note that as special case, the Farrell-Debreu radial distances can be recovered; e.g. if $d = (-x, 0)$, $\delta(x, y \mid d_x, d_y) = 1 - \theta(x, y)^{-1}$. Another interesting feature is that directional distances are additive measures, hence they permit negative values of x and y (e.g., in finance, an output y may be the return of a fund, which can be, and often is, negative). Long debates have been discussed in the economic literature on the directions should be chosen, and many choices are possible (e.g., a common one for all firms, or a specific direction for each firm); see Färe et al. (2008) for discussion.

It should be noticed that all these efficiency measures characterize the efficient boundary by measuring distance from a known, fixed point (x, y) to the unobserved boundary Ψ^{∂}; the only difference among the measures in (2.6)–(2.9) is in the direction in which distance is measured.

2.2 Returns to scale

Returns to scale is an important property of the technology Ψ^{∂}, and determines what happens as the scale of production is increased. Let $\mathcal{V}(\Psi)$ denote the convex cone of Ψ, and define

$$\mathcal{V}(\Psi)^{\partial} = \left\{(x, y) \in \mathcal{V}(\Psi) \mid (\gamma^{-1}x, \gamma y) \notin \mathcal{V}(\Psi) \text{ for any } \gamma > 1\right\} \tag{2.10}$$

analogous to (2.2). If $\mathcal{V}(\Psi) = \Psi$, then $\mathcal{V}(\Psi)^{\partial} = \Psi^{\partial}$ and Ψ^{∂} exhibits *globally* constant returns to scale (CRS).

Alternatively, for Ψ convex, if $\Psi \subset \mathcal{V}(\Psi)$, then the different regions of Ψ^{∂} may display *locally* either constant, increasing, or decreasing returns to scale. In this case, the subset of Ψ^{∂} given by $\left\{(x, y) \mid (x, y) \in \mathcal{P}^{\partial}\ (x, y) \in \mathcal{V}(\Psi)^{\partial}\right\}$ exhibits locally CRS (this subset may include a single point or perhaps many points). The subset of Ψ^{∂} given by

$$\left\{(x, y) \mid (x, y) \in \Psi^{\partial},\ (\alpha x, \alpha y) \in \Psi,\ (\alpha x, \alpha y) \notin \Psi^{\partial} \text{ for some } \alpha \in (1, \infty)\right\} \tag{2.11}$$

exhibits locally increasing returns to scale (IRS), while the subset of Ψ^{∂} given by

$$\left\{(x, y) \mid (x, y) \in \Psi^{\partial},\ (\alpha x, \alpha y) \in \Psi,\ (\alpha x, \alpha y) \notin \Psi^{\partial} \text{ for some } \alpha \in [0, 1)\right\} \tag{2.12}$$

exhibits locally decreasing returns to scale (DRS). If Ψ^{∂} has different regions that display IRS, CRS, and DRS, then Ψ^{∂} is said to be of varying returns to scale (VRS).

Along the IRS portion of a technology, a small increase in input usage allows a greater-than-proportionate increase in output quantities produced. By contrast, along the DRS portion of a technology, a small decrease in input usage requires only a less-than-proportionate decrease in output quantities. In this sense, absent other considerations, the CRS portion of Ψ^{∂} is the optimal part of the technology; i.e., the CRS portion of Ψ^{∂} corresponds to the most productive scale of operation (see Banker, 1984 for discussion).

2.3 Statistical modeling

The concepts introduced above are useful in theory, but in practice, the attainable set Ψ and its boundary Ψ^{∂} are unknown, as are the efficiency scores. The best an empirical researcher can hope to do is to estimate these from a random sample of input-output combinations $\mathcal{X}_n = \{(X_i, Y_i)\}_{i=1}^{n}$. Of course, a well-defined statistical model—a description of the data generating process (DGP)—describing how the random sample is generated is needed before anything can be estimated.

There are two main streams of thought in the extant literature: (i) so-called *deterministic frontier models* (this wording is unfortunate because nothing is deterministic), and (ii) *stochastic frontier models.* In deterministic models, it is assumed that all observations in the sample belong to the attainable set:

$$\Pr\left((X, Y) \in \Psi\right) = 1. \tag{2.13}$$

This may be seen as a reasonable assumption, but it implies that no noise (e.g., measurement error) is admitted in the DGP. In these models, distance to the frontier is interpreted as pure inefficiency. Alternatively, stochastic frontier models permit noise in the DGP, so some observations may lie outside Ψ. This is appealing for its flexibility, but the result is that distance to the frontier has two components, noise and inefficiency, and hence identification becomes problematic, requiring additional assumptions that may reduce flexibility.

Another classification of the approaches concerns the chosen level of modeling: either parametric models or nonparametric models may be used. Parametric models are rather restrictive since they rely on particular functional form both for the frontier and for the two elements of the stochastic parts of the model (i.e. the distribution of the inefficiency and, for stochastic frontier models, the distribution of the noise). Basic references for parametric, deterministic models include Aigner and Chu (1968) and Greene (1980); for parametric, stochastic models, see Aigner et al. (1977), Battese and Corra (1977), Meeusen and van den Broeck (1977), Olson et al. (1980), and Jondrow et al. (1982).

To illustrate, consider the case of one output Y and a vector of inputs X. One of the simplest specification for the production frontier function may be

$$Y_i = \beta_0 + \beta' X_i + V_i - U_i, \tag{2.14}$$

where $V_i \sim N(0, \sigma_V^2)$ is random noise and $U_i \sim N^+(0, \sigma_U^2)$ represents inefficiency ($U_i \geq 0$; in a deterministic model, V_i would not appear), with $V_i \perp\!\!\!\perp U_i \ \forall \ i = 1, \ \ldots, \ n$. If the variables are in the log scale, (2.14) gives the familiar Cobb-Douglas production function. Clearly, this model rests on some very specific hypotheses. Maximum likelihood or modified ordinary least squares (OLS) can be used to estimate the parameters, but inefficiency U_i is not identified. Once parameters have been estimated, only the convoluted residual $\widehat{\varepsilon}_i = Y_i - \widehat{\beta}_0 + \widehat{\beta}' X_i$ is observed, presenting a deconvolution problem to estimate the inefficiency and the noise parts of $\widehat{\varepsilon}_i$. Jondrow et al. (1982) suggest estimating individual efficiency by an estimate of $E(U_i \mid \widehat{\varepsilon}_i)$. Various numerical problems arise in the estimation of these models, and inference about inefficiency presents additional problems. Simar and Wilson (2010) suggest bootstrap and bagging methods for making inference on the parameters and on the individual efficiencies in such models.

The parametric routes in this Guided Tour are not pursued; the interested reader can refer to Kumbhakar and Lovell (2000) and Greene (2008). Hereafter the focus is on nonparametric frontier models, which share the very attractive property of relying only on mild assumptions suggested by economic theory, such as those in (2.3), (2.4), and in some cases in (2.5). In addition, except for some mild regularity conditions, no parametric restrictions will be imposed on the distribution of (X, Y) on Ψ. To date, mainly deterministic frontier models have been developed in these nonparametric approaches, because the identification problem that arises when noise is added becomes much more difficult to handle in a nonparametric framework. Later sections will describe how nonparametric deterministic models can be modified to introduce some noise, and to be resistant to outliers or extreme data points. In addition, Section 7.2 briefly discusses some new lines of research aimed at allowing for noise in nonparametric frontier models.

3 The Nonparametric Envelopment Estimators

Historically, Farrell (1957) is the pioneering, first empirical work to estimate an attainable set enveloping the cloud of data points and the resulting efficiency scores. This has been popularized by by Charnes et al. (1978) and Banker et al. (1984) using linear programming techniques. These works rely on the convexity assumption (2.5) for Ψ and various returns-to-scale assumptions. Estimates without imposing convexity on Ψ came later in Afriat (1972) and Deprins et al. (1984). For simplification, the presentation below starts with the latter. For many years, estimators of efficiency revolved around the Farrell-Debreu radial measures. As seen below, these have more recently been extended to hyperbolic and directional distances. Introductory textbooks on these topics include Thanassoulis (2001) and Cooper et al. (2011). Fried et al. (2008) present a more advanced, comprehensive picture of the topic, including parametric approaches.

3.1 Free Disposal Hull (FDH) Estimators

The FDH estimator, relying only on the free disposal assumption (2.3), was proposed by Deprins et al. (1984). The FDH estimator of Deprins et al. (1984) is able to handle full multivariate inputs and outputs and can be used to estimate distances functions in any direction.[1]

[1] Afriat (1972) used similar ideas earlier to define a left-continuous monotone production function, but only for the case of univariate output and freely disposable inputs. In addition, his approach only allowed efficiency to be measured in the output direction.

The idea is very simple; an estimator

$$\widehat{\Psi}_{FDH}(\mathcal{X}_n) = \left\{(x,y) \in \mathbb{R}_+^{p+q} \mid y \leq Y_i,\ x \geq X_i,\ (X_i,Y_i) \in \mathcal{X}_n\right\}. \tag{3.1}$$

of Ψ is defined by considering the union of all the orthants (positive in x and negative in y) having their vertex at the observed data points. FDH estimators of efficiency are obtained by plugging the estimator (3.1) into the definitions (2.6)–(2.9). For example, in the case of a particular production plan (x,y), substituting $\widehat{\Psi}_{FDH}(\mathcal{X}_n)$ for Ψ in (2.6) yields

$$\widehat{\theta}_{\text{FDH}}(x,y) = \inf\left\{\theta \mid (\theta x, y) \in \widehat{\Psi}_{FDH}(\mathcal{X}_n)\right\}. \tag{3.2}$$

Estimators $\widehat{\lambda}_{\text{FDH}}(x,y)$, $\widehat{\gamma}_{\text{FDH}}(x,y)$, and $\widehat{\delta}_{\text{FDH}}(x,y)$ are obtained by similar substitutions. Note that the notation $\widehat{\Psi}_{FDH}(\mathcal{X}_n)$ in (3.1) makes clear that this estimator depends on the random sample $\mathcal{X}_n$.

FDH estimators are fast and easy to compute using only sorting algorithms. A point $(\widetilde{x},\widetilde{y}) \in \Psi$ is said to *dominate* another point $(x,y) \in \Psi$ if $\widetilde{x} \leq x$ and $\widetilde{y} \geq y$. Let $D_{x,y}$ denote the set indices of points in $\mathcal{X}_n$ dominating (x,y), i.e., $D_{x,y} = \{i \mid (X_i,Y_i) \in \mathcal{X}_n,\ X_i \leq x,\ Y_i \geq y\}$. Then

$$\widehat{\theta}(x,y) = \min_{i \in D_{x,y}} \max_{j=1,\ldots,p} \left(\frac{X_i^j}{x^j}\right), \tag{3.3}$$

where for a vector a, a^j denotes its j-th component. The estimator $\widehat{\lambda}_{\text{FDH}}(x,y)$ is computed similarly; see Simar and Wilson (2013) for details. Wilson (2011) gives an algorithm for computing $\widehat{\gamma}_{\text{FDH}}(x,y)$, and Simar and Vanhems (2012) provide a simple way for computing $\widehat{\delta}_{\text{FDH}}(x,y \mid d_x,d_y)$ when all elements of the direction vectors are strictly positive; if this is not the case, algorithms given by Daraio and Simar (2013) can be used.

Jeong and Simar (2006) propose a Linearized version of the FDH estimator (LFDH), denoted by $\widehat{\Psi}_{\text{LFDH}}(\mathcal{X}_n)$, to avoid the unpleasant step-function nature of the FDH estimator. The neighboring vertices of the FDH solution are interpolated by an hyperplane, so that $\widehat{\Psi}_{\text{FDH}}(\mathcal{X}_n) \subseteq \widehat{\Psi}_{\text{LFDH}}(\mathcal{X}_n)$. The vertices to be interpolated are identified using Delaunay triangulation methods (or tessellation; e.g., see Brown, 1979), then linear programming methods identify the supporting hyperplane. Asymptotic properties of the LFDH estimators are the same as those of the FDH estimators described below.

3.2 Data Envelopment Analysis (DEA) Estimators

The DEA estimator introduced by Farrell (1957) and popularized by Charnes et al. (1978) is based on the additional assumption that Ψ is convex and allows for VRS. Farrell's estimator of Ψ is the convex hull of the FDH estimator $\widehat{\Psi}_{FDH}(\mathcal{X}_n)$ given by

$$\widehat{\Psi}_{\text{VRS}}(\mathcal{X}_n) = \left\{(x, y) \in \mathbb{R}_+^{p+q} \mid y \leq \sum_{i=1}^{n} \gamma_i Y_i,\ x \geq \sum_{i=1}^{n} \gamma_i X_i,\ \sum_{i=1}^{n} \gamma_i = 1,\ \gamma_i \geq 0\ \forall\ i = 1,\ \ldots,\ n\right\}. \tag{3.4}$$

The multipliers γ_i and the constraint $\sum_{i=1}^{n} \gamma_i = 1$ serve to "convexify" the free disposal hull of the sample observations (X_i, Y_i).

VRS-DEA estimators of the efficiency measures defined in (2.6)–(2.9) are obtained by substituting $\widehat{\Psi}_{VRS}(\mathcal{X}_n)$ for Ψ in the definitions; in the case of the input-oriented measure $\theta(x, y)$, the resulting VRS-DEA estimator is given by the linear program

$$\widehat{\theta}_{\text{VRS}}(x, y) = \min_{\substack{\theta \\ \gamma_1,\ \ldots,\ \gamma_n}} \left\{\theta \mid y \leq \sum_{i=1}^{n} \gamma_i Y_i,\ \theta x \geq \sum_{i=1}^{n} \gamma_i X_i,\ \sum_{i=1}^{n} \gamma_i = 1,\ \gamma_i \geq 0\ \forall\ i = 1,\ \ldots,\ n\right\}. \tag{3.5}$$

The VRS-DEA estimator of $\lambda(x, y)$ is given by the linear program

$$\widehat{\lambda}_{\text{VRS}}(x, y) = \max_{\substack{\lambda \\ \gamma_1,\ \ldots,\ \gamma_n}} \left\{\lambda \mid \lambda y \leq \sum_{i=1}^{n} \gamma_i Y_i,\ x \geq \sum_{i=1}^{n} \gamma_i X_i,\ \sum_{i=1}^{n} \gamma_i = 1,\ \gamma_i \geq 0\ \forall\ i = 1,\ \ldots,\ n\right\}. \tag{3.6}$$

The VRS-DEA estimator $\widehat{\delta}_{\text{VRS}}(x, y \mid d_x, d_y)$ of $\delta(x, y \mid d_x, d_y)$ can also be computed as the solution to a linear program; see Simar et al. (2012) for details. Wilson (2011) gives a numerical algorithm for computing the hyperbolic estimator $\widehat{\gamma}_{\text{VRS}}(x, y)$, which cannot be written as a linear program.

If Ψ^{∂} is globally CRS, then Ψ can be estimated by the convex cone of $\widehat{\Psi}_{\text{FDH}}(\mathcal{X})$ obtained by dropping the constraint $\sum_{i=1}^{n} \gamma_i = 1$ in (3.4); denote this estimator by $\widehat{\Psi}_{\text{CRS}}(\mathcal{X})$. Substituting $\widehat{\Psi}_{\text{CRS}}(\mathcal{X})$ for Ψ in (2.6)–(2.9) yields CRS-DEA estimators of the efficiency measures defined there. Computation of the corresponding CRS-DEA estimators is similar to computation of their VRS-DEA counterparts; the input- and output-oriented estimators $\widehat{\theta}_{\text{CRS}}(x, y)$ and $\widehat{\lambda}_{\text{CRS}}(x, y)$ are computed by dropping the constraint $\sum_{i=1}^{n} \gamma_i = 1$ in (3.5)–(3.6). The CRS-DEA directional estimator $\widehat{\delta}_{\text{CRS}}(x, y \mid d_x, d_y)$ is computed by dropping the same constraint from the linear program that appears in Simar et al. (2012), an the algorithm appearing in Wilson (2011) is easily modified to compute the hyperbolic estimator $\widehat{\gamma}_{\text{CRS}}(x, y)$.

The various DEA estimators that have been discussed can be adapted to handle cases where Ψ^{∂} exhibits either increasing and constant, but not decreasing returns to scale, or constant and decreasing, but not increasing returns to scale. Details are given in Simar and Wilson (2002, 2013) and Banker et al. (2004). For purposes of this review, only the VRS-DEA and CRS-DEA estimators are needed.

Software for computing the nonparametric estimators—both FDH and DEA—is available in the *FEAR* library described by Wilson (2008) for use with *R*. The software is freely available for academic use as described in the license that accompanies *FEAR*.

4 Statistical Inference using DEA/FDH Estimators

The nonparametric envelopment estimators described above in Section 3 have been used in several thousands of papers without acknowledging the fact that the resulting DEA/FDH efficiency scores computed in these papers are in fact estimators.[2] For a given sample $\mathcal{X}_n$, the various FDH and DEA efficiency estimators discussed above provide only point estimates of unknown quantities. It seems that until recently, most researchers using these estimators were unaware, and unconcerned, by the statistical properties of these estimators.

Until recently, much of the DEA/FDH literature was concentrated in the field of Operations Research; DEA/FDH techniques were considered as non-statistical or non-econometric by those working in Statistics or Econometrics, who focused primarily on parametric approaches to efficiency analysis. By now, however, this has changed; the "two worlds" have been largely unified by recent theoretical results. After the original empirical work of Farrell (1957), Afriat (1972) and Deprins et al. (1984), Banker (1993), Korostelev et al. (1995a, 1995b) and Simar (1992, 1996) were the first to consider the FDH/DEA procedures from a statistical viewpoint (including presentation of a well-defined statistical model). This section summarizes most of the results available today. To streamline the discussion, only results for the radial, input-oriented Farrell-Debreu efficiency measures are presented. These results extend trivially to the output-orientation with some (perhaps tedious) changes in notation. Extension to the hyperbolic and directional measures is given by Wilson (2011), Simar and Vanhems (2012), and Simar et al. (2012). A comprehensive and informative survey covering all these cases can be found in Simar and Wilson (2013).

4.1 Asymptotic properties

As noted earlier, the notation introduced in Section 3 makes clear that $\widehat{\Psi}_{FDH}(\mathcal{X}_n)$ and $\widehat{\Psi}_{DEA}(\mathcal{X}_n)$ are function of the random sample $\mathcal{X}_n$. Consequently, all of the efficiency estimators discussed so far necessarily measure efficient relative to the boundary of an *estimate* of the attainable set.

[2] Seiford (1996) provides a survey of more than 700 published papers using DEA/FDH techniques; Cooper et al. (2000) similarly cite roughly 1,500 references; and Gattoufi et al. (2004) provide more than 1,800 references. A search on Google Scholar using the keywords ”efficiency,” ”production,” and either ”dea” or ”fdh” yielded about 102,000 papers (many are unpublished working papers) on 1 November 2013. Almost all of these papers ignore any statistical considerations.

Consistency

The first results concern minimal, but essential properties: statistical consistency and achieved rates of convergence. For a long time, the only available results were for cases where where either inputs or outputs were unidimensional. For $p = 1$ and $q \geq 1$, Banker (1993) establishes consistency of $\widehat{\theta}_{DEA}(x, y)$ for convex sets Ψ, but provides no information on the rate of convergence. The first systematic analysis of convergence of envelopment estimators appears in Korostelev et al. (1995a, 1995b).[3]

For the case $p = 1$ and $q \geq 1$, Korostelev et al. (1995a) prove that under the free disposability assumption, and the hypothesis that the joint density of (X, Y) on Ψ is uniform, the FDH estimator of Ψ is the maximum likelihood estimator, is consistent, and achieves the optimal rate

$$d_H(\widehat{\Psi}_{FDH}(\mathcal{X}_n), \Psi) = O_p\left((n/\log n)^{-1/(1+q)}\right), \tag{4.1}$$

where $d_H(\cdot, \cdot)$ denotes the Hausdorff metric between the two sets. Korostelev et al. also describe a "blown-up" version of the FDH estimator reaching asymptotically the optimal mini-max risk.

Korostelev et al. (1995b) relax the uniform distribution and consider both FDH and DEA estimators, again with $p = 1$ and $q \geq 1$. Here the risk of the estimators are defined in terms of $d_{\mathcal{L}}$, the Lebesgue measure of the symmetric difference between sets. Under free disposability assumption (but not convexity of Ψ), Korostelev et al. (1995b) obtain

$$d_{\mathcal{L}}(\widehat{\Psi}_{FDH}(\mathcal{X}_n), \Psi) = O_p\left(n^{-1/(1+q)}\right), \tag{4.2}$$

where $d_{\mathcal{L}}$ denotes the Lebesgue measure of the symmetric difference between two sets. Adding the assumption of convexity of Ψ, they obtain

$$d_{\mathcal{L}}(\widehat{\Psi}_{DEA}(\mathcal{X}_n), \Psi) = O_p\left(n^{-2/(2+q)}\right). \tag{4.3}$$

Korostelev et al. (1995b) show also that under their respective frameworks, both the FDH and the DEA estimators converge with the best possible rates in the class of monotone boundaries for FDH and the class of monotone and concave boundaries for DEA.

These results also reveal for the first time that the nonparametric envelopment estimators suffer from the "curse of dimensionality" shared by most nonparametric techniques. This was new, even if not a surprise in the statistical world; FDH and DEA are consistent and share the best possible rates in their respective class, but ever more data are needed as the the dimensionality of the problem increases.

[3] Note that Afriat (1972), in the univariate output case, also introduced some statistical methods in his study, by considering a parametric, beta distribution for inefficiency, and using maximum likelihood for parameter estimation. However, no statistical properties are established in this paper.

A bit later, Kneip et al. (1998) (for the DEA case) and Park et al. (2000) (for the FDH case) derive the rates of convergence for the efficiency estimators in the more general setting of multivariate inputs and outputs. For the FDH estimators, only the free disposability assumption is needed; for the DEA estimators, the convexity assumption is required. Under some regularity assumptions (e.g., smoothness of the frontier, continuity of the density of (X, Y) near the boundary, and $f(x, y)$ strictly positive on the frontier), they obtain for any fixed point $(x, y) \in \Psi$

$$\widehat{\theta}_{\bullet}(x,y) - \theta(x,y) = O_p\left(n^{-\tau}\right), \tag{4.4}$$

where "$\bullet$" represents either FDH or VRS and $\tau = 1/(p+q)$ for the FDH case (only under the free disposability assumption) and $\tau = 2/(p+q+1)$ for the VRS-DEA case (adding the convexity assumption). Adding the assumption that Ψ^∂ is globally CRS, Park et al. (2010) prove that the rate of the CRS-DEA estimator improves to $\tau = 2/(p+q)$. Interestingly, in this case, whenever $p + q \leq 4$, the rate of the nonparametric estimator $\widehat{\theta}_{\text{CRS}}(x, y)$ is faster than the usual $\sqrt{n}$ parametric rate, but only under globally CRS. Even more recently, Kneip et al. (2013b) prove that the VRS-DEA estimator also achieves the rate with $\tau = 2/(p+q)$ when Ψ^∂ is globally CRS.

The rate in (4.4) for the input-oriented Farrell-Debreu radial score also holds for the output-oriented case after straightforward changes in notation. Wilson (2011) establishes the same rates for the hyperbolic case, and Simar and Vanhems (2012) and Simar et al. (2012) establish the same rates for the directional case.

Asymptotic law

Statistical consistency is a fundamental, essential property of any estimator, but for inference, sampling distributions or their approximations are needed. The FDH case is easier because it is linked to the estimation of a maximum or a minimum of the support of some appropriate random variable. The first result is due to Park et al. (2000), who shown that under mild regularity conditions,

$$n^{1/(p+q}\left(\widehat{\theta}_{FDH}(x,y) - \theta(x,y)\right) \xrightarrow{\mathcal{L}} \text{Weibull}\left(\mu^{p+q}, p+q\right). \tag{4.5}$$

Park et al. describe the parameter μ of the limiting Weibull and show that μ^{p+q} is the probability of observing a firm dominating the point $(x^\partial + \zeta, y)$ for small ζ, where (x^∂, y) as the reference frontier point of (x, y) in the input direction such that $x^\partial = \theta(x, y)x \leq x$. Park et al. suggest a consistent estimator of μ and a way of selecting the "smoothing" parameter ζ in simulated samples, but in practice this parameter is difficult to implement. Hence bootstrap techniques are an attractive alternative. Bădin and Simar (2009) propose a simple way to correct the inherent bias of the FDH estimator in finite samples.

In the particular case of one input in the input orientation (or of one output in the output orientation), Daouia et al. (2010) use results from Extreme Value Theory (EVT) to extend the above result in cases where the density of (X, Y) smoothly tends to zero as, for example, in the univariate, input orientation, $x \to x^{\partial}$. The result is similar, but as expected the rate of convergence deteriorates by the speed (the number of derivatives converging to zero) with which this density approaches zero at the frontier point. Daouia et al. also examine the case where the density of (X, Y) tends to infinity when approaching the frontier, with the reverse effect (i.e., improving the rate of convergence). Recently, Daouia et al. (2013) generalize these results from EVT to the full multivariate setting by considering the behavior of the joint density of X and Y near the boundary along the appropriate ray (e.g., x for the radial oriented measure), and obtain similar results.

The DEA cases are much more difficult to analyze, because the efficient frontier is determined by a facet of a convex polyhedra (for the VRS case) or of a convex cone (for the CRS case). Results from EVT cannot be directly applied. The main results are Kneip et al. (2008) for the VRS-DEA case, and Park et al. (2010) for the CRS-DEA case. As usual, for the VRS-DEA case, both free disposability and convexity of Ψ are needed; for the CRS-DEA case, the additional assumptions of global CRS is required. Under analog mild regularity conditions (in particular, strict positivity of the joint density on the frontier),

$$n^{\tau}\left(\widehat{\theta}_{\bullet}(x,y) - \theta(x,y)\right) \xrightarrow{\mathcal{L}} Q_{\bullet}(\eta), \tag{4.6}$$

where "$\bullet$" now denotes either VRS or CRS, and where $Q_{\bullet}(\cdot)$ is some regular, non-degenerate distribution with unknown parameters η depending on unknown quantities characterizing the DGP. Kneip et al. (2008) analyze the VRS case, with $\tau = 2/(p+q+1)$, and Park et al. (2010) examine the CRS case, with the faster rate $\tau = 2/(p+q)$. There are no explicit closed form for the limiting distribution in either case. Jeong (2004) and Jeong and Park (2006) suggest a way to simulate the distribution in the univariate input case, and Park et al. (2010) provide a way to simulate the distribution in the multivariate CRS-DEA case, but these methods involve some smoothing parameters and are difficult to implement in practice. See Jeong and Park (2011) for a theoretical survey of such methods. Clearly, for the DEA estimators, bootstrap techniques are very useful.

Note that in the particular case of one input and one output, Gijbels et al. (1999) provide a closed-form expression of the asymptotic distribution of the VRS-DEA estimator; in the same simple setting, Park et al. (2010) show that the CRS-DEA estimator has a simple exponential distribution. Interestingly, Kuosmanen (2008) finds that for the univariate output case, the DEA estimator can be obtained by solving a convex nonparametric least-squares problem which as an equivalent representation in term of a quadratic programming problem, where the objective is quadratic but with linear constraints.

Note that other approaches (e.g., Hall et al., 1998) use estimators of boundaries without imposing the natural economic constraints (like monotonicity or concavity); hence these are less popular than DEA/FDH estimators in the field of productivity and efficiency analysis.

4.2 Bootstrap techniques

As noted above, there are several difficulties associated with the practical use of the asymptotic results described in Section 4.1 for making inference. So far, bootstrap methods seem to be the only viable alternative for making inference on $\theta(x, y)$.

The first suggestion to use of bootstrap methods in the context of production and efficiency analysis appears in Simar (1992) in a panel-data setting, but without any theoretical justification. The first theoretical results for using bootstrap in a frontier setup appear in Hall et al. (1995), where consistency of the bootstrap is established for the particular case of a semiparametric panel model; a double bootstrap is recommended to improve performance of the approximation.

The first study suggesting bootstrap techniques for assessing the sampling variability of the VRS-DEA efficiency estimator in a fully nonparametric frontier model is Simar and Wilson (1998). The procedure is rather simple: the idea is to generate a bootstrap sample $\mathcal{X}_n^*$ from $\mathcal{X}_n$ in an appropriate way. Then, efficiency for any point (x, y) of interest is evaluated relative to a bootstrap estimate $\widehat{\Psi}_{VRS}(\mathcal{X}_n^*)$ to obtain the corresponding bootstrap value of the efficiency score, say $\widehat{\theta}^*_{VRS}(x, y)$ for the (input) radial distance to the boundary of $\widehat{\Psi}_{DEA}(\mathcal{X}_n^*)$. If the bootstrap is consistent, then as $n \to \infty$,

$$n^{2/(p+q+1)} \left(\widehat{\theta}^*_{VRS}(x, y) - \widehat{\theta}_{VRS}(x, y) \right) \overset{\text{approx.}}{\sim} n^{2/(p+q+1)} \left(\widehat{\theta}_{VRS}(x, y) - \theta(x, y) \right), \tag{4.7}$$

where Monte-Carlo replications of the left hand side can be used to approximate the unknown right hand side.

The main problem is how $\mathcal{X}_n^*$, i.e., the bootstrap sample of size n, should be generated so that (4.7) holds. It is well-known from the statistical literature (e.g., Bickel and Freedman, 1981) that the naive bootstrap (i.e., resampling with replacement from the pairs (X_i, Y_i) in $\mathcal{X}_n$) is not consistent due to the unknown boundary of Ψ which is the support of (X, Y) (see, for example, Simar and Wilson, 2011a for a pedagogical explanation of the problem). This fact was not recognized by some in the frontier literature as indicated by the debate in Simar and Wilson (1999c, 1999b).

Rather than using the inconsistent naive bootstrap, Simar and Wilson (1998) propose using a smooth bootstrap. In this first study, Simar and Wilson implement the smoothed bootstrap in a simple model under the assumption that the distribution of the inefficiencies along the chosen direction (input rays or output rays) is homogeneous in the input-output space. Hence the smoothing operates only on the estimation of the univariate density of the efficiencies, making the problem much easier to handle. Simar and Wilson (2000) extend this idea to a more general heterogeneous case where the distribution of efficiency is allowed to vary over Ψ. This requires more complication than the original procedure, and involves the estimation of a smoothed density of (X, Y) with unknown support in a $(p+q)$-dimensional space. No theoretical justification was given for either approach, but results from intensive Monte-Carlo experiments described in both papers suggest that these bootstrap procedures give reasonable approximations for correcting the bias of the efficiency estimates and for building individual confidence intervals for the efficiency of any fixed point (x, y).

The full theory on the asymptotic properties of the VRS-DEA estimator and of the bootstrap is established in Kneip et al. (2008). Here, two bootstrap techniques are proven to be consistent: (i) a double-smooth bootstrap where in addition to smoothing the empirical distribution of the data, the support of Ψ is estimated by a smoothed version of the VRS-DEA estimator; and (ii) a subsampling bootstrap.

The double-smooth bootstrap developed by Kneip et al. (2008) involves numerical difficulties making it difficult to implement and computationally demanding. Kneip et al. (2011) provide a simplified, consistent, and computationally-efficient version of the double-smooth bootstrap. The idea is rather simple. It is well-known that the naive bootstrap does not work, but the problem is localized to points near the boundary. The idea behind the simplified Kneip et al. (2011) method is to draw naively among observations which are "far" from the frontier and draw the remaining points from a uniform distribution with support "near" the frontier. This neighborhood of the frontier is tuned by a smoothing parameter that can be selected by simple rule of thumb. For obtaining consistency, the VRS-DEA frontier estimate must be smoothed, and here a second bandwidth parameter is selected by cross-validation methods.

The subsampling approach is much more simple to implement, since a bootstrap sample $\mathcal{X}_m^*$, where $m = n^\gamma$ for some $\gamma \in (0, 1)$, is obtained by drawing with (or without) replacement m pairs (X_i, Y_i) from the original sample $\mathcal{X}_n$. Kneip et al. (2008) prove consistency of subsampling, but do not provide suggestions for how a value for m might be selected in practice. Their simulation results indicate that performance of the subsampling bootstrap in terms of achieved coverages of estimated confidence intervals is quite sensitive to the choice of m.

Unless the convexity assumption is imposed on Ψ, the FDH or the LFDH estimators must be used, as the DEA estimators are inconsistent without convexity of Ψ. The limiting Weibull distribution is not easy to use since it contains an unknown parameter that is not easy to estimate in practice. Jeong and Simar (2006) prove that the subsampling provides a consistent approximation of the sampling distribution of FDH and LFDH estimators, but here again, no practical advice is offered on how to select an appropriate subsample size.

Simar and Wilson (2011a), using results from Politis et al. (2001) and Bickel and Sakov (2008), provide a data-based algorithm for selecting an appropriate value of the subsample size m, for both the FDH and DEA cases. The idea is to compute the object of interest (e.g., bounds of a confidence interval, or bias estimate) for various values of m on some selected grid. Then the value of m where the results show the smallest volatility is selected. This volatility can be computed for each value of m in the grid by computing, for example, the standard deviation between the 3 or 5 values found for the adjacent values of m. Simar and Wilson (2011a) investigate the performance of their method (in terms of achieved coverages of individual confidence intervals for efficiency scores) by intensive Monte-Carlo experiments, for both FDH and DEA estimators. The results indicate that the method works well for moderate sample sizes faced in practice, providing reasonable approximations of the sampling distribution of the estimators.[4]

5 Robust Versions of Envelopment Estimators

5.1 Probabilistic formulation of the production process

In an innovative paper, Cazals et al. (2002) focus on the probabilistic structure of the DGP. Doing so permits the efficiency measures defined in (2.6)–(2.9) to be reformulated and suggests alternative features that may be estimated. Clearly, the DGP is completely characterized by the bounded joint density of (X, Y), or any one-to-one transformation of it. It is convenient to characterize the joint probability law of (X, Y) by

$$H_{XY}(x, y) = \Pr(X \leq x,\ Y \geq y), \tag{5.1}$$

which is the probability of observing a firm dominating the production plan (x, y). The attainable set Ψ is the support of H_{XY}. This joint distribution can be decomposed by writing

$$H_{XY}(x, y) = \Pr(X \leq x \mid Y \geq y)\ \Pr(Y \geq y) = F_{X|Y}(x \mid y)\ S_Y(y), \tag{5.2}$$

while noting that the conditional distribution function (DF) $F_{X|Y}(x \mid y)$ is non-standard since the conditioning on Y is not $Y = y$ or $Y \leq y$ but instead $Y \geq y$.

Remarkably, as shown by Daraio and Simar (2007b) by extending Cazals et al. (2002) to multivariate settings, under the free disposability assumption the Farrell-Debreu input-oriented efficiency measure in (2.6) can be defined equivalently as

$$\theta(x, y) = \inf\left\{\theta \mid F_{X|Y}(\theta x \mid y) > 0\right\}. \tag{5.3}$$

[4] Note that Jeong and Park (2011) suggest an alternative way to select a subsample size, for special case of a univariate output when output-efficiency is estimated.

Then obvious nonparametric estimator of $\theta(x, y)$ is obtained by replacing $F_{X|Y}(\theta x \mid y)$ in (5.3) with its empirical analog, i.e.,

$$\widehat{F}_{X|Y,n}(x \mid y) = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i \leq x,\ Y_i \geq y)}{\sum_{i=1}^{n} \mathbb{1}(Y_i \geq y)}. \tag{5.4}$$

Then for any y in the support of Y, simple manipulations reveal that

$$\widehat{\theta}(x, y) = \inf \left\{ \theta \mid \widehat{F}_{X|Y,n}(\theta x \mid y) > 0 \right\} = \min_{i|Y_i \geq y} \ \max_{j=1,\ldots,p} \left(X_i^j / x^j \right), \tag{5.5}$$

which can be shown to be equal to the expression given in (3.3) for the FDH estimator. This gives an additional natural motivation for the FDH estimators. Similar results are obtained for the output orientation by writing $H_{XY}(x, y) = S_{Y|X}(y \mid x) F_X(x)$ where the conditional survival function $S_{Y|X}(y \mid x)$ is also non-standard since the conditioning event is $X \leq x$ instead of $X = x$ or $X > x$.

This nice and simple idea opens the door to a number of other developments such as the robust versions of envelopment estimators described below as well as the conditional measures of efficiency introduced later in Section 6. Wheelock and Wilson (2008) and Wilson (2011) extend this approach to hyperbolic measures, and Simar and Vanhems (2012) extend the approach to directional measures.

5.2 Order-m partial frontiers

Both the FDH and the DEA estimators fully envelop the sample observations in $\mathcal{X}_n$. As a result, FDH and DEA estimators are very sensitive to outliers or extreme data points. Several methods exist (e.g., Wilson (1993, 1995); Simar, 2003, Porembski et al., 2005) for detecting outliers in this setting, but determining what constitutes an "outlier" necessarily involves some subjectivity on the part of the researcher.

Alternatively, Cazals et al. (2002) introduce a concept of a "partial" frontier (as opposed to the "full" frontier Ψ^{∂}) that provides a less-extreme benchmark than the support of the random variable (X, Y) and has its own economic interpretation. The concept is presented here in the input orientation, but extension to the output, hyperbolic, and directional cases is straightforward.

To begin, consider a single input (or cost) x. So here the full frontier can be represented by a function $\varphi(y) = \inf \left\{ x \mid F_{X|Y}(x \mid y) > 0 \right\}$, where the conditional DF is defined above in (5.2) with conditioning on $Y \geq y$. The order-m frontier for an integer $m \geq 1$ is defined by

$$\varphi_m(y) = \mathbb{E}\left[\min(X_1,\ \ldots,\ X_m \mid Y \geq y\right] = \int_0^{\infty} \left[1 - F_{X|Y}(x \mid y)\right]^m dx \tag{5.6}$$

provides a less-extreme benchmark for (x, y) than the full frontier; the first equality in (5.6) defines the concept, and the second is a property derived by Cazals et al. (2002). So, the benchmark for a unit (x, y) producing level y of outputs is the expected minimum input level among m firms drawn at random from the population of firms producing *at least* output level y. For finite m, this is clearly less extreme than the full frontier. Cazals et al. (2002) show that $\varphi_m(y) \to \varphi(y)$ as $m \to \infty$. The order-m efficiency score can be defined as $\theta_m(x, y) = \varphi_m(y)/x$. Note that for finite m, $\theta_m(x, y)$ is not bounded above by 1, in contrast to $\theta(x, y)$ defined in (2.6).

Cazals et al. (2002) extend the order-m efficiency score to multivariate settings as follows. Consider m random draws of random variables X_i, $i = 1, \ldots, m$ from $F_{X|Y}(x \mid y)$. Define the random set $\Psi_m(y) = \left\{(u, v) \in \mathbb{R}_+^{p+q} \mid u \geq X_i, v \geq y\right\}$. Then the Farrell-Debreu input oriented efficiency score of (x, y) with respect to the attainable set $\Psi_m(y)$ is given by

$$\widetilde{\theta}_m(x, y) = \min\left\{\theta \mid (\theta x, y) \in \Psi_m(y)\right\} = \min_{i=1,\ldots,m} \max_{j=1,\ldots,p} \left(X_i^j / x^j\right). \tag{5.7}$$

Since $\Psi_m(y)$ is random, $\widetilde{\theta}_m(x, y)$ is a random variable. Cazals et al. define the order-m input efficiency score as the expectation of this random variable, i.e.,

$$\theta_m(x, y) = \mathbb{E}\left[\widetilde{\theta}_m(x, y) \mid Y \geq y\right] = \int_0^\infty \left[1 - F_{X|Y}(\eta x \mid y)\right]^m d\eta. \tag{5.8}$$

This can be easily computed by a simple Monte Carlo method.

Extension to the output direction is simple; see Cazals et al. (2002) for details. Extension to hyperbolic and directional distances is somewhat more complicated due to the nature of the order-m concept in the multivariate framework, and requires some additional work. Results are given by Wilson (2011) for the hyperbolic case, and by Simar and Vanhems (2012) for directional cases.

A nonparametric estimator of the order-m input-efficiency score is obtained by plugging the empirical distribution $\widehat{F}_{X|Y,n}(\theta x \mid y)$ into (5.8) to replace the unknown $F_{X|Y}(x \mid y)$. Cazals et al. derive a remarkable property for the resulting estimator, i.e.,

$$\sqrt{n}\left(\widehat{\theta}_m(x, y) - \theta_m(x, y)\right) \xrightarrow{\mathcal{L}} N(0, \sigma_m^2(x, y)), \tag{5.9}$$

where an explicit expression for $\sigma_m^2(x, y)$ is given by Cazals et al.. The $\sqrt{n}$ rate of convergence is rather unusual in nonparametric settings; the asymptotic normality facilitates easy construction of confidence intervals. Of course, similar properties hold in the output orientation. In addition, all the properties of the order-m radial distances and their estimators have been extended to hyperbolic and directional distances by Wilson (2011) and Simar and Vanhems (2012), respectively.

5.3 Order-α quantile frontiers

An alternative partial frontier concept for defining a less-extreme benchmark than the full frontier is related to the concept of conditional quantiles, though different from the usual conditional quantile. Aragon et al. (2005) introduce the idea for the case of a univariate input (for an input-oriented measure) or a univariate output (for the output orientation) by using quantiles of a nonstandard, univariate DF. These ideas are extended to the full multivariate setting by Daouia and Simar (2007), who derive quantiles along the radial distances.

Working in the input direction, the central idea is to benchmark the unit operating at (x, y) against the input level not exceeded by $(1 - \alpha) \times 100$-percent of firms among the population of units producing at least output level y. The resulting efficiency measure is defined by

$$\theta_\alpha(x, y) = \inf \left\{ \theta \mid F_{X|Y}(\theta x \mid y) > 1 - \alpha \right\}, \tag{5.10}$$

where it is important to recall that the conditioning is on $Y \geq y$.[5] The quantity $\theta_\alpha(x, y)$ is called the "input efficiency at level $\alpha \times 100\%$." If $\theta_\alpha(x, y) = 1$, then the unit operating at (x, y) is the said to be input efficient at the level $\alpha \times 100\%$ since it is dominated by firms producing at least the level of output y with probability $1 - \alpha$. Similar to the order-m measure, it is clear that $\theta_\alpha(x, y) \to \theta(x, y)$ as $\alpha \to 1$; i.e., the full frontier efficiency measure is recovered as $\alpha \to 1$.

A nonparametric estimator of $\theta_\alpha(x, y)$ is obtained by using the empirical DF $\widehat{F}_{X|Y,n}(\theta x \mid y)$ to replace $F_{X|Y,n}(\theta x \mid y)$ in (5.10); a simple computational algorithm is provided in Daouia and Simar (2007), where the corresponding output-oriented measure and its estimator are also presented. Wheelock and Wilson (2008) extend the order-α concept to hyperbolic measures, and provide a fast numerical algorithm for computing estimates. Simar and Vanhems (2012) extend the method to directional distances.

The properties of the nonparametric order-α estimators are similar to those of the order-m estimators; e.g., in the input orientation,

$$\sqrt{n} \left(\widehat{\theta}_{\alpha,n}(x, y) - \theta_\alpha(x, y) \right) \xrightarrow{\mathcal{L}} N(0, \sigma^2_\alpha(x, y)), \tag{5.11}$$

where again an explicit expression is given by Daouia and Simar (2007) for $\sigma^2_\alpha(x, y)$. Similar results hold in the output, hyperbolic, and directional cases.[6]

[5] Note that the approach here is quite different from traditional nonparametric quantile regression (e.g., Fan et al., 1994; Li and Racine, 2007), where the conditioning is on the event $Y = y$. The non-standard conditioning on $Y \geq y$ in (5.10) guarantees monotonicity properties of the resulting frontier estimates, providing a better economic interpretation (as explained in the next subsection). In addition, conditioning on $Y \geq y$ allows an estimation procedure that avoids smoothing techniques, leading to the $\sqrt{n}$ convergence rate of the resulting nonparametric estimators.

[6] Note that Martins-Filho and Yao (2007) suggest, in the case of a univariate output in an output orien-

5.4 Further extensions

Cazals et al. (2002, Theorem 2.4) and Daouia and Simar (2007, Proposition 2.5) show that the order-m and order-α partial efficiency scores are monotonic under the assumption of tail monotonicity of the implied nonstandard, conditional DF. In other words, So, for example, the desired monotonicity property for $\theta_m(x, y)$ and $\theta_\alpha(x, y)$ are monotone, nondecreasing with y under the tail monotonicity assumption. Similarly, the corresponding output measures $\lambda_m(x, y)$ and $\lambda_\alpha(x, y)$ are monotone nondecreasing with x. In the input-oriented case, tail monototonicity amounts to assuming that for all $y' \geq y$, $F_{X|Y}(x \mid y') \leq F_{X|Y}(x \mid y)$. This is not too restrictive in practice; roughly speaking, it requires that the probability of using less that a fixed input level x decreases as the production level increases.

Unfortunately, even with the tail monotonicity assumption, in finite samples the partial efficiency estimators do not share this monotonicity property. For the univariate case (e.g., with one input in the input orientation), Daouia and Simar (2005) propose an easy way to monotonize the estimated frontiers and show that the modified estimators retain the same asymptotic properties as the original estimators. In a more recent study, Daouia et al. (2013) propose an alternative way for defining the order-α efficiency scores, in a full multivariate setting. The resulting estimators share the desirable monotonicity property and have superior robustness properties, even if the tail monotonicity assumption does not hold. This new quantile approach is obtained from the directional distance estimator of order-α described in Simar and Vanhems (2012). In the input orientation it involves a vector of zero directions for the outputs; in the output orientation, a vector of zero directions is used for the inputs. Analog results for the order-m case, should be available soon.

5.5 Robust estimation of the full frontier

Both the order-m and the order-α partial frontiers can be used to provide robust estimation of the full frontier itself, or of the corresponding full-efficiency scores defined in (2.6)–(2.9). Cazals et al. (2002) show that the estimator $\widehat{\theta}_m(x, y)$ converges to $\widehat{\theta}_{FDH}(x, y)$ as $m \to \infty$. If the convergence is fast enough (i.e., $m = O(n)$), then $\widehat{\theta}_m(x, y)$ converges also to the full frontier efficiency score $\theta(x, y)$, but with the same limiting distribution and the same nonparametric rate of the FDH estimator when $n \to \infty$. However, for finite n, m will be finite and will not envelop all the data points; hence the order-m estimator remains more robust than the FDH estimator in the presence of outliers.

More recently Daouia et al. (2012) have shown, for case of a univariate input in the input

tation, smoothing the estimator of the conditional DF before determining its quantile-frontier. It is not clear that this smoothing, which requires a bandwidth, adds any substantial gain over the simpler procedure using the $\sqrt{n}$-consistent empirical conditional DF as described above.

orientation, by letting m converge to ∞ slowly enough, the order-m estimator provides an asymptotically normally-distributed estimator of distance to the full frontier. The necessary rate for m is roughly $m = O(n^{1/3})$; see Daouia et al., 2012 for a precise formulation.

Not surprisingly, very similar properties hold for the order-α frontiers. Clearly, as $\alpha \to 1$, the order-α estimator $\widehat{\theta}_\alpha(x, y)$ converges to the FDH estimator. But, as shown in Daouia and Simar (2007), if $\alpha = \alpha(n) \to 1$ fast enough, i.e., if $n^{(p+q+1)/(p+q)}(1 - \alpha(n)) \to 0$ as $n \to \infty$, the order-$\alpha(n)$ estimator can be used to estimate full efficiency, obtaining the properties of the FDH estimator (with nonparametric rate of convergence and limiting Weibull distribution). Of course in practice, n is finite, so the order-α frontier will not envelop all the data points, and will also be more robust with respect to extreme or outliers than the ordinary FDH estimator.

Using results from EVT, Daouia et al. (2010) show for the univariate input case in the input orientation that choosing $\alpha = \alpha(n)$ converging to 1 slowly enough, an estimator of distance to the full frontier with a normal limiting distribution is obtained.

Order-m and order-α efficiency estimators are compared and investigated from the Robustness Theory perspective in Daouia and Ruiz-Gazen (2006) and Daouia and Gijbels (2011b, 2011a). They demonstrate relations between the two concepts and analyze their respective advantages and limitations. Daouia and Gijbels (2011b) formalize a data-driven procedure to detect outliers and select appropriate values of the orders, providing a theoretical background for some of ideas in Simar (2003) for detecting outliers. As noted at the beginning of Section 5.2, detection of outliers is critical when using envelopment estimators, and several methods exist for detecting outliers in the context of efficiency estimation. Careful applied researchers should use several of these, as any one method is unlikely to detect all outliers in every situation.

6 Introducing Environmental Factors

The analysis of productive efficiency has in general two components: (i) estimation of a benchmark frontier that serves to evaluate performance of firms; and (ii) investigation of the influence of outside, environmental factors on the production process. These factors, denoted below by $Z \in \mathcal{Z} \subset \mathbb{R}^r$, may reflect difference in ownership, regulatory constraints, business environment, etc. Such factors are neither inputs nor outputs, and are typically not under control of the manager, but nonetheless they may influence the production process. The effect of Z may affect the range of attainable values for the inputs and outputs (X, Y), and hence the shape of the boundary of the attainable set; or Z may affect only the distribution of inefficiencies inside the attainable set; or, in some cases, Z may affect both. Of course Z might also be completely independent of (X, Y). The effect of Z is unknown, and must be estimated appropriately.

There has been dozens of paper in the nonparametric literature suggesting ways to introduce Z into the analysis of the production process. Some are rather simple but quite restrictive (e.g., the one-stage approaches discussed below), while others are valid only under very peculiar, restrictive conditions that are rarely tested. Examples of the latter, including the two-stage approaches discussed later, can be found in hundreds of published applied papers. Cazals et al. (2002) show a natural way to introduce environmental variables by extending their probabilistic formulation of the production process. This lead to the conditional efficiency scores defined in Daraio and Simar (2007b). The various approaches are briefly reviewed below.

6.1 One-stage approaches

The earliest attempts to incorporate environmental variables into the analysis of production were one-stage approaches along the lines of Banker and Morey (1986) and Färe et al. (1989). In this approach, the $Z \in \mathbb{R}^r$ is treated as a vector of r freely disposable inputs or outputs that contribute to the definition of an augmented attainable set $\Psi \subset \mathbb{R}_+^p \times \mathbb{R}_+^q \times \mathbb{R}^r$. Then efficiency scores are defined relative to the boundary of this new set; e.g., in the input orientation, one might define

$$\theta(x, y, z) = \inf \{\theta \mid (\theta x, y, z) \in \Psi\}. \tag{6.1}$$

Then nonparametric DEA or FDH estimators of Ψ treat Z as a freely disposable input if it is favorable to production of output, or as a freely disposable, undesirable output if it is detrimental to output production.

This approach has its own merits and is particularly easy to implement, but it has three important drawbacks. First, the researcher must know in a priori whether Z is favorable or detrimental. Second, the approach only allows monotone effects of Z on the process (in many cases, effects may be either U-shaped or inverted U-shaped). Finally, one must assume free disposability, and in addition convexity if DEA estimators are used, of the augmented set Ψ. For these reasons, this approach has been used less than others in recent years.

6.2 Two-stage approaches

A large part of the literature focuses on two-stage approaches for including environmental factors. One can find perhaps hundreds of papers using this approach, but as Simar and Wilson (2007) discuss, in most of these studies statistical models are ill-defined, inappropriate estimators are used, and inference is inconsistent. The basic idea is to estimate efficiency scores in a first stage considering only the space of inputs and outputs (X, Y), ignoring Z.

Then in a second stage, the estimated efficiencies are regressed on Z. Although DEA and FDH efficiency estimates are truncated at one by construction, many examples exist in the literature where researchers have either ignored this, or have confused truncation with censoring. By formalizing this procedure, Simar and Wilson (2007 2011b) have shown that (i) this approach is meaningful only if a "separability condition" between Z and (X, Y) holds; and (ii) even if the second stage regression is meaningful, traditional inference is flawed by problems linked to the fact that the true efficiency scores are not observed, but instead must be replaced by biased estimators that are not independent. Simar and Wilson suggest use of bootstrap methods for addressing issue (ii), but the separability assumption remains a restrictive assumption in need of testing.

The problem can be formalized as follows. Consider the the random variables (X, Y, Z) defined on an appropriate probability space with support $\mathcal{P} \subseteq \mathbb{R}_+^p \times \mathbb{R}_+^q \times \mathbb{R}^r$. Consider also the conditional distribution of (X, Y), conditional on $Z = z$, described by

$$H_{XY|Z}(x, y \mid z) = \Pr(X \leq x, Y \geq y \mid Z = z). \tag{6.2}$$

This is the probability that a firm facing environmental conditions z will dominate the point (x, y). Given that $Z = z$, the attainable set of combinations of inputs and outputs is

$$\Psi^z = \left\{(x, y) \in \mathbb{R}_+^{p+q} \mid Z = z \text{ and } x \text{ can produce } y\right\}, \tag{6.3}$$

the support of $H_{XY|Z}(x, y \mid z)$.

Denoting as above in (5.2) The unconditional probability of (x, y) being dominated is given by by (5.2); then obviously

$$H_{XY}(x, y) = \int_{\mathcal{Z}} H_{XY|Z}(x, y \mid z) f_Z(z) dz \tag{6.4}$$

where $f_Z(z)$ is the marginal density of Z. The support of $H_{XY}(x, y)$ is as usual denoted by Ψ, the marginal attainable set. These attainable sets are related by

$$\Psi = \left\{(x, y) \in \mathbb{R}_+^{p+q} \mid x \text{ can produce } y\right\} = \bigcup_{z \in \mathcal{Z}} \Psi^z. \tag{6.5}$$

Of course, for all $z \in \mathcal{Z}$, $\Psi^z \subseteq \Psi$.

Now situations where the two-stage approach is meaningful can be described. In particular, suppose that the shape of the attainable sets Ψ^z change with z, which is quite natural in many applications). Then no economic meaning for a firm facing environmental conditions z can be given to the marginal measure of efficiency (e.g., $\theta(x, y)$) with respect to the boundary of the *marginal* set Ψ, because the boundary of this set may not be reachable for the unit facing condition z. The only situation where any meaning can be attached to these marginal measures is the case where the shape of the boundary of the conditional attainable sets is independent of z. This is the "separability condition" in Simar and Wilson (2007):

$$\Psi^z = \Psi, \quad \text{for all } z \in \mathcal{Z}. \tag{6.6}$$

If this condition is verified, then it is reasonable to use an appropriate second-stage regression to investigate whether Z has some impact on the distribution of the efficiencies inside the unique attainable set Ψ, provided one uses methods to make inference consistently.

Additional methodological difficulties are described in Simar and Wilson (2007, 2011b). Even if the specified second-stage regression model is "true" (or say, appropriate) for the true values of $\theta(x, y)$, these are latent variables and in practice are replaced by nonparametric estimates, which are biased, suffer form the curse of dimensionality, and are not independent. The statistical model specified by Simar and Wilson (2007) suggests using a truncated normal regression, transforming the estimated efficiencies so that they are bounded below by 1 in the case of input-efficiency estimates. Results from Monte-Carlo experiments presented by Simar and Wilson indicate that bootstrap algorithms provide reasonable approximation for making inference in the second stage regression. Park et al. (2008) suggest use of a nonparametric truncated regression model in the second stage, using local likelihood methods.

Kneip et al. (2013a) analyze the consequence of replacing true, unknown efficiency scores by DEA or FDH estimators in a second-stage regression, and suggest alternative methods for obtaining valid inference. Banker and Natarajan (2008) propose a different model where the two-stage approach can be applied. However, as discussed in Simar and Wilson (2011b), their model is rather restrictive, and the conventional inference they suggest in the second stage (using simple OLS techniques) is incorrect due to the problems enumerated in Kneip et al. (2013a).

6.3 Conditional frontiers

Since the separability condition (6.6) may be problematic, the safest approach for introducing environmental variables is rely on the conditional model (6.2) along the lines of Cazals et al. (2002). Daraio and Simar (2007b) define the conditional Farrell-Debreu input efficiency measure as

$$\theta(x, y \mid z) = \inf \left\{ \theta > 0 \mid (\theta x, y) \in \Psi^z \right\}, \tag{6.7}$$

i.e., the radial distance (in the input space) from (x, y) to the efficient boundary of units facing environmental conditions z. Adaptation to the output orientation is straightforward. Along the lines of the probabilist formulation of efficiency scores in Section 5.1, it can be shown that

$$\theta(x, y \mid z) = \inf \left\{ \theta \mid F_{X|Y,Z}(\theta x \mid y, z) > 0 \right\}, \tag{6.8}$$

which can be compared with (5.3). Note also that here, in $F_{X|Y,Z}(x \mid y, z)$, the conditioning event is $(Y \geq y, Z = z)$. This conditional DF is given by $F_{X|Y,Z}(x \mid y, z) = H_{XY|Z}(x, y \mid z)/H_{XY|Z}(\infty, y \mid z)$.

Nonparametric estimators of the $\theta(x, y \mid z)$ are obtained from a sample $\mathcal{S}_n = \{(X_i, Y_i, Z_i)\}_{i=1}^n$ by plugging a nonparametric estimator of $F_{X|Y,Z}(x \mid y, z)$ into (6.8). Such estimator may be obtained by standard nonparametric kernel smoothing, e.g.,

$$\widehat{F}_{X|Y,Z,n}(x \mid y, z) = \frac{\sum_{i=1}^n \mathbb{1}(X_i \leq x, Y_i \geq y) K((Z_i - z)/h)}{\sum_{i=1}^n \mathbb{1}(Y_i \geq y) K((Z_i - z)/h)}, \tag{6.9}$$

where $K((Z_i - z)/h)$ is a familiar shorthand notation in case of multivariate Z (using product kernels and a vector of bandwidths $h = (h_1, \ldots, h_r)$). As noted in Daraio and Simar (2007b), the kernels must have a compact support. Optimal bandwidth selection procedures for standard conditional distributions have been proposed by Hall et al. (2004). The procedure has been adapted to the setup here by Bădin et al. (2010).

The resulting estimator of $\theta(x, y \mid z)$ is called the conditional FDH estimator. Daraio and Simar (2005, 2007a) show that the estimator is given by the simple expression

$$\widehat{\theta}_{FDH}(x, y \mid z) = \inf\left\{\theta \mid \widehat{F}_{X|Y,Z,n}(\theta x \mid y, z) > 0\right\} = \min_{\{i:\, ||Z_i - z|| \leq h\}} \max_{j=1,\, \ldots\, ,p} X_i^j / x^j, \tag{6.10}$$

where $||a||$ denotes the Euclidean norm of a vector a. This can be interpreted as a localized version of the FDH estimator, with localization for data point such that Z_i is in an h-neighborhood of z (compare this with the unconditional FDH given in (3.3)). In fact, the conditional attainable set is estimated by

$$\widehat{\Psi}^z_{FDH} = \left\{(x, y) \in \mathbb{R}_+^{p+q} \mid x \geq x_i,\ y \leq y_i \text{ for } i \text{ such that } ||Z_i - z|| \leq h\right\}. \tag{6.11}$$

Daraio and Simar (2007a 2007b) suggest convexifying this set to obtain an estimator of Ψ^z under the additional assumption of local convexity (i.e., assuming that Ψ^z is convex):

$$\widehat{\Psi}^z_{DEA} = \left\{(x, y) \in \mathbb{R}_+^{p+q} \mid x \geq \sum_{i \in \mathcal{J}_z} \gamma_i x_i,\ y \leq \sum_{i \in \mathcal{J}_z} \gamma_i y_i \text{ for } \gamma_i \geq 0 \text{ such that } \sum_{i \in \mathcal{J}_z} \gamma_i = 1\right\}, \tag{6.12}$$

where $\mathcal{J}_z = \{i \colon ||Z_i - z|| \leq h\}$. The corresponding conditional DEA estimator of $\theta(x, y \mid z)$ is obtained as

$$\widehat{\theta}_{DEA}(x, y \mid z) = \inf\left\{\theta \mid (\theta x, y) \in \widehat{\Psi}^z_{DEA}\right\}, \tag{6.13}$$

which can be written as a local version of the linear program in (3.5).

The asymptotic properties of these nonparametric conditional efficiency estimators are derived in Jeong et al. (2010). To summarize, the properties are similar to those of the unconditional DEA and FDH estimators, except that the effective number of observations used for building the estimators is now $nh_1 \ \ldots, \ h_r$. When optimal bandwidths are selected as in Bădin et al. (2010), this gives an effective number of observations of the order $n^{4/(4+r)}$ in place of n for the unconditional estimators. The rates of convergence are thus deteriorated by this factor.

The conditional efficiency scores have also their robust versions; see Cazals et al. (2002) and Daraio and Simar (2007b) for the order-m version, and Daouia and Simar (2007) for the order-α analog. Also, conditional measures have been extended to hyperbolic distances in Wheelock and Wilson (2008), and to hyperbolic distances by Simar and Vanhems (2012).

Bădin et al. (2012, 2013) suggest useful tools for analyzing the impact of Z on the production process, by exploiting the comparison between the conditional and unconditional measures. These tools (graphical and nonparametric regressions) allow one to disentangle the impact of Z on any potential shift of the frontier or potential shift of the inefficiency distributions. Daraio and Simar (2013) provide also a bootstrap test for testing the significance of environmental factors on the conditional efficiency scores.

Recently Florens et al. (2013) propose an alternative approach for estimating the conditional efficiency scores. The approach avoids explicit estimation of a nonstandard conditional distribution (e.g., $F_{X|Y,Z}(x \mid y, z)$). The approach is less sensitive to the curse of the dimensionality described above. It is based on very flexible nonparametric location-scale regression models for pre-whitening the inputs and the outputs to eliminate their dependence on Z. This allows one to define "pure" inputs and outputs, and hence a "pure" measure of efficiency. The method permits returning in a second stage to the original units and evaluating the conditional efficiency scores, but without explicitly estimating a conditional DF. The paper proposes also a bootstrap procedure for testing the validity of the location-scale hypothesis.

7 Other Aspects and Open Issues

7.1 Parametric approximations of nonparametric frontier

Nonparametric methods are attractive because they impose few restrictions on either the functional form of the frontier or on the stochastic part of the model (i.e., the distribution of the inefficiencies below the frontier). On the other hand, parametric models for the frontier are appealing because they offer a richer economic interpretation of the production process, e.g., sensitivity of the production of output to particular inputs, etc. For purposes of illustration in the discussion below, consider a single output and a production function $Y_i = \beta_0 + \beta' X_i - U_i$, where $U_i \geq 0$ (alternatively, for a cost function, $U_i \leq 0$); Y and X could be measure on a log scale.

The first approach in the parametric world is due to Aigner and Chu (1968), who proposed enveloping the data in a parametric way by solving some mathematical programs to minimize $\sum_{i=1}^{n} |U_i|$ or $\sum_{i=1}^{n} U_i^2$ with respect to β_0, β, subject to $U_i \geq 0$. One of the drawbacks of the approach is that very few useful statistical properties of the resulting estimator of the βs have been developed (some consistency results have been obtained by Knight, 2006). In addition, in the mathematical programs, observations far from the frontier have excessive weight (in particular, when using a quadratic objective) in determining the shape of the optimal frontier. Finally, since the resulting frontier will envelop all the data points, the estimators will be very sensitive to outliers.

The other stream of approaches in the parametric framework follows the lines of Greene (1980) and is based on (shifted) regression ideas. As explained in details in Florens and Simar (2005), these approaches have several drawbacks. First, they require in most cases specification of the density of U. Second, independence between the inputs X and the stochastic inefficiency term U (or at least, $\mathbb{E}(U|X) = \mu$ for a constant μ) is required for all the proposed estimators. Most importantly, however, since $\mathbb{E}(Y_i|X_i) = \beta_0 + \beta' X_i - \mu$, the frontier function is, by construction, a shift of the regression of Y on X, i.e, the conditional mean function. Hence any estimation procedure (by Modified OLS, or MLE) will capture the shape of the "middle" of the cloud of data points. This is not natural for capturing the shape of the frontier function, which may differ from the shape of the conditional mean function.

Florens and Simar (2005) address all these drawbacks. A parametric frontier model is estimated by using a two-step procedure. In the first step, a nonparametric method is used to estimate where the production frontier is located, and all the data points are projected onto this frontier. Then in a second step, these projections are adjusted using simple methods to fit a specified parametric model. Since the production frontier is the locus of optimal production plans, one should expect a better parametric fit if only efficient units are used to estimate it. This idea appears in Simar (1992), but with no theoretical justification nor results on the statistical properties of the estimators of the βs. When using the FDH estimator in the first stage, Florens and Simar (2005) prove the consistency of the method, and in the order-m case (which is more robust to outliers), they obtain $\sqrt{n}$-consistency and asymptotic normality, with an explicit expression for the variance. Daouia et al. (2008) obtain similar results when using the robust order-α estimators in the first stage. Daraio and Simar (2007a) indicate how to extend these ideas to a full multivariate setting by using a parametric model for a distance function. These approaches are very appealing and suggest that bridges can built between the two worlds (i.e., parametric and nonparametric).

7.2 Nonparametric Stochastic Frontier

One of the limitations of the nonparametric envelopment estimators described above is that, as with all estimators in "deterministic" frontier models, they do not allow for noise in the DGP. It is frequently argued that parametric approaches are superior to the nonparametric approaches because they admit "stochastic" frontier models such as (2.14). Of course this neglects the fact that in some applications the parametric restrictions, both on the frontier function and on the stochastic part of the model, may be irrelevant. A number of recent papers attempt to introduce noise into nonparametric boundary and frontier estimation. The robust estimators of the full frontier described above can be viewed as allowing the presence of noise, but still the underlying model is a "deterministic" one.

Hall and Simar (2002) show that even if the noise is symmetric and the inefficiency distribution has a jump at the frontier, a fully nonparametric model with both noise and inefficiency is not identified. They provide a strategy that allows introduction of noise into the model and consistent estimation of the unknown boundary of support of a random variable reflecting inefficiency. Consistency and identification are obtained by letting the variance of the noise converge to zero as $n \to \infty$. Monte-Carlo experiments indicate that the procedure works well for a signal-to-noise ratio (measured by ratio of the respective standard deviations) of 5, and the authors argue that the signal-to-noise ratio is likely to be high in many applications. They apply the procedure to estimate nonparametrically a production function, and hence provide a first nonparametric alternative to classical parametric stochastic frontier models such as (2.14). Simar (2007) extends these ideas to fully multivariate settings, confirming good behavior of the resulting modified DEA estimators in the presence of noise of moderate size.

To solve the basic identifiability issue, some structure on the model is required. One approach is to leave the production function unspecified while specifying a fully parametric model for the stochastic part (i.e., specifying a parametric density for the inefficiency term U and for the independent noise V). The simplest approach is to assume homoskedasticity of both components. This is investigated by Fan et al. (1996) and Kuosmanen and Kortelainen (2012). These semiparametric approaches are interesting, but they retain both the homoskedasticity assumption of the stochastic terms and the parametric assumptions for the inefficiency component, and hence are likely to introduce misspecification errors into the model, leaving statistical consistency in doubt.

Kumbhakar et al. (2007) propose an alternative based on local maximum likelihood techniques. Their model is

$$Y_i = r(X_i) + V_i - U_i, \tag{7.1}$$

where $U \mid X \sim |N(0, \sigma_u^2(X))|$ and $V \mid X \sim N(0, \sigma_v^2(X))$, with U and V independent conditionally on X. The functions $r(X), \sigma_u^2(X)$ and $\sigma_v^2(X)$ are unspecified, as are unknown functional parameters. Estimation uses local maximum likelihood techniques where the unknown functional parameters are approximated by local polynomials (either linear or quadratic). Kumbhakar et al. (2007) provide asymptotic properties, which involve a limiting normal distribution. The procedure requires selection of bandwidths, which is done by the usual likelihood cross-validation. Simar and Zelenyuk (2011) extend the procedure to a fully multivariate setup by using an analog of the model in (7.1) to characterize the univariate distance to a multivariate boundary surface. By doing so, they provide stochastic versions of DEA and FDH estimators. The resulting estimates show encouraging results, such as adaptation to the size of noise (i.e., the stochastic FDH/DEA estimators collapse to the usual FDH/DEA estimators in the absence of noise), robustness with respect to outliers, and other properties. However, the method is computationally intensive. Recent work by Simar et al. (2013) avoids much of the computational burden by proposing a nonparametric version of the modified OLS technique. Asymptotic properties of this nonparametric, modified least-squares estimator are also given.

The latter approaches are appealing and very flexible, but still they require some “local” parametric assumptions on the distribution of the inefficiency term in order to obtain statistical consistency. Kneip et al. (2012b) overcome this limitation, obtaining identification of the model by assuming independent Gaussian noise, but with unknown variance. The inefficiency distribution is left unspecified, and is estimated by simple histograms. Then the model is estimated by the penalized likelihood method, where the penalization controls the smoothness of the histogram. Kneip et al. first address estimation of a univariate boundary, then adapt the procedure to estimate a production (or cost) function. Although only $\log(n)$ convergence rates are achieved in theory, in practice the resulting estimates behave well in finite samples of sizes commonly faced in practice, as revealed by Monte-Carlo experiments.

7.3 Testing Issues

For the practitioner, it is important to be able to test empirically hypothesis relating to the DGP and having to do with the shape of the frontier (e.g., convexity, returns to scale, etc.). This is important not only for economic considerations, but also for statistical reasons, since as shown above there is much to be gained in terms of statistical precision by assuming convexity of Ψ or CRS for Ψ^∂ *if such assumptions are appropriate.* Reducing the dimensionality of the problem by testing the relevance of certain inputs and outputs or the possibility of aggregating inputs or outputs is of interest. Conversely, too-restrictive assumptions (e.g., assuming Ψ^∂ is CRS when it is in fact only VRS) lead to inconsistent estimation.

In typical testing problems, a test statistic $T(\mathcal{X}_n)$ is defined to discriminate between a null and an alternative hypothesis; the primary difficulty is in providing a critical value for a test of a given size. This has been done in several studies using bootstrap approximations; e.g., Simar and Wilson (2001)) use bootstrap methods to test whether inputs or outputs can be aggregated. Similarly, bootstrap methods are used by Simar and Wilson (2002) to test CRS versus VRS; by Simar and Zelenyuk (2006, 2007) to test whether efficiency distributions and their means differ across two samples; by Simar and Wilson (2011a) to test convexity of Ψ; and by Daraio et al. (2010)) to testing "separability" condition discussed in Section 6.2 in the presence of environmental factors. These studies used either the restrictive, homogeneous bootstrap ideas of Simar and Wilson (1998) or the subsampling ideas from Simar and Wilson (2011a), but lack convincing theoretical justification. Intensive Monte-Carlo simulations in several of the studies indicate that the suggested procedures achieve reasonable size and power properties in samples of moderate size, but theoretical results are lacking in these studies. This gap is filled by Kneip et al. (2013a).

In each case listed above, test statistics involving comparisons of FDH, CRS-DEA, or VRS-DEA estimates are used. The difficulty for providing theoretical justification for general hypothesis tests is that the properties of FDH/DEA estimators described in Section 4 hold only for a given, fixed point of interest, e.g., for $\widehat{\theta}_{\text{VRS}}(x, y \mid \mathcal{X}_n)$, an estimator of $\theta(x, y)$, where (x, y) is non-stochastic. Note that here the notation for the VRS-DEA estimator of $\theta(x, y)$ has been modified to explicitly show that the estimator is based on a random sample $\mathcal{X}_n$ of firms' input-output combinations. But when constructing test statistics for tests of returns to scale and other model features, the FDH or DEA estimators are evaluated at random points $(X_i, Y_i) \in \mathcal{X}_n$.

To illustrate the problem, consider perhaps the simplest statistic, the sample mean of the efficiency scores

$$\widehat{\mu}_n = n^{-1} \sum_{i=1}^{n} \widehat{\theta}(X_i, Y_i \mid \mathcal{X}_n), \tag{7.2}$$

where the efficiency estimators under the summation sign could be FDH, CRS-DEA, or VRS-DEA estimators. The statistic $\widehat{\mu}_n$ might be used to make inference on the (population) mean efficiency $\mu_\theta = \mathbb{E}(\theta(X, Y))$. Denote by $\sigma_\theta^2 = \text{VAR}(\theta(X, Y))$ the population variance of efficiency scores. If the *true* efficiencies were observable, one might be happy to use

$$\overline{\theta}_n = n^{-1} \sum_{i=1}^{n} \theta(X_i, Y_i) \tag{7.3}$$

to make inference since under mild regularity conditions, $\sqrt{n}(\overline{\theta}_n - \mu_\theta) \xrightarrow{\mathcal{L}} N(0, \sigma_\theta^2)$. But of course $\overline{\theta}_n$ is a latent variable, and must be replaced by $\widehat{\mu}_n$, which is observable.

Unfortunately, DEA and FDH estimators are biased and correlated. The bias is of order $n^{-\kappa}$, the same order as the convergence rate, where κ is defined above for FDH, VRS-DEA, and CRS-DEA estimators and depends on the dimension $p+q$ of the input-output space. The bias does not disappear when averaging in (7.2), but the variance tends to zero, as shown in Kneip et al. (2013a). In fact, Theorem 4.1 of Kneip et al. reveals that under mild regularity conditions and assumptions appropriate for the given estimator (i.e., free disposability for FDH, plus convexity for VRS-DEA or constant returns to scale for the CRS-DEA),

$$\sqrt{n}(\widehat{\mu}_n - \mu_\theta - Cn^{-\kappa} - R_{n,\kappa}) \xrightarrow{\mathcal{L}} N(0, \sigma_\theta^2), \tag{7.4}$$

where C is some constant and $R_{n,\kappa} = o_p(n^{-\kappa})$. This result shows clearly that the inherent bias of the envelopment estimators "kills" the variance whenever $\kappa \leq 1/2$. Kneip et al. (2013a) solve the problem by estimating the leading terms of the bias by a kind of generalized jackknife. Then by using a simple, consistent estimator of the variance, they provide a way for making inference on μ_θ using normal approximations. If the dimension $p+q$ is too large ($p+q > 4$ for the VRS case;, $p+q > 5$ for the CRS case; or $p+q > 3$ for the FDH case), this bias correction is not enough. In these cases, a solution is provided by computing the estimator of μ by an average of $\widehat{\theta}(X_j, Y_j \mid \mathcal{X}_n)$ over random subsample $j = 1, \ldots, n_\kappa$ of observations, where $n_\kappa = [n^{2\kappa}] \leq n$. Kneip et al. (2013a) show that the resulting statistic has a corresponding central limit, normal approximation, but with a lower rate.

Kneip et al. (2013b) extend these basic theoretical results to testing problems such as testing whether mean efficiencies differ across two groups of producers, testing CRS versus VRS, testing convexity of Ψ. The proposed tests avoid the need for bootstrap methods, although the bootstrap remains a valid alternative and is perhaps useful in some instances. The tests are used in an empirical setting by Apon et al. (2013) to first test whether Ψ is convex, then to test CRS versus VRS in cases where convexity of Ψ is not rejected, and finally to test for differences in mean efficiency across two groups; the estimators (i.e., FDH, VRS-DEA, or CRS-DEA) for the last test are chosen according to the outcomes of the first two tests.

7.4 Nonparametric models for panel data

Many approaches have been proposed for dealing with panel data in the parametric world of productivity analysis; see Kumbhakar and Lovell (2000) and Greene (2008) and the references therein. By contrast, although a large literature on using panel data and FDH/DEA estimators to examine changes in production processes over time, the literature is largely astatistical. Most of this nonparametric literature revolves around Malmquist indices defined to measure changes in productivity over time. These changes can be decomposed using various identities to attribute changes in productivity to changes in efficiency, shifts in the technology Ψ^∂, and other changes from one period to the next; see Färe et al., 2008 for details and a comprehensive survey.

To illustrate, consider two periods $t_1 < t_2$, with corresponding attainable sets Ψ_{t_1}, Ψ_{t_2}. Let (x^{t_j}, y^{t_j}) denote the production plan of a firm at time t_j, $j = 1, 2$, and let $\Delta_o^{t_i}(x^{t_j}, y^{t_j})$ denote the output-oriented Shephard (1970) distance (i.e., the inverse of the output-oriented Farrell efficiency measure) of the point x^{t_j}, y^{t_j}, relative to the conical hull of Ψ_{t_i}, $i = 1, 2$. Then the output-oriented Malmquist productivity index is defined by

$$M_o = \left(\frac{\Delta_o^{t_1}(x^{t_2}, y^{t_2})}{\Delta_o^{t_1}(x^{t_1}, y^{t_1})} \times \frac{\Delta_o^{t_2}(x^{t_2}, y^{t_2})}{\Delta_o^{t_2}(x^{t_1}, y^{t_1})} \right)^{1/2} . \tag{7.5}$$

This gives the geometrical mean of the gain in productivity of the firm moving from (x^{t_1}, y^{t_1}) in period t_1 to (x^{t_2}, y^{t_2}) in period t_2, measured relative to the conical hulls of the attainable sets in each of the two periods. In practice, the terms $\Delta_o^{t_i}(x^{t_j}, y^{t_j})$ can be estimated by CRS-DEA estimators.

Simar and Wilson (1999a) adapt the smooth, homogeneous bootstrap of Simar and Wilson (1998) to make inference about the measure defined by (7.5), bootstrapping on pairs of observations (from each of two time periods) to preserve the time dependence. Daskovska et al. (2010) extend the idea, and introduce some dynamics to provide forecasts of Malmquist indexes.

Kneip and Simar (1996) use kernel methods to estimate individual production functions (for each firm) in a panel-data setting, and define the production frontier as the envelope of the these individual functions. Asymptotic theory is provided, but the method rests on $T \to \infty$ as well as $n \to \infty$, where T and n denote the number of time periods and number of firms, respectively. Note that by enveloping the individual production frontiers, Kneip and Simar anticipate the idea of "metafrontiers" presented by O'Donnell et al. (2008).

Several semiparametric approaches are possible in the context of panel data. Park and Simar (1994) suggest linear models for the production function with a nonparametric random firm effect, whose support (i.e., upper boundary) determines the frontier level. Park et al. (1998) analyze the case where this random effect is correlated with some regressors (inputs). Park et al. (2003a, 2003, 2007) extend these results to various forms of dynamic models. Kneip et al. (2012a) extend these ideas to still more general semiparametric models by analyzing a model having a nonparametric long time trend and a firm-specific technical efficiency term varying non parametrically with time. The latter is estimated by factor models. This approach can be viewed as a compromise between the fully parametric approach of Kneip and Simar (1996) and the somewhat restrictive semiparametric model of Park and Simar (1994). Kneip and Sickles (2011) provide a comprehensive survey of the various approaches, including the ones mentioned here.

8 Conclusion

The field of frontier estimation is fascinating, because it is a nonstandard econometric problem and is not easy. Both parametric and nonparametric approaches have their own advantages and disadvantages, but clearly both approaches are statistical in nature. The only real differences are in the assumptions the researcher willing to make.

Nonparametric methods for efficiency estimation bring together a wide variety of mathematical tools from mathematical statistics, econometrics, and operations research. As this Guided Tour shows, a large number of statistical results are available today, but a number of challenges remain to be solved. These include the nonparametric treatment of endogeneity and latent heterogeneity in frontier settings, as well as relevant, flexible nonparametric models for panel data. These and other unresolved issues are currently being pursued by the authors of the Guided Tour.

References

Afriat, S. (1972), Efficiency estimation of production functions, *International Economic Review* 13, 568–598.

Aigner, D., C. A. K. Lovell, and P. Schmidt (1977), Formulation and estimation of stochastic frontier production function models, *Journal of Econometrics* 6, 21–37.

Aigner, D. J. and S. F. Chu (1968), On estimating the industry production function, *American Economic Review* 58, 826–839.

Apon, A. W., L. B. Ngo, M. E. Payne, and P. W. Wilson (2013), Assessing the effect of high performance computing capabilities on academic research output. Unpublished working paper, Department of Economics and School of Computing, Clemson University, Clemson, SC 29634 USA.

Aragon, Y., A. Daouia, and C. Thomas-Agnan (2005), Nonparametric frontier estimation: A conditional quantile-based approach, *Econometric Theory* 21, 358–389.

Banker, R. (1984), Estimating most productive scale size using data envelopment analysis, *European Journal of Operations Research* 17, 35–44.

Banker, R. D. (1993), Maximum likelihood, consistency and data envelopment analysis: a statistical foundation, *Management Science* 39, 1265–1273.

Banker, R. D., A. Charnes, and W. W. Cooper (1984), Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science* 30, 1078–1092.

Banker, R. D., W. W. Cooper, L. M. Seiford, R. M. Thrall, and J. Zhu (2004), Returns to scale in different dea models, *European Journal of Operational Research* 154, 345–362.

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

Banker, R. D. and R. Natarajan (2008), Evaluating contextual variables affecting productivity using data envelopment analysis, *Operations Research* 56, 48–58.

Battese, G. E. and G. S. Corra (1977), Estimation of a production frontier model: With application to the pastoral zone off eastern australia, *Australian Journal of Agricultural Economics* 21, 169–179.

Bickel, P. J. and D. A. Freedman (1981), Some asymptotic theory for the bootstrap, *Annals of Statistics* 9, 1196–1217.

Bickel, P. J. and A. Sakov (2008), On the choice of m in the m out of n bootstrap and confidence bounds for extrema, *Statistica Sinica* 18, 967–985.

Brown, D. F. (1979), Voronoi diagrams from convex hulls, *Information Processing Letters* 9, 223–228.

Bădin, L., C. Daraio, and L. Simar (2010), Optimal bandwidth selection for conditional efficiency measures: A data-driven approach, *European Journal of Operational Research* 201, 633–664.

— (2012), How to measure the impact of environmental factors in a nonparametric production model, *European Journal of Operational Research* 223, 818–833.

— (2013), Explaining inefficiency in nonparametric production models: The state of the art, *Annals of Operations Research* Forthcoming.

Bădin, L. and L. Simar (2009), A bias-corrected nonparametric envelopment estimator of frontiers, *Econometric Theory* 25, 1289–1318.

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

Chambers, R. G., Y. Chung, and R. Färe (1998), Profit, directional distance functions, and nerlovian efficiency, *Journal of Optimization Theory and Applications* 98, 351–364.

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

Cooper, W. W., L. M. Seiford, and K. Tone (2000), *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*, Boston, MA: Kluwer Academic Publishers.

Cooper, W. W., L. M. Seiford, and J. Zhu, eds. (2011), *Handbook on Data Envelopement Analysis*, International Series in Operations Research and Management Sciences, New York: Springer-Verlag, 2nd edition.

Daouia, A., J. P. Florens, and L. Simar (2008), Functional convergence of quantile-type frontiers with application to parametric approximations, *Journal of Statistical Planning and Inference* 138, 708–725.

— (2010), Frontier estimation and extreme value theory, *Bernoulli* 16, 1039–1063.

— (2012), Regularization of non-parametric frontier estimators, *Journal of Econometrics* 168, 285–299.

Daouia, A. and I. Gijbels (2011a), Estimating frontier cost models using extremiles, in I. Van Keilegom and P. W. Wilson, eds., *Exploring Research Frontiers in Contemporary Statistics and Econometrics*, Berlin: Springer-Verlag, pp. 65–81.

— (2011b), Robustness and inference in nonparametric partial frontier modeling, *Journal of Econometrics* 161, 147–165.

Daouia, A. and A. Ruiz-Gazen (2006), Robust nonparametric frontier estimators: Qualitative robustness and influence function, *Statistica Sinica* 16, 1233–1253.

Daouia, A. and L. Simar (2005), Robust nonparametric estimators of monotone boundaries, *Journal of Multivariate Analysis* 96, 311–331.

— (2007), Nonparametric efficiency analysis: A multivariate conditional quantile approach, *Journal of Econometrics* 140, 375–400.

Daouia, A., L. Simar, and P. W. Wilson (2013), Measuring firm performance using nonparametric quantile-type distances. Discussion paper, Institut de Statistique Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Daraio, C. and L. Simar (2005), Introducing environmental variables in nonparametric frontier models: A probabilistic approach, *Journal of Productivity Analysis* 24, 93–121.

— (2007a), *Advanced Robust and Nonparametric Methods in Efficiency Analysis*, New York: Springer Science+Business Media, LLC.

— (2007b), Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach, *Journal of Productivity Analysis* 28, 13–32.

— (2013), Directional distances and their robust versions: Computational and testing issues. Discussion paper #2013/38, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Daraio, C., L. Simar, and P. W. Wilson (2010), Testing whether two-stage estimation is meaningful in non-parametric models of production. Discussion paper #1031, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Daskovska, A., L. Simar, and S. Van Bellegem (2010), Forecasting the Malmquist productivity index, *Journal of Productivity Analysis* 33, 97–107.

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

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

Fan, J., T. C. Hu, and Y. K. Truong (1994), Robust nonparametric function estimation, *The Scandinavian Journal of Statistics* 21, 433–446.

Fan, Y., Q. Li, and A. Weersink (1996), Semiparametric estimation of stochastic production frontier models, *Journal of Business and Economic Statistics* 14, 460–468.

Färe, R. and S. Grosskopf (2000), Theory and application of directional distance functions, *Journal of Productivity Analysis* 13, 93–103.

— (2004), *Efficiency and Productivity: New Directions*, Boston, MA: Kluwer Academic Publishers.

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

Färe, R., S. Grosskopf, C. A. K. Lovell, and C. Pasurka (1989), Multilateral productivity comparisons when some outputs are undesirable: A nonparametric approach, *Review of Economics and Statistics* 71, 90–98.

Färe, R., S. Grosskopf, and D. Margaritis (2008), Productivity and efficiency: Malmquist and more, in H. Fried, C. A. K. Lovell, and S. Schmidt, eds., *The Measurement of Productive Efficiency*, chapter 5, Oxford: Oxford University Press, 2nd edition, pp. 522–621.

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

Florens, J. P. and L. Simar (2005), Parametric approximations of nonparametric frontiers, *Journal of Economietrics* 124, 91–116.

Florens, J. P., L. Simar, and I. Van Keilegom (2013), Frontier estimation in nonparametric location-scale models, *Journal of Econometrics* Forthcoming.

Fried, H. O., C. A. K. Lovell, and S. S. Schmidt, eds. (2008), *The Measurement of Productive Efficiency*, Oxford: Oxford University Press, 2nd edition.

Gattoufi, S., M. Oral, and A. Reisman (2004), Data envelopment analysis literature: A bibliography update (1951–2001), *Socio-Economic Planning Sciences* 38, 159–229.

Gijbels, I., E. Mammen, B. U. Park, and L. Simar (1999), On estimation of monotone and concave frontier functions, *Journal of the American Statistical Association* 94, 220–228.

Greene, W. H. (1980), Maximum likelihood estimation of econometric frontier functions, *Journal of Econometrics* 13, 27–56.

— (2008), The econometric approach to efficiency analysis, in H. O. Fried, C. A. K. Lovell, and S. S. Schmidt, eds., *The Measurement of Productive Efficiency and Productivity Growth*, Oxford: Oxford University Press, Inc., pp. 92–250.

Hall, P., W. Härdle, and L. Simar (1995), Iterated bootstrap with application to frontier models, *The Journal of Productivity Analysis* 6, 63–76.

Hall, P., B. U. Park, and S. E. Stern (1998), On polynomial estiators of frontiers and boundaries, *Journal of Multivariate Analysis* 43, 71–98.

Hall, P., J. S. Racine, and Q. Li (2004), Cross-validation and the estimation of conditional probability densities, *Journal of the American Statistical Association* 99, 1015–1026.

Hall, P. and L. Simar (2002), Estimating a changepoint, boundary or frontier in the presence of observation error, *Journal of the American Statistical Association* 97, 523–534.

Jeong, S. O. (2004), Asymptotic distribution of DEA efficiency scores, *Journal of the Korean Statistical Society* 33, 449–458.

Jeong, S. O. and B. U. Park (2006), Large sample approximation of the distribution for convex-hull estimators of boundaries, *Scandinavian Journal of Statistics* 33, 139–151.

— (2011), On convex boundary estimation, in I. Van Keilegom and P. W. Wilson, eds., *Exploring Research Frontiers in Contemporary Statistics and Econometrics*, Berlin: Springer-Verlag, pp. 115–150.

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

Jeong, S. O. and L. Simar (2006), Linearly interpolated FDH efficiency score for nonconvex frontiers, *Journal of Multivariate Analysis* 97, 2141–2161.

Jondrow, J., C. A. K. Lovell, I. S. Materov, and P. Schmidt (1982), On the estimation of technical inefficiency in the stochastic frontier production model, *Journal of Econometrics* 19, 233–238.

Kneip, A., B. Park, and L. Simar (1998), A note on the convergence of nonparametric DEA efficiency measures, *Econometric Theory* 14, 783–793.

Kneip, A. and R. Sickles (2011), Panel data, factor models and the Solow residual, in I. Van Keilegom and P. W. Wilson, eds., *Exploring Research Frontiers in Contemporary Statistics and Econometrics*, Berlin: Springer-Verlag, pp. 83–114.

Kneip, A., R. Sickles, and W. Song (2012a), A new panel data treatment for heterogeneity in time trends, *Econometric Theory* 28, 590–628.

Kneip, A. and L. Simar (1996), A general framework for frontier estimation with panel data, *Journal of Productivity Analysis* 7, 187–212.

Kneip, A., L. Simar, and I. Van Keilegom (2012b), Boundary estimation in the presence of measurement error with unknown variance. Discussion paper #2012/02, Institut de Statistique Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

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.

— (2011), A computationally efficient, consistent bootstrap for inference with non-parametric DEA estimators, *Computational Economics* 38, 483–515.

— (2013a), Central limit theorems for DEA scores: When bias can kill the variance. Discussion paper #2013/12, Institut de Statistique Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

— (2013b), Testing issues in nonparametric frontier models. Discussion paper #2013/41, Institut de Statistique Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Knight, K. (2006), Asymptotic theory for m-estimators of boundaries, in S. Sperlich, W. Härdle, and G. Aydinli, eds., *The Art of Semiparametrics*, Berlin: Physica-Verlag, pp. 1–21.

Koopmans, T. C. (1951), An analysis of production as an efficient combination of activities, in T. C. Koopmans, ed., *Activity Analysis of Production and Allocation*, New York: John-Wiley and Sons, Inc., pp. 33–97. Cowles Commission for Research in Economics, Monograph 13.

Korostelev, A., L. Simar, and A. B. Tsybakov (1995a), Efficient estimation of monotone boundaries, *The Annals of Statistics* 23, 476–489.

— (1995b), On estimation of monotone and convex boundaries, *Publications de l'Institut de Statistique de l'Université de Paris XXXIX* 1, 3–18.

Kumbhakar, S. C. and C. A. K. Lovell (2000), *Stochastic Frontier Analysis*, Cambridge: Cambridge University Press.

Kumbhakar, S. C., B. U. Park, L. Simar, and E. G. Tsionas (2007), Nonparametric stochastic frontiers: A local likelihood approach, *Journal of Econometrics* 137, 1–27.

Kuosmanen, T. (2008), Representation thoerems for convex nonparametric least-squares, *Econometrics Journal* 11, 308–325.

Kuosmanen, T. and M. Kortelainen (2012), Stochastic non-smooth envelopment of data: Semi-parametric frontier estimation subject to shape constraints, *Journal of Productivity Analysis* 38, 11–28.

Li, Q. and J. Racine (2007), *Nonparametric Econometrics*, Princeton, NJ: Princeton University Press.

Martins-Filho, C. and F. Yao (2007), A smooth nonparametric conditional quantile frontier estimator, *Journal of Econometrics* 143, 317–333.

Meeusen, W. and J. van den Broeck (1977), Efficiency estimation from Cobb-Douglas production functions with composed error, *International Economic Review* 18, 435–444.

O'Donnell, C. J., D. S. P. Rao, and G. E. Battese (2008), Metafrontier frameworks for the study of firm-level efficiencies and technology ratios, *Empirical Economics* 34, 231–255.

Olson, J. A., P. Schmidt, and D. M. Waldman (1980), A monte carlo study of estimators of stochastic frontier production functions, *Journal of Econometrics* 13, 67–82.

Park, B. U., S.-O. Jeong, and L. Simar (2010), Asymptotic distribution of conical-hull estimators of directional edges, *Annals of Statistics* 38, 1320–1340.

Park, B. U., R. Sickles, and L. Simar (1998), Stochastic panel frontiers: A semiparametric approach, *Journal of Econometrics* 84, 273–301.

— (2003), Corrigendum to 'Semiparametric efficient estmatior of AR(1) panel data models', *Journal of Econometrics* 117, 311.

— (2003a), Semiparametric efficient estmatior of AR(1) panel data models, *Journal of Econometrics* 117, 279–309.

— (2007), Semiparametric efficient estimation in dynamic panel data models, *Journal of Econometrics* 136, 281–301.

Park, B. U. and L. Simar (1994), Efficient semiparametric estimation in a stochastic frontier model, *Journal of the American Statistical Association* 89, 929–936.

Park, B. U., L. Simar, and C. Weiner (2000), FDH efficiency scores from a stochastic point of view, *Econometric Theory* 16, 855–877.

Park, B. U., L. Simar, and V. Zelenyuk (2008), Local likelihood estimation of truncated regression and its partial derivative: Theory and application, *Journal of Econometrics* 146, 185–2008.

Politis, D. N., J. P. Romano, and M. Wolf (2001), On the asymptotic theory of subsampling, *Statistica Sinica* 11, 1105–1124.

Porembski, M., K. Breitenstein, and P. Alpar (2005), Visualizing efficiency and reference relations in data envelopment analysis with an application to the branches of a German bank, *Journal of Productivity Analysis* 23, 203–221.

Seiford, L. M. (1996), Data envelopment analysis: The evolution of the state-of-the-art (1978–1995), *Journal of Productivity Analysis* 7, 99–138.

Shephard, R. W. (1970), *Theory of Cost and Production Functions*, Princeton: Princeton University Press.

Simar, L. (1992), Estimating efficiencies from frontier models with panel data: a comparison of parametric, non-parametric and semi-parametric methods with bootstrapping, *Journal of Productivity Analysis* 3, 171–203.

— (1996), Aspects of statistical analysis in DEA-type frontier models, *Journal of Productivity Analysis* 7, 177–185.

— (2003), Detecting outliers in frontier models: A simple approach, *Journal of Productivity Analysis* 20, 391–424.

— (2007), How to improve the performances of DEA/FDH estimators in the presence of noise, *Journal of Productivity Analysis* 28, 183–201.

Simar, L., I. Van Keilegom, and V. Zelenyuk (2013), Nonparametric least squares methods for stochastic frontier models. Unpublished working paper, Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Simar, L. and A. Vanhems (2012), Probabilistic characterization of directional distances and their robust versions, *Journal of Econometrics* 166, 342–354.

Simar, L., A. Vanhems, and P. W. Wilson (2012), Statistical inference for dea estimators of directional distances, *European Journal of Operational Research* 220, 853–864.

Simar, L. and P. W. Wilson (1998), Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models, *Management Science* 44, 49–61.

— (1999a), Estimating and bootstrapping Malmquist indices, *European Journal of Operational Research* 115, 459–471.

— (1999b), Of course we can bootstrap DEA scores! but does it mean anything? logic trumps wishful thinking, *Journal of Productivity Analysis* 11, 93–97.

— (1999c), Some problems with the Ferrier/Hirschberg bootstrap idea, *Journal of Productivity Analysis* 11, 67–80.

— (2000), A general methodology for bootstrapping in non-parametric frontier models, *Journal of Applied Statistics* 27, 779–802.

— (2001), Testing restrictions in nonparametric efficiency models, *Communications in Statistics* 30, 159–184.

— (2002), Nonparametric tests of returns to scale, *European Journal of Operational Research* 139, 115–132.

— (2007), Estimation and inference in two-stage, semi-parametric models of productive efficiency, *Journal of Econometrics* 136, 31–64.

— (2010), Estimation and inference in cross-sectional, stochastic frontier models, *Econometric Reviews* 29, 62–98.

— (2011a), Inference by the m out of n bootstrap in nonparametric frontier models, *Journal of Productivity Analysis* 36, 33–53.

— (2011b), Two-Stage DEA: Caveat emptor, *Journal of Productivity Analysis* 36, 205–218.

— (2013), Estimation and inference in nonparametric frontier models: Recent developments and perspectives, *Foundations and Trends in Econometrics* 5, 183–337.

Simar, L. and V. Zelenyuk (2006), Statistical inference for aggregates of farrell-type efficiencies, *Journal of Applied Econometrics* Forthcoming.

— (2007), Statistical inference for aggregates of farrell-type efficiencies, *Journal of Applied Econometrics* 27, 1367–1394.

— (2011), Stochastic FDH/DEA estimators for frontier analysis, *Journal of Productivity Analysis* 36, 1–20.

Thanassoulis, E. (2001), *Introduction to the Theory and Application of Data Envelopement Analysis: A Foundation Text with Integrated Software*, New York: Springer.

Wheelock, D. C. and P. W. Wilson (2008), Non-parametric, unconditional quantile estimation for efficiency analysis with an application to Federal Reserve check processing operations, *Journal of Econometrics* 145, 209–225.

Wilson, P. W. (1993), Detecting outliers in deterministic nonparametric frontier models with multiple outputs, *Journal of Business and Economic Statistics* 11, 319–323.

— (1995), Detecting influential observations in data envelopment analysis, *Journal of Productivity Analysis* 6, 27–45.

— (2008), FEAR: A software package for frontier efficiency analysis with R, *Socio-Economic Planning Sciences* 42, 247–254.

— (2011), Asymptotic properties of some non-parametric hyperbolic efficiency estimators, in I. Van Keilegom and P. W. Wilson, eds., *Exploring Research Frontiers in Contemporary Statistics and Econometrics*, Berlin: Springer-Verlag, pp. 115–150.