Estimation and Inference in Nonparametric Frontier Models: Recent Developments and Perspectives

By Léopold Simar and Paul W. Wilson

Contents

1 Nonparametric Statistical Models of Production: Combining Economics and Statistics 185
   1.1 Economic Considerations 185
   1.2 Statistical Considerations 193

2 The Nonparametric Envelopment Estimators 199
   2.1 The Statistical Model 199
   2.2 The FDH Estimator 201
   2.3 The DEA Estimators 207
   2.4 An Alternative Probabilistic Formulation of the DGP 210
   2.5 Properties of FDH and DEA Estimators 216

3 Bootstrap Inference Using DEA and FDH Estimators 229
   3.1 General Idea 229
   3.2 Confidence Intervals for Efficiency of a Particular Point 243
## 4 Robust Order-$m$ Estimators

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>The Need for Robustness</td>
</tr>
<tr>
<td>4.2</td>
<td>Definitions and Basic Ideas</td>
</tr>
<tr>
<td>4.3</td>
<td>Nonparametric Order-$m$ Estimators</td>
</tr>
<tr>
<td>4.4</td>
<td>Statistical Properties</td>
</tr>
<tr>
<td>4.5</td>
<td>Empirical Examples</td>
</tr>
</tbody>
</table>

## 5 Robust Order-$\alpha$ Estimators

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Definition and Basic Ideas</td>
</tr>
<tr>
<td>5.2</td>
<td>Nonparametric Order-$\alpha$ Estimators</td>
</tr>
<tr>
<td>5.3</td>
<td>Statistical Properties</td>
</tr>
<tr>
<td>5.4</td>
<td>Empirical Examples</td>
</tr>
</tbody>
</table>

## 6 Outlier Detection

## 7 Explaining Inefficiency

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Introducing Environmental Variables</td>
</tr>
<tr>
<td>7.2</td>
<td>Two-Stage Regression Approach</td>
</tr>
<tr>
<td>7.3</td>
<td>Conditional Efficiency Measures</td>
</tr>
</tbody>
</table>

## 8 Unanswered Questions, Promising Ideas

Acknowledgments

References
Estimation and Inference in Nonparametric Frontier Models: Recent Developments and Perspectives

Léopold Simar\(^1\) and Paul W. Wilson\(^2\)

\(^1\) Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium, leopold.simar@uclouvain.be

\(^2\) Department of Economics and School of Computing, 228 Sirrine Hall, Clemson University, Clemson, South Carolina, 29634–1309, USA, pww@clemson.edu

Abstract

Nonparametric estimators are widely used to estimate the productive efficiency of firms and other organizations, but often without any attempt to make statistical inference. Recent work has provided statistical properties of these estimators as well as methods for making statistical inference, and a link between frontier estimation and extreme value theory has been established. New estimators that avoid many of the problems inherent with traditional efficiency estimators have also been developed; these new estimators are robust with respect to outliers and avoid the well-known curse of dimensionality. Statistical properties, including asymptotic distributions, of the new estimators
have been uncovered. Finally, several approaches exist for introducing environmental variables into production models; both two-stage approaches, in which estimated efficiencies are regressed on environmental variables, and conditional efficiency measures, as well as the underlying assumptions required for either approach, are examined.
The economic theory underlying analysis of efficiency in production dates at least to the work of Koopmans (1951), Debreu (1951), and Farrell (1957). Farrell made the first attempt to estimate efficiency from a set of observed production units, but the statistical properties of his estimator were only considered much later.

The discussion that follows introduces basic concepts and notation; *Economic Considerations* introduces an economic framework, to which *Statistical Considerations* adds a statistical paradigm.

### 1.1 Economic Considerations

Producers transform inputs into outputs; for example, in manufacturing, inputs typically include labor, capital, energy, materials, and perhaps other things, while outputs are the products produced. There may be one, several, perhaps many different products that are produced. Of course, production is constrained by what is possible or feasible. Let \( x \in \mathbb{R}_+^p \) and \( y \in \mathbb{R}_+^q \) denote vectors of input and output quantities, respectively, and let

\[
P = \{(x, y) \mid x \text{ can produce } y\}
\]  

(1.1)
denote the set of feasible combinations of inputs and outputs, i.e., the production set. Any output quantities \( y \) can be produced using input quantities \( x \) if and only if \((x, y) \in \mathcal{P}\). However, the points in \( \mathcal{P} \) are not equally desirable.

The following three assumptions regarding \( \mathcal{P} \) are standard in microeconomic theory of the firm; see, for example, Shephard (1970) and Färe (1988).

**Assumption 1.1.** \( \mathcal{P} \) is closed.

**Assumption 1.2.** All production requires use of some inputs: \((x, y) \not\in \mathcal{P}\) if \( x = 0 \) and \( y \geq 0, y \neq 0\).\(^1\)

**Assumption 1.3.** Both inputs and outputs are freely disposable: if \((x, y) \in \mathcal{P}\), then for any \((x', y')\) such that \( x' \geq x \) and \( y' \leq y \), \((x', y') \in \mathcal{P}\).

Assumption 1.1 ensures that the boundary of \( \mathcal{P} \) is included in \( \mathcal{P} \). Assumption 1.2 means that there are no “free lunches.” The free disposability assumption is sometimes called strong disposability and is equivalent to an assumption of monotonicity of the technology. This property also characterizes the technical possibility of wasting resources (i.e., the possibility of producing less with more resources).

For purposes of efficiency measurement, the upper boundary of \( \mathcal{P} \) is relevant. The efficient subset of points in \( \mathcal{P} \) is the upper boundary (frontier) of \( \mathcal{P} \), i.e., the locus of optimal production plans (e.g., minimal achievable input level for a given output, or maximal achievable output given the level of the inputs). The upper boundary of \( \mathcal{P} \),

\[
\mathcal{P}^\partial = \{(x, y) \in \mathcal{P} \mid (\gamma^{-1} x, \gamma y) \not\in \mathcal{P} \forall \gamma \in (1, \infty)\} \tag{1.2}
\]

is sometimes referred to as the *technology* or the *production frontier*, and is given by the intersection of \( \mathcal{P} \) and the closure of its complement.

---

\(^1\)Throughout, inequalities involving vectors are assumed to hold element by element; e.g., \( a \leq b \) denotes \( a_j \leq b_j \) for each \( j = 1, \ldots, k \), where \( k \) is the length of \( a \) and \( b \).
Firms that are technically inefficient operate at points in the interior of $P$, while those that are technically efficient operate somewhere along the technology defined by $P^\partial$.

Various features of the production set $P$ and its frontier $P^\partial$ are often of interest to applied researchers. One such feature is returns to scale. Strictly speaking, returns to scale is a feature of the frontier, $P^\partial$, but it is common to ascribe such features to the set $P$. There are several possibilities.

**Definition 1.1.** The frontier $P^\partial$ displays globally constant returns to scale (CRS) if and only if $(\alpha x, \alpha y) \in P \forall (x, y) \in P$ and $\alpha \in [0, \infty)$.

**Definition 1.2.** The frontier $P^\partial$ displays globally variable returns to scale (VRS) if and only if for any $(x, y) \in P$, there exist constants $a(x, y) \in \mathbb{R}_+^1$, $b(x, y) \in \mathbb{R}_+^{1+}$ such that $a(x, y) \leq b(x, y)$ and $(\alpha x, \alpha y) \in P \forall \alpha \in [a(x, y), b(x, y)]$.

**Definition 1.3.** The frontier $P^\partial$ displays globally nonincreasing returns to scale (NIRS) if and only if $(\alpha x, \alpha y) \in P \forall (x, y) \in P$ and $\alpha \in [0, 1]$.

Note that Definition 1.2 encompasses Definitions 1.1 and 1.3; i.e., CRS and NIRS are special cases of VRS. In the same way, Definition 1.3 encompasses Definition 1.1 in that CRS is a special case of NIRS. In other words, assuming either CRS or NIRS is more restrictive than assuming VRS; assuming CRS is more restrictive that assuming NIRS.

The production set $P$ can also be described by its sections or level sets. For instance, the input requirement set for some $y \in \mathbb{R}^q_+$ is given by

$$X(y) = \{x \in \mathbb{R}^p_+ \mid (x, y) \in P\},$$

i.e., the set of all input vectors $x$ that can produce output vector $y$. The boundary of this set, i.e., the (input-oriented) efficiency boundary $X^\partial(y)$, is defined for a given $y \in \mathbb{R}^q_+$ by

$$X^\partial(y) = \{x \mid x \in X(y), \theta x \notin X(y), \forall \theta \in (0, 1)\}.$$
Alternatively, the output feasibility set for some \( x \in \mathbb{R}^p_+ \) is defined by
\[
\mathcal{Y}(x) = \{ y \in \mathbb{R}^q_+ | (x, y) \in \mathcal{P} \},
\]
which gives the set of all output vectors \( y \) than can be produced with given input quantities \( x \). The (output-oriented) efficiency boundary \( \mathcal{Y}^\partial(x) \) is defined, for a given \( x \in \mathbb{R}^p_+ \), as
\[
\mathcal{Y}^\partial(x) = \{ y | y \in \mathcal{Y}(x), \lambda y \notin \mathcal{Y}(x), \forall \lambda > 1 \}.
\]

Then the production set \( \mathcal{P} \) corresponds to the union of all sets \( x(y) \) over all \( y \in \mathbb{R}^q_+ \), or to the union of all sets \( \mathcal{Y}(x) \) over all \( x \in \mathbb{R}^p_+ \).

The Debreu–Farrell input measure of technical efficiency for a given point \( (x, y) \in \mathbb{R}^{p+q}_+ \) is given by
\[
\theta(x, y | \mathcal{P}) = \inf \{ \theta | \theta x \in \mathcal{X}(y) \}
= \inf \{ \theta | (\theta x, y) \in \mathcal{P} \}.
\]

Note that this measure is defined for some points in \( \mathbb{R}^{p+q}_+ \) not necessarily in \( \mathcal{P} \) (i.e., points for which a solution exists in (1.7)). Given an output level \( y \), and an input mix (a direction) given by the vector \( x \), the corresponding efficient level of input is given by
\[
x^\partial(y) = \theta(x, y | \mathcal{P}) x,
\]
which is the projection of \( (x, y) \) onto the efficient boundary \( \mathcal{P}^\partial \), along the ray \( x \) and orthogonal to the vector \( y \).

Figure 1.1 illustrates a point \( (x_0, y_0) \in \mathcal{P} \) for \( p = q = 1 \). The level set \( x^\partial(y) \) defined in (1.8) contains just one point in Figure 1.1; in terms of the labels on the horizontal axis, \( \theta(x_0, y_0 | \mathcal{P}) = x^\partial(y_0)/x_0 < 1 \).

In general, for \( (x, y) \in \mathcal{P} \), \( \theta(x, y | \mathcal{P}) \) gives the feasible proportionate reduction of inputs that a unit located at \( (x, y) \) could undertake to become technically efficient. By construction, for all \( (x, y) \in \mathcal{P} \), \( \theta(x, y | \mathcal{P}) \in (0, 1] \); \( (x, y) \) is technically efficient if and only if \( \theta(x, y | \mathcal{P}) = 1 \). This measure is the reciprocal of the Shephard (1970) input distance function.

Similarly, in the output direction, the Debreu–Farrell output measure of technical efficiency is given by
\[
\lambda(x, y | \mathcal{P}) = \sup \{ \lambda | \lambda y \in \mathcal{Y}(x) \}
= \sup \{ \lambda | (x, \lambda y) \in \mathcal{P} \}
\]
1.1 Economic Considerations

for \((x, y) \in \mathbb{R}^{p+q}_+\). Analogous to the input-oriented case described above, 
\(\lambda(x, y \mid P)\) gives the feasible proportionate increase in outputs for a 
unit operating at \((x, y) \in P\) that would achieve technical efficiency. 
By construction, for all \((x, y) \in P\), \(\lambda(x, y \mid P) \in [1, \infty)\) and \((x, y)\) is 
technically efficient if and only if \(\lambda(x, y \mid P) = 1\).

The output efficiency measure \(\lambda(x, y \mid P)\) is the reciprocal of the 
Shephard (1970) output distance function. The efficient level of output, for the input level \(x\) and for the direction of the output vector 
determined by \(y\), is given by

\[
y^{\beta}(x) = \lambda(x, y \mid P)y. \tag{1.10}
\]

Figure 1.2 illustrates the same point \((x_0, y_0) \in P\) shown in 
Figure 1.1. Here, the set \(y^{\beta}(x_0)\) defined by (1.10) also contains a single 
point, and in terms of the labels on the vertical axis in Figure 1.2, 
\(\lambda(x_0, y_0 \mid P) = y^{\beta}(x_0)/y_0 > 1\).

Efficiency can be measured in other directions, although care should 
be taken to avoid having efficiency measures depend on units of measurement for inputs or outputs. For example, a hyperbolic measure of efficiency is given by

\[
\gamma(x, y \mid P) = \sup\{\gamma \mid (\gamma^{-1}x, \gamma y) \in P\} \tag{1.11}
\]
Fig. 1.2 Output efficiency measure.

for \((x, y) \in \mathbb{R}_+^{p+q}\). This hyperbolic measure of efficiency gives the simultaneous proportionate, feasible reduction in input levels and the proportionate, feasible increase in output levels for a unit operating at \((x, y) \in \mathcal{P}\) that would result in technical efficiency, and is the reciprocal of the hyperbolic graph measure of efficiency defined by Färe et al. (1985).

In terms of the illustration in Figure 1.3, the firm operating at the point \((x_0, y_0) \in \mathcal{P}\) can become technically efficient by moving along the curved (hyperbolic) path from \((x_0, y_0)\) to \((x_0^\gamma(x_0, y_0), y_0^\gamma(x_0, y_0))\). By construction, \(\gamma(x, y | \mathcal{P}) \in [1, \infty)\) for all \((x, y) \in \mathcal{P}\); in addition, \((x, y)\) is technically efficient if and only if \(\gamma(x, y | \mathcal{P}) = 1\). For \((x, y)\) in the interior of \(\mathcal{P}\), the corresponding hyperbolic-efficient levels of inputs are outputs that are given by

\[
\begin{align*}
(x_0^\gamma(x, y), y_0^\gamma(x, y)) &= (\gamma(x, y | \mathcal{P})^{-1}x, \gamma(x, y | \mathcal{P})y).
\end{align*}
\]

Alternatively, Chambers et al. (1996) introduced the directional efficiency measure defined by

\[
\delta(x, y | u, v, \mathcal{P}) = \sup\{\delta \mid (x - \delta u, y + \delta v) \in \mathcal{P}\},
\]

where \(u\) and \(v\) are direction vectors with \(u \in \mathbb{R}_+^p\), \(v \in \mathbb{R}_+^q\), and \([u' v'] \neq 0\). This distance function projects a point \((x, y)\) onto the
1.1 Economic Considerations

Fig. 1.3 Hyperbolic efficiency measure. The frontier $P^\partial$ in the direction $(-u, v)$, with

$$
\left( x_0^\partial(x, y \mid u, v), y_0^\partial(x, y \mid u, v) \right) =
(x - \delta(x, y \mid u, v)u, y + \delta(x, y \mid u, v)v)
$$

(1.14)
giving the directionally-efficient (in the direction $(u, v)$) levels of inputs and outputs.

The directional distance function is illustrated in Figure 1.4. Setting $u = x_0, v = y_0$, the firm operating at $(x_0, y_0)$ becomes technically efficient when it moves to $(x_0^\partial(x_0, y_0), y_0^\partial(x_0, y_0))$ defined by (1.14) (the notation indicating dependence on direction vectors is suppressed here and in Figure 1.4 to conserve space). By construction, $\delta(x, y \mid u, v, P) \in [0, \infty)$ for all $(x, y) \in P$; a point $(x, y) \in P$ is technically efficient if and only if $\delta(x, y \mid u, v, P) = 0$.

Färe et al. (2008, p. 534) state that the directional distance function is independent of units of measurement in the sense that

$$
\delta(\alpha_x \circ x, \alpha_y \circ y \mid \alpha_x \circ u, \alpha_y \circ v, P) = \delta(x, y \mid u, v, P),
$$

(1.15)

where $\alpha_x \in \mathbb{R}^{p+}_+, \alpha_y \in \mathbb{R}^{q+}_+$, and $\circ$ denotes the Hadamard product.\footnote{The Hadamard product of two arrays $A = [a_{ij}]$ and $B = [b_{ij}]$ with the same dimensions is given by the array $C = [c_{ij}]$ having the same dimensions as $A$ and $B$ where $c_{ij} = a_{ij}b_{ij}$;}

However, while (1.15) is true, it also indicates that if units of
measurement for inputs or outputs are changed, the corresponding direction vector must be rescaled to avoid changing the value of the directional distance function. Instead of being homogeneous of degree zero with respect to inputs and outputs, the directional distance function is only homogeneous of degree zero with respect to inputs, outputs, and direction vectors.

This feature of the directional distance function makes the range of reasonable choices for the direction vectors less broad than has been suggested in the literature. For example, Färe et al. (2008, p. 533) note that the direction vectors should be specified in the same units as the inputs and outputs, but then go on to suggest choosing $u = 1$, $v = 1$ or to optimize $u$ and $v$ to minimize distance to the (estimated) frontier. But, if one specifies $u = 1$, $v = 1$, and then changes the units of measurement, this will require re-scaling also $u$ and $v$ so that they no longer equal unity in order to avoid changing the value of the distance function. Hence the choice of $(1, 1)$ for $(u, v)$ is arbitrary, and therefore rather meaningless. Moreover, if the direction vectors are optimized to

c.e.g., see Marcus and Kahn (1959), Marcus and Thompson (1963), and Johnson (1974a, 1974b).
1.2 Statistical Considerations

minimize distance to the estimated frontier, then the results will be sensitive to the units of measurement that are used.

It is easy to show that if $u = x$ and $v = 0$, then

$$\delta(x, y \mid u = x, v = 0, P) = 1 - \theta(x, y \mid P).$$

(1.16)

Similarly, if $u = 0$ and $v = y$, then

$$\delta(x, y \mid u = 0, v = y, P) = \lambda(x, y \mid P) - 1.$$  

(1.17)

A common choice, when $x \in \mathbb{R}^p_{++}$ and $y \in \mathbb{R}^q_{++}$, is to set $u = x$ and $v = y$. One can also set the direction vectors equal to the sample means of inputs and outputs in order to use a common direction for all observations. For additional properties of the directional distance function, see Chambers et al. (1996).

Both the hyperbolic and the directional measures are measures of technical efficiency, as are the input- and output-oriented measures discussed above. Technical efficiency refers to what is possible, but as suggested earlier, not everything that is possible is desirable. Firms may want to maximize profits, which requires considering the prices of inputs and outputs in addition to their quantities. In the case of government provision of goods and services, or in regulated industries, producers’ goals may be cost minimization or perhaps revenue maximization.

A variety of assumptions on $P$ are found in the literature (e.g., free disposability, convexity, etc.; see Shephard, 1970 for examples). The assumptions about $P$ determine the appropriate estimator that should be used to estimate $P^\partial$, $\theta(x, y \mid P)$, $\lambda(x, y \mid P)$, $\gamma(x, y)$, or $\delta(x, y)$. This issue will be discussed next.

1.2 Statistical Considerations

In real-world research problems, the attainable set $P$, as well as $X(y)$, $X^\partial(y)$, $Y(x)$, and $Y^\partial(x)$ are unknown to the analyst. Consequently, the efficiency scores $\theta(x, y \mid P)$, $\lambda(x, y \mid P)$, $\gamma(x, y \mid P)$, and $\delta(x, y \mid u, v, P)$ corresponding to a particular unit operating at $(x, y) \in P$ are also unknown.
In ordinary settings, the only information available to the researcher is a sample

\[ S_n = \{(X_i, Y_i), i = 1, \ldots, n\} \]  

\[ (1.18) \]
of observations on input and output levels for a set of production units engaged in the activity of interest.\(^3\) The statistical paradigm raises the following question that must be answered: what can be learned by observing \( S_n \)? In other words, how can the information in \( S_n \) be used to estimate \( \theta(x, y | \mathcal{P}) \), \( \lambda(x, y | \mathcal{P}) \), \( \gamma(x, y | \mathcal{P}) \), \( \delta(x, y | u, v, \mathcal{P}) \), or other things of interest?

Answering these questions involves much more than reading the data in \( S_n \) into a computer program and pushing some buttons on the keyboard to solve some linear programs. A relevant question is, “what is learned from an estimate of \( \theta(x, y | \mathcal{P}) \), \( \lambda(x, y | \mathcal{P}) \), or other numbers computed from \( S_n \)?” The answer is clear and certain: almost nothing. One might learn, for example, that unit A uses less input quantities while producing greater output quantities than unit B, but little else can be learned from estimates of the efficiency measures introduced above without doing some additional work.

Before anything can be learned about \( \theta(x, y | \mathcal{P}) \), \( \lambda(x, y | \mathcal{P}) \), \( \gamma(x, y | \mathcal{P}) \), \( \delta(x, y | u, v, \mathcal{P}) \), or by extension about \( \mathcal{P} \) and its various features, one must use methods of statistical analysis to understand the properties of whatever estimators have been used to obtain estimates of the things of interest.\(^4\) This raises the following questions: Is the estimator consistent? Is the estimator biased? If the estimator is biased, does the bias disappear as the sample size tends toward infinity? If the estimator is biased, can the bias be corrected, and at what cost; i.e., does correcting the bias introduce too much noise? Can confidence intervals for the values of interest be estimated, and if so, how? How might one test interesting hypotheses about the production process? Notions of statistical consistency, etc. are discussed below.

---

\(^3\)Following standard notation, random variables are denoted by upper-case letters, and realizations of random variables and other nonstochastic quantities by lower-case letters.

\(^4\)Note that an estimator is a random variable, while an estimate is a realization of an estimator (random variable). An estimator can take perhaps infinitely many values with different probabilities, while an estimate is merely a known, nonrandom value.
Before these questions can be answered, a statistical model must be defined; without a statistical model, one cannot know what is estimated. Statistical models consist of two parts: (i) a probability model, which in the present case includes assumptions on the production set $\mathcal{P}$ and the distribution of input and output vectors $(\mathbf{x}, \mathbf{y})$ over $\mathcal{P}$; and (ii) a sampling model describing how data are obtained from the probability model. The statistical model provides a theoretical description of the mechanism that yields the data in the sample $\mathcal{S}_n$, and is sometimes called the data-generating process (DGP). In typical research settings, the task is to use the data in $\mathcal{S}_n$ to learn something about the features of the DGP.

In cases where a group of productive units are observed at the same point in time, i.e., where cross-sectional data are observed, it is convenient and often reasonable to assume the sampling process involves independent draws from the probability distribution defined in the DGP’s probability model. With regard to the probability model, one must attempt reasonable assumptions. Of course, there are trade-offs here; the assumptions on the probability model must be strong enough to permit estimation using estimators that have useful properties, and to allow those properties to be deduced, yet not so strong as to impose conditions on the DGP that do not reflect reality. The goal should be, in all cases, to make minimal, flexible assumptions in order to let the data reveal as much as possible about the underlying DGP, as opposed to making strong, untested assumptions that might influence the results of estimation and inference in perhaps large and misleading ways. The assumptions defining the statistical model are of crucial importance, since any inference that might be made will typically be valid only if the assumptions are in fact true.

The above considerations apply equally to parametric as well as nonparametric approaches to estimation and inference. One can imagine a spectrum of estimation approaches, ranging from fully parametric (most restrictive) to fully nonparametric (least restrictive). Fully parametric estimation strategies necessarily involve stronger assumptions on the probability model, which is completely specified in terms of a specific probability distribution function, structural equations, etc. Semi-parametric strategies are less restrictive; in these approaches, some (but
not all) features of the probability model are left unspecified (for example, in a regression setting one might specify parametric forms for some, but not all, of the moments of a distribution function in the probability model). Fully nonparametric approaches assume no parametric forms for any features of the probability model. Instead, only (relatively) mild assumptions on broad features of the probability distribution are made, usually involving assumptions of various types of continuity, degrees of smoothness, etc.

With fully nonparametric approaches to efficiency estimation, no specific analytical function describing the frontier is assumed. In addition, possibly restrictive assumptions on the stochastic part of the model, describing the probabilistic behavior of the observations in the sample with respect to the efficient boundary of $\mathcal{P}$, are also avoided. There is, however, a cost for this flexibility; in particular, all observed input–output pairs $(X_i, Y_i)$ are assumed to be technically attainable; observations $(X_i, Y_i)$ on input, output vectors are assumed to be drawn randomly and independently from a population of firms whose input–output vectors are distributed on the attainable set $\mathcal{P}$ according to some unknown probability law described by a probability density function $f(x, y)$ or the corresponding distribution function $F(x, y) = \Pr(X \leq x, Y \leq y)$, with

$$\Pr(X_i, Y_i) \in \mathcal{P} = 1.$$  \hspace{1cm} (1.19)

By contrast, fully parametric approaches to efficiency estimation developed by Aigner et al. (1977), Meeusen and van den Broeck (1977), Battese and Corra (1977), Jondrow et al. (1982), and others allow some observations to lie outside the production set $\mathcal{P}$ by incorporating a (two-sided) stochastic term reflecting measurement error or other noise in addition to a (one-sided) stochastic term reflecting inefficiency. Introduction of the stochastic noise term, however, incurs a cost: some parametric structure is required for such models to be identified, which in turn requires assumptions that may or may not be supported by data. In addition, such models typically allow for only a single response variable, i.e., a single output variable in a production framework; researchers typically work in a cost framework when there are multiple outputs, but this in turn requires data on input-prices.
To the extent that the fully parametric approach allows for measurement error, it only does so for the response variable, and not for any of the explanatory variables. Perhaps most problematic, writing the model in a regression framework introduces issues of causality and exogeneity that do not arise in the fully nonparametric approach, which more closely resembles an exercise in density estimation.

The fully parametric approaches are often called stochastic frontier analysis, while the fully nonparametric approaches are frequently called deterministic frontier analysis. This terminology is unfortunate since it is misleading and has created a good bit of confusion in the literature. In both approaches, there is only one frontier, and it is fixed, not stochastic. In both approaches, the location of the frontier is unknown, and this is what necessitates estimation and gives rise to uncertainty. In both approaches, the distance from a given observation \((X_i, Y_i)\) to the frontier (in any direction) is unknown, and must be estimated.

The most popular nonparametric efficiency estimators are based on the idea of estimating the attainable set \(P\) by the smallest set \(\hat{P}\) within some class of sets that envelop the observed data. Depending on assumptions made on \(P\), this idea leads to the Free Disposal Hull (FDH) estimator of Deprins et al. (1984), which relies only on an assumption of free disposability, and the Data Envelopment Analysis (DEA) estimators which incorporate additional assumptions. Farrell (1957) was the first to use a DEA estimator in an empirical application, but the idea remained obscure until it was popularized by Charnes et al. (1978) and Banker et al. (1984). Charnes et al. estimated \(P\) by the convex cone of the FDH estimator of \(P\), thus imposing an assumption of constant returns to scale, while Banker et al. used the convex hull of the FDH estimator of \(P\), thereby allowing for VRS.

The primary advantage of nonparametric models and estimators lies in their great flexibility (as opposed to parametric, deterministic frontier models). In addition, the nonparametric estimators are easy to compute, and today most of their statistical properties are well-established. As will be discussed below, inference is available using bootstrap methods.

The main drawbacks of the fully nonparametric DEA and FDH estimators is that they are very sensitive to outliers and extreme values,
and that noisy data are not allowed. Fortunately, robust alternatives to DEA and FDH estimators are available for use in fully nonparametric models; these alternative approaches will be described later. Also, as discussed below in the last section of this survey, “stochastic” versions of DEA and FDH estimators are the object of current research.

It should be noted that allowing for noise in frontier models presents difficult problems even in a fully parametric framework where one can rely on the assumed parametric structure. In fully parametric models where the DGP involves a one-sided error process reflecting inefficiency and a two-sided error process reflecting statistical noise, numerical identification of the statistical model’s features is sometimes highly problematic even with large (but finite) samples; see Ritter and Simar (1997) for examples.

Apart from the issue of numerical identification, fully parametric frontier models that incorporate a noise term present other difficulties. Efficiency estimates in these models are based on residual terms that are unidentified. Researchers instead base efficiency estimates on an expectation, conditional on a composite residual; estimating an expected inefficiency is rather different from estimating actual inefficiency. An additional problem arises from the fact that, even if the fully parametric, stochastic frontier model is correctly specified, there is typically a nontrivial probability of drawing samples with the “wrong” skewness (e.g., when estimating cost functions, one would expect composite residuals with right-skewness, but it is certainly possible to draw finite samples with left-skewness — the probability of doing so depends on the sample size and the mean of the composite errors). Since there are apparently no published studies, and also apparently no working papers in circulation, where researchers report composite residuals with the “wrong” skewness when fully parametric, stochastic frontier models are estimated, it appears that estimates are sometimes, perhaps often, conditioned (i) on either drawing observations until the desired skewness is obtained or (ii) on model specifications that result in the desired skewness. This raises formidable questions for inference; see Simar and Wilson (2010) for discussion.
2

The Nonparametric Envelopment Estimators

2.1 The Statistical Model

The assumptions listed below are adapted from Kneip et al. (1998), Park et al. (2000), Kneip et al. (2008), and Park et al. (2010). These assumptions define a statistical model (i.e., a DGP), are very flexible, and seem quite reasonable in many practical situations.

The first assumption below reflects a deterministic frontier model as discussed earlier, and implies the condition in (1.19). In addition, for sake of simplicity, the standard independence hypothesis is assumed, meaning that the observed firms are considered as being drawn randomly and independently from a population of firms.

Assumption 2.1. The sample observations \((X_i, Y_i)\) in \(S_n\) are realizations of identically, independently distributed (iid) random variables \((X, Y)\) with probability density function \(f(x, y)\), which has compact support over \(D \subset \mathcal{P} \subset \mathbb{R}_+^{n+q}\).

The compact set \(D\) is introduced in Assumption 2.1 for technical reasons and is used in proofs of consistency of DEA and FDH
estimators; essentially, the assumption rules out use of infinite quantities of one or more inputs.

The next assumption is a regularity condition sufficient for proving the consistency of all the nonparametric estimators described in this section. It says that the probability of observing firms in any open neighborhood of the frontier is strictly positive — quite a reasonable property since microeconomic theory indicates that with competitive input and output markets, firms which are inefficient will, in a long run, be driven from the market.

**Assumption 2.2.** The density \( f(x, y) \) is strictly positive on the boundary \( P^b \) of the production set \( P \) and is continuous in any direction toward the interior of \( P \).

This assumption is needed to establish consistency of DEA and FDH estimators along the lines of Kneip et al. (1998) and Park et al. (2000). In the case of output-oriented FDH estimators with \( q = 1 \), Daouia et al. (2010) relax this assumption, allowing \( f(x, y) \) to approach zero at the frontier, with a consequent change in the rate of convergence.

In order to prove consistency of the estimators, the production frontier must be sufficiently smooth.

**Assumption 2.3.** For all \((x, y)\) in the interior of \( P \), the function \( \theta(x, y \mid P) \) and \( \lambda(x, y \mid P) \) are differentiable in both their arguments.

The characterization of smoothness in Assumption 2.3 is stronger than required for the consistency of the nonparametric estimators. Kneip et al. (1998) require only Lipschitz continuity of the efficiency scores, which is implied by the simpler, but stronger requirement presented here. However, derivation of limiting distributions of the nonparametric estimators has been obtained only with the stronger assumption made here.

Finally, the following assumption is sometimes made.

**Assumption 2.4.** \( P \) is convex: if \((x_1, y_1), (x_2, y_2) \in P\), then \((x, y) \in P\) for \((x, y) = \alpha(x_1, y_1) + (1 - \alpha)(x_2, y_2)\), for all \( \alpha \in [0, 1] \).
This convexity assumption may be dubious in many situations. Several recent studies focus on the convexity assumption in frontier models (e.g., Bogetoft, 1996; Bogetoft et al., 2000; Briec et al., 2004). Assumption 2.4 will be relaxed at various points below.

2.2 The FDH Estimator

The FDH estimator was first proposed by Deprins et al. (1984). It relies on the free disposability assumed in Assumption 1.3, and does not require Assumption 2.4. The FDH estimator \( \hat{P}_{FDH} \) of the attainable set \( \mathcal{P} \) is simply the free disposal hull of the observed sample \( S_n \), given by

\[
\hat{P}_{FDH} = \bigcup\limits_{(X_i, Y_i) \in S_n} \{ (x, y) \in \mathbb{R}^{p+q}_+ \mid x \geq X_i, y \leq Y_i \},
\]

and is the union of \( n \) southeast-orthants with vertices \( (X_i, Y_i) \), where \( n \) is the number of input–output pairs in \( S_n \). Figure 2.1 illustrates the idea when \( p = q = 1 \).

A nonparametric estimator of the input efficiency for a given point \( (x, y) \in \mathbb{R}^{p+q}_+ \) is obtained by replacing the true production set \( \mathcal{P} \) in the definition of \( \theta(x, y \mid \mathcal{P}) \) given by (1.7) with the estimator \( \hat{P}_{FDH} \),
yielding

$$\theta(x, y | \hat{P}_{FDH}) = \inf \{ \theta | (\theta x, y) \in \hat{P}_{FDH} \}. \quad (2.2)$$

Estimates can be computed in two steps: first, identify the set of observed points dominating \((x, y)\):

$$D(x, y) = \{ i | (X_i, Y_i) \in S_n, X_i \leq x, Y_i \geq y \}. \quad (2.3)$$

Then \(\theta(x, y | \hat{P}_{FDH})\) can be computed simply by evaluating

$$\theta(x, y | \hat{P}_{FDH}) = \min_{i \in D(x, y)} \max_{j=1, \ldots, p} \left( \frac{X_j^i}{x^j} \right), \quad (2.4)$$

where for a vector \(a\), \(a^j\) denotes the \(j\)th element of \(a\). Evaluating (2.4) requires only some sorting and some simple logical comparisons; hence the estimator can be computed quickly and easily.

An estimate of the efficient levels of inputs for a given output levels \(y\) and a given input direction determined by the input vector \(x\) is given by

$$\hat{x}^\theta(y) = \theta(x, y | \hat{P}_{FDH})x. \quad (2.5)$$

By construction, \(\hat{P}_{FDH} \subseteq P\), and so \(\hat{P}_{FDH}\) is a downward- (i.e., inward)-biased estimator of \(P\), and \(\hat{X}^\theta(y)\) is an upward-biased estimator of \(X^\theta(y)\). Hence \(\theta(x, y | \hat{P}_{FDH})\) is an upward-biased estimator of \(\theta(x, y | P)\); i.e., \(\theta(x, y | \hat{P}_{FDH}) \geq \theta(x, y | P)\). In other words, technical efficiency is over-stated by \(\theta(x, y | \hat{P}_{FDH})\).

Estimators for the output-orientation can be constructed similarly. The FDH estimator of \(\lambda(x, y | P)\) is defined by

$$\lambda(x, y | \hat{P}_{FDH}) = \sup \{ \lambda | (x, \lambda y) \in \hat{P}_{FDH} \}. \quad (2.6)$$

This can be computed quickly and easily by evaluating

$$\lambda(x, y | \hat{P}_{FDH}) = \max_{i \in D(x, y)} \min_{j=1, \ldots, q} \left( \frac{Y_j^i}{y^j} \right). \quad (2.7)$$

Efficient output levels for given input levels \(x\) and given an output mix (direction) described by the vector \(y\) are estimated by

$$\hat{y}^\theta(x) = \lambda(x, y | \hat{P}_{FDH})y. \quad (2.8)$$
2.2 The FDH Estimator

By construction, \( \lambda(x, y \mid \hat{P}_{FDH}) \) is a downward-biased estimator of \( \lambda(x, y \mid P) \), and hence the estimator over-states technical efficiency.

Using the same plug-in principle used above to define estimators of the input- and output-oriented efficiency measures, the hyperbolic measure of technical efficiency \( \gamma(x, y \mid P) \) can be estimated by replacing \( P \) with \( \hat{P}_{FDH} \) in (1.11) to obtain

\[
\gamma(x, y \mid \hat{P}_{FDH}) = \sup \{ \gamma \mid (\gamma^{-1} x, \gamma y) \in \hat{P}_{FDH} \}.
\] (2.9)

Wilson (2011) shows that this can be computed by solving

\[
\gamma(x, y \mid \hat{P}_{FDH}) = \left[ \min_{i \in D(x, y)} \left( \max_{j=1, \ldots, p} \left( \frac{X^j}{x^j}, \frac{Y^k}{y^k} \right) \right) \right]^{-1}
\]

\[
= \max_{i \in D(x, y)} \left( \min_{j=1, \ldots, p} \left( \frac{X^j}{x^j}, \frac{Y^k}{y^k} \right) \right)
\] (2.10)

(note that the hyperbolic measure defined in Wilson, 2011 is the inverse of the measure defined in (1.11)). Wilson (2011) also introduces the mapping \( \phi : \mathbb{R}^{p+q} \rightarrow \mathbb{R}^{p+q} \) such that \( \phi(x, y) = (x, y^{-1}) \equiv w \), where \( y^{-1} \) is the vector whose elements are the inverses of the corresponding elements of \( y \). Then

\[
\gamma(x, y \mid \hat{P}_{FDH}) = \left[ \min_{i \in D(x, y)} \left( \max_{j=1, \ldots, (p+q)} \left( \frac{W^j}{w^j} \right) \right) \right]^{-1}.
\] (2.11)

This is identical in form to the inverse of an input-oriented estimator \( \hat{\theta}(w) \) in (2.4) in \( w \)-space with no output dimensions.\(^1\)

Working in the hyperbolic orientation, efficient input–output levels corresponding to given input levels \( x \) and given output levels \( y \) are estimated by

\[
\left( \hat{x}^\gamma(x, y), \hat{y}^\gamma(x, y) \right) = \left( \gamma(x, y \mid \hat{P}_{FDH})^{-1} x, \gamma(x, y \mid \hat{P}_{FDH}) y \right).
\] (2.12)

By construction, \( \gamma(x, y \mid \hat{P}_{FDH}) \) is a downward-biased estimator of \( \gamma(x, y \mid P) \), and hence \( \gamma(x, y \mid \hat{P}_{FDH}) \leq \gamma(x, y \mid P) \). Consequently,

\(^1\)To see this, replace \( x \) in (2.4) with \( w \), set \( q = 1, y = 0 \), and replace \( p \) with \( p + q \). The resulting expression is equivalent to (2.11).
\( \hat{x}_\gamma(x, y) \) is biased upward while \( \hat{y}_\gamma(x, y) \) is biased downward in the sense that \( \hat{x}_\gamma(x, y) \geq \gamma(x, y \mid P)^{-1} x \) and \( \hat{y}_\gamma(x, y) \leq \gamma(x, y \mid P) y \).

The same plug-in approach used to define the distance function estimators in (2.2), (2.6), and (2.9) can also be used to define an estimator of the directional distance function given in (1.13) by writing

\[
\delta(x, y \mid u, v, \hat{P}_{FDH}) = \max_{\delta} \{ \delta \mid (x - \delta u, y + \delta v) \in \hat{P}_{FDH} \}. \tag{2.13}
\]

If \( u = x > 0 \) and \( v = 0 \) in (1.13), then the resulting directional distance function can be estimated by replacing \( \theta(x, y \mid P) \) in (1.16) with \( \theta(x, y \mid \hat{P}_{FDH}) \). Similarly, if \( u = 0 \) and \( v = y > 0 \) in (1.13), then the resulting directional distance function can be estimated by replacing \( \lambda(x, y \mid P) \) in (1.17) with \( \lambda(x, y \mid \hat{P}_{FDH}) \).

In general, (2.13) can be computed using linear programming methods to solve

\[
\delta(x, y \mid u, v, \hat{P}_{FDH}) = \max_{\delta, \omega} \{ \delta \mid y + \delta v \leq Y \omega, \quad x - \delta u \geq X \omega, \quad \omega \in \{0, 1\}^n \}. \tag{2.14}
\]

However, simpler and faster approaches are available. If all elements of the direction vectors \( u \) and \( v \) in (1.13) are strictly positive, estimates of the directional distance function can be computed easily using a simple transformation suggested by Simar and Vanhems (2012). For \( u \in \mathbb{R}^p_+ \) and \( v \in \mathbb{R}^q_+ \), let

\[
x^\dagger = \exp \left( x \circ u^{(-1)} \right) \tag{2.15}
\]

and

\[
y^\dagger = \exp \left( y \circ v^{(-1)} \right), \tag{2.16}
\]

where the superscript \((-1)\) denotes element-wise inverse of the direction vectors \( u, v \). Define the transformed production set

\[
P^\dagger := \left\{(x^\dagger, y^\dagger) \mid (x, y) \in P\right\}. \tag{2.17}
\]

Simar and Vanhems (2012) show that

\[
\delta(x, y \mid u, v, P) = \log \gamma(x^\dagger, y^\dagger \mid P^\dagger), \tag{2.18}
\]
again provided that the direction vectors are strictly positive. Hence transforming the observations in $S_n$ and computing

$$
\delta(x, y \mid u, v, \hat{P}_{FDH}) = \log \gamma(x^\dagger, y^\dagger \mid \hat{P}_{FDH})
$$

(2.19)

by using (2.10) gives an estimate of the directional distance function defined in (1.13).

Of course, the transformations in (2.15) and (2.16) are defined only when all elements of the direction vectors $u$ and $v$ are strictly positive. Simar and Vanhems (2012) propose an alternative to the transformation in (2.15) and (2.16) to allow zero elements in the direction vectors. Suppose that $0 \leq p_1 \leq p$ elements of $u$ and $0 \leq q_1 \leq q$ elements of $v$ are strictly positive, and that other elements equal zero. Then if $p_1 = 0$, all elements of $u$ are zero; similarly, if $q_1 = q$, then all elements of $v$ are zero. Without loss of generality, rearrange the elements of $u$ and $v$ so that $u = [u'_1 \ 0'_{p-p_1}]^\top$ and $v = [v'_1 \ 0'_{q-q_1}]^\top$. Let elements of $x$ and $y$ be rearranged and partitioned accordingly, so that $x = [x'_1 \ x'_2]^\top$ and $y = [y'_1 \ y'_2]^\top$. Note that some of the $x_k$, $y_k$, $k \in \{1, 2\}$ could be empty vectors; e.g., if $p_1 = p$, then $x_2$ is empty, and if $p_1 = 0$, then $x_1$ is empty; assume $p_1 + q_1 \geq 1$.

Now consider the following transformation:

$$
x_1^\dagger = \exp (x_1 \circ u_1^{(-1)}),
$$

(2.20)

$$
x_2^\dagger = x_2,
$$

(2.21)

$$
y_1^\dagger = \exp (y_1 \circ v_1^{(-1)}),
$$

(2.22)

and

$$
y_2^\dagger = y_2.
$$

(2.23)

Analogous to (2.17), define the transformed production set

$$
P^\dagger = \{ (x^\dagger, y^\dagger) \mid (x, y) \in P \}.
$$

(2.24)

By definition,

$$
\delta(x, y \mid u, v, P) = \sup \{ \delta \mid ([x_1 - \delta u_1 \ x_2], [y_1 + \delta v_1 \ y_2]) \in P \}
$$

$$
= \log (\gamma^\dagger(x^\dagger, y^\dagger))
$$

(2.25)
The Nonparametric Envelopment Estimators

where

$$\gamma^\dagger(x^\dagger, y^\dagger) = \sup \left\{ \gamma \mid \left( \gamma^{-1} x^\dagger_1 \ 0 \ x^\dagger_2 \right), \left[ \gamma y^\dagger_1 \ 0 \ y^\dagger_2 \right] \in \mathcal{P}^\dagger \right\}. \ (2.26)$$

The last two lines of (2.25) resembles the hyperbolic measure defined in (1.11), except that here, the optimization is over only the dimensions for which the direction vectors have nonzero elements.

The FDH estimator of $$\gamma^\dagger(x^\dagger, y^\dagger)$$ is easy to compute; working in the space of $$(x^\dagger, y^\dagger)$$, let

$$D^\dagger(x^\dagger, y^\dagger) = \{ i \mid X^\dagger_i \leq x^\dagger, \ Y^\dagger_i \geq y^\dagger, \ i = 1, \ldots, n \}, \ (2.27)$$

analogous to (2.3). Then the FDH estimator of $$\gamma^\dagger(x^\dagger, y^\dagger)$$ is

$$\hat{\gamma}^\dagger(x^\dagger, y^\dagger) = \max_{i \in D^\dagger(x^\dagger, y^\dagger)} \left( \min_{j=1, \ldots, p_1, k=1, \ldots, q_1} \left( \frac{x^\dagger_{1,j}}{X^\dagger_{1,i}}, \frac{Y^\dagger_{1,k}}{y^\dagger_{1,i}} \right) \right). \ (2.28)$$

Then the FDH estimate of (2.25) is

$$\delta(x, y \mid u, v, \hat{P}_{FDH}) = \log \hat{\gamma}^\dagger(x^\dagger, y^\dagger).$$

Throughout, reference will be made to the transformation (2.15)–(2.16), but the reader should understand that this should be replaced by the transformation in (2.20)–(2.23) whenever the direction vectors $$u, v$$ contains elements equal to zero.

Using estimates of the directional distance function in (1.13) to project points onto the estimated frontier $$\hat{P}_{FDH}$$, efficient input–output levels corresponding to given input levels $$x$$ and given output levels $$y$$ are estimated by $$\left( \hat{x}_\delta^\dagger(x, y \mid \hat{P}_{FDH}), (\hat{y}_\delta^\dagger(x, y \mid \hat{P}_{FDH}) \right)$$ where

$$\hat{x}_\delta^\dagger(x, y \mid \hat{P}_{FDH}) = x - \delta(x, y \mid u, v, \hat{P}_{FDH})u $$ \ (2.29)

and

$$\hat{y}_\delta^\dagger(x, y \mid \hat{P}_{FDH}) = y + \delta(x, y \mid u, v, \hat{P}_{FDH})v. \ (2.30)$$

Since $$\hat{P}_{FDH}$$ is biased inward, $$\hat{\delta}_{FDH}(x, y \mid u, v, S_n)$$ is necessarily biased downward. Hence, relative to $$x - \delta(x, y \mid u, v, \mathcal{P})x, \hat{x}_\delta^\dagger(x, y \mid \delta(x, y \mid u, v, \hat{P}_{FDH})$$ is biased upward, while $$\hat{y}_\delta^\dagger(x, y \mid \delta(x, y \mid u, v, \hat{P}_{FDH})$$ is biased downward relative to $$y + \delta(x, y \mid u, v, \mathcal{P})v$$. 

2.3 The DEA Estimators

Although DEA estimators were first used by Farrell (1957) to measure technical efficiency for a set of observed firms, the idea did not gain wide acceptance until the paper by Charnes et al. (1978) appeared 21 years later. Charnes et al. used the convex cone (rather than the convex hull) of $\hat{P}_{FDH}$ to estimate $P$, which would be appropriate only if returns to scale are everywhere constant. Later, Banker et al. (1984) used the convex hull of $\hat{P}_{FDH}$ to estimate $P$, thus allowing variable returns to scale. Here, “DEA” refers to both of these approaches, as well as other approaches that involve definition of a convex set enveloping the FDH estimator $\hat{P}_{FDH}$ to estimate $P$.

The most general DEA estimator of the attainable set $P$ is simply the convex hull of $\hat{P}_{FDH}$, i.e.,

$$\hat{P}_{VRS} = \{(x, y) \in \mathbb{R}^{p+q} \mid y \leq Yq, x \geq Xq, i'_nq = 1, q \in \mathbb{R}^n_+\}, \tag{2.31}$$

where $X = [X_1, \ldots, X_n]$ and $Y = [Y_1, \ldots, Y_n]$ are $(p \times n)$ and $(q \times n)$ matrices (respectively) whose columns are the input–output combinations in $S_n$, $q$ is an $(n \times 1)$ vector of weights, and $i_n$ is an $(n \times 1)$ vector of ones. Alternatively, the conical hull of the FDH estimator, $\hat{P}_{CRS}$, used by Charnes et al. (1978), is obtained by dropping the constraint $i_nq = 1$ in (2.31); i.e.,

$$\hat{P}_{CRS} = \{(x, y) \in \mathbb{R}^{p+q} \mid y \leq Yq, x \geq Xq, q \in \mathbb{R}^n_+\}. \tag{2.32}$$

Other DEA estimators of $P$ can be defined by modifying the constraint $i_nq = 1$ in (2.31). For example, the estimator

$$\hat{P}_{NIRS} = \{(x, y) \in \mathbb{R}^{p+q} \mid y \leq Yq, x \geq Xq, i'_nq \leq 1, q \in \mathbb{R}^n_+\} \tag{2.33}$$

allows for nonincreasing returns to scale. In other words, returns to scale along the boundary of $\hat{P}_{NIRS}$ are either constant or decreasing, but not increasing. By contrast, returns to scale along the boundary of $\hat{P}_{VRS}$ are either increasing, constant, or decreasing, while returns to scale along the boundary of $\hat{P}_{CRS}$ are constant everywhere.

Figure 2.2 illustrates the DEA estimator $\hat{P}_{VRS}$ for the case of one input and one output ($p = q = 1$). The FDH estimator $\hat{P}_{FDH}$ is depicted by dotted lines where it does not correspond to $\hat{P}_{VRS}$. 
As with the FDH estimators, DEA estimators of the efficiency scores \( \theta(x, y \mid P) \), \( \lambda(x, y \mid P) \), \( \gamma(x, y \mid P) \), and \( \delta(x, y \mid u, v, P) \) defined in (1.7), (1.9), (1.11), and (1.13) can be obtained using the plug-in method by replacing the true, but unknown, production set \( P \) with one of the estimators \( \hat{P}_{VRS} \), \( \hat{P}_{CRS} \), or \( \hat{P}_{NIRS} \). For example, in the input-orientation, with varying returns to scale, using \( \hat{P}_{VRS} \) to replace \( P \) in (1.7) leads to the estimator
\[
\theta(x, y \mid \hat{P}_{VRS}) = \inf \{ \theta \mid (\theta x, y) \in \hat{P}_{VRS} \}.
\]

As a practical matter, \( \theta(x, y \mid \hat{P}_{VRS}) \) can be computed by solving the linear program
\[
\theta(x, y \mid \hat{P}_{VRS}) = \min \{ \theta > 0 \mid y \leq Y_q, \theta x \geq X_q, i'_n q = 1, q \in \mathbb{R}_+^n \}.
\]

A number of algorithms exist to solve linear programs such as the one in (2.35); in principle, solutions can be obtained easily.\(^2\) However, the computational burden represented by (2.35) is typically greater than that posed by the FDH estimators. In terms of Figure 2.2, for the observation \((X_i, Y_i)\) at point \( C \), \( \theta(X_i, Y_i \mid \hat{P}_{VRS}) = \frac{AB}{AC} \).

\(^2\) The FEAR package (Wilson, 2008) uses a simplex method described by Hadley (1962). One can also use interior-point methods or other algorithms.
These ideas extend naturally to the output orientation. For example, replacing \( \mathcal{P} \) with \( \hat{\mathcal{P}}_{VRS} \) in (1.9) yields

\[
\lambda(x, y \mid \hat{\mathcal{P}}_{VRS}) = \sup \left\{ \lambda \mid (x, \lambda y) \in \hat{\mathcal{P}}_{VRS} \right\},
\]

(2.36)

which can be computed by solving the linear program

\[
\lambda(x, y \mid \hat{\mathcal{P}}_{VRS}) = \{ \lambda \mid \lambda y \leq Y q, x \geq X q, i'_n q = 1, q \in \mathbb{R}^n_+ \}. \quad (2.37)
\]

In either the input or output orientation, estimators of \( \theta(x, y \mid \mathcal{P}) \) and \( \lambda(x, y \mid \mathcal{P}) \) using either \( \hat{\mathcal{P}}_{CRS} \) or \( \hat{\mathcal{P}}_{NIRS} \) can be constructed similarly using the plug-in method. Specifically, estimators \( \theta(x, y \mid \hat{\mathcal{P}}_{CRS}) \) and \( \lambda(x, y \mid \hat{\mathcal{P}}_{CRS}) \) are obtained by dropping the constraint \( i'_n q = 1 \) on the right-hand sides of (2.35) and (2.37), while estimators of \( \theta(x, y \mid \hat{\mathcal{P}}_{NIRS}) \) and \( \lambda(x, y \mid \hat{\mathcal{P}}_{NIRS}) \) are obtained by changing the constraint \( i'_n q = 1 \) on the right-hand sides of (2.35) and (2.37) to \( i'_n q \leq 1 \).

Under CRS, estimation of the hyperbolic efficiency measure \( \gamma(x, y \mid \mathcal{P}) \) in (1.11) is straightforward since it is easy to show that

\[
\gamma(x, y \mid \mathcal{P}) = \theta(x, y \mid \mathcal{P})^{-1/2} = \lambda(x, y \mid \mathcal{P})^{1/2}.
\]

(2.38)

Hence, under CRS, the DEA estimator of \( \gamma(x, y \mid \mathcal{P}) \) can be defined by either \( \theta(x, y \mid \hat{\mathcal{P}}_{CRS})^{-1/2} \) or \( \lambda(x, y \mid \hat{\mathcal{P}}_{CRS})^{1/2} \). But under VRS or NIRS, no such relationship exists, and substituting \( \hat{\mathcal{P}}_{VRS} \) or \( \hat{\mathcal{P}}_{NIRS} \) into (1.11) results in a difficult-to-solve nonlinear programming problem. Wilson (2011) describes a numerical method for computing \( \gamma(x, y \mid \hat{\mathcal{P}}_{VRS}) \) or \( \gamma(x, y \mid \hat{\mathcal{P}}_{NIRS}) \) that can be made accurate to an arbitrary degree by choosing a priori the convergence tolerance to be appropriately small. This method is used in the FEAR software package described by Wilson (2008).

DEA estimators of the directional efficiency measure defined in (1.13) can similarly be constructed using the now-familiar plug-in method. Replacing \( \mathcal{P} \) in (1.13) with \( \hat{\mathcal{P}}_{VRS} \) results in an estimator that can be computed by solving the linear program

\[
\delta(x, y \mid u, v, \hat{\mathcal{P}}_{VRS}) = \max \left\{ \delta \mid y + \delta v \leq Y q, x - \delta u \geq X q, \right. \\
\left. i'_n q = 1, q \in \mathbb{R}^n_+ \right\}
\]

(2.39)

As in the case of the input- and output-oriented measures, estimators \( \delta(x, y \mid u, v, \hat{\mathcal{P}}_{CRS}) \) or \( \delta(x, y \mid u, v, \hat{\mathcal{P}}_{NIRS}) \) can be obtained by either
dropping the constraint $i_n'q = 1$ in (2.39) or changing it to $i_n'q \leq 1$ (respectively).

Technically efficient levels of inputs and outputs can be estimated by plugging the appropriate DEA efficiency estimator among those described above into either (1.8), (1.10), (1.12), or (1.14). For example, under VRS, in the input orientation the technically efficient level of inputs for a given level of outputs $y$, is estimated by $\theta(x, y | \hat{P}_{VRS})x$, while in the output orientation, the technically efficient level of outputs for a given level of inputs $x$ is estimated by $\lambda(x, y | \hat{P}_{VRS})y$.

All of the FDH and DEA estimators are biased by construction since $\hat{P}_{FDH} \subseteq \hat{P}_{VRS} \subseteq P$. Moreover, $\hat{P}_{VRS} \subseteq \hat{P}_{NIRS} \subseteq \hat{P}_{CRS}$. If the technology $P^O$ exhibits globally constant returns to scale, then $\hat{P}_{CRS} \subseteq P$; otherwise, $\hat{P}_{CRS}$ will not be a statistically consistent estimator of $P$. Of course, if $P$ is not convex, then $\hat{P}_{VRS}$ will also be inconsistent. These relations further imply that $\theta(x, y | \hat{P}_{FDH}) \geq \theta(x, y | \hat{P}_{VRS}) \geq \theta(x, y | P)$ and $\theta(x, y | \hat{P}_{VRS}) \geq \theta(x, y | \hat{P}_{NIRS}) \geq \theta(x, y | \hat{P}_{CRS})$. Alternatively, in the output orientation, $\lambda(x, y | \hat{P}_{FDH}) \leq \lambda(x, y | \hat{P}_{VRS}) \leq \lambda(x, y | P)$ and $\lambda(x, y | \hat{P}_{VRS}) \leq \lambda(x, y | \hat{P}_{NIRS}) \leq \lambda(x, y | \hat{P}_{CRS})$. Similar relations hold for estimators of the hyperbolic and directional efficiency measures.

### 2.4 An Alternative Probabilistic Formulation of the DGP

The description above of the DGP is traditional. However, the DGP can also be described in terms that allow a probabilistic interpretation of the Debreu–Farrell efficiency scores, providing a new way of describing the nonparametric FDH and DEA estimators. This new formulation is useful for introducing extensions of the FDH and DEA estimators described above, as will be seen later, and for linking frontier estimation to extreme value theory as explained by Daouia et al. (2010). The presentation here follows that of Daraio and Simar (2005), who extend the ideas of Cazals et al. (2002).

The stochastic part of the DGP introduced above through the probability density function $f(x, y)$ (or the corresponding distribution function $F(x, y)$) is completely characterized by the following
probability function:

\[ H_{XY}(x, y) = \Pr(X \leq x, Y \geq y). \]  \hspace{1cm} (2.40)

Note that this is not a standard distribution function, since the cumulative form is used for the inputs \( x \) and the survival form is used for the outputs \( y \). Nonetheless, \( H_{XY}(x, y) \) is well-defined.

The distribution function in (2.40) has a nice interpretation and some interesting properties. First, \( H_{XY}(x, y) \) gives the probability that a unit operating at input, output levels \((x, y)\) is dominated, i.e., that another unit produces at least as much output while using no more of any input than the unit operating at \((x, y)\). Second, \( H_{XY}(x, y) \) is monotone, nondecreasing in \( x \) and monotone nonincreasing in \( y \). Third, the support of the distribution function \( H_{XY}(\cdot, \cdot) \) is the attainable set \( \mathcal{P} \); i.e.,

\[ H_{XY}(x, y) = 0 \quad \forall \ (x, y) \notin \mathcal{P}. \]  \hspace{1cm} (2.41)

The joint probability \( H_{XY}(x, y) \) can be decomposed using Bayes’ rule by writing

\[ H_{XY}(x, y) = \underbrace{\Pr(X \leq x \mid Y \geq y)}_{=F_{X|Y}(x \mid y)} \Pr(Y \geq y) = \underbrace{S_Y(y)}_{\text{survivor function of } Y} \]  \hspace{1cm} (2.42)

\[ = \Pr(Y \geq y \mid X \leq x) \Pr(X \leq x), \]  \hspace{1cm} (2.43)

where \( S_Y(y) = \Pr(Y \geq y) \) denotes the survivor function of \( Y \), \( S_{Y|X}(y \mid x) = \Pr(Y \geq y \mid X \leq x) \) denotes the conditional survivor function of \( Y \), and the conditional distribution and survivor functions are assumed to exist whenever used (i.e., when needed, \( S_Y(y) > 0 \) and \( F_X(x) > 0 \)). Since the support of the joint distribution is the attainable set, boundaries of \( \mathcal{P} \) can be defined in terms of the conditional distributions defined above by (2.42) and (2.43). This allows definition of some new concepts of efficiency.

For the input-oriented case, assuming \( S_Y(y) > 0 \), define

\[ \bar{\theta}(x, y \mid H_{XY}) = \inf \{ \theta \mid F_{X|Y}(\theta x \mid y) > 0 \} \]

\[ = \inf \{ \theta \mid H_{XY}(\theta x, y) > 0 \}. \]  \hspace{1cm} (2.44)
Similarly, for the output-oriented case, assuming $F_X(x) > 0$, define
\[
\tilde{\lambda}(x, y \mid H_{XY}) = \sup \{\lambda \mid S_{Y\mid X}(\lambda y \mid x) > 0\}
\]
\[
= \sup \{\lambda \mid H_{XY}(x, \lambda y) > 0\}. \quad (2.45)
\]
The input efficiency score $\tilde{\theta}(x, y \mid H_{XY})$ may be interpreted as the proportionate reduction of inputs (holding output levels fixed) required for a unit operating at $(x, y) \in \mathcal{P}$ to achieve zero probability of being dominated by a randomly chosen unit. Analogously, the output efficiency score $\lambda(x, y \mid H_{XY})$ gives the proportionate increase in outputs required for the same unit to have zero probability of being dominated by a randomly chosen unit, holding input levels fixed. Note that in a multivariate framework, the radial nature of the Debreu–Farrell measures is preserved.

Similar definitions are possible in the hyperbolic and directional orientations. In particular, define
\[
\tilde{\gamma}(x, y \mid H_{XY}) = \inf \{\gamma \mid H_{XY}(\gamma^{-1}x, \gamma y) > 0\} \quad (2.46)
\]
and
\[
\tilde{\delta} = (x, y \mid u, v, H_{XY}) = \sup \{\delta \mid H_{XY}(x - \delta u, y + \delta v) > 0\}. \quad (2.47)
\]
The hyperbolic efficiency measure $\tilde{\gamma}(x, y \mid H_{XY})$ gives the simultaneous, proportionate reduction in input quantities and increase in output quantities required for a unit operating at $(x, y) \in \mathcal{P}$ to reduce its probability of being dominated by a randomly chosen unit to zero. Similarly, for a unit operating at $(x, y) \in \mathcal{P}$, the directional measure in (2.47) gives the amounts by which input quantities must be reduced and by which output quantities must be increased (simultaneously) to reduce the unit’s probability of being dominated by a randomly chosen unit to zero.

From the properties of the distribution function $H_{XY}(x, y)$, it is clear that the new efficiency scores defined in (2.44)–(2.47) have some interesting, reasonable properties. First, $\tilde{\theta}(x, y \mid H_{XY})$ and $\tilde{\gamma}(x, y \mid H_{XY})$ are monotone, nonincreasing in $x$ and monotone, nondecreasing in $y$. Second, $\tilde{\lambda}(x, y \mid H_{XY})$ and $\tilde{\delta}(x, y \mid H_{XY})$ are monotone, nondecreasing in $x$ and monotone, nonincreasing in $y$. Third, and most
importantly, if $\mathcal{P}$ is free disposal (an assumption maintained throughout this section), it is trivial to show that
\begin{align}
\tilde{\theta}(x, y | H_{XY}) &= \theta(x, y | \mathcal{P}), \\
\tilde{\lambda}(x, y | H_{XY}) &= \lambda(x, y | \mathcal{P}), \\
\tilde{\gamma}(x, y | H_{XY}) &= \gamma(x, y | \mathcal{P}),
\end{align}
and
\begin{equation}
\tilde{\delta}(x, y | u, v, H_{XY}) = \delta(x, y | u, v, \mathcal{P}).
\end{equation}
Therefore, under Assumption 1.3, the probabilistic formulation presented here leads to an alternative representation of the traditional Debreu–Farrell efficiency scores.

For a given output vector $y$, the efficient frontier of $\mathcal{P}$ can be characterized as noted above by $x^\partial(y)$ defined in (1.8). If $x$ is univariate, then $x^\partial(y)$ determines a frontier function $\phi(y): \mathbb{R}^q_+ \mapsto \mathbb{R}^1_+$ such that
\begin{equation}
\phi(y) = \inf \{x | F_X(x | y) > 0\} \forall y \in \mathbb{R}^q_+.
\end{equation}
The intersection of the horizontal line at output level $y$ and the curve representing $\phi(y)$ gives the minimum input level $\phi(y)$ than can produce output level $y$. Similarly, if $y$ is univariate, then working in the output direction one can define a production function $\psi(x): \mathbb{R}^p_+ \mapsto \mathbb{R}^1_+$ such that
\begin{equation}
\psi(x) = \sup \{y | S_X(y | x)\} \forall x \in \mathbb{R}^p_+.
\end{equation}
The functions $\phi(y)$ and $\psi(x)$ are illustrated in Figure 2.3 for the simple case where $p = q = 1$.

Nonparametric estimators of the efficiency scores $\tilde{\theta}(x, y | \mathcal{P})$ and $\tilde{\lambda}(x, y | \mathcal{P})$ (and hence also of $\theta(x, y | \mathcal{P})$ and $\lambda(x, y | \mathcal{P})$, provided Assumption 1.3 holds), can be obtained using the same plug-in approach used earlier. Here, this involves substituting the empirical distribution function $\hat{H}_{XY,n}(x, y)$ for the true distribution function $H_{XY}(x, y)$ in (2.44)–(2.47). The empirical analog of $H_{XY}(x, y)$ is
\begin{equation}
\hat{H}_{XY,n}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x, Y_i \geq y),
\end{equation}
The Nonparametric Envelopment Estimators

Fig. 2.3 Input-oriented frontier function \( \phi(y) \) for the bivariate case \((p = q = 1)\).

where \( \mathbb{1}(\cdot) \) is the indicator function (i.e., \( \mathbb{1}(A) = 1 \) if \( A \); otherwise, \( \mathbb{1}(A) = 0 \)). Note that at any point \((x, y) \in \mathbb{R}^{p+q} \), \( \hat{H}_{XY,n}(x, y) \) gives the proportion of sample observations in \( S_n \) with values \( X_i \leq x \) and \( Y_i \geq y \); in other words, \( \hat{H}_{XY,n}(x, y) \) gives the proportion of points in the sample \( S_n \) that (weakly) dominate \((x, y)\).

Now nonparametric estimators of the efficiency measures given in (2.44)–(2.47) can be defined. In the input-oriented case, \( \theta(x, y \mid \mathcal{P}) \) is estimated by

\[
\tilde{\theta}(x, y \mid \hat{H}_{XY,n}(x, y)) = \inf \{ \theta \mid \hat{F}_{X \mid Y}(\theta x \mid y) > 0 \} = \inf \{ \theta \mid \hat{H}_{XY}(\theta x, y) > 0 \}. \tag{2.55}
\]

where

\[
\hat{F}_{X \mid Y,n}(x \mid y) = \frac{\hat{H}_{XY,n}(x, y)}{\hat{H}_{XY,n}(\infty, y)} \tag{2.56}
\]

is the empirical conditional distribution function that is the empirical analog of \( F_{X \mid Y}(x \mid y) \) appearing in (2.42). Similarly, in the output orientation, plugging in \( \hat{H}_{XY,n}(x, y) \) to replace \( H_{XY}(x, y) \) in (2.45) yields the estimator

\[
\tilde{\lambda}(x, y \mid \hat{H}_{XY,n}) = \sup \{ \lambda \mid \hat{S}_{Y \mid X,n}(\lambda y \mid x) > 0 \} = \sup \{ \lambda \mid \hat{H}_{XY,n}(x, \lambda y) > 0 \}, \tag{2.57}
\]
2.4 An Alternative Probabilistic Formulation of the DGP

where \( \hat{S}_{Y|X,n}(y | x) \) is the empirical conditional survival function given by

\[
\hat{S}_{Y|X,n}(y | x) = \frac{\hat{H}_{XY,n}(x, y)}{\hat{H}_{XY,n}(x, 0)}.
\] (2.58)

In the hyperbolic and directional orientations, similar substitutions lead to estimators of \( \tilde{\gamma}(x, y | H_{XY}) \) and \( \tilde{\delta}(x, y | H_{XY}) \) given by

\[
\tilde{\gamma}(x, y | \hat{H}_{XY,n}) = \inf \{ \gamma | \hat{H}_{XY,n}(\gamma x, \gamma^{-1} y) > 0 \}
\] (2.59)

and

\[
\tilde{\delta} = (x, y | u, v, \hat{H}_{XY,n}) = \sup \{ \delta | \hat{H}_{XY,n}(x - \delta u, y + \delta v) > 0 \}. \] (2.60)

Under Assumption 1.3, it is easy to show that these estimators coincide with the FDH estimators of the Debreu–Farrell efficiency scores; i.e.,

\[
\tilde{\theta}(x, y | H_{XY,n}) = \theta(x, y | \hat{P}_{FDH}), \] (2.61)
\[
\tilde{\lambda}(x, y | H_{XY,n}) = \lambda(x, y | \hat{P}_{FDH}), \] (2.62)
\[
\tilde{\gamma}(x, y | H_{XY,n}) = \gamma(x, y | \hat{P}_{FDH}), \] (2.63)

and

\[
\tilde{\delta}(x, y | H_{XY,n}) = \delta(x, y | \hat{P}_{FDH}). \] (2.64)

Hence the FDH efficiency estimators can be seen as intuitive, empirical versions of the Debreu–Farrell efficiency scores, without explicit reference to the free disposal hull of the observed cloud of points. This interpretation of the FDH estimators suggests that the FDH efficiency estimators are arguably the most natural nonparametric estimators of the Debreu–Farrell efficiency measures using the plug-in principle.

The probabilistic, plug-in presentation in this section is very useful for providing extensions of the FDH estimators. Bădin and Simar (2009) used the probabilistic interpretation to develop a bias-corrected version of the input- and output-oriented FDH efficiency estimators. Below, in Section 4, the probabilistic interpretation is used to define some robust nonparametric estimators, while Section 7 shows how this framework can be used to investigate the effect of environmental variables on the production process.
2.5 Properties of FDH and DEA Estimators

As has already been noted, a well-defined, coherent statistical model is necessary in order to know what is estimated. But this is not all that is needed. In order to know what, if anything, any particular estimator may be able to reveal about the underlying quantity or model feature being estimated, some knowledge of the estimator’s underlying statistical properties is required. An understanding of the properties of an estimator is necessary in order to make statistical inference. Earlier, the FDH and DEA estimators were shown to be biased. Here, the questions of whether, and to what extent, these estimators can reveal useful information about efficiency, and under what conditions are considered. Simar and Wilson (2000b, 2008) surveyed the statistical properties of FDH and DEA estimators, but more recent results have been obtained.

2.5.1 Stochastic Convergence and Rates of Convergence

Perhaps the most fundamental property that an estimator should possess is that of consistency. Loosely speaking, if an estimator \( \hat{\zeta}_n \) of an unknown parameter \( \zeta \) is consistent, then the estimator converges (in some sense) to \( \zeta \) as the sample size \( n \) increases toward infinity (a subscript \( n \) is often attached to notation for estimators to remind the reader that one can think of an infinite sequence of estimators, each based on a different sample size). In other words, if an estimator is consistent, then more data should be helpful — quite a sensible property. If an estimator is inconsistent, then even an arbitrarily large amount of data would offer no particular hope or guarantee of getting “close” to the true value that one wishes to estimate. It is this sense in which consistency is the most fundamental property that an estimator might have; if an estimator is not consistent, there is little reason to consider what other properties the estimator might have, nor is there typically any reason to use such an estimator.

To be more precise, first consider the notion of convergence in probability, denoted \( \hat{\zeta}_n \xrightarrow{p} \zeta \). Convergence in probability occurs whenever

\[
\lim_{n \to \infty} \Pr(|\hat{\zeta}_n - \zeta| > \varepsilon) = 0 \quad \text{for any } \varepsilon > 0.
\]
An estimator that converges in probability (to the quantity of interest) is said to be weakly consistent; other types of consistency can also be defined (e.g., see Serfling, 1980). Convergence in probability means that, for any arbitrarily small (but strictly positive) $\varepsilon$, the probability of obtaining an estimate different from $\zeta$ by more than $\varepsilon$ in either direction tends to 0 as $n \to \infty$.

Note that consistency does not mean that it is impossible to obtain an estimate very different from $\zeta$ using a consistent estimator with a very large sample size. Rather, consistency is an asymptotic property; it only describes what happens in the limit. Although consistency is a fundamental property, it is also a minimal property in this sense. Depending on the rate, or speed, with which $\hat{\zeta}_n$ converges to $\zeta$, a particular sample size may or may not offer much hope of obtaining an accurate, useful estimate.

In nonparametric statistics, it is often difficult to prove convergence of an estimator and to obtain its rate of convergence. Often, convergence and its rate are expressed in terms of the stochastic order of the error of estimation. The weakest notion of the stochastic order is related to the notion of “bounded in probability.” A sequence of random variables $A_n$ is said to be bounded in probability if and only if there exists $B_\varepsilon$ and $n_\varepsilon$ such that for all $n > n_\varepsilon$ and $\varepsilon > 0$, $\Pr(A_n > B_\varepsilon) < \varepsilon$; such cases are denoted by writing $A_n = O_p(1)$. This means that when $n$ is large, the random variable $A_n$ is bounded, with probability tending to one. The notation $A_n = O_p(n^{-\alpha})$, where $\alpha > 0$, denotes that the sequence $A_n/n^{-\alpha} = n^\alpha A_n$ is $O_p(1)$. In this case, $A_n$ is said to converge to a small quantity of order $n^{-\alpha}$. With some abuse of language, one can say also that $A_n$ converges at the rate $n^{\alpha}$ because when $A_n$ is multiplied by $n^{\alpha}$ (which can be rather large if $n$ is large and if $\alpha$ is not too small), the sequence $n^\alpha A_n$ remains bounded in probability. Consequently, if $\alpha$ is small (near zero), the rate of convergence is considered to be very slow, because $n^{-\alpha}$ is not so small even when $n$ is large (i.e., $\lim_{\alpha \downarrow 0} n^{-\alpha} = 1$).

This type of convergence is weaker than convergence in probability. If $\hat{A}_n$ converges in probability at the rate $n^{\alpha}$ ($\alpha > 0$), then $A_n/n^{-\alpha} = n^\alpha A_n \overset{p}{\to} 0$, which can be denoted by writing $A_n = o_p(n^{-\alpha})$ or $n^\alpha A_n = o_p(1)$ (this is sometimes called big-$O$, little-$o$ notation).
Writing $A_n = O_p(n^{-\alpha})$ means that $n^\alpha A_n$ remains bounded when $n \to \infty$ but writing $A_n = o_p(n^{-\alpha})$ means that $n^\alpha A_n \xrightarrow{p} 0$ when $n \to \infty$.

In terms of the convergence of an estimator $\hat{\zeta}_n$ of $\zeta$, writing $\hat{\zeta}_n - \zeta = o_p(n^{-\alpha})$ means that $\hat{\zeta}_n$ converges in probability at rate $n^\alpha$. Writing $\hat{\zeta}_n - \zeta = O_p(n^{-\alpha})$ implies the weaker form of convergence, but the rate is still said to be $n^\alpha$. Standard, parametric estimation problems usually yield estimators that converge in probability at the rate $\sqrt{n}$ (corresponding to $\alpha = 1/2$) and are said to be root-$n$ consistent in such cases; this provides a familiar benchmark to which the rates of convergence of other, nonparametric estimators can be compared.

In all cases, the value of $\alpha$ plays a crucial role by indicating the “stochastic order” of the error of estimation. Since in many nonparametric problems $\alpha$ will be small (i.e., smaller than 1/2), it is important to consider whether one has enough data to obtain meaningful estimates. When the method of ordinary least squares (OLS) is used to estimate a parametric, linear regression model, researchers are sometimes willing to consider results from perhaps only 10 observations, though most researchers would retain substantial skepticism about the results. One might be more comfortable with, say, 30 observations; with 100 observations, many would not worry about having enough data. With 500 observations, few would doubt that very accurate estimates might be obtained. Ordinary least squares estimators achieve the classical parametric rate of convergence, i.e., $n^{1/2}$. Faced with an estimator that converges at rate $n^\alpha$, and $n$ observations, some estimation error must be incurred. One can find the number of observations $n_{\text{OLS}}$ that would achieve the same order of estimation error when OLS is used to estimate a linear regression model by writing $n_{\text{OLS}}^{1/2} = n^\alpha$ and solving for $n_{\text{OLS}}$ to obtain $n_{\text{OLS}} = n^{2\alpha}$.

Table 2.1 presents such calculations for $n = 25, 50, 75, 100, 1,000,$ and 10,000. For example, using 100 observations and an estimator that converges at rate $n^{2/3}$ would yield estimation error of the same order that one attains with 464 observations in an OLS regression. But as noted above, most nonparametric estimators have convergence rates slower than the typical root-$n$ parametric rate. Using 100 observations and an estimator that converges at rate $n^{1/3}$ yields estimation error of the same order as OLS with only 22 observations; with an estimator
Table 2.1. Effective number of observations when convergence rate is $n^\alpha$.

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>100</th>
<th>1000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>625</td>
<td>2500</td>
<td>5625</td>
<td>10000</td>
<td>10000000</td>
<td>100000000</td>
</tr>
<tr>
<td>2/3</td>
<td>73</td>
<td>184</td>
<td>316</td>
<td>464</td>
<td>215443</td>
<td></td>
</tr>
<tr>
<td>1/2</td>
<td>25</td>
<td>50</td>
<td>75</td>
<td>100</td>
<td>10000000</td>
<td>100000000</td>
</tr>
<tr>
<td>2/5</td>
<td>13</td>
<td>23</td>
<td>32</td>
<td>40</td>
<td>1585</td>
<td></td>
</tr>
<tr>
<td>1/3</td>
<td>9</td>
<td>14</td>
<td>18</td>
<td>22</td>
<td>100</td>
<td>464</td>
</tr>
<tr>
<td>2/7</td>
<td>6</td>
<td>9</td>
<td>12</td>
<td>14</td>
<td>52</td>
<td>193</td>
</tr>
<tr>
<td>1/4</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>10</td>
<td>32</td>
<td>100</td>
</tr>
<tr>
<td>2/9</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>22</td>
<td>60</td>
</tr>
<tr>
<td>1/5</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>16</td>
<td>40</td>
</tr>
<tr>
<td>2/11</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>12</td>
<td>28</td>
</tr>
<tr>
<td>1/6</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td>2/13</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>1/7</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>14</td>
</tr>
</tbody>
</table>

converging at rate $n^{2/9}$, using 100 observations would yield estimation error of the same order that one would achieve with OLS using only 8 observations; one would need well more than 10,000 observations to reduce the order of the estimation error to that attained by OLS with 100 observations.

As noted above, parametric estimators typically achieve a convergence rate of $n^{1/2}$, while many nonparametric estimators achieve only a (sometimes far) slower rate of convergence. The tradeoffs are clear: parametric estimators offer fast convergence, and hence it is possible to obtain meaningful estimates with smaller amounts of data than would be required by nonparametric estimators with slower convergence rates. But this is valid only if the parametric model that is estimated accurately reflects the true DGP; if not, there is specification error, calling into question consistency (and perhaps other properties) of the parametric estimators. Robinson (1988, p. 933) refers to estimators in models as "$n^{1/2}$-inconsistent"; a root-$n$ consistent estimator in a misspecified model will converge quickly to something that is perhaps crazy or meaningless. By contrast, the nonparametric estimators discussed in this survey largely avoid the risk of specification error, but (in some, but not all cases) at the cost of slower convergence rates and hence larger data requirements. What might constitute a
The Nonparametric Envelopment Estimators

“large sample” varies, depending on the stochastic order of the estimation error for the estimator that one chooses. Unfortunately, as will be shown below, many published applications of DEA estimators have used far fewer data than what might reasonably be required to obtain statistically meaningful results.

2.5.2 Consistency of FDH and DEA Estimators

This section summarizes the consistency results for the nonparametric envelopment estimators under the Assumptions 2.1 to 2.3 given above. For the DEA estimators, convexity of $\mathcal{P}$ (Assumption 2.4) is needed, but for the FDH case, the results are valid with or without this assumption.

Research on consistency and rates of convergence of efficiency estimators first examined the simpler cases where either inputs or outputs were unidimensional. For $p = 1$ and $q \geq 1$, Banker (1993) showed consistency of the input efficiency estimator $\theta(x, y | \hat{\mathcal{P}}_{VRS})$ for convex $\mathcal{P}$, but obtained no information on the rate of convergence.

The first systematic analysis of the convergence properties of the envelopment estimators ($\hat{\mathcal{P}}_{FDH}$ and $\hat{\mathcal{P}}_{VRS}$) appeared in Korostelev et al. (1995a, 1995b). For the case $p = 1$ and $q \geq 1$, Korostelev et al. (1995b) proved that when $\mathcal{P}$ satisfies free disposability, but not convexity,

$$ d_\Delta(\hat{\mathcal{P}}_{FDH}, \mathcal{P}) = O_p(n^{-\frac{1}{q+1}}), $$

(2.65)

where $d_\Delta(\cdot, \cdot)$ is the Lebesgue measure (giving the volume) of the difference between the two sets. Korostelev et al. (1995b) proved that when $\mathcal{P}$ satisfies both free disposability and convexity,

$$ d_\Delta(\hat{\mathcal{P}}_{VRS}, \mathcal{P}) = O_p(n^{-\frac{2}{q+2}}), $$

(2.66)

Although the rates of convergence for $\hat{\mathcal{P}}_{FDH}$ and $\hat{\mathcal{P}}_{VRS}$ are not very different, it remains true that for all $q \geq 1$, $n^{-1/(q+1)} > n^{-2/(q+2)}$; i.e., for all $q \geq 1$, $\hat{\mathcal{P}}_{FDH}$ converges at a slower rate than does $\hat{\mathcal{P}}_{VRS}$. Moreover, the difference is larger for small values of $q$ as revealed by Table 2.1. Clearly, incorporation of the convexity assumption into the estimator of $\mathcal{P}$ (as done by $\hat{\mathcal{P}}_{VRS}$) improves the rate of convergence. However, as
2.5 Properties of FDH and DEA Estimators

Noted earlier, if \( \mathcal{P} \) is not convex, the DEA estimator is not consistent and hence does not converge, whereas the FDH estimator converges regardless of whether \( \mathcal{P} \) is convex. In addition, the rates of convergence depend on the dimensionality of the problem, i.e., on the number of outputs \( q \) (when \( p = 1 \)). This is yet another manifestation of the “curse of dimensionality” shared by most nonparametric approaches in statistics and econometrics. Additional discussion on this issue is given below.

Despite the curse of dimensionality, Korostelev et al. (1995a, 1995b) show that the FDH and DEA estimators share some optimality properties. In particular, under free disposability (but not convexity), \( \hat{\mathcal{P}}_{FDH} \) is the most efficient estimator of \( \mathcal{P} \) (in term of the minimax risk over the set of estimators sharing the free disposability assumption, where the loss function is \( d_{\Delta} \)). Where \( \mathcal{P} \) is both free-disposal and convex, \( \hat{\mathcal{P}}_{VRS} \) becomes the most efficient estimator over the class of all the estimators sharing the free disposal and the convexity assumption. These are quite important results and suggest the inherent quality of the envelopment estimators, even if they are imprecise in small samples where \( q \) large.

Because of the difficulty of proving results about \( \hat{\mathcal{P}}_{FDH} \) and \( \hat{\mathcal{P}}_{VRS} \), Korostelev et al. (1995a, 1995b) were forced to limit their analysis to cases where \( p = 1 \). Focusing instead on the DEA efficiency estimator under VRS, Kneip et al. (1998) proved that

\[
\theta(x, y \mid \hat{\mathcal{P}}_{VRS}) - \theta(x, y \mid \mathcal{P}) = O_p(n^{-\frac{2}{p+q+1}}). \tag{2.67}
\]

Straightforward, but tedious changes in notation extend this result to the output oriented estimator \( \lambda(x, y \mid \hat{\mathcal{P}}_{VRS}) \). Wilson (2011) established consistency and the same rate of convergence for the hyperbolic estimator \( \gamma(x, y \mid \hat{\mathcal{P}}_{VRS}) \), while Simar et al. (2012) established consistency and the same rate of convergence for the directional estimator \( \delta(x, y \mid \hat{\mathcal{P}}_{VRS}) \).

Park et al. (2010) proved, under the additional assumption that \( \mathcal{P}^\theta \) is globally CRS, that

\[
\theta(x, y \mid \hat{\mathcal{P}}_{CRS}) - \theta(x, y \mid \mathcal{P}) = O_p(n^{-\frac{2}{p+q+1}}). \tag{2.68}
\]

This result also extends to the output-oriented estimator \( \lambda(x, y \mid \hat{\mathcal{P}}_{CRS}) \) after straightforward but tedious changes in notation. Recalling that
under CRS, \( \gamma(x, y \mid \hat{P}_{\text{CRS}}) = \theta(x, y \mid \hat{P}_{\text{CRS}})^2 \), it is obvious that the estimator of the hyperbolic measure \( \gamma(x, y \mid P) \) also achieves the same \( n^{2/(p+q)} \) rate of convergence. Consequently, the discussion surrounding the transformation in (2.15)–(2.16) and the result in (2.18) makes clear that the estimator \( \delta(x, y \mid u, v, \hat{P}_{\text{CRS}}) \) also achieves the \( n^{2/(p+q)} \) rate of convergence.

Turning to the FDH efficiency estimators, Park et al. (2000) established that

\[
\theta(x, y \mid \hat{P}_{\text{FDH}}) - \theta(x, y \mid P) = O_p(n^{-\frac{1}{p+q}}). \tag{2.69}
\]

As with the DEA estimators, this result extends to the output-oriented case after straightforward changes in notation. Wheelock and Wilson (2008) extended the result to the case of the hyperbolic FDH efficiency estimator, and Simar and Vanhems (2012) extended the result to the case of directional FDH efficiency estimators. In all cases, the convergence rate is \( n^{1/(p+q)} \).

The convergence rates obtained by Korostelev et al. (1995a, 1995b) and by Gijbels et al. (1999) when \( p = 1 \) are special cases of the results described above. The curse of dimensionality acts symmetrically in both input and output spaces; i.e., the rate of convergence declines by the same degree regardless of whether either \( p \) or \( q \) is increased.

Note that the convergence rate for \( \theta(x, y \mid \hat{P}_{VRS}) \) is slightly faster than that for \( \theta(x, y \mid \hat{P}_{FDH}) \), provided Assumption 2.4 is satisfied; otherwise, \( \theta(x, y \mid \hat{P}_{VRS}) \) is inconsistent. The faster convergence rate for the DEA estimator is due to the fact that \( \hat{P}_{FDH} \subseteq \hat{P}_{VRS} \subseteq P \).

### 2.5.3 Curse of Dimensionality: Parametric versus Nonparametric Inference

Table 2.1 makes clear that as \( p + q \) increases, the effective sample size (in terms of the sample size that would yield estimation error of the same order when OLS is used to estimate a linear regression) decreases. Alternatively, a much larger sample size is needed to reach the precision obtained in the simplest case where \( p = q = 1 \), in which case the parametric rate \( n^{1/2} \) is attained by the FDH estimators, or where an even better rate \( n^{2/3} \) is obtained by the DEA–VRS estimator. One can...
find published examples where authors have used 100 or fewer observations with six or more dimensions. For example, Table 2.1 indicates that with 50 observations and \( p + q = 8 \), the estimation error incurred by the DEA–VRS estimator is equivalent to what one would obtain using OLS with only 6 observations. Few would take seriously results from a regression with only 6 observations.

When \( p + q \) is large, unless a very large quantity of data are available, the resulting imprecision will manifest itself in the form of large bias, large variance, and consequently, very wide confidence intervals. This has been confirmed in Monte-Carlo experiments. In fact, as the number of dimensions is increased, the number of observations must increase at an exponential rate to maintain a given mean-square error with the nonparametric estimators of \( \mathcal{P} \). General statements on the number of observations required to achieve a given level of mean-square error are not possible, since the exact convergence of the nonparametric estimators depends on unknown constants related to the features of the unobserved \( \mathcal{P} \). Nonetheless, for estimation purposes, it is always true that more data are better than fewer data (recall that this is a consequence of consistency). In the case of nonparametric estimators such as \( \hat{\mathcal{P}}_{\text{FDH}} \) and \( \hat{\mathcal{P}}_{\text{VRS}} \), this statement is more than doubly true — it is exponentially true. In the context of nonparametric regression or density estimation, dimension-reduction techniques are sometimes used (e.g., see Wheelock and Wilson, 2011); Daraio and Simar (2007a) discuss similar ideas that can be used with DEA or FDH estimators.

The curse of dimensionality results from the fact that as a given set of \( n \) observations are projected in an increasing number of orthogonal directions, the Euclidean distance between the observations necessarily must increase. Moreover, for a given sample size, increasing the number of dimensions will result in more observations lying on the boundaries of the estimators \( \hat{\mathcal{P}}_{\text{FDH}} \) and \( \hat{\mathcal{P}}_{\text{VRS}} \). The FDH estimator is particularly affected by this problem. The DEA estimator is also affected, often times to a lesser degree due to its incorporation of the convexity assumption (recall that \( \hat{\mathcal{P}}_{\text{VRS}} \) is merely the convex hull of \( \hat{\mathcal{P}}_{\text{FDH}} \); consequently, fewer points will lie on the boundary of \( \hat{\mathcal{P}}_{\text{VRS}} \) than on the boundary of \( \hat{\mathcal{P}}_{\text{FDH}} \). Wheelock and Wilson (2003) and Wilson (2004) have found cases where all or nearly all observations in samples of
The Nonparametric Envelopment Estimators

several thousand observations lie on the boundary of $\hat{P}_{FDH}$, while relatively few observations lie on the boundary of $\hat{P}_{VRS}$. Both papers argue that the FDH estimator should be used as a diagnostic to check whether it might be reasonable to employ the DEA estimator in a given application; large numbers of observations falling on the boundary of $\hat{P}_{FDH}$ may indicate problems due to the curse of dimensionality.

Parametric estimators suffer little from this phenomenon in the sense that their rate of convergence typically does not depend on dimensionality of the problem. The parametric structure incorporates information from all of the observations in a sample, regardless of the dimensionality of $P$. This also explains why parametric estimators are usually (but not always) more efficient in a statistical sense than their nonparametric counterparts — the parametric estimators extract more information from the data, assuming of course, that the parametric assumptions that have been made are correct. This is frequently a big assumption.

Additional insight is provided by comparing parametric maximum likelihood estimators and nonparametric FDH and DEA estimators. Parametric maximum likelihood estimation is the most frequently used parametric estimation method. Under some regularity conditions that do not involve dimensionality of the problem at hand (see, for example, Spanos, 1999), maximum likelihood estimators are root-$n$ consistent. Maximum likelihood estimation involves maximizing a likelihood function in which each observation is weighted equally. Hence, maximum likelihood estimators are global estimators, as opposed to FDH and DEA estimators, which are local estimators. With FDH or DEA estimators of the frontier, only observations near the point where the frontier is being estimated contribute to the estimate at that point; far-away observations contribute little or nothing to estimation at the point of interest.

It remains true, however, that when data are projected in an increasing number of orthogonal directions, Euclidean distance between the observations necessarily increases. This is problematic for FDH and DEA estimators, because it means that increasing dimensionality results in fewer near-by observations that can impart information about the frontier at a particular point of interest. But increasing distance
between observations means also that parametric, maximum likelihood estimators are combining information, and weighting it equally, from observations that are increasingly far apart (with increasing dimensionality). Hence, increasing dimensionality means that the researcher must rely increasingly on the parametric assumptions of the model. Again, these are often big assumptions, and should be tested, though often they are not.

The points made here go to the heart of the tradeoff between nonparametric and parametric estimators. Parametric estimators incur the risk of misspecification, which typically results in inconsistency, but are almost always statistically more efficient than nonparametric estimators if properly specified. Nonparametric estimators avoid the risk of misspecification, but usually involve more noise than parametric estimators. Lunch is not free, and the world is full of tradeoffs, which creates employment opportunities for economists and statisticians.

2.5.4 Asymptotic Sampling Distributions

As discussed above, consistency is an essential property for any estimator. However, consistency is also a minimal theoretical property. The preceding discussion indicates that DEA or FDH efficiency estimators converge as the sample size increases, although perhaps at a slow rate. However, by themselves, these results have little practical use other than to confirm that the DEA or FDH estimators are possibly reasonable to use for efficiency estimation.

For empirical applications, more is needed — in particular, the applied researcher must have some knowledge of the sampling distributions in order to make inferences about the true levels of efficiency or inefficiency (correction for the bias and construction of confidence intervals, for instance). This is particularly important in situations where point estimates of efficiency might be highly variable due to the curse of dimensionality or other problems. In nonlinear settings, and especially in nonparametric settings, typically only asymptotic distributional results are available, and this is the case in the nonparametric framework of FDH and DEA estimators.
In the case of the input-oriented FDH efficiency estimator, Park et al. (2000) established
\[
n^{p+q} \left( \theta(\mathbf{x}, \mathbf{y} | \hat{P}_{FDH}) - \theta(\mathbf{x}, \mathbf{y} | P) \right) \overset{d}{\rightarrow} \text{Weibull}(\mu_{x,y}, p + q), \tag{2.70}
\]
where \( \mu_{x,y} \) is a constant depending on the DGP. This constant is proportional to the probability of observing a firm dominating a point on the ray \( \mathbf{x} \), in a neighborhood of the frontier point \( \mathbf{x}(\mathbf{y}) \). The constant \( \mu_{x,y} \) is larger (smaller) as the density \( f(\mathbf{x}, \mathbf{y}) \) provides more (less) mass in the neighborhood of the frontier point. The bias and standard deviation of the FDH estimator are of the order \( n^{-1/(p+q)} \), and are proportional to \( \mu_{x,y}^{-1/(p+q)} \). Also, the ratio of the mean to the standard deviation of \( \theta(\mathbf{x}, \mathbf{y} | \hat{P}_{FDH}) - \theta(\mathbf{x}, \mathbf{y} | P) \) does not depend on \( \mu_{x,y} \) nor on \( n \), and increases as the dimensionality \( p + q \) increases. Thus, the curse of dimensionality here is two fold: both the rate of convergence and the bias worsen with increasing dimensionality. These results suggest a strong need for a bias-corrected version of the FDH estimator.

Park et al. (2000) propose a consistent estimator of \( \mu_{x,y} \) in order to obtain a bias-corrected estimator and asymptotic confidence intervals. However, Monte-Carlo studies indicate that the noise introduced by estimating \( \mu_{x,y} \) reduces the quality of inference when \( p + q \) is large with moderate sample sizes (e.g., \( n \leq 1000 \) with \( p + q \geq 5 \)). Here again the bootstrap might provide a useful alternative, but some smoothing of the FDH estimator might even be more appropriate than it is in the case of the bootstrap with DEA estimators (Jeong and Simar, 2006).

All of the results obtained by Park et al. (2000) extend to the output-oriented estimator \( \lambda(\mathbf{x}, \mathbf{y} | \hat{P}_{FDH}) \) after straightforward changes in notation. Simar and Vanhems (2012) extend these results to the FDH estimator \( \delta(\mathbf{x}, \mathbf{y} | \mathbf{u}, \mathbf{v}, \hat{P}_{FDH}) \) of the directional distance function, and Wilson (2011) establishes similar results for the FDH estimator \( \gamma(\mathbf{x}, \mathbf{y} | \mathbf{u}, \mathbf{v}, \hat{P}_{FDH}) \) of the hyperbolic distance function.

For the special case of one input and one output \((p = q = 1)\), Gijbels et al. (1999) obtain for the input-oriented DEA-VRS estimator
\[
n^{2/3} \left( \theta(\mathbf{x}, \mathbf{y} | \hat{P}_{VRS}) - \theta(\mathbf{x}, \mathbf{y} | P) \right) \overset{d}{\rightarrow} Q_1(\cdot), \tag{2.71}
\]
as well as an analytical expression for the limiting distribution \( Q_1(\cdot) \). The limiting distribution is a regular distribution function known up
to some constants that depend on features of the DGP involving (i) the curvature of the frontier, and (ii) the magnitude of the density $f(x, y)$ at the true frontier point $(\theta(x, y | \mathcal{P})x, y)$. Expressions for the asymptotic bias and variance are also provided. As expected for the DEA estimator, the bias is of order $n^{-2/3}$, and is larger for greater curvature of the true frontier. The DEA-VRS frontier estimator is piecewise linear, and so increasing curvature of the true frontier makes estimation more difficult. The bias decreases (increases) as the density at the true frontier point $(\theta(x, y | \mathcal{P})x, y)$ increases (decreases). Note that large density at $(\theta(x, y | \mathcal{P})x, y)$ implies greater chances of observing observations near the point of interest that can impart information useful for estimating $\theta(x, y)$; DEA estimators are local estimators. The variance of the DEA estimator behaves similarly. The results of Gijbels et al. also suggest that in most cases, the bias will be much larger than the variance.

Using simple estimators of the two constants in $Q_1(\cdot)$, Gijbels et al. provide a bias-corrected estimator and a procedure for building confidence intervals for $\theta(x, y | \mathcal{P})$. Monte-Carlo experiments indicate that even for moderate sample sizes (e.g., $n = 100$), the procedure works reasonably well.

Although the results of Gijbels et al. are limited to the case where $p = q = 1$, they provide insight for inference by identifying which features of the DGP affect the quality of inference-making. One should expect that in a more general multivariate setting, the same features should play similar roles, but perhaps in a more complicated fashion.

Derivation of analytical results for DEA estimators in the general, multivariate setting with $p, q \geq 1$ is more complicated, and the results are less edifying, though still necessary in order to have any hope for inference. Kneip et al. (2008) obtain an important key result, namely

$$
\frac{1}{n^{p+q+1}} \left( \frac{\theta(x, y | \hat{\mathcal{P}}_{\text{VRS}})}{\theta(x, y | \mathcal{P})} - 1 \right) \xrightarrow{d} Q_2(\cdot),
$$

where $Q_2(\cdot)$ is a regular distribution function known up to some constants. However, in this case, no closed analytical form for $Q_2(\cdot)$ is available, and hence the result is of little direct use for inference. In particular, the moments and the quantiles of $Q_2(\cdot)$ are not available,
so that neither bias-correction nor confidence intervals can be provided easily using only the result in (2.72).

The results described above extend to the output-oriented estimator \( \lambda(x, y \mid \hat{P}_{VRS}) \) after straightforward (but perhaps tedious) changes in notation. Wilson (2011) extends these results for the DEA estimator \( \gamma(x, y \mid \hat{P}_{VRS}) \) of the hyperbolic efficiency measure, and Simar et al. (2012) extends the results to the DEA estimator \( \delta(x, y \mid u, v, \hat{P}_{VRS}) \) of the directional efficiency measure. In each case, the convergence rate is \( n^{2/(p+q+1)} \).

In the case of CRS, Park et al. (2010) prove that

\[
\frac{1}{n^{p+q}} \left( \frac{\theta(x, y \mid \hat{P}_{CRS})}{\theta(x, y \mid P)} - 1 \right) \overset{d}{\rightarrow} Q_3(\cdot), \tag{2.73}
\]

where \( Q_3(\cdot) \) is a regular distribution function known up to some constants. As before, this result extends trivially to the output-oriented estimator \( \lambda(x, y \mid \hat{P}_{CRS}) \). Due to (2.38), the result also extends to the hyperbolic estimator \( \gamma(x, y \mid \hat{P}_{CRS}) \). Finally, it is straightforward to use the analysis of Simar et al. (2012) to obtain similar results for the directional efficiency estimator \( \delta(x, y \mid u, v, \hat{P}_{CRS}) \). In each case, the rate of convergence is \( n^{2/(p+q)} \).

Although the limiting distributions of DEA efficiency estimators are not directly applicable for inference, the above results establish existence of stable, limiting distributions for DEA efficiency estimators; this is required for inference to be possible. In particular, existence of a limiting distribution is necessary to establish consistency of bootstrap methods for inference as discussed below in Bootstrap Inference using DEA and FDH Estimators. To date, bootstrap methods provide the only practical approach to inference using FDH and DEA estimators.
Bootstrap Inference Using DEA and FDH Estimators

3.1 General Idea

3.1.1 Naive Bootstrap

Bootstrap methods for inference were introduced by Efron (1979), and have been discussed by Bickel and Freedman (1981), Efron (1982), Hall (1992), Efron and Tibshirani (1993), and others. The idea is intuitively appealing, and is based on replication of the DGP. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which the (perhaps vector-valued) random variable $Z$ is defined, and consider a random sample $Z_n = \{Z_1, \ldots, Z_n\}$ of draws of the random variable $Z$. Let $\mathcal{P} \in \mathbb{P}$ denote the DGP. Suppose that some quantity $\zeta \in \mathbb{R}$, is estimated by $\hat{\zeta}_n = \psi(Z_n)$, where $\psi(\cdot)$ is a real-valued function. In addition, suppose that $n^{-\tau}(\hat{\zeta}_n - \zeta) \sim G_n$ for some distribution function $G_n$.

If $G_n$ is known, then it is easy to see that a $(1 - \alpha) \times 100$-percent confidence interval for $\zeta$ is given by

$$\hat{\zeta}_n - n^{-\tau}G_n^{-1}\left(1 - \frac{\alpha}{2}\right) \leq \zeta \leq \hat{\zeta}_n - n^{-\tau}G_n^{-1}\left(\frac{\alpha}{2}\right), \quad (3.1)$$

where $G_n^{-1}(\cdot)$ denotes the quantile function corresponding to $G_n$. By construction, the statement in (3.1) is true with probability $(1 - \alpha)$.  

Of course, $G_n$ is seldom known exactly in applications. If $G_n$ is known up to some unknown constants (i.e., if the functional form of $G_n$ is known), then $G_n$ can be estimated by replacing the unknown constants with consistent estimates to obtain an estimator $\hat{G}_n$ of $G_n$ and corresponding quantile function $\hat{G}_n^{-1}(\cdot)$. Then the confidence interval in (3.1) is estimated by replacing $G_n^{-1}(\cdot)$ with $\hat{G}_n^{-1}(\cdot)$ to obtain an estimated confidence interval.

In many cases, the finite-sample distribution function $G_n$ may be completely unknown. However, if

$$n^\tau \left( \hat{\zeta}_n - \zeta \right) \mid \mathcal{P} \xrightarrow{d} G, \tag{3.2}$$

i.e., if the sequence of functions $\{G_n\}_{n=1}^\infty$ converges to a function $G$, and $G$ is known, then $G_n$ can be estimated by $G$. If $G$ exists but is estimated by $\hat{G}$, then $\hat{G}$ can also be used to estimate $G_n$. In the latter case, there are two sources of error: the asymptotic approximation of $G_n$ by $G$ introduces approximation error to the extent that $G$ and $G_n$ differ, and estimation of $G$ by $\hat{G}$ introduces estimation error.

As with $G_n$, if $G$ is known up to some unknown constants, then it may be possible to estimate $G$ by first estimating the constants and then substituting estimates for the unknown constants in $G$. In some cases, however, there may not be a closed-form expression for $G$, or the unknown parameters may be difficult to estimate. This is the case with the limiting distributions of the FDH and DEA estimators discussed earlier. In such cases, bootstrap methods may be useful.

Bootstrap methods typically involve drawing a pseudo-sample $Z_{m}^* = \{Z_{m}^1, \ldots, Z_{m}^m\}$ of size $m$ from a suitable estimate $\hat{\mathcal{P}}$ of the DGP $\mathcal{P}$. In many cases, $m$ will be set equal to $n$, but in making inference about efficiency using FDH and DEA estimators, it will be useful to set $m < n$ for reasons that will become clear below. Given the pseudo-sample $Z_{m}^*$, it should be straightforward to apply the same function $\psi$ to the data in $Z_{m}^*$ that was used to construct the original estimator of $\zeta$; doing so yields a pseudo-estimate $\hat{\zeta}_m^* = \psi(Z_{m}^*)$. Moreover, it should also be straightforward (but perhaps computationally demanding) to repeat this exercise $B$ times to obtain a set of $B$ pseudo-estimates $\{\hat{\zeta}_{m1}^*, \ldots, \hat{\zeta}_{mB}^*\}$. If $m$ and $\hat{\mathcal{P}}$ are chosen appropriately, then in many cases the unknown distribution $G$ can be approximated well by the
empirical distribution of \( m^\tau (\hat{\zeta}_m^* - \hat{\zeta}_n) \); in other words, the goal is to select \( m \) and \( \hat{P} \) so that as \( m \to \infty \),

\[
m^\tau (\hat{\zeta}_m^* - \hat{\zeta}_n) \mid Z_n, \hat{P} \xrightarrow{d} G. \tag{3.3}
\]

To illustrate, consider the random variable \( Z \) with probability density function \( f(z) = \rho \mathcal{N}(\mu_1, 1) + (1 - \rho) \mathcal{N}(\mu_2, 1) \), \( \rho \in [0, 1] \). This is a two-component mixture model; with probability \( \rho \), \( Z \) is a \( \mathcal{N}(\mu_1, 1) \) random variable, and with probability \((1 - \rho)\), \( Z \) is a \( \mathcal{N}(\mu_2, 1) \) random variable. Suppose \( n = 20 \), \( \rho = 0.95 \), \( \mu_1 = -1 \), and \( \mu_2 = -\frac{\rho}{1-\rho} \mu_1 \); then \( E(Z) = \mu_Z = 0 \). Let \( \hat{\mu}_n \) denote the sample mean of \( n \) draws from \( f(z) \). By the Lindeberg–Levy central limit theorem, \( n^{1/2}(\hat{\mu}_n - \mu_Z) \xrightarrow{d} N(0, \sigma^2_Z) \), where \( \text{VAR}(Z) = \sigma^2_Z \). However, because the distribution has a good deal of skewness (by construction), one should not expect the sampling distribution of \( \hat{\mu}_n \) to closely resemble a normal distribution unless the sample size \( n \) is quite large. One could estimate \( \sigma^2_Z \) by maximum likelihood to obtain \( \hat{\sigma}^2_Z = n^{-1} \sum_{i=1}^n (Z_i - \hat{\mu}_n)^2 \) and then use \( N(0, \hat{\sigma}^2_Z) \) to construct an estimated confidence interval. This is easy to do using R:

```r
> set.seed(900001)
> rho=0.95
> mu1=-1
> mu2=-(rho/(1-rho))*mu1
> n=20
> #
> rnmix <- function(n,rho,mu1,mu2) {
+   t1=runif(n)
+   i1=which(t1<=rho)
+   i2=which(t1>rho)
+   ran=vector(length=n)
+   ran[i1]=rnorm(n=length(i1),mean=mu1)
+   ran[i2]=rnorm(n=length(i2),mean=mu2)
+   return(ran)
+ }
> #
> z=rnmix(n,rho,mu1,mu2)
> zbar=mean(z)
> print(zbar)
[1] 0.6945251
> zsig=sd(z)
```
> print(zsig)
[1] 6.181361
> qq=qnorm(c(0.025,0.975),mean=0,sd=zsig)
> print(qq)
[1] -12.11524 12.11524
> ci.norm=c(zbar-qq[2]/sqrt(n),zbar-qq[1]/sqrt(n))
> print(ci.norm)
[1] -2.014526 3.403576
>
For the sample drawn by the *R* code, the sample mean is found to be about 0.6945, although the true mean is zero (note that in this example, the sample mean is unbiased due to the Lindeberg–Levy central limit theorem, but in the sample drawn here, the realized sample mean lies to the right of zero). The estimated 95-percent confidence interval, \([-2.0145, 3.4036]\), includes zero, but the interval is rather wide. By construction, the quantiles in *qq* are symmetric around zero, and hence the estimated interval is symmetric around the sample mean by construction.

As noted earlier, the distribution of *Z* is asymmetric by construction. Yet the conventional approach to inference implemented in the *R* code above enforces symmetry since quantiles are taken from the limiting normal distribution. Alternatively, one might use the bootstrap ideas discussed above to make inference about \(\mu_z\), as in the following exercise using *R*:

> nrep=1000
> zbar.star=vector(length=1000)
> for (b in 1:nrep) {
+   ind=floor(n*runif(n))+1
+   zbar.star[b]=mean(z[ind])
+ }
> g.star=sqrt(n)*(zbar.star-zbar)
> summary(g.star)

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>-11.4700</td>
<td>-4.5000</td>
<td>-0.1893</td>
<td>0.2717</td>
<td>4.2920</td>
<td>24.6900</td>
</tr>
</tbody>
</table>
> qq=quantile(g.star,probs=c(0.025,0.975))
> print(qq)

<table>
<thead>
<tr>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>-9.395732</td>
<td>13.520819</td>
</tr>
</tbody>
</table>
> ci=c(zbar-qq[2]/sqrt(n),zbar-qq[1]/sqrt(n))
Here, bootstrap pseudo-samples are formed in a loop by drawing independently and with replacement from the empirical distribution of the original sample; with each pass through the loop, the sample mean of the observations in the pseudo-sample is computed and stored in the vector \texttt{zbar.star}. After 1,000 passes through the loop, each bootstrap estimate \( \hat{\mu}_n^* \) is used to compute \( n^{1/2} (\hat{\mu}_n^* - \mu) \), and these are stored in the vector \texttt{g.star}. The \texttt{summary} command reveals that the distribution of the 1,000 values in \texttt{g.star} is skewed to the right, since the mean exceeds the median; consequently, the quantiles in \texttt{qq} are now asymmetric around zero. In addition, the difference between the quantiles in \texttt{qq} is now less than in the previous case. The empirical distribution of the values stored in \texttt{g.star} are used to approximate the distribution of \( n^{1/2} (\hat{\mu}_n - \mu) \), and the resulting 95-percent confidence interval estimate is \([-2.3288, 2.7955]\), which is considerably tighter (i.e., more narrow) than in the previous case. Moreover, the confidence interval estimate is no longer symmetric around the sample mean, but instead is almost symmetric around zero (i.e., the true value that is the object of estimation).

In the example above, bootstrap pseudo-samples are constructed by drawing identically and with replacement from the empirical distribution of the original sample. While this “works” in many problems, if “fails” in some problems, as the next example illustrates. Suppose now that \( Z \) is distributed uniform on the \([0, \zeta]\) interval. Given a sample \( Z_n = \{Z_1, \ldots, Z_n\} \) of size \( n \), the maximum likelihood estimate of \( \zeta \) is \( \hat{\zeta}_n = Z_{(n)} = \max\{Z_1, \ldots, Z_n\} \). It is well known that \( n\zeta^{-1}(\zeta - \hat{\zeta}_n) \sim \text{Exp}(1) \), where \( \text{Exp}(1) \) denotes an exponential distribution with parameter 1; hence in terms of the notation used above, \( G_n \) is known exactly (i.e., the distribution does not depend upon unknown parameters). An asymptotic \((1 - \alpha) \times 100\)-percent confidence interval for \( \zeta \) is given by \( \left[ \hat{\zeta}_n \left( 1 - n^{-1}G_{\exp}^{-1}(\frac{\alpha}{2}) \right)^{-1} , \hat{\zeta}_n \left( 1 - n^{-1}G_{\exp}^{-1}(1 - \frac{\alpha}{2}) \right)^{-1} \right] \), where \( G_{\exp}^{-1}(\cdot) \) denotes the quantile function corresponding to the \( \text{Exp}(1) \)
distribution. For sample size $n = 20$, such a confidence interval is estimated by the following $R$ code:

```r
> set.seed(900001)
> n=20
> #
> #
> z=runif(n)
> zeta.hat=max(z)
> print(zeta.hat)
[1] 0.992571
> qq=qexp(c(0.025,0.975),rate=1)
> print(qq)
[1] 0.02531781 3.68887945
> ci.exp=c(zeta.hat/(1-qq[1]/n),zeta.hat/(1-qq[2]/n))
> print(ci.exp)
[1] 0.9938291 1.2170482
>
Note that for this example, $\hat{\zeta}_n$ (stored in zeta.hat) lies to the left of the estimated confidence interval. This is as should be expected, since $\hat{\zeta}_n$ is biased downward. The bootstrap method can also be used to estimate a 95-percent confidence interval for $\zeta$:

```r
> nrep=1000
> zeta.star=vector(length=1000)
> for (b in 1:nrep) {
+   ind=floor(n*runif(n))+1
+   zeta.star[b]=max(z[ind])
+ }
> g.star=(n/zeta.hat)*(zeta.hat-zeta.star)
> summary(g.star)
   Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
 0.0000 0.0000 0.0000 0.2572 0.1505 3.6700
> qq=quantile(g.star,probs=c(0.025,0.975))
> print(qq)
   2.5%   97.5%
0.000000 1.870576
> ci=c(zeta.hat/(1-qq[1]/n),zeta.hat/(1-qq[2]/n))
> print(ci)
   2.5%   97.5%
0.992571 1.094983
>```
3.1 General Idea

At first glance, the bootstrap estimated confidence interval — [0.9926, 1.0949] — appears to be more informative than the previous estimated confidence interval; the bootstrap estimate covers one, and is considerably more narrow than the previous interval estimate. However, its coverage is likely to be poor in the sense that if a large number of samples of size \( n \) are drawn, and for each sample a confidence interval is estimated by the bootstrap method used above, the proportion of these intervals that include one will be different from 95-percent. This is confirmed by continuing the experiment as follows:

```r
> coverage=0
> for (j in 1:1000) {
+   z=runif(n)
+   zeta.hat=max(z)
+   for (b in 1:nrep) {
+     ind=floor(n*runif(n))+1
+     zeta.star[b]=max(z[ind])
+   }
+   g.star=(n/zeta.hat)*(zeta.hat-zeta.star)
+   qq=quantile(g.star,probs=c(0.025,0.975))
+   ci=c(zeta.hat/(1-qq[1]/n),zeta.hat/(1-qq[2]/n))
+   if (ci[1]<=1 & ci[2]>=1) coverage=coverage + 1
+ }
> coverage=coverage/1000
> print(coverage)
[1] 0.881
```

In fact, the bootstrap estimated confidence intervals cover the true value one in only about 88.1 percent of all trials. The reason for this “failure” of the bootstrap method is easy to see. Consider a single bootstrap pseudo-sample of size \( n \); on each draw from \( Z_n \), the probability of drawing \( Z_n \) is \( 1/n \). Hence the probability of not drawing \( Z_n \) on a single draw is \( 1 - n^{-1} \), and the probability of not drawing \( Z_n \) on any of \( n \) draws is \( (1 - n^{-1})^n \). Therefore, the probability that a particular bootstrap sample includes \( Z_n \) is \( 1 - (1 - n^{-1})^n \). Moreover,

\[
\lim_{n \to \infty} 1 - (1 - n^{-1})^n = 1 - e^{-1} \\
\approx 0.6321. \tag{3.4}
\]
So, over $B$ bootstrap replications, almost two-thirds of the bootstrap samples will contain at least one instance of $Z(n)$; for each of these samples, the bootstrap estimate $\hat{\zeta}_n^*$ will be identical to the original estimate $\hat{\zeta}_n$.

Among the bootstrap samples that do not contain $Z(n)$, many will contain the second-largest value in $Z$. The empirical distribution of $n\hat{\zeta}_n^{-1}(\hat{\zeta}_n - \zeta^*_n)$ — with which one hopes to approximate the distribution of $n\zeta^{-1}(\zeta - \hat{\zeta}_n)$ — can be expected to have support at only a few discrete points. Consequently, it does not approximate the Exp(1) distribution, even as $n \to \infty$. In the experiment above, this is confirmed by the following exercise using R:

```R
> length(which(zeta.star==zeta.hat))
[1] 660
> length(which(g.star==0))
[1] 660
> tmp=sort(unique(g.star))
> print(tmp)
[1] 0.0000000 0.1504549 1.7748937 1.8705764 3.6695468
```

This indicates that in 660 of 1,000 bootstrap replications, the bootstrap sample contains the original sample maximum. As a result, there are only 5 unique values in g.star. The distribution of values in g.star should resemble the Exp(1) distribution, but it does not. Since the Exp(1) distribution has support that is continuous, the probability of drawing a value identically equal to zero is 0; yet, in the bootstrap procedure used here, 66 percent of the values in g.star are identically zero. Moreover, the problem does not go away as $n$ becomes larger. Setting $n = 10,000$ and re-running the code above, the bootstrap confidence intervals cover 1 in 865 trials, less than when $n = 20$. In this case, the naive bootstrap (based on drawing $n$ times from the original data) is inconsistent in the sense that as $n \to \infty$, there is no guarantee that coverages of estimated confidence intervals will approach the nominal coverage (in this example, 95-percent).

The example considered here has been discussed previously by Bickel and Freedman (1981), Beran and Ducharme (1991), Simar and
Wilson (1999c), and others. The problem becomes much worse in higher-dimensional problems such as those encountered in efficiency estimation, where instead of only one dimension as in the present example, researchers must work with \((p + q)\) dimensions. The “failure” in the preceding example is not a failure of the bootstrap per se; rather, the problem is in construction of the bootstrap samples. Instead of resampling from the empirical distribution of the data as was done above, one must either (i) draw from a smoothed version of the empirical distribution as discussed by Bickel and Freedman (1981), or (ii) draw bootstrap samples of size \(m < n\) as discussed by Bretagnolle (1983), Swanepoel (1986), Beran and Srivastava (1985), and Athreya (1987). These two approaches are examined below.

3.1.2 A Smooth Bootstrap

Smoothing of the empirical distribution can be accomplished using kernel methods for density estimation; see Silverman (1986) or Scott (1992) for discussion. However, kernel density estimators are inconsistent near support boundaries. The easiest method to restore consistency is the reflection method described by Silverman (1986) and used by Simar and Wilson (1998). Continuing with the example above, reflect the data in \(Z_n\) around \(Z_{(n)}\) to form the set of \(2n\) elements \(Z_n^R = \{Z_1, \ldots, Z_n, 2Z_{(n)} - Z_1, \ldots, 2Z_{(n)} - Z_n\}\). Then use the kernel density estimator

\[
\hat{f}_R(t) = \frac{1}{2nh} \sum_{i=1}^{n} \left[ \phi \left( \frac{Z_i - t}{h} \right) + \phi \left( \frac{(2Z_{(n)} - Z_i) - t}{h} \right) \right]
\]  

(3.5)

with standard normal density function \(\phi(\cdot)\) as the kernel function and bandwidth \(h\) to estimate the density of the \(2n\) elements in \(Z_n^R\). The bandwidth parameter can be chosen using least-squares cross-validation (e.g., Silverman, 1978, 1986) or plug-in methods (e.g., Sheather and Jones, 1991) with the reflected data, but should then be rescaled by multiplying by \(2^{1/5}\) since the optimal bandwidth is of order \(O(n^{-1/5})\), and there are only \(n\) (as opposed to \(2n\)) real observations. Next, define

\[
\hat{f}_h(t) = \begin{cases} 
2\hat{f}_R^R(t) & \forall \, t \leq Z_{(n)}; \\
0 & \text{otherwise.}
\end{cases}
\]

(3.6)
Bootstrap Inference Using DEA and FDH Estimators

It can be proven that \( \hat{f}_h(t) \) is a consistent estimator of the true density of the \( Z_i \) in \( Z_n \) (e.g., see Schuster, 1985).

Drawing a bootstrap pseudo-sample from \( \hat{f}_h(t) \) is straightforward. Let \( \{Z_1^*, \ldots, Z_n^*\} \) be a naive bootstrap sample drawn from the empirical distribution of the reflected sample \( Z_n^R \), obtained by drawing \( n \) times independently, uniformly, and with replacement \( n \) times from \( Z_n^R \). Let \( \hat{\mu}^* = \frac{1}{n} \sum_{i=1}^{n} Z_i^* \) denote the sample mean of this naive bootstrap sample, and let \( \{\varepsilon_1, \ldots, \varepsilon_n\} \) be a set of \( n \) independent draws from \( N(0,1) \). Then compute

\[
Z_{i}^{**} = \begin{cases} 
Z_i^* + h\epsilon_i & \text{if } Z_i^* + h\epsilon_i \leq Z(n); \\
2Z(n) - Z_i^* - h\epsilon_i & \text{otherwise}, 
\end{cases} 
\]

for \( i = 1, \ldots, n \). Let \( \tilde{\sigma}_Z^2 \) denote the sample variance of the \( Z_i^{**} \). Finally, compute for each \( i = 1, \ldots, n \)

\[
Z_{i}^{***} = \hat{\mu}^* + (Z_i^{**} - \hat{\mu}^*) \left(1 + h^2 \tilde{\sigma}_Z^{-2}\right)^{-1/2}. 
\]

It can be shown that the \( Z_i^{***} \) behave as draws from \( \hat{f}_h(t) \) with

\[
E(Z_{i}^{***} | Z_n) = \hat{\mu}_n 
\]

and

\[
\text{VAR}(Z_{i}^{***} | Z_n) = \tilde{\sigma}_n^2 \left(1 + h^2 \left(n(\tilde{\sigma}_n^2 + h^2)^{-1}\right)\right), 
\]

where \( \hat{\mu}_n \) and \( \tilde{\sigma}_n^2 \) denote the sample mean and sample variance of the original \( n \) observations in \( Z_n \). See Simar and Wilson (1998, pp. 55–56) for additional discussion.

The smooth bootstrap is illustrated in the following R session:

```r
> set.seed(900001)
> require(KernSmooth)
Loading required package: KernSmooth
KernSmooth 2.23 loaded
Copyright M. P. Wand 1997–2009
> n=20
> nrep=1000
> coverage=0
> n2=n*2
> for (j in 1:1000) {
```
In this example, R’s \texttt{dpik} command (in the KernSmooth package) is used to implement the Sheather and Jones (1991) plug-in method for optimizing the kernel density estimator’s bandwidth, which is then adjusted by multiplying by $2^{1/5}$ to account for the fact that there are only $n$ observations, even though there are $2n$ elements after reflecting the data. The realized coverage of the 95-percent confidence intervals for $\zeta$ that are estimated on each of 1,000 Monte-Carlo trials is 0.942, which is not significantly different from the nominal value of 0.95.

### 3.1.3 Subsampling and the $m$-Bootstrap

The previous example illustrates that the smooth bootstrap can yield confidence intervals with coverage close to nominal values with a sample size of only 20 observations in a problem with only one dimension. Alternatively, sub-sampling methods can be used when support boundaries are estimated.
Recall from the discussion above that the naive bootstrap fails to provide consistent inference in the example of the univariate problem where \( Z \) is uniform on \([0, \zeta]\), and \( \zeta \) is estimated by the maximum likelihood estimator \( \hat{\zeta}_n \) because the bootstrap distribution fails to mimic the true, limiting \( \text{Exp}(1) \) distribution of the statistic \( R_n = n\zeta^{-1}\left(\zeta - \hat{\zeta}_n\right) \). Rather than having continuous support on \((0, \infty)\) as does the \( \text{Exp}(1) \) distribution, the bootstrap distribution of \( R^*_n = n\hat{\zeta}_n^{-1}\left(\hat{\zeta}_n - \hat{\zeta}_m^*\right) \) has support on a discrete, countable set of points. From the earlier discussion, it is clear that the problem arises because the sample maximum \( \zeta(n) \) is drawn too often; i.e., when bootstrap samples are formed by drawing \( n \) times uniformly, independently, and with replacement from \( Z_n \), a given bootstrap sample will contain \( \zeta(n) \) with probability approaching about 0.632 as seen in (3.4).

The idea of the \( m \)-bootstrap is to reduce the proportion of bootstrap samples that contain \( \zeta(n) \) by drawing (uniformly, independently, and with replacement) from \( Z_n \) only \( m \) times, with \( m < n \). Subsampling is similar, except that draws from \( Z_n \) are made without replacement. In either case, drawing bootstrap samples of size \( m < n \) reduces the probability that a given bootstrap sample will contain \( \zeta(n) \). In the one-dimensional setting, the \( m \)-bootstrap has been studied by Bretagnolle (1983), Swanepoel (1986), Beran and Srivastava (1985), Athreya (1987), Beran and Ducharme (1991), Bickel and Sakov (2008), and others. The subsampling method has been considered by Politis and Romano (1994, 1999), Politis et al. (2001), and others. In either case, the resulting confidence interval estimates have coverage approaching their nominal coverage rate provided \( m < n \), \( m \to \infty \), and \( m/n \to 0 \) as \( n \to \infty \).

Now consider a bootstrap sample \( \{Z^*_1, \ldots, Z^*_m\} \) obtained by sampling either with or without replacement. Denote the bootstrap estimate corresponding to this sample by \( \hat{\zeta}_m^* = Z^*_m \). For a bootstrap sample containing \( Z(n) \), \( \hat{\zeta}_m^* = Z(n) = \hat{\zeta}_n \), and hence \( R^*_m = m\hat{\zeta}_n^{-1}\left(\hat{\zeta}_n - \hat{\zeta}_m^*\right) = 0 \). In the case of the \( m \)-bootstrap, following the logic that led to (3.4) reveals that

\[
\Pr(R^*_m = 0) = 1 - (1 - n^{-1})^m. \tag{3.11}
\]
In the case of subsampling, where bootstrap samples are created by drawing without replacement,

\[ \Pr(R^* = 0) = 1 - \prod_{j=1}^{m} \left( 1 - \frac{1}{1 + n - j} \right). \tag{3.12} \]

It is easy to show that under the conditions on \( m \) given above, the probabilities in (3.11) and (3.12) converge to 0 as \( n \to \infty \). This solves the problem of the naive bootstrap.

The \( m \)-bootstrap is illustrated by the following simple \( R \) session:

```R
> bs.sub <- function(y,m,irep,nrep) {
+   n=length(y)
+   ir=irep==0,FALSE,TRUE)
+   zeta.hat=max(y)
+   zeta.star=vector(length=nrep)
+   for (irep in 1:nrep) {
+     zeta.star[irep]=max(sample(y,size=m,replace=ir))
+   }
+   g.star=(n/zeta.hat)*(zeta.hat-zeta.star)
+   qq=quantile(g.star,probs=c(0.025,0.975))
+   ci=c(zeta.hat/(1-qq[1]/n),zeta.hat/(1-qq[2]/n))
+   res=list(t1=zeta.hat,t2=ci)
+ }
>
> set.seed(900001)
> n=20
> ntrial=1000
> coverage=rep(0,20)
> p=vector(length=20)
> for (im in 1:20) {
+   for (itrial in 1:ntrial) {
+     y=runif(n)
+     res=bs.sub(y,m=im,irep=1,nrep=2000)
+     if (res$t2[1]<=1 & 1<=res$t2[2]) {
+       coverage[im]=coverage[im]+1
+     }
+   }
+   p[im]=1-(1-1/n)**im
+ }
> coverage=coverage/ntrial
> print(coverage)
[1] 1.000 1.000 1.000 1.000 0.999 0.997 1.000 0.992 0.993 0.990 0.976 0.981 0.952 0.956 0.959 0.937 0.902 0.897 0.885 0.913
```
Bootstrap Inference Using DEA and FDH Estimators

In this example, a function (bs.sub) is first created to estimate confidence intervals on a given Monte-Carlo trial, with sample size \( n = 20 \). In this function, \( R \)'s `sample` command is used to draw bootstrap samples of size \( m \). When the function is called from the main program, `irep` is set to 1 so that bootstrap samples are drawn with replacement. On a given Monte-Carlo trial, if the estimated 95-percent confidence interval covers \( \zeta \) (i.e., 1), the variable `coverage` is incremented by 1. At the end, `coverage` is divided by the number of Monte-Carlo trials, and the resulting values are printed. In addition, `p` is a vector giving the probabilities in (3.11) corresponding to each subsample size \( m = 1, 2, \ldots, 20 \), and these values are also printed.

This simple Monte-Carlo exercise shows that the realized coverage of the estimated 95-percent confidence intervals is closest to 0.95 when \( m = 13 \). With \( m = 13 \), the probability of a given bootstrap sample containing \( \hat{\zeta}_n \) is about 0.4867, which is less than the limiting value given in (3.4). Note that this probability can be reduced by using smaller values of \( m \), but this results in more information being lost, and in coverages that are too large.

Using again the `bs.sub` function defined above, the following \( R \) commands illustrate estimation of confidence intervals by subsampling, i.e., using resampling without replacement:

```r
> set.seed(900001)
> n=20
> ntrial=1000
> coverage=rep(0,20)
> p=vector(length=20)
> for (im in 1:20) {
+   for (itrial in 1:ntrial) {
+     y=runif(n)
+     res=bs.sub(y,m=im,irep=0,nrep=2000)
+     if (res$t2[1]<=1 & 1<=res$t2[2]) {
+       coverage[im]=coverage[im]+1
+     }
+   }
+ }
```
Comparing the results of this experiment with those of the previous experiment reveals some differences. The value of the subsample size \( m \) that yields coverage closest to 95-percent for the estimate confidence intervals is \( m = 10 \), which is smaller than for the \( m \)-bootstrap. At this level of \( m \), and with \( n = 20 \), the probability that a given bootstrap sample contains \( Z_n \) is 0.5, which is again smaller than the limiting value given in (3.4). Nonetheless, the point is to reduce the size of the bootstrap samples in order to avoid too-frequently drawing bootstrap samples that contain \( \zeta_n \), and both the \( m \)-bootstrap and the subsampling methods accomplish this.

Of course, things become more complicated in more than one dimension, but not prohibitively so as will be seen below. In addition, the examples here make clear that while the \( m \)-bootstrap and subsampling method give consistent inference in theory as long as \( m \to \infty \) and \( m/n \to 0 \) as \( n \to \infty \), in practice the choice of \( m \) is critical in finite-sample applications. This too will be addressed below.

### 3.2 Confidence Intervals for Efficiency of a Particular Point

#### 3.2.1 Smooth Bootstrap

The preceding examples and discussion should make clear that the naive bootstrap cannot give consistent inference when estimating support boundaries. This is precisely the case with frontier estimation. Note that in the univariate example given above, the goal was to estimate and make inference about the upper bound of a uniform
distribution. For $Z \sim \text{uniform on } [0, \zeta]$, nothing is changed by estimating the distance from a fixed point $z$ to the upper bound $\zeta$; in fact, one could define $\lambda = \frac{Z}{\zeta}$, and then estimate $\lambda$ by $\hat{\lambda} = \frac{Z(n)}{\zeta}$. Since $z$ is fixed and known (i.e., the analyst gets to choose $z$), it is clear that estimating $\zeta$ is no different from estimating $\lambda$. Moreover, it should be clear that $\lambda$ is simply an output efficiency measure with $q = 1$, and with all inputs fixed at the same values. Moving to more complicated situations, by increasing the number of dimensions, does not change the fact that the naive bootstrap cannot yield consistent inference when estimating support boundaries; see Simar and Wilson (1999c, 1999b) for additional discussion.

Simar and Wilson (1998) proposed a smooth bootstrap for making inference about firms’ efficiency in general settings with $p$ inputs and $q$ outputs, when efficiency is estimated by DEA estimators. The approach is a direct extension of the smooth bootstrap considered in the univariate example above. Given a sample $S_n = \{(X_i, Y_i)\}_{i=1}^n$ of observed input–output vectors, one can compute, for example, output efficiency estimates $\hat{\lambda}_i = \lambda(X_i, Y_i | \hat{P}_{VRS}) \geq 1$ for $i = 1, \ldots, n$. A naive bootstrap sample would be obtained by drawing from the pairs in $S_n$ $n$ times, independently, with replacement, and with probability equal to $\frac{1}{n}$ of selecting any given element on a particular draw. Instead, Simar and Wilson (1998) construct a bootstrap sample $S_n^* = \{(X_i^*, Y_i^*)\}_{i=1}^n$ by independently drawing $n$ values $\lambda_i^*$ from a smooth kernel estimate of the probability density of the $\hat{\lambda}_i$. Since the $\hat{\lambda}_i$ are bounded from below at 1, this requires use of the reflection method discussed above. In addition, a bandwidth must be chosen, but this can be optimized using least-squares cross-validation as discussed by Simar and Wilson (1998).

After drawing a set of values $\{\lambda_i^*\}_{i=1}^n$, the bootstrap sample is constructed by setting, for each $i = 1, \ldots, n$, $X_i^* = X_i$ and $Y_i^* = \hat{\lambda}_i Y_i / \lambda_i^*$. Note that multiplying $Y_i$ by $\hat{\lambda}_i$ projects the observation $(X_i, Y_i)$ onto the frontier estimate $\hat{P}_{VRS}$ in the output direction; dividing by $\lambda_i^*$ subsequently projects this point away from the estimated frontier, again along the output dimensions, by a random distance.

Given the bootstrap sample $S_n^*$, one can compute bootstrap estimates $\hat{\lambda}_i^* = \lambda(X_i, Y_i | \hat{P}_{VRS}^*)$, where $\hat{P}_{VRS}^*$ is the convex hull of the
free-disposal hull of the input–output vectors in the bootstrap sample $S^*_n$. Repeating this a large number $B$ times results in sets of bootstrap estimates $B^* = \{\hat{\lambda}_{i,k}^*\}_{b=1}^B$. For each $i = 1, \ldots, n$, these can be used with the corresponding estimates $\hat{\lambda}_i$ to construct estimated confidence intervals for the output efficiencies $\lambda_i$.

Although Simar and Wilson (1998) did not provide a formal proof of the consistency of their bootstrap, simulation results indicate that the method appears to work well in the sense of providing confidence interval estimates whose coverages are reasonably close to nominal values. The method is implemented by the `boot.sw98` command in FEAR (Wilson, 2008) for making inference about $\theta(x, y \mid P)$, $\lambda(x, y \mid P)$, and $\gamma(x, y \mid P)$; given the link between $\gamma(x, y \mid P)$ and $\delta(x, y \mid P)$, the command can also be used to make inference about the directional distance function.

Taking draws from an estimate of the marginal density of the $\hat{\lambda}_i$ in the Simar and Wilson (1998) method involves an implicit assumption of homogeneity, i.e., that inefficiency is distributed the same throughout $P$. Simar and Wilson (2000a) provide an alternative bootstrap method that avoids this assumption, and which permits the distribution of efficiency to vary, or to be heterogeneous. The method involves drawing input–output pairs from a smooth estimate of the density of inputs and outputs, which has bounded support over $P$. Again, due to the support boundary at $P^0$, this necessitates use of the reflection method, but here, the reflection is along multiple dimensions, and around an unknown and possibly nonlinear support boundary. Simar and Wilson (2000a) implement the required reflection by transforming the data from Cartesian coordinates to cylindrical coordinates, and exploiting the fact that $\lambda_i$ and $\hat{\lambda}_i$ measure distance from the frontier and its estimate. See Simar and Wilson (2000a) for details. The method is more complicate than the Simar and Wilson (1998) method, but this is the price one must pay for relaxing the homogeneity assumption.

### 3.2.2 Double-Smooth Bootstrap

As noted earlier, Kneip et al. (2008) provided the first results on the asymptotic distribution of $\theta(x, y \mid \hat{P}_{VRS})$, and proved consistency of two
Bootstrap methods for making inference about \( \theta(x, y \mid P) \). The results extend to \( \lambda(x, y \mid \hat{P}_{\text{VRS}}) \) after some changes in notation. Wilson (2011) extended these results to the hyperbolic DEA estimator \( \gamma(x, y \mid \hat{P}_{\text{VRS}}) \), and Simar et al. (2012) extended the results to the directional DEA estimator \( \delta(x, y \mid u, v, P_{\text{VRS}}) \).

The first bootstrap method for which Kneip et al. (2008) proved consistency is based on subsampling; however, Kneip et al. did not provide any guidance for choosing the size \( m < n \) of the subsamples, and simulation results showed that the choice of \( m \) in finite samples is critical. The second bootstrap method for which Kneip et al. (2008) proved consistency involves smoothing not only the \((p + q)\)-dimensional density of inputs and outputs, but also smoothing the initial DEA-VRS frontier estimate, while preserving strict convexity. Simulation results indicate the method works well in terms of yielding estimated confidence intervals whose coverage is reasonably close to nominal values in sample sizes typically encountered in applied work. However, the double-smooth method is difficult to program, involves numerical difficulties, and is computationally burdensome. Nonetheless, Kneip et al. (2008) provided the first proof of consistency of a bootstrap for efficiency estimated by DEA; moreover, by providing the asymptotic distribution of the DEA-VRS estimator, Kneip et al. opened the door for further research aimed at establishing consistency of other bootstrap methods.

### 3.2.3 Computationally Efficient Smooth Bootstrap

Kneip et al. (2011) provided a computationally-efficient bootstrap based on some of the ideas of Kneip et al. (2008) that avoids some of the numerical difficulties of the double-smooth bootstrap. The idea is simple: naive resampling does not work, but the problem is at the boundary. In other words, only naive resampling of points lying near the frontier prevents the naive bootstrap from providing consistent inference; naive resampling of points far from the frontier has no such consequence. Consequently, one should be able to draw naively among observations lying “far” from the frontier, and draw from a uniform distribution with support “near” the frontier.
Distinguishing between “near” and “far” from the frontier involves establishing a neighborhood along the DEA-VRS frontier estimate; the size of this neighborhood (i.e., its extend from the initial frontier estimate), is controlled by a tuning parameter (i.e., a bandwidth) that can be selected using a simple rule-of-thumb. To construct bootstrap samples, input–output pairs are resampled naively; whenever a pair lying in the neighborhood of the initial frontier estimate, this draw is replaced by a draw from a uniform distribution. In order to control the depth of the neighborhood of the initial frontier estimate, data are transformed using an orthonormal basis of the input or output vector (depending on whether one is working in an input or output orientation). Although the transformation must be recomputed for each point about which inference is made, the method remains computationally efficient.

To prove consistency of the method, Kneip et al. (2011) required smoothing of the initial DEA-VRS frontier estimate, which requires a second bandwidth. Smoothing the DEA-VRS frontier estimate is facilitated by the transformation of the data mentioned above. Kneip et al. describe a cross-validation method for optimizing the second bandwidth, and in addition provide simulation evidence that shows the method yields confidence interval estimates with achieved coverages close to nominal values.

3.2 Subsampling

As noted above, working with the input-oriented DEA-VRS estimator, Kneip et al. (2008) established consistency of both (i) a double-smooth bootstrap in which both the efficiency distribution and the frontier estimate are smoothed, as well as (ii) a subsampling bootstrap, in which \( m < n \) input–output pairs are resampled naively from the original sample \( S_n \). Also as noted above, the double-smooth bootstrap is computationally burdensome and difficult to program. Of the two alternatives, the subsampling method is much more promising for practitioners, although Kneip et al. (2008) did not provide a method for optimizing the subsample size \( m \), and simulation results reported by Kneip et al. show that the method’s performance in terms of coverages
achieved by estimated confidence intervals depends critically on the choice of the subsample size $m$, and simulation results reported by Kneip et al. show that the method’s performance in terms of coverages achieved by estimated confidence intervals depends critically on the choice of $m$.

In more than one dimension, one must account for various assumptions that might be made about the shape of the frontier, and hence the different estimators of distance to the frontier that might be used. Where the production set $P$ is assumed to be convex, for a fixed point $(x, y) \in P$, one might use either $\theta(x, y | \hat{P}_{VRS})$, $\lambda(x, y | \hat{P}_{VRS})$, $\gamma(x, y | \hat{P}_{VRS})$ or $\delta(x, y | u, v, \hat{P}_{VRS})$ to estimate distance from $(x, y)$ to the frontier $P^\partial$. If one assumes $P^\partial$ is globally CRS, one might estimate efficiency using either $\theta(x, y | \hat{P}_{CRS})$, $\lambda(x, y | \hat{P}_{CRS})$, $\gamma(x, y | \hat{P}_{CRS})$ or $\delta(x, y | u, v, \hat{P}_{CRS})$. On the other hand, if one is not willing to assume $P$ is convex, efficiency can be estimated by $\theta(x, y | \hat{P}_{FDH})$, $\lambda(x, y | \hat{P}_{FDH})$, $\gamma(x, y | \hat{P}_{FDH})$ or $\delta(x, y | u, v, \hat{P}_{FDH})$.

In the case of the input and output-oriented DEA-VRS estimators, Kneip et al. (2008) develop the asymptotic theory needed for using subsampling to estimate confidence intervals for the efficiency of a particular point $(x, y) \in P$, and in addition give an algorithm with computational details. Wilson (2011) extends this theory to the hyperbolic estimator $\gamma(x, y | \hat{P}_{VRS})$, while Simar et al. (2012) extend the theory to the directional distance function estimator $\delta(x, y | u, v, \hat{P}_{VRS})$. As mentioned previously, analogous results due to Park et al. (2010), Wilson (2011), and Simar and Vanhems (2012) are available for the DEA-CRS and FDH estimators.

Results on the asymptotic distributions of the various DEA and FDH estimators along with results from Politis et al. (2001) were used by Simar and Wilson (2011a) to establish consistency of bootstrap subsampling to estimate confidence intervals for individual firms’ efficiencies in all the cases listed above. In addition, Simar and Wilson (2011a) provide a data-driven method for optimizing the choice of subsample size $m$, and in addition provide simulation results indicating that the method works well for estimating confidence intervals with good coverages.
In each case, the bootstrap principle is based on approximations of the form
\[ m^\kappa \left( \frac{\hat{\tau}}{\hat{\tau}^*} - 1 \right) \stackrel{\text{approx.}}{\sim} n^\kappa \left( \frac{\tau}{\tau} - 1 \right), \tag{3.13} \]
where \( n, m \to \infty, m/n \to 0; \) \( \tau \) represents (for a fixed point \((x, y)\)) one of the distance functions \( \theta(x, y \mid P), \lambda(x, y \mid P), \gamma(x, y \mid P) \), or \( \delta(x, y \mid P); \) \( \hat{\tau} \) represents either the corresponding DEA-VRS, DEA-CRS, or FDH estimator; and \( \hat{\tau}^* \) represents the corresponding bootstrap estimate obtained by applying the same estimator for the fixed point \((x, y)\) but using a bootstrap sample of size \( m \) to define the frontier estimate. The exponent \( \kappa \) in (3.13) is \( \frac{2}{p+q+1} \) for the DEA-VRS estimators, \( \frac{2}{p+q} \) for the DEA-CRS estimators, or \( \frac{1}{p+q} \) for the FDH estimators. Then a bootstrap estimate of a \( (1 - \alpha) \) confidence interval for \( \tau \) is given by
\[ \left[ \hat{\tau} \left( 1 + n^{-\kappa \hat{\psi}_{\alpha/2,m}} \right), \hat{\tau} \left( 1 + n^{-\kappa (1-\alpha)_{\hat{\psi}_{1-\alpha/2,m}}} \right) \right], \tag{3.14} \]
where \( \hat{\psi}_{\alpha,m} \) is the \( \alpha \)-quantile of the bootstrap distribution of \( m^\kappa \left( \frac{\hat{\tau}}{\hat{\tau}^*} - 1 \right). \)

As in the univariate example considered earlier, when the subsampling is done without replacement, the bootstrap distribution in (3.13) will become too concentrated as \( m \to n; \) if fact, if \( m = n \), the bootstrap distribution collapses to a single probability mass. On the other hand, as \( m \to 0 \), the resulting confidence interval estimates will either under- or over-cover \( \theta(x, y) \) since too much information is lost. An optimal value of \( m \) will lie between these extremes, and Politis et al. (2001) show that the approximation should work well for a wide range of values for \( m \); thus, the idea is to choose a value of \( m \) that yields “stable” estimates for confidence intervals.

Kneip et al. (2008) proved, for the input-oriented DEA-VRS estimator, that the approximation in (3.13) is consistent for any choice \( m = n^\omega \) with \( \omega \in (0, 1) \). However, as noted previously, simulation results in Kneip et al. make clear that the quality of the approximation in finite samples depends crucially on \( \omega \). Both Bickel and Sakov (2008) and Politis et al. (2001) proposed computing the object of interest (e.g., a
confidence interval estimate or critical value of a test) for various values of \( m \), and then choosing the value of \( m \) that minimizes some measure of volatility of the object of interest.

In particular, once can estimate confidence intervals of size \( \alpha \) using various bootstrap subsample sizes \( m_1 < m_2 < \ldots < m_J \), and then measure volatility corresponding to \( m_j \) by computing the standard deviations of the bounds of the estimated confidence intervals corresponding to \( m_{j-k}, \ldots, m_j, \ldots, m_{j+k} \), where \( k \) is a small integer (e.g., \( k = 1, 2, \) or \( 3 \)) and \( j = (k+1), \ldots, (J-k) \). The subsample size \( m \) would then be chosen as the \( m_j \) yielding the smallest measure of volatility; explicit details are given in Simar and Wilson (2011a), and this method has been implemented in the `dea.subsample` command in `FEAR` (Wilson, 2008).
4

Robust Order-\( m \) Estimators

4.1 The Need for Robustness

Both DEA and FDH estimators of \( P \) fully envelop all of the sample observations; for this reason they are called “full-envelopment estimators.” Full-envelopment estimators are particularly sensitive to outliers, or extreme observations. These extreme points may disproportionately, and perhaps misleadingly, influence the evaluation of the performance of other firms; a single outlier, depending on its location, can distort efficiency estimates for one or more firms to an arbitrarily large degree. This feature is a consequence of fully enveloping the sample observations, and also plagues parametric frontier estimators when deterministic frontier models are considered. One approach to this problem would be to identify any outliers in the data, and then perhaps delete them if they result from corrupted data. A number of techniques exist for finding outliers in frontier settings (e.g., Wilson, 1993, 1995; Simar, 2003; Porembski et al., 2005), Alternatively, one can use robust, partial frontier estimators, which also offer other advantages.

Recent papers by Cazals et al. (2002), Daraio and Simar (2005), Aragon et al. (2005), and Daouia and Simar (2005, 2007), Wheelock
Robust Order-m Estimators

and Wilson (2008), Wilson (2011), Simar and Vanhems (2012), and Simar et al. (2012) have developed robust alternatives to the traditional FDH and DEA estimators. These new estimators involve the concept of a “partial” frontier, as opposed to the traditional idea of a “full” frontier that envelops all the data. The new ideas replace the goals of estimating the absolute lowest (uppermost) technically achievable level of input (output) for a given level of output (input) with the idea of estimating something “close” to these quantities.

As such, the partial frontiers considered in this section are “less extreme” than the full frontier. Since estimation of partial frontiers avoids many of the statistical problems inherent in estimating a full frontier, partial frontiers provide a more useful benchmark against which firms can be compared. Natural nonparametric estimators of these partial frontiers will be proposed that are very easy and fast to compute. Daouia and Gijbels (2011b) analyze the properties of these estimators from the viewpoint of robustness theory; see also Daouia and Ruiz-Gazen (2006).

Two classes of partial frontiers have been proposed: (i) order-m frontiers, where \( m \) can be viewed as a trimming parameter, and (ii) order-\( \alpha \) quantile frontiers, analogous to traditional quantile functions but adapted to the frontier problem. It turns out that the resulting nonparametric estimators have much better properties than the usual nonparametric (DEA or FDH) frontier estimators; the new estimators do not suffer from the curse of dimensionality, and the standard parametric root-\( n \) rate of convergence is achieved along with asymptotic normality.

Another interesting feature is that both estimators of these partial frontiers are also consistent estimators of the full frontier, by allowing the order of the frontier (\( m \) or \( \alpha \)) to grow (at an appropriate rate) with increasing sample size. These new estimators of the full frontier share the same asymptotic properties as FDH estimators. But, in finite samples, the new estimators do not envelop all the data, and so are much more robust with respect to outliers and extreme values in the sample than the usual FDH or DEA estimators. As a side benefit, these “partial-order” frontiers and their estimators provide very useful tools for detecting outliers (e.g., Simar, 2003). The basic ideas are
based on the probabilistic formulation of a production process discussed earlier.

### 4.2 Definitions and Basic Ideas

#### 4.2.1 Input Orientation

Recall that the production process can be modeled in terms of the probability distribution function $H_{XY}(x, y)$ defined in (2.40), allowing the input, output, hyperbolic, and directional efficiency measures to be expressed in terms of this distribution function in (2.48)–(2.51). For the input orientation, the full frontier and the resulting Debreu–Farrell efficiency scores are characterized by properties of the conditional distribution $F_{X|Y}(x | y) = \Pr(X \leq x | Y \geq y)$ appearing in (2.42) that describes the behavior of firms which produce at least level $y$ of output. In terms of the input efficiency score,

$$
\theta(x, y | \mathcal{P}) = \inf\{\theta | F_{X|Y}(\theta x | y) > 0\},
$$

assuming free disposability; see (2.44).

The full frontier $\mathcal{P}^\theta$ can be viewed as the boundary of support of $f(x, y)$, or equivalently, of $H_{XY}(x, y)$. The full frontier can also be viewed as the lower radial boundary of $F_{X|Y}(x | y)$, i.e., as the minimum achievable lower boundary of inputs for all possible firms producing at least level $y$ of output. This is a rather extreme and absolute theoretical concept; it gives the full minimum achievable level of input over all production plans that are technically feasible. However, the full frontier $\mathcal{P}^\theta$ is just one of perhaps many benchmarks by which firms’ performances might be measured. In the input orientation, an alternative benchmark is obtained by defining the expected minimum input achieved by any $m$ firms chosen randomly from the population of firms producing at least output level $y$. As will be seen below, for finite values of $m$, this idea provides potentially more useful benchmarks than the full frontier $\mathcal{P}^\theta$ for comparing firms in terms of their efficiencies. It will also become apparent that if $m$ goes to infinity, the problem becomes identical to FDH estimation of the full frontier $\mathcal{P}^\theta$.

To be more precise, suppose an output level $y$ is given. Consider $m$ iid random variables $X_i$, $i = 1, \ldots, m$ drawn from the conditional
p-variate distribution function $F_X(\cdot \mid y)$, and define the set

$$
P_m(y) = \bigcup_{j=1}^{m} \{(\bar{x}, \bar{y}) \in \mathbb{R}_+^{p+q} \mid \bar{x} \geq X_j, \bar{y} \geq y\}. \quad (4.2)
$$

This random set is the free disposal hull of $m$ randomly-chosen firms that produce at least the level $y$ of output. Then, for any $x$ and the given $y$, the Debreu–Farrell input efficiency score relative to the set $P_m(y)$ is

$$
\tilde{\theta}_m(x, y) = \inf \{\theta \mid (\theta x, y) \in P_m(y)\}. \quad (4.3)
$$

The set $P_m(y)$ is random, since it depends on the random variables $X_i$ with (conditional) distribution function $F_X(\cdot \mid y)$. Hence the efficiency score $\tilde{\theta}_m(x, y)$ is also random. For a given realization of the $m$ values $X_i$, a realization of $\tilde{\theta}_m(x, y)$ is obtained by computing

$$
\tilde{\theta}_m(x, y) = \min_{i=1, \ldots, m} \left\{ \max_{j=1, \ldots, p} \left( \frac{X_i^j}{x^j} \right) \right\}. \quad (4.4)
$$

The order-$m$ input efficiency score is defined as follows:

**Definition 4.1.** For all $\underline{y}$ such that $S_Y(\underline{y}) = \Pr(Y \geq \underline{y}) > 0$, the expected order-$m$ input efficiency measure, referred to as the order-$m$ input efficiency score, is given by

$$
\theta_m(x, y) \equiv E(\tilde{\theta}_m(x, y) \mid Y \geq y). \quad (4.5)
$$

The order-$m$ input efficiency score benchmarks a unit operating at $(x, y)$ against the expected minimum input among $m$ peers randomly drawn from the population of units producing at least $y$. This efficiency measure in turn defines an order-$m$ input efficient frontier. For any $(x, y) \in \mathcal{P}$, the expected minimum level of inputs of order-$m$ for a unit producing output level $y$ and for an input mix determined by the vector $x$ is given by

$$
x^\partial_m(y) = \theta_m(x, y)x. \quad (4.6)
$$

By contrast, recall that the full frontier $\mathcal{P}^\partial$ is defined (at output level $y$) by $x^\partial(y) = \theta(x, y \mid \mathcal{P})x$ in (1.8).
**Definition 4.2.** The expected input efficient frontier of order \(m\), known as the input order-\(m\) efficient frontier and denoted by \(P^\partial_m\), is defined as

\[
P^\partial_m \equiv \left\{ (\tilde{x}, \tilde{y}) \mid \tilde{x} = x_{\theta m}(x, y), \tilde{y} = y, (x, y) \in \mathcal{P} \right\}.
\]

(4.7)

If \(x\) is univariate, the order-\(m\) input frontier can be described by an input function of order-\(m\):

\[
x^\partial_m(y) = \phi_m(y) = E\left[\min(X^1, \ldots, X^m) \mid Y \geq y\right]
= \int_0^\infty \left[1 - F_{X|Y}(x \mid Y \geq y)\right]^m \, dx.
\]

(4.8)

This concept is illustrated in the following example. In cases where the integral in (4.8) has no closed-form solution, it can be computed easily using Gaussian quadrature.

**Example 4.1.** Suppose \(p = q = 1\), and

\[
f(x, y) = \begin{cases} 
2 & \forall x \in [0, 1] \text{ and } y \in (0, x]; \\
0 & \text{otherwise}.
\end{cases}
\]

(4.9)

It is easy to show that \(f(x \mid Y \geq y) = 2(1 - y)^{-2}(x - y) \forall x \in (y, 1]\) and hence

\[
F_{X|Y}(x \mid Y \geq y) = \begin{cases} 
1 & \forall x > 1; \\
(x - y)^2(1 - y)^{-2} & \forall x \in (y, 1]; \\
0 & \forall x \leq y.
\end{cases}
\]

(4.10)

Substituting for \(F_{X|Y}\) in (4.8), the order-\(m\) input frontier can be computed easily.

Figure 4.1 shows \(P^\partial\) as a solid line, and the order-\(m\) input frontier with \(m = 50\) as a dashed line. The order-\(m\) input frontier lies inside the production set \(\mathcal{P}\), and has steeper slope than the full frontier. Table 4.1 shows values of \(\phi_m(y)\) corresponding to various values of \(m\) and \(y = 0.3\), confirming that as \(m\) increase toward infinity, \(\phi_m(y)\) approaches \(\phi(y)\).
Robust Order-\(m\) Estimators

Fig. 4.1 Order-\(m\) input frontier (\(m = 50\)).

Table 4.1. Order-\(m\) input frontier function for \(y = 0.3\).

<table>
<thead>
<tr>
<th>(m)</th>
<th>(\phi_m(y))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10^1)</td>
<td>0.48918</td>
</tr>
<tr>
<td>(10^2)</td>
<td>0.36180</td>
</tr>
<tr>
<td>(10^3)</td>
<td>0.31961</td>
</tr>
<tr>
<td>(10^4)</td>
<td>0.30620</td>
</tr>
<tr>
<td>(10^5)</td>
<td>0.30196</td>
</tr>
<tr>
<td>(10^6)</td>
<td>0.30062</td>
</tr>
<tr>
<td>(10^7)</td>
<td>0.30020</td>
</tr>
<tr>
<td>(10^8)</td>
<td>0.30000</td>
</tr>
</tbody>
</table>

Figure 4.2, which resembles Figure 2.3, provides an additional illustration. Figure 4.2 shows the input-oriented order-\(m\) frontier for a particular value \(y\) in the bivariate case. For this (or for any other) \(y\), 

\[ \phi_m(y) = E \left[ \min(X^1, \ldots, X^m) \mid Y \geq y \right]. \]

The stars represent \(m\) draws \(X^1, \ldots, X^m\) for \(Y \geq y_0\). For finite \(m\), the input-oriented order-\(m\) frontier function is under (to the right of) the full frontier function \(\phi(y)\) as defined in (2.52) and shown by the upward-sloping curve in Figure 4.2.
As a matter of fact, the order-m efficiency score has an explicit expression which only depends on the distribution \( F_{X|Y}(x | y) \). It can be proved that

\[
\theta_m(x, y) = \int_0^\infty (1 - F_{X|Y}(ux | y))^m du,
\]

\[
= \theta(x, y | P) + \int_{\theta(x, y | P)}^\infty (1 - F_{X|Y}(ux | y))^m du. \quad (4.11)
\]

Note that the integral is univariate, even in a full multivariate setting where \((p, q) \geq 1\). As expected, as \( m \to \infty \), the order-m frontier tends to the full frontier. From (4.11), it is clear that

\[
\lim_{m \to \infty} \theta_m(x, y) = \theta(x, y | P) \quad (4.12)
\]

and

\[
\lim_{m \to \infty} x_m^\beta(y) = x^\beta(y) \quad (4.13)
\]

since the integrand in (4.11) is less or equal to 1.

### 4.2.2 Output Orientation

Working in the output-orientation, suppose input levels \( x \) in the interior of the support of \( X \) is given, and consider \( m \) iid random variables \( Y_i \),...
Robust Order-$m$ Estimators

$i = 1, \ldots, m$ drawn from the conditional $q$-variate distribution function $F_{Y \mid X}(y \mid x) = \text{Prob}(Y \leq y \mid X \leq x)$. Analogous to (4.2), define the random set

$$\mathcal{P}_m(x) = \bigcup_{j=1}^{m} \{(\tilde{x}, \tilde{y}) \in \mathbb{R}_+^{p+q} \mid \tilde{x} \leq x, Y_j \leq \tilde{y}\}. \quad (4.14)$$

Analogous to the set $\mathcal{P}(y)$ defined in (4.2), the random set $\mathcal{P}_m(y)$ is the free disposal hull of $m$ randomly-chosen firms that use no more than $x$ levels of the $p$ inputs. For any $y$ and the given $x$, the Debreu–Farrell output efficiency score relative to the set $\mathcal{P}_m(x)$ is

$$\tilde{\lambda}_m(x, y) = \sup \{\lambda \mid (x, \lambda y) \in \mathcal{P}_m(x)\}. \quad (4.15)$$

Of course, the efficiency score $\tilde{\lambda}_m(x, y)$ is random, since the set $\mathcal{P}_m(x)$ is random. For a given realization of the $m$ values $Y_i$, a realization of $\tilde{\lambda}_m(x, y)$ us determined by

$$\tilde{\lambda}_m(x, y) = \max_{i=1, \ldots, m} \left\{ \min_{j=1, \ldots, p} \left( \frac{Y_j}{y^j} \right) \right\}. \quad (4.16)$$

The order-$m$ output efficiency measure is defined as follows:

**Definition 4.3.** For any $y \in \mathbb{R}_+^q$, the (expected) order-$m$ output efficiency measure, denoted by $\lambda_m(x, y)$, is defined for all $x$ in the interior of the support of $X$ by

$$\lambda_m(x, y) \equiv E(\tilde{\lambda}_m(x, y) \mid X \leq x), \quad (4.17)$$

where the expectation is assumed to exist.

The order-$m$ output efficiency score benchmarks a unit operating at $(x, y)$ against the expected maximum output among $m$ peers randomly drawn from the population of units using no more than $x$ amounts of inputs. This efficiency measure in turn defines an order-$m$ output efficient frontier, different from the order-$m$ input efficient frontier. For any $(x, y) \in \mathcal{P}$, the expected maximum level of outputs of order-$m$ for a unit using input level $x$ and for an output mix determined by the vector $y$ is given by

$$y_m^0(x) = \lambda_m(x, y) y. \quad (4.18)$$
To draw a contrast, recall that the full frontier $\mathcal{P}^\partial$ is defined (at input level $x$) by $y^\partial(x) = \lambda(x, y \mid \mathcal{P})y$ in (1.10).

**Definition 4.4.** The expected output efficient frontier of order $m$, known as the output order-$m$ efficient frontier and denoted by $\mathcal{P}^\partial_{\text{out}}$, is as

$$
\mathcal{P}^\partial_{\text{out}} \equiv \{(\bar{x}, \bar{y}) \mid \bar{x} = x, \bar{y} = y\lambda_m(x, y), (x, y) \in \mathcal{P}\}. \quad (4.19)
$$

If $y$ is univariate, the order-$m$ output frontier can be described by an output function of order-$m$:

$$
y^\partial_m(x) = \psi_m(x) = E[\max(Y^1, \ldots, Y^m) \mid X \leq x]
= \int_0^\infty [1 - (1 - S_{Y|X}(y \mid x))^m] \, dy. \quad (4.20)
$$

This concept is illustrated in the following example.

**Example 4.2.** Suppose $p = q = 1$, and let $f(x, y)$ be given by (4.9) in Example 4.1. Then it is straightforward to show that $f(y \mid X \leq x) = 2x^{-1}[1 - x^{-1}y] \forall y \in [0, x]$ and hence

$$
F_{Y|X}(y \mid X \leq x) = \begin{cases} 
1 & \forall y > x; \\
2x^{-1}y - x^{-2}y^2 & \forall y \in (0, x]; \\
0 & \forall y \leq 0.
\end{cases} \quad (4.21)
$$

Noting that $S_{Y|X}(y \mid x) = 1 - F_{Y|X}(y \mid x)$ when $q = 1$ and substituting into (4.20) allows one to compute the order-$m$ output frontier.

Figure 4.3 shows $\mathcal{P}^\partial$ as a solid line, and the order-$m$ output frontier with $m = 50$ as a dashed line. As with the order-$m$ output frontier, the order-$m$ output frontier lies inside the production set $\mathcal{P}$, but in contrast to the order-$m$ input frontier, the order-$m$ output frontier has less steep slope than the full frontier. Table 4.2 shows values of $\psi_m(x)$ corresponding to various values of $m$ and $x = 0.7$, confirming that $\psi_m(x) \to \psi(x)$ as $m \to \infty$.

Figure 4.4 gives an additional illustration of the order-$m$ output frontier, showing the output-oriented order-$m$ frontier for a particular
value \( x \) in the bivariate case with \( p = q = 1 \). For this and any other \( x \),
\[
    \psi_m(x) = E \left[ \max(Y^1, \ldots, Y^m) \mid X \leq x \right].
\]
The figure shows \( m \) draws \((X_i, Y_i)\) such that \( X_i \leq x \). For finite \( m \), the order-\( m \) output frontier function \( \psi_m(x) \) lies below the full frontier function \( \psi(x) \) defined in (2.53) and shown by the solid, upward sloping curve in Figure 4.4.

Analogous to the input-oriented case, the output-oriented order-\( m \) efficiency score has an explicit expression that depends only on the
4.2 Definitions and Basic Ideas

Fig. 4.4 Output-oriented order-$m$ frontier for the bivariate case.

Consider a set of $m$ iid random variables $\{(X_j, Y_j)\}_{j=1}^m$ drawn from the density $f(x, y)$ defined in Assumption 2.1 with bounded support over $\mathcal{P}$. Define the random set

$$\mathcal{P}_m \equiv \bigcup_{j=1}^m \left\{ (x, y) \in \mathbb{R}_+^{p+q} \mid x \geq X_j, y \leq Y_j \right\}. \quad (4.24)$$

This random set is the free disposal hull of $m$ randomly chosen firms; note there is no conditioning here on input or output levels, which is
different from the input- and output-oriented cases considered above. Then, for any \((x, y) \in \mathbb{R}^{p+q}_+\), the hyperbolic efficiency score relative to the set \(P_m\) is

\[
\tilde{\gamma}_m(x, y) \equiv \inf\{\gamma \mid (\gamma^{-1}x, \gamma y) \in \mathcal{P}_m\}.
\] (4.25)

Since the set \(\mathcal{P}_m\) is random, the efficiency score \(\tilde{\gamma}_m(x, y)\) is also random. For a given draw of \(m\) input–output pairs \((X_i, Y_i), i = 1, \ldots, m\), the corresponding realization of \(\tilde{\gamma}_m(x, y)\) can be computed by

\[
\tilde{\gamma}_{m,n}(x, y) = \max_{i=1, \ldots, m} \left( \min_{j=1, \ldots, p} \left( \frac{x_j X_i^j}{Y_i^k} \right) \right).
\] (4.26)

The hyperbolic order-\(m\) efficiency score is defined as follows:

**Definition 4.5.** For all \((x, y) \in \mathbb{R}^{p+q}_+\), the expected hyperbolic order-\(m\) efficiency measure, known as the hyperbolic order-\(m\) efficiency score, is given by

\[
\gamma_m(x, y) \equiv E(\tilde{\gamma}_m(x, y))
\] (4.27)

provided \(E(\tilde{\gamma}_m(x, y))\) exists.

Wilson (2011, Theorem 2.1) establishes that

\[
\gamma_m(x, y) = \int_0^\infty \left[ 1 - H_{XY}(u^{-1}x, u y) \right]^m du.
\] (4.28)

The hyperbolic order-\(m\) efficiency score benchmarks a unit operating at \((x, y)\) against the expected best performance among \(m\) randomly chosen peers, along a hyperbolic path passing through the point \((x, y)\), subject to free-disposability of the production possibilities set \(\mathcal{P}\). This efficiency measure in turn defines a hyperbolic order-\(m\) efficient frontier.

**Definition 4.6.** The expected hyperbolic, efficient frontier of order \(m\), known as the hyperbolic order-\(m\) efficient frontier and denoted by \(\mathcal{P}_m^\partial\), is defined as

\[
\mathcal{P}_m^\partial \equiv \left\{ (\tilde{x}, \tilde{y}) \mid \tilde{x} = x\gamma_m(x, y)^{-1}, \tilde{y} = y\gamma_m(x, y), (x, y) \in \mathcal{P} \right\}.
\] (4.29)
**Definition 4.7.** For any \((x, y) \in \mathbb{R}^{p+q}_+\), the expected minimum input, maximum output combination of order \(m\), denoted by \((x_m^\partial(x, y), y_m^\partial(x, y))\), is given by

\[
x_m^\partial(x, y) \equiv x^\gamma_m(x, y)^{-1}
\]

and

\[
y_m^\partial(x, y) \equiv y^\gamma_m(x, y).
\]

The limiting behavior of the expected minimum input, maximum output combination of order \(m\), as \(m \to \infty\), is similar to that established by Cazals et al. (2002) for the expected minimum input of order \(m\) (conditional on \(y\)) given in (4.12) and (4.13). In particular, provided \(E(\tilde{\gamma}_m(x, y))\) exists,

\[
\lim_{m \to \infty} E(\tilde{\gamma}_m(x, y)) = \lim_{m \to \infty} \gamma_m(x, y) = \gamma(x, y) \quad (4.32)
\]

and

\[
\lim_{m \to \infty} \left( x_m^\partial(x, y), y_m^\partial(x, y) \right) = \left( x^\partial(x, y), y^\partial(x, y) \right). \quad (4.33)
\]

Details and proofs are given in Wilson (2011).

**Example 4.3.** Again suppose that \(p = q = 1\), and let \(f(x, y)\) be given by (4.9) in Example 4.1. It is clear that

\[
H_{XY}(x, y) = F_{Y|X}(y \mid X \leq x) [1 - F_Y(y)]; \quad (4.34)
\]

\(F_{Y|X}(y \mid X \leq x)\) is given in (4.10), and it is easy to show that

\[
F_Y(y) = \begin{cases} 
1 & \forall y > 1; \\
2y - y^2 & \forall y \in (0, \infty]; \\
0 & \forall y \leq 0.
\end{cases} \quad (4.35)
\]

For \(f(x, y)\) given by (4.9), the full frontier \(\mathcal{P}^\partial\) is a 45-degree line from \((0,0)\) to \((1,1)\) as shown by the solid line in Figure 4.5. Figure 4.5 also shows the input and output order-\(m\) frontiers from
Examples 4.1 and 4.2, for $m = 50$, as dotted lines. The hyperbolic order-$m$ frontier (with $m = 50$) can be found by substituting (4.9) and (4.35) into (4.34), and then substituting the resulting expression into (4.28) to compute $\gamma_m(x, y)$ for points along the frontier, and then, for $(x, y) \in \mathcal{P}^\delta$, computing $(\gamma_m(x, y)x, \gamma_m(x, y)^{-1}y)$ to trace out the hyperbolic order-$m$ frontier. Doing so yields the dashed curve in Figure 4.5. Figure 4.5 illustrates that whereas the input order-$m$ frontier diverges (from left to right) from the full frontier, and the output order-$m$ frontier converges (from left to right) from the full frontier, the hyperbolic order-$m$ frontier neither diverges nor converges, at least not to the extent of the input- or output-oriented order-$m$ frontiers. The hyperbolic order-$m$ frontier shown in Figure 4.5 is roughly parallel to the full frontier; this is a consequence of the uniform distribution assumed in (4.9) in Example 4.1.

Fig. 4.5 Hyperbolic order-$m$ frontier ($m = 50$).
4.2 Definitions and Basic Ideas

4.2.4 Directional Distances

Simar and Vanhems (2012) use the ideas of Cazals et al. (2002) and Wilson (2011) to extend the order-\(m\) concept to directional distances. As before, consider \(m\) iid random variables \(S_m = \{(X_i, Y_i)\}_{i=1}^m\) drawn from the joint density \(f(x, y)\) with bounded support over \(P\), and recall the random set \(P_m\) defined in (4.24), but with \(\mathbb{R}^{p+q}\) replacing \(\mathbb{R}_+^{p+q}\). Then for any \((x, y) \in \mathbb{R}^{p+q}\),

\[
\tilde{\delta}_m(x, y \mid u, v, P_m) = \sup \{\delta \mid (x - \delta u, y + \delta v) \in P_m\} \tag{4.36}
\]
defines a random, directional distance along the lines of (1.13). For a given draw of \(m\) input-output pairs \((X_i, Y_i), i = 1, \ldots, m\), the corresponding realization of \(\tilde{\delta}_m(x, y \mid u, v, P_m)\) can be computed by

\[
\tilde{\delta}_m(x, y \mid u, v, P_m) = \log \left[ \max_{i=1,\ldots,m} \left( \min_{j=1,\ldots,p} \left( \frac{x^i_j Y^i_k}{X^i_j Y^i_k} \right) \right) \right] \tag{4.37}
\]

provided the direction vectors \(u, v\) are strictly positive; otherwise, \(\tilde{\delta}_m(x, y \mid u, v, P_m)\) can be computed after applying the transformation in (2.20)–(2.23).

Reasoning along the lines of that used for the input, output, and hyperbolic orientations leads to the following:

**Definition 4.8.** For any \((x, y) \in P\) and for any integer \(m \geq 1\), the expected order-\(m\) directional distance corresponding to the point \((x, y)\) with directional vector \((u, v)\), known as the directional order-\(m\) efficiency score, is given by

\[
\delta_m(x, y \mid u, v, P) = E(\tilde{\delta}_m(x, y \mid u, v, P_m)) \tag{4.38}
\]

provided the expectation exists.

As with the hyperbolic measures, the directional measures of order-\(m\) are unconditional on particular values of inputs or outputs.

The directional order-\(m\) efficiency score benchmarks a unit operating at \((x, y)\) against the expected best performance among \(m\)
randomly chosen peers, along a linear path in the direction \((-\mathbf{u}, \mathbf{v})\) passing through the point \((x, y)\), subject to free-disposability of the production possibilities set \(\mathcal{P}\). This efficiency measure in turn defines a directional order-\(m\) efficient frontier.

**Definition 4.9.** The expected directional, efficient frontier of order \(m\), known as the directional order-\(m\) efficient frontier and denoted by \(\mathcal{P}_m^\partial (\mathbf{u}, \mathbf{v})\), is defined as

\[
\mathcal{P}_m^\partial (\mathbf{u}, \mathbf{v}) \equiv \left\{ (\bar{x}, \bar{y}) \mid \bar{x} = x - u \delta_m(x, y \mid \mathbf{u}, \mathbf{v}, \mathcal{P}), \bar{y} = y + v \delta_m(x, y \mid \mathbf{u}, \mathbf{v}, \mathcal{P}), (x, y) \in \mathcal{P} \right\}. \tag{4.39}
\]

Note that the directional order-\(m\) efficient frontier depends on the direction vectors \((\mathbf{u}, \mathbf{v})\); in principle, for a given finite value of \(m\), there are as many directional order-\(m\) efficient frontiers as there are pairs \((\mathbf{u}, \mathbf{v})\) of direction vectors.

**Definition 4.10.** For any \((x, y) \in \mathbb{R}^{p+q}_+\), the expected directional minimum input, maximum output combination of order \(m\), denoted by \((x_m^\partial(x, y \mid \mathbf{u}, \mathbf{v}), y_m^\partial(x, y \mid \mathbf{u}, \mathbf{v}))\), is given by

\[
x_m^\partial(x, y \mid \mathbf{u}, \mathbf{v}) \equiv x - u \delta_m(x, y \mid \mathbf{u}, \mathbf{v}, \mathcal{P}_m) \tag{4.40}
\]

and

\[
y_m^\partial(x, y \mid \mathbf{u}, \mathbf{v}) \equiv y + v \delta_m(x, y \mid \mathbf{u}, \mathbf{v}, \mathcal{P}_m). \tag{4.41}
\]

As in the output-oriented and hyperbolic cases, the limiting behavior of the expected directional minimum input, maximum output combination of order \(m\), as \(m \to \infty\), is similar to that established by Cazals et al. (2002) for the expected minimum input of order \(m\) (conditional on \(y\)) given in (4.12) and (4.13). In particular, provided \(\delta_m(x, y \mid \mathbf{u}, \mathbf{v}, \mathcal{P}_m)\) exists,

\[
\lim_{m \to \infty} \delta_m(x, y \mid \mathbf{u}, \mathbf{v}, \mathcal{P}_m) = \delta(x, y \mid \mathbf{u}, \mathbf{v}, \mathcal{P}) \tag{4.42}
\]
4.2 Definitions and Basic Ideas

and

\[ \lim_{m \to \infty} \left( x_m^\beta(x, y \mid u, v), y_m^\beta(x, y \mid u, v) \right) = \left( x^\beta(x, y \mid u, v), y^\beta(x, y \mid u, v) \right). \]  

(4.43)

See Simar and Vanhems (2012) for details and proofs; in particular, they establish that

\[ \delta_m(x, y \mid u, v, P) = - \int_0^1 i(\log w) \left[ 1 - H_{X^\dagger Y^\dagger}(w^{-1}x^\dagger, wy^\dagger) \right]^m dw. \]  

(4.44)

Example 4.4. Again suppose that \( p = q = 1 \), and let \( f(x, y) \) be given by (4.9) in Example 4.1. Then it is easy to show that the marginal densities are given by

\[ f_X(x) = \begin{cases} 
2x & \forall x \in [0, 1]; \\
0 & \text{otherwise} 
\end{cases} \]  

(4.45)

and

\[ f_Y(y) = \begin{cases} 
2 - 2y & \forall y \in [0, 1]; \\
0 & \text{otherwise.} 
\end{cases} \]  

(4.46)

Set the (scalar) direction vectors \( u = E(X) = 2/3 \) and \( v = E(Y) = 1/3 \), and use these to apply the transformation in (2.15) and (2.16) to construct random variables \((X^\dagger, Y^\dagger)\). Using (2.16) and (4.35), it is easy to show that

\[ F_{Y^\dagger}(y^\dagger) = \begin{cases} 
1 & \forall y^\dagger > e^{1/v}; \\
2v\log(y^\dagger) - (v\log(y^\dagger))^2 & \forall y^\dagger \in (1, e^{1/v}]; \\
0 & \forall y^\dagger \leq 1. 
\end{cases} \]  

(4.47)

Moreover, (2.15), (2.16) and (4.10) can be used to show that the distribution function for \( X^\dagger \), conditional on \( Y^\dagger \geq y^\dagger \), is given by

\[ F_{X^\dagger \mid Y^\dagger}(x^\dagger \mid Y^\dagger \geq y^\dagger) = \begin{cases} 
1 & \forall x^\dagger > e^{1/u}; \\
\left( u\log(x^\dagger) - v\log(y^\dagger) \right)^2 \left( 1 - v\log(y^\dagger) \right)^{-2} & \forall x^\dagger \in (y^\dagger e^{v/u}, e^{1/u}]; \\
0 & \forall x^\dagger \leq y^\dagger e^{v/u}. 
\end{cases} \]  

(4.48)
Then
\[ H_{X\mid Y^+}(x^+, y^+) = F_{X\mid Y^+}(x^+ \mid Y^+ \geq y^+) \left[ 1 - F_{Y^+}(y^+) \right] . \] (4.49)

Solving the integral in (4.44) using quadrature methods presents numerical difficulties. However, Monte-Carlo methods can be used to compute \( \delta_m(x, y \mid u, v, \mathcal{P}) \). Consider a set of fixed points \( x_1 = y_1 = 0.01, x_2 = y_2 = 0.02, \ldots, x_{99} = y_{99} = 0.99 \). Then the directional order-\( m \) efficient frontier corresponding the the direction vectors \((u, v)\) can be traced out by computing \( (x_i - u \delta_m(x, y \mid u, v, \mathcal{P}), y_i + v \delta_m(x, y \mid u, v, \mathcal{P})) \) for each \( i = 1, \ldots, 99 \). Figure 4.6 shows the results of this exercise, with \( m = 50 \), where the full frontier is depicted by a solid line and the directional order-\( m \) efficient frontier is shown by a dashed line.

As with the hyperbolic order-\( m \) frontier, the directional order-\( m \) frontier is seen in Figure 4.6 to lie roughly parallel with the full frontier. Comparing Figures 4.5 and 4.6, it is evident that the directional order-\( m \) frontier lies farther from the full frontier than the hyperbolic frontier. In addition to the additive nature of the directional concept,
as opposed to the multiplicative nature of the hyperbolic concept, the
directional order-$m$ frontier shown in Figure 4.6 corresponds to a spe-
cific combination of direction vectors.

4.3 Nonparametric Order-$m$ Estimators

As shown above for the full frontier approach, the probabilistic for-
mulation using the distribution function $H_{XY}(x, y)$ allows the plug-in
principle to be used to provide an appealing, alternative characteriza-
tion of FDH estimators. The same idea leads to an intuitive, nonpara-
metric estimator of $\theta_m(x, y)$ obtained by plugging the empirical version
of $F_X(x \mid y)$ into (4.11) to yield

$$
\hat{\theta}_{m,n}(x, y) = \hat{E}(\hat{\theta}_m(x, y) \mid Y \geq y)
= \int_0^\infty (1 - \hat{F}_{X,n}(ux \mid y))^m du,
= \hat{\theta}_{FDH}(x, y) + \int_0^\infty (1 - \hat{F}_{X,n}(ux \mid y))^m du. \tag{4.50}
$$

This estimator involves the computation of a univariate integral that
is easy to solve using numerical integration methods. Even for large
values of $p$, $q$ and $m$, the estimator is very fast to compute because the
integral remains one-dimensional. When $p = 1$, an explicit analytical
solution is available; see Cazals et al. (2002).

In the full multivariate setting, Daraio and Simar (2005) propose a
Monte-Carlo algorithm for approximating the expectation in (4.50),
thereby avoiding the need for numerical integration. However, for
large $m$, computing the integral in (4.50) is likely to be faster.

Analogous to (4.6), for a firm operating at $(x, y) \in \mathcal{P}$, an estimate
of its expected minimum input level of order-$m$ is

$$
\widehat{x}_m^\partial(y) = \hat{\theta}_{m,n}(x, y)x. \tag{4.51}
$$

Just as (4.6) can be evaluated over the range of possible output levels $y$
to trace out the expected minimum input frontier of order $m$, denoted
$\mathcal{P}_m^\text{din}$, (4.51) can also be evaluated for the possible values of $y$ to trace
out an estimate $\widehat{\mathcal{P}}_m^\text{din}$ of $\mathcal{P}_m^\text{din}$.
In the output orientation, a nonparametric estimator of \( \lambda_m(x, y) \) is given by
\[
\hat{\lambda}_m(x, y) = \int_0^\infty \left[ 1 - (1 - \hat{S}_{Y,n}(uy|x))^m \right] du = \hat{\lambda}_n(x, y) - \int_0^{\hat{\lambda}_n(x, y)} (1 - \hat{S}_{Y,n}(uy|x))^m du. \tag{4.52}
\]
As in the input-orientation, the integral in (4.52) is univariate and can be solved using numerical integration. For a firm operating at \((x, y) \in P\), an estimate of its expected maximum output level of order-\(m\) is
\[
\hat{y}_m^o(x) = \hat{\lambda}_{m,n}(x, y)y, \tag{4.53}
\]
analogous to (4.18).
Wilson (2011) gives an estimator of the expected hyperbolic order-\(m\) efficiency defined in (4.27) by substituting the empirical distribution function given in (2.54) into (4.27) to obtain
\[
\hat{\gamma}_{m,n}(x, y) = \hat{E}(\gamma_m(x, y)) = \int_0^\infty \left[ 1 - \hat{H}_{XY,n}(u^{-1}x, uy) \right]^m du \tag{4.54}
\]
(again, Wilson, 2011 works with the reciprocal of the expression used here). The integral here is also univariate, and can be computed using numerical methods. Alternatively, because of the lack of conditioning on either inputs or outputs, \(\hat{\gamma}_{m,n}(x, y)\) can also be computed very quickly using Monte-Carlo Methods along the lines of those described by Cazals et al. (2002) for computing \(\hat{\theta}_{m,n}(x, y)\).
Consider a Monte-Carlo sample \(S_{n}^{b,m} = \{(X_{bi}, Y_{bi})\}_{i=1}^{m}\) obtained by drawing \(m\) times, independently and with replacement, from input–output pairs in \(S_n\). Then compute
\[
\hat{\gamma}_{m,n,b}(x, y) = \max_{i=1, \ldots, m} \left( \min_{k=1, \ldots, q} \left( \frac{x_j}{X_{bi}} \frac{Y_{bi}^k}{y^k} \right) \right). \tag{4.55}
\]
Repeat this exercise \(B\) times, for \(b = 1 \ldots, B\). Then
\[
\hat{\gamma}_{m,n}(x, y) \approx B^{-1} \sum_{b=1}^{B} \hat{\gamma}_{m,n,b}(x, y), \tag{4.56}
\]
with the approximation improving as \(B\) increases.
Again using the plug-in principle, the expected minimum input, maximum output combination of order $m$ for any $(x, y) \in \mathbb{R}_+^{p+q}$ is estimated by

$$\hat{x}_m^\theta(x, y) = x \hat{\gamma}_{m,n,b}(x, y)$$ (4.57)

and

$$\hat{y}_m^\theta(x, y) = y \hat{\gamma}_{m,n,b}(x, y).$$ (4.58)

The directional order-$m$ efficiency measure can be estimated similarly, due to the link between directional and hyperbolic distances that has been discussed previously, using the transformation in (2.15) and (2.16). After applying this transformation to the observations in the sample $S_n$, draw $m$ times, independently and with replacement, from the $n$ transformed input–output pair observations. Then compute

$$\hat{\delta}_{m,n,b}(x, y \mid u, v) = \log \left[ \min_{i=1, \ldots, m} \left( \max_{j=1, \ldots, p} \left( \frac{X^j_{mi}}{x^j_{i}}, \frac{y^k_{mj}}{Y^k_{i}} \right) \right) \right].$$ (4.59)

Repeat this exercise $B$ times, for $b = 1 \ldots, B$. Then

$$\hat{\delta}_{m,n}(x, y \mid u, v) \approx B^{-1} \sum_{b=1}^{B} \hat{\delta}_{m,n,b}(x, y \mid u, v),$$ (4.60)

with the approximation improving as $B$ increases.

For any $(x, y) \in \mathbb{R}_+^{p+q}$, the expected directional minimum input, maximum output combination of order $m$ for any $(x, y) \in \mathbb{R}_+^{p+q}$ is estimated by

$$\hat{x}_m^\theta(x, y \mid u, v) \equiv x - u \hat{\delta}_{m,n,b}(x, y \mid u, v)$$ (4.61)

and

$$\hat{y}_m^\theta(x, y \mid u, v) \equiv y + v \hat{\delta}_{m,n,b}(x, y \mid u, v).$$ (4.62)

### 4.4 Statistical Properties

#### 4.4.1 Input Orientation

The nonparametric order-$m$ efficiency estimators described above have some interesting and useful properties. By construction, the estimator $\hat{\delta}_{m,n,b}^\text{min}$ with finite $m$ does not envelop all the observations in the
sample $S_n$. Consequently, the estimator is less sensitive than either the FDH or DEA frontier estimators to extreme points or outliers. Again, the estimator $\hat{\mathcal{P}}_{\text{in}}^m$ shares the same property as its population counterpart, $\mathcal{P}_{\text{in}}^m$; as $m$ increases, for a fixed sample size $n$, $\hat{\mathcal{P}}_{\text{in}}^m$ converges to the usual FDH estimator of $\mathcal{P}^\partial$, while $\mathcal{P}_{\text{in}}^m$ converges to $\mathcal{P}^\partial$; i.e.,

$$\hat{\theta}_{m,n}(x, y) \to \hat{\theta}_{\text{FDH}}(x, y), \text{ as } m \to \infty.$$ \hspace{1cm} (4.63)

The proof is trivial; see Cazals et al. (2002).

Cazals et al. (2002) also demonstrate that $\hat{\theta}_{m,n}(x, y)$ is strongly consistent; i.e.,

$$\hat{\theta}_{m,n}(x, y) \xrightarrow{a.s.} \theta_{m}(x, y)$$ \hspace{1cm} (4.64)

for any $(x, y) \in \mathcal{P}$ such that $\theta_{m}(x, y)$ exists and for any $m \geq 1$. As noted earlier, however, while consistency is perhaps the most fundamental property of an estimator, it is also, by itself, a very weak property. Perhaps the most intriguing property of the input-oriented order-$m$ efficiency estimator is that for any fixed value of $m$, $\hat{\theta}_{m,n}(x, y)$ converges to $\theta_{m}(x, y)$ at the rate $\sqrt{n}$, and its asymptotic distribution is normal; i.e.,

$$\sqrt{n} \left[ \hat{\theta}_{m,n}(x, y) - \theta_{m}(x, y) \right] \xrightarrow{d} N(0, \sigma^2_{\text{in}}(x, y)).$$ \hspace{1cm} (4.65)

The root-$n$ convergence rate attained by the nonparametric conditional order-$m$ input efficiency estimator is quite unusual among nonparametric estimators. Unlike FDH and DEA estimators, there is no curse of dimensionality here; i.e., the convergence rate of the order-$m$ estimator does not become slower as the dimensionality $(p + q)$ increases.

The variance $\sigma^2_{\text{in}}(x, y)$ in (4.65) depends on $F_X(x \mid y)$; see Cazals et al. (2002) for details. This variance can be estimated consistently using the plug-in principle, where $F_X(x \mid y)$ is replaced by its empirical analog $\hat{F}_{X,n}(x \mid y)$ in the expression for $\sigma^2_{\text{in}}(x, y)$ given by Cazals et al. (2002). Alternatively, $\sigma^2_{\text{in}}$ can be estimated using a naive bootstrap (due in part to the asymptotic normality), where pseudo-samples are obtained by drawing from the empirical distribution of the observations in $S_n$. The naive bootstrap can be used (provided $m$ is finite) since the boundaries of support for $f(x, y)$ are not estimated; i.e., both $\mathcal{P}_{\text{in}}^m$ and $\hat{\mathcal{P}}_{\text{in}}^m$ lie in the interior of $\mathcal{P}$, away from $\mathcal{P}^\partial$. Estimates of confidence intervals can also be obtained easily using a naive bootstrap approach.
Since $\theta_m(x, y) \to \theta(x, y \mid P)$ as $m \to \infty$, one might be tempted to use $\hat{\theta}_{m,n}(x, y)$ as an estimator of $\theta(x, y \mid P)$ itself, by using large values of $m$. Cazals et al. (2002) show that if $m = m(n) \to \infty$ as $n \to \infty$ (with $m$ increasing at rate $n \log n S_Y(y)$ as $n \to \infty$),

$$\frac{1}{n^{p+q}} \left[ \hat{\theta}_{m,n}(x, y) - \theta(x, y \mid P) \right] \xrightarrow{d} \text{Weibull}(\mu_{x,y}, p + q),$$

where the parameter of the Weibull is the same as in the case of the FDH estimator. Hence if $n \to \infty$ and $m = m(n) \to \infty$, $\hat{\theta}_{m,n}(x, y)$ shares the same properties as the FDH estimator. In finite samples, however, the order-$m$ estimator will be more robust with respect to outliers and extreme values since it does not envelop all the (extreme) observations in the sample.

This latter property gives another appealing feature of the order-$m$ estimators. In practice, the choice of a particular value of $m$ should be guided by this robustness property. Since the estimator involves little computational burden, it can be computed for several values of $m$ (e.g., $m = 25, 50, 100, 150, 200, \ldots$). For each value of $m$, one can observe the percentage of observations in the sample that lie outside the resulting order-$m$ frontier estimate (such observations are indicated by a value $\hat{\theta}_{m,n}(X_i, Y_i) > 1$). This percentage will decrease as $m$ increases; as described below, this property can also be used to detect outliers as explained below in Outlier Detection. The final value of $m$ can be chosen in terms of the desired level of robustness; for example, one might choose a value for $m$ that results in approximately 5 percent of the observations in $S_{m}$ lying outside $\hat{\Omega}_m$. Experience has shown that in many applications, qualitative conclusions are little affected by particular choices of $m$, provided the values of $m$ are somewhat less than the sample size, $n$.

### 4.4.2 Output Orientation

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

$$\tilde{\lambda}_m(x, y) = \int_0^{\infty} \left[ 1 - (1 - \tilde{S}_{Y, n}(w y \mid x))^m \right] du$$

$$= \tilde{\lambda}_n(x, y) - \int_0^{\tilde{\lambda}_n(x, y)} (1 - \tilde{S}_{Y, n}(w y \mid x))^m du. \quad (4.67)$$
This estimator shares properties analogous to those of the input-oriented estimator. In particular, the output-oriented estimator also achieves $\sqrt{n}$-consistency and asymptotic normality; i.e.,

$$\sqrt{n} \left[ \hat{\lambda}_{m,n}(x, y) - \lambda_m(x, y) \right] \xrightarrow{d} N(0, \sigma_{out}^2(x, y)).$$ (4.68)

In addition, as $m \to \infty$, $\lambda_m(x, y) \to \lambda(x, y | \mathcal{P})$ and $\hat{\lambda}_{m,n}(x, y) \to \hat{\lambda}_{FDH}(x, y)$. Analogous to (4.66), if $m = m(n) \to \infty$ at rate $n \log n F_X(x)$ as $n \to \infty$,

$$n^{1/p} \left[ \hat{\lambda}_{m,n}(x, y) - \lambda(x, y | \mathcal{P}) \right] \xrightarrow{d} \text{Weibull}(\mu_{x,y}, p + q) \text{ as } n \to \infty.$$ (4.69)

As $m = m(n) \to \infty$, the estimator $\hat{\lambda}_{m,n}(x, y)$ shares the same properties as the FDH estimator. But, in finite samples, it will be more robust to outliers and extreme values since it will not envelop all the observations in the sample.

**4.4.3 Hyperbolic and Directional Orientations**

Wilson (2011) establishes asymptotic properties for the hyperbolic order-$m$ efficiency estimator defined in (4.54). In particular, under mild assumptions, for any $(x, y)$ for which there exists $\gamma$ such that $(\gamma x, \gamma^{-1} y) \in \mathcal{P}$, and for any $m \geq 1$,

$$\tilde{\gamma}_{m,n}(x, y) \xrightarrow{a.s.} \gamma_m(x, y)$$ (4.70)

and

$$n^{1/2} (\tilde{\gamma}_{m,n}(x, y) - \gamma_m(x, y)) \xrightarrow{d} N(0, \sigma_{\gamma,m}^2(x, y)).$$ (4.71)

The limiting form of the hyperbolic order-$m$ estimator (as $m \to \infty$) is the hyperbolic FDH estimator; i.e., for fixed $n$,

$$\lim_{m \to \infty} \tilde{\gamma}_{m,n}(x, y) = \tilde{\gamma}_{FDH,n}(x, y).$$ (4.72)

As with the input- and output-oriented order-$m$ efficiency estimators, by letting $m \to \infty$ at an appropriate rate as $n \to \infty$, $\tilde{\gamma}_{m,n}(x, y)$
can be used to estimate $\gamma(x, y)$. This is apparent when one considers the result in (4.72) in conjunction with Corollary 4.1 in Wheelock and Wilson (2008), which demonstrates that $\widehat{\gamma}_{FDH,n}(x, y)$ converges to $\gamma(x, y)$ as $n \to \infty$. As remarked by Cazals et al., $m$ can be viewed as a “trimming” parameter. Wilson (2011) proves that, under mild assumptions, for $m = m(n)$ a sequence in $n$ such that $m(n) = O(\beta n \log(n) H_{XY}(x, y))$ where $\beta > \frac{1}{p+q}$,

$$
\frac{m}{n} \left( \widehat{\gamma}_{m(n),n}(x, y) - \gamma(x, y) \right) \xrightarrow{d} \text{Weibull} \left( \mu_{H,0}^{p+q}, p + q \right),
$$

where $\mu_{H,0}$ is a constant given in Wheelock and Wilson (2008, Equation A.7).

Given the relation between directional and hyperbolic distances, it should not be surprising that the directional order-$m$ efficiency estimator $\widehat{\delta}_{m}(x, y \mid u, v)$ in (4.60) possesses properties analogous to those of $\widehat{\gamma}_{m}(x, y)$. In particular, Simar and Vanhems (2012) establish strong consistency, asymptotic normality, and root-$n$ convergence of $\widehat{\delta}_{m,n}(x, y \mid u, v)$.

### 4.5 Empirical Examples

To illustrate the order-$m$ estimators, consider a simple example where $p = q = 1$ and the upper boundary of the production set $\mathcal{P}$ is described by

$$
y = g(x) = (2x - x^2)^{1/2}. \tag{4.74}
$$

Suppose also that the joint probability density of inputs and outputs, $f(x, y)$, is uniform over $\mathcal{P}$; hence

$$
f(x, y) = \begin{cases} 
4\pi^{-1} & \forall x \in [0, 1], \ y \in [0, g(x)]; \\
0 & \text{otherwise.}
\end{cases} \tag{4.75}
$$

The marginal distribution functions corresponding to the probability density function in (4.75) are given in Wilson (2011), and allow recovery of the true order-$m$ frontiers (input, output, hyperbolic, and directional orientations). These are shown by the dashed curves in the four panels of Figure 4.7 for $m = 75$; in the case of the directional orientation, the direction vectors used are $u = E(X) \approx 0.5756$. 

and $v = E(Y) \approx 0.4244$. The full frontier is represented by the smooth, solid curve in each panel.

Of course, the true order-$m$ frontiers cannot be observed in typical applied settings, and must be estimated from data. Each panel of Figure 4.7 shows a sample of $n = 300$ observations drawn independently from the density in (4.75); the same sample is used in each case. For each observation, efficiency estimates were computed using each of the estimators $\hat{\theta}_{m,n}(x,y)$, $\hat{\lambda}_{m,n}(x,y)$, $\hat{\gamma}_{m,n}(x,y)$, and $\hat{\delta}_{m,n}(x,y \mid u,v)$. The resulting estimates were then used to project each sample observation onto the estimate of the corresponding order-$m$ frontier (again,
with $m = 75$). The estimated order-$m$ frontiers appear as solid but jagged curves in each panel of Figure 4.7. In each case, the estimated order-$m$ frontiers are rough, jagged, and variable. As discussed below in *Unanswered Questions, Promising Ideas*, smoothing techniques can be applied to the estimated order-$m$ frontiers to reduce this variability.
5

Robust Order-α Estimators

5.1 Definition and Basic Ideas

5.1.1 Why Use Quantiles?

As noted above, the quantity $m$ in order-$m$ frontier estimation serves as a trimming parameter which determines the percentage of points that will lie above the order-$m$ frontier. The idea underlying order-α quantile frontiers is to reverse this causation, and choose the proportion of the data lying above the partial frontier directly.

Quantile estimation in regression contexts is an old idea. In the case of linear (or nonlinear) parametric regression where $y = g(x | \beta) + \epsilon$, instead of minimizing the sum of squared errors to estimate parameters, one can minimize the sum of absolute values of the errors to obtain a least-absolute deviation (LAD) estimator of $\beta$; the estimated regression equation is then interpreted as an estimate of the median, or 50-percent quantile function, conditional on the regressors in $x$. More generally, Koenker and Bassett (1978) proposed minimizing (with
5.1 Definition and Basic Ideas

To estimate the $\alpha \times 100$-percent quantile function (again, conditional on $x$). In the case of production relationships, the Koenker and Bassett (1978) approach presents two new problems: (i) the conditional quantile function must be specified parametrically, a priori, with the risk of mis-specification; and (ii) it is apparently not possible to constrain estimates of the conditional quantile function to be monotonic in multivariate settings, suggesting a loss of statistical efficiency at a minimum, and perhaps worse problems for interpretation of any estimates that are obtained.

In the framework of production frontiers, using the probabilistic formulation of the DGP that has been developed, it is straightforward to adapt the order-$m$ ideas to order-$\alpha$ quantile estimation. These ideas were developed for the univariate case in the input and output orientations by Aragon et al. (2005), and extended to the multivariate setting by Daouia and Simar (2007). Wheelock and Wilson (2008) extended the ideas to the hyperbolic orientation, and Simar and Vanhems (2012) extended the ideas to directional measures. Each of these is considered separately below.

5.1.2 Input Orientation

As discussed above, using order-$m$ partial frontiers, a unit operating at $(x, y)$ is benchmarked against the expected minimum input among $m$ peers drawn randomly from the population of units producing output levels of at least $y$. The idea underlying order-$\alpha$ quantiles is similar: instead of benchmarking firms’ performances against the full frontier $\mathcal{P}^0$, which may be difficult to estimate precisely, one can benchmark firms’ performances against a quantile lying “close” to the full frontier. In particular, input-oriented order-$\alpha$ quantile frontiers 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 output levels of at least $y$. 

\[
\sum_{y_i \geq g(x_i | \beta)} \alpha |y_i - g(x_i | \beta)| + \sum_{y_i < g(x_i | \beta)} (1 - \alpha) |y_i - g(x_i | \beta)| \quad (5.1)
\]
Definition 5.1. For all $y$ such that $S_Y(y) > 0$ and for $\alpha \in (0, 1]$, the $\alpha$-quantile input efficiency score for the unit operating at $(x, y) \in P$ is defined by

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

(5.2)

Recalling (2.44) and (2.48), it is clear that

$$
\theta(x, y \mid P) = \inf \{ \theta \mid F_{X \mid Y}(\theta x \mid y) > 0; \}
$$

(5.3)

the only difference between $\theta(x, y \mid P)$ and $\theta_\alpha(x, y \mid P)$ is that the zero in (5.3) has been replaced with $(1 - \alpha)$ in (5.2). This simple substitution is the basis underlying order-$\alpha$ efficiency scores in the output, hyperbolic, and directional orientations, as well as here in the input orientation.

The $\alpha$-quantile input efficiency score can be interpreted as follows. First, suppose that $\theta_\alpha(x, y) = 1$; then the unit is said to be input efficient at the level $\alpha \times 100\%$ since it is dominated by firms producing more than $y$ with a probability of $1 - \alpha$. Then more generally, if $\theta_\alpha(x, y)(<, >)1$, the firm at $(x, y)$ can (reduce, increase) its input usage to $\theta_\alpha(x, y)x$ to become input-efficient at the level $\alpha \times 100\%$. The quantity $\theta_\alpha(x, y)$ is called the “input efficiency at level $\alpha \times 100\%$.”

The idea of the order-$\alpha$ conditional quantile is illustrated in Figure 5.1 for the case of univariate $x$. The figure plots the conditional distribution function $F(x \mid Y \geq y_0)$, which equals zero at $x = \phi(y_0)$, defining the full input-frontier for a unit producing output level $y_0$. The $(1 - \alpha)$ quantile input-frontier is given by $x = \phi_\alpha(y_0)$. For a unit operating at levels $(x_0, y_0)$, its $\alpha$-quantile input efficiency is given by $\theta_\alpha(x_0, y_0) = \phi_\alpha(y_0)/x_0$, whereas its Farrell input efficiency score is given as usual by $\theta(x_0, y_0) = \phi(y_0)/x_0$.

The concept of efficiency of order-$\alpha$ allows definition of the corresponding efficient frontier at the level $\alpha \times 100\%$. This production frontier is called the $\alpha$-quantile efficient frontier. It may be defined as the efficient input at the level $\alpha \times 100\%$ for a given level of output $y$. 
5.1 Definition and Basic Ideas

and for an input mix given by an input vector $x$:

$$x^\alpha_\theta(y) = \theta_\alpha(x, y)x.$$  

(5.4)

By definition, a unit operating at the point $(x^\alpha_\theta(y), y) \in \mathcal{P}$ has a probability $H_{XY}(x^\alpha_\theta(y), y) = (1 - \alpha)S_Y(y) \leq 1 - \alpha$ of being dominated. Analogous to $\mathcal{P}^{\text{fin}}_m$, $x^\alpha_\theta(y)$ can be evaluated for all possible $y$ to trace out an order-$\alpha$ input frontier, denoted $\mathcal{P}^{\text{fin}}_\alpha$. This is illustrated below in Example 5.1.

It is clear from the definition that if $\alpha = 1$, the full frontier $\mathcal{P}^{\theta}$ is recovered; in this case, $\theta_1(x, y) \equiv \theta(x, y | \mathcal{P})$ is the Debreu–Farrell input measure of efficiency. In fact, it can be proved that the convergence of $\theta_\alpha(x, y)$ to $\theta(x, y | \mathcal{P})$ when $\alpha \to 1$ is monotone; i.e.,

$$\lim_{\alpha \to 1} \setminus_\alpha \theta_\alpha(x, y) = \theta(x, y | \mathcal{P}),$$  

(5.5)

where “$\setminus_\alpha$” denotes monotone convergence from above.

The order-$\alpha$ efficiency concept comes with an important existence property. It can be proved that if $F_{X|Y}$ is continuous and monotonically increasing in $x$ then, for all $(x, y)$ in the interior of $\mathcal{P}$, there exists an $\alpha \in (0, 1]$ such that $\theta_\alpha(x, y) = 1$, where $\alpha = 1 - F_{X|Y}(x | y)$. In other words, any point $(x, y)$ in the interior of $\mathcal{P}$ belongs to some $\alpha$-quantile input efficient frontier. Therefore, the value $\alpha = \alpha(x, y)$ such
Robust Order-α Estimators

that \((x, y) \in P^\partial_{\alpha}\) (i.e., the value of \(\alpha\) that defines an order-\(\alpha\) frontier containing \((x, y)\)) can be used as a new measure of input efficiency. Order-\(m\) frontiers do not have an analogous property, due to the discrete, integer-valued parameter \(m\).

**Example 5.1.** Again suppose that \(p = q = 1\), and let \(f(x, y)\) be given by (4.9) in Example 4.1. Using the expression for \(F_{X|Y}(x \mid Y \geq y)\) in (4.10), it is straightforward to approximate, for a given \((x, y)\), \(\theta_\alpha(x, y)\) in this example by setting

\[
(\theta_\alpha(x, y)x - y)^2(1 - y)^{-2} \approx (1 - \alpha)
\]

and then solving for \(\theta_\alpha(x, y)\). Doing so for a grid of points on \(P^\partial\), represented by the solid line in Figure 5.2, allows one to trace out the input order-\(\alpha\) frontier, represented by the dashed line in Figure 5.2 (for \(\alpha = 0.95\)). As with the input order-\(m\) frontier, the input order-\(\alpha\) frontier has slope steeper than the slope of the full frontier \(P^\partial\).

---

**Fig. 5.2** Input order-\(\alpha\) frontier (\(\alpha = 0.95\)).
5.1.3 Output Orientation

For the output oriented case, the order-\(\alpha\) efficiency measure can be defined as follows:

**Definition 5.2.** For all \(x\) such that \(F_X(x) > 0\) and for \(\alpha \in (0,1]\), the output \(\alpha\)-quantile efficiency score for the unit operating at \((x,y) \in \mathcal{P}\) is defined as

\[
\lambda_\alpha(x,y) = \sup \{ \lambda \mid S_{Y|X}(\lambda y \mid x) > 1 - \alpha \}. \tag{5.7}
\]

To illustrate this measure, suppose that \(\lambda_\alpha(x,y) = 1\). Then the unit operating at \((x,y) \in \mathcal{P}\) is said to be output efficient at the level \((\alpha \times 100)-\)percent, meaning that the unit is dominated with probability \((1 - \alpha)\) by firms using no more than \(x\) level of inputs. More generally, if \(\lambda_\alpha(x,y)(<,>)1\), the firm at \((x,y)\) can (decrease, increase) its output to \(\lambda_\alpha(x,y) y\) to become output-efficient at level \((\alpha \times 100)-\)percent, i.e., to be dominated by firms using weakly less input (than the level \(x\)) with probability \(1 - \alpha\).

The concept of order-\(\alpha\) output efficiency allows definition of the corresponding efficient frontier at the level \(\alpha \times 100\). For a given \((x,y)\), the order-\(\alpha\) output efficiency level of outputs is given by

\[
y_\alpha^\partial(x) = \theta_\alpha(x,y) y. \tag{5.8}
\]

By construction, a unit operating at the point \((x,y_\alpha^\partial(s)) \in \mathcal{P}\) has a probability \(H_{XY}(x,y_\alpha^\partial(x)) = (1 - \alpha)F_X(x) \leq 1 - \alpha\) of being dominated. Analogous to \(\mathcal{P}_m^{\partial_{\text{out}}}, y_\alpha^\partial(x)\) can be evaluated for all possible \(x\) to trace out an order-\(\alpha\) output frontier, denoted \(\mathcal{P}_\alpha^{\partial_{\text{out}}}\) as illustrated below in Example 5.2.

Analogous to the input-oriented order-\(\alpha\) efficiency measure, \(\lambda_\alpha(x,y)\) converges monotonically to the Debreu–Farrell output efficiency measure; i.e.,

\[
\lim_{\alpha \to 1} \nearrow \lambda_\alpha(x,y) = \lambda(x,y \mid \mathcal{P}), \tag{5.9}
\]

where \(\nearrow\) denotes monotonic convergence from below. Moreover, for all \((x,y) \in \mathcal{P}, (x,y) \notin \mathcal{P}_{\partial},\) there exists an \(\alpha \in (0,1]\) such that \(\lambda_\alpha(x,y) = 1\), where \(\alpha = 1 - S_{Y|X}(y \mid x)\).
Example 5.2. Again suppose that $p = q = 1$, and let $f(x, y)$ be given by (4.9) in Example 4.1. Using the expression for $S_{Y|X}(y \mid X \leq x)$ in Example 4.2, it is straightforward to approximate, for a given $(x, y)$, $\lambda_\alpha(x, y)$ in this example by setting

$$1 - 2x^{-1}\lambda_\alpha(x, y)y - x^{-2}\lambda_\alpha(x, y)^2y^2 = (1 - x^{-1}\lambda_\alpha(x, y))^2 \approx (1 - \alpha)$$

and then rearranging terms to obtain

$$\lambda_\alpha(x, y)y = x(1 - \sqrt{1 - \alpha}).$$

(5.11)

Evaluating this for a grid of points $(x, y)$ lying on $\mathcal{P}^\theta$, represented by the solid line in Figure 5.3, traces out the order-$\alpha$ output frontier, represented by the dashed line in Figure 5.3 for $\alpha = 0.95$. As with the order-$m$ output frontier, the order-$\alpha$ output frontier has slope less steep than the slope of the full frontier $\mathcal{P}^\theta$; in fact, from (5.11), the order-$\alpha$ output frontier has slope $(1 - \sqrt{1 - \alpha})$, while the slope of $\mathcal{P}^\theta$ is 1.

Fig. 5.3 Output order-$\alpha$ frontier ($\alpha = 0.95$).
5.1.4 Hyperbolic Orientation

In both the input and output orientations, the order-\( \alpha \) efficiency measures involve conditioning either on the firm’s existing level of outputs (in the input orientation), or on the firm’s existing level of inputs (in the output orientation). This results in order-\( \alpha \) frontiers that diverge from (or converge to) the full frontier as one moves from left to right in the production set, as illustrated in Examples 5.1 and 5.2 above. To avoid this feature of the input and output oriented measures, Wheelock and Wilson (2008) extended the ideas of Daouia and Simar (2007) to develop an hyperbolic order-\( \alpha \) efficiency measure given by the following.

**Definition 5.3.** For \( \alpha \in (0, 1] \), the unconditional, hyperbolic \( \alpha \)-quantile efficiency score for the unit operating at \((x, y) \in \mathcal{P}\) is defined by

\[
\gamma_{\alpha}(x, y) = \sup \{ \gamma > 0 \mid H_{XY}(\gamma^{-1}x, \gamma y) > (1 - \alpha) \}. \tag{5.12}
\]

This measure has a simple interpretation. Suppose \( \gamma_{\alpha}(x, y) = 1 \); then the unit at \((x, y)\) is said to be hyperbolic efficient at the level \((\alpha \times 100)\)-percent, and has probability \((1 - \alpha)\) of being dominated by any randomly chosen firm. Alternatively, if \( \gamma_{\alpha}(x, y)(<, >)1 \), then the firm operating at \((x, y) \in \mathcal{P}\) can (increase, decrease) its input usage to \( \gamma_{\alpha}(x, y)^{-1}x \) while simultaneously (decreasing, increasing) its output production to \( \gamma_{\alpha}(x, y)y \) to become hyperbolic-efficient at the level \((\alpha \times 100)\)-percent. The quantity \( \gamma_{\alpha}(x, y) \) is called the “hyperbolic efficiency at level \((\alpha \times 100)\)-percent.”

As with the input and output oriented order-\( \alpha \) measures, the full frontier is recovered when \( \alpha = 1 \); in particular,

\[
\lim_{\alpha \to 1} \gamma_{\alpha}(x, y) = \gamma(x, y \mid \mathcal{P}). \tag{5.13}
\]

The convergence is monotonic.

The hyperbolic \( \alpha \)-quantile frontier is defined by

\[
\mathcal{P}_{\alpha}^{\theta} = \{ (\gamma_{\alpha}(x, y)^{-1}x, \gamma_{\alpha}(x, y)y) \mid (x, y) \in \mathcal{P} \}. \tag{5.14}
\]

Using Assumption 2.2 and the fact that \( H_{XY}(x_0, y_0) \) is monotone, non-decreasing in \( x_0 \) and monotone, nonincreasing in \( y_0 \), it is easy to show
that $\mathcal{P}_\alpha^0$ is monotone in the sense that if $(x_0, y_0) \in \mathcal{P}_\alpha^0$, $(\tilde{x}, \tilde{y}) \in \mathcal{P}_\alpha^0$, and $\tilde{x} \geq x_0$, then $\tilde{y} \geq y_0$.

**Example 5.3.** Again suppose that $p = q = 1$, and let $f(x, y)$ be given by (4.9) in Example 4.1. Using the expression for $H_{XY}(x, y)$ appearing in Example 4.3, it is straightforward to approximate, for a given $(x, y)$, $\gamma_{\alpha}(x, y)$ by setting

$$
(\gamma_{\alpha}(x, y)^{-1}x - \gamma_{\alpha}(x, y)y)^2 \approx (1 - \alpha)
$$

and then solving for $\gamma_{\alpha}(x, y)$. Doing so for a grid of points on $\mathcal{P}^0$, represented by the solid line in Figure 5.3, traces out the hyperbolic order-$\alpha$ frontier $\mathcal{P}_\alpha^0$, represented by the dashed line in Figure 5.4 for $\alpha = 0.95$. Note that the hyperbolic order-alpha frontier parallel to the full frontier $\mathcal{P}^0$, due to the fact that $f(x, y)$ is uniform. By construction, Any point along $\mathcal{P}_\alpha^0$ shown in Figure 5.3 has probability $(1 - \alpha)$ of being weakly dominated by a randomly chosen firm drawn from $f(x, y)$.

---

![Fig. 5.4 Hyperbolic order-\(\alpha\) frontier (\(\alpha = 0.95\)).](image)
5.2 Nonparametric Order-\(\alpha\) Estimators

5.1.5 Directional Orientation


**Definition 5.4.** For all \((x, y) \in \mathcal{P}\) and \(\alpha \in (0, 1]\), the directional order-\(\alpha\) efficiency of \((x, y)\) with direction vector \((u, v)\) is given by

\[
\delta_{\alpha}(x, y \mid u, v) = \sup \{\delta \mid H_{XY}(x - \delta u, y + \delta v) > 1 - \alpha\}. \tag{5.16}
\]

Not surprisingly, this measure has properties similar to those of the hyperbolic order-\(\alpha\) efficiency measure. In particular, \(\delta_{\alpha}(x, y \mid u, v)\) converges monotonically to the directional measure \(\delta(x, y \mid u, v)\); i.e.,

\[
\lim_{\alpha \to 1} \delta_{\alpha}(x, y \mid u, v) = \delta(x, y \mid u, v). \tag{5.17}
\]

Using the transformation in (2.15) and (2.16), the directional order-\(\alpha\) efficiency measure can be written as

\[
\delta_{\alpha}(x, y \mid u, v) = \log(\gamma_{\alpha}(x^\dagger, y^\dagger)), \tag{5.18}
\]

where

\[
\gamma_{\alpha}(x^\dagger, y^\dagger) = \sup \{\gamma > 0 \mid H_{X^\dagger Y^\dagger}(\gamma^{-1}x^\dagger, \gamma y^\dagger) > 1 - \alpha\}. \tag{5.19}
\]

In addition, for a given value of \(\alpha \in (0, 1]\),

\[
\{(x, y) \mid \delta_{\alpha}(x, y \mid u, v) = 0, \ u \in \mathbb{R}^p_+, \ v \in \mathbb{R}^q_+\} = \mathcal{P}^0_{\alpha}; \tag{5.20}
\]

i.e., the partial frontier corresponding to the directional order-\(\alpha\) partial frontier is identical to the hyperbolic order-\(\alpha\) frontier for the same value of \(\alpha\). This is due to the fact that neither the hyperbolic nor the directional order-\(\alpha\) efficiency scores involve conditioning on either input or output levels.

5.2 Nonparametric Order-\(\alpha\) Estimators

Here again the plug-in principle can be used to obtain an intuitive, nonparametric order-\(\alpha\) efficiency estimators. In the input orientation,
an estimator of $\theta_\alpha(x, y)$ is obtained by replacing $F_{X|Y}(\cdot | \cdot)$ in (5.2) with its empirical counterpart, yielding

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

(5.21)

Computation of this estimator is very fast. An explicit analytical formula is available for the case $p = 1$ (see Aragon et al., 2005), and the following simple algorithm for the multivariate case is given in Daouia and Simar (2007).

Let $n_y = nS_{Y,n}(y) > 0$, which gives the number of observations in $S_n$ with $Y_i \geq y$, and define

$$X_i = \max_{k=1, \ldots, p} \frac{X_i^k}{x^k},$$  

(5.22)

$i = 1, \ldots, n$. For $j = 1, \ldots, n_y$, let $X_{(j)}^y$ denote the $j$th order statistic of the observations $X_i$ such that $Y_i \geq y$: $X_{(1)}^y \leq X_{(2)}^y \leq \cdots \leq X_{(n_y)}^y$. Then

$$\hat{F}_{X|Y,n}(\theta x \mid y) = \frac{\sum_{i : y_i \geq y} \mathbb{1}(X_i \leq \theta)}{n_y} = \frac{\sum_{j=1}^{n_y} \mathbb{1}(X_{(j)}^y \leq \theta)}{n_y}$$

$$= \begin{cases} 
0 & \text{if } \theta < X_{(1)}^y; \\
\frac{j}{n_y} & \text{if } X_{(j)}^y \leq \theta < X_{(j+1)}^y, j = 1, \ldots, n_y - 1; \\
1 & \text{if } \theta \geq X_{(n_y)}^y. 
\end{cases}$$

(5.23)

It follows that

$$\hat{\theta}_{\alpha,n}(x, y) = \begin{cases} 
X_{(\lceil(1-\alpha)n_y\rceil)}^y & \text{if } (1 - \alpha)n_y \in \mathbb{N}_+ \\
X_{(\lceil(1-\alpha)n_y\rceil+1)}^y & \text{otherwise}, 
\end{cases}$$

(5.24)

where $\mathbb{N}_+$ denotes the set of all nonnegative integers.

As with the estimators of the various order-$m$ efficiencies, this estimator does not envelop all the observed data in $S_n$ when $\alpha < 1$. Hence, it is less sensitive to extreme points and to outliers than traditional FDH and DEA estimators. This estimator also shares the same property as the corresponding population concept i.e., for a fixed sample size $n$, as $\alpha \to 1$, the order-$\alpha$ estimator converges to the FDH estimator.
of the full frontier; in particular,
\[
\lim_{\alpha \to 1} \sqrt[n]{\hat{\theta}_{\alpha,n}(x,y)} = \hat{\theta}_{FDH}(x,y).
\] (5.25)

A nonparametric estimator of \(\lambda_{\alpha}(x,y)\) is obtained by substituting
the empirical counterpart of \(S_{Y|X,n}(\cdot | \cdot)\) defined previously into (5.7)
to obtain
\[
\hat{\lambda}_{\alpha,n}(x,y) = \sup\{\lambda \mid \hat{S}_{Y|X,n}(\lambda y | x) > 1 - \alpha\}.
\] (5.26)

This estimator is also easy to compute using the algorithm given by
Daouia and Simar (2007).

Define
\[
Y_i = \min_{k=1, \ldots, q} Y_{kX_i},
\] (5.27)
i = 1, \ldots, n, and let \(n_x = nF_X(x) > 0\). For \(j = 1, \ldots, n_x\), let \(\gamma_{(j)}^{x}\)
denote the \(j\)th order statistic of the observations \(Y_i\) such that \(X_i \leq x\),
so that \(\gamma_{(1)}^{x} \leq \gamma_{(2)}^{x} \leq \ldots \leq \gamma_{(n_x)}^{x}\). Reasoning as in (5.23) and (5.24)
leads to
\[
\hat{\lambda}_{\alpha,n} = \begin{cases} \gamma_{(\alpha n_x)}^{x} & \text{if } \alpha n_x \in \mathbb{N}_+; \\ \gamma_{([\alpha n_x]+1)}^{x} & \text{otherwise} \end{cases}
\] (5.28)

where \([\alpha n_x]\) denotes the integer part of \(\alpha n_x\).

Estimation of \(\gamma_{\alpha}(x,y)\) is also straightforward. Using the now-
familiar plug-in principle, an estimator of \(\gamma_{\alpha}(x,y)\) is obtained by replacing
\(H_{XX}(\cdot,\cdot)\) in (5.12) with \(\hat{H}_{XY,n}(\cdot,\cdot)\) to obtain
\[
\hat{\gamma}_{\alpha,n}(x,y) = \sup\{\gamma > 0 \mid \hat{H}_{XY,n}(\gamma^{-1}x,\gamma y) > (1 - \alpha)\}.
\] (5.29)

Wheelock and Wilson (2008) considered two methods for computing
\(\hat{\gamma}_{\alpha,n}(x,y)\). The first method works along the lines of the methods used
by Daouia and Simar (2007) to compute \(\theta_{\alpha,n}(x,y)\) and \(\lambda_{\alpha,n}(x,y)\).

Given a sample \(S_n\) and the fixed point of interest \((x,y)\), for each
\(i = 1, \ldots, n\), define
\[
\chi_i = \min_{j=1, \ldots, p} \frac{x_{ij}}{X_{ij}}
\] (5.30)
Robust Order-α Estimators

ψ_i = \min_{j=1, \ldots, q} \frac{Y_{i,j}}{y^j}. \tag{5.31}

Let \( \mathcal{A} = \{i \mid y \psi_i \leq Y_i\} \), \( \mathcal{B} = \{i \mid x \chi_i^{-1} \geq X_i\} \setminus \mathcal{A} \), and \( \mathcal{C} = \{\omega_j\} = \{\chi_i \mid i \in \mathcal{A}\} \cup \{\psi_i \mid i \in \mathcal{B}\} \). Note that removing elements of \( \mathcal{A} \) in the definition of \( \mathcal{B} \) is required in the event of a tie, which can occur when the path \((\gamma^{-1}x, \gamma y)\), \( \gamma > 0 \), passes through one of the observations in \( S_n \). Then \#\( \mathcal{C} = n \). A nonparametric estimator of the hyperbolic α-quantile distance function \( \gamma_\alpha(x, y) \) defined in (5.12) is given by

\[
\hat{\gamma}_{\alpha,n}(x, y) = \begin{cases} 
\omega(\alpha n) & \text{if } \alpha n \in \mathbb{N}_{++}, \\
\omega(\alpha n + 1) & \text{otherwise};
\end{cases}
\tag{5.32}
\]

where \( \mathbb{N}_{++} \) denotes the set of strictly positive integers and \( \omega(j) \) denotes the \( j \)th largest element of the set \( \mathcal{C} \), i.e., \( \omega(1) \leq \omega(2) \leq \ldots \leq \omega(n) \).

The second method considered by Wheelock and Wilson (2008) exploits the fact that the optimization in (5.29) is a univariate problem. Given the point of interest \((x, y)\), it is easy to find initial values \( \gamma_a, \gamma_b \) that bracket the solution so that \( \tilde{H}_{XY,n}(\gamma^{-1}_a x, \gamma_a y) > (1 - \alpha) \) and \( \tilde{H}_{XY,n}(\gamma^{-1}_b x, \gamma_b y) < (1 - \alpha) \), and then solve for \( \hat{\gamma}_{\alpha,n}(x, y) \) using the bisection method. This method can be made accurate to an arbitrarily small degree; details are given in Wheelock and Wilson (2008).

Wheelock and Wilson noted that their second algorithm, based on the bisection method, was faster by a factor of about 70 than their first method. The second method is implemented in the hquan command in versions 1.11 (and later) of Wilson’s 2008 FEAR package.

Finally, the directional order-α efficiency score can be computed by transforming the data using (2.15) and (2.16), applying the hyperbolic order-α estimator to the transformed data, and then taking the log of this estimator.

5.3 Statistical Properties

5.3.1 Input Orientation

The statistical properties of the order-α quantile estimators are very similar to those of the order-\( m \) estimators, but with a few important
differences. In particular,
\[ \tilde{\theta}_{\alpha,n}(x, y) \xrightarrow{c} \theta_{\alpha}(x, y), \tag{5.33} \]
where \( \xrightarrow{c} \) denotes complete convergence, which implies (but is stronger than) almost-sure convergence.\(^1\) The complete convergence in (5.33) implies
\[ \lim_{n \to \infty} \sum_{j=1}^{n} \Pr \left( |\tilde{\theta}_{\alpha,j}(x, y) - \theta_{\alpha}(x, y)| \geq \epsilon \right) < \infty \tag{5.34} \]
for all \( \epsilon > 0 \). Moreover,
\[ \sqrt{n} \left[ \tilde{\theta}_{\alpha,n}(x, y) - \theta_{\alpha}(x, y) \right] \xrightarrow{d} N(0, \sigma_{\alpha}^2(x, y)). \tag{5.35} \]

An expression is for the variance term \( \sigma_{\alpha}^2(x, y) \) is given by Daouia and Simar (2007). As with the order-\( m \) efficiency estimators, the non-parametric order-\( \alpha \) input efficiency estimator is root-\( n \) consistent with an asymptotic normal distribution.

In addition, analogous to the order-\( m \) estimators, when \( \alpha = \alpha(n) \to 1 \) at an appropriate rate as \( n \to \infty \) (see Daouia and Simar, 2007),
\[ n^{-1/q} \left[ \tilde{\theta}_{\alpha(n),n}(x, y) - \theta(x, y | P) \right] \xrightarrow{d} \text{Weibull}(\mu_{x,y}, p + q). \tag{5.36} \]
Hence \( \tilde{\theta}_{\alpha(n),n}(x, y) \) shares the same asymptotic property as the FDH estimator but, since the corresponding frontier does not envelop all the observed data points, \( \tilde{\theta}_{\alpha(n),n}(x, y) \) is more robust to outliers in finite sample than the FDH estimator. Note that as \( \alpha(n) \) tends to 1, the root-\( n \) convergence rate is lost; lunch is not free, and estimating the full frontier comes with a cost.

5.3.2 Output Orientation

Similar results obtain for the order-\( \alpha \) output efficiency estimator. Provided \( \alpha < 1 \), the estimator converges completely, i.e.,
\[ \hat{\lambda}_{\alpha,n}(x, y) \xrightarrow{c} \lambda_{\alpha}(x, y), \tag{5.37} \]
\(^1\)Complete convergence implies, and is a stronger form of convergence than almost-sure convergence. A sequence of random variables \( \{\zeta_n\}_{n=1}^{\infty} \) converges completely to a random variable \( \zeta \), denoted by \( \zeta_n \xrightarrow{c} \zeta \), if \( \lim_{n \to \infty} \sum_{j=1}^{\infty} \Pr(|\zeta_j - \zeta| \geq \epsilon) < \infty \forall \epsilon > 0 \). This type of convergence was introduced by Hsu and Robbins (1947).
and is asymptotically normally distributed, with root-$n$ convergence rate:

$$\sqrt{n}(\hat{\lambda}_{\alpha,n}(x, y) - \lambda_{\alpha}(x, y)) \xrightarrow{d} N(0, \sigma_{\text{out}}^2(x, y)),$$  \hspace{1cm} (5.38)

As with the order-$m$ input efficiency estimator, the order-$m$ output efficiency estimator converges to the FDH estimator as $\alpha \to 1$:

$$\lim_{\alpha \to 1} \sqrt{n}(\hat{\lambda}_{\alpha,n}(x, y) - \lambda_{\alpha}(x, y)) = \hat{\lambda}_{\text{FDH}}(x, y).$$  \hspace{1cm} (5.39)

Finally, $\hat{\lambda}_{\alpha,n}(x, y)$ provides a robust estimator of the full-frontier efficiency scores: as $\alpha = \alpha(n) \to 1$ at an appropriate rate given by Daouia and Simar (2007) as $n \to \infty$,

$$\frac{1}{n^{p+q}} \left[ \hat{\lambda}_{\alpha(n),n}(x, y) - \lambda(x, y | P) \right] \xrightarrow{d} \text{Weibull}(\mu_{x,y}, p + q).$$  \hspace{1cm} (5.40)

### 5.3.3 Hyperbolic Orientation

Wheelock and Wilson (2008) obtain similar results for the hyperbolic order-$m$ efficiency estimator. In particular, for $\alpha < 1$,

$$\hat{\gamma}_{\alpha,n}(x, y) \xrightarrow{c} \gamma_{\alpha}(x, y).$$  \hspace{1cm} (5.41)

In addition, under mild assumptions,

$$\sqrt{n}(\hat{\gamma}_{\alpha,n}(x, y) - \gamma_{\alpha}(x, y)) \xrightarrow{d} N(0, \sigma_{\alpha}^2(x, y)).$$  \hspace{1cm} (5.42)

An expression for the variance $\sigma_{\alpha}(x, y)^2$ is given in Wheelock and Wilson (2008). This estimator also achieves the root-$n$ convergence rate.

As with the input and output order-$m$ efficiency estimators, the hyperbolic order-$m$ efficiency estimator converges to the FDH estimator as $\alpha \to 1$:

$$\lim_{\alpha \to 1} \sqrt{n}(\hat{\gamma}_{\alpha,n}(x, y) - \gamma_{\alpha}(x, y)) = \hat{\gamma}_{\text{FDH}}(x, y).$$  \hspace{1cm} (5.43)

Furthermore, Wheelock and Wilson (2008) prove, under mild assumptions, for $\alpha = \alpha(n) \to 1$ with the order of $\alpha(n) > 0$ such that $n^{(p+q+1)/(p+q)}(1 - \alpha(n)) \to 0$ as $n \to \infty$, for any $(x, y) \in P$,

$$n^{1/(p+q)} \left( \gamma(x, y) - \hat{\gamma}_{\alpha(n),n}(x, y) \right) \xrightarrow{d} \text{Weibull}(\mu_{H,0}^{p+q}, p + q),$$  \hspace{1cm} (5.44)

where $\mu_{H,0}$ is a constant for which an expression is given in the appendix of Wheelock and Wilson (2008). The required rate for $\alpha(n)$ is the same rate required by Daouia and Simar (2007) for the similar results regarding the order-$\alpha$ input and output efficiency estimators.
5.3.4 Directional Orientation

Simar and Vanhems (2012) establish properties for the directional order-$\alpha$ efficiency estimator similar to properties of the other order-$m$ estimators considered above. For $\alpha < 1$,

$$\hat{\delta}_{\alpha,n}(x, y \mid u, v) \xrightarrow{c} \delta_{\alpha}(x, y \mid u, v).$$  \hfill (5.45)

In addition, under mild assumptions,

$$\sqrt{n} \left( \hat{\delta}_{\alpha,n}(x, y \mid u, v) - \delta_{\alpha}(x, y \mid u, v) \right) \xrightarrow{d} N \left( 0, \sigma_{\alpha}^2(x, y \mid u, v) \right);$$  \hfill (5.46)

an expression for the variance $\sigma_{\alpha}(x, y)^2$ appears in Simar and Vanhems (2012). This estimator also achieves the root-$n$ convergence rate.

As with the input, output, and hyperbolic order-$\alpha$ efficiency estimators, the directional order-$\alpha$ efficiency estimator converges to the corresponding FDH estimator as $\alpha \to 1$:

$$\lim_{\alpha \to 1} \hat{\delta}_{\alpha,n}(x, y \mid u, v) = \hat{\delta}_{FDH}(x, y \mid u, v, \mathcal{P}).$$  \hfill (5.47)

Furthermore, Simar and Vanhems (2012) prove, under mild assumptions, for $\alpha = \alpha(n) \to 1$ with the order of $\alpha(n) > 0$ such that $n^{(p+q+1)/(p+q)}(1 - \alpha(n)) \to 0$ as $n \to \infty$, for any $(x, y) \in \mathcal{P},$

$$n^{1/(p+q)} \left( \delta(x, y \mid u, v) - \hat{\delta}_{\alpha(n),n}(x, y \mid u, v) \right) \xrightarrow{d} \text{Weibull} \left( \mu_{\mathcal{H},0}^{p+q}, p + q \right),$$  \hfill (5.48)

where $\mu_{\mathcal{H},0}^{p+q}$ is the constant which also appears in (5.44); see the appendix of Wheelock and Wilson (2008) for an explicit expression. The required rate for $\alpha(n)$ is the same rate required by Daouia and Simar (2007) for the similar results regarding the order-$\alpha$ input and output efficiency estimators.

5.3.5 Practical Implications of Order-$\alpha$

and Order-$m$ Estimators

In principle, the asymptotic normality results obtained for both the order-$\alpha$ as well as the order-$m$ efficiency estimators when $\alpha < 1$ and
$m < 1$ could be used for statistical inference about order-$\alpha$ or order-$m$ efficiencies. However, this would require estimating the corresponding variance terms, and these depend on derivative of the distribution function $H_{XY}(\cdot, \cdot)$; estimation of these derivatives is difficult. Alternatively, it is easy to implement a bootstrap procedure for purposes of testing and inference. Since partial frontiers (rather than the boundary of the production set) are estimated, a naive bootstrap based on resampling from the empirical distribution of the inputs and outputs could be implemented for inference-making purposes. However, although valid asymptotically, the naive bootstrap may result in poor coverages because of the discrete nature of data in small samples. Wheelock and Wilson (2008) used a smooth bootstrap procedure to overcome this problem in their application examining U.S. Federal Reserve check-processing operations.

Results obtained for both the order-$\alpha$ as well as the order-$m$ efficiency estimators permit two interpretations of any of these estimators — either as estimators of distance to a partial frontier (when $\alpha$ or $m$ is fixed), or as estimators of distance to the full frontier (when $\alpha$ or $m$ tend to one at appropriate rates). In the latter case, the root-$n$ convergence and asymptotic normality properties are lost, and the curse of dimensionality re-appears. There are no free lunches here. However, compared to the FDH efficiency estimators, order-$\alpha$ and order-$m$ efficiency estimators will necessarily be far less sensitive to outliers.

5.4 Empirical Examples

To illustrate the order-$\alpha$ estimators, consider the simple scenario given at the end of Robust Order-$m$ Estimators, where $p = q = 1$ and the upper boundary of the production set $P$ is given by (4.74) and the joint probability density of inputs and outputs is uniform over $P$. As noted earlier, the marginal and conditional distribution functions corresponding to this density are given in Wilson (2011); these allow derivation of the true order-$\alpha$ frontiers, which are depicted by the dashed curves in Figure 5.5 for $\alpha = 0.98$. The four panels of Figure 5.5 correspond to the input and output order-$\alpha$ frontiers, and the hyperbolic and directional order-$\alpha$ frontiers, respectively. The direction vectors
used in the directional orientation are again $u = E(X) \approx 0.5756$ and $v = E(Y) \approx 0.4244$. The full frontier is represented by the smooth, solid curve in each panel.

As in the order-$m$ case, the true order-$\alpha$ frontiers cannot be observed in typical applied settings, and must be estimated from data. Each panel of Figure 5.5 shows a sample of $n = 300$ observations drawn independently from the density in (4.75); the same sample is used in each case, and this sample is identical to the sample in the example given at the end of Robust Order-$m$ Estimators. For each observation, efficiency estimates were computed using each of the estimators $\hat{\theta}_{\alpha,n}(x, y)$,
\( \hat{\lambda}_{a,n}(x, y), \hat{\gamma}_{a,n}(x, y), \text{ and } \hat{\delta}_{a,n}(x, y \mid u, v) \). The resulting estimates were then used to project each sample observation onto the estimate of the corresponding order-\( \alpha \) frontier (again, with \( \alpha = 0.98 \)). The estimated order-\( \alpha \) frontiers appear as solid but jagged curves in each panel of Figure 4.7. In each case, the estimated order-\( \alpha \) frontiers are rough, jagged, and variable. As discussed below in Unanswered Questions, Promising Ideas, smoothing techniques can be applied to the estimated order-\( \alpha \) frontiers to reduce this variability.
Full-envelopment estimators such as FDH and DEA frontier estimators are appealing because they rely on only a few mild assumptions. FDH and DEA estimators have been in use for many years, and continue to be widely used. Moreover, they estimate the full frontier $\mathcal{P}^0$, rather than the partial frontiers described above. The cost of this, however, is that FDH and DEA estimators are highly sensitive to outliers.

Outliers are unusual or atypical observations. Some outliers may result from recording or measurement errors and should be corrected (if possible) or deleted from the data. Outliers might also appear in a given sample if, inadvertently, observations are drawn from different DGPs. On the other hand, when data are viewed as having come from a probability distribution, then it is quite possible to observe points with low probability, i.e., from the tail of a distribution. One would not expect to observe many such points given their low probability, and hence they appear as outliers. Cook and Weisberg (1982) observe that outliers of this type may lead to the recognition of important phenomena that might otherwise have gone unnoticed; in some cases such outliers may be the most interesting part of the data.
Since FDH and DEA estimators continue to be used, and sense they are sensitive to outliers, outlier detection is of vital importance. Of course, in simple, bivariate scenarios where \( p = q = 1 \), one could merely look at a scatter plot to find any outliers that might exist. Unfortunately, in higher dimensions, visualization is more problematic (Porembski et al., 2005 proposed a method for visualizing high-dimensional data, but the method has not been widely used to date, perhaps in part because of its complexity), and detecting outliers is much more difficult. Although there is a large literature in statistics and econometrics on outlier detection, most methods that have been proposed were designed for the problem of conditional mean estimation and do not take aspects of frontier models into account; see Wilson (1993) for discussion. In the frontier setting, the most worrisome outliers are those that have undue influence on the estimated efficiencies of other observations. Many of the standard methods for outlier detection are also very computer intensive in multivariate settings.

A relatively small number of outlier-detection methods have been proposed specifically for the context of frontier estimation. It is unlikely that a single method will be able to find all outliers in all instances when the number of dimensions is large. Consequently, many have argued that several methods should be employed on a given data set. All outlier-detection methods should be viewed as diagnostic tools; the goal is to first identify a (hopefully) small set of observations for further scrutiny. See Wilson (1993), Olesen et al. (1996), Simar (2003), and Porembski et al. (2005) for further discussion.

Wilson (1995) proposed a procedure similar to jackknife estimation, wherein observations are deleted one-at-a-time, with efficiency for the remaining \((n - 1)\) observations re-estimated with each deletion to gauge the influence of each observation on the estimated efficiencies of other observations in the sample. The process is then repeated, deleting pairs of observations, then repeated again while deleting triplets, etc. in order to uncover masking effects. Masking occurs when two (or more) outliers lie close to each other; removing only one of these will not reveal undue influence, which can only be discovered if the entire group of near-by outliers is removed. The method is effective, but the number of linear program solutions that must be obtained to implement
the Wilson (1995) procedure if pairs, triplets, etc., are deleted grows combinatorially with sample size. The procedure eventually becomes very costly in terms of required computational time as the sample size becomes large, and consequently is not practical for use with samples larger than a few hundred or perhaps a few thousand observations. Nonetheless, few applied studies have used more than a few hundred observations, and the increasing availability of high-performance computing systems has in recent years mitigated the concerns over computational expense.

Wilson (1993) proposed another method employing an influence function based on the geometric volume spanned by the sample observations, and the sensitivity of this volume with respect to deletions of singletons, pairs, triplets, etc. from the sample. Some notation is required to explain the method. For a value $\xi$ computed from a set of observations indexed by $S = \{1, \ldots, n\}$, let $D_L^\ell(\xi)$ denote the value computed similarly from observations with indices $S - L$, where $L \subset S$ and $L$ contains $\ell$ elements, $\ell < n$. Let $\bar{D}_L^\ell(\xi)$ denote the value computed similarly from observations only in $L$; e.g., $D_L^\ell(\xi) = \bar{D}_L^\ell(S - L)$.

For any matrix $A$ of full column rank, let $D_L^\ell(A)$ denote the determinant of the matrix $D_L^\ell(A)$. Let $X$ be an $(p \times n)$ matrix of observations on $n$ input vectors, and $Y$ be an $(q \times n)$ matrix of observations on the corresponding output vectors. Then add a row of ones to $X$ by defining the $(p + 1 \times n)$ partitioned matrix $Z = [1 \ X']$. Define the $(p + q + 1 \times n)$ partitioned matrix $Z^* = [Z' \ Y']'$. For the case of $q$ outputs, $q \geq 1$, $|Z^*| = |ZZ'| \cdot |YY' - YZ'B|$ with $B = (ZZ')^{-1}ZY'$. For $j, k = 1, \ldots, q$, $Y_j Y'_k - Y_j Z'b_k = Y_j(Y'_k - Z'b_k) = Y_j e'_k = Y_j M_Z Y'_k = Y_j M_Z' M_Z Y'_k = e_j e'_k$, where $M_Z$ is the idempotent matrix $(I - Z'(ZZ')^{-1}Z)$ and $e_j, e_k$ are ordinary least squares (OLS) residuals from the regressions of $Y'_j, Y'_k$ on $Z'$.

Then $|Z^*| = |ZZ'| \times |YY' - YZ'B| = |ZZ'| \times |\Omega|$, where $\Omega = [e_j e'_k], \ j, k = 1, \ldots, q$. The statistic

$$R_L^\ell(Z') = D_L^\ell(|Z'Z|) \cdot |Z'Z|^{-1} [D_L^\ell(|\Omega|)] |\Omega|^{-1} \quad (6.1)$$

represents the proportion of the geometric volume in $(p + q + 1)$-space spanned by a subset of the data obtained by deleting the $\ell$ observations.
with indices in the set $L$, relative to the volume spanned by the entire set of $n$ observations.

Wilson (1993) discusses a statistical procedure for deriving significance levels for the statistic defined in (6.1), and also describes a graphical analysis where log-ratios $\log \left[ R^{(\ell)}_L (Z^{*}) / R^{(\ell)}_{\min} \right]$, where $R^{(\ell)}_{\min} = \min_L \left\{ R^{(\ell)}_L (Z^{*}) \right\}$, are plotted as a function of $\ell$, the number of observations deleted. This method is implemented in the FEAR software package described by Wilson (2008).

Wilson (1993) provides further details and empirical examples. However, results for the second example in Section 2 of Wilson (1993), based on data from Charnes et al. (1981), inadvertently ignored the first row of $Z^*$ (results for the first example in Wilson, 1993, based on data from Wood, 1993, are unaffected). Consequently, the results reported in Table 2 and in Figure 2 of Wilson (1993) are wrong; corrected results are given here in Table 6.1 and Figure 6.1. After the correction, it is clear that observation 33 is identified as an outlier when $i \geq 3$; in the original table, observation 33 first appeared as an outlier when $i = 3$, but did not appear as an outlier when $i = 4$, though the observation was selected as an outlier for $i \geq 5$. In addition, Table 2 in Wilson (1993) indicated that observation 46 appeared as an outlier when $i = 11$ or 12, followed by observation 68 when $i = 12$. In the corrected results, observation 68 appears as an outlier when $i = 11$ or 12, followed by observation 46 when $i = 12$. 

---

<table>
<thead>
<tr>
<th>$i$</th>
<th>Observations</th>
<th>$R^{(i)}_{\min}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>59</td>
<td>0.3396</td>
</tr>
<tr>
<td>2</td>
<td>59, 44</td>
<td>0.1436</td>
</tr>
<tr>
<td>3</td>
<td>59, 44, 33</td>
<td>0.08258</td>
</tr>
<tr>
<td>4</td>
<td>59, 44, 33, 35</td>
<td>0.05092</td>
</tr>
<tr>
<td>5</td>
<td>59, 44, 33, 35, 66</td>
<td>0.03091</td>
</tr>
<tr>
<td>6</td>
<td>59, 44, 33, 35, 66, 67</td>
<td>0.01779</td>
</tr>
<tr>
<td>7</td>
<td>59, 44, 33, 35, 66, 67, 68</td>
<td>0.01192</td>
</tr>
<tr>
<td>8</td>
<td>59, 44, 33, 35, 66, 67, 54, 50</td>
<td>0.007481</td>
</tr>
<tr>
<td>9</td>
<td>59, 44, 33, 35, 66, 67, 54, 50, 1</td>
<td>0.004578</td>
</tr>
<tr>
<td>10</td>
<td>59, 44, 33, 35, 66, 67, 54, 50, 1, 10</td>
<td>0.002874</td>
</tr>
<tr>
<td>11</td>
<td>59, 44, 33, 35, 66, 67, 54, 50, 1, 10, 68</td>
<td>0.001874</td>
</tr>
<tr>
<td>12</td>
<td>59, 44, 33, 35, 66, 67, 54, 50, 1, 10, 68, 46</td>
<td>0.001238</td>
</tr>
</tbody>
</table>
Turning to the graphical analysis described in Wilson (1993), the corrected log-ratio plot shown in Figure 6.1 has three peaks as did the original Figure 2 in Wilson (1993). In both cases, the peaks correspond to $i = 2$, 6, and 9, and identify three groups of outliers: observations 44 and 59 (corresponding to the first peak), observations 33, 35, 66, and 67 (corresponding to the second peak), and observations 1, 50, and 54 (corresponding to the third peak). As in the original results, observation 68 appears as an outlier when $i = 7$, but not when $i = 8$ or 9. While observations 50 and 54 appear as outliers in the third group of outliers, it is apparent that they mask each other’s effect since then are only identified as outliers when both can be simultaneously deleted.

Wilson (1993) noted that the example using data from Wood (1973) with $i_{\text{max}} = 12$ required “less than 161 minutes of processor time on a Sun Microsystems 4/70 (SPARCstation 2) desktop workstation.” Running a single thread on a more modern 64-bit Intel XEON 5060 processor, the example using the Wood (1973) data with $i_{\text{max}} = 12$ requires 52.8 seconds of CPU time, while the example using the Charnes et al. (1981) data (again with $i_{\text{max}} = 12$) requires 52.84 seconds of CPU time. With advances in computer technology, the outlier detection
Outlier Detection

method described by Wilson (1993) has become increasingly practical in terms of computational requirements. The outlier diagnostics have been implemented in the `ap` and `ap.plot` commands in the `FEAR` package for use with R (see Wilson, 2008 for details).

Simar (2003) proposed an outlier detection strategy using the ideas of partial frontiers developed in the preceding sections. If a sample observation remains outside the order-$m$ frontier as $m$ increases, or alternatively outside the order-$\alpha$ frontier as $\alpha$ increases, then such an observation may be an outlier. The procedure is implemented by first computing, for each sample observation $(X_i, Y_i)$, $i = 1, \ldots, n$, the leave-one-out order-$m$ efficiencies $\hat{\theta}^{(i)}_{m,n}(X_i, Y_i)$ for several values of $m$, e.g., $m \in \{10, 25, 50, 75, 100, 150, \ldots\}$, where $\hat{\theta}^{(i)}_{m,n}(X_i, Y_i)$ denotes the order-$m$ efficiency estimate for the $i$th observation computed after deleting the $i$th observation. Next, the percentage of points lying outside the order-$m$ frontier (identified by values $\hat{\theta}^{(i)}_{m,n}(X_i, Y_i) > 1$ in the input-orientation, or $\hat{\lambda}^{(i)}_{m,n}(X_i, Y_i) < 1$ in the output-orientation) can be plotted as a function of $m$; the resulting points can be connected by line segments to form a piecewise-linear relationship that is necessarily downward sloping due to the properties of the order-$m$ estimator. If there are no outliers in the data set, the plotted relationship should be nearly linear. If the graph shows an elbow effect (i.e., a sharp bend, with large negative slope becoming much less negative at a particular value of $m$), firms remaining outside the order-$m$ frontier for the value of $m$ where the sharp bend occurs are likely outliers and should be carefully scrutinized for data errors, etc. The analysis can be performed using both input- and output-oriented perspectives to detect potential outliers in both directions.

This method requires that threshold values be chosen for deciding when the values $\hat{\theta}^{(i)}_{m,n}(X_i, Y_i)$ or $\hat{\lambda}^{(i)}_{m,n}(X_i, Y_i)$ are significantly greater or less than one. Ideally, these threshold values should be computed from the asymptotic standard deviations of the estimators. However, to make the procedure simple and easy to compute, Simar (2003) suggests producing several diagnostic plots as described above using percentages of points with $\hat{\theta}^{(i)}_{m,n}(X_i, Y_i) > 1 + \epsilon$ or $\hat{\lambda}^{(i)}_{m,n}(X_i, Y_i) > 1 - \epsilon$ with several small, arbitrary values of $\epsilon$; e.g., $\epsilon \in \{0, 0.05, 0.10, 0.15, 0.20, \ldots\}$.
See also Daouia and Gijbels (2011b). Daouia and Gijbels (2011a) use a similar idea in tuning an extremile-based estimator.

Although the method proposed by Simar (2003b) does not explicitly address the possibility of masking among outliers, if only a small number of outliers lie close to each other, the robustness properties of the order-$m$ estimator makes it likely these will be detected. Once an initial set of outliers has been identified, the analysis could be repeated after deleting the initial set of outliers. The method is implemented in the *FEAR* package (Wilson, 2008); in addition, Matlab code for the procedure as well as several examples are provided in Simar (2003).
Explaining Inefficiency

7.1 Introducing Environmental Variables

Researchers have sought to explain differences in estimated efficiencies across firms for many years. Typically, researchers have in mind a set of environmental factors that might be related to efficiency differences; these factors might reflect differences in ownership type or structure, regulatory constraints, business environment, competition, etc. among the firms under analysis. Typically, such factors are viewed as possibly affecting the production process, but not under the control of firms’ managers. Understanding how such factors might be related to efficiency is important for determining how firms’ performances might be improved. In addition, from a public policy perspective, understanding these relationships is important for assessing the costs of regulation.

One approach (e.g., Coelli et al., 1997) is to augment the basic statistical model introduced earlier by treating the \( r \) (continuous) environmental factors \( Z \) as free disposal inputs or outputs that contribute to defining the attainable set \( \mathcal{P} \subset \mathbb{R}_+^p \times \mathbb{R}_+^q \times \mathbb{R}^r \), and to modify
the efficiency scores to reflect the view that the environmental variables are beyond managers’ control. For example, the analog of (1.9) would be

$$\lambda(x, y | z) = \inf \{ \lambda \mid (x, \lambda y, z) \in \mathcal{P} \}.$$  \hspace{1cm} (7.1)

The FDH and DEA estimators of $\mathcal{P}$ in this case would be defined as before after adding the variables $Z$ and treating them as either inputs or outputs (if $Z$ is a vector of variables, some elements of $Z$ might be treated as inputs, while others might be treated as outputs).

There are at least two problems this approach. First, one must decide \textit{a priori} what the role of $Z$ is; in other words, do the variables contained in $Z$ contribute to, or hinder production of quantities $Y$? Second, free-disposability is assumed when FDH or DEA estimators are used; in addition, convexity of $\mathcal{P}$ is also assumed when DEA estimators are employed. For many environmental variables, it is doubtful such assumptions might be appropriate.

When discrete, categorical factors are considered (e.g., for-profit versus non-profit organizations; private versus public ownership; etc.), the categorical variables effectively divide producers into different groups. If there are $\tilde{r}$ binary, discrete categorical variables, then there are potentially $2^{\tilde{r}}$ different groups represented among the $n$ sample observations. In this case, one might be tempted to test hypothesis about differences in average efficiency, etc. across groups to make comparisons among the $2^{\tilde{r}}$ implied groups of firms. One might test, for example, whether the various groups face the same technology, or whether the distributions of efficiencies are the same across groups if they do face the same technology. However, the statistical theory needed to implement such tests remains unsettled.

When the environmental variables $Z$ are continuous, observations are not grouped into a small number of categories. Two approaches for investigating the effects of continuous environmental variables on efficiency are discussed below. The first is based on a two-stage approach that has been widely used in the literature, but too often with dubious specifications and invalid inference. The second approach was recently proposed by Daraio and Simar (2005), and involves an extension of the probabilistic formulation introduced earlier using (2.40).
7.2 Two-Stage Regression Approach

Two-stage estimation procedures wherein technical efficiency is estimated by DEA or FDH estimators in the first stage, and the resulting efficiency estimates are regressed on some environmental variables in a second stage, are common in the literature on efficiency analysis. The Google Scholar search engine returned about 20,000 items after a search on “efficiency,” “two-stage,” and “dea” on August 14, 2012. Replacing “dea” with “fdh” returned 743 hits. A large number of papers employing a two-stage approach use either ordinary least squares (OLS) or tobit regression in the second stage and rely on conventional methods for inference.

Simar and Wilson (2007) presented a well-defined, coherent statistical model in which a second-stage regression is meaningful in the sense that the form of the second-stage regression equation is determined by the structure of the model in the first stage where the initial DEA estimates are obtained. In an attempt to rationalize studies where second-stage regressions have been estimated but no statistical model has been specified, Simar and Wilson introduced assumptions that lead to a truncated regression in the second stage that can be estimated consistently using the maximum likelihood (ML) method. As discussed below, the assumption leading to a truncated regression in the second stage can be easily replaced to obtain a logistic or other parametric regression equation, or even a fully non-parametric regression equation. However, irregardless of the specification of the second-stage regression (i.e., either parametric or nonparametric), conventional inference methods fail to give valid inference due to the fact that in the second-stage, true efficiency remains unobserved and must be replaced with DEA estimates of efficiency, and these are correlated by construction. Simar and Wilson showed how bootstrap methods can be used for inference in the case of a truncated regression, and these methods are easy to extend to cases where different assumptions, leading to different forms of the second-stage regression equation in their model, are made.

Simar and Wilson (2007) cited 48 published papers that regressed DEA efficiency estimates on some environmental variables in a second stage, and commented that “as far as we have been able to determine,
none of the studies that employ this two-stage approach have described the underlying data-generating process.” Simar and Wilson (2007) went on to

(1) define a statistical model where truncated (but not censored, i.e., tobit, nor OLS) regression yields consistent estimation of model features;
(2) demonstrated that conventional, likelihood-based approaches to inference are invalid; and
(3) developed a bootstrap approach that yields valid inference in the second-stage regression when such regressions are appropriate.

It is important to note that Simar and Wilson did not recommend that anyone should use two-stage procedures; rather, the point of the paper was (i) to rationalize what has been done in the literature by providing a coherent, well-defined statistical model where a second-stage regression would be appropriate; and (ii) to show how valid inference could be made in the second-stage regression. With regard to the first point, the model provided by Simar and Wilson was apparently the first complete description of a DGP where second-stage regression would be appropriate. Simar and Wilson (2007) did not claim that this was the only such model; in fact, Banker and Natarajan (2008) proposed an alternative, though restrictive, model as discussed by Simar and Wilson (2011b).

To be precise, 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. In addition, let \( Z \in \mathcal{Z} \subseteq \mathbb{R}^r \) denote a vector of \( r \) environmental variables with domain \( \mathcal{Z} \). Firms transform quantities of inputs into various quantities of outputs, but environmental variables may affect how well firms do this on average, or possibly the range of possibilities for production. Given \( Z \), a firm becomes more technically efficient if it increases at least some of its output levels without increasing its input levels (output orientation), or alternatively if it reduces its use of at least some inputs without decreasing output levels (input orientation). In the real world, the analyst observes a set of observations \( S_n = \{(X_i, Y_i, Z_i)\}_i^n \).
Assumption 7.1. The sample observations \((X_i, Y_i, Z_i)\) in \(S_n\) are realizations of identically, independently distributed random variables \((X, Y, Z)\) with probability density function \(f(x, y, z)\) which has support over a compact set \(P \subseteq \mathbb{R}_+^{p+q} \times \mathbb{R}^r\) with level sets \(P(z)\) defined by

\[
P(z) = \{(X, Y) \mid Z = z, X \text{ can produce } Y\}. \tag{7.2}
\]

Let

\[
\Psi = \bigcup_{z \in \mathcal{Z}} P(z) \subseteq \mathbb{R}_+^{p+q}. \tag{7.3}
\]

The model developed by Simar and Wilson (2007) to rationalize regression of technical efficiency estimates on environmental variables in a second-stage regression involves the following, additional assumption on the production set \(P\).

Assumption 7.2. \(P(z) = \Psi \forall z \in \mathcal{Z}\).

Under Assumption 7.2, \(P = \Psi \times \mathcal{Z}\); this is the “separability” condition described by Simar and Wilson (2007). Whether the assumption is satisfied is ultimately an empirical question which, as Simar and Wilson noted (pages 35 and 36), should be tested. Unfortunately, however, no formal test is available, although work is progressing on this problem.

Note that \(y \in \mathbb{R}_+^q\) can be described in terms of polar coordinates; Simar and Wilson (2007) demonstrate that the modulus is related to the Farrell output efficiency measure in (1.9), and so \(y\) can be described by \((\eta, \lambda)\), where \(\eta\) is a vector of \((q - 1)\) angles and \(\lambda\) is the Farrell output efficiency corresponding to \(Y\). Then the joint density \(f(x, y, z)\) can be described by a series of conditional densities and in terms of cylindrical coordinates:

\[
f(x, \eta, \lambda, z) = f(x, \eta \mid \lambda, z)f(\lambda \mid z)f(z). \tag{7.4}
\]

The order of the conditioning on the right-hand side of (7.4) reflects the sequential nature of the DGP. Firm \(i\) is faced with environmental
variables $Z_i$ drawn from $f(z)$. Given this $Z_i$, an efficiency level $\lambda_i$ is
drawn from $f(\lambda | Z_i)$, and then $X_i$ and $\eta_i$ are drawn from $f(x, \eta | \lambda_i, Z_i)$, resulting in a realization
$(X_i, Y_i, Z_i)$ from the joint density $f(x, y, z)$ after transforming the polar coordinates $(\eta_i, \lambda_i)$ to Cartesian
coordinates $Y_i$.

Assumptions 7.1 and 7.2, together with the next assumption, comprise the nucleus of the model developed by Simar and Wilson (2007).

**Assumption 7.3.** The conditioning in $f(\lambda | z)$ in (7.4) operates
through the following mechanism:

$$\lambda_i = \psi(Z_i) + \varepsilon_i \geq 1,$$

(7.5)

where $\psi$ is a smooth, continuous, function and $\varepsilon_i$ is a continuous iid random variable, independent of $Z_i$.

Assumptions 7.1–7.3 impose a “separability” condition on the model
of Simar and Wilson (2007). In this model, $P(z)$ is unaffected by $z$, but
the distribution of technical efficiency depends on $z$ through Assump-
tion 7.3.\(^1\) As Simar and Wilson (2007, p. 36) noted, however, “one
might reasonably wonder whether the implied separability condition is
supported by the data.”

The next two assumptions are similar to Assumptions 1.2 and 1.3,
which have been modified to account for the environmental variables $Z$.

**Assumption 7.4.** For any $z \in Z$, $(x, y) \notin P(z)$ if $x = 0, y \neq 0$; i.e.,
all production requires use of some inputs.

**Assumption 7.5.** For any $z \in Z$, $\bar{x} \geq x$, and $\bar{y} \leq y$, if $(x, y) \in P(z)$
then $(\bar{x}, y) \in P(z)$ and $(x, \bar{y}) \in P(z)$, i.e., both inputs and outputs are
strongly disposable.

\(^1\)In the empirical literature, researchers have typically assumed
$\psi(Z_i) = Z_i \beta$, where $\beta$ is
an $(r \times 1)$ vector of parameters. In addition to Assumptions 7.1–7.3, Simar and Wilson
(2007) assume the error $\varepsilon$ in (7.5) is distributed (truncated) normal in order to reflect the
empirical literature. Alternatively, $\psi(Z_i)$ and the distribution of $\varepsilon$ can be assumed to be
nonparametric; see Park et al. (2008) for details.
Assumption 7.5 is equivalent to an assumption of monotonicity of the technology.

The next assumptions define a DGP; the framework here is similar to that in Simar (1996), Kneip et al. (1998), Simar and Wilson (1998; 2000a), Kneip et al. (2008), and Jeong et al. (2010).

**Assumption 7.6.** The $n$ observations in $S_n$ are iid random variables on the attainable set $P$.

**Assumption 7.7.**

(a) Conditional on $Z = z$, the $(X, Y)$ possess a joint density $f_{XY|Z}$ with support $P(z)$; (b) $f_{XY|Z}$ is continuous on $P(z)$; and (c) $f_{XY|Z}(x, \lambda(x, y)) > 0 \forall (x, y)$ in the interior of $P(z)$.

Assumption 7.7(c) imposes a discontinuity in $f$ at points on the boundary of $P(z)$, ensuring a strictly positive, non-negligible probability of observing production units close to the production frontier. For points lying outside $P(z)$, $f \equiv 0$.

**Assumption 7.8.** The function $\lambda(x, y | z)$ is twice continuously differentiable for all $z \in Z$ and $(x, y) \in P(z)$.

Assumption 7.8 imposes some smoothness on the boundary of $P$. This assumption is slightly stronger, but simpler, than a corresponding assumption needed by Kneip et al. (1998) to establish consistency of the DEA estimators. Assumption 7.8 is adapted from Kneip et al. (2008) and Jeong et al. (2010), where additional discussion is given; the assumption is similar to Assumption 2.3 given earlier.

The separability condition in Assumption 7.2 implies that the support of the output variables does not depend on the environmental variables in $Z$. To illustrate this condition, consider the two DGPs given by

\[ Y^* = g(X)e^{\beta ZU} \]  \hspace{1cm} (7.6)

and

\[ Y^{**} = g(X)e^{\beta Z e^{-U}} \]  \hspace{1cm} (7.7)
where $\beta = 0.15$, $g(X) = \left(1 - (X - 1)^2\right)^{1/2}$, $X \in [0, 1]$, $Z \in [0, 4]$, and $U \geq 0$ is a one-sided inefficiency process. Setting $U = 0$ in (7.6) and (7.7) gives the frontiers for the two DGPs, as illustrated in Figure 7.1, where the frontier corresponding to (7.6) is shown in the left panel, and the frontier corresponding to (7.7) is shown in the right panel. Clearly, the frontiers are very different; it is clear that for a given level of the input variable $X$, the maximal output level $Y^*$ in (7.6) does not vary with $Z$, as reflected in panel (a) of Figure 7.1. However, the maximal output level $Y^{**}$ in (7.7) does vary with $Z$, and the frontier shown in panel (b) of Figure 7.1 is seen to fall as $Z$ increases. The “separability” condition discussed by Simar and Wilson (2007) is satisfied by the DGP in (7.6), but not by the DGP in (7.7).

Continuing with this example, now consider the observations $(0.6, 0.5, 1.0)$, $(0.6, 0.5, 2.0)$, and $(0.6, 0.5, 4.0)$ for $(X,Y,Z)$. If the true model is (7.6), then some calculations reveal that the output efficiency corresponding to each of these observations is (approximately) $0.9165/0.5 \approx 1.8330$. By contrast, if the true model is given by (7.7), then the output efficiencies are (approximately) $0.7889/0.5 \approx 1.5777$, $0.6790/0.5 \approx 1.3579$, and $0.5030/0.5 \approx 1.0060$, respectively. As noted in the previous paragraph, the frontier corresponding to (7.6) is invariant with respect to $Z$, while the frontier corresponding to (7.7) is not. It is clear that whether the “separability” condition holds has an impact.
on the underlying, true efficiency levels, and this impact may be large. In the example considered here, the output efficiency level for the third observation is about 82.2 percent larger if the DGP is given by (7.6), where the “separability” condition is satisfied, as opposed to the case where the DGP is given by (7.7), where the “separability” condition is not satisfied.

Continuing with the example illustrated in Figure 7.1, given a sample \( \{(X_i, Y_i, Z_i)\}_{i=1}^n \), consider what would it mean to estimate efficiency with DEA using the observations \( \{(X_i, Y_i)\}_{i=1}^n \) if the underlying technology is the one in the right-hand panel. The preceding argument makes clear that for a particular observation \((X_i, Y_i)\), DEA would estimate the distance not to the frontier \( P(Z_i) \), but to the boundary of the set \( \Psi \) described in (7.3). In terms of the right-hand panel in Figure 7.1, the frontier of the corresponding set \( \Psi \) is identical to the frontier shown in the left-hand panel of Figure 7.1. Hence the DEA estimator, for a point \((X_i, Y_i)\), measures distance not to the technology, but to a frontier that is very different from the frontier shown in the right-hand panel of Figure 7.1. Clearly, if the separability condition is not satisfied, as in the right-hand panel of Figure 7.1, measuring efficiency while ignoring this fact leads to meaningless results in the first stage of any two-stage estimation procedure.

Now let \( \lambda_i = \lambda(X_i, Y_i) \) denote (true) output efficiencies for points \((X_i, Y_i)\) when the separability condition is satisfied. If the \( \lambda_i \) were observed, it would be straightforward to estimate (7.5). One might assume \( \psi(Z, \beta) = Z\beta \) and estimate the model by the maximum likelihood method using standard software; note, however, that in the model given by (7.6), the relation between Farrell output efficiency and \( Z \) is given by\(^{7.8}\)

\[
\lambda = \frac{g(X)}{Y^*} = e^{-(Z-2)^2U}
\]

and hence is nonlinear in \( Z \). One could model this explicitly if the true DGP in (7.6) were known, or alternatively, one could allow \( \psi(Z, \beta) \) and the distribution of \( \epsilon \) to be nonparametric and estimate \( \psi(\cdot) \) using the local likelihood method discussed by Park et al. (2008).

Unfortunately, however, the \( \lambda_i \) are not observed. SW present two approaches for dealing with this problem. In the typical two-stage
estimation approach, DEA estimates $\hat{\lambda}_i$ from the first stage estimation are used to replace the unobserved $\lambda_i$ in (7.5) with $\psi(Z_i, \beta) = Z_i \beta$. Since the DEA estimates are consistent under the assumptions of the model in Simar and Wilson (2007), maximum likelihood estimation of the truncated regression

$$\hat{\lambda}_i = Z_i \beta + \xi_i \geq 1 \quad (7.9)$$

corresponding to Equation (13) of Simar and Wilson (2007) yields consistent estimates of $\beta$. However, as Simar and Wilson note, inference is not straightforward due to the fact that $\hat{\lambda}_i$ has replaced the unobserved $\lambda_i$, and while the $\hat{\lambda}_i$ consistently estimate the $\lambda_i$, the DEA estimators converge slowly, at rate $n^{-2/(p+q+1)}$, and are biased. Consequently, the inverse of the negative Hessian of the log-likelihood corresponding to (7.9) does not consistently estimate the variance of the ML estimator of $\beta$. Simar and Wilson (2007) give two bootstrap procedures to overcome this problem, and in addition provide simulation evidence suggesting that the methods yield confidence intervals with coverages close to nominal levels. See Simar and Wilson (2007) for specific details.

It is important to note that Simar and Wilson (2007) did not, for reasons given here, recommend using two-stage methods. The goal in Simar and Wilson (2007) was to provide a well-defined statistical model that could rationalize what has been done in the literature. In the end, the model in Simar and Wilson (2007) requires truncated regression in the second stage. Within the assumptions of the model in Simar and Wilson, tobit regression constitutes a mis-specification. Moreover, the simulation results presented by Simar and Wilson confirm that under the assumptions of their model, tobit estimation in the second stage yields biased and inconsistent estimates.

To date, no well-defined statistical model in which second-stage tobit regression of DEA (or FDH) efficiency estimates on some environmental variables would produce consistent estimates has been presented in the literature. Despite this, and despite the results in Simar and Wilson (2007), papers that employ tobit-regression and other ad-hoc methods in second-stage regressions continue to appear, although perhaps typically not in serious journals. Simar and Wilson (2011b)
discuss some of these pseudo-scientific, ad-hoc approaches and provide further details.

### 7.3 Conditional Efficiency Measures

In situations where the separability condition in Assumptions 7.2 does not hold, Assumptions 7.2 and 7.3 may be replaced with the following assumption.

**Assumption 7.9.** \( \mathcal{P}(z) \neq \Psi \) for some \( z \in \mathcal{Z} \); i.e., \( \mathcal{P}(z) \neq \mathcal{P}(\tilde{z}) \) for some \( z \neq \tilde{z}, z, \tilde{z} \in \mathcal{Z} \).

If the sets \( \mathcal{P}(z) \) are convex for all \( z \in \mathcal{Z} \), then the set \( \Psi \) is convex under Assumption 7.2. However, if Assumption 7.9 holds, \( \Psi \) is not in general convex, even if the sets \( \mathcal{P}(z) \) are convex for all \( z \in \mathcal{Z} \). Moreover, under Assumption 7.9, some of the input-output combinations in \( \Psi \) are not attainable. Consequently, \( \theta(x, y \mid \Psi) \), \( \lambda(x, y \mid \Psi) \), etc. measure distance from the point \((x, y)\) to the frontier of an unattainable set when Assumption 7.9 holds, and thus have no meaning in familiar economic terms; moreover, if Assumption 7.9 holds, the usual DEA estimators of \( \Psi \) the various efficiency measures given in (1.7), (1.9), (1.11), and (1.13), are statistically inconsistent. This further implies that Assumption 7.3 is meaningless under Assumption 7.9; hence, regressing estimated efficiencies on environmental variables is also meaningless under Assumption 7.9. In cases where Assumption 7.9 holds, *conditional* measures of technical efficiency developed by Daraio and Simar (2005) are needed.

Daraio and Simar (2005) proposed an intuitive way to introduce environmental factors into the production process by defining a conditional efficiency measure; the idea is an extension of the ideas of Cazals et al. (2002). The approach does not require an a priori assumption on the effect of \( Z \) on efficiency as in existing one-stage approaches (e.g., the approach of Coelli et al., 1997), nor does it require a separability condition as with the two-stage approach discussed above. The approach by Daraio and Simar (2005) follows from the probabilistic formulation of the production process used for defining the partial frontiers in *Robust*
Order-\(m\) Estimators and Robust Order-\(\alpha\) Estimators, and is appropriate in situations where Assumption 7.9 holds.

Recall that the production process is characterized by a joint probability measure of \((X, Y)\) on \(\mathbb{R}_+^p \times \mathbb{R}_+^q\) generating observations \(\{(X_i, Y_i)\}, \ i = 1, \ldots, n\), where the support of \((X, Y)\) is the attainable set \(\mathcal{P}\). Recall also that under the free-disposability assumption, the efficiency measures defined in (1.7), (1.9), (1.11), and (1.13) can be defined in terms of this joint probability measure as in (2.44)–(2.47); e.g., the input-oriented Debreu–Farrell efficiency measure defined in (1.7) can be defined in terms of the conditional distribution function where

\[
F_{X|Y}(x | y) = \Pr(X \leq x \mid Y \geq y)
\]

as in (2.44). The corresponding FDH estimators can be defined in terms of the empirical analogs of (2.44)–(2.47), as was done in (2.55), (2.57), (2.59), and (2.60).

The joint density introduced in Assumption 7.1 implies a conditional distribution function

\[
H_{XY|Z}(x, y | z) = \Pr(X \leq x, Y \geq y \mid Z = z)
\]

\[
= \frac{\Pr(X \leq x \mid Y \geq y, Z = z) \Pr(Y \geq y \mid Z = z)}{S_{Y|Z}(y | z) F_{X|YZ}(x | y, z)}
\]

\[
= \frac{\Pr(Y \geq y \mid X \leq x, Z = z) \Pr(X \leq x \mid Z = z)}{-S_{Y|XZ}(y | x, z) - F_{X|Z}(x | z)}
\]

(7.10)

(7.11)

where \(F_{X|YZ}(x | y, z)\) is the conditional (on \(Y \geq y\) and \(Z = z\)) distribution function of \(X\); \(S_{Y|Z}(y | z)\) is the conditional (on \(Z = z\)) survivor function for \(Y\); \(S(y | x, z)\) is the conditional (on \(X \leq x\) and \(Z = z\)) survivor function of \(Y\), and \(F_{X|Z}\) is the conditional (on \(Z = z\)) distribution function for \(X\). Note the different treatment of \(Z\) and \(Y\) in the conditioning in (7.10) and (7.11). Since \(Y(X)\) is an output (input) vector, the conditioning is on \(Y \geq y\) (\(X \leq x\)). By contrast, \(Z\) is a vector of environmental variables whose effect on output is of unknown sign; hence the conditioning is on \(Z = z\).
Conditional versions of the efficiency measures defined in (1.7), (1.9), (1.11), and (1.13) can be written as

\[ \theta(x, y \mid P(z)) = \inf\{\theta \mid F_{X|Y,Z}(\theta x \mid y, z) > 0\} = \inf\{\theta \mid H_{XY|Z}(\theta x, y \mid z) > 0\}, \]

(7.12)

\[ \lambda(x, y \mid P(z)) = \sup\{\lambda \mid S_{Y|X}(\lambda y \mid x, z) > 0\} = \sup\{\lambda \mid H_{XY|Z}(x, \lambda y \mid z) > 0\}, \]

(7.13)

\[ \gamma(x, y \mid P(z)) = \inf\{\gamma \mid H_{XY|Z}(\gamma^{-1}x, \gamma y \mid z) > 0\}, \]

(7.14)

and

\[ \delta = (x, y \mid u, v, P(z)) = \sup\{\delta \mid H_{XY|Z}(x - \delta u, y + \delta v \mid z) > 0\}. \]

(7.15)

Note that there are no \textit{a priori} assumptions on \(Z\) (such as \(Z\) acting as a free disposal input, or as an undesired, freely disposable output).

Nonparametric estimators of the conditional efficiency measures defined by (7.12)--(7.15) are obtained using the familiar plug-in principle. First, consider (7.12); the idea is to replace \(H_{XY|Z}(x, y \mid z)\) with a nonparametric estimator. Due to the conditioning on \(Z = z\), this requires some smoothing along the dimensions of \(z\); this is accomplished by adding a kernel-type estimator to the empirical distribution function to obtain

\[ \hat{H}_{XY|Z,n}(x, y \mid z) = \sum_{i=1}^{n} I(X_i \leq x, Y_i \geq y)K_h(Z_i - z), \]

(7.16)

where \(K_h(\cdot) = K(\cdot/h)\) and \(K(\cdot)\) is an \(r\)-variate kernel function and \(h\) is a bandwidth of appropriate size. If \(r > 1\), a multivariate kernel function may be constructed from a univariate kernel function \(\tilde{K}(\cdot)\), either as a product kernel by setting

\[ K(z) = \prod_{\ell=1}^{r} \tilde{K}(z^\ell), \]

(7.17)

or as a spherically-symmetric kernel by setting

\[ K(z) = \tilde{K}(z'z), \]

(7.18)
where $z$ is ($r \times 1$).

The conditional FDH efficiency estimators of (7.12)–(7.15) can now be defined as

$$
\hat{\theta}_n(x, y \mid P(z)) = \inf\{\theta \mid \hat{H}_{XY|Z,n}(\theta x, y \mid z) > 0\},
$$

(7.19)

$$
\hat{\lambda}_n(x, y \mid P(z)) = \sup\{\lambda \mid \hat{H}_{XY|Z,n}(x, \lambda y \mid z) > 0\},
$$

(7.20)

$$
\hat{\gamma}_n(x, y \mid P(z)) = \inf\{\gamma \mid \hat{H}_{XY|Z,n}(\gamma^{-1}x, \gamma y \mid z) > 0\},
$$

(7.21)

and

$$
\hat{\delta}_n(x, y \mid u, x \mid P(z)) = \inf\{\delta \mid \hat{H}_{XY|Z,n}(x - \delta u, y + \delta v \mid z) > 0\},
$$

(7.22)

As observed by Daraio and Simar (2005), when kernel functions with unbounded support (e.g., the Gaussian kernel) are used, the estimators in (7.19)–(7.22) will equal the corresponding FDH efficiency estimators; hence only kernels that have compact support (e.g., $K(u) = 0 \forall u \notin [0, 1]$, such as the Epanechnikov kernel

$$
\tilde{K}(u) = \frac{3}{4} (1 - u^2) \mathbb{1}(|u| \leq 1)
$$

(7.23)

can be used. Nonetheless, computing estimates from (7.19)–(7.22) is straightforward and easy, given a bandwidth $h$. For example, $\hat{\theta}_n(x, y \mid \hat{P}_{FDH}(z))$ can be computed as

$$
\hat{\theta}(x, y \mid \hat{P}_{FDH}(z)) = \min_{\{i \mid Y_i \geq y, |Z_i - z| \leq h\}} \left( \max_{j=1,\ldots,p} \frac{X_j}{x_j} \right).
$$

(7.24)

Similarly, the estimators in (7.20) and (7.21) can be computed as

$$
\lambda(x, y \mid \hat{P}_{FDH}(z)) = \max_{\{i \mid x_i \leq x, |Z_i - z| \leq h\}} \left( \min_{j=1,\ldots,q} \left( \frac{Y_j}{y_j} \right) \right)
$$

(7.25)

and

$$
\gamma(x, y \mid \hat{P}_{FDH}(z)) = \max_{\{i \mid |Z_i - z| \leq h\}} \left( \min_{k=1,\ldots,p} \left( \frac{X_j}{X_j^k}, \frac{Y_k}{y_k} \right) \right).
$$

(7.26)

Finally, (7.22) can be computed by transforming the data using (2.15) and (2.16) and then using (7.26) on the transformed data.
Daraio and Simar (2007b) give a localized estimator of $P$ when environmental variables are present; this leads to localized, conditional versions of the DEA efficiency estimators. In the input orientation, the conditional DEA estimator of $\hat{\theta}(x, y \mid z)$ is

$$\hat{\theta}(x, y \mid \hat{P}_{VRS}(z)) = \max_{\theta, \omega_1, \ldots, \omega_n} \left\{ \theta > 0 \mid y \leq \sum_{i=1}^{n} \omega_i Y_i, \quad \theta x \geq \sum_{i=1}^{n} \omega_i X_i, \quad \omega_i = 1, \quad \omega_i \geq 0 \quad \forall \ i \text{ such that } |Z_i - z| \leq h \right\},$$

(7.27)

where $h$ is again an $r$-vector of bandwidth parameters. In the output orientation, the conditional DEA estimator of $\hat{\lambda}(x, y \mid z)$ is

$$\hat{\lambda}(x, y \mid \hat{P}_{VRS}(z)) = \max_{\lambda, \omega_1, \ldots, \omega_n} \left\{ \lambda > 0 \mid \lambda y \leq \sum_{i=1}^{n} \omega_i Y_i, \quad x \geq \sum_{i=1}^{n} \omega_i X_i, \quad \omega_i = 1, \quad \omega_i \geq 0 \quad \forall \ i \text{ such that } |Z_i - z| \leq h \right\}.$$

(7.28)

Asymptotic results for both the conditional FDH and conditional DEA estimators (in the input and output orientations are given by Jeong et al. (2010)). With appropriate size of the bandwidth, the convergence rates of the conditional estimators are slower than their unconditional counterparts by a factor $n^{-4/(4+r)}$. The bandwidth parameter $h$ can be optimized using the least-squares cross-validation technique discussed by Bädin et al. (2010), provided the elements of $Z$ are continuous. If $Z$ contains qualitative variables, then the sample observations must first be divided into groups defined by the qualitative variables. For example, if $Z$ includes $r_c$ continuous variables and $r_d$ binary dummy variables, $Z$ can be partitioned by writing $Z = [Z^c \ Z^d]$. There are
7.3 Conditional Efficiency Measures

potentially $2^{r_d}$ groups of observations where the $r_d$ binary variables have the same values within each group (note some cells may be empty). For each group, either $\hat{\lambda}(x, y | \hat{P}_{FDH}(z))$ or $\hat{\lambda}(x, y | \hat{P}_{VRS}(z))$ can be computed using only the continuous elements $Z^c$ of $Z$, optimizing the bandwidth for continuous components of $Z$ using only observations within the given group. By computing estimates separately for each group, one necessarily conditions on $Z^d = z^d$. Dividing into groups is necessary since the bandwidths in (7.19)–(7.22) and (7.27)–(7.28) merely determine which observations fall into the reference set that defines the estimators. This contrasts with nonparametric regression estimators such as the Nadaraya–Watson estimator, where bandwidths determine kernel weights; here, however, no smoothing is performed since only the support of $H_{XY|Z}(x, y | z)$ is estimated. As a result, the ideas of Racine and Li (2004) for smoothing across discrete categories cannot be used here.

The conditional estimators presented here provide tools for investigating the effect of a particular factor $Z$ on the production process by comparing $\hat{\theta}(x, y | \hat{P}_{FDH})$ with $\hat{\theta}(x, y | \hat{P}_{FDH}(z))$. Daraio and Simar (2005) propose regressing the ratios $\hat{\theta}(x, y | \hat{P}_{FDH}(z))/\hat{\theta}(x, y | \hat{P}_{FDH})$ nonparametrically on the observed values $z$ (e.g., using a Nadaraya–Watson or local linear estimator); if the resulting regression curve is increasing, $Z$ is detrimental (unfavorable) to efficiency, and if the regression curve is decreasing, $Z$ is conducive (favorable) to efficiency.

Indeed, if $Z$ is unfavorable, the environmental variable acts as an “extra,” undesired output requiring the use of more inputs in production activity, and hence $Z$ has an adverse effect on the production process. By construction, $\hat{\theta}(x, y | \hat{P}_{FDH}(z))$ is always larger than $\hat{\theta}(x, y | \hat{P}_{FDH})$, but if $Z$ is unfavorable, $\hat{\theta}(x, y | \hat{P}_{FDH}(z))$ will be much larger than $\hat{\theta}(x, y | \hat{P}_{FDH})$ for large values of $Z$ than for small values of $Z$. Hence, if $Z$ is unfavorable, the ratios $\hat{\theta}(x, y | \hat{P}_{FDH}(z))/\hat{\theta}(x, y | \hat{P}_{FDH})$ will on average increase with $Z$.

On the other hand, if $Z$ is favorable to the production process, then the environmental variable plays the role of a “substitutive” input in the production process, providing an opportunity to conserve input quantities $x$ while producing $y$. In this case, $Z$ has a positive effect on
the production process, and the ratios \( \hat{\theta}(x, y \mid \hat{P}_{FDH}(z))/\hat{\theta}(x, y \mid \hat{P}_{FDH}) \) will, on average, decrease when \( Z \) increases.

Conceivably, one might discover situations where the ratios are on average increasing with \( z \) for some range of values of \( z \), neutral for another range, and then perhaps decreasing for still another range of values of \( z \). Because the method is fully nonparametric, it is flexible enough to detect various possibilities. Daraio and Simar (2005) illustrate the method with simulated data, confirming the method’s usefulness and its ability to detect the true effect of \( Z \) on the production process.

Of course, the ideas here easily adapt to the output-oriented case. Daraio and Simar (2005) have adapted the method to the robust order-\( m \) efficiency scores, and Daraio and Simar (2007b) have extended the idea to convex technologies, introducing conditional DEA efficiency scores along the same lines as the conditional FDH scores discussed here. It is equally easy to adapt the ideas presented here to the hyperbolic and directional orientations.

Analogs of the conditional measures discussed above can also be defined for the partial measures (order-\( m \) and order-\( \alpha \)), as suggested by Daraio and Simar (2007a, 2007b). Bădin et al. (2012b, 2012a) note that this has two advantages: first, by choosing high values for \( m \) and \( \alpha \), robust versions of the full frontier conditional measures are obtained; and second, by choosing low values for \( m \) or \( \alpha \) (e.g., \( m = 1 \) or \( \alpha = 0.5 \)), an idea of the effect of the environmental factors on the distribution of the inefficiencies when comparing with their unconditional counterparts (e.g., the mean when \( m = 1 \) or the median if \( \alpha = 0.5 \)) is obtained. To summarize, full frontier ratios as described above (or their robust versions) permit checking whether the environmental factors have an influence on the shape of the frontier (i.e., the support of the distribution), whereas the ratios obtained with the conditional partial frontiers (with low values of \( m \) or \( \alpha \)) facilitate investigation the impact of the environmental variables on the distribution of the inefficiencies.
The preceding discussion makes clear that nonparametric frontier models are statistical (or econometric) models. The FDH and DEA tools are estimators of unknown frontiers or efficiency levels, just as least-squares or maximum likelihood estimators are used to estimate unknown parameters in fully parametric models. As with any model, some assumptions are needed to establish asymptotic properties of estimators; the asymptotic properties allow one to make inferences about the efficiency levels of individual firms, i.e., the distance between observed firms and the true frontier estimated by FDH or DEA. The discussion above also indicates that the direction in which this distance is estimated does not matter too much in determining the asymptotic properties of the various efficiency estimators. As a practical matter, inference requires use of bootstrap methods, but today, computationally efficient methods and software are available.

The great advantage of nonparametric estimators is their flexibility and consistency properties under very general (i.e., mild) regularity conditions, as opposed to parametric estimators which typically require much stronger assumptions. Of course, many (but not all) nonparametric estimators incur the curse of dimensionality, which
causes convergence rates to be slower than the root-$n$ rate typically obtained with parametric estimators. However, as noted by Cameron and Trivedi (2005), misspecification of parametric models typically leads to inconsistent estimation. Hence, as noted above, parametric estimators can be viewed as “root-$n$ inconsistent,” meaning that if some of the restrictive parametric assumptions are not satisfied, the resulting estimators converge with very fast rate to some meaningless quantity. FDH and DEA estimators are sensitive to outliers or extreme observations, but this limitation can be addressed by using order-$m$ and order-$\alpha$ partial frontier estimators and the corresponding efficiency estimators. The order-$m$ and order-$\alpha$ efficiency estimators, while fully nonparametric, achieve root-$n$ rates of convergence and have Gaussian limiting distributions.

It is easy to introduce environmental factors into analyses of production processes. When the separability condition described earlier is satisfied, two-stage procedures can be used. More generally, one can estimate conditional frontiers and conditional efficiency scores (in input, output, hyperbolic, or directional orientations). The latter approach allows one to not only check whether the environmental factors might influence the shape of the frontier, but also, by using conditional partial frontiers, to check whether these factors have some impact on the distribution of the inefficiencies.

It is important to stress that parametric and nonparametric approaches are not in opposition; the main difference between the two approaches lies in what one is willing to assume about the underlying statistical model. In many cases, semi-parametric approaches might be a useful compromise; for example, one might use nonparametric methods for deriving better (i.e., less variable) estimators of a parametric frontier, under much less restrictive assumptions than would typically be required. Florens and Simar (2005) and Daouia et al. (2008) develop such a method in which nonparametric methods are first used to filter the data by eliminating inefficient firms, after which the desired parametric model is estimated. The procedure involves two-stages: first, all observations are projected onto some nonparametric estimate of the frontier (e.g., an FDH or partial order frontier estimate); then, in the second stage, a parametric model is then fitted by least squares
using the projected, “efficient” firms. The underlying idea is intuitively appealing: to learn about the shape of the frontier function, it is better to use an artificial sample of \( n \) efficient firms than the original sample of \( n \) firms containing primarily inefficient firms. Florens and Simar (2005) provide asymptotic properties of the resulting estimators when FDH and order-\( m \) estimators are used in the first stage, while Daouia et al. (2008) give asymptotic properties when order-\( \alpha \) estimators are used to filter the data.

Although great strides have been made in nonparametric and semiparametric approaches to efficiency estimation in recent years, a number of important, unanswered questions remain and are the focus of current, ongoing research. Only a few of these problems, where results should be available soon, are discussed below.

Figures 4.7 and 5.5 indicate that partial frontier estimates are rather rough and jagged. Jeong and Simar (2006) have shown how to linearize FDH estimators, and these ideas could be used to smooth the partial order frontiers. Daouia and Simar (2005) provide a simple device to to monotonize the estimators, which provides some smoothing of the resulting frontier estimates while retaining the asymptotic properties of the original partial frontier estimates. But more general smoothing techniques that keep the nice asymptotic properties of the robust estimators would be very useful for practitioners; these could be viewed as an extension of Florens and Simar (2005).

Use of the partial frontiers and their estimators requires choosing the order, \( m \) or \( \alpha \). By choosing these appropriately as a function of the sample size, it is possible to obtain estimators of the full frontier that share the robustness properties of the partial frontiers and retain Gaussian limiting distributions for the estimator of the full frontier. The idea is to define sequences \( m = m(n) \to \infty \) or \( \alpha = \alpha(n) \to 1 \) as \( n \to \infty \), but not too quickly (to avoid convergence to the FDH estimator), so that the asymptotic normality of the estimators does not disappear. This permits use of normal quantiles to provide (robust) confidence intervals for frontier points. Useful results are obtained for the order-\( \alpha \) estimator by Daouia et al. (2010) and for the order-\( m \) estimator by Daouia et al. (2012); these results are derived by adapting results from extreme value theory. To date, results have been obtained only for the
cases of univariate output when estimating a production frontier or univariate input when estimating a minimal input frontier. Work to extend these results to the full multivariate case is in progress.

A few small steps have been made toward introducing noise into nonparametric production models. The idea is to allow distance to the frontier to be contaminated by random noise, as in the case of parametric frontier models along the lines of Aigner et al. (1977) and others, so that the distance of a given observation to the frontier is the convolution of an one-sided inefficiency term and a two-sided noise term. A fully nonparametric model allowing convolution of inefficiency and two-sided noise is not identifiable, as shown by Hall and Simar (2002). Consequently, some amount of structure must be assumed to allow for identification. One approach is to leave the production function unspecified, while specifying a parametric density for the inefficiency term and an independent Gaussian process for the noise. The simplest (and perhaps most naive approach) would be to add a homoskedasticity assumption for both the inefficiency and noise processes. This has been investigated by Fan et al. (1996) and a few other papers in this vein (e.g., see Kuosmanen and Kortelainen, 2012). This semi-parametric approach has some appeal, but it is very restrictive: both the homoskedasticity assumption and the choice of a particular, parametric density for the inefficiency term seem problematic in the sense that whatever choices are made are likely to introduce parametric misspecification, and consequently statistical inconsistency, into the analysis.

An alternative approach is to use localization ideas. For example, when estimating a production function, for a given level of the inputs $\mathbf{x} \in \mathbb{R}_+^p$, a half-normal distribution with parameters $\mu_u(\mathbf{x})$ and $\sigma_u^2(\mathbf{x})$ might be assumed locally (i.e., in a neighborhood of $\mathbf{x}$) for the inefficiency process, while assuming (again, locally) a normal distribution with mean zero and variance $\sigma_v^2(\mathbf{x})$, where conditionally on $\mathbf{x}$, the inefficiency and noise terms are independent. Kumbhakar et al. (2007) use local polynomials for estimating the 3 functional parameters and maximize the local likelihood for such a model and derive asymptotic properties. The numerical and computational burdens involved in maximizing the local log-likelihood function are formidable, but feasible on
modern computers. The method is almost fully nonparametric since the inefficiency and noise processes are allowed to vary with \( x \).

Simar and Zelenyuk (2011) use the multivariate nonparametric model suggested by Simar (2007) to generalize the Kumbhakar et al. (2007) method to a fully multivariate framework, and propose “stochastic” versions of the FDH and DEA estimators (i.e., allowing for convolution of noise and inefficiency). The results are encouraging, and simulation results indicate that the procedure works well. The method provides modifications of the traditional FDH and DEA estimators that are robust with respect to outliers. Work remains to reveal the properties of these estimators, and to reduce the computational burden, and is progressing.

Yet another stream of research based on deconvolution methods is underway. In this approach, identification is achieved with only a minimal assumption of normality with unknown variance for the noise process; the remainder of the model, including the inefficiency process, is nonparametric. Kneip et al. (2012a) provide a first step in this direction.

The limiting distributions of FDH and DEA efficiency estimators, for individual efficiency scores, have been derived by Park et al. (2000) and Kneip et al. (2008). For purposes of testing hypotheses, test statistics could be based on functions of these scores estimated at random points \((X, Y) \in \mathcal{P}\). For example, to make inference about the average level of inefficiency, a natural statistic would be the sample mean of the efficiencies estimated at the (random) sample observations \((X_i, Y_i), i = 1, \ldots, n\). On the surface, this seems a very simple problem; but, it is not, because the bias of the FDH and DEA estimators cannot be tuned to balance their variance, and the estimators at points \((X_i, Y_i)\) are correlated. Kneip et al. (2012b) give a detailed analysis of this problem, and provide new central limit theorems that are useful for making inferences about mean efficiency.

A next step, currently underway, is to adapt the results from Kneip et al. (2012b) to develop more elaborate test statistics for testing more complicated hypotheses about the underlying model structure. For example, one might want to test whether the production set \( \mathcal{P} \) is convex, whether the frontier \( \mathcal{P}^g \) displays globally CRS, or
whether the separability condition discussed earlier holds. One might also want to test hypotheses regarding changes in productivity, efficiency, technology, etc., or hypotheses about the distribution of inefficiency (e.g., whether it is homogeneous, with constant variance), or whether inputs or outputs can be aggregated. Preliminary work on such testing issues can be found in Simar and Wilson (1999a, 2001b, and 2001a). Monte-Carlo evidence given in these papers give encouraging results, but theoretical justification based on the results in Kneip et al. (2012b) is needed, and is the target of ongoing research efforts.
Acknowledgments

We are grateful to Jean-Pierre Florens for comments on a draft version, and for support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) and the U.S. National Science Foundation (grant no. SMA-1243436). Any remaining errors are solely our responsibility.


Bickel, J. P. 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.


*Canadian Mathematical Bulletin* 2, 81–93.


