

A general methodology for bootstrapping in non-parametric frontier models

LÉOPOLD SIMAR¹ & PAUL W. WILSON², ¹*Institut de Statistique, Université Catholique de Louvain, Belgium* and ²*Department of Economics, University of Texas, Austin, USA*

ABSTRACT *The Data Envelopment Analysis method has been extensively used in the literature to provide measures of firms' technical efficiency. These measures allow rankings of firms by their apparent performance. The underlying frontier model is non-parametric since no particular functional form is assumed for the frontier model. Since the observations result from some data-generating process, the statistical properties of the estimated efficiency measures are essential for their interpretations. In the general multi-output multi-input framework, the bootstrap seems to offer the only means of inferring these properties (i.e. to estimate the bias and variance, and to construct confidence intervals). This paper proposes a general methodology for bootstrapping in frontier models, extending the more restrictive method proposed in Simar & Wilson (1998) by allowing for heterogeneity in the structure of efficiency. A numerical illustration with real data is provided to illustrate the methodology.*

1 Introduction

An extensive literature concerning the measurement of efficiency in production has developed since Debreu (1951) and Farrell (1957) provided basic definitions for technical and allocative efficiency in production. One large section of this literature focuses on linear-programming based measures of efficiency along the lines of Charnes *et al.* (1978, 1979), Deprins *et al.* (1984), and Färe *et al.* (1985).¹ Among this part of the literature, those approaches that rely on convexity assumptions are known as Data Envelopment Analysis (DEA).

DEA models measure efficiency relative to a non-parametric, maximum likelihood *estimate* of an unobserved *true* frontier, conditional on observed data resulting

Correspondence: Léopold Simar, Institut de Statistique, Université Catholique de Louvain, Voie du Roman Pays 20, Louvain-la-Neuve, Belgium.

from an underlying (and unobserved) data-generating process (DGP).² These methods have been widely applied to examining technical and allocative efficiency in a variety of industries; Lovell (1993) and Seiford (1996, 1997) provide lengthy bibliographies of these applications. Aside from the production setting, the problem of estimating monotone concave boundaries also naturally occurs in portfolio management. In capital asset pricing models (CAPM), the objective is to analyse the performance of investment portfolios. Risk and average return on a portfolio are analogous to inputs and outputs in models of production; in CAPM, the attainable set of portfolios is naturally convex and the boundary of this set gives a benchmark relative to which the efficiency of a portfolio can be measured. These models were developed by Markovitz (1959) and others; Sengupta (1991) and Sengupta & Park (1993) provide links between CAPM and non-parametric estimation of frontiers as in DEA.

Lovell (1993) and others have labelled DEA and similar approaches to efficiency measurement as *deterministic*, as if to suggest that DEA models have no statistical underpinnings. Typical DEA applications invariably present point estimates of inefficiency, with no measure and only slight or no discussion of uncertainty surrounding these estimates. Yet, since efficiency is measured relative to an estimate of the frontier, estimates of efficiency from DEA models are subject to uncertainty due to sampling variation.

Banker (1993) proved the consistency of the *output*-oriented efficiency scores in the case of a *single* output, but gave no indication of the achieved rate of convergence. Korostelev *et al.* (1995a, 1995b) also analysed the single-output problem and derived the speed of convergence for the estimated attainable production set (using either the Hausdorff metric or the Lebesgue measure of symmetric differences between the true and the estimated production sets), but not for the estimated measures of efficiency. The theory of statistical consistency in DEA models has been extended to the general multi-input *and* multi-output case for both input- and output-oriented efficiency measures in Kneip *et al.* (1998), where the rates of convergence are also derived.

Due to the complexity and multidimensional nature of DEA estimators, the sampling distributions of the estimators are not easily available. In the very particular case of one-input and one-output, Gijbels *et al.* (1999) derived the asymptotic sampling distribution of the DEA estimator, with an expression for its asymptotic bias and variance. However, in the more useful multi-output and multi-input case, the bootstrap methodology seems, so far, to be the only way to investigate sampling properties of DEA estimators. Simar & Wilson (1998) proposed a bootstrap strategy for analysing the sensitivity of the efficiency measures to sampling variation, providing confidence intervals and corrections for the bias inherent in the DEA procedure. The method has been applied to time dependence structures for estimating Malmquist indices (Simar & Wilson, 1999). However, the methodology in both these papers relies on some restrictive homogeneity assumptions on the distribution of efficiency among firms.

This paper presents an extension of the method to allow for more general DGPs; in particular, for less restrictive efficiency structures. Section 2 introduces some basic concepts and notation along with a brief introduction to DEA estimators. In Section 3 we analyse the underlying statistical model and explicitly state our assumptions regarding the DGP. In Section 4 we propose a general bootstrap procedure. An empirical illustration is given in Section 5, and conclusions are discussed in Section 6. Although the discussion throughout is in terms of the

Shephard (1970) input distance functions, it is straightforward to adapt the techniques presented below to output distance functions or to other non-parametric models (e.g. FDH models), including DEA measures of allocative efficiency and overall efficiency such as those described by Färe *et al.* (1985).

2 Frontier analysis and DEA estimators

2.1 The frontier model

Given column vectors of p inputs (denoted by $x \in \mathbb{R}^p_+$) and of q outputs (denoted by $y \in \mathbb{R}^q_+$), the activity of a productive organization can be described by means of the production set Ψ of physically attainable points (x, y) :

$$\Psi = \{(x, y) \in \mathbb{R}^{p+q}_+ \mid x \text{ can produce } y\} \tag{1}$$

This set can be described by its sections, either an input requirement set defined $\forall y \in \Psi$,

$$X(y) = \{x \in \mathbb{R}^p_+ \mid (x, y) \in \Psi\} \tag{2}$$

or an output correspondence set defined $\forall x \in \Psi$,

$$Y(x) = \{y \in \mathbb{R}^q_+ \mid (x, y) \in \Psi\} \tag{3}$$

Clearly,

$$x \in X(y) \Leftrightarrow y \in Y(x) \tag{4}$$

The relations between these two sets, along with standard assumptions one may reasonably make on them, are discussed in Section 9.1 of Shephard (1970). The convexity of $X(y)$ for all y (and of $Y(x)$ for all x) and the disposability of inputs and outputs are the most usual. The disposability assumptions correspond to monotonicity of the frontier; i.e.

$$y_1 \leq y_2 \Rightarrow X(y_2) \subseteq X(y_1) \tag{5}$$

$$x_1 \leq x_2 \Rightarrow Y(x_1) \subseteq Y(x_2) \tag{6}$$

where the inequality between the vectors is understood to be component-wise.

The Farrell efficiency boundaries are subsets of $X(y)$ and $Y(x)$, denoted by $\partial X(y)$ and $\partial Y(x)$, respectively:

$$\partial X(y) = \{x \mid x \in X(y), \theta x \notin X(y) \quad \forall 0 < \theta < 1\} \tag{7}$$

$$\partial Y(x) = \{y \mid y \in Y(x), \beta y \notin Y(x) \quad \forall \beta > 1\} \tag{8}$$

These may be used to define the Farrell input and output measures of efficiency (respectively) for a given point (x, y) in Ψ :

$$\theta(x, y) = \inf\{\theta \mid \theta x \in X(y)\} \tag{9}$$

$$\lambda(x, y) = \sup\{\lambda \mid \lambda y \in Y(x)\} \tag{10}$$

We will discuss only the input-oriented case to conserve space; the output-oriented case largely involves a straightforward translation of the notation in what follows.

Equivalently, Farrell’s input efficiency may be described by the Shephard input distance function

$$\delta(x, y) = (\theta(x, y))^{-1} = \sup \left\{ \delta \mid \frac{x}{\delta} \in X(y) \right\} \tag{11}$$

The input distance function δ gives a normalized measure of the distance from a point (x, y) to the frontier $\partial X(y)$, holding output and the direction of the input vector fixed. Since the direction of the input vector is held fixed, δ is said to be a *radial* measure; it gives the maximum feasible, proportionate reduction of inputs for a firm operating at $(x, y) \in \Psi$. Clearly, $\delta(x, y) \geq 1$ if and only if $x \in X(y)$; if $\delta(x, y) = 1$, then $(x, y) \in \partial X(y)$ and the point (x, y) is said to be *input-efficient*. It will be useful later to denote the efficient level of input, corresponding to the output level y and the input vector direction determined by x , as

$$x^{\theta}(y) = \frac{x}{\delta(x, y)} \tag{12}$$

Note that $x^{\theta}(y)$ is the intersection of $\partial X(y)$ and the ray $(\theta x, y), \theta \in [0, \infty]$.

Since radial distances are used, we will often refer to the polar coordinates of x defined by its modulus $\omega = \omega(x) \in \mathbb{R}_+$

where $\omega(x) = \sqrt{(x'x)}$, and its angle $\eta = \eta(x) \in \left[0, \frac{\pi}{2} \right]^{p-1}$

where, for $j = 1, \dots, p - 1, \eta_j = \arctan\left(\frac{x_{j+1}}{x_1}\right)$ if $x_1 > 0$ or $\eta_j = \frac{\pi}{2}$ if $x_1 = 0$.

Typically, $\Psi, X(y)$ and $\partial X(y)$ are unknown; hence, for a firm producing at (x, y) , $\delta(x, y)$ is also unknown. The DEA technique provides a consistent estimator of $\delta(x, y)$ from a random sample $\mathcal{X} = \{(x_i, y_i) \mid i = 1, \dots, n\}$.

2.2 The DEA approach

The DEA approach involves measurement of efficiency for a given firm at (x, y) , relative to the boundary of either the convex or conical hulls of the data $\mathcal{X} = \{(x_i, y_i), i = 1, \dots, n\}$ intersected with the free-disposal hull. The intersection of the convex and free-disposal hulls is given by³

$$\Psi = \{(x, y) \in \mathbb{R}^{p+q} \mid y \leq \sum_{i=1}^n \gamma_i y_i, x \geq \sum_{i=1}^n \gamma_i x_i, \sum_{i=1}^n \gamma_i = 1, \gamma_i \geq 0, i = 1, \dots, n\} \tag{13}$$

Replacing Ψ in (2) and (7) with Ψ yields estimates of the input requirement set and the input-efficient boundary for the output level y , respectively:

$$\hat{X}(y) = \{x \in \mathbb{R}^p \mid (x, y) \in \Psi\} \tag{14}$$

$$\partial \hat{X}(y) = \{x \in \mathbb{R}^p \mid x \in \hat{X}(y), \theta x \notin \hat{X}(y) \quad \forall 0 < \theta < 1\} \tag{15}$$

Finally, for any given point (x, y) , the estimator $\hat{\delta}(x, y)$ of $\delta(x, y)$ is obtained by substituting $\hat{X}(y)$ from (14) for $X(y)$ in (11), yielding

$$\hat{\delta}(x, y) = \sup\{\delta \mid x/\delta \in \hat{X}(y)\} \tag{16}$$

To make this operational, we rewrite (16) as a linear program:

$$(\hat{\delta}(x, y))^{-1} = \min\{\theta > 0 \mid y \leq \sum_{i=1}^n \gamma_i y_i, \theta x \geq \sum_{i=1}^n \gamma_i x_i, \sum_{i=1}^n \gamma_i = 1, \gamma_i \geq 0, i = 1, \dots, n\} \tag{17}$$

Analogous to equation (12) we have an estimator of the input-efficient level of inputs,

$$\hat{x}^\theta(y) = \frac{x}{\hat{\delta}(x, y)} \tag{18}$$

Note that $\Psi \subseteq \hat{\Psi}$, and so $\partial \hat{X}(y)$ is an upward-biased estimator of $\partial X(y)$, and $\hat{\delta}(x, y)$ is a downward-biased estimator of $\delta(x, y)$ so that

$$\hat{\delta}(x, y) \leq \delta(x, y) \quad \forall (x, y) \in \Psi \tag{19}$$

Moreover,

$$\hat{\delta}(x, y) \geq 1 \quad \forall (x, y) \in \hat{\Psi} \tag{20}$$

Thus, for each of the sample points in \mathcal{X} , we have

$$\hat{\delta}(x_i, y_i) \geq 1 \quad \forall i = 1, \dots, n \tag{21}$$

In order to perform statistical inference on the estimated input distances, we must analyse the behaviour of the difference $(\hat{\delta}(x, y) - \delta(x, y))$ for a given (x, y) by investigating its sampling distribution.

3 The statistical model and consistency of DEA

Our statistical model is defined through the following assumptions, which were used by Kneip *et al.* (1998). These assumptions serve to characterize the DGP, and augment Shephard’s (1970) assumptions, which are more concerned with the nature of the underlying production set.

Assumption A1: $\{(x_i, y_i), i = 1, \dots, n\}$ are i.i.d. random variables on the convex production set Ψ .

For the input-oriented case presented here, consider a point $(x, y) \in \hat{\Psi}$. Then for a given output level y , the corresponding input level x falls on the ray $(\delta x^\theta(y), y)$, $\delta \in [1, \infty]$; the deviation of (x, y) away from $\partial X(y)$ is assumed to result from technical inefficiency.

Assumption A2: Outputs y possess a density $f(\cdot)$ whose bounded support $\mathcal{Y} \subseteq \mathbb{R}^q_+$ is compact.

Due to the radial nature of the inefficiency, the conditional p.d.f. of x for a given y is more naturally introduced through the polar coordinates of x .

Assumption A3: For all $y \in \mathcal{Y}$, $\eta = (\eta_1, \dots, \eta_{p-1})$ has a conditional p.d.f. $f(\eta \mid y)$ on $[0, \pi/2]^{p-1}$ and conditional on (y, η) , the modulus ω has a density $f(\omega \mid y, \eta)$.

Note that, for a given (y, η) the efficient input-level $x^a(y)$ defined by (12) has modulus

$$\omega(x^a(y)) = \inf\{\omega \in \mathbb{R}^+ \mid f(\omega \mid y, \eta) > 0\} \tag{22}$$

The relation between $\omega(x)$ and the input distance function $\delta(x, y)$ is given by

$$\delta(x, y) = \frac{\omega(x)}{\omega(x^a(y))} \tag{23}$$

Assumption A3 induces, by (23), a conditional p.d.f. for $\delta(x, y)$ given (y, η) , namely $f(\delta \mid y, \eta)$, with support $[1, \infty)$.⁴

In order to achieve consistency in any non-parametric estimation of spatial boundaries, the DGP must ensure that points will be observed near the frontier when n is sufficiently large. This imposes, in our case, an assumption on the conditional p.d.f. of ω given (y, η) .

Assumption A4: For all $y \in \mathcal{Y}$ and all $\eta \in [0, \pi/2]^{p-1}$, there exist constants $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $\forall \omega \in [\omega(x^a(y)), \omega(x^a(y)) + \varepsilon_2], f(\omega \mid y, \eta) \geq \varepsilon_1$.

Again, this implies by (23), that $f(\delta \mid y, \eta) \geq \varepsilon_1 \forall \delta \in [1, 1 + \varepsilon_2]$.

In addition, Kneip *et al.* (1998) make two additional assumptions regarding the smoothness of the frontier in order to derive the rates of convergence; these may be written in terms of $\delta(x, y)$. For sake of simplicity, we assume the following here.⁵

Assumption A5: The distance function $\delta(x, y)$ is differentiable in both its arguments.

Consider now a fixed point (x, y) and a sample $\mathcal{X} = \{(x_i, y_i) \mid i = 1, \dots, n\}$ generated by the DGP defined by A1–A5 in terms of polar coordinates, where the modulus is defined through the distance function with respect to the frontier by equation (23). An observation $(x_i, y_i) \in \mathbb{R}^{p+q}$ has as its polar coordinate representation

$$(y_i, \eta_i, \delta_i) \in \mathcal{A} = \mathbb{R}^q_+ \times [0, \pi/2]^{p-1} \times [1, \infty) \tag{24}$$

where

$$\delta_i = \frac{\omega(x_i)}{\omega(x^a(y_i))} \tag{25}$$

The DGP is completely defined through the density of (y_i, η_i, δ_i) on \mathcal{A} :

$$f(y_i, \eta_i, \delta_i) = f(\delta_i \mid y_i, \eta_i) f(\eta_i \mid y_i) f(y_i) \tag{26}$$

Kneip *et al.* (1998) proved that for a fixed point (x, y) ,

$$\hat{\delta}(x, y) - \delta(x, y) = \mathcal{O}_P(n^{-2/(p+q+1)}) \tag{27}$$

Thus, $\hat{\delta}(x, y)$ is a consistent estimator of $\delta(x, y)$, but the rate of convergence is low; furthermore, by construction, $\hat{\delta}(x, y)$ is downward-biased. Unfortunately, very few results exist on the sampling distribution of $\hat{\delta}(x, y)$. Gijbels *et al.* (1999) derived the asymptotic distribution of $\hat{\delta}(x, y)$ in the special case of one input and one output ($p = q = 1$) along with an analytic expression for its large sample bias and variance. This would allow one to construct a bias-corrected estimator and confidence intervals for this special case. Unfortunately, in the more general multivariate setting, the radial nature of the distance function and the complexity of the estimated frontier complicates the derivations; so far, the bootstrap appears

to offer the only way to approximate this asymptotic distribution. The next section proposes a general methodology for bootstrapping the distribution of $\hat{\delta}(x, y) - \delta(x, y)$.

4 The bootstrap

4.1 The principles

Let \mathcal{P} denote the DGP defined by assumptions A1–A5, from which the random sample $\mathcal{B} = \{(x_i, y_i) \mid i = 1, \dots, n\}$ is obtained. Consider again a fixed point (x, y) . From (16) we can obtain an estimate $\hat{\delta}(x, y)$ of $\delta(x, y)$. Typically, we would choose this fixed point to correspond to one of the points in \mathcal{B} , but this is not necessary.

Given a consistent estimator $\hat{\mathcal{P}}$ of \mathcal{P} estimated from the data \mathcal{B} , consider now a new data set $\mathcal{B}^* = \{(x_i^*, y_i^*), i = 1, \dots, n\}$ drawn from $\hat{\mathcal{P}}$. The convex hull of \mathcal{B}^* gives an estimator Ψ^* of Ψ , which from the perspective of \mathcal{B}^* , is the true set of possible values in \mathcal{B}^* , and which in our original setting was an estimate of Ψ defined in (1). Specifically, we have

$$\Psi^* = \{(x, y) \in \mathbb{R}^{p+q} \mid y \leq \sum_{i=1}^n \gamma_i y_i^*, x \geq \sum_{i=1}^n \gamma_i x_i^*, \sum_{i=1}^n \gamma_i = 1, \gamma_i \geq 0, i = 1, \dots, n\} \tag{28}$$

Analogous to (14)–(15), corresponding to Ψ^* we have

$$\hat{X}^*(y) = \{x \in \mathbb{R}^p \mid (x, y) \in \Psi^*\} \tag{29}$$

and

$$\partial \hat{X}^*(y) = \{x \in \mathbb{R}^p \mid x \in \hat{X}^*(y), \theta x \notin \hat{X}^*(y) \forall 0 < \theta < 1\} \tag{30}$$

Replacing $\hat{X}(y)$ with $\hat{X}^*(y)$ in (16), we have

$$\hat{\delta}^*(x, y) = \sup\{\delta \mid x/\delta \in \hat{X}^*(y)\} \tag{31}$$

which may be evaluated by solving the linear program

$$\begin{aligned} (\hat{\delta}^*(x, y))^{-1} &= \min\{\theta > 0 \mid y \leq \sum_{i=1}^n \gamma_i y_i^* \theta x \geq \sum_{i=1}^n \gamma_i x_i^*, \\ &\sum_{i=1}^n \gamma_i = 1, \gamma_i \geq 0, i = 1, \dots, n\} \end{aligned} \tag{32}$$

Note that conditional on \mathcal{B} , the sampling distribution of $\hat{\delta}^*(x, y)$ is (in principle) completely known since $\hat{\mathcal{P}}$ is known, although it may be difficult to compute analytically. However, the sampling distributions are easily approximated by Monte Carlo methods. Using $\hat{\mathcal{P}}$ to generate B samples $\mathcal{B}_b^*, b = 1, \dots, B$, yields a set of pseudo estimates $\hat{\delta}_b^*(x, y), b = 1, \dots, B$. The empirical density function of these bootstrap values gives a Monte Carlo approximation of the sampling distribution of $\hat{\delta}^*(x, y)$, conditional on \mathcal{P} .

The bootstrap method introduced by Efron (1979) (see also Efron, 1982, or Efron & Tibshirani, 1993) is based on the idea that if $\hat{\mathcal{P}}$ is a consistent estimator of \mathcal{P} , the known bootstrap distributions will mimic the original unknown sampling distributions of the estimators of interest. More specifically,

$$(\hat{\delta}^*(x, y) - \hat{\delta}(x, y)) \mid \hat{\mathcal{P}} \overset{\text{approx}}{\sim} (\hat{\delta}(x, y) - \delta(x, y)) \mid \mathcal{P} \tag{33}$$

A naive bootstrap would consist of sampling the pairs (x_i^*, y_i^*) with replacement from the original pairs in \mathcal{X} . In boundary problems such as this, the naive bootstrap yields inconsistent estimates (see Bickel & Freedman, 1981 or Beran & Ducharme, 1991 for discussions of this problem in the context of estimating the support of a univariate density; see Gijbels *et al.*, 1999 for a two-dimensional case). To illustrate, consider a fixed point (x, y) . With non-zero probability, the naive bootstrap estimate $\hat{\delta}^*(x, y)$ will equal $\hat{\delta}(x, y)$. This can be proven as follows. In \mathbb{R}^{p+q} , $(\hat{X}(y), y)$ is a polyhedron defined by all the dominating facets of (x_i, y_i) , $i = 1, \dots, n$, where the dominating facets are given by $(\partial\hat{X}(y), y)$ and are determined by $(p+q)$ observed efficient points. The probability that the bootstrap sample \mathcal{X}^* contains the original dominating facet of (x, y) is then

$$\sum_{j=0}^{p+q} \binom{p+q}{j} (-1)^j \left(1 - \frac{j}{n}\right)^n,$$

which has the limit $(1 - e^{-1})^{p+q} > 0$ as $n \rightarrow \infty$. Consequently, in the naive bootstrap $\hat{\delta}^*(x, y) = \hat{\delta}(x, y)$ with non-zero probability; moreover, this problem does not go away as $n \rightarrow \infty$, and so the naive bootstrap is inconsistent.

One way to overcome this difficulty is to use a smoothed bootstrap, i.e. to draw i.i.d. bootstrap samples (x_i^*, y_i^*) , $i = 1, \dots, n$, from a density $\hat{f}(x, y)$ on Ψ where $\hat{f}(x, y)$ is a smooth, consistent estimator of the joint density of (x, y) on Ψ . In terms of the polar coordinates introduced in equations (24)–(25), this is equivalent to estimating the density $f(y, \eta, \delta)$ and drawing bootstrap samples $(y_i^*, \eta_i^*, \delta_i^*)$ from the estimated density. Unfortunately, $f(y, \eta, \delta)$ has bounded support as described by (24), and ordinary kernel density estimates are known to be biased and inconsistent near boundaries. Scott (1992) proposes the use of boundary kernels to deal with this problem in the univariate case, but it is not clear how this method could be extended to higher-dimensional spaces such as ours. Instead, we adopt the reflection method proposed by Silverman (1986) to estimate $f(y, \eta, \delta)$.⁶ Since the frontier is unknown, δ is not directly observable in our sample. Therefore, we will use the estimator of $f(y, \eta, \hat{\delta})$ from the set of points $\{(y_i, \eta_i, \hat{\delta}_i), i = 1, \dots, n\}$, where $\hat{\delta}_i$ is the consistent DEA estimator of δ_i .

To estimate $f(y, \eta, \hat{\delta})$ consistently, recall from equation (24) that $\hat{\delta} \in [1, \infty]$. As noted earlier, in a given sample of size n we will likely observe many values of $\hat{\delta}_i = 1$. To ensure consistency of our estimator \hat{f} of $f(y, \eta, \hat{\delta})$, we reflect each point $(y_i, \eta_i, \hat{\delta}_i)$ through the boundary at unity for $\hat{\delta}_i$.⁷ Let \mathcal{Z} denote the $n \times (p+q)$ matrix

$$\mathcal{Z} = [y_i \quad \eta_i \quad \hat{\delta}_i] \tag{34}$$

so that the i th row of \mathcal{Z} contains the observation for the i th firm expressed in polar coordinates. Then the matrix of points reflected about the boundary $\hat{\delta}_i = 1$ is

$$\mathcal{Z}_R = [y_i \quad \eta_i \quad 2 - \hat{\delta}_i] \tag{35}$$

Let z_i denote the i th row of \mathcal{Z} , and z_{Ri} denote the i th row of \mathcal{Z}_R . The ‘augmented’ set of points is now given by the $2n \times (p+q)$ matrix

$$\mathcal{Z} = \begin{bmatrix} \mathcal{Z} \\ \mathcal{Z}_R \end{bmatrix} \tag{36}$$

We wish to estimate the unknown density of the $2n$ unbounded points represented by \mathcal{L} . We use a multivariate Gaussian kernel function scaled to have the same shape as the cloud of points in $(p + q)$ -space represented in \mathcal{L} . Specifically, let Σ_1 be an estimate of the covariance matrix of \mathcal{L} , which can be written in partitioned form as

$$\Sigma_1 = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \tag{37}$$

where S_{11} is $(p + q - 1) \times (p + q - 1)$, $S_{12} = S'_{21}$ is $(p + q - 1) \times 1$, and S_{22} is scalar.⁸ Accordingly, the corresponding estimate of the covariance matrix of the reflected data is

$$\Sigma_2 = \begin{bmatrix} S_{11} & -S_{12} \\ -S_{21} & S_{22} \end{bmatrix} \tag{38}$$

Finally, let $K_\ell(\cdot)$ be a $(p + q)$ -variate Gaussian density with zero mean and shape Σ_ℓ for $\ell = 1, 2$; i.e.

$$K_\ell(x) = (2\pi)^{-(p+q)/2} (\det(\Sigma_\ell))^{-1/2} \exp\left(-\frac{1}{2}x'\Sigma_\ell^{-1}x\right), \quad \ell = 1, 2 \tag{39}$$

A consistent kernel estimator of the density of $z = (y, \eta, \delta)$ over the $2n$ observations in \mathcal{L} is given by

$$\tilde{f}_h(z) = \frac{1}{2nh^{(p+q)}} \sum_{i=1}^n \left[K_1\left(\frac{z - z_i}{h}\right) + K_2\left(\frac{z - z_{Ri}}{h}\right) \right] \tag{40}$$

where h is the bandwidth.

Then, a consistent estimator \hat{f}_h of f , with bounded support \mathcal{A} defined in equation (24), is obtained by setting

$$\hat{f}_h(z) = \begin{cases} 2\tilde{f}_h(z) & \text{if } z \in \mathcal{A} \\ 0 & \text{otherwise} \end{cases} \tag{41}$$

The remaining question regards the proper choice of the bandwidth h . Consistency of $\hat{f}_h(z)$, and hence $\hat{f}_h(z)$, requires $h \rightarrow 0$ as $n \rightarrow \infty$, but not too quickly; in particular, we need $h = o(n^{-1/(p+q+4)})$. One possibility is to use the normal reference rule (e.g. see Scott, 1992), which assigns

$$\hat{h} = \left(\frac{4}{p + q + 2}\right)^{1/(p+q+4)} n^{-1/(p+q+4)} \tag{42}$$

This bandwidth minimizes the asymptotic mean integrated square error (AMISE) when the data are normally distributed, and have been prewhitened to have unit variance and zero covariance. Unfortunately, the data in \mathcal{L} will almost certainly be highly non-normal, and so using equation (42) to determine the bandwidth will almost certainly fail to minimize the AMISE. A better approach is to use the data to choose a value \hat{h} that minimizes an estimate of mean integrated square error (MISE); we provide such a method in the appendix.

To generate a random sample of size n from $\hat{f}_h(z)$ in equation (41) is very easy;

an algorithm is given in the next section. This provides pseudo observations $\{(y_i^*, \eta_i^*, \delta_i^*) \in \mathcal{A}, i = 1, \dots, n\}$, which can be translated back to the rectangular coordinates $\{(x_i^*, y_i^*), i = 1, \dots, n\}$. To demonstrate this, consider the draw $(y_i^*, \eta_i^*, \delta_i^*)$ from $\hat{f}_h(z)$ in equation (41). Given the output level y_i^* and the angle η_i^* , we must determine the location of the original estimated frontier $\widehat{\partial X}(y_i^*)$ in order to compute x_i^* . This is accomplished by solving an additional linear program.

Let \tilde{x} be any point in the x -space on the ray with angle η_i^* (for instance, take $\tilde{x}_1 = 1$ and $\tilde{x}_{j+1} = \tan(\eta_j^*)$ for $j = 1, \dots, p - 1$). For this point (\tilde{x}, y_i^*) , use equation (17) to compute $\hat{\delta}(\tilde{x}, y_i^*)$. Given the output level y_i^* and the angle η_i^* contained in our draw from $\hat{f}_h(z)$, we can replace $\hat{\delta}(x, y)$ with $\hat{\delta}(\tilde{x}, y_i^*)$ in equation (18) and the x appearing in the numerator of (18) with \tilde{x} to obtain

$$\hat{x}^{a*}(y_i^*) = \frac{\tilde{x}}{\hat{\delta}(\tilde{x}, y_i^*)} \tag{43}$$

which gives the (true) input-efficient level of inputs in the bootstrap world.⁹ In other words, $\hat{x}^{a*}(y_i^*)$ gives the value of the input vector with angle η_i^* , corresponding to output level y_i^* , and lying on the estimated frontier $\widehat{\partial X}(y_i^*)$.

Finally we can define the Cartesian coordinates x_i^* by computing

$$x_i^* = \delta_i^* \hat{x}^{a*}(y_i^*) \tag{44}$$

Thus, x_i^* is formed by taking a random deviation away from the input vector $\hat{x}^{a*}(y_i^*)$ lying on the estimated frontier $\widehat{\partial X}(y_i^*)$, consistent with our assumptions A1–A5 regarding the underlying DGP.

Now, having constructed the pseudo-sample $\mathcal{B}^* = \{(x_i^*, y_i^*), i = 1, \dots, n\}$ we can compute the bootstrap value $\hat{\delta}^*(x, y)$ by solving the linear program (32).¹⁰ To clarify, we detail the various steps we have described above in algorithmic form in the next section.

4.2 The algorithm

As seen from the previous section, the critical element of our bootstrap procedure involves generation of pseudo-data $(y_i^*, \eta_i^*, \delta_i^*) \in \mathcal{A}$ from the density estimate $\hat{f}_h(\cdot)$ given by equation (41). Our procedure consists of eleven steps, as detailed below.

- Step 1. From the original data set \mathcal{X} , compute $\hat{\delta}_i = \hat{\delta}(x_i, y_i) \forall i = 1, \dots, n$ using equation (16).
- Step 2. Translate the data into polar coordinates: $(y_i, \eta_i, \hat{\delta}_i) \forall i = 1, \dots, n$, and form the augmented matrix \mathcal{L} as in (34)–(36).
- Step 3. Compute the estimated covariance matrices Σ_1, Σ_2 as in (37)–(38), and the lower triangular matrices L_1 and L_2 such that $\Sigma_1 = L_1 L_1'$ and $\Sigma_2 = L_2 L_2'$ via the Cholesky decomposition.
- Step 4. Choose an appropriate bandwidth h as described in the appendix using the information in \mathcal{L}, Σ_1 and Σ_2 .
- Step 5. Draw n rows randomly, with replacement from the augmented matrix \mathcal{L} and denote the result by the $n \times (p + q)$ matrix \mathcal{L}^* ; compute \bar{z}^* , the $1 \times (p + q)$ row vector containing the means of each column of \mathcal{L}^* .
- Step 6. Use a random number generator to generate an $n \times (p + q)$ matrix ε of i.i.d. standard normal pseudo-random variates; let $\varepsilon_i \cdot$ denote the i th row of this matrix. Then compute the $n \times (p + q)$ matrix ε^* with the i th row ε_i^* given by

$$\varepsilon_i^* = \varepsilon_i \cdot L' \tag{45}$$

so that $\varepsilon_i^* \sim N_{p+q}(0, \Sigma_\ell)$ where $\ell = 1$ if the i th row of \mathcal{Z}^* was drawn from rows $1, \dots, n$ of \mathcal{Z} , or $\ell = 2$ if the i th row of \mathcal{Z}^* was drawn from rows $(n + 1), \dots, 2n$ of \mathcal{Z} .

Step 7. Compute the $n \times (p + q)$ matrix

$$\Gamma = (1 + h^2)^{-1/2} (M\mathcal{Z}^* + h\varepsilon^*) + i_n \otimes \bar{z}^* \tag{46}$$

where $M = I_n - (1/n) i_n i_n'$ is the usual $n \times n$ centring matrix with I_n denoting an identity matrix of order n , i_n an $n \times 1$ vector of ones, and \otimes denotes the Kronecker product.¹¹

Step 8. Partition Γ so that $\Gamma = [\gamma_{i1}, \gamma_{i2}, \gamma_{i3}]$, where $\gamma_{i1} \in \mathbb{R}^q$, $\gamma_{i2} \in [0, \pi/2]^{p-1}$, and $\gamma_{i3} \in (-\infty, \infty) \forall i = 1, \dots, n$. Define the $n \times (p + q)$ matrix of bootstrap pseudo-data \mathcal{Z}^* such that the i th row z_i^* of \mathcal{Z}^* is given by

$$z_i^* = \begin{cases} (\gamma_{i1}, \gamma_{i2}, \gamma_{i3}) & \text{if } \gamma_{i3} \geq 1 \\ (\gamma_{i1}, \gamma_{i2}, 2 - \gamma_{i3}) & \text{otherwise.} \end{cases} \tag{47}$$

Thus, values $\gamma_{i3} < 1$ are reflected back across the boundary $\hat{\delta}_i = 1$, ensuring that $z_i^* \in \mathcal{A} \forall i = 1, \dots, n$.

Step 9. Translate the polar coordinates in \mathcal{Z}^* to Cartesian coordinates using equations (43) and (44); note that this requires solving n linear programs similar to (17), as discussed in the previous section and in footnote 9. This yields the bootstrap sample $\mathcal{Z}^* = \{(x_i^*, y_i^*), i = 1, \dots, n\}$. For observations where this results in linear programs with infeasible solutions, repeat Steps 5–8.¹²

Step 10. For the given point (x, y) , compute $\hat{\delta}^*(x, y)$ by solving the DEA program (32).

Step 11. Repeat Steps 5–10 B times to obtain a set of bootstrap estimates $\{\hat{\delta}_b^*(x, y) | b = 1, \dots, B\}$.

Step 11 amounts to a Monte-Carlo simulation. Unfortunately, the computational burden is not negligible, particularly for large n . Step 1 requires that n linear programs be solved; Step 9 requires solution of an additional n linear programs, and Step 10 requires solution of a linear program for each point (x, y) being considered. Moreover, Steps 9 and 10 are repeated B times. If one were to consider the efficiency of each of the n points in the original sample \mathcal{Z} , the total number of linear programs for such an application would be $n(2B + 1)$, which is of order $O(nB)$. For very large values of n , however, using the bootstrap to estimate confidence intervals, variance, bias, etc, for every observation might yield an overwhelming amount of information. One might gain more insight by identifying groups of similar observations, perhaps through a cluster analysis, and then bootstrapping from a representative point for each of these groups, perhaps located at the centre of each group. Alternatively, the researcher might be interested in the performance of only a few firms within a large sample, in which case the computational costs will be much lower than indicated above.

Once the bootstrap values $\hat{\delta}_b^*(x, y)$, $b = 1, \dots, B$ have been obtained, we can compute the bootstrap bias estimate for the original estimator $\hat{\delta}(x, y)$ as

$$\widehat{\text{bias}}_B[\hat{\delta}(x, y)] = B^{-1} \sum_{b=1}^B \hat{\delta}_b^*(x, y) - \hat{\delta}(x, y) \tag{48}$$

which is merely the empirical bootstrap analogue of the definition of bias, i.e. $E[\hat{\delta}(x, y)] - \delta(x, y)$. Then a bias-corrected estimator of $\delta(x, y)$ can be computed as

$$\begin{aligned} \hat{\delta}(x, y) &= \hat{\delta}(x, y) - \text{bias}_B[\hat{\delta}(x, y)] \\ &= 2\hat{\delta}(x, y) - B^{-1} \sum_{b=1}^B \hat{\delta}_b^*(x, y) \end{aligned} \tag{49}$$

Of course, one should avoid using $\hat{\delta}(x, y)$ if it has a higher mean-square error than $\hat{\delta}(x, y)$. The variance of the summation term on the right-hand side (r.h.s.) of the second line of equation (49) can be made arbitrarily small by increasing B . Yet, even if $B \rightarrow \infty$, the bias-corrected estimator $\hat{\delta}(x, y)$ will have variance equal to four times that of the original, uncorrected estimator, $\hat{\delta}(x, y)$. The sample variance of the bootstrap values $\hat{\delta}_b^*(x, y)$ gives an estimate of the variance of $\hat{\delta}(x, y)$; call this estimate $\hat{\sigma}^2$. Then the estimated mean square error of $\hat{\delta}(x, y)$ is $4\hat{\sigma}^2$ if $B \rightarrow \infty$, and $[\hat{\sigma}^2 + (\text{bias}_B[\hat{\delta}(x, y)])^2]$ for $\hat{\delta}(x, y)$. Hence, the bias-correction should not be used unless

$$\hat{\sigma}^2 < \frac{1}{3} (\widehat{\text{bias}}_B[\hat{\delta}(x, y)])^2 \tag{50}$$

Moreover, since equation (50) contains only estimates of variance and bias rather than the true values, caution would indicate that the bias correction in equation (49) be used only if the ratio $\frac{1}{3} (\widehat{\text{bias}}_B[\hat{\delta}(x, y)])^2 / \hat{\sigma}^2$ is well above unity.

The bootstrap values $\hat{\delta}_b^*(x, y)$ can also be used to construct confidence intervals for the true value $\delta(x, y)$, along the lines of Simar & Wilson (1999). Recall that the idea behind the bootstrap is to approximate the unknown distribution of $(\hat{\delta}(x, y) - \delta(x, y))$ by the distribution of $(\hat{\delta}^*(x, y) - \hat{\delta}(x, y))$ conditioned on the original data \mathcal{X} . The bootstrap values $\hat{\delta}_b^*(x, y), b = 1, \dots, B$, together with the original estimate $\hat{\delta}(x, y)$, can be used to obtain an empirical approximation to the second distribution.

If we knew the distribution of $(\hat{\delta}(x, y) - \delta(x, y))$, then it would be trivial to find values a_α, b_α such that

$$\text{Prob}(-b_\alpha \leq \hat{\delta}(x, y) - \delta(x, y) \leq -a_\alpha) = 1 - \alpha \tag{51}$$

for some small value of α , say 0.10 or 0.05. Since we do not know this distribution, we can use the bootstrap values to find values a_α^*, b_α^* such that the statement

$$\text{Prob}(-b_\alpha^* \leq \hat{\delta}^*(x, y) - \hat{\delta}(x, y) \leq -a_\alpha^* | \mathcal{X}) = 1 - \alpha \tag{52}$$

is true with high probability.¹³ Mechanically, this involves sorting the values

$$(\hat{\delta}^*(x, y) - \hat{\delta}(x, y)), \quad b = 1, \dots, B$$

by algebraic value, deleting $(\alpha/2 \times 100)$ -percent of the elements at either end of this sorted array, and then setting $-b_\alpha^*$ and $-a_\alpha^*$ equal to the endpoints of the resulting (sorted) array, with $a_\alpha^* \leq b_\alpha^*$. When we say that equation (52) is ‘true with high probability’, we mean that this can be made so by making the number of bootstrap replications, B , large enough; as $B \rightarrow \infty$, the probability that equation (52) is true approaches 1.

Since

$$[\hat{\delta}(x, y) - \delta(x, y)] | \mathcal{P} \stackrel{\text{approx}}{\sim} [\hat{\delta}^*(x, y) - \hat{\delta}(x, y)] | \mathcal{P} \tag{53}$$

we obtain the bootstrap approximation

$$\text{Prob}(-b_a^* \leq \hat{\delta}(x, y) - \delta(x, y) \leq -a_a^* | \mathcal{B}) \approx 1 - \alpha \tag{54}$$

by substituting a_a^* and b_a^* for a_a and b_a in equation (51) and noting the conditioning in (52). Rearranging the terms in parentheses in equation (54) yields an estimated $(1 - \alpha)$ -percent confidence interval¹⁴

$$\hat{\delta}(x, y) + a_a^* \leq \delta(x, y) \leq \hat{\delta}(x, y) + b_a^* \tag{55}$$

As a final note, we remark that, by construction, $\delta(x, y) \geq \hat{\delta}(x, y) \geq 1$. Similarly, we have $\hat{\delta}(x, y) \geq \hat{\delta}_b^*(x, y) \forall b = 1, \dots, B$. Necessarily then, $0 \leq a_a^* \leq b_a^*$ and so the estimated confidence interval in equation (55) will only include the original estimate $\hat{\delta}(x, y)$ on its lower boundary if $a_a^* = 0$; more typically, $\hat{\delta}(x, y)$ will fall outside the estimated confidence interval. This indeed should not be surprising, since $\hat{\delta}(x, y)$ is necessarily a downward-biased estimator of $\delta(x, y)$. This result merely indicates that $\hat{\delta}(x, y)$ and the bootstrap values $\hat{\delta}_b^*(x, y)$ provide enough information about the true value $\delta(x, y)$ to suggest that $\delta(x, y)$ lies in some range above $\hat{\delta}(x, y)$. In any case given an estimated confidence interval of size $1 - \alpha$ as in (55), it is not clear why the applied researcher would still be concerned with a point estimate.

In the next section, we provide an empirical example to illustrate the algorithm proposed above.

5 Empirical example

As an illustration of our methods, we examine data reported by Charnes *et al.* (1981) on Program Follow Through (PFT), an experimental education program administered in US schools. In particular, Charnes *et al.* displayed data on five inputs and three outputs for 49 schools that implemented PFT, and 21 schools that did not, for a total of 70 observations.¹⁵

Step (4) in our bootstrap algorithm involves choosing the bandwidth parameter h as described in the Appendix. For the Charnes *et al.* data, the normal reference rule in equation (42) gives a bandwidth of 0.65025. Using the least-squares cross-validation procedure described in the Appendix, we chose a bandwidth of 0.87946, which we used in our bootstrap estimation. The cross-validation procedure required just over 34 minutes elapsed time on a SUN Sparcstation 20.

Table 1 shows our original distance function estimates, $\hat{\delta}(x_i, y_i)$, the corresponding bias estimates, and the estimated standard deviations across bootstrap replications for each observation ($\hat{\sigma}_i$). The fifth column of the table gives the ratio $r_i = \frac{1}{3} (\text{bias}_B[\hat{\delta}(x_i, y_i)])^2 / \hat{\sigma}^2$ discussed in Section 4, which may be used to assess whether the bias correction might increase mean-square error. The sixth column shows the bias-corrected distance function estimate, while the last two columns give the estimated 95% confidence bounds. Results were produced using $B = 2000$ bootstrap replications using Fortran code written by the authors, requiring approximately 47 minutes elapsed computation time on a SUN Sparcstation 20.

Bootstrap estimates could not be determined for observations 44 and 59. These observations were identified as outliers in Wilson (1993); indeed, on every bootstrap replication, these observations lie above the bootstrap frontier, making solutions to the linear program (32) infeasible.

Of 70 observations in the data, 27 appear ostensibly efficient as indicated by distance function estimates $\hat{\delta}(x_i, y_i) = 1$. The remaining 43 observations have distance function estimates ranging from 1.0006 to 1.2611. The average distance

TABLE 1. Estimates from Charnes *et al.* (1981) Data

i	$\hat{\delta}(x_i, y_i)$	$\widehat{\text{bias}}_B$ ($\hat{\delta}(x_i, y_i)$)	$\hat{\sigma}$	r_i	$\hat{\delta}(x_i, y_i)$	95% <i>conf. int.</i>	
1	1.0393	-0.3660	0.09983	4.4800	1.4053	1.2979	1.6658
2	1.1098	-0.1298	0.02755	7.4010	1.2396	1.1920	1.2988
3	1.0697	-0.1044	0.02917	4.2700	1.1741	1.1244	1.2364
4	1.1091	-0.0796	0.01493	9.4590	1.1887	1.1626	1.2199
5	1.0000	-0.4555	0.04279	37.7600	1.4555	1.3849	1.5520
6	1.0990	-0.2148	0.02752	20.3100	1.3138	1.2668	1.3727
7	1.1218	-0.0739	0.01246	11.7200	1.1957	1.1739	1.2222
8	1.1049	-0.2146	0.09555	1.6820	1.3195	1.1862	1.5587
9	1.1647	-0.0583	0.01623	4.3030	1.2230	1.1940	1.2586
10	1.0629	-0.1912	0.07785	2.0110	1.2541	1.1427	1.4365
11	1.0000	-0.1852	0.04536	5.5570	1.1852	1.1166	1.2902
12	1.0000	-0.2263	0.05568	5.5040	1.2263	1.1295	1.3492
13	1.1596	-0.0577	0.01222	7.4360	1.2173	1.1964	1.2446
14	1.0104	-0.3282	0.03583	27.9700	1.3386	1.2707	1.4102
15	1.0000	-0.5242	0.06214	23.7200	1.5242	1.3992	1.6348
16	1.0524	-0.1876	0.12940	0.6999	1.2400	1.0903	1.5755
17	1.0000	-0.4295	0.03129	62.8200	1.4295	1.3758	1.4965
18	1.0000	-0.1628	0.03022	9.6760	1.1628	1.1086	1.2273
19	1.0498	-0.1136	0.06259	1.0980	1.1634	1.0899	1.3208
20	1.0000	-0.2803	0.04940	10.7300	1.2803	1.1865	1.3745
21	1.0000	-0.2745	0.06869	5.3220	1.2745	1.1762	1.4426
22	1.0000	-0.1517	0.02412	13.1900	1.1517	1.1055	1.2005
23	1.0258	-0.0818	0.08330	0.3216	1.1076	1.0452	1.3559
24	1.0000	-0.2797	0.04325	13.9400	1.2797	1.1926	1.3607
25	1.0217	-0.0472	0.01016	7.2050	1.0689	1.0511	1.0905
26	1.0609	-0.0586	0.01968	2.9580	1.1195	1.0908	1.1605
27	1.0000	-0.2055	0.03653	10.5500	1.2055	1.1379	1.2809
28	1.0097	-0.2101	0.04340	7.8100	1.2198	1.1374	1.3101
29	1.1321	-0.3840	0.05839	14.4200	1.5161	1.4065	1.6416
30	1.1193	-0.1371	0.03092	6.5520	1.2564	1.2034	1.3240
31	1.1949	-0.1108	0.02143	8.9130	1.3057	1.2676	1.3511
32	1.0000	-0.3943	0.05993	14.4300	1.3943	1.2755	1.5158
33	1.0503	-0.0995	0.09790	0.3444	1.1498	1.0683	1.4155
34	1.1640	-0.0556	0.01957	2.6870	1.2196	1.1903	1.2643
35	1.0000	-0.1989	0.06898	2.7720	1.1989	1.1276	1.3640
36	1.2611	-0.2452	0.02787	25.8100	1.5063	1.4610	1.5666
37	1.1914	-0.2136	0.02556	23.2600	1.4050	1.3605	1.4566
38	1.0000	-0.4555	0.01724	232.7000	1.4555	1.4302	1.4963
39	1.0621	-0.1587	0.02787	10.8000	1.2208	1.1737	1.2822
40	1.0528	-0.1237	0.02077	11.8100	1.1765	1.1349	1.2188
41	1.0500	-0.0419	0.00911	7.0520	1.0919	1.0761	1.1117
42	1.0491	-0.2513	0.04766	9.2660	1.3004	1.2103	1.3987
43	1.1564	-0.1302	0.03012	6.2300	1.2866	1.2341	1.3534
44	1.0000	-	-	-	-	-	-
45	1.0000	-0.4512	0.01796	210.4000	1.4512	1.4238	1.4932
46	1.0954	-0.0453	0.02795	0.8764	1.1407	1.1108	1.2011
47	1.0000	-0.2209	0.02121	36.1400	1.2209	1.1867	1.2676
48	1.0000	-0.5234	0.08306	13.2400	1.5234	1.3451	1.6637
49	1.0000	-0.2589	0.06305	5.6180	1.2589	1.1514	1.3912
50	1.0431	-0.1373	0.04753	2.7820	1.1804	1.0989	1.2882
51	1.0871	-0.2310	0.04088	10.6500	1.3181	1.2381	1.4025
52	1.0000	-0.2859	0.06394	6.6650	1.2859	1.1805	1.4317
53	1.1498	-0.1050	0.02130	8.0960	1.2548	1.2169	1.2999
54	1.0000	-0.4867	0.10150	7.6630	1.4867	1.3971	1.7701
55	1.0006	-0.1763	0.01990	26.1400	1.1769	1.1430	1.2199

TABLE 1.—(Continued)

i	$\hat{\delta}(x_i, y_i)$	$\widehat{\text{bias}}_B$ ($\hat{\delta}(x_i, y_i)$)	$\hat{\sigma}$	r_i	$\hat{\delta}(x_i, y_i)$	95% <i>conf. int.</i>	
56	1.0000	-0.4798	0.07058	15.4000	1.4798	1.3393	1.6047
57	1.0788	-0.1390	0.02660	9.1010	1.2178	1.1676	1.2718
58	1.0000	-0.3884	0.02700	68.9800	1.3884	1.3357	1.4410
59	1.0000	-	-	-	-	-	-
60	1.0199	-0.1105	0.02123	9.0370	1.1304	1.0908	1.1743
61	1.1202	-0.4258	0.05587	19.3700	1.5460	1.4464	1.6654
62	1.0000	-0.5512	0.02615	148.0000	1.5512	1.5145	1.6131
63	1.0379	-0.2103	0.03244	14.0100	1.2482	1.1898	1.3135
64	1.0749	-0.0490	0.01276	4.9210	1.1239	1.1009	1.1507
65	1.0252	-0.2276	0.03026	18.8700	1.2528	1.1980	1.3150
66	1.0687	-0.0516	0.01264	5.5470	1.1203	1.0971	1.1472
67	1.0568	-0.0481	0.01353	4.2090	1.1049	1.0808	1.1353
68	1.0000	-0.3085	0.07885	5.1030	1.3085	1.1707	1.4780
69	1.0000	-0.5641	0.03237	101.2000	1.5641	1.5084	1.6354
70	1.0373	-0.1484	0.03230	7.0310	1.1857	1.1314	1.2540

function estimate over the PFT schools (observations 1–49) is 1.0384, while the average estimate over the non-PFT schools (observations 50–70) is 1.0300, indicating only a negligible difference; Charnes *et al.* concluded that PFT schools were no different from non-PFT schools.

Additional insight is gained, however, by considering the stochastic nature of the estimation problem using our bootstrap method. Table 1 reveals that the estimated biases are negative, as expected, and in numerous cases quite large. In fact, among the observations whose initial distance function estimate is unity, lower bounds for the estimated 95% confidence intervals range from 1.1055 to 1.5145. The estimated variances, on the other hand, are frequently quite small, so that the ratios r_i frequently exceed unity, in many cases by a great deal. Thus, for example, observation 5, which appears ostensibly efficient if only views only the original distance function estimate, has a bias-corrected distance function estimate of 1.4555, suggesting that the same outputs could have been produced while scaling inputs back by more than 45%. The estimated 95% confidence interval for this observation suggests that inputs could have been reduced by between 38.5% and 55.2%.

The PFT observations have an average bias-corrected distance function estimate of 1.2717, while the non-PFT schools have a corresponding average of 1.2972. Similarly, the estimated lower confidence bound averages are 1.2055 for the PFT schools, and 1.2322 for the non-PFT schools. The estimated upper 95% confidence bounds average 1.3693 and 1.3806 for the PFT and non-PFT schools, respectively. These results seem to support the conclusion that there is no significant difference between PFT and non-PFT schools. However, our results also suggest that the 70 schools in the sample are considerable more inefficient than revealed by the initial point estimates of inefficiency.

To check the robustness of our results with respect to the choice of bandwidth in Step 4 of our bootstrap algorithm in Section 4.2, we repeated our analysis of these data while setting the bandwidth first at 0.5 times the value chosen by our cross-validation procedure, then at 1.5 times the cross-validated bandwidth. Table 2 shows the 95% confidence interval estimates for both these cases, as well as for the original case for comparison. These results indicate that our bootstrap procedure is very robust with respect to choices of the bandwidth parameter used in the kernel

TABLE 2. Estimated 95% confidence intervals with alternative bandwidths

i	$h = 0.87946$		$h = 0.43973$		$h = 1.3192$	
1	1.2979	1.6658	1.2982	1.7050	1.3046	1.7675
2	1.1920	1.2988	1.1921	1.2992	1.1920	1.3010
3	1.1244	1.2364	1.1270	1.2376	1.1265	1.2355
4	1.1626	1.2199	1.1632	1.2217	1.1637	1.2208
5	1.3849	1.5520	1.3886	1.5571	1.3923	1.5603
6	1.2668	1.3727	1.2666	1.3782	1.2678	1.3826
7	1.1739	1.2222	1.1733	1.2224	1.1736	1.2223
8	1.1862	1.5587	1.1910	1.5489	1.1929	1.5390
9	1.1940	1.2586	1.1944	1.2578	1.1939	1.2570
10	1.1427	1.4365	1.1526	1.4452	1.1568	1.4764
11	1.1166	1.2902	1.1183	1.2928	1.1208	1.2947
12	1.1295	1.3492	1.1383	1.3449	1.1492	1.3446
13	1.1964	1.2446	1.1970	1.2457	1.1959	1.2447
14	1.2707	1.4102	1.2771	1.4160	1.2800	1.4213
15	1.3992	1.6348	1.4157	1.6486	1.4263	1.6529
16	1.0903	1.5755	1.0969	1.5606	1.1010	1.5714
17	1.3758	1.4965	1.3774	1.5000	1.3790	1.5046
18	1.1086	1.2273	1.1130	1.2303	1.1188	1.2348
19	1.0899	1.3208	1.0927	1.3099	1.0927	1.2964
20	1.1865	1.3745	1.1967	1.3783	1.2069	1.3797
21	1.1762	1.4426	1.1861	1.4453	1.1922	1.4360
22	1.1055	1.2005	1.1087	1.2048	1.1111	1.2057
23	1.0452	1.3559	1.0457	1.3436	1.0460	1.3306
24	1.1926	1.3607	1.2020	1.3649	1.2103	1.3698
25	1.0511	1.0905	1.0511	1.0913	1.0513	1.0901
26	1.0908	1.1605	1.0917	1.1625	1.0911	1.1632
27	1.1379	1.2809	1.1437	1.2862	1.1456	1.2886
28	1.1374	1.3101	1.1409	1.3141	1.1453	1.3180
29	1.4065	1.6416	1.4085	1.6492	1.4159	1.6573
30	1.2034	1.3240	1.2027	1.3316	1.2050	1.3334
31	1.2676	1.3511	1.2682	1.3536	1.2690	1.3531
32	1.2755	1.5158	1.2876	1.5232	1.2904	1.5289
33	1.0683	1.4155	1.0686	1.4205	1.0700	1.3882
34	1.1903	1.2643	1.1907	1.2616	1.1902	1.2600
35	1.1276	1.3640	1.1284	1.3824	1.1300	1.3725
36	1.4610	1.5666	1.4595	1.5695	1.4609	1.5736
37	1.3605	1.4566	1.3620	1.4643	1.3634	1.4664
38	1.4302	1.4963	1.4312	1.5002	1.4318	1.5018
39	1.1737	1.2822	1.1755	1.2852	1.1768	1.2850
40	1.1349	1.2188	1.1359	1.2191	1.1383	1.2197
41	1.0761	1.1117	1.0759	1.1102	1.0762	1.1100
42	1.2103	1.3987	1.2161	1.4011	1.2153	1.4088
43	1.2341	1.3534	1.2355	1.3523	1.2340	1.3492
44	–	–	–	–	–	–
45	1.4238	1.4932	1.4252	1.4975	1.4252	1.4981
46	1.1108	1.2011	1.1097	1.1919	1.1099	1.1872
47	1.1867	1.2676	1.1886	1.2703	1.1883	1.2732
48	1.3451	1.6637	1.3689	1.6790	1.3945	1.6785
49	1.1514	1.3912	1.1621	1.4125	1.1701	1.4261
50	1.0989	1.2882	1.1032	1.2887	1.1065	1.2878
51	1.2381	1.4025	1.2429	1.4086	1.2472	1.4109
52	1.1805	1.4317	1.1845	1.4368	1.1968	1.4452
53	1.2169	1.2999	1.2179	1.3038	1.2193	1.3054
54	1.3971	1.7701	1.3995	1.8317	1.3971	1.8561
55	1.1430	1.2199	1.1422	1.2188	1.1433	1.2213
56	1.3393	1.6047	1.3497	1.6168	1.3753	1.6278

TABLE 2.—(Continued)

<i>i</i>	<i>h</i> = 0.87946		<i>h</i> = 0.43973		<i>h</i> = 1.3192	
57	1.1676	1.2718	1.1707	1.2732	1.1727	1.2758
58	1.3357	1.4410	1.3402	1.4434	1.3439	1.4467
59	—	—	—	—	—	—
60	1.0908	1.1743	1.0936	1.1781	1.0947	1.1794
61	1.4464	1.6654	1.4524	1.6783	1.4598	1.6799
62	1.5145	1.6131	1.5168	1.6202	1.5171	1.6257
63	1.1898	1.3135	1.1935	1.3186	1.1958	1.3230
64	1.1009	1.1507	1.1000	1.1477	1.0994	1.1484
65	1.1980	1.3150	1.2013	1.3215	1.2020	1.3262
66	1.0971	1.1472	1.0976	1.1459	1.0978	1.1459
67	1.0808	1.1353	1.0813	1.1327	1.0813	1.1321
68	1.1707	1.4780	1.1808	1.4786	1.1928	1.4761
69	1.5084	1.6354	1.5123	1.6409	1.5139	1.6475
70	1.1314	1.2540	1.1340	1.2608	1.1356	1.2670

smoothing; missing the cross-validated bandwidth by as much as 50% in either direction has only negligible effect on the estimated confidence intervals for each observation in our sample. Note also that the normal reference rule for this sample gives a bandwidth of 0.65025, which is 0.739 times the cross-validated bandwidth for this sample. Thus, our results suggest that one might use the normal reference rule to avoid the computational burden of the cross-validation procedure at little cost in terms of the estimated confidence intervals, at least with the sample we have examined.

6 Conclusions

As noted in the introduction, DEA methods have been widely applied to examine efficiency in a variety of production settings. Yet, despite a small but growing literature on the statistical properties of DEA estimators, most researchers have used these methods while ignoring the sampling noise in the resulting efficiency estimators, and continue to do so. As our empirical example demonstrates, ignoring the statistical properties of DEA estimators and the uncertainty surrounding DEA estimates can lead to erroneous conclusions.

We have provided a general, computationally tractable method for adapting bootstrap methods to the problem of non-parametric efficiency estimation. While it is true that with samples containing many hundreds or thousands of observations it will likely be infeasible to obtain bootstrap estimated confidence intervals for each observation, as we have done in our empirical example, it is not clear why one would want to do so when the resulting volume of information would overwhelm any mortal practitioner. Rather, our method can be used to assess uncertainty about distance to the true production frontier from a relatively small number of points in the production set, cleverly chosen to reflect the location of most of the data, or of at least the most interesting part of the data.

Acknowledgements

Research support from the contract ‘Projet d’Actions de Recherche Concertées’ (PARC no. 94/98-164 and no. 98/03-217) of the Belgian Government is gratefully acknowledged. Research support from the Management Science Group, US Department of Veterans Affairs, is also gratefully acknowledged.

Notes

1. Estimation of efficiency using the free disposal hull (FDH) method of Deprins *et al.* is typically accomplished without solving linear programs, although the problem is sometimes formulated in terms of linear programs.
2. We discuss the nature of the DGP implicit in DEA models below; see also Simar (1996) and Kneip *et al.* (1998).
3. Use of the convex hull to estimate Ψ allows for the possibility of varying returns to scale. If the underlying technology is one of constant returns to scale, we could delete the constraint $\sum_{i=1}^n \gamma_i = 1$ in equation (13) to obtain the conical hull (intersected with the free-disposal hull) as an estimate of Ψ . Alternatively, replacing the equality constraint in (13) by the inequality $\sum_{i=1}^n \gamma_i \leq 1$ assumes non-increasing returns to scale. Since we do not know the true production set Ψ , we use the less-constrained convex hull to estimate Ψ .
4. Even with this non-restrictive assumption, the consistency of $\hat{\delta}(x, y)$ for a given (x, y) is achieved. Simar & Wilson (1998, 1999) imposed independence between δ and (y, η) so that $f(\delta|y, \eta) = f(\delta)$. We avoid this restriction here, allowing the distribution of radial inefficiency to vary over the production set.
5. For more details on the role of the smoothness of the frontier on the rates of convergence, see Kneip *et al.* (1998). Assumption A5 here is slightly stronger than assumptions A5–A6 in Kneip *et al.*, but can be relaxed easily; the only cost is an increase in notational complexity.
6. This idea was used in Simar & Wilson (1998, 1999) but there, the estimator of \mathcal{P} incorporates homogeneity restrictions as described previously in footnote 4.
7. The boundary conditions $y \geq 0$ and $\eta_j \in [0, \pi/2]$, $j = 1, \dots, p-1$ are not taken into account here for sake of simplicity and because, in practice, these constraints are seldom binding. The problem mentioned above is related to the fact that for a given (x, y) , the naive bootstrap frontier $\hat{\Delta}X^*(y)$ could be equal to the estimated frontier $\hat{\Delta}X(y)$ with positive probability. So only reflection about this boundary is necessary in practice. Our method could be adapted, if desired, to account for the boundary constraints on y and η but at a cost of complexity: y_k , $k = 1, \dots, q$ would be reflected about 0 and $\eta_j \in [0, \pi/2]$, $j = 1, \dots, p-1$ would be reflected about zero and about $\pi/2$. This would lead to an augmented data matrix \mathcal{Z} corresponding to equation (36) of size $12nq(p-1) \times (p+q)$. We prefer, for simplicity, to delete any negative bootstrap value of y_k^* and any infeasible value of η_j^* below, if by chance, such values should appear in the bootstrap algorithm.
8. The data represented by the rows of \mathcal{Z} may not fall in a regular ellipsoid due to the typical large number of cases where $\hat{\delta}_i = 1$. From the viewpoint of the sample covariance matrix, this situation resembles the problem of outliers; furthermore, the sample covariance matrix is quite sensitive to outliers. A robust estimator with high breakdown point would be a better choice for $\hat{\Sigma}_1$. We use the M-estimator of the covariance matrix described by Campbell (1980). Alternatively, one could use the minimum volume ellipsoid estimator proposed by Rousseeuw (1985) when $(p+q)$ is large, since the breakdown point of M-estimators decreases as the dimensionality increases (see Hampel *et al.*, 1986, chapter 5, for a discussion of this point). Unfortunately, the minimum volume ellipsoid estimator is more difficult to compute.
9. Replace (x, y) with (\hat{x}, y_i^*) in equation (17) and solve the resulting linear program to obtain $\hat{\delta}(\hat{x}, y_i^*)$.
10. The linear program (32) will not have a solution if $y_i > \max_j(y_j^*)$; the solution is infeasible in this case since (x, y) lies above the bootstrap frontier. This is a problem of finite samples.
11. The rescaling factor $(1+h^2)^{-1/2}$ is required for the rows of Γ to have approximately the covariance structure $\hat{\Sigma}_1$ of the original data in \mathcal{Z} ; the centering correction is to keep the correct mean of \mathcal{Z} .
12. In some cases, the linear program described in footnote 9 may have infeasible solutions; this will happen when $y_i^* > \max_j(y_j)$. One approach would be to impose an additional boundary condition, namely, $y_i^* \leq \max_j(y_j)$, but this would result in a great increase in complexity and computational burden. We adopt the simpler, innocuous approach of deleting bootstrap values for which $y_i^* > \max_j(y_j)$, and then redrawing. Note that the inequalities here are understood to be element-wise. In our empirical example in Section 5, this problem arises in approximately 2.3% of bootstrap draws.
13. Note that we have omitted the subscript b from $\hat{\delta}^*$ in (52), to signify the random variable $\hat{\delta}^*$, as opposed to one of its realizations, $\hat{\delta}_b^*$.
14. In Simar & Wilson (1998), we constructed confidence intervals for distance functions using an estimate of bias analogous to (48). The approach in this paper avoids introducing the extra noise contained in this estimate. The bias inherent in the distance function estimates is implicitly accounted for here since we use the bootstrap values to construct an empirical distribution of differences as in equation (53).

15. See Charnes *et al.* for the actual data and a detailed discussion, including definitions of the input and output variables. Charnes *et al.* indicate that schools with similar circumstances were chosen for their data to allow comparison of those implementing PFT and those not doing so.

REFERENCES

- BANKER, R. D. (1993) Maximum likelihood, consistency and data envelopment analysis: a statistical foundation, *Management Science*, 39(10), pp. 1265–1273.
- BERAN, R. & DUCHARME, G. (1991) *Asymptotic Theory for Bootstrap Methods in Statistics* (Montreal, Centre de Recherches Mathematiques, University of Montreal).
- BICKEL, P. J. & FREDMAN, D. A. (1981) Some asymptotic theory for the bootstrap, *Annals of Statistics* 9, pp. 1196–1217.
- CAMPBELL, N. A. (1980) Robust procedures in multivariate analysis I: robust covariance estimation, *Applied Statistics*, 29, pp. 231–237.
- CHARNES, A., COOPER, W. W. & RHODES, E. (1978) Measuring the inefficiency of decision making units, *European Journal of Operational Research*, 2, pp. 429–444.
- CHARNES, A., COOPER, W. W. & RHODES, E. (1979) Measuring the efficiency of decision making units, *European Journal of Operational Research*, 3, p. 339.
- CHARNES, A., COOPER, W. W. & RHODES, E. (1981) Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through, *Management Science*, 27, pp. 668–697.
- DEBREU, G. (1951) The coefficient of resource utilization, *Econometrica*, 19, pp. 273–292.
- DEPRINS, D., SIMAR, L. & TULKENS, H. (1984) Measuring labor inefficiency in post offices. In: M. MARCHAND, P. PESTIEAU & H. TULKENS (Eds) *The Performance of Public Enterprises: Concepts and Measurements* (Amsterdam, North-Holland), pp. 243–267.
- EFRON, B. (1979) Bootstrap methods: another look at the jackknife, *Annals of Statistics* 7, pp. 1–16.
- EFRON, B. (1982) *The Jackknife, the Bootstrap, and Other Resampling Plans*, DBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38, Society for Industrial and Applied Mathematics, Philadelphia.
- EFRON, B. & TIBSHIRANI, R. J. (1993) *An Introduction to the Bootstrap* (New York, Chapman and Hall).
- FAN, J. & GIJBELS, I. (1996) *Local Polynomial Modeling and its Applications* (New York, Chapman and Hall).
- FÄRE, R., GROSSKOPF, S. & LOVELL, C. A. K. (1985) *The Measurement of Efficiency of Production* (Boston, Kluwer-Nijhoff Publishing).
- FARRELL, M. J. (1957) The measurement of productive efficiency, *Journal of the Royal Statistical Society*, A(120), pp. 253–281.
- GIJBELS, I., MAMMEN, E., PARK, B. U. & SIMAR, L. (1999) On estimation of monotone and concave frontier functions, *Journal of the American Statistical Association*, 94, pp. 220–228.
- HÄRDLE, W. (1990) *Applied Nonparametric Regression* (Cambridge, Cambridge University Press).
- HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. & STAHEL, W. A. (1986) *Robust Statistics: The Approach Based on Influence Functions* (New York, Wiley).
- KNEIP, A., PARK, B. & SIMAR, L. (1998) A note on the convergence of nonparametric DEA efficiency measures, *Econometric Theory*, 14, pp. 783–793.
- KNEIP, A. & SIMAR, L. (1996) A general framework for frontier estimation with panel data, *Journal of Productivity Analysis*, 7(2/3), pp. 187–212.
- KOROSTELEV, A., SIMAR, L. & TSYBAKOV, A. (1995a) Efficient estimation of monotone boundaries, *The Annals of Statistics*, 23, pp. 476–489.
- KOROSTELEV, A., SIMAR, L. & TSYBAKOV, A. (1995b) On estimation of monotone and convex boundaries, *Publications de l'Institut de Statistique de l'Université de Paris XXXIX*, 1, pp. 3–18.
- LOVELL, C. A. K. (1993) Production frontiers and productive efficiency. In: HAL FRIED, C. A. KNOX LOVELL & SHELTON S. SCHMIDT (Eds) *The Measurement of Productive Efficiency: Techniques and Applications* (Oxford, Oxford University Press), pp. 3–67.
- MARKOVITZ, H. M. (1959) *Portfolio Selection: Efficient Diversification of Investments* (New York, Wiley).
- ROUSSEEUW, P. J. (1985) Multivariate estimation with high breakdown point. In: W. GROSSMAN, G. PLUG, I. VINCZE & W. WERTZ (Eds) *Mathematical Statistics and Applications, Vol B* (Dordrecht, Reidel Publishing), pp. 283–297.
- SCOTT, D. W. (1992) *Multivariate Density Estimation* (New York, Wiley).
- SEIFORD, L. M. (1996) Data envelopment analysis: the evolution of the state-of-the-art (1978–1995), *Journal of Productivity Analysis*, 7(2/3), pp. 99–138.

SEIFORD, L. M. (1997) A bibliography for data envelopment analysis (1978–1996), *Annals of Operations Research*, 73, pp. 393–438.

SENGUPTA, J. K. (1991) Maximum probability dominance and portfolio theory, *Journal of Optimization Theory and Applications*, 71, pp. 341–357.

SENGUPTA, J. K. & PARK, H. S. (1993) Portfolio efficiency tests based on stochastic dominance and cointegration, *International Journal of Systems Science*, 24, pp. 2135–2158.

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

SILVERMAN, B. W. (1986) *Density Estimation for Statistics and Data Analysis* (London, Chapman and Hall).

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

SIMAR, L. & WILSON, P. (1998) Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models, *Management Science*, 44, pp. 49–61.

SIMAR, L. & WILSON, P. (1999) Estimating and bootstrapping Malmquist indices, *European Journal of Operational Research*, 115, pp. 459–471.

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

Appendix. Choice of bandwidth

Least-squares cross-validation (e.g. see Silverman, 1986) provides a data-driven criterion for choosing the bandwidth h in kernel density estimation. As Silverman (1986) demonstrates, the usual cross-validation function gives an approximation to MISE; the goal is to choose h to minimize this approximation. As noted previously, the usual cross-validation function suffers from degenerate behaviour when data have been discretized, allowing the cross-validation function to approach $-\infty$ as $h \rightarrow 0$.

We deal with this problem by minimizing an approximation to the mean weighted integrated square error (MWISE) (see for example Härdle, 1990). The MWISE is defined by

$$\begin{aligned}
 & E \left[\int w(z) (f(z) - \hat{f}_h(z))^2 dz \right] \\
 &= E \left[\int w(z) \hat{f}_h^2(z) dz - 2 \int w(z) f(z) \hat{f}_h(z) dz + \int w(z) f^2(z) dz \right]
 \end{aligned}
 \tag{A1}$$

where $w(z)$ is some predefined weight function.

The last term of equation (A1) does not depend on h , so the optimal choice of the bandwidth (in the sense of minimizing MWISE) minimizes the following criterion:

$$C(h) = E \left[\int w(z) \hat{f}_h^2(z) dz - 2 \int w(z) f(z) \hat{f}_h(z) dz \right]
 \tag{A2}$$

The idea of cross validation is to construct an estimate of $C(h)$ from the data $\{z_i, i = 1, \dots, n\}$. Following Silverman, an unbiased estimator of $C(h)$ is given by

$$CV(h) = \int w(z) \hat{f}_h^2(z) dz - \frac{2}{n} \sum_{i=1}^n w(z_i) \hat{f}_{h,-i}(z_i)
 \tag{A3}$$

where $\hat{f}_{h,-i}(z_i)$ is the leave-one out estimator of $f(z_i)$ based on all the original observations except z_i , with bandwidth h .

In order to avoid the discretization problem, we define the weight function $w(z)$ as

$$w(z) = \begin{cases} 0 & \text{if } \delta \in [1, 1 + \varepsilon] \\ 1 & \text{otherwise,} \end{cases} \tag{A4}$$

where ε is small, say 10^{-6} . By choosing $\varepsilon > 0$ but very small, we avoid the discretization problem by preventing observations where $\hat{\delta}_i = 1$ from influencing the $CV(h)$ in (A3). The value of ε should be chosen small enough so that only observations where $\hat{\delta}_i = 1$ are eliminated.

In minimizing $CV(h)$ with respect to h , clearly there is no problem in computing the second term on the r.h.s. of (A3), but the first term is more difficult. We can write this first term as

$$\int w(z) \hat{f}_h^2(z) dz = \int_{z \in \mathcal{A}} \hat{f}_h^2(z) dz - \int_{z \in \mathcal{A}_\varepsilon} \hat{f}_h^2(z) dz \tag{A5}$$

where $\mathcal{A}_\varepsilon = \mathbb{R}^q \times [0, \pi/2]^{p-1} \times [1, 1 + \varepsilon]$.

Consider the first term of the r.h.s. of (A5):

$$\begin{aligned} \int_{z \in \mathcal{A}} \hat{f}_h^2(z) dz &= \int_{z \in \mathcal{A}} \frac{1}{nh^{(p+q)}} \sum_{i=1}^n \left[K_1 \left(\frac{z - z_i}{h} \right) + K_2 \left(\frac{z - z_{Ri}}{h} \right) \right] \\ &\quad \times \frac{1}{nh^{(p+q)}} \sum_{j=1}^n \left[K_1 \left(\frac{z - z_j}{h} \right) + K_2 \left(\frac{z - z_{Rj}}{h} \right) \right] dz \\ &= \frac{1}{n^2 h^{2(p+q)}} \sum_{i=1}^n \sum_{j=1}^n \int_{z \in \mathcal{A}} \left[K_1 \left(\frac{z - z_i}{h} \right) K_1 \left(\frac{z - z_j}{h} \right) + K_1 \left(\frac{z - z_i}{h} \right) K_2 \left(\frac{z - z_{Rj}}{h} \right) \right. \\ &\quad \left. + K_2 \left(\frac{z - z_{Ri}}{h} \right) K_1 \left(\frac{z - z_j}{h} \right) + K_2 \left(\frac{z - z_{Ri}}{h} \right) K_2 \left(\frac{z - z_{Rj}}{h} \right) \right] dz \tag{A6} \\ &= \frac{1}{n^2 h^{2(p+q)-1}} \sum_{i=1}^n \sum_{j=1}^n \int_{hu \in \mathcal{A}} [K_1(h^{-1}z_i - u)K_1(u - h^{-1}z_j) \\ &\quad + K_1(h^{-1}z_i - u)K_2(u - h^{-1}z_{Rj}) + K_2(h^{-1}z_{Ri} - u)K_1(u - h^{-1}z_j) \\ &\quad + K_2(h^{-1}z_{Ri} - u)K_2(u - h^{-1}z_{Rj})] du \\ &= \frac{1}{n^2 h^{2(p+q)-1}} \sum_{i=1}^n \sum_{j=1}^n \left[K_{11} \left(\frac{z_i - z_j}{h} \right) + 2K_{12} \left(\frac{z_i - z_{Rj}}{h} \right) + K_{22} \left(\frac{z_{Ri} - z_{Rj}}{h} \right) \right] \end{aligned}$$

where $u = h^{-1}z$ (implying $dz = h du$), $K_{11}(\cdot)$ is the convolution of $K_1(\cdot)$ with itself and hence is $N(0, 2\Sigma_1)$, $K_{12}(\cdot)$ is the convolution of $K_1(\cdot)$ with $K_2(\cdot)$ and hence is $N(0, \Sigma_1 + \Sigma_2)$ and $K_{22}(\cdot)$ is the convolution of $K_2(\cdot)$ with itself and hence is $N(0, 2\Sigma_2)$.

Consider now the second term on the r.h.s. of (A5). Let $\beta = (\gamma, \eta)$, $\mathcal{B} = \mathbb{R}^q_+ \times [0, \pi/2]^{p-1}$, and $\mathcal{C} = [1, 1 + \varepsilon]$. Then

$$\int_{z \in \mathcal{A}_c} \hat{f}_h^2(z) dz = \int_{\delta \in \mathcal{C}} \int_{\beta \in \mathcal{B}} \hat{f}_h^2(\beta | \delta) \hat{f}_h^2(\delta) d\beta d\delta \tag{A7}$$

$$= \int_{\delta \in \mathcal{C}} \left[\int_{\beta \in \mathcal{B}} \hat{f}_h^2(\beta | \delta) d\beta \right] \hat{f}_h^2(\delta) d\delta$$

Now consider the bracketed term in the last part of (A7). Partition z_i and z_{Ri} so that $z_i = (\beta_i, \hat{\delta}_i)$ and $z_{Ri} = (\beta_i, 2 - \hat{\delta}_i)$. Then conditioning on δ in our definition of $\hat{f}_h(z)$ in equation (41) yields

$$\hat{f}_h(\beta | \delta) = \frac{1}{nh^{(\varphi+q)}} \sum_{i=1}^n \left[K_1 \left(\frac{\beta - \beta_i}{h} \middle| \frac{\delta - \hat{\delta}_i}{h} \right) + K_2 \left(\frac{\beta - \beta_i}{h} \middle| \frac{\delta - (2 - \hat{\delta}_i)}{h} \right) \right] \tag{A8}$$

Squaring both sides of this yields

$$\hat{f}_h^2(\beta | \delta) = \frac{1}{n^2 h^{2(\varphi+q)}} \sum_{i=1}^n \sum_{j=1}^n \left[K_1 \left(\frac{\beta - \beta_i}{h} \middle| \frac{\delta - \hat{\delta}_i}{h} \right) K_1 \left(\frac{\beta - \beta_j}{h} \middle| \frac{\delta - \hat{\delta}_j}{h} \right) \right. \\ \left. + 2K_1 \left(\frac{\beta - \beta_i}{h} \middle| \frac{\delta - \hat{\delta}_i}{h} \right) K_2 \left(\frac{\beta - \beta_j}{h} \middle| \frac{\delta - (2 - \hat{\delta}_j)}{h} \right) \right. \\ \left. K_2 \left(\frac{\beta - \beta_i}{h} \middle| \frac{\delta - (2 - \hat{\delta}_i)}{h} \right) K_2 \left(\frac{\beta - \beta_j}{h} \middle| \frac{\delta - (2 - \hat{\delta}_j)}{h} \right) \right] \tag{A9}$$

We must integrate both sides of this expression to obtain the bracketed term in (A7). Since $K_1(\cdot)$ is multivariate $N(0, \Sigma_1)$, $K_1 \left(\frac{\beta - \beta_i}{h} \middle| \frac{\delta - \hat{\delta}_i}{h} \right)$ must be

$$N \left(S_{12} S_{22}^{-1} \frac{\delta - \hat{\delta}_i}{h}, S_{11} - S_{12} S_{22}^{-1} S_{21} \right)$$

Similarly, $K_2(\cdot)$ is multivariate $N(0, \Sigma_2)$, and hence $K_2 \left(\frac{\beta - \beta_i}{h} \middle| \frac{\delta - (2 - \hat{\delta}_i)}{h} \right)$ must be

$$N \left(-S_{12} S_{22}^{-1} \frac{\delta - (2 - \hat{\delta}_i)}{h}, S_{11} - S_{12} S_{22}^{-1} S_{21} \right)$$

Let $S_{11.2} = S_{11} - S_{12} S_{22}^{-1} S_{21}$, $\mu_{1i} = S_{12} S_{22}^{-1} \frac{\delta - \hat{\delta}_i}{h}$ and $\mu_{2i} = -S_{12} S_{22}^{-1} \frac{\delta - (2 - \hat{\delta}_i)}{h}$.

Then we can rewrite $K_1 \left(\frac{\beta - \beta_i}{h} \middle| \frac{\delta - \hat{\delta}_i}{h} \right)$ as $\phi_{1i} \left(\frac{\beta - \beta_i}{h} \right)$, where $\phi_{1i}(\cdot)$ is $N(\mu_{1i}, S_{11.2})$.

Similarly, $K_2\left(\frac{\beta - \beta_i}{h} \middle| \frac{\delta - (2 - \hat{\delta}_i)}{h}\right)$ can be rewritten as $\phi_{2i}\left(\frac{\beta - \beta_i}{h}\right)$, where $\phi_{2i}(\cdot)$ is $N(\mu_{2i}, S_{11.2})$. Then, integrating (A9), we have

$$\int_{\beta \in \mathcal{B}} \hat{f}_h^2(\beta | \delta) d\beta = \frac{1}{n^2 h^{2(p+q)}} \sum_{i=1}^n \sum_{j=1}^n \int_{\beta \in \mathcal{B}} \left[\phi_{1i}\left(\frac{\beta - \beta_i}{h}\right) \phi_{1j}\left(\frac{\beta - \beta_j}{h}\right) + 2\phi_{1i}\left(\frac{\beta - \beta_i}{h}\right) \phi_{2j}\left(\frac{\beta - \beta_j}{h}\right) + \phi_{2i}\left(\frac{\beta - \beta_i}{h}\right) \phi_{2j}\left(\frac{\beta - \beta_j}{h}\right) \right] d\beta \tag{A10}$$

Using the same convolution argument as in (A6), we obtain the bracketed term of (A7):

$$\int_{\beta \in \mathcal{B}} \hat{f}_h^2(\beta | \delta) d\beta = \frac{1}{n^2 h^{2(p+q)-1}} \sum_{i=1}^n \sum_{j=1}^n \left[\phi_{11ij}\left(\frac{\beta_i - \beta_j}{h}\right) + 2\phi_{12ij}\left(\frac{\beta_i - \beta_j}{h}\right) + \phi_{22ij}\left(\frac{\beta_i - \beta_j}{h}\right) \right] \tag{A11}$$

where for $k, \ell = 1, 2$, $\phi_{k\ell ij}(\cdot)$ is the convolution of $\phi_{ki}(\cdot)$ and $\phi_{\ell j}(\cdot)$. Hence ϕ_{11ij} is $N(\mu_{1i} + \mu_{1j}, 2S_{11.2})$, ϕ_{12ij} is $N(\mu_{1i} + \mu_{2j}, 2S_{11.2})$, and ϕ_{22ij} is $N(\mu_{2i} + \mu_{2j}, 2S_{11.2})$.

Substituting (A11) into (A7), we have, for the second term on the r.h.s. of (A5),

$$\int_{z \in \mathcal{A}_z} \hat{f}_h^2(z) dz = \frac{1}{n^2 h^{2(p+q)-1}} \sum_{i=1}^n \sum_{j=1}^n \int_{\delta \in \mathcal{E}} \left[\phi_{11ij}\left(\frac{\beta_i - \beta_j}{h}\right) + 2\phi_{12ij}\left(\frac{\beta_i - \beta_j}{h}\right) + \phi_{22ij}\left(\frac{\beta_i - \beta_j}{h}\right) \right] \hat{f}_h^2(\delta) d\delta \tag{A12}$$

We now need $\hat{f}_h^2(\delta)$. Note that from the definition of our multivariate density estimator (equation (41)), and the fact that the marginal of a multivariate normal is also normal, we have

$$\hat{f}_h(\delta) = \begin{cases} \frac{1}{nh} \sum_{i=1}^n \left[K_\star\left(\frac{\delta - \hat{\delta}_i}{h}\right) + K_\star\left(\frac{\delta - (2 - \hat{\delta}_i)}{h}\right) \right] & \text{if } \delta \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{A13}$$

where $K_\star(\cdot)$ is the univariate normal density with mean 0 and variance S_{22} .

Squaring both sides of (A13) yields

$$\hat{f}_h^2(\delta) = \frac{1}{n^2 h^2} \sum_{k=1}^n \sum_{l=1}^n \left[K_\star\left(\frac{\delta - \hat{\delta}_k}{h}\right) K_\star\left(\frac{\delta - \hat{\delta}_l}{h}\right) + 2K_\star\left(\frac{\delta - \hat{\delta}_k}{h}\right) K_\star\left(\frac{\delta - (2 - \hat{\delta}_l)}{h}\right) + K_\star\left(\frac{\delta - (2 - \hat{\delta}_k)}{h}\right) K_\star\left(\frac{\delta - (2 - \hat{\delta}_l)}{h}\right) \right] \tag{A14}$$

Finally, substituting (A14) into (A12) yields a form for the second term on the r.h.s. of (A5) that we can compute:

$$\begin{aligned}
 \int_{z \in \mathcal{A}_\varepsilon} \hat{f}_h^2(z) dz &= \frac{1}{n^4 h^{2(p+q)+1}} \int_{\delta \in \mathcal{C}} \\
 &\times \left\{ \sum_{i=1}^n \sum_{j=1}^n \left[\phi_{11ij} \left(\frac{\beta_i - \beta_j}{h} \right) + 2\phi_{12ij} \left(\frac{\beta_i - \beta_j}{h} \right) + \phi_{22ij} \left(\frac{\beta_i - \beta_j}{h} \right) \right] \right\} \\
 &\times \left\{ \sum_{k=1}^n \sum_{l=1}^n \left[K_\star \left(\frac{\delta - \hat{\delta}_k}{h} \right) K_\star \left(\frac{\delta - \hat{\delta}_l}{h} \right) + 2K_\star \left(\frac{\delta - \hat{\delta}_k}{h} \right) K_\star \left(\frac{\delta - (2 - \hat{\delta}_l)}{h} \right) \right. \right. \\
 &\left. \left. + K_\star \left(\frac{\delta - (2 - \hat{\delta}_k)}{h} \right) K_\star \left(\frac{\delta - (2 - \hat{\delta}_l)}{h} \right) \right] \right\} d\delta \tag{A15}
 \end{aligned}$$

This can be solved by any one of several one-dimensional numerical integration techniques on δ over $\mathcal{C} = [1, 1 + \varepsilon]$. We use ten-point Gauss–Legendre quadrature in our empirical example discussed in Section 5. The computation cost in solving (A15) should not be too great, since the integral is over a short interval in one dimension. In addition, the integral must be evaluated only once each time $CV(h)$ is evaluated. Note that ϕ_{11ij} , ϕ_{12ij} and ϕ_{22ij} depend on δ through their mean terms, and so remain under the integral sign.

Finally, the second term in $CV(h)$ in (A3) can be written

$$\begin{aligned}
 \frac{2}{n} \sum_{i=1}^n w(z_i) \hat{f}_{h,-i}(z_i) &= \frac{2}{n} \sum_{i=1}^n w(z_i) \frac{1}{(n-1)h^{(p+q)}} \sum_{\substack{j=1 \\ j \neq i}}^n \left[K_1 \left(\frac{z_i - z_j}{h} \right) + K_2 \left(\frac{z_i - z_{Ri}}{h} \right) \right] \\
 &= \frac{2}{n(n-1)h^{(p+q)}} \sum_{i=1}^n w(z_i) \sum_{\substack{j=1 \\ j \neq i}}^n \left[K_1 \left(\frac{z_i - z_j}{h} \right) + K_2 \left(\frac{z_i - z_{Ri}}{h} \right) \right] \tag{A16}
 \end{aligned}$$

which involves straightforward computations.

Substituting (A6) and (A15) into (A5), and then substituting the resulting expression together with (A16) into (A3) yields an expression for $CV(h)$ that can be computed numerically. Of course, the resulting expression of $CV(h)$ will be rather complicated; consequently, minimization with respect to h will typically require a grid search over a range of values $h \in [h_{\min}, h_{\max}]$. The range of values could be determined by taking factors of, say 0.25 and 2.0, of the bandwidth suggested by the normal reference rule in equation (42). Note that due to the nature of kernel functions, the expression comprising $CV(h)$ will be easier to compute for small values of h than for large values. In a similar problem, Fan & Gijbels (1996) suggest evaluating $CV(h = h_{\min})$, then successively increasing h by small increments, evaluating $CV(h)$ each time. The process terminated after some number of successive increases in $CV(h)$ have been observed, and the value of h that yielded the smallest value for $CV(h)$ is then chosen as the optimum.