



ELSEVIER

European Journal of Operational Research 139 (2002) 115–132

EUROPEAN  
JOURNAL  
OF OPERATIONAL  
RESEARCH

www.elsevier.com/locate/dsw

Stochastics and Statistics

## Non-parametric tests of returns to scale

Léopold Simar<sup>a,1</sup>, Paul W. Wilson<sup>b,\*,2</sup>

<sup>a</sup> *Institut de Statistique, Université Catholique de Louvain, Voie du Roman Pays 20, Louvain-la-Neuve, Belgium*

<sup>b</sup> *Department of Economics, University of Texas, Austin, TX 78712, USA*

Received 10 May 1999; accepted 28 February 2001

---

### Abstract

This paper discusses various statistics for testing hypotheses regarding returns to scale in the context of non-parametric models of technical efficiency. In addition, the paper presents bootstrap estimation procedures which yield appropriate critical values for the test statistics. Evidence on the true sizes and power of the various proposed tests is obtained from Monte-Carlo experiments. This paper is an extension of earlier work in [Manage. Sci. 44 (1998) 49; J. Appl. Statist. 27 (2000b) 779]. © 2002 Elsevier Science B.V. All rights reserved.

*Keywords:* Returns to scale; Data envelopment analysis; Bootstrap; Distance function; Efficiency; Frontier models

---

### 1. Introduction

Non-parametric methods have been widely used in management science and economics to estimate the efficiency of production units.<sup>3</sup> These methods are based on definitions of technical and allocative efficiency by Debreu (1951) and Farrell

(1957). A variety of non-parametric approaches for efficiency estimation have been proposed; those which rely on convexity assumptions have been termed Data Envelopment Analysis (DEA), and typically rely on linear programming (LP) techniques for solution.

Whether the underlying technology exhibits increasing, constant, or decreasing returns to scale is a crucial question in any study of productive efficiency. Färe and Grosskopf (1985) suggested an approach for determining local returns to scale in the estimated frontier which involves comparing different DEA efficiency estimates obtained under the alternative assumptions of constant, variable, or non-increasing returns to scale, but did not provide a formal statistical test of returns to scale. Banker (1996) proposed two rather ad hoc, semi-parametric statistical tests, although we see no

---

\* Corresponding author. Tel.: +1-512-475-8527; fax: +1-512-471-3510.

E-mail address: wilson@eco.utexas.edu (P.W. Wilson).

<sup>1</sup> Research support from the contract “Projet d’Actions de Recherche Concertées” (PARC No. 93/98–164) of the Belgian Government is gratefully acknowledged.

<sup>2</sup> Research support from the Management Science Group, US Department of Veterans Affairs, is gratefully acknowledged.

<sup>3</sup> Seiford (1997) provides an extensive bibliography of this literature; see also the survey by Lovell (1993).

particular reason why one should expect these tests to have the correct size. In addition, Banker proposed using a Kolmogorov–Smirnov test to examine whether DEA efficiency estimates obtained under alternative assumptions regarding returns to scale have different distributions. Löthgren and Tambour (1999) suggest a bootstrap method for implementing statistical tests based on either ratios of distance function estimates along the lines of Färe and Grosskopf (1985) or on estimates of scale elasticity along the lines of Löthgren and Tambour (1996). Unfortunately, the bootstrap proposed by Löthgren and Tambour is seriously flawed, as we have explained in some detail in Simar and Wilson (2000a); in particular, their method yields inconsistent estimates and is based on a complete misunderstanding of what is known and what must be estimated.

We propose a bootstrap procedure for testing hypotheses regarding returns to scale, avoiding the ad hoc assumptions of Banker (1996) as well as the problems in Löthgren and Tambour (1999). In addition, we provide Monte-Carlo evidence indicating that our procedure yields tests whose sizes are much closer to the nominal size than Banker's tests. The development of a reliable test procedure for examining returns to scale is important for both economic and statistical reasons. The question of whether a technology exhibits constant returns to scale everywhere typically has important economic implications – if the technology does not exhibit constant returns to scale, then some production units may be found to be either too large or too small. Some authors have a priori imposed the rather restrictive assumption of constant returns to scale while using DEA methods; this may seriously distort measures of efficiency if the true technology displays non-constant returns to scale. Indeed, as we discuss below, this scenario results in statistically inconsistent estimates of efficiency. Alternatively, if one assumes variable returns to scale when returns are actually constant everywhere, there may be a loss of statistical efficiency. Using the methods we propose below, researchers can first test whether returns to scale are constant, and then choose the appropriate estimator to measure efficiency.

DEA methods have frequently been characterized as *deterministic* in the literature (e.g.,

Lovell, 1993), as if to suggest that non-parametric models of efficiency have no statistical underpinnings. Yet, efficiency is measured relative to an *estimate* of the boundary of the underlying *true* production set, conditional on observed data resulting from an underlying (and unknown) data-generating process (DGP). Banker (1993) was one of the first to discuss the statistical nature of DEA estimators. Korostelev et al. (1995), Kneip et al. (1998), and Gijbels et al. (1999) examined the statistical properties of DEA estimators. Curiously, papers continue to appear in respectable journals which ignore the statistical nature of DEA problems. Simar and Wilson (1998) provide a bootstrap methodology for estimating confidence intervals for individual production units' efficiency; this methodology is adapted here to the problem of testing hypotheses regarding returns to scale.

In the following section we discuss methods for measuring returns to scale in the *true* (but unknown) technology using distance functions, and in Section 3 we discuss *estimation* of these distance functions. The distinction between these two sections is important, and continues to be overlooked in many papers currently circulating. In Section 4 we propose several statistics for statistically testing hypotheses regarding returns to scale of the technology, and in Section 5 we show how critical values for our statistics can be estimated using bootstrap methods. Monte-Carlo estimates of the true sizes of our tests at various nominal sizes are provided in Section 7. Conclusions are discussed in the final section.

## 2. Measuring returns to scale

To establish notation and define the things we wish to test, let  $\mathbf{x} \in \mathbb{R}_+^p$  denote a vector of  $p$  inputs and  $\mathbf{y} \in \mathbb{R}_+^q$  denote a vector of  $q$  outputs. Define the production set

$$\mathcal{P} \equiv \{(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \text{ can produce } \mathbf{y}\}, \quad (2.1)$$

which is merely the set of feasible combinations of  $\mathbf{x}$  and  $\mathbf{y}$ . The boundary of  $\mathcal{P}$  is sometimes referred to as the *technology* or *production frontier*, and is

given by the intersection of  $\mathcal{P}$  and the closure of its complement, denoted  $\mathcal{P}^B$ . Various assumptions regarding properties of the production set  $\mathcal{P}$  are possible; we adopt those of Shephard (1970) and Färe (1988), although our bootstrap methodology does not depend on these assumptions.

Now define the set  $\mathcal{V}$  as the convex cone (with vertex at the origin) spanned by  $\mathcal{P}$  so that  $\mathcal{P} \subseteq \mathcal{V}$ . If  $\mathcal{P}^B$  exhibits constant returns to scale (CRS) everywhere, then the technology  $\mathcal{P}^B$  implies a mapping  $\mathbf{x} \rightarrow \mathbf{y}$  that is homogeneous of degree 1; i.e.,  $(\mathbf{x}, \mathbf{y}) \in \mathcal{P}^B$  implies  $(\lambda\mathbf{x}, \lambda\mathbf{y}) \in \mathcal{P}^B$  for all  $\lambda > 0$ , and  $\mathcal{P} = \mathcal{V}$ .

Alternatively, if  $\mathcal{P}^B$  does not exhibit CRS everywhere, then  $\mathcal{P} \subset \mathcal{V}$ . In this case, we say that  $\mathcal{P}^B \cap \mathcal{V}$  gives the region of  $\mathcal{P}^B$  that exhibits CRS. The intersection  $\mathcal{P}^B \cap \mathcal{V}$  may be as small as a single point, but by construction cannot be the null set. Decreasing returns to scale in the neighborhood of a point  $(\mathbf{x}, \mathbf{y}) \in \mathcal{P}^B$  implies  $(\lambda\mathbf{x}, \lambda\mathbf{y}) \notin \mathcal{P}$  for  $\lambda > 1$ . Similarly, increasing returns to scale in the neighborhood of a point  $(\mathbf{x}, \mathbf{y}) \in \mathcal{P}^B$  implies  $(\lambda\mathbf{x}, \lambda\mathbf{y}) \notin \mathcal{P}$  for  $\lambda < 1$ . A technology  $\mathcal{P}^B$  that exhibits increasing, constant, and decreasing returns to scale in different regions is one of variable returns to scale (VRS) in the sense of Afriat (1972).<sup>4</sup>

The Shephard (1970) output distance function provides a normalized measure of Euclidean distance from a point  $(\mathbf{x}, \mathbf{y}) \in \mathcal{P}$  to  $\mathcal{P}^B$  in a direction orthogonal to  $\mathbf{x}$ , and may be defined as

$$D(\mathbf{x}, \mathbf{y}) \equiv \inf\{\theta > 0 \mid (\mathbf{x}, \theta^{-1}\mathbf{y}) \in \mathcal{P}\}. \quad (2.2)$$

Clearly,  $D(\mathbf{x}, \mathbf{y}) \leq 1$  for  $(\mathbf{x}, \mathbf{y}) \in \mathcal{P}$ . If  $D(\mathbf{x}, \mathbf{y}) = 1$ , then  $(\mathbf{x}, \mathbf{y}) \in \mathcal{P}^B$ ; i.e., that the point  $(\mathbf{x}, \mathbf{y})$  lies on the boundary of  $\mathcal{P}$ , and the firm is technically efficient.<sup>5</sup>

<sup>4</sup> Grosskopf (1986) gives additional discussion of returns to scale.

<sup>5</sup> One can similarly define the Shephard (1970) input distance function, measures of hyperbolic graph efficiency defined by Färe et al. (1985), directional distance functions discussed by Färe et al. (1997) and Färe and Grosskopf (2000), or perhaps other measures. We present our methodology only in terms of the output orientation to conserve space, but the methods we propose can be adapted to the other cases by straightforward changes in our notation.

Analogous to (2.2), define

$$D^{\text{crs}}(\mathbf{x}, \mathbf{y}) \equiv \inf\{\theta > 0 \mid (\mathbf{x}, \theta^{-1}\mathbf{y}) \in \mathcal{V}\}. \quad (2.3)$$

For  $(\mathbf{x}, \mathbf{y}) \in \mathcal{P}$ ,  $D^{\text{crs}}(\mathbf{x}, \mathbf{y}) \leq D(\mathbf{x}, \mathbf{y}) \leq 1$ . If  $D(\mathbf{x}, \mathbf{y}) > D^{\text{crs}}(\mathbf{x}, \mathbf{y})$ , then  $\mathcal{P} \neq \mathcal{V}$  and  $\mathcal{P}^B$  does not exhibit CRS everywhere. If  $D(\mathbf{x}, \mathbf{y}) = D^{\text{crs}}(\mathbf{x}, \mathbf{y})$ , then the projection of  $(\mathbf{x}, \mathbf{y})$  onto  $\mathcal{P}^B$  along the path  $(\mathbf{x}, \lambda\mathbf{y})$ ,  $\lambda > 1$ , yields a point on  $\mathcal{P}^B$  where  $\mathcal{P}^B$  exhibits CRS in the sense defined above. The same may not be true for another point, however; i.e., it does not follow that  $\mathcal{P} = \mathcal{V}$ .

Consider the set

$$\mathcal{V}^{\text{nirs}} = \mathcal{P} \cup \{(\mathbf{x}, \mathbf{y}) \mid (\mathbf{x}, \mathbf{y}) \notin \mathcal{P}, (\mathbf{x}, \mathbf{y}) \in \mathcal{V}, (\lambda\mathbf{x}, \lambda\mathbf{y}) \in \mathcal{P} \text{ for some } \lambda > 1\}. \quad (2.4)$$

Then  $\mathcal{P} \subseteq \mathcal{V}^{\text{nirs}} \subseteq \mathcal{V}$ . If  $\mathcal{P} = \mathcal{V}^{\text{nirs}} \neq \mathcal{V}$ , then  $\mathcal{P}^B$  exhibits non-increasing returns to scale (NIRS), i.e., either constant or decreasing returns to scale. Analogous to (2.2), distance from a point  $(\mathbf{x}, \mathbf{y}) \in \mathcal{P}$  to the boundary of  $\mathcal{V}^{\text{nirs}}$  in the direction orthogonal to  $\mathbf{x}$  is given by

$$D^{\text{nirs}}(\mathbf{x}, \mathbf{y}) \equiv \inf\{\theta > 0 \mid (\mathbf{x}, \theta^{-1}\mathbf{y}) \in \mathcal{V}^{\text{nirs}}\}. \quad (2.5)$$

The distance functions defined by (2.2), (2.3) and (2.5) may be used to construct measures of scale efficiency along the lines of Färe and Grosskopf (1985), namely

$$s(\mathbf{x}, \mathbf{y}) \equiv D^{\text{crs}}(\mathbf{x}, \mathbf{y})/D(\mathbf{x}, \mathbf{y}) \leq 1 \quad (2.6)$$

and

$$\eta(\mathbf{x}, \mathbf{y}) \equiv D^{\text{nirs}}(\mathbf{x}, \mathbf{y})/D(\mathbf{x}, \mathbf{y}) \leq 1. \quad (2.7)$$

From the preceding discussion, if  $s(\mathbf{x}, \mathbf{y}) = 1$ , then  $\mathcal{P}^B$  exhibits CRS at the point where  $(\mathbf{x}, \mathbf{y}) \in \mathcal{P}$  is projected onto  $\mathcal{P}^B$  in a direction orthogonal to  $\mathbf{x}$ . A firm operating at  $(\mathbf{x}, \mathbf{y})$  is then said to be *output scale-efficient*. On the other hand, if  $s(\mathbf{x}, \mathbf{y}) < 1$  and the technology exhibits increasing, constant, and decreasing returns at different locations, then the firm at  $(\mathbf{x}, \mathbf{y})$  operates under the decreasing returns portion of the technology if  $\eta(\mathbf{x}, \mathbf{y}) = 1$ , or under the increasing returns portion of the technology if  $\eta(\mathbf{x}, \mathbf{y}) < 1$ .

### 3. Estimating distance functions

None of the quantities discussed in the previous section are observed, including the scale efficiency measures  $s(x, y)$  and  $\eta(x, y)$ ; rather, they must be estimated from data. Given a sample  $\mathcal{S}_n = \{(x_i, y_i)\}_{i=1}^n$  with  $n$  observations, we can construct an estimator  $\widehat{\mathcal{P}}$  of  $\mathcal{P}$ , which also implies an estimator  $\widehat{\mathcal{P}}^B$  of  $\mathcal{P}^B$ . Substituting  $\widehat{\mathcal{P}}$  for  $\mathcal{P}$  in (2.2) yields a distance function estimator  $\widehat{D}(x, y)$  of  $D(x, y)$ . Whereas  $D(x, y)$  measures distance in the output direction from the point  $(x, y)$  to  $\mathcal{P}^B$ ,  $\widehat{D}(x, y)$  measures distance from the same point to  $\widehat{\mathcal{P}}^B$ , in the same direction. Much of the DEA literature obscures the distinction between measures of true efficiency such as  $D(x, y)$ , and estimators such as  $\widehat{D}(x, y)$  which, when computed from observed data, yield estimates of the true efficiency measures. Since the true production set is unobserved, all we can hope for is an estimate of efficiency, rather than any direct measure.

Several estimators of  $\mathcal{P}$  are possible. The union of the free disposal and convex hulls of  $\mathcal{S}_n$  provides one estimator,  $\widehat{\mathcal{P}}_n^{\text{vts}}$ , of  $\mathcal{P}$ ; Korostelev et al. (1995) prove consistency and establish rates of convergence. Alternatively, the convex polyhedral cone (with vortex at the origin) spanned by  $\widehat{\mathcal{P}}_n^{\text{vts}}$ , denoted  $\widehat{\mathcal{V}}_n$ , gives an estimator of  $\mathcal{V}$ . Clearly,  $\widehat{\mathcal{P}}_n^{\text{vts}} \subseteq \widehat{\mathcal{V}}_n$ . Results in Korostelev et al. (1995) may be extended to prove that  $\widehat{\mathcal{V}}_n$  is a consistent estimator of  $\mathcal{V}$ , and therefore a consistent estimator of  $\mathcal{P}$  if  $\mathcal{P} = \mathcal{V}$ , i.e., if  $\mathcal{P}^B$  exhibits CRS everywhere. But, if  $\mathcal{P}$  is convex and  $\mathcal{P}^B$  does not exhibit CRS everywhere, then  $\widehat{\mathcal{P}}_n^{\text{vts}}$  is a consistent estimator of  $\mathcal{P}$ , but  $\widehat{\mathcal{V}}_n$  is not. Therefore we say that  $\widehat{\mathcal{V}}_n$  is constrained relative to  $\widehat{\mathcal{P}}_n^{\text{vts}}$ ; i.e.,  $\widehat{\mathcal{V}}_n$  is a more restrictive estimator of  $\mathcal{P}$  than  $\widehat{\mathcal{P}}_n^{\text{vts}}$  in the sense that  $\widehat{\mathcal{P}}_n^{\text{vts}}$  will be consistent regardless of the shape of the convex set  $\mathcal{P}$ , while  $\widehat{\mathcal{V}}_n$  will be consistent only if  $\mathcal{P}^B$  exhibits constant returns to scale everywhere, i.e., if  $\mathcal{P} = \mathcal{V}$ . This point is important in establishing null hypotheses about returns to scale which can be tested statistically.

An estimator of  $\mathcal{V}^{\text{nirs}}$  is provided by

$$\widehat{\mathcal{V}}_n^{\text{nirs}} = \{(x, y) | y \leq Y \tau, x \geq X \tau, i\tau \leq 1, \tau \in \mathbb{R}_+^n\}, \tag{3.1}$$

where  $Y = [y_1 \cdots y_n]$ ,  $X = [x_1 \cdots x_n]$ , with each  $x_i, y_i$   $i = 1, \dots, n$  denoting the  $(p \times 1)$  and  $(q \times 1)$  vectors of observed inputs and outputs (respectively),  $i$  is a  $(1 \times n)$  vector of ones, and  $\tau = [\tau_1 \cdots \tau_n]'$  is a  $(n \times 1)$  vector of intensity variables. By construction,  $\widehat{\mathcal{P}}_n^{\text{vts}} \subseteq \widehat{\mathcal{V}}_n^{\text{nirs}} \subseteq \widehat{\mathcal{V}}_n$ . Once again, results from Korostelev et al. (1995) can be extended to prove that  $\widehat{\mathcal{V}}_n^{\text{nirs}}$  will be a consistent estimator of  $\mathcal{P}$  if  $\mathcal{P}^B$  exhibits only constant or decreasing returns to scale (i.e., non-increasing returns to scale). Similar reasoning suggests that  $\widehat{\mathcal{P}}_n^{\text{vts}}$ ,  $\widehat{\mathcal{V}}_n^{\text{nirs}}$ , and  $\widehat{\mathcal{V}}_n$  will consistently estimate  $\mathcal{P}$  if  $\mathcal{P}^B$  exhibits CRS everywhere, although  $\widehat{\mathcal{V}}_n$  may have a higher convergence rate in this case.  $\widehat{\mathcal{V}}_n^{\text{nirs}}$  and  $\widehat{\mathcal{V}}_n$  will fail to estimate  $\mathcal{P}$  consistently if  $\mathcal{P}^B$  exhibits both increasing and decreasing returns to scale at different locations, but  $\widehat{\mathcal{P}}_n^{\text{vts}}$  remains consistent in this case.

Estimators of the distance functions defined by (2.2), (2.3), and (2.5) may be obtained by substituting the estimators  $\widehat{\mathcal{P}}_n^{\text{vts}}$ ,  $\widehat{\mathcal{V}}_n$ , and  $\widehat{\mathcal{V}}_n^{\text{nirs}}$  for  $\mathcal{P}$ ,  $\mathcal{V}$ , and  $\mathcal{V}^{\text{nirs}}$ . Doing so yields

$$\left[ \widehat{D}_n^{\text{vts}}(x, y) \right]^{-1} = \max \{ \theta | \theta y \leq Y \tau, x \geq X \tau, i\tau = 1, \tau \in \mathbb{R}_+^n \}, \tag{3.2}$$

$$\left[ \widehat{D}_n^{\text{crs}}(x, y) \right]^{-1} = \max \{ \theta | \theta y \leq Y \tau, x \geq X \tau, \tau \in \mathbb{R}_+^n \}, \tag{3.3}$$

and

$$\left[ \widehat{D}_n^{\text{nirs}}(x, y) \right]^{-1} = \max \{ \theta | \theta y \leq Y \tau, x \geq X \tau, i\tau \leq 1, \tau \in \mathbb{R}_+^n \}, \tag{3.4}$$

respectively.

### 4. Testing hypotheses regarding returns to scale

Aside from the economic implications of whether  $\mathcal{P}^B$  exhibits CRS, there are also statistical issues relating to this question. If  $\mathcal{P}^B$  exhibits CRS everywhere, then both  $\widehat{D}_n^{\text{crs}}(x, y)$  and  $\widehat{D}_n^{\text{vts}}(x, y)$  are consistent estimators of  $D(x, y)$ , but  $\widehat{D}_n^{\text{vts}}(x, y)$  may be less efficient in a statistical sense than  $\widehat{D}_n^{\text{crs}}(x, y)$  due to slower convergence. On the other hand, if

$\mathcal{P}^B$  exhibits non-constant returns to scale at some locations, then  $\widehat{D}_n^{\text{CRS}}(\mathbf{x}, \mathbf{y})$  will be an inconsistent estimator of  $D(\mathbf{x}, \mathbf{y})$ .<sup>6</sup> Therefore, researchers need to know whether the technology is one of constant returns to scale before estimating technical efficiency. Moreover, even if  $\mathcal{P}^B$  is known to exhibit non-constant returns to scale, for policy purposes one might need to know whether a particular region of the technology displays increasing, constant, or decreasing returns to scale.

A priori assuming a CRS technology without investigating the possibility that returns to scale are not constant incurs the risk of inconsistently estimating technical efficiency. A number of papers have analyzed returns to scale along the lines of Färe and Grosskopf (1985) by computing ratios  $\widehat{D}_n^{\text{CRS}}(\mathbf{x}_i, \mathbf{y}_i) / \widehat{D}_n^{\text{VRS}}(\mathbf{x}_i, \mathbf{y}_i) \leq 1$  for each observation  $i = 1, \dots, n$  in  $\mathcal{S}_n$ . If  $\widehat{D}_n^{\text{CRS}}(\mathbf{x}_i, \mathbf{y}_i) / \widehat{D}_n^{\text{VRS}}(\mathbf{x}_i, \mathbf{y}_i) = 1$ , then the estimated technology is assumed to exhibit CRS at the point  $(\mathbf{x}_i, \mathbf{y}_i / \widehat{D}_n^{\text{VRS}})$  (i.e., where  $(\mathbf{x}_i, \mathbf{y}_i)$  is projected onto  $\widehat{\mathcal{P}}^B$  in the output direction); otherwise, the estimated technology is assumed to exhibit either increasing or decreasing returns to scale at this point. Examples of this approach include Byrnes et al. (1986), Grosskopf and Valdmanis (1987), Dusansky and Wilson (1994), and Ferrier (1994).

With this type of analysis, definitive statements regarding the *estimated* technology  $\widehat{\mathcal{P}}^B$  are possible, but how these translate into *inferences* regarding the *true*, but *unknown* technology  $\mathcal{P}^B$  has been unclear, due to a lack of a formal statistical testing procedure. In other words, if for a particular observation one finds  $\widehat{D}_n^{\text{CRS}}(\mathbf{x}_i, \mathbf{y}_i) / \widehat{D}_n^{\text{VRS}}(\mathbf{x}_i, \mathbf{y}_i) < 1$ , then without a formal testing procedure, it is impossible to determine whether this is due to non-constant returns to scale or merely due to sampling variation. We redress this deficiency below.

<sup>6</sup> We state this without formal proof, although results from Kneip et al. (1998) can be extended to give a formal proof. The result is intuitively obvious. If  $\mathcal{P}^B$  exhibits non-constant returns to scale at some locations, then for points  $(\mathbf{x}, \mathbf{y})$  located below the non-constant returns to scale regions of  $\mathcal{P}^B$ ,  $\widehat{D}_n^{\text{CRS}}(\mathbf{x}, \mathbf{y}) < D(\mathbf{x}, \mathbf{y})$  even when the sample size  $n$  approaches infinity.

From the viewpoint of statistical hypothesis testing, the first issue is whether  $\mathcal{P}^B$  represents a CRS technology. If there is no evidence to the contrary, there seems to be little reason to examine scale efficiency. As discussed in the previous section, the assumption of CRS embodied in the estimator  $\widehat{\mathcal{V}}_n$  is more restrictive than the assumption of variable returns to scale embodied by  $\widehat{\mathcal{P}}_n^{\text{VRS}}$ . This suggests

Test #1 :  $H_0 : \mathcal{P}^B$  is globally CRS  
 versus  $H_1 : \mathcal{P}^B$  is VRS.

If  $H_0$  is rejected, we may wish to perform another test with a less restrictive null hypothesis, namely

Test #2 :  $H'_0 : \mathcal{P}^B$  is globally NIRS  
 versus  $H_1 : \mathcal{P}^B$  is VRS.

Numerous test statistics are possible for Test #1. First, consider an estimator of  $s(\mathbf{x}, \mathbf{y})$ , defined in (2.6):

$$\widehat{s} = \widehat{D}_n^{\text{CRS}}(\mathbf{x}, \mathbf{y}) / \widehat{D}_n^{\text{VRS}}(\mathbf{x}, \mathbf{y}). \tag{4.1}$$

This statistic gives an estimate of the distance between the CRS and VRS estimated frontiers in the output direction at the point where  $(\mathbf{x}, \mathbf{y})$  is projected onto the estimated frontiers. Evaluating this at each observation  $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{S}_n$  yields  $n$  estimates  $\widehat{s}_i, i = 1, \dots, n$ , of scale efficiency corresponding to each observation in  $\mathcal{S}_n$ . Suppose that for each  $i = 1, \dots, n$  we compare  $\widehat{s}_i$  with the appropriate critical value, rejecting the null hypothesis of CRS whenever  $\widehat{s}_i$  falls below an appropriate critical value. Such a procedure amounts to  $n$  applications of Test #1; if each is independent of the other applications, then under the null  $H_0$  in Test #1, the number of rejections of the null hypothesis  $H_0$  over the  $n$  tests would be binomially distributed.

Applying the above test  $n$  times with nominal size  $\alpha$  in each case, if the  $n$  individual tests are independent and the true size of the tests is  $\alpha$ , then the probability of  $r \leq n$  or more rejections among the  $n$  individual tests, when  $H_0$  is true, is given by the incomplete beta function

$$I_\alpha(r, n - r + 1) = \sum_{j=r}^n \binom{n}{j} \alpha^j (1 - \alpha)^{n-j}. \tag{4.2}$$

With a nominal size of 5%, we would reject  $H_0$  if  $I_{0.05}(r, n - r + 1) < 0.05$ .<sup>7</sup>

Alternatively, a binomial test can be constructed where we reject  $H_0$  if only one (or more) of the  $n$  individual tests results in a rejection. This approach requires that the size of each of the  $n$  individual tests,  $\alpha_\ell$ , be chosen much smaller than the global or overall size of the test,  $\alpha_g$ . Using (4.2), we have

$$\Pr(r \geq 1) = 1 - \Pr(r = 0) = 1 - (1 - \alpha_\ell)^n. \quad (4.3)$$

The null hypothesis  $H_0$  will be rejected when this probability is less than or equal to  $\alpha_g$ . Thus, setting  $\Pr(r \geq 1) = \alpha_g$  yields

$$\alpha_\ell = 1 - (1 - \alpha_g)^{1/n}. \quad (4.4)$$

For example, if  $n = 20$  and the overall size of the test is set at  $\alpha_g = 0.05$ , then the size of the individual tests should be set at  $\alpha_\ell = 0.00256$ .

The binomial tests illustrate the fundamental importance of Test #1. It can only make sense to examine scale efficiency of individual firms if the underlying technology is one of non-constant returns to scale along some regions. Yet, we do not know this – we must first test the null hypothesis  $H_0$  in Test #1. If  $H_0$  is rejected, one could then in principle use the statistic defined in (4.1) to test hypotheses regarding the scale efficiency of individual firms. Yet, any such tests would necessarily be conditional on the outcome in Test #1, which would affect how one might interpret the results of tests for scale efficiency of individual firms.

Other test statistics for Test #1 are also possible. An obvious candidate is the mean of ratios  $\widehat{D}_n^{\text{CRS}}(\mathbf{x}_i, \mathbf{y}_i) / \widehat{D}_n^{\text{VRS}}(\mathbf{x}_i, \mathbf{y}_i)$ , i.e.,

$$\widehat{S}_{1n}^{\text{CRS}} = n^{-1} \sum_{i=1}^n \widehat{D}_n^{\text{CRS}}(\mathbf{x}_i, \mathbf{y}_i) / \widehat{D}_n^{\text{VRS}}(\mathbf{x}_i, \mathbf{y}_i), \quad (4.5)$$

<sup>7</sup> If the  $n$  individual tests are not independent, then the binomial formula is incorrect and the binomial test is questionable. In particular, since each test involves the same estimate of the production set, this may be a crucial issue. The bootstrap methodology we propose in Section 5 ensures independence among the individual tests.

which is an estimator of  $S_1^{\text{CRS}} = n^{-1} \sum_{i=1}^n [D_n^{\text{CRS}}(\mathbf{x}_i, \mathbf{y}_i) / D_n^{\text{VRS}}(\mathbf{x}_i, \mathbf{y}_i)]$ . If  $H_0$  is true, then  $D_n^{\text{CRS}}(\mathbf{x}_i, \mathbf{y}_i) = D_n^{\text{VRS}}(\mathbf{x}_i, \mathbf{y}_i)$  for all  $i = 1, \dots, n$ , and  $S_1^{\text{CRS}} = 1$ ; otherwise,  $S_1^{\text{CRS}} < 1$ . By construction,  $\widehat{S}_{1n}^{\text{CRS}} \leq 1$ ; the null hypothesis  $H_0$  will be rejected when  $\widehat{S}_{1n}^{\text{CRS}}$  is significantly less than unity.

Instead of using the mean of ratios as in (4.5), we might use a ratio of means as our test statistic in Test #1:

$$\widehat{S}_{2n}^{\text{CRS}} = \sum_{i=1}^n \widehat{D}_n^{\text{CRS}}(\mathbf{x}_i, \mathbf{y}_i) / \sum_{i=1}^n \widehat{D}_n^{\text{VRS}}(\mathbf{x}_i, \mathbf{y}_i). \quad (4.6)$$

Other possibilities include the median of ratios, the ratio of medians, the 10% trimmed mean of ratios, or the ratio of 10% trimmed means:<sup>8</sup>

$$\widehat{S}_{3n}^{\text{CRS}} = \text{Med} \left\{ \widehat{D}_n^{\text{CRS}}(\mathbf{x}_i, \mathbf{y}_i) / \widehat{D}_n^{\text{VRS}}(\mathbf{x}_i, \mathbf{y}_i) \right\}_{i=1}^n, \quad (4.7)$$

$$\widehat{S}_{4n}^{\text{CRS}} = \text{Med} \left\{ \widehat{D}_n^{\text{CRS}}(\mathbf{x}_i, \mathbf{y}_i) \right\}_{i=1}^n / \text{Med} \left\{ \widehat{D}_n^{\text{VRS}}(\mathbf{x}_i, \mathbf{y}_i) \right\}_{i=1}^n, \quad (4.8)$$

$$\widehat{S}_{5n}^{\text{CRS}} = \text{Trim}_{10} \left\{ \widehat{D}_n^{\text{CRS}}(\mathbf{x}_i, \mathbf{y}_i) / \widehat{D}_n^{\text{VRS}}(\mathbf{x}_i, \mathbf{y}_i) \right\}_{i=1}^n, \quad (4.9)$$

$$\widehat{S}_{6n}^{\text{CRS}} = \text{Trim}_{10} \left\{ \widehat{D}_n^{\text{CRS}}(\mathbf{x}_i, \mathbf{y}_i) \right\}_{i=1}^n / \text{Trim}_{10} \left\{ \widehat{D}_n^{\text{VRS}}(\mathbf{x}_i, \mathbf{y}_i) \right\}_{i=1}^n. \quad (4.10)$$

The mean of ratios in (4.5) has an intuitive geometric interpretation. To see this, consider the ratio  $\widehat{D}_n^{\text{CRS}}(\mathbf{x}_i, \mathbf{y}_i) / \widehat{D}_n^{\text{VRS}}(\mathbf{x}_i, \mathbf{y}_i)$ . At the point where  $(\mathbf{x}_i, \mathbf{y}_i)$  is projected onto the estimated frontiers in the output direction, this ratio gives the distance between the estimated VRS frontier and the estimated CRS frontier. If this distance is small on average, we would not want to reject the null hypothesis  $H_0$  in Test #1. The ratio of means in (4.6) is a natural variation on this idea. The test statistics

<sup>8</sup> Given a set of  $n$  scalars  $A = \{a_1, \dots, a_n\}$ , we define  $\text{Med } A$  as the median of the elements in  $A$ . In addition, we define  $\text{Trim}_{10} A$  as the 10% trimmed mean of the elements in  $A$  obtained by arranging the elements of  $A$  in an algebraic order, deleting 5% of the elements from either end of the sorted array, and computing the mean of the remaining 90% of the elements of  $A$ . The median and trimmed mean are more robust measures of location than the sample mean when the underlying distribution is skewed, as in the present case.

defined in (4.7)–(4.10) are analogous to those defined in (4.5) and (4.6), except that robust measures of location (namely, medians and trimmed means) are used rather than arithmetic means.

If, after employing one of the tests described above, the null hypothesis of global constant returns to scale is rejected, the researcher may wish to perform Test #2 as discussed above. For Test #2, test statistics  $\widehat{\eta}_i, \widehat{S}_{1n}^{\text{nirs}}, \dots, \widehat{S}_{6n}^{\text{nirs}}$  analogous to  $\widehat{s}_i, \widehat{S}_{1n}^{\text{crs}}, \dots, \widehat{S}_{6n}^{\text{crs}}$  can be defined by replacing  $\widehat{D}_n^{\text{crs}}(\mathbf{x}_i, \mathbf{y}_i)$  with  $\widehat{D}_n^{\text{nirs}}(\mathbf{x}_i, \mathbf{y}_i)$  in (4.1) and (4.5)–(4.10), respectively, where  $\widehat{\eta}_i$  is an estimator of  $\eta(\mathbf{x}_i, \mathbf{y}_i)$  defined in (2.7).

To employ either of these tests, we must obtain appropriate critical values or estimate *p*-values. To estimate the appropriate critical values, we extend the bootstrap methodology developed in Simar and Wilson (1998) as discussed in the following section.

### 5. Bootstrapping the test statistics

In each of our testing problems, we have a univariate parameter represented here by  $\omega$ . In each of our tests, if the null hypothesis  $H_0$  is true, then  $\omega = \omega_0$ . Otherwise, if the alternative hypothesis  $H_1$  is true, then  $\omega < \omega_0$ . In addition, in each of our tests, we know  $\omega_0 = 1$  when the null hypothesis is true, but  $\omega$  is unknown. For each test we also have a consistent estimator of  $\omega$ ,  $\widehat{\omega}$ . For example, in the case of Test #1,  $\omega$  might represent

$$S_1^{\text{crs}} = n^{-1} \sum_{i=1}^n [D_n^{\text{crs}}(\mathbf{x}_i, \mathbf{y}_i) / D_n^{\text{vrs}}(\mathbf{x}_i, \mathbf{y}_i)],$$

which is estimated by  $\widehat{S}_{1n}^{\text{crs}}$  defined in (4.5).

For each of our tests, we reject the null hypothesis if  $\widehat{\omega} \leq \omega_0 - c_\alpha$ , where for a test with size equal to  $\alpha$ ,  $c_\alpha > 0$  is such that

$$\Pr(\widehat{\omega} \leq \omega_0 - c_\alpha | H_0) = \alpha. \tag{5.1}$$

Unfortunately, however, we do not know the distribution of our test statistic  $\widehat{\omega}$  under  $H_0$ , which would allow us to determine the appropriate value for  $c_\alpha$ . Efron’s (1979) bootstrap offers a solution to this problem by replicating the DGP which yielded the original sample  $\mathcal{S}_n$  to generate *B* pseudo sam-

ples  $\mathcal{S}_{bn}^*$ ,  $b = 1, \dots, B$ , each with *n* observations. Then, the original estimation method is applied to each of the pseudo samples to yield bootstrap estimates  $\widehat{\omega}_b^*$ . This yields an observed empirical distribution for  $(\widehat{\omega}^* - \widehat{\omega})$ ; the idea behind the bootstrap is to use this distribution to approximate the unknown distribution of  $(\widehat{\omega} - \omega_0)$  under  $H_0$ .

The crucial point is that the pseudo samples  $\mathcal{S}_{bn}^*$  must be generated such that

$$(\widehat{\omega} - \omega) | H_0 \sim (\widehat{\omega}^* - \widehat{\omega}) | H_0, \mathcal{S}_n; \tag{5.2}$$

since  $\omega = \omega_0 = 1$  under the null hypothesis, (5.2) is equivalent to

$$(\widehat{\omega} - 1) | H_0 \sim (\widehat{\omega}^* - \widehat{\omega}) | H_0, \mathcal{S}_n. \tag{5.3}$$

In particular, (5.2) requires that the bootstrap pseudo samples be generated under the null hypothesis. In addition, since for each of our tests the test statistic represented by  $\widehat{\omega}$  is composed of distance function estimates, the pseudo samples must be generated so that the empirical bootstrap distributions of the bootstrap distance function estimates consistently estimate the sampling distributions of the original distance function estimates. As demonstrated in Simar and Wilson (1998, 1999a,b, 2000b), the naive bootstrap based on constructing pseudo samples by resampling from the empirical distribution of the  $(\mathbf{x}, \mathbf{y})$  pairs does not accomplish this, but a smoothing procedure along the lines of Simar and Wilson (1998, 2000b) does.

Given the bootstrap values  $\widehat{\omega}_b^*$ ,  $b = 1, \dots, B$ , we can approximate  $c_\alpha$  by  $c_\alpha^*$ , defined by

$$\Pr(\widehat{\omega}^* \leq \widehat{\omega} - c_\alpha^* | H_0, \mathcal{S}_n) = \alpha, \tag{5.4}$$

which is the bootstrap analog of (5.1). The conditioning on  $\mathcal{S}_n$  in (5.4) is due to the fact that the distribution of  $(\widehat{\omega}^* - \widehat{\omega})$  in (5.2) is conditioned on  $\mathcal{S}_n$ . Mechanically, this involves sorting the values  $(\widehat{\omega}_b^* - \widehat{\omega})$ ,  $b = 1, \dots, B$ , by algebraic value, deleting  $(1 - \alpha) \times 100$  percent of the elements at the right-hand end of this sorted array, and then setting  $-c_\alpha^*$  equal to the right-hand endpoint of the resulting (sorted) array. Given (5.2) and (5.3), we obtain the bootstrap approximation

$$\Pr(\widehat{\omega} \leq \omega_0 - c_\alpha^* | H_0, \mathcal{S}_n) \approx \alpha \tag{5.5}$$

by substituting  $c_\alpha^*$  for  $c_\alpha$  in (5.1). Recalling that  $\omega_0 = 1$  under the null hypothesis in our tests, we reject the null if

$$\widehat{\omega} \leq 1 - c_\alpha^* \tag{5.6}$$

for a test of size  $\alpha$ .

Alternatively, the testing problem can be approached in terms of  $p$ -values (e.g., see Efron and Tibshirani, 1993, Chapters 15–16). Let  $\widehat{\omega}_{\text{obs}}$  be the observed value of our test statistic  $\widehat{\omega}$  in a particular application. Then the  $p$ -value for  $H_0$  is

$$p = \Pr(\widehat{\omega} \leq \widehat{\omega}_{\text{obs}} | H_0); \tag{5.7}$$

if  $p$  were known, we would reject the null hypothesis  $H_0$  when  $p$  is sufficiently small, say less than 0.05. Under  $H_0$ , (5.7) is equivalent to

$$p = \Pr(\widehat{\omega} - \omega_0 \leq \widehat{\omega}_{\text{obs}} - \omega_0 | H_0). \tag{5.8}$$

Using the bootstrap as described above, this can be approximated by

$$\widehat{p} = \Pr(\widehat{\omega}^* - \widehat{\omega} \leq \widehat{\omega}_{\text{obs}} - \omega_0 | H_0, \mathcal{S}_n). \tag{5.9}$$

Since conditionally on  $\mathcal{S}_n$ ,  $\widehat{\omega} = \widehat{\omega}_{\text{obs}}$ , (5.9) is equivalent to

$$\widehat{p} = \Pr(\widehat{\omega}^* \leq 2\widehat{\omega}_{\text{obs}} - \omega_0 | H_0, \mathcal{S}_n), \tag{5.10}$$

which can be easily evaluated from the empirical bootstrap distribution obtained from the Monte-Carlo simulation. Note that asymptotically, since  $\widehat{\omega}$  is consistent, the probability statement in (5.10) is asymptotically equivalent to

$$\widehat{p} = \Pr(\widehat{\omega}^* \leq \widehat{\omega}_{\text{obs}} | H_0, \mathcal{S}_n). \tag{5.11}$$

This expression is used in Efron and Tibshirani (1993, Chapter 16). In each of our Monte-Carlo experiments,  $\widehat{p}$  defined in (5.11) gives tests with size closer to the nominal size than does (5.10); consequently, all of our results in the next section use (5.11) rather than (5.10). If  $\alpha$  is the nominal size of our test, then we reject the null hypothesis  $H_0$  when  $\widehat{p} \leq \alpha$ .

Each of the test statistics proposed in the previous section is a function of the distance function estimators defined in (3.2)–(3.4). Therefore, bootstrap estimates of the distance function

estimators are needed to bootstrap the test statistics defined above. These can be obtained using the methods in Simar and Wilson (1998, 2000b) with resampling from a smooth estimate of the distribution of the  $\widehat{D}_n^{\text{crs}}$  so that the null hypothesis is maintained. For the test statistic  $\widehat{S}_{1n}^{\text{crs}}$ ,  $B$  bootstrap replications yield  $B$  bootstrap estimates

$$\widehat{S}_{1nb}^{\text{crs}*} = n^{-1} \sum_{i=1}^n \widehat{D}_{nb}^{\text{crs}*}(\mathbf{x}_i, \mathbf{y}_i) / \widehat{D}_{nb}^{\text{vrs}*}(\mathbf{x}_i, \mathbf{y}_i), \tag{5.12}$$

$b = 1, \dots, B$ . Following the procedures outlined by Eqs. (5.1)–(5.6) or (5.7)–(5.11), we can use these bootstrap values, together with the original estimate  $\widehat{S}_{1n}^{\text{crs}}$ , to determine  $c_\alpha^*$  or  $\widehat{p}$ .<sup>9</sup>

If the null hypothesis of constant returns to scale in Test #1 is rejected, we may adapt the above bootstrap procedure to Test #2 by resampling from an estimate of the distribution of the  $\widehat{D}_n^{\text{nirs}}(\mathbf{x}_i, \mathbf{y}_i)$  to maintain the null hypothesis in Test #2.

The binomial test suggested by (4.2) requires independence. This can be ensured by re-ordering the loops in the algorithms in Simar and Wilson (1998, 2000b). Rather than looping over  $n$  firms on each bootstrap replication, the binomial test requires that we perform  $B$  bootstrap replications for firm 1, then  $B$  bootstrap replications for firm 2, ..., and finally  $B$  bootstrap replications for firm  $n$ . The  $B$  bootstrap estimates for each firm are then used to determine the number  $r$  of rejections among  $n$  trials as described previously. The binomial probability in (4.2) can then be used to determine if enough rejections have occurred to reject  $H_0$ , the hypothesis of global CRS. Or, the bootstrap values  $s_{ib}^*$  for each  $i = 1, \dots, n$  can be used to perform  $n$  individual tests of size  $\alpha_\ell$  determined from (4.4), and then  $H_0$  can be rejected if these  $n$  tests produce one or more rejections. As before, the binomial test can easily be extended to Test #2.

<sup>9</sup> While we have used the statistic in (4.5) to illustrate our bootstrap procedure, the algorithm can be extended to estimate critical values for the test statistics given by (4.6)–(4.10) by merely rewriting (5.12) to reflect the chosen statistic.

## 6. Banker's tests

Banker (1996) proposed three statistics for testing the null hypotheses of CRS or NIRS against the alternative hypothesis of VRS. In terms of our notation, for the case of testing the null hypothesis of CRS, the first of Banker's statistics is given by

$$\widehat{F}_{1n}^{\text{crs}} = \sum_{i=1}^n \left[ \left( \widehat{D}^{\text{crs}}(\mathbf{x}_i, \mathbf{y}_i) \right)^{-1} - 1 \right] \bigg/ \sum_{i=1}^n \left[ \left( \widehat{D}^{\text{vrs}}(\mathbf{x}_i, \mathbf{y}_i) \right)^{-1} - 1 \right]. \quad (6.1)$$

Banker's second statistic amounts to

$$\widehat{F}_{2n}^{\text{crs}} = \sum_{i=1}^n \left[ \left( \widehat{D}^{\text{crs}}(\mathbf{x}_i, \mathbf{y}_i) \right)^{-1} - 1 \right]^2 \bigg/ \sum_{i=1}^n \left[ \left( \widehat{D}^{\text{vrs}}(\mathbf{x}_i, \mathbf{y}_i) \right)^{-1} - 1 \right]^2. \quad (6.2)$$

Banker argued that if  $D(\mathbf{x}_i, \mathbf{y}_i) \sim$  exponential, then  $\widehat{F}_{1n}^{\text{crs}} \xrightarrow{d} F_{2n, 2n}$ . Alternatively, if one assumes  $D(\mathbf{x}_i, \mathbf{y}_i) \sim$  half-normal, then  $\widehat{F}_{2n}^{\text{crs}} \xrightarrow{d} F_{n, n}$ .

Banker offered no guidance on how one might decide whether the true distance function values are distributed exponentially or half-normally across production units. Moreover, it is unclear why one should be willing to assume *any* distribution for purposes of hypothesis testing if the distance functions have been estimated nonparametrically. If the researcher were willing to make such an assumption for purposes of testing hypotheses, then why not make use of the distributional assumption in the estimation process? If the distributional assumption is correct, but not used in the estimation process, then presumably the resulting efficiency estimates will be statistically inefficient relative to an estimation procedure that makes use of the distributional information. On the other hand, if the distributional assumption is incorrect, then there is no reason to believe that tests based on the statistics in (6.1) and (6.2) would perform well in terms of size or power.

Kittelsen (1997) suggests two additional problems with Banker's approach. The distance func-

tion estimators used in (6.1) and (6.2) are biased in finite samples, and have very low convergence rates in high dimensions. Thus, even if the distributional assumptions are correct, the resulting statistics are not likely to be  $F$ -distributed in finite samples. Moreover, the numerators are correlated with the denominators, compounding the problem. Kittelsen presents Monte-Carlo evidence showing that these tests perform poorly in terms of size and power with  $p = q = 1$  and  $n = 100$ , even when the distributional assumptions are correct.

Banker's third statistic is given by

$$\widehat{K}_n^{\text{crs}} = \max \left[ \left( \widehat{D}^{\text{crs}}(\mathbf{x}_i, \mathbf{y}_i) \right)^{-1} - \left( \widehat{D}^{\text{vrs}}(\mathbf{x}_i, \mathbf{y}_i) \right)^{-1} \mid i = 1, \dots, n \right], \quad (6.3)$$

which is the Kolmogorov–Smirnov test statistic. Kim and Jennrich (1973) give an algorithm for computing the exact sampling distribution for this statistic when  $n^2 \leq 10^4$ , as well as a very good approximation for cases where  $n^2 > 10^4$ .

Each of the statistics defined by (6.1)–(6.3) can be adapted to the case of Test #2, where the null hypothesis is NIRS, by merely replacing  $\widehat{D}^{\text{crs}}(\mathbf{x}_i, \mathbf{y}_i)$  with  $\widehat{D}^{\text{nirs}}(\mathbf{x}_i, \mathbf{y}_i)$ . The approach suggested by the third statistic seems more reasonable than the first two, since no distributional assumptions are necessary. Moreover, using  $\widehat{K}_n^{\text{crs}}$  has a potential advantage over those defined in Section 4, since (possibly computationally burdensome) bootstrapping can be avoided. Unfortunately, however, the distance function estimators in (6.3) are not independent, as observed earlier, violating a key assumption in deriving the sampling distribution for the Kolmogorov–Smirnov statistic. Our Monte-Carlo experiments, as well as Kittelsen (1997), indicate that tests based on this statistic have a grossly incorrect size.

## 7. Evidence from Monte-Carlo experiments

As in any statistical testing problem based on asymptotics, the bootstrap procedures discussed in

Section 5 may involve errors (either type I or type II) in finite samples due to sampling variation in the distance function estimators as well as additional noise introduced by the resampling process itself.<sup>10</sup> As a result, the true sizes of the tests described in Section 5 may differ from the nominal value  $\alpha$ . In order to examine the performance of our bootstrap tests of returns to scale, we ran a series of Monte-Carlo experiments.

For each experiment, we used  $B = 2000$  bootstrap replications, and performed 1000 Monte-Carlo trials.<sup>11</sup> In addition, we used the homogeneous bootstrap described in Simar and Wilson (1998). For each Monte-Carlo trial, after generating the input and output data, we computed the relevant test statistics. Then, we applied the bootstrap procedures outlined in Section 5 to obtain critical values for each test statistic at various nominal significance levels. On each Monte-Carlo trial, we used a standard normal kernel function with a bandwidth that minimized mean weighted integrated square error.<sup>12</sup> We estimated the true size of each test for each experiment by recording the number of times the null hypothesis was rejected at each significance level over all Monte-Carlo trials, then dividing by the number of Monte-Carlo trials. In each of our experiments, the number of outputs,  $q$ , was set equal to one to simplify the data-simulation process and to reduce the computational burden. All input data ( $x_i$ ) were simulated by generating identically, independently

distributed (i.i.d.) pseudo random uniform deviates on the interval  $(1, 9)$ .<sup>13</sup>

We first examined the performance of the bootstrap in the context of Test #1. We ran 12 experiments, with one output ( $q = 1$ ), either one or two inputs ( $p \in \{1, 2\}$ ), and  $n \in \{20, 40, 60\}$ . To simulate the output data ( $y_i$ ) under the null hypothesis that  $\mathcal{P}^B$  exhibits CRS everywhere, we first generated i.i.d. pseudo random standard normal deviates  $v_i$ . Then the output for the  $i$ th simulated firm was computed as

$$y_i = \prod_{j=1}^p x_{ij}^{1/p} e^{-0.1|v_i|}, \quad i = 1, \dots, n. \quad (7.1)$$

The Monte-Carlo results for Test #1, with one input ( $p = 1$ ), appear in Table 1. Similarly, Monte-Carlo results for Test #1 with two inputs ( $p = 2$ ) appear in Table 2.

The first binomial test performs quite poorly in terms of its size, which is larger than the size of any of the other tests in Tables 1 and 2 in almost every instance. The second binomial test also performs quite poorly, but its size is far smaller than the nominal size in every instance, with the true size becoming smaller (for a given nominal size) as the sample size  $n$  increases. This seemingly perverse behavior is explained by considering the expression in (4.4). With  $n = 60$  and an overall nominal test size of 0.1, the individual tests comprising the second binomial test procedure should have size 0.00175 – apparently requiring a higher degree of accuracy in estimating the tail of the sampling distribution than is achieved with  $B = 2000$  bootstrap replications. The problem is worse when the overall nominal test size is 0.05 or less. While the second binomial test might be more accurate if a much larger number of bootstrap replications were used, doing so would substantially increase the computational burden of the test, and make our Monte-Carlo experiments infeasible.

<sup>10</sup> In particular, kernel estimators, while consistent, are slow to converge. Resampling from kernel estimates of the density of distance function estimates might be a significant source of noise in the bootstrap process.

<sup>11</sup> In effect, our bootstrap procedures are attempting to estimate the tails of sampling distributions. Researchers typically set  $B \approx 100$  when computing variances, and  $B \approx 1000$  when computing confidence intervals. Since estimating the tail of a distribution constitutes a more difficult problem, we chose  $B = 2000$  to ensure adequate coverage. For the experiment represented in Table 1 with  $n = 20$ , we also used 4000 and then 10,000 replications to check whether the choice of  $B$  would influence our results for the statistics  $\hat{S}_{1n}^{\text{CRS}}, \dots, \hat{S}_{6n}^{\text{CRS}}$ . In both instances, we obtained estimates of true size very close to those reported in Table 1 with  $B = 2000$ .

<sup>12</sup> Details of this are given in a technical appendix, available from the authors on request.

<sup>13</sup> Pseudo random uniform  $(0,1)$  deviates were generated using the multiplicative congruential generator with modulus  $(2^{31} - 1)$  and multiplier 16807 (see Lewis et al., 1969). Pseudo random normal deviates were generated via the Box–Muller method (see Press et al., 1986, pp. 202–203).

Table 1  
Monte-Carlo estimates of size – Test #1 ( $H_0$  : CRS) one input, one output ( $p = 1, q = 1$ )

| Statistic                | Nominal size |       |       |       |       |       |       |       |
|--------------------------|--------------|-------|-------|-------|-------|-------|-------|-------|
|                          | 0.3          | 0.25  | 0.2   | 0.15  | 0.1   | 0.05  | 0.02  | 0.01  |
| <i>n</i> = 20            |              |       |       |       |       |       |       |       |
| $\widehat{S}_{1n}^{crs}$ | 0.384        | 0.324 | 0.260 | 0.204 | 0.144 | 0.070 | 0.026 | 0.015 |
| $\widehat{S}_{2n}^{crs}$ | 0.368        | 0.305 | 0.249 | 0.195 | 0.130 | 0.070 | 0.025 | 0.011 |
| $\widehat{S}_{3n}^{crs}$ | 0.357        | 0.298 | 0.227 | 0.179 | 0.117 | 0.061 | 0.021 | 0.012 |
| $\widehat{S}_{4n}^{crs}$ | 0.562        | 0.504 | 0.441 | 0.363 | 0.279 | 0.176 | 0.090 | 0.062 |
| $\widehat{S}_{5n}^{crs}$ | 0.378        | 0.323 | 0.251 | 0.195 | 0.138 | 0.066 | 0.023 | 0.011 |
| $\widehat{S}_{6n}^{crs}$ | 0.522        | 0.465 | 0.383 | 0.304 | 0.223 | 0.124 | 0.054 | 0.025 |
| Binomial #1              | 0.504        | 0.492 | 0.458 | 0.361 | 0.291 | 0.203 | 0.101 | 0.045 |
| Binomial #2              | 0.165        | 0.136 | 0.104 | 0.077 | 0.049 | 0.028 | 0.013 | 0.008 |
| $\widehat{F}_{1n}^{crs}$ | 0.690        | 0.551 | 0.415 | 0.252 | 0.127 | 0.043 | 0.008 | 0.004 |
| $\widehat{F}_{2n}^{crs}$ | 0.999        | 0.041 | 0.016 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 |
| $\widehat{K}_n^{crs}$    | 0.389        | 0.389 | 0.389 | 0.184 | 0.184 | 0.076 | 0.028 | 0.007 |
| <i>n</i> = 40            |              |       |       |       |       |       |       |       |
| $\widehat{S}_{1n}^{crs}$ | 0.320        | 0.268 | 0.220 | 0.171 | 0.110 | 0.045 | 0.014 | 0.010 |
| $\widehat{S}_{2n}^{crs}$ | 0.313        | 0.256 | 0.210 | 0.160 | 0.102 | 0.040 | 0.012 | 0.010 |
| $\widehat{S}_{3n}^{crs}$ | 0.323        | 0.281 | 0.224 | 0.168 | 0.121 | 0.060 | 0.024 | 0.007 |
| $\widehat{S}_{4n}^{crs}$ | 0.471        | 0.400 | 0.343 | 0.281 | 0.206 | 0.109 | 0.049 | 0.025 |
| $\widehat{S}_{5n}^{crs}$ | 0.325        | 0.276 | 0.223 | 0.168 | 0.110 | 0.054 | 0.012 | 0.006 |
| $\widehat{S}_{6n}^{crs}$ | 0.477        | 0.416 | 0.340 | 0.264 | 0.191 | 0.106 | 0.035 | 0.014 |
| Binomial #1              | 0.513        | 0.433 | 0.392 | 0.351 | 0.273 | 0.172 | 0.070 | 0.041 |
| Binomial #2              | 0.067        | 0.058 | 0.046 | 0.032 | 0.024 | 0.007 | 0.006 | 0.000 |
| $\widehat{F}_{1n}^{crs}$ | 0.597        | 0.404 | 0.271 | 0.138 | 0.053 | 0.011 | 0.002 | 0.001 |
| $\widehat{F}_{2n}^{crs}$ | 0.676        | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| $\widehat{K}_n^{crs}$    | 0.321        | 0.186 | 0.186 | 0.186 | 0.092 | 0.013 | 0.007 | 0.004 |
| <i>n</i> = 60            |              |       |       |       |       |       |       |       |
| $\widehat{S}_{1n}^{crs}$ | 0.294        | 0.237 | 0.192 | 0.141 | 0.098 | 0.049 | 0.020 | 0.005 |
| $\widehat{S}_{2n}^{crs}$ | 0.278        | 0.229 | 0.182 | 0.138 | 0.095 | 0.046 | 0.015 | 0.004 |
| $\widehat{S}_{3n}^{crs}$ | 0.312        | 0.267 | 0.213 | 0.162 | 0.106 | 0.051 | 0.025 | 0.010 |
| $\widehat{S}_{4n}^{crs}$ | 0.411        | 0.359 | 0.300 | 0.253 | 0.184 | 0.095 | 0.038 | 0.022 |
| $\widehat{S}_{5n}^{crs}$ | 0.290        | 0.245 | 0.188 | 0.140 | 0.092 | 0.051 | 0.017 | 0.009 |
| $\widehat{S}_{6n}^{crs}$ | 0.435        | 0.376 | 0.315 | 0.235 | 0.166 | 0.092 | 0.046 | 0.018 |
| Binomial #1              | 0.466        | 0.453 | 0.404 | 0.346 | 0.276 | 0.181 | 0.083 | 0.052 |
| Binomial #2              | 0.057        | 0.045 | 0.037 | 0.029 | 0.022 | 0.009 | 0.000 | 0.000 |
| $\widehat{F}_{1n}^{crs}$ | 0.518        | 0.300 | 0.156 | 0.064 | 0.016 | 0.005 | 0.001 | 0.000 |
| $\widehat{F}_{2n}^{crs}$ | 0.001        | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| $\widehat{K}_n^{crs}$    | 0.166        | 0.088 | 0.088 | 0.053 | 0.027 | 0.011 | 0.003 | 0.002 |

Banker’s test (Banker, 1996) statistics defined in (6.1) and (6.2) also perform poorly. This is to be expected for the first two of Banker’s statistics, since the DGPs in (7.1) used for our Monte-Carlo experiments in Tables 1 and 2 do not reflect the distributional assumptions underlying these tests.<sup>14</sup>

Although  $\widehat{F}_{1n}^{crs}$  results in a test with true size close to the nominal size for nominal sizes in the range 0.1–0.01 and for  $n = 20$  in Table 1, its performance deteriorates in Table 1 as the sample size is increased. In Table 2, the true sizes of tests based on this statistic are much too large in every instance. The second of Banker’s statistics,  $\widehat{F}_{2n}^{crs}$ , yields tests with true size that is near zero for small nominal sizes. But, the true size of these tests jumps suddenly to values far in excess of the nominal size as the nominal size is increased. This behavior again

<sup>14</sup> The specification in (7.1) ensures that, for  $p = 1$ , we have  $D(x_i, y_i) = x_i e^{-0.1|v_i|} / x_i = e^{-0.1|v_i|}$ , and hence  $\log D(x_i, y_i) \sim$  half-normal.

Table 2  
 Monte-Carlo estimates of size – Test #1 ( $H_0$  : CRS) two inputs, one output ( $p = 2, q = 1$ )

| Statistic                | Nominal size |       |       |       |       |       |       |       |
|--------------------------|--------------|-------|-------|-------|-------|-------|-------|-------|
|                          | 0.3          | 0.25  | 0.2   | 0.15  | 0.1   | 0.05  | 0.02  | 0.01  |
| <i>n</i> = 20            |              |       |       |       |       |       |       |       |
| $\widehat{S}_{1n}^{crs}$ | 0.689        | 0.639 | 0.561 | 0.477 | 0.379 | 0.243 | 0.133 | 0.083 |
| $\widehat{S}_{2n}^{crs}$ | 0.660        | 0.589 | 0.516 | 0.431 | 0.330 | 0.202 | 0.113 | 0.070 |
| $\widehat{S}_{3n}^{crs}$ | 0.698        | 0.631 | 0.558 | 0.454 | 0.352 | 0.218 | 0.122 | 0.075 |
| $\widehat{S}_{4n}^{crs}$ | 0.963        | 0.938 | 0.919 | 0.894 | 0.839 | 0.701 | 0.485 | 0.345 |
| $\widehat{S}_{5n}^{crs}$ | 0.697        | 0.637 | 0.568 | 0.487 | 0.386 | 0.240 | 0.130 | 0.090 |
| $\widehat{S}_{6n}^{crs}$ | 0.754        | 0.702 | 0.629 | 0.544 | 0.437 | 0.286 | 0.153 | 0.095 |
| Binomial #1              | 0.657        | 0.616 | 0.585 | 0.390 | 0.325 | 0.224 | 0.144 | 0.072 |
| Binomial #2              | 0.441        | 0.381 | 0.334 | 0.264 | 0.195 | 0.100 | 0.042 | 0.020 |
| $\widehat{F}_{1n}^{crs}$ | 0.975        | 0.945 | 0.882 | 0.795 | 0.669 | 0.482 | 0.294 | 0.202 |
| $\widehat{F}_{2n}^{crs}$ | 1.000        | 1.000 | 1.000 | 1.000 | 0.517 | 0.004 | 0.000 | 0.000 |
| $\widehat{K}_n^{crs}$    | 0.614        | 0.614 | 0.614 | 0.357 | 0.357 | 0.164 | 0.069 | 0.014 |
| <i>n</i> = 40            |              |       |       |       |       |       |       |       |
| $\widehat{S}_{1n}^{crs}$ | 0.580        | 0.513 | 0.447 | 0.375 | 0.287 | 0.174 | 0.094 | 0.045 |
| $\widehat{S}_{2n}^{crs}$ | 0.526        | 0.459 | 0.399 | 0.324 | 0.234 | 0.141 | 0.065 | 0.033 |
| $\widehat{S}_{3n}^{crs}$ | 0.607        | 0.551 | 0.476 | 0.403 | 0.312 | 0.188 | 0.099 | 0.060 |
| $\widehat{S}_{4n}^{crs}$ | 0.893        | 0.867 | 0.829 | 0.785 | 0.702 | 0.546 | 0.411 | 0.318 |
| $\widehat{S}_{5n}^{crs}$ | 0.603        | 0.539 | 0.478 | 0.404 | 0.309 | 0.179 | 0.090 | 0.060 |
| $\widehat{S}_{6n}^{crs}$ | 0.705        | 0.651 | 0.590 | 0.495 | 0.399 | 0.269 | 0.153 | 0.086 |
| Binomial #1              | 0.732        | 0.614 | 0.580 | 0.549 | 0.490 | 0.367 | 0.193 | 0.154 |
| Binomial #2              | 0.334        | 0.281 | 0.225 | 0.194 | 0.134 | 0.076 | 0.042 | 0.000 |
| $\widehat{F}_{1n}^{crs}$ | 0.995        | 0.985 | 0.948 | 0.877 | 0.721 | 0.463 | 0.207 | 0.123 |
| $\widehat{F}_{2n}^{crs}$ | 1.000        | 1.000 | 1.000 | 1.000 | 0.761 | 0.001 | 0.001 | 0.001 |
| $\widehat{K}_n^{crs}$    | 0.873        | 0.747 | 0.747 | 0.573 | 0.573 | 0.224 | 0.132 | 0.065 |
| <i>n</i> = 60            |              |       |       |       |       |       |       |       |
| $\widehat{S}_{1n}^{crs}$ | 0.593        | 0.484 | 0.417 | 0.329 | 0.235 | 0.154 | 0.076 | 0.040 |
| $\widehat{S}_{2n}^{crs}$ | 0.509        | 0.427 | 0.349 | 0.291 | 0.193 | 0.115 | 0.060 | 0.036 |
| $\widehat{S}_{3n}^{crs}$ | 0.640        | 0.564 | 0.487 | 0.386 | 0.284 | 0.174 | 0.088 | 0.072 |
| $\widehat{S}_{4n}^{crs}$ | 0.861        | 0.841 | 0.800 | 0.749 | 0.632 | 0.479 | 0.313 | 0.244 |
| $\widehat{S}_{5n}^{crs}$ | 0.625        | 0.552 | 0.459 | 0.365 | 0.261 | 0.143 | 0.099 | 0.060 |
| $\widehat{S}_{6n}^{crs}$ | 0.788        | 0.729 | 0.647 | 0.566 | 0.438 | 0.265 | 0.142 | 0.087 |
| Binomial #1              | 0.697        | 0.682 | 0.572 | 0.584 | 0.425 | 0.293 | 0.169 | 0.091 |
| Binomial #2              | 0.180        | 0.180 | 0.157 | 0.123 | 0.078 | 0.033 | 0.000 | 0.000 |
| $\widehat{F}_{1n}^{crs}$ | 1.000        | 0.998 | 0.970 | 0.895 | 0.709 | 0.387 | 0.157 | 0.070 |
| $\widehat{F}_{2n}^{crs}$ | 1.000        | 1.000 | 0.998 | 0.987 | 0.000 | 0.000 | 0.000 | 0.000 |
| $\widehat{K}_n^{crs}$    | 0.902        | 0.784 | 0.784 | 0.644 | 0.500 | 0.351 | 0.142 | 0.888 |

merely reflects the misspecification of the distributional assumptions underlying these tests. While our Monte Carlo experiments are designed in such a way that it ensures that tests based on  $\widehat{F}_{1n}^{crs}$  and  $\widehat{F}_{2n}^{crs}$  will have incorrect size, we would expect similar results in real applications since there is no convincing reason why distance function values should be distributed either exponentially or half-normally. The Kolmogorov–Smirnov statistic in (6.3) suggested by Banker avoids the distributional assumptions underlying the other tests suggested by

Banker, but this test also performs poorly in our Monte-Carlo experiments, yielding tests with true size that is either much too large or too small relative to nominal size.

Of the remaining statistics,  $\widehat{S}_{2n}^{crs}$  consistently performs best in terms of size. In Table 1 with one input and one output, the true size is quite close to the nominal size with only 40 observations. In Table 2 where we consider two inputs, the true size is larger, but not as large as the other tests. Moreover, as sample size increases, we see that the

size of tests based on  $\widehat{S}_{2n}^{crs}$  decreases toward the nominal size more rapidly than for other tests we consider. The increased size in Table 2, relative to Table 1, merely reflect the curse of dimensionality inherent in non-parametric estimators of technical efficiency; with more than 60 observations, we would expect the size of our bootstrap tests to improve, but it is computationally infeasible to demonstrate this in Monte-Carlo experiments.

Among the statistics  $\widehat{S}_{1n}^{crs}, \dots, \widehat{S}_{6n}^{crs}, \widehat{S}_{4n}^{crs}$  performs most poorly in both Tables 1 and 2, although its performance improves with sample size. The next-poorest performance in Tables 1 and 2 is by  $\widehat{S}_{6n}^{crs}$ . There are no compelling differences in the performance of  $\widehat{S}_{1n}^{crs}, \widehat{S}_{3n}^{crs}$ , and  $\widehat{S}_{5n}^{crs}$  in Tables 1 and 2, but these are dominated in almost every instance by the performance of  $\widehat{S}_{2n}^{crs}$ .

In evaluating the performance of our test statistics and bootstrap procedure for Test #2, we must simulate the data under the null hypothesis of NIRS. We simulate the input data as before. For the case of one input, we compute the output for the  $i$ th simulated firm as

$$y_i = \begin{cases} x_i e^{-0.1|v_i|} & \text{if } x_i \leq 5; \\ (\frac{5}{2} + \frac{1}{2}x_i)e^{-0.1|v_i|} & \text{otherwise.} \end{cases} \quad (7.2)$$

For the case of two inputs, we compute the output for the  $i$ th simulated firm as

$$y_i = \begin{cases} \sqrt{x_{i1}}\sqrt{x_{i2}}e^{-0.1|v_i|} & \text{if } x_{i1} + x_{i2} \leq 11, \\ \left\{ \sqrt{\frac{x_{i2}}{x_{i1}}}(11)\left(\frac{x_{i2}}{x_{i1}} + 1\right)^{-1} \right. \\ \quad \left. + \left[ x_{i1} - \left(\frac{x_{i2}}{x_{i1}}\right)(11)\left(\frac{x_{i2}}{x_{i1}} + 1\right)^{-1} \right]^{1/4} \right. \\ \quad \left. \times \left[ x_{i2} - (11)\left(\frac{x_{i2}}{x_{i1}} + 1\right)^{-1} \right]^{1/4} \right\} e^{-0.1|v_i|} & \text{otherwise.} \end{cases} \quad (7.3)$$

This ensures continuity in the simulated technology. We ran 12 experiments, analogous to those for Test #1. Results for the case of one input are shown in Table 3, while results for the case of two inputs are shown in Table 4.

The results in Tables 3 and 4 are qualitatively similar to those in Tables 1 and 2. The binomial

tests perform poorly in every case, as do Banker’s tests. The binomial tests and Banker’s exponential and half-normal tests do not improve with sample size; the Kolmogorov–Smirnov test improves only very slightly with sample size. Among the remaining statistics,  $\widehat{S}_{2n}^{nirs}$  has estimated true size closer to nominal size than the other statistics in almost every case. Surprisingly, in view of the previous results,  $\widehat{S}_{4n}^{nirs}$  does not stand out as having exceptionally bad performance in Table 3, but  $\widehat{S}_{3n}^{nirs}$  does. Note that in Table 4, as the sample size increases from  $n = 20$  to  $n = 60$ , the estimated sizes diverge from the nominal sizes for the first six test statistics in many cases. Statistical consistency does not necessarily imply any type of monotonic convergence; the phenomena in Table 4 merely reflects the curse of dimensionality. Apparently, for the test of the null hypothesis of NIRS, with  $p + q = 3$ , even  $n = 60$  is a very small sample size.

There is, of course, a trade-off between size and power in hypothesis testing. Therefore, we also examined the power of tests based on the statistics  $\widehat{S}_{1n}^{crs}, \dots, \widehat{S}_{6n}^{crs}$ , which performed best in our Monte-Carlo experiments in terms of size. Since it is typically impossible to analytically derive power functions when either the null or alternative hypotheses are composite, examining the power of our tests requires additional Monte-Carlo experiments. Since these experiments are computationally burdensome, we focus only on Test #1 for the case of one input and one output, and sample size  $n = 20$ .

To examine the power function for our test, data must be generated such that the null hypothesis of CRS is false. Let  $\varphi \in (\frac{\pi}{2}, \pi]$ . We simulated inputs  $x_i$  by generating i.i.d. pseudo random uniform deviates on the interval  $(-5[1 - \tan(\frac{3\pi}{4} - \frac{\varphi}{2})]/\tan(\frac{3\pi}{4} - \frac{\varphi}{2}), 9)$ . Output data ( $y_i$ ) were simulated by first generating i.i.d. pseudo random standard normal deviates  $v_i$  as before, and then setting

$$y_i = \begin{cases} \left\{ 5\left[1 - \tan\left(\frac{3\pi}{4} - \frac{\varphi}{2}\right)\right] + \tan\left(\frac{3\pi}{4} - \frac{\varphi}{2}\right)x_i \right\} e^{-0.1|v_i|} & \text{if } x < 5, \\ \left\{ 5\left[1 - \tan\left(\frac{\varphi}{2} - \frac{\pi}{4}\right)\right] + \tan\left(\frac{\varphi}{2} - \frac{\pi}{4}\right)x_i \right\} e^{-0.1|v_i|} & \text{otherwise} \end{cases} \quad (7.4)$$

for  $i = 1, \dots, n$ . For  $\varphi = \pi$ , (7.4) reduces to  $y_i = x_i e^{-0.1|v_i|}$ , where the technology exhibits CRS

Table 3  
 Monte-Carlo estimates of size – Test #2 ( $H_0 : \text{NIRS}$ ) one input, one output ( $p = 1, q = 1$ )

| Statistic                        | Nominal size |       |       |       |       |       |       |       |
|----------------------------------|--------------|-------|-------|-------|-------|-------|-------|-------|
|                                  | 0.3          | 0.25  | 0.2   | 0.15  | 0.1   | 0.05  | 0.02  | 0.01  |
| <i>n</i> = 20                    |              |       |       |       |       |       |       |       |
| $\widehat{S}_{1n}^{\text{nirs}}$ | 0.481        | 0.436 | 0.393 | 0.330 | 0.264 | 0.164 | 0.078 | 0.050 |
| $\widehat{S}_{2n}^{\text{nirs}}$ | 0.475        | 0.429 | 0.384 | 0.323 | 0.258 | 0.154 | 0.076 | 0.049 |
| $\widehat{S}_{3n}^{\text{nirs}}$ | 0.990        | 0.976 | 0.943 | 0.879 | 0.741 | 0.506 | 0.261 | 0.136 |
| $\widehat{S}_{4n}^{\text{nirs}}$ | 0.622        | 0.539 | 0.454 | 0.380 | 0.285 | 0.174 | 0.077 | 0.045 |
| $\widehat{S}_{5n}^{\text{nirs}}$ | 0.561        | 0.505 | 0.445 | 0.374 | 0.294 | 0.190 | 0.100 | 0.059 |
| $\widehat{S}_{6n}^{\text{nirs}}$ | 0.516        | 0.470 | 0.424 | 0.371 | 0.301 | 0.195 | 0.100 | 0.056 |
| Binomial #1                      | 0.995        | 0.991 | 0.982 | 0.778 | 0.393 | 0.288 | 0.233 | 0.127 |
| Binomial #2                      | 0.704        | 0.665 | 0.598 | 0.525 | 0.391 | 0.186 | 0.078 | 0.040 |
| $\widehat{F}_{1n}^{\text{nirs}}$ | 0.221        | 0.142 | 0.083 | 0.040 | 0.012 | 0.000 | 0.000 | 0.000 |
| $\widehat{F}_{2n}^{\text{nirs}}$ | 0.000        | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| $\widehat{K}_n^{\text{nirs}}$    | 0.006        | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| <i>n</i> = 40                    |              |       |       |       |       |       |       |       |
| $\widehat{S}_{1n}^{\text{nirs}}$ | 0.457        | 0.404 | 0.336 | 0.270 | 0.204 | 0.116 | 0.053 | 0.031 |
| $\widehat{S}_{2n}^{\text{nirs}}$ | 0.448        | 0.391 | 0.320 | 0.253 | 0.188 | 0.112 | 0.049 | 0.030 |
| $\widehat{S}_{3n}^{\text{nirs}}$ | 0.997        | 0.985 | 0.958 | 0.916 | 0.836 | 0.641 | 0.421 | 0.290 |
| $\widehat{S}_{4n}^{\text{nirs}}$ | 0.549        | 0.465 | 0.384 | 0.311 | 0.219 | 0.142 | 0.070 | 0.036 |
| $\widehat{S}_{5n}^{\text{nirs}}$ | 0.537        | 0.488 | 0.427 | 0.374 | 0.282 | 0.173 | 0.076 | 0.044 |
| $\widehat{S}_{6n}^{\text{nirs}}$ | 0.542        | 0.493 | 0.432 | 0.358 | 0.279 | 0.164 | 0.086 | 0.048 |
| Binomial #1                      | 1.000        | 1.000 | 1.000 | 0.969 | 0.572 | 0.477 | 0.375 | 0.348 |
| Binomial #2                      | 0.801        | 0.764 | 0.700 | 0.597 | 0.395 | 0.150 | 0.084 | 0.000 |
| $\widehat{F}_{1n}^{\text{nirs}}$ | 0.142        | 0.071 | 0.027 | 0.005 | 0.001 | 0.000 | 0.000 | 0.000 |
| $\widehat{F}_{2n}^{\text{nirs}}$ | 0.001        | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| $\widehat{K}_n^{\text{nirs}}$    | 0.006        | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| <i>n</i> = 60                    |              |       |       |       |       |       |       |       |
| $\widehat{S}_{1n}^{\text{nirs}}$ | 0.403        | 0.348 | 0.285 | 0.236 | 0.185 | 0.108 | 0.057 | 0.033 |
| $\widehat{S}_{2n}^{\text{nirs}}$ | 0.389        | 0.331 | 0.271 | 0.230 | 0.175 | 0.099 | 0.051 | 0.031 |
| $\widehat{S}_{3n}^{\text{nirs}}$ | 0.994        | 0.990 | 0.975 | 0.950 | 0.886 | 0.723 | 0.481 | 0.344 |
| $\widehat{S}_{4n}^{\text{nirs}}$ | 0.466        | 0.395 | 0.346 | 0.277 | 0.198 | 0.119 | 0.054 | 0.026 |
| $\widehat{S}_{5n}^{\text{nirs}}$ | 0.472        | 0.424 | 0.373 | 0.307 | 0.236 | 0.144 | 0.075 | 0.040 |
| $\widehat{S}_{6n}^{\text{nirs}}$ | 0.510        | 0.448 | 0.394 | 0.325 | 0.254 | 0.155 | 0.090 | 0.048 |
| Binomial #1                      | 1.000        | 1.000 | 1.000 | 0.987 | 0.557 | 0.491 | 0.446 | 0.409 |
| Binomial #2                      | 0.844        | 0.792 | 0.697 | 0.556 | 0.344 | 0.107 | 0.000 | 0.000 |
| $\widehat{F}_{1n}^{\text{nirs}}$ | 0.091        | 0.035 | 0.013 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 |
| $\widehat{F}_{2n}^{\text{nirs}}$ | 0.000        | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| $\widehat{K}_n^{\text{nirs}}$    | 0.001        | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

everywhere. For  $\varphi < \pi$ , the technology has a kink at the point (5,5) in  $(x,y)$ -space, so that the technology form angles of  $\frac{\pi-\varphi}{2}$  radians on either side of the point (5,5). The technology itself forms a right angle at (5,5) when  $\varphi = \frac{\pi}{2}$ . By varying the value of  $\varphi$  in our Monte-Carlo experiments, we can control the degree of departure from the null hypothesis of CRS.

Using the DGP represented by (7.4), we performed Monte-Carlo experiments with  $\varphi \in \{\pi - \kappa$

$\frac{\pi}{2} | \kappa = 0.01, 0.02, \dots, 0.09, 0.1, 0.2, 0.3, 0.4\}$  to estimate the power function  $\mathcal{P}(\varphi)$  for each statistic  $\widehat{S}_{1n}^{\text{crs}}, \dots, \widehat{S}_{6n}^{\text{crs}}$ . Each experiment consisted of  $M = 1000$  Monte-Carlo trials, with  $B = 2000$  bootstrap replications on each of these trials. For nominal size equal to 0.05,  $n = 20$ , and  $p = q = 1$ , this yielded the results shown in Fig. 1. These results indicate that tests based on  $\widehat{S}_{2n}^{\text{crs}}$  have good power even for small departures from the null hypothesis of CRS. For instance, with  $\kappa = 0.01$ ,  $\mathcal{P}(\varphi) = 0.619$ ; for  $\kappa = 0.02$ ,  $(\varphi) = 0.866$ , and for

Table 4  
 Monte-Carlo estimates of size – Test #2 ( $H_0$  : NIRS) one input, one output ( $p = 2, q = 1$ )

| Statistic                 | Nominal size |       |       |       |       |       |       |       |
|---------------------------|--------------|-------|-------|-------|-------|-------|-------|-------|
|                           | 0.3          | 0.25  | 0.2   | 0.15  | 0.1   | 0.05  | 0.02  | 0.01  |
| <i>n</i> = 20             |              |       |       |       |       |       |       |       |
| $\widehat{S}_{1n}^{nirs}$ | 0.755        | 0.698 | 0.642 | 0.577 | 0.495 | 0.366 | 0.246 | 0.172 |
| $\widehat{S}_{2n}^{nirs}$ | 0.740        | 0.682 | 0.625 | 0.553 | 0.460 | 0.336 | 0.222 | 0.151 |
| $\widehat{S}_{3n}^{nirs}$ | 0.919        | 0.843 | 0.721 | 0.559 | 0.397 | 0.256 | 0.183 | 0.127 |
| $\widehat{S}_{4n}^{nirs}$ | 0.859        | 0.833 | 0.798 | 0.765 | 0.715 | 0.597 | 0.444 | 0.329 |
| $\widehat{S}_{5n}^{nirs}$ | 0.812        | 0.766 | 0.697 | 0.621 | 0.557 | 0.416 | 0.260 | 0.173 |
| $\widehat{S}_{6n}^{nirs}$ | 0.792        | 0.748 | 0.696 | 0.615 | 0.530 | 0.402 | 0.262 | 0.187 |
| Binomial #1               | 0.907        | 0.752 | 0.367 | 0.088 | 0.044 | 0.009 | 0.007 | 0.000 |
| Binomial #2               | 0.156        | 0.123 | 0.097 | 0.067 | 0.033 | 0.018 | 0.009 | 0.003 |
| $\widehat{F}_{1n}^{nirs}$ | 0.731        | 0.623 | 0.527 | 0.411 | 0.294 | 0.160 | 0.082 | 0.053 |
| $\widehat{F}_{2n}^{nirs}$ | 1.000        | 0.999 | 0.417 | 0.024 | 0.001 | 0.000 | 0.000 | 0.000 |
| $\widehat{K}_n^{nirs}$    | 0.081        | 0.081 | 0.081 | 0.016 | 0.016 | 0.004 | 0.002 | 0.000 |
| <i>n</i> = 40             |              |       |       |       |       |       |       |       |
| $\widehat{S}_{1n}^{nirs}$ | 0.847        | 0.806 | 0.755 | 0.689 | 0.580 | 0.426 | 0.298 | 0.211 |
| $\widehat{S}_{2n}^{nirs}$ | 0.820        | 0.776 | 0.722 | 0.635 | 0.533 | 0.389 | 0.255 | 0.168 |
| $\widehat{S}_{3n}^{nirs}$ | 0.984        | 0.959 | 0.910 | 0.808 | 0.669 | 0.547 | 0.466 | 0.380 |
| $\widehat{S}_{4n}^{nirs}$ | 0.902        | 0.875 | 0.852 | 0.779 | 0.689 | 0.538 | 0.376 | 0.268 |
| $\widehat{S}_{5n}^{nirs}$ | 0.911        | 0.887 | 0.854 | 0.800 | 0.709 | 0.562 | 0.391 | 0.292 |
| $\widehat{S}_{6n}^{nirs}$ | 0.893        | 0.865 | 0.838 | 0.773 | 0.695 | 0.546 | 0.374 | 0.281 |
| Binomial #1               | 1.000        | 0.946 | 0.385 | 0.238 | 0.245 | 0.065 | 0.012 | 0.003 |
| Binomial #2               | 0.158        | 0.102 | 0.075 | 0.051 | 0.023 | 0.012 | 0.006 | 0.000 |
| $\widehat{F}_{1n}^{nirs}$ | 0.830        | 0.715 | 0.582 | 0.409 | 0.256 | 0.093 | 0.021 | 0.008 |
| $\widehat{F}_{2n}^{nirs}$ | 1.000        | 0.996 | 0.001 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 |
| $\widehat{K}_n^{nirs}$    | 0.180        | 0.076 | 0.076 | 0.035 | 0.035 | 0.003 | 0.000 | 0.000 |
| <i>n</i> = 60             |              |       |       |       |       |       |       |       |
| $\widehat{S}_{1n}^{nirs}$ | 0.840        | 0.818 | 0.773 | 0.729 | 0.674 | 0.542 | 0.353 | 0.287 |
| $\widehat{S}_{2n}^{nirs}$ | 0.818        | 0.784 | 0.740 | 0.707 | 0.619 | 0.443 | 0.342 | 0.199 |
| $\widehat{S}_{3n}^{nirs}$ | 0.994        | 0.994 | 0.960 | 0.872 | 0.784 | 0.707 | 0.674 | 0.619 |
| $\widehat{S}_{4n}^{nirs}$ | 0.895        | 0.862 | 0.796 | 0.752 | 0.653 | 0.553 | 0.344 | 0.278 |
| $\widehat{S}_{5n}^{nirs}$ | 0.961        | 0.950 | 0.939 | 0.873 | 0.807 | 0.686 | 0.498 | 0.443 |
| $\widehat{S}_{6n}^{nirs}$ | 0.950        | 0.939 | 0.884 | 0.818 | 0.774 | 0.663 | 0.520 | 0.397 |
| Binomial #1               | 1.000        | 1.000 | 0.551 | 0.276 | 0.234 | 0.147 | 0.008 | 0.000 |
| Binomial #2               | 0.098        | 0.089 | 0.035 | 0.017 | 0.009 | 0.009 | 0.000 | 0.000 |
| $\widehat{F}_{1n}^{nirs}$ | 0.902        | 0.775 | 0.617 | 0.437 | 0.238 | 0.076 | 0.016 | 0.003 |
| $\widehat{F}_{2n}^{nirs}$ | 1.000        | 0.978 | 0.616 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| $\widehat{K}_n^{nirs}$    | 0.263        | 0.144 | 0.144 | 0.073 | 0.032 | 0.014 | 0.002 | 0.000 |

$\kappa = 0.03$ , we find  $\mathcal{P}(\varphi) = 0.913$ . By contrast, for tests based on  $\widehat{S}_{3n}^{crs}$ , the power is much lower as indicated by Fig. 1. Tests based on  $\widehat{S}_{4n}^{crs}$  also have low power.

### 8. Conclusions

As with most Monte-Carlo studies, the scope of our experiments is rather limited in terms of the number of cases considered due to the computa-

tional burden. Each experiment requires solution of  $[2 \times M \times n \times (B + 1)]$  LPs, where  $M$  is the number of Monte-Carlo trials, with the time required for each LP increasing with the number of observations and the dimensionality of the input/output space. Nonetheless, our results provide some insight into the likely performance of non-parametric tests regarding returns to scale in small-sample applications.

In most instances, we find that the true sizes of our tests are reasonably close to the nominal sizes,

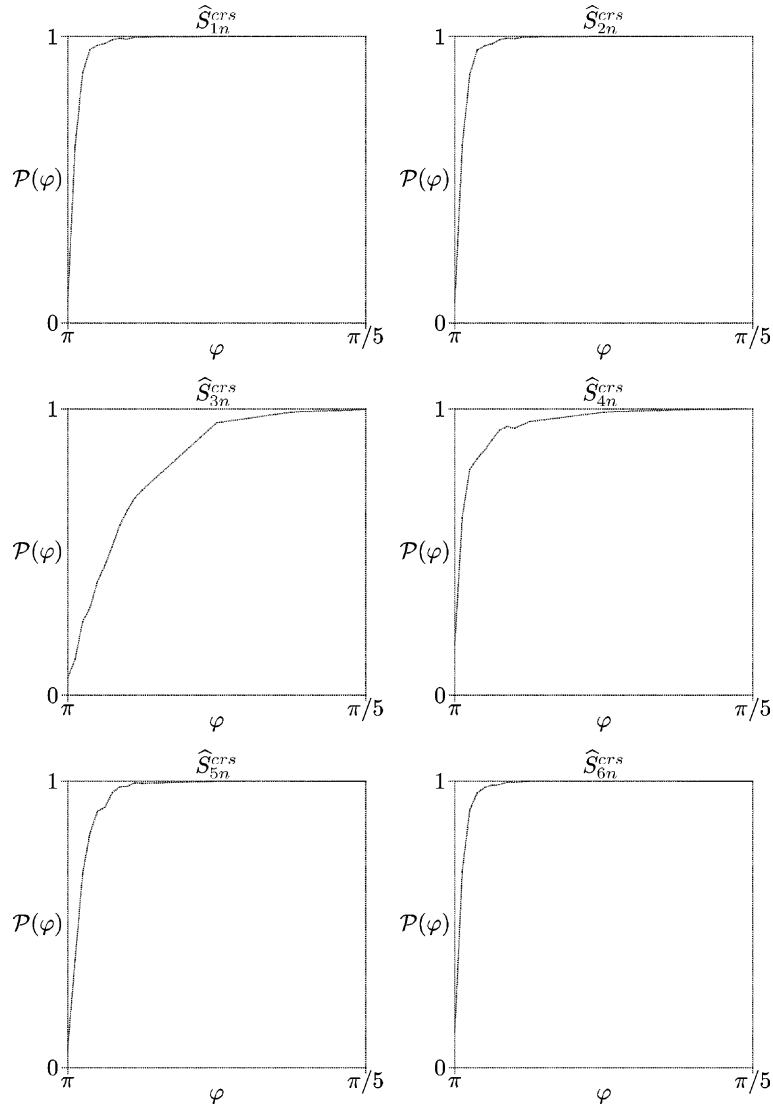


Fig. 1. Power functions ( $\alpha = 0.05, n = 20, p = q = 1$ ).

or that the difference is comparable to corresponding differences that have been reported for other statistics used by econometricians in testing, for example, hypotheses regarding unit roots in time series data. Indeed, since there seem to be no other formal statistical tests of regarding returns to scale in the context of DEA models at this time, our tests seem to offer the best alternative. In our experiments, we find that the true sizes of our tests typically exceed the nominal sizes, which would

tend to cause excessive type-I errors in small samples.

Whether this is a great problem may depend on the purpose of the tests; if the purpose is to decide which estimator of the production set to use in estimating distance functions, then our results indicate that the choice will be too often in favor of estimators that do not assume constant returns to scale. This means that distance function estimates may be less statistically efficient than they could be

when the true technology is CRS. But, if the true technology is not CRS everywhere and the CRS estimator of the production set is used, distance function estimates will be statistically inconsistent. So, for this purpose, type-I errors may be less serious than type-II errors, since statistical inconsistency is a worse problem than statistical inefficiency. If, on the other hand, the researcher is trying to infer returns to scale along the technology to answer economic policy questions, then armed with the information provided by this study, careful researchers should perhaps choose a smaller nominal size than they otherwise would, particularly as the dimensionality of the input/output space increases for a given number of observations.

## References

- Afriat, S., 1972. Efficiency estimation of production functions. *International Economic Review* 13, 568–598.
- Banker, R.D., 1993. Maximum likelihood, consistency and data envelopment analysis: A statistical foundation. *Management Science* 39, 1265–1273.
- Banker, R.D., 1996. Hypothesis tests using data envelopment analysis. *Journal of Productivity Analysis* 7, 139–159.
- Byrnes, P., Grosskopf, S., Hayes, K., 1986. Efficiency and ownership: Further evidence. *Review of Economics and Statistics* 68, 337–341.
- Debreu, G., 1951. The coefficient of resource utilization. *Econometrica* 19, 273–292.
- Dusansky, R., Wilson, P.W., 1994. Technical efficiency in the decentralized care of the developmentally disabled. *Review of Economics and Statistics* 76, 340–345.
- Efron, B., 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7, 1–16.
- Efron, B., Tibshirani, R.J., 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- Färe, R., 1988. *Fundamentals of Production Theory*. Springer, Berlin.
- Färe, R., Grosskopf, S., 1985. A nonparametric cost approach to scale efficiency. *Scandinavian Journal of Economics* 87, 594–604.
- Färe, R., Grosskopf, S., 2000. Theory and application of directional distance functions. *Journal of Productivity Analysis* 13, 93–103.
- Färe, R., Grosskopf, S., Lovell, C.A.K., 1985. *The Measurement of Efficiency of Production*. Kluwer-Nijhoff Publishing, Boston.
- Färe, R., Grosskopf, S., Roos, P., 1997. Profit productivity and quality: A directional distance function approach, unpublished working paper. Department of Economics, Southern Illinois University, Carbondale, IL, 62901.
- Farrell, M.J., 1957. The measurement of productive efficiency. *Journal of the Royal Statistical Society A* 120, 253–281.
- Ferrier, G.D., 1994. Ownership type, property rights, and relative efficiency. In: Charnes, A., Cooper, W., Lewin, A.Y., Seiford, L.M. (Eds.), *Data Envelopment Analysis: Theory, Methodology and Applications*. Kluwer Academic Publishers, Boston, pp. 273–283.
- 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, 220–228.
- Grosskopf, S., 1986. The role of the reference technology in measuring productive efficiency. *The Economic Journal* 96, 499–513.
- Grosskopf, S., Valdmanis, V., 1987. Measuring hospital performance: A nonparametric approach. *Journal of Health Economics* 6, 89–107.
- Kim, P.J., Jennrich, R.I., 1973. Tables of the exact sampling distribution of the two sample Kolmogorov–Smirnov criterion  $D_{mn}$  ( $m < n$ ). In: Harter, H.L., Owen, D.B. (Eds.), *Selected Tables in Mathematical Statistics*, vol. 1, American Mathematical Society, Providence, RI.
- Kittelsen, S.A.C., 1997. Monte Carlo simulations of DEA efficiency measures and hypothesis tests, Unpublished working paper. SNF Oslo, Gaustadalleen 21, N-0371 Oslo, Norway.
- Kneip, A., Park, B.U., Simar, L., 1998. A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econometric Theory* 14, 783–793.
- Korostelev, A., Simar, L., Tsybakov, A., 1995. On estimation of monotone and convex boundaries. *Publications de l'Institut de l'Université de Paris XXXIX* 1, 3–18.
- Lewis, P.A., Goodman, A.S., Miller, J.M., 1969. A pseudo-random number generator for the System/360. *IBM Systems Journal* 8, 136–146.
- Löthgren, M., Tambour, M., 1996. Alternative Approaches to Estimate Returns to Scale in DEA Models, working paper no. 90. Stockholm School of Economics.
- Löthgren, M., Tambour, M., 1999. Scale efficiency and scale elasticity in DEA models – A bootstrapping approach. *Applied Economics* 31, 1231–1237.
- Lovell, C.A.K., 1993. Production frontiers and productive efficiency. In: Fried, H., Knox Lovell, C.A., Schmidt, S.S. (Eds.), *The Measurement of Productive Efficiency: Techniques and Applications*. Oxford University Press, Oxford, pp. 3–67.
- Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1986. *Numerical Recipes*. Cambridge University Press, Cambridge.
- Seiford, L.M., 1997. A bibliography for data envelopment analysis (1978–1996). *Annals of Operations Research* 73, 393–438.
- Shephard, R.W., 1970. *Theory of Cost and Production Functions*. Princeton University Press, Princeton, NJ.
- Simar, L., Wilson, P.W., 1998. Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science* 44, 49–61.

- Simar, L., Wilson, P.W., 1999a. Some problems with the Ferrier/Hirschberg bootstrap idea. *Journal of Productivity Analysis* 11, 67–80.
- Simar, L., Wilson, P.W., 1999b. Of course we can bootstrap DEA scores! But does it mean anything? Logic trumps wishful thinking. *Journal of Productivity Analysis* 11, 93–97.
- Simar, L., Wilson, P.W., 2000a. Statistical inference in nonparametric frontier models: The state of the art. *Journal of Productivity Analysis* 13, 49–78.
- Simar, L., Wilson, P.W., 2000b. A general methodology for bootstrapping in nonparametric frontier models. *Journal of Applied Statistics* 27, 779–802.