INSTITUT DE
STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

0933

INFERENCE BY SUBSAMPLING IN NONPARAMETRIC FRONTIER MODELS

SIMAR, L. and P.W. WILSON

Inference by Subsampling in Nonparametric Frontier Models

Léopold Simar Paul W. Wilson*

December 2009

Abstract

This paper provides a simple, tractable bootstrap for use with Data Envelopment Analysis (DEA) estimators in nonparametric frontier models. It is well-known that a naive bootstrap yields inconsistent inference in this context. However, subsampling—where for a sample of size n bootstrap pseudo-samples of size $m < n$ are drawn from the empirical distribution of pairs of observed input-output vectors—provides consistent inference, although coverages are quite sensitive to the choice of subsample size m. We show that a simple, data-based rule for selecting m gives confidence interval estimates with good coverage properties. In addition, we show that subsampling performs well for testing hypotheses about returns to scale and other features of the model when a similar data-based rule is used to select m. Our methods (i) allow for heterogeneity in the inefficiency process, and unlike previous methods, (ii) do not require multivariate kernel smoothing, and (iii) avoid the need for solutions of intermediate linear programs.

Keywords: nonparametric frontier, efficiency, bootstrap, nonparametric testing, testing convexity, testing returns to scale.

*Simar: Institut de Statistique, Université Catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium and IDEI, Toulouse School of Economics, F 31000 Toulouse, France; email leopold.simar@uclouvain.be. Wilson: The John E. Walker Department of Economics, 222 Sirrine Hall, Clemson University, Clemson, South Carolina 29634–1309, USA; email wilson@clemson.edu. Financial support from the "Interuniversity Attraction Pole", Phase VI (No. P6/03) from the Belgian Government (Belgian Science Policy) and from the Chair of Excellency "Pierre de Fermat", Région Midi-Pyrénées, France are gratefully acknowledged. This work was made possible by the Palmetto cluster maintained by Clemson Computing and Information Technology (CCIT) at Clemson University; we are grateful for technical support by the staff of CCIT. The usual caveats apply.

1 Introduction

This paper develops testing procedures based on sub-sampling while using the ideas developed by Bickel and Sakov (2008) and Politis et al. (2001) for choosing the appropriate size of the sub-samples. We provide evidence from extensive Monte Carlo experiments showing that these procedures work rather well in finite samples, both in terms of the achieved level of the tests as well as power of the tests. The computational burden of our procedure is modest, and is comparable to the computational requirements of the method proposed by Kneip et al. (2009) for estimating confidence intervals for efficiency of a particular point. Although the methods proposed here can be used to estimate confidence intervals, their main usefulness is for testing hypotheses about the structure of the underlying non-parametric model.

Non-parametric data envelopment analysis (DEA) estimators have been widely used in studies of productive efficiency by firms, government agencies, national economies, and other decision-making units; Gattoufi et al. (2004) cite more than 1,800 published articles appearing in more than 400 journals. DEA estimators rely on linear programming methods along the lines of Charnes et al. (1978, 1979) and Färe et al. (1985) to estimate efficiency measures proposed by Debreu (1951), Farrell (1957), Shephard (1970), and others. DEA estimators measure efficiency relative to an *estimate* of an unobserved *true* frontier, conditional on observed data resulting from an underlying data-generating process (DGP). Under certain assumptions the DEA *frontier* estimator is a consistent, maximum likelihood estimator (Banker, 1993), with rates of convergence given by Korostelev et al. (1995). Consistency and convergence rates of DEA *efficiency* estimators have been established by Kneip et al. (1998); see Simar and Wilson (2000b) for a survey of the statistical properties of DEA estimators.

Although DEA estimators have been widely used, inference about the underlying model structure or the efficiencies that are estimated remains problematic. Gijbels et al. (1999) derived the asymptotic distribution of a DEA efficiency estimators in the case of one input and one output, permitting classical inference in this special, limited case. However, one of the attractive features of DEA estimators is that they simultaneously allow both multiple inputs as well as multiple outputs. Simar and Wilson (1998, 2000a) proposed bootstrap methods for inference about efficiency based on DEA estimators in a multivariate framework,

and Simar and Wilson (2001a, 2001b) proposed bootstrap methods for testing hypotheses about the structure of the underlying nonparametric model of production, but consistency of these procedures has not been established. Banker (1993, 1996) proposed tests of model structure based on ad-hoc distributional assumptions, but simulation results obtained by Kittelsen (1999) and Simar and Wilson (2001a) show that these tests perform poorly in terms of both size and power.

Jeong (2004) derived the limiting distribution of DEA efficiency estimators under variable returns to scale for the special case $p = 1,\ q \geq 1$ in the input orientation (or $p \geq 1,\ q = 1$ in the output orientation), where p and q denote the numbers of inputs and outputs, respectively. Kneip et al. (2008) and Park et al. (2009) derived the limiting distributions of DEA efficiency estimators under variable returns to scale and constant returns to scale (respectively), with arbitrary numbers of inputs and outputs. These distributions contain several unknown quantities, and are not useful in a practical sense for inference. Kneip et al. (2008) also proposed two bootstrap procedures for inference about efficiency, and proved consistency of both methods. The first approach uses sub-sampling, where bootstrap samples of size $m < n$ are drawn (independently, with replacement) from the empirical distribution of the original n sample observations. Simulation results provided by Kneip et al. (2008) indicate that in finite-sample scenarios, coverages of confidence intervals for efficiency estimated by bootstrap sub-sampling are quite sensitive to the choice of the sub-sample size m; Kneip et al. (2008) did not provide a method for choosing m in applied work.

The second, full-sample bootstrap procedure described by Kneip et al. (2008) requires for consistency not only smoothing of the distribution of the observations as proposed in Simar and Wilson (1998, 2000a) but also smoothing of the initial DEA estimate of the frontier itself. This double-smoothing necessitates choosing values for two smoothing parameters. One of these can be optimized using existing methods from kernel density estimation, while a simple rule-of-thumb is provided for selecting the bandwidth used to smooth the frontier estimate. Simulation results presented in Kneip et al. indicate that the method works moderately well if smoothing parameters are chosen appropriately. However, the method requires solving n auxiliary linear programs (each with $(p + q + 1)$ constraints and $(n + 1)$ weights, where $(p + q)$ is the sum of input and output dimensions and n represents sample

size) for *each* of B bootstrap replications, leading to a formidable computational burden.

The naive bootstrap—based on re-sampling from the empirical distribution of the data—is attractive for its simplicity and low computational burden, but is inconsistent in situations where DEA efficiency estimators are used. As discussed by Simar and Wilson (1999a, 1999b), the inconsistency arises in part from the fact that when drawing from the empirical distribution, observations lying on the initial DEA frontier estimate are too-frequently selected. Kneip et al. (2009) developed a consistent bootstrap method that retains the simple features of the naive bootstrap to construct the part of a bootstrap sample lying "far" from the estimated frontier, while drawing from a smooth, uniform distribution to construct the part of the bootstrap sample lying "near" the estimated frontier. The distinction between "near" and "far" is controlled by a smoothing parameter, while a second smoothing parameter controls the degree of smoothing applied to the estimated frontier. Since no distributions are estimated, and no auxiliary linear programs are needed, the speed of the procedure is comparable to that of the naive bootstrap. However, the method of Kneip et al. (2009) requires complicated coding and is not appropriate for approximating the sampling distribution of a test statistic or of a function of efficiency estimators corresponding to different points in the sample space (which is needed for testing features of the model such as convexity of the production set, returns to scale, etc.).

The remainder of the paper unfolds as follows. In Section 2 we introduce a statistical model of a generic production process along with notation useful for describing features of the model that one might want to test. Section 3 describes the relevant estimators and their properties. In Section 4 we explain the sub-sampling method for testing hypotheses about the model and for estimating confidence intervals for the efficiencies of individual points. In Section 5 we present results from our Monte Carlo experiments and give practical advice for empirical researchers. Summary and conclusions are given in Section 6.

2 A Statistical Model of Production

Let $\boldsymbol{x} \in \mathbb{R}^p_+$ denote a vector of p input quantities, and let $\boldsymbol{y} \in \mathbb{R}^q_+$ denote a vector of q output quantities. Firms transform quantities of inputs into various quantities of outputs; a firm becomes more technically efficient if it increases at least some of its output levels without increasing its input levels (output orientation), or alternatively if it reduces its use of at least

some inputs without decreasing output levels (input orientation).

The set of feasible combinations of input and output vectors is given by the production set

$$\mathcal{P} = \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^p_+ \times \mathbb{R}^q_+ \mid \boldsymbol{x} \text{ can produce } \boldsymbol{y}\}. \tag{2.1}$$

The technical efficiency of a given point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ is determined by the distance from the point to the boundary, or efficient frontier,

$$\mathcal{P}^\partial = \left\{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P} \mid \left(\gamma \boldsymbol{x}, \gamma^{-1} \boldsymbol{y}\right) \notin \mathcal{P} \text{ for any } \gamma < 1\right\} \tag{2.2}$$

of the attainable set $\mathcal{P}$.

The boundary $\mathcal{P}^\partial$ of $\mathcal{P}$ constitutes the *technology*. Microeconomic theory of the firm suggests that in perfectly competitive markets, firms operating in the interior of $\mathcal{P}$ will be driven from the market, but makes no prediction of how long this might take; moreover, a firm that is inefficient today might become efficient tomorrow. The following assumptions on $\mathcal{P}$ are standard in microeconomics; e.g., see Shephard (1970) and Färe (1988).

Assumption 2.1. *$\mathcal{P}$ is compact and convex.*

Assumption 2.2. *$(\boldsymbol{x}, \boldsymbol{y}) \notin \mathcal{P}$ if $\boldsymbol{x} = 0$, $\boldsymbol{y} \gneq 0$; i.e., all production requires use of some inputs.*

Assumption 2.3. *for $\widetilde{\boldsymbol{x}} \geq \boldsymbol{x}$, $\widetilde{\boldsymbol{y}} \leq \boldsymbol{y}$, if $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ then $(\widetilde{\boldsymbol{x}}, \boldsymbol{y}) \in \mathcal{P}$ and $(\boldsymbol{x}, \widetilde{\boldsymbol{y}}) \in \mathcal{P}$, i.e., both inputs and outputs are strongly disposable.*

Here and throughout, inequalities involving vectors are defined on an element-by-element basis; e.g., for $\widetilde{\boldsymbol{x}}$, $\boldsymbol{x} \in \mathbb{R}^p_+$, $\widetilde{\boldsymbol{x}} \geq \boldsymbol{x}$ means that some number $\ell \in \{0, 1, \ldots, p\}$ of the corresponding elements of $\widetilde{\boldsymbol{x}}$ and $\boldsymbol{x}$ are equal, while $(p - \ell)$ of the elements of $\widetilde{\boldsymbol{x}}$ are greater than the corresponding elements of $\boldsymbol{x}$. Assumption 2.3 is equivalent to an assumption of monotonicity of the technology.

The Shephard (1970) input distance function

$$\theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}) \equiv \sup\left\{\theta > 0 \mid (\theta^{-1}\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}\right\} \tag{2.3}$$

measures technical efficiency in the input direction, i.e., in the direction parallel to the vector $\boldsymbol{x}$ and orthogonal to $\boldsymbol{y}$. This measure is "radial" in the sense that efficiency of a point

$(\boldsymbol{x}, \boldsymbol{y})$ is defined in terms of how much all input quantities can be contracted, by the same proportion, without altering output levels to arrive at the boundary $\mathcal{P}^{\partial}$. By construction, $\theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}) > 1$ for all $(\boldsymbol{x}, \boldsymbol{y})$ in the interior of $\mathcal{P}$, and $\theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}) = 1$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}^{\partial}$. For a given point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$, $(\theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P})^{-1}\boldsymbol{x}, \boldsymbol{y})$ is its projection onto $\mathcal{P}^{\partial}$ in the input direction.

Alternatively, the Shephard (1970) output distance function

$$\lambda(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}) \equiv \inf \left\{ \lambda > 0 \mid (\boldsymbol{x}, \lambda^{-1}\boldsymbol{y}) \in \mathcal{P} \right\} \tag{2.4}$$

measures technical efficiency in the output direction, i.e., in the direction orthogonal to $\boldsymbol{x}$ and parallel to $\boldsymbol{y}$; $\lambda(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P})$ gives the maximum, proportionate, feasible expansion of $\boldsymbol{y}$, holding input quantities $\boldsymbol{x}_0$ fixed. By construction, $\lambda(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}) < 1$ for $(\boldsymbol{x}, \boldsymbol{y})$ in the interior of $\mathcal{P}$, and $\lambda(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}) = 1$ for $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}^{\partial}$. For a given point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$, $(\boldsymbol{x}, \lambda(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P})^{-1}\boldsymbol{y})$ is its projection onto $\mathcal{P}^{\partial}$ in the output direction.

Since $\mathcal{P}$ (and hence $\mathcal{P}^{\partial}$) is unknown, it must be estimated from an observed sample $\mathcal{S}_n = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ of data on firms' input and output quantities. The next assumptions define a DGP; the framework here is similar to that in Simar (1996), Kneip et al. (1998), Simar and Wilson (1998, 2000a), and Kneip et al. (2008).

Assumption 2.4. *The n observations in $\mathcal{S}_n$ are identically, independently distributed (iid) random variables on the convex attainable set $\mathcal{P}$.*

Assumption 2.5. *(a) The $(\boldsymbol{x}, \boldsymbol{y})$ possess a joint density f with support $\mathcal{P}$; (b) f is continuous on $\mathcal{P}$; and (c) $f\left(\theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P})^{-1}\boldsymbol{x}, \boldsymbol{y}\right) > 0$ and $f\left(\boldsymbol{x}, \lambda(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P})^{-1}\boldsymbol{y}\right) > 0$ for all $(\boldsymbol{x}, \boldsymbol{y})$ in the interior of $\mathcal{P}$.*

Assumption 2.5(c) imposes a discontinuity in f at points in $\mathcal{P}^{\partial}$ ensuring a strictly positive, non-negligible probability of observing production units close to the production frontier. For points lying outside $\mathcal{P}$, $f \equiv 0$.

Assumption 2.6. *The functions $\theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P})$ and $\lambda(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P})$ are twice continuously differentiable for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$.*

Assumption 2.6 imposes some smoothness on the boundary $\mathcal{P}^{\partial}$. This assumption is slightly stronger, but simpler, than a corresponding assumption needed by Kneip et al. (1998) to establish consistency of the DEA estimators. We have adopted Assumption 2.6 from Kneip et al. (2008), where additional discussion is given.

In order to consider testing of hypotheses about the shape of $\mathcal{P}$ or $\mathcal{P}^\partial$, some additional notation is needed. Let the operator $\mathcal{F}(\cdot)$ denote the free-disposal hull of a set in $\mathbb{R}_+^{p+q}$ so that

$$\mathcal{F}(\mathcal{P}) = \bigcup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{P}} \{(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}}) \in \mathbb{R}_+^{p+q} \mid \widetilde{\boldsymbol{y}} \leq \boldsymbol{y},\ \widetilde{\boldsymbol{x}} \geq \boldsymbol{x}\}. \tag{2.5}$$

The assumption of disposability of inputs and outputs (Assumption 2.3) ensures $\mathcal{P} = \mathcal{F}(\mathcal{P})$. Now let $\mathcal{C}(\mathcal{P})$ denote the convex hull of $\mathcal{F}(\mathcal{P})$, and let $\mathcal{V}(\mathcal{P})$ denote the conical hull of $\mathcal{F}(\mathcal{P})$. In general, if the convexity assumption is dropped and only disposability of inputs and outputs is assumed, then

$$\mathcal{P} = \mathcal{F}(\mathcal{P}) \subseteq \mathcal{C}(\mathcal{P}) \subseteq \mathcal{V}(\mathcal{P}). \tag{2.6}$$

Under the additional assumption of convexity given by Assumption 2.1, we have

$$\mathcal{P} = \mathcal{F}(\mathcal{P}) = \mathcal{C}(\mathcal{P}) \subseteq \mathcal{V}(\mathcal{P}). \tag{2.7}$$

If, in addition, we assume that returns to scale are globally constant, then

$$\mathcal{P} = \mathcal{F}(\mathcal{P}) = \mathcal{C}(\mathcal{P}) = \mathcal{V}(\mathcal{P}). \tag{2.8}$$

Among (2.6)–(2.8), the latter is the most restrictive or constrained model. Alternatively, assuming $\mathcal{P}$ is convex with $\mathcal{P}^\delta$ exhibiting varying returns to scale,

$$\mathcal{P} = \mathcal{F}(\mathcal{P}) = \mathcal{C}(\mathcal{P}) \subset \mathcal{V}(\mathcal{P}). \tag{2.9}$$

If the empirical researcher accepts Assumption 2.3, implying $\mathcal{P} = \mathcal{F}(\mathcal{P})$, he may wish to test the assumption of convexity in Assumption 2.1 by testing the null hypothesis $H_0\colon \mathcal{P} = \mathcal{C}(\mathcal{P}) \subseteq \mathcal{V}(\mathcal{P})$ versus the alternative $H_1\colon \mathcal{P} = \mathcal{F}(\mathcal{P}) \subset \mathcal{C}(\mathcal{P})$. Alternatively, if the assumption of convexity is accepted, one might test $H_0'\colon \mathcal{P} = \mathcal{C}(\mathcal{P}) = \mathcal{V}(\mathcal{P})$ versus $H_1'\colon \mathcal{P} = \mathcal{C}(\mathcal{P}) \subset \mathcal{V}(\mathcal{P})$, which amounts to a test of globally constant returns to scale of the technology $\mathcal{P}^\partial$ versus variable returns to scale.

Other hypotheses may also be of interest. For example, one might wish to test whether a subset of inputs (or outputs) can be aggregated, whether an input or output is irrelevant, whether returns to scale are non-increasing, etc. Or, one might want to estimate confidence intervals for $\theta(\boldsymbol{x},\boldsymbol{y} \mid \mathcal{P})$ or $\lambda(\boldsymbol{x},\boldsymbol{y} \mid \mathcal{P})$. We show in the following sections how subsampling

can be used to test the null hypotheses of convexity versus non-convexity or constant returns to scale versus variable returns; it is easy to extend these ideas to test other hypotheses about the production process. Sub-sampling can also be used to consistently estimate confidence intervals for technical efficiency measures and perhaps other quantities of interest.

3 Non-parametric Efficiency Estimators

The distance functions in (2.3)–(2.4) are defined in terms of the *unknown, true* production set $\mathcal{P}$, and must be *estimated* from a set $\mathcal{S}_n = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$ of observed input/output combinations. Traditional non-parametric approaches used in analyses of efficiency and production typically assume $\Pr((\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{P}) = 1 \; \forall \; i = 1, \; \ldots, \; n$ and replace $\mathcal{P}$ in (2.3)–(2.4) with an estimator of the production set to obtain estimators of the Shephard input- and output-oriented distance functions. Several possibilities exist.

Deprins et al. (1984) proposed estimating $\mathcal{P}$ by the free-disposal hull (FDH) of the observations in $\mathcal{S}_n$, i.e.,

$$\widehat{\mathcal{P}}_{\text{FDH}}(\mathcal{S}_n) = \mathcal{F}(\mathcal{S}_n) = \bigcup_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{S}_n} \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}_+^{p+q} \mid \boldsymbol{y} \leq \boldsymbol{y}_i, \; \boldsymbol{x} \geq \boldsymbol{x}_i\}. \tag{3.1}$$

This estimator is consistent under Assumptions 2.1–2.6, but also remains consistent when the assumption of convexity is dropped. Alternatively, the convex hull of $\widehat{\mathcal{P}}_{\text{FDH}}$,

$$\widehat{\mathcal{P}}_{\text{VRS}}(\mathcal{S}_n) = \mathcal{C}(\mathcal{F}(\mathcal{S}_n)) = \Big\{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}_+^{p+q} \mid \boldsymbol{y} \leq \sum_{i=1}^n \omega_i \boldsymbol{y}_i, \; \boldsymbol{x} \geq \sum_{i=1}^n \omega_i \boldsymbol{x}_i, \; \sum_{i=1}^n \omega_i = 1, \; \omega_i \geq 0 \; \forall \; i = 1, \; \ldots, \; n\Big\}, \tag{3.2}$$

can be used to estimate $\mathcal{P}$. If we want to estimate the more restricted model with globally constant returns to scale ($\mathcal{P} = \mathcal{V}(\mathcal{P})$), then $\mathcal{P}$ can be estimated consistently by the conical hull of $\widehat{\mathcal{P}}_{\text{VRS}}(\mathcal{S}_n)$ (or, equivalently, the conical hull of $\widehat{\mathcal{P}}_{\text{FDH}}(\mathcal{S}_n)$), denoted $\widehat{\mathcal{P}}_{\text{CRS}}(\mathcal{S}_n)$ and obtained by dropping the constraint $\sum_{i=1}^n \omega_i = 1$ in (3.2).

As a practical matter, DEA estimates of input or output distance functions are obtained by solving the resulting familiar linear programs obtained after substituting $\widehat{\mathcal{P}}_{\text{VRS}}(\mathcal{S}_n)$ or $\widehat{\mathcal{P}}_{\text{CRS}}(\mathcal{S}_n)$ for $\mathcal{P}$ in (2.3) or (2.4). For example, when $\widehat{\mathcal{P}}_{\text{VRS}}(\mathcal{S}_n)$ is substituted for $\mathcal{P}$ in (2.3),

one obtains the estimator

$$\widehat{\theta}_{\mathrm{VRS}}(\boldsymbol{x},\boldsymbol{y}\mid\mathcal{S}_n)=\min_{\theta,\omega_1,\ldots,\omega_n}\Big\{\theta>0\mid \boldsymbol{y}\le\sum_{i=1}^{n}\omega_i\boldsymbol{y}_i,\ \theta\boldsymbol{x}\ge\sum_{i=1}^{n}\omega_i\boldsymbol{x}_i,\ \sum_{i=1}^{n}\omega_i=1, n\ \omega_i\ge 0\ \forall\ i=1,\ \ldots,\ n\Big\}. \tag{3.3}$$

Similarly, substituting $\widehat{\mathcal{P}}_{\mathrm{CRS}}(\mathcal{S}_n)$ for $\mathcal{P}$ in (2.3) leads to the estimator

$$\widehat{\theta}_{\mathrm{CRS}}(\boldsymbol{x},\boldsymbol{y}\mid\mathcal{S}_n)=\min_{\theta,\omega_1,\ldots,\omega_n}\Big\{\theta>0\mid \boldsymbol{y}\le\sum_{i=1}^{n}\omega_i\boldsymbol{y}_i,\ \theta\boldsymbol{x}\ge\sum_{i=1}^{n}\omega_i\boldsymbol{x}_i,\ \omega_i\ge 0\ \forall\ i=1,\ \ldots,\ n\Big\}, \tag{3.4}$$

which resembles $\widehat{\theta}_{\mathrm{VRS}}(\boldsymbol{x},\boldsymbol{y}\mid\mathcal{S}_n)$ defined in (3.3), except that the constraint $\sum_{i=1}^{n}\omega_i=1$ does not appear on the right-hand side of (3.4).

Although FDH efficiency estimators can be written in terms of integer programming problems, estimates based on (3.1) can be obtained using simple numerical calculations. In particular, in the input orientation, one can compute

$$\widehat{\theta}_{\mathrm{FDH}}(\boldsymbol{x},\boldsymbol{y}\mid\mathcal{S}_n)=\min_{\substack{i=1,\ldots,n\\ \mid \boldsymbol{y}_i\ge\boldsymbol{y}}}\left(\max_{j=1,\ldots,p}\left(\frac{\boldsymbol{x}^j}{\boldsymbol{x}_i^j}\right)\right), \tag{3.5}$$

where $\boldsymbol{x}^j$, $\boldsymbol{x}_i^j$ denote the jth elements of $\boldsymbol{x}$ (i.e., the input vector corresponding to the fixed point of interest) and $\boldsymbol{x}_i$ (i.e., the input vector corresponding to the ith observation in $\mathcal{S}_n$).

Asymptotic properties of estimators of the input and output distance functions in (2.3)–(2.4) based on $\widehat{\mathcal{P}}_{\mathrm{FDH}}(\mathcal{S}_n)$ and $\widehat{\mathcal{P}}_{\mathrm{VRS}}(\mathcal{S}_n)$, as well as the assumptions needed to establish consistency of the estimators, are summarized in Simar and Wilson (2000b). In particular, under Assumptions 2.1–2.6, $\widehat{\theta}_{\mathrm{VRS}}(\boldsymbol{x},\boldsymbol{y})$ is a consistent estimator of $\theta(\boldsymbol{x},\boldsymbol{y}\mid\mathcal{P})$, with convergence rate $n^{-2/(p+q+1)}$ (Kneip et al., 1998). If in addition $\mathcal{P}=\mathcal{V}(\mathcal{P})$, then $\widehat{\theta}_{\mathrm{CRS}}(\boldsymbol{x},\boldsymbol{y})$ is a consistent estimator of $\theta(\boldsymbol{x},\boldsymbol{y}\mid\mathcal{P})$, with convergence rate $n^{-2/(p+q)}$ (Park et al., 2009). Finally, if $\mathcal{P}$ is compact (but perhaps not convex), then under Assumptions 2.2–2.6, $\widehat{\theta}_{\mathrm{FDH}}(\boldsymbol{x},\boldsymbol{y})$ is a consistent estimator of $\theta(\boldsymbol{x},\boldsymbol{y}\mid\mathcal{P})$, but with convergence rate $n^{-1/(p+q)}$ (Park et al., 2000).

As noted in Section 1, Kneip et al. (2008) derived the limiting distribution for the DEA estimator $\widehat{\theta}_{\mathrm{VRS}}(\boldsymbol{x},\boldsymbol{y})$ and proved consistency of two bootstrap procedures for inference about $\theta(\boldsymbol{x},\boldsymbol{y}\mid\mathcal{P})$. One procedure requires smoothing not only the density of the observations in $\mathcal{S}_n$, but also the initial frontier estimate; consequently, the method is computationally

burdensome. Kneip et al. (2009) offer a simpler bootstrap based on naively resampling from the empirical distribution of the estimated efficiencies, and replacing any draws in a neighborhood of the frontier estimated with a draw from a uniform distribution over this neighborhood. This procedure avoids some of the computational difficulty of the earlier double-smooth bootstrap in Kneip et al., but nonetheless remains complicated and is not suitable for approximating the sampling distribution of test statistics involving efficiency measures at several data points.

The second bootstrap procedure suggested by Kneip et al. (2008) is based on subsampling, but no guidance was given for choosing the size of the subsamples. For purposes of testing hypotheses about the structure of $\mathcal{P}^{\partial}$ or other features of the model, there is to date no viable alternative to using subsampling techniques—the smoothing methods proposed by Kneip et al. (2008) and Kneip et al. (2009), due to their focus on a single point, cannot be adapted to more general testing situations. Recent papers by Politis et al. (2001) and Bickel and Sakov (2008) provide theoretical results and practical suggestions for choosing the size of subsamples when using a subsampling bootstrap for inference. In the next section, we provide additional results needed to adapt their results to the particular circumstances of the nonparametric production model presented in Section 2 in order to make inference about the shape of $\mathcal{P}$ or $\mathcal{P}^{\delta}$, or to make inference about the efficiency of a particular point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$.

4 The Subsampling Bootstrap

4.1 Confidence intervals for efficiency of a particular point

For situations where DEA estimators are used while maintaining the convexity assumption, Kneip et al. (2008) develop the asymptotic theory needed for using subsampling to estimate confidence intervals for the efficiency of a particular point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ and in addition give an algorithm with computational details. Jeong and Simar (2006) provide similar theory for the case of FDH estimators, where the convexity assumption may be relaxed. In the case of VRS estimators, the bootstrap principle is based on the following approximation: as $n, m \to \infty$ with $m/n \to 0$,

$$m^{2/(p+q+1)} \left(\frac{\widehat{\theta}_{\mathrm{VRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)}{\widehat{\theta}_{\mathrm{VRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_m^*)} - 1 \right) \overset{\text{approx.}}{\sim} n^{2/(p+q+1)} \left(\frac{\theta(\boldsymbol{x}, \boldsymbol{y})}{\widehat{\theta}_{\mathrm{VRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)} - 1 \right), \tag{4.1}$$

where $\mathcal{S}_m^*$ is a naive bootstrap sample of size m drawn from $\mathcal{S}_n$. Note that the resampling can be done either with or without replacement. Then a $(1-\alpha)$ confidence interval for $\theta(\boldsymbol{x}, \boldsymbol{y})$ is given by

$$\left[\widehat{\theta}_{\mathrm{VRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)\left(1 + n^{-2/(p+q+1)}\psi_{\alpha/2,m}\right),\ \widehat{\theta}_{\mathrm{VRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)\left(1 + n^{-2/(p+q+1)}\psi_{1-\alpha/2,m}\right)\right], \quad (4.2)$$

where $\psi_{\alpha,m}$ is the a-quantile of the bootstrap distribution of $m^{2/(p+q+1)}\left(\frac{\widehat{\theta}_{\mathrm{VRS}}(\boldsymbol{x},\boldsymbol{y}|\mathcal{S}_n)}{\widehat{\theta}_{\mathrm{VRS}}(\boldsymbol{x},\boldsymbol{y}|\mathcal{S}_m^*)} - 1\right)$.

Kneip et al. (2008) proved that the approximation in (4.1) is consistent for any choice $m = n^\gamma$ with $\gamma \in (0, 1)$; however, the quality of the approximation in finite samples depends crucially on γ. Below, in Section 5.4, we discuss results from extensive Monte Carlo experiments designed to determine whether ideas discussed by Bickel and Sakov (2008) and Politis et al. (2001) for choosing m (or equivalently, choosing γ) yield confidence interval estimates with reasonable coverage properties. Both Bickel and Sakov and Politis et al. proposed computing the object of interest (e.g., a confidence interval estimate or critical value of a test) for various values of m, and then choosing the value of m that minimizes some measure of volatility of the object of interest. In the case of confidence intervals for $\theta(\boldsymbol{x}, \boldsymbol{y})$, one might estimate confidence intervals of size α using various bootstrap sub-sample sizes $m_1 < m_2 < \ldots < m_J$, and then measure volatility corresponding to m_j by computing the standard deviations of the bounds of the estimated confidence intervals corresponding to $m_{j-k},\ \ldots, m_j,\ \ldots, m_{j+k}$ where k is a small integer (e.g., $k = 1$, 2, or 3) and $j = (k+1),\ \ldots,\ (J-k)$. The sub-sample size m would then be chosen as the m_j yielding the smallest measure of volatility; explicit details are given below in Section 5.

When the sub-sampling is done without replacement, the bootstrap distribution in (4.1) will become too concentrated as $m \to n$; if fact, if $m = n$, the bootstrap distribution collapses to a single probability mass. On the other hand, as $m \to 0$, the resulting confidence interval estimates will either under- or over-cover $\theta(\boldsymbol{x}, \boldsymbol{y})$ since too much information is lost. An optimal value of m will lie between these extremes; the idea is to choose a value of m that yields "stable" estimates for confidence intervals.

Politis et al. (2001) also discussed how these ideas can be used for hypothesis testing. In the remainder of this section, we expand their ideas to incorporate the particular features of the model and estimators presented above in Sections 2 and 3.

4.2 A Probabilistic Framework for Testing

In order to test hypotheses about the shape of the frontier $\mathcal{P}^\partial$ defined in (2.2), we must first define a probabilistic framework within which the model characteristic to be tested can be described. This allows us to define test statistics that discriminate between the conditions of null and alternative hypotheses. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be the probability space on which the random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are defined; by Assumption 2.5, $\mathcal{P}$ is the support of the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y})$. Denote the DGP by $P \in \mathbb{P}$. Let $\mathbb{P}_0$ denote the restricted DGPs where the null hypothesis is true, and let $\mathbb{P}_1$ denote the complement of $\mathbb{P}_0$, so that $\mathbb{P}_0 \cap \mathbb{P}_1 = \emptyset$ and $\mathbb{P} = \mathbb{P}_0 \cup \mathbb{P}_1$. Under the null, $P \in \mathbb{P}_0$.

Now consider a particular model $P \in \mathbb{P}$ with model characteristic $\tau(P)$ defined as

$$\tau(P) = E\left(h(g_0(\boldsymbol{X}, \boldsymbol{Y}), g(\boldsymbol{X}, \boldsymbol{Y}))\right) \tag{4.3}$$

where $h\colon \mathbb{R}^2 \to \mathbb{R}^1$ is a given smooth (differentiable) function of $g_0(\cdot)$, $g(\cdot)$, and where $g(\cdot)\colon \mathbb{R}_+^{p+q} \to \mathbb{R}$ and $g_0(\cdot)\colon \mathbb{R}_+^{p+q} \to \mathbb{R}$. We assume all these functions are Borel (measurable) functions, so that the expectation $\tau(P)$ is well-defined. We will also assume that the variance of $h(g_0(\boldsymbol{X}, \boldsymbol{Y}), g(\boldsymbol{X}, \boldsymbol{Y}))$, denoted by $\sigma^2(P)$, is finite.

The choice of $h(\cdot)$ depends on the hypothesis to be tested; in the cases we consider, $g_0(\cdot)$ will be some Shephard distance function measuring distance from $(\boldsymbol{X}, \boldsymbol{Y})$ to the frontier $\mathcal{P}^\partial$ under the null (i.e., when $P \in \mathbb{P}_0$), whereas $g(\cdot)$ will be a less-restrictive distance measure appropriate for $P \in \mathbb{P}$. As will become apparent below, in all of the testing situations we consider, it will be possible to define the function $h(\cdot)$ such that for all $P \in \mathbb{P}$, $h(g_0(\boldsymbol{X}, \boldsymbol{Y}), g(\boldsymbol{X}, \boldsymbol{Y})) \overset{a.s.}{\geq} 0$, while if $P \in \mathbb{P}_0$, then $h(g_0(\boldsymbol{X}, \boldsymbol{Y}), g(\boldsymbol{X}, \boldsymbol{Y})) \overset{a.s.}{=} 0$.

To be explicit, for purposes of testing convexity, i.e., for testing $H_0\colon \mathcal{P} = \mathcal{C}(\mathcal{P})$ versus $H_1\colon \mathcal{P} = \mathcal{F}(\mathcal{P}) \subset \mathcal{C}(\mathcal{P})$, we might consider

$$\tau(P) = E\left(\frac{g_0(\boldsymbol{X}, \boldsymbol{Y})}{g(\boldsymbol{X}, \boldsymbol{Y})} - 1\right), \tag{4.4}$$

where $g(\boldsymbol{X}, \boldsymbol{Y}) := \theta(\boldsymbol{X}, \boldsymbol{Y} \mid \mathcal{P})$ and $g_0(\boldsymbol{X}, \boldsymbol{Y}) := \theta(\boldsymbol{X}, \boldsymbol{Y} \mid \mathcal{C}(\mathcal{P}))$ with $\theta(\boldsymbol{X}, \boldsymbol{Y} \mid \cdot)$ defined by (2.3). Alternatively, for testing globally constant returns to scale versus non-constant, variable returns to scale, i.e., for testing $H_0'\colon \mathcal{P} = \mathcal{V}(\mathcal{P})$ versus $H_1'\colon \mathcal{P} = \mathcal{C}(\mathcal{P}) \subset \mathcal{V}(\mathcal{P})$, we might use the expression for $\tau(P)$ in (4.4) while defining $g(\boldsymbol{X}, \boldsymbol{Y}) := \theta(\boldsymbol{X}, \boldsymbol{Y} \mid \mathcal{C}(\mathcal{P}))$ and $g_0(\boldsymbol{X}, \boldsymbol{Y}) := \theta(\boldsymbol{X}, \boldsymbol{Y} \mid \mathcal{V}(\mathcal{P}))$. In all situations we define $\tau(P)$ so that $\tau(P) \geq 0 \;\forall\; P \in \mathbb{P}$,

but $\tau(P) = 0$ if $P \in \mathbb{P}_0$ and $\tau(P) > 0$ if $P \in \mathbb{P}_1$. Hence testing the null amounts to testing $H_0 \colon \tau(P) = 0$ versus $H_1 \colon \tau(P) > 0$.

A consistent estimator of $\tau(P)$ is easy to derive. Let the sample empirical mean replace of the expectation in (4.3) and replace the unknown functions $g(\cdot)$ and $g_0(\cdot)$ with their appropriate estimators (e.g., depending on the framework, the DEA or FDH estimators defined in (3.3), (3.4), or (3.5)). Given a random sample $\mathcal{S}_n = \{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i=1}^n$, we obtain

$$\tau_n(\mathcal{S}_n) = n^{-1} \sum_{i=1}^{n} \left(h(\widehat{g}_0(\boldsymbol{X}_i, \boldsymbol{Y}_i), \widehat{g}(\boldsymbol{X}_i, \boldsymbol{Y}_i)\right). \tag{4.5}$$

Note that $\widehat{g}_0(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ is abbreviated notation for $\widehat{g}_0(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}_n)$, and similarly for $\widehat{g}(\boldsymbol{X}_i, \boldsymbol{Y}_i)$; the estimators are evaluated at the point $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ using a reference sample $\mathcal{S}_n$. Below, it will be useful to use this explicit notation, in particular when using the bootstrap.

To simplify notation, let $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})$ denote a generic observation. Define

$$T(\boldsymbol{Z}) = h(g_0(\boldsymbol{Z}), g(\boldsymbol{Z})) \tag{4.6}$$

and

$$\widehat{T}(\boldsymbol{Z} \mid \mathcal{S}_n) = h(\widehat{g}_0(\boldsymbol{Z}), \widehat{g}(\boldsymbol{Z})), \tag{4.7}$$

where $\mathcal{S}_n = \{\boldsymbol{Z}_i\}_{i=1}^n$. Then

$$\tau(P) = E(T(\boldsymbol{Z})) \tag{4.8}$$

and

$$\tau_n(\mathcal{S}_n) = n^{-1} \sum_{i=1}^{n} \widehat{T}(\boldsymbol{Z}_i \mid \mathcal{S}_n). \tag{4.9}$$

4.3 Asymptotic Behavior of $\widehat{T}(\boldsymbol{Z} \mid \mathcal{S}_n)$

The framework introduced above ensures that for all P, $T(\boldsymbol{Z}) \overset{a.s.}{\geq} 0$ and if $P \in \mathbb{P}_0$, then $T(\boldsymbol{Z}) \overset{a.s.}{=} 0$. In addition, under regularity conditions (i.e., Assumptions 2.1–2.6), for all fixed $\boldsymbol{z} \in \mathcal{P}$,

$$n^{\kappa} \left(\widehat{T}(\boldsymbol{z} \mid \mathcal{S}_n) - T(\boldsymbol{z})\right) \xrightarrow{\mathcal{L}} G(\cdot \mid \boldsymbol{z}), \tag{4.10}$$

where $G(. \mid \boldsymbol{z})$ is a nondegenerate distribution whose characteristics depends on $\boldsymbol{z}$. The value of κ is known and depends on the problem at hand. This rate is governed by the smallest rate of convergence of the DEA or FDH estimators used to define $T(\boldsymbol{Z})$; for example,

$\kappa = 2/(p+q+1)$ when testing constant returns to scale, and $\kappa = 1/(p+q)$ when testing convexity. This implies that

$$\lim_{n\to\infty} \Pr\left[n^{\kappa}\left(\widehat{T}(\boldsymbol{z} \mid \mathcal{S}_n) - T(\boldsymbol{z})\right) \leq a\right] = G(a \mid \boldsymbol{z}). \tag{4.11}$$

Since $\widehat{T}(\boldsymbol{Z} \mid \mathcal{S}_n)$ and $T(\boldsymbol{Z})$ are well-defined random variables on $(\Omega, \mathcal{A})$, (4.11) can be considered as a conditional statement, with conditioning on $\boldsymbol{Z} = \boldsymbol{z}$; hence

$$\lim_{n\to\infty} \Pr\left[n^{\kappa}\left(\widehat{T}(\boldsymbol{Z} \mid \mathcal{S}_n) - T(\boldsymbol{Z})\right) \leq a \mid \boldsymbol{Z} = \boldsymbol{z}\right] = G(a \mid \boldsymbol{z}). \tag{4.12}$$

By marginalizing on $\boldsymbol{Z}$, we have

$$\lim_{n\to\infty} \Pr\left[n^{\kappa}\left(\widehat{T}(\boldsymbol{Z} \mid \mathcal{S}_n) - T(\boldsymbol{Z})\right) \leq a\right] = \int_{\mathcal{P}} G(a \mid \boldsymbol{z}) f_Z(\boldsymbol{z}) d\boldsymbol{z} = Q(a). \tag{4.13}$$

Note that the density introduced in Assumption 2.5 has been re-written here as $f_Z(\cdot)$. Since $G(\cdot \mid \boldsymbol{z})$ and $f_Z(\cdot)$ are nondegenerate, $Q(\cdot)$ is a nondegenerate distribution. It follows that

$$n^{\kappa}\big(\widehat{T}(\boldsymbol{Z} \mid \mathcal{S}_n) - T(\boldsymbol{Z})\big) \xrightarrow{\mathcal{L}} Q(\cdot). \tag{4.14}$$

Now let μ_Q and σ_Q^2 denote the finite mean and strictly positive variance of $Q(\cdot)$. Since $\widehat{T}(\boldsymbol{Z} \mid \mathcal{S}_n)) = T(\boldsymbol{Z}) + n^{-\kappa} W(\boldsymbol{Z})$, $W(\boldsymbol{Z})$ must have limiting distribution $Q(\cdot)$ as $n \to \infty$. Combining this result with (4.8), we have

$$E(\widehat{T}(\boldsymbol{Z} \mid \mathcal{S}_n)) = \tau(P) + \mu_Q/n^{\kappa} \tag{4.15}$$

and

$$\mathrm{VAR}(\widehat{T}(\boldsymbol{Z} \mid \mathcal{S}_n)) = \sigma^2(P) + \sigma(P)O(n^{-\kappa}) + \sigma_Q^2/n^{2\kappa}, \tag{4.16}$$

where the second term in (4.16) accounts for the covariance between $T(\boldsymbol{Z})$ and $n^{-\kappa}W(\boldsymbol{Z})$ (which is bounded by the product of their standard deviations). Note that when the null is true, i.e., $P \in \mathbb{P}_0$, $\tau(P) = \sigma^2(P) = 0$ and the formulae (4.15) and (4.16) simplify accordingly.

4.4 Asymptotic Behavior of $\tau_n(\mathcal{S}_n)$

From the results in (4.15) and (4.16), it is easy to derive the asymptotic mean and the variance of $\tau_n(\mathcal{S}_n)$. For the latter, we have to consider the asymptotic covariance between $\widehat{T}(\boldsymbol{Z}_j; \mathcal{S}_n)$ and $\widehat{T}(\boldsymbol{Z}_k; \mathcal{S}_n)$, for $j \neq k$. The local nature of the asymptotic distribution of DEA

efficiency estimators is given by Theorem 1(i) in Kneip et al. (2008) and Theorem 4.1 in Kneip et al. (2009). The value of the DEA estimator at a point is essentially determined by those observations which fall into a small neighborhood of the projection of this point onto the frontier. Using the reasoning in the proof of Theorem 4.1 in Kneip et al. (2009), consider a point $\boldsymbol{z} \in \mathcal{P}$ where the DEA score (i.e., efficiency estimate) is evaluated and let $C_z(h)$ be a neighborhood of the frontier point $\boldsymbol{z}^{\partial}$ determined by the projection of the point $\boldsymbol{z}$ on the true frontier; h is a bandwidth that controls the size of this neighborhood. If $h^2 = O\big(n^{-\kappa}\big)$, $C_z(h)$ will contain the DEA estimate of the frontier at $\boldsymbol{z}$ with probability 1. Since $h \to 0$ as $n \to \infty$, the probability of an observation $\boldsymbol{Z}_i$ falling in $C_z(h)$ is approximated by

$$\pi_n = \Pr\big(\boldsymbol{Z} \in C_z(h)\big) \approx f_Z(\boldsymbol{z}^{\partial})(2h)^{p+q-1}h^2 = O\big(n^{-1}\big). \tag{4.17}$$

For large n, the distribution of the number of points $\boldsymbol{Z}_i$ falling in $C_z(h)$ follows approximately a Poisson distribution with parameter $n\pi_n = O(1)$. As shown in Kneip et al. (2009), when $n \to \infty$, only points falling in this neighborhood influence the distribution of the DEA estimator at the point $\boldsymbol{z}$. The number of such points is $O(1)$. Consequently, the covariances between the DEA estimator at $\boldsymbol{Z}_j$ and the $(n-1)$ DEA estimators at the other points $\boldsymbol{Z}_k$ is nonzero for at most $O(1)$ of these $(n-1)$ estimators. Moreover, each of the nonzero covariances is bounded by the product of the standard deviations derived from (4.16); therefore, the n covariance terms sum to $nO(1)[\sigma^2(P) + \sigma(P)O(n^{-\kappa}) + \sigma_Q^2/n^{2\kappa}]$. Combining these results, we obtain

$$E(\tau_n(\mathcal{S}_n)) = \tau(P) + \frac{\mu_Q}{n^{\kappa}} \tag{4.18}$$

and

$$\mathrm{VAR}(\tau_n(\mathcal{S}_n)) = \frac{1}{n^2}\left\{n \times [\sigma_Q^2/n^{2\kappa} + O(n^{-\kappa})\sigma(P) + \sigma^2(P)]\right\} = O(n^{-1}). \tag{4.19}$$

Hence for all $P \in \mathbb{P}$, $\tau_n(\mathcal{S}_n) \xrightarrow{P} \tau(P)$ and $\tau_n(\mathcal{S}_n)$ is a consistent estimator of $\tau(P)$. From (4.18) we also see that μ_Q/n^{κ} acts as a bias term that disappears asymptotically. Under the null, since $\tau(P) = \sigma(P) = 0$, we obtain for all $P \in \mathbb{P}_0$,

$$E(\tau_n(\mathcal{S}_n)) = \mu_Q/n^{\kappa} \tag{4.20}$$

and

$$\mathrm{VAR}(\tau_n(\mathcal{S}_n)) = \sigma_Q^2/n^{1+2\kappa} = O(n^{-(1+2\kappa)}), \tag{4.21}$$

indicating that the rate of convergence of $\tau_n(\mathcal{S}_n)$ is faster when the null is true as opposed to when it is false.

Consistency of the subsampling approximation in (4.1) follows from Theorem 3.1 of Politis et al. (2001), which requires that under the null, $n^\kappa\sqrt{n}\tau_n(\mathcal{S}_n)$ converge to a nondegenerate distribution. It is sufficient to assume an additional technical regularity condition on $f_Z(\cdot)$, in order to obtain a normal limiting distribution.

Proposition 4.1. *If the joint density $f(\boldsymbol{x},\boldsymbol{y})$ of $(\boldsymbol{X},\boldsymbol{Y})$ is such that the moments of $Q(\cdot)$ exist up to the fourth order, then under the null hypothesis $H_0\colon P\in\mathbb{P}_0$,*

$$n^\kappa\sqrt{n}\,\left(\tau_n(\mathcal{S}_n)-\mu_Q/n^\kappa\right)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\sigma_Q^2). \tag{4.22}$$

This proposition follows directly when considering the triangular array (see e.g. Serfling, 1980, Section 1.9.3, p.31)

$$\begin{array}{llll}
\widehat{T}(\boldsymbol{Z}_1;\mathcal{S}_1); & & & \\
\widehat{T}(\boldsymbol{Z}_1;\mathcal{S}_2) & \widehat{T}(\boldsymbol{Z}_2;\mathcal{S}_2); & & \\
\quad\vdots & & & \\
\widehat{T}(\boldsymbol{Z}_1;\mathcal{S}_n) & \widehat{T}(\boldsymbol{Z}_2;\mathcal{S}_n) & \ldots & \widehat{T}(\boldsymbol{Z}_n;\mathcal{S}_n); \\
\quad\vdots & & &
\end{array}$$

The mean and the variance of the sums were derived above. Proposition 4.1 follows from the Lyapunov condition with $\nu=3$ (see the corollary in Section 1.9.3 of Serfling), i.e.,

$$\frac{nE\left|\widehat{T}(\boldsymbol{Z}_j;\mathcal{S}_n)-\mu_Q/n^\kappa\right|^3}{\left(n\sigma_Q^2/n^{2\kappa}\right)^{3/2}}=o(1), \tag{4.23}$$

which holds provided moments of $Q(\cdot)$ exist up to fourth order.[1]

[1]The argument used here is standard; in addition, it is straightforward to verify that the result also holds in the unrestricted case where $P\in\mathbb{P}$. The only difference will be in the expressions for the mean and the variance of $\tau_n(\mathcal{S}_n)$ that appear in (4.18) and (4.19). Of course, to satisfy the Lyapunov condition an additional technical regularity condition on the random variable $T(\boldsymbol{Z})$ is needed. In particular, $T(\boldsymbol{Z})$ must have finite moments up to order 4 (when H_0 is true, $P\in\mathbb{P}_0$ and $T(\boldsymbol{Z})$ is a degenerate random variable equal to zero). Hence, for $P\in\mathbb{P}$,

$$\sqrt{n}\left(\tau_n(\mathcal{S}_n)-(\tau(P)+\mu_Q/n^\kappa)\right)\xrightarrow{\mathcal{L}}\mathcal{N}\left(0,\sigma_Q^2/n^{2\alpha}+O(n^{-\kappa})\sigma(P)+\sigma^2(P)\right).$$

Note that the rate of convergence is slower when the null is false.

4.5 Testing by Subsampling

Since $\tau_n(\mathcal{S}_n)$ is a consistent estimator of $\tau(P)$, we will reject the null if $\tau_n(\mathcal{S}_n)$ is "too large." For $m < n$, let $\tau_m(\mathcal{S}_m^*)$ denote the test statistic evaluated using the pseudo data set $\mathcal{S}_m^*$ obtained by drawing m observations from $\mathcal{S}_n$ without replacement. Due to the results derived above, for a test of level α we reject the null hypothesis H_0 if and only if $n^\kappa\sqrt{n}\tau_n(\mathcal{S}_n) > q_{m,n}(1-\alpha)$, where $q_{m,n}(1-\alpha)$ is the $(1-\alpha)$ quantile of the bootstrap distribution of $m^\kappa\sqrt{m}\,\tau_m(\mathcal{S}_m^*)$ approximated by

$$\widehat{G}_{m,n}(a) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left(m^\kappa\sqrt{m}\,\tau_m(\mathcal{S}_m^{*,b}) \leq a\right), \tag{4.24}$$

where $\mathbb{1}()$ denotes the indicator function, $B \leq \binom{n}{m}$ is the number of bootstrap replications, and $\left\{\tau_m(\mathcal{S}_m^{*,b})\right\}_{b=1}^{B}$ is the set of bootstrap estimates, each computed from different random subsamples of size m. Theorem 3.1 of Politis et al. (2001) ensures that this testing procedure is asymptotically of size α and is consistent (i.e., the probability of rejecting the null when it is false converges to 1), provided $m, n \to \infty$ with $m/n \to 0$.

Note that in the procedure proposed here, we neglect the bias term μ_Q/n^κ appearing in (4.22). This bias term could be estimated while performing the bootstrap computations, but results from our Monte-Carlo experiments suggest that this introduces substantial noise; results (both in term of achieved level and of power) are better when the bias term is simply ignored.

The procedure for selecting the subsample size m is in practice very similar to the idea explained above in Section 4.1 in connection with estimation of confidence intervals. In testing situations, the "optimal" m can be selected by minimizing the volatility of the quantiles $q_{m,n}(1-\alpha)$, viewed as a function of m.

5 Monte Carlo Evidence

5.1 Experimental Framework

We perform three sets of Monte Carlo experiments to examine the performance of the subsampling bootstrap is situations faced by applied researchers under real-world conditions. In the first two sets of experiments, we consider size and power properties of tests of convexity

of the production set $\mathcal{P}$ and returns to scale of the technology $\mathcal{P}^{\partial}$. In the third set of experiments, we examine the coverages of confidence intervals for technical efficiency of a fixed point.

In each of the three sets of experiments, we consider two sample sizes, $n \in \{100,\ 1000\}$, and DGPs with either two dimensions (with $p = q = 1$) or four dimensions (with $p = 3$, $q = 1$).[2] In each experiment, we perform 1,024 Monte Carlo trials. On each Monte Carlo trial, we perform 2,000 bootstrap replications for each of 49 sub-sample sizes $m \in \mathbb{M}_n = \left\{\frac{n}{50},\ \frac{2n}{50},\ \frac{3n}{50},\ \ldots,\ \frac{49n}{50}\right\}$. For each sub-sample size m, we use $k \in \{1,\ 2,\ 3\}$ to select the "optimal" sub-sample size as described above in Section 4. We conduct experiments using resampling without replacement as well as resampling with replacement.

In the first two experiments where we test a null hypothesis H_0 against an alternative hypothesis H_1, on a particular Monte Carlo trial, we generate n observations and then compute the relevant test statistic. Next, for each sub-sample size $m_j \in \mathbb{M}_n$, we perform 2,000 bootstrap replications and compute corresponding critical values $\{c_1,\ c_2,\ \ldots,\ c_{49}\}$ for (one-sided) tests of size $\alpha \in \{.1,\ .05,\ .01\}$. Then, for a given test size α, we minimize critical value volatility along the lines of Politis et al. (2001) using the following steps:

[i] For $j \in \{J_{\text{lo}},\ \ldots,\ J_{\text{hi}}\}$ and for a small integer value k, compute volatility indices given by the standard deviations $\widehat{s}_j$ of the critical values $\{c_{j-k},\ \ldots,\ c_{j+k}\}$.

[ii] Choose $\widehat{j}$ corresponding to the smallest volatility index, and take $c_{\widehat{j}}$ as the final critical value, with corresponding sub-sample size $\widehat{m} = m_{\widehat{j}}$.

The procedure for estimating confidence intervals is similar. On a particular Monte Carlo trial, we generate data from the relevant DGP, and compute an estimate $\widehat{\theta}$ corresponding to the point of interest $(\boldsymbol{x}_0, \boldsymbol{y}_0)$, where $\widehat{\theta}$ is either $\widehat{\theta}_{VRS}(\boldsymbol{x}_0, \boldsymbol{y}_0 \mid \mathcal{S}_n)$ defined in (3.3) or $\widehat{\theta}_{FDH}(\boldsymbol{x}_0, \boldsymbol{y}_0 \mid \mathcal{S}_n)$ defined in (3.5). Then for each sub-sample size $m_j \in \mathbb{M}_n$, we perform 2,000 bootstrap replications yielding bootstrap values $\{\widehat{\theta}^*_{mb}\}_{b=1}^{2000}$ corresponding to the initial estimate $\widehat{\theta}$. For confidence intervals of size α, we next compute the $\psi_{\alpha/2,m_j}$ and $\psi_{1-\alpha/2,m_j}$ percentiles of the empirical distribution of the bootstrap values $m_j^{\kappa}\left(\frac{\widehat{\theta}}{\widehat{\theta}^*_{mb}} - 1\right)$, where κ equals either $2/(p+q+1)$ if $\widehat{\theta}$ is the VRS-DEA estimator defined in (3.3), or $1/(p+q)$ if $\widehat{\theta}$ is the

[2] Of course, situations involving more than one output can be easily handled using our methods; here, we use only one output to simplify the process of simulating data.

FDH estimator defined in (3.5). The confidence interval estimate of nominal size α is then $\left[\widehat{\theta}\left(1+n^{-\kappa}\psi_{\alpha/2,m_j}\right),\ \widehat{\theta}\left(1+n^{-\kappa}\psi_{1-\alpha/2,m_j}\right)\right]$.[3]

After performing the bootstrap for each subsample size $m_j \in \mathbb{M}_n$, we have 49 confidence interval estimates $\{(\widehat{c}_{\mathrm{lo},j}(\alpha), \widehat{c}_{\mathrm{hi},j}(\alpha))\}$ for a particular size α. We then choose among the various confidence interval estimates by minimizing volatility as in Algorithm 6.1 appearing in Politis et al. (2001); in particular, we use the following steps:

[i] For each $j \in \{J_{\mathrm{lo}},\ \ldots,\ J_{\mathrm{hi}}\}$ and for a small integer value k, compute the volatility index $\widehat{s}_j$ given by the sum of the standard deviations of $\{\widehat{c}_{\mathrm{lo},j-k}(\alpha),\ \ldots,\ \widehat{c}_{\mathrm{lo},j+k}(\alpha)\}$ and $\{\widehat{c}_{\mathrm{hi},j-k}(\alpha),\ \ldots,\ \widehat{c}_{\mathrm{hi},j+k}(\alpha)\}$.

[ii] Choose $\widehat{j}$ corresponding to the smallest volatility index, and take $\left[\widehat{c}_{\mathrm{lo},\widehat{j}}, \widehat{c}_{\mathrm{hi},\widehat{j}}\right]$ as the final confidence interval estimate, with corresponding sub-sample size $\widehat{m} = m_{\widehat{j}}$.

The remainder of this section describes results from specific Monte Carlo experiments designed to gage the performance of the sub-sampling bootstrap for testing convexity and returns to scale, as well as for estimating confidence intervals for technical efficiency of a given point.

5.2 Testing Convexity

Suppose that a sample $\mathcal{S}_n$ of n input-output vectors is observed. For purposes of testing convexity, i.e., testing $H_0\colon \mathcal{P} = \mathcal{F}(\mathcal{P}) = \mathcal{C}(\mathcal{P})$ versus $H_1\colon \mathcal{P} = \mathcal{F}(\mathcal{P}) \subset \mathcal{C}(\mathcal{P})$, we consider two different test statistics, namely

$$\widehat{\tau}_1(\mathcal{S}_n) = n^{-1}\sum_{i=1}^{n}\left(\frac{\widehat{\theta}_{\mathrm{VRS}}(\boldsymbol{x}_i, \boldsymbol{y}_i \mid \mathcal{S}_n)}{\widehat{\theta}_{\mathrm{FDH}}(\boldsymbol{x}_i, \boldsymbol{y}_i \mid \mathcal{S}_n)} - 1\right) \geq 0 \tag{5.1}$$

and

$$\widehat{\tau}_2(\mathcal{S}_n) = n^{-1}\sum_{i=1}^{n}\boldsymbol{D}_{2i}{}'\boldsymbol{D}_{2i} \geq 0, \tag{5.2}$$

where $\boldsymbol{D}_{2i} = \left(\boldsymbol{x}_i\widehat{\theta}_{\mathrm{VRS}}(\boldsymbol{x}_i, \boldsymbol{y}_i \mid \mathcal{S}_n)^{-1} - \boldsymbol{x}_i\widehat{\theta}_{\mathrm{FDH}}(\boldsymbol{x}_i, \boldsymbol{y}_i \mid \mathcal{S}_n)^{-1}\right)$ is a $(p\times 1)$ vector. The statistic $\widehat{\tau}_1(\mathcal{S}_n)$ exploits the multiplicative structure of the non-parametric efficiency estimators. The statistic $\widehat{\tau}_2(\mathcal{S}_n)$ gives an estimate of the mean integrated square difference between the VRS

[3]The interval given by (4.2) is a special case of this, where $\widehat{\theta}$ is the VRS-DEA estimator.

and FDH frontier estimates; a similar statistic was proposed by Härdle and Mammen (1993) to test parametric regression model fits against non-parametric alternatives. In terms of the discussion in Section 4, the statistics defined in (5.1) and (5.2) are estimators of the population quantities τ_1 and τ_2, respectively, obtained by replacing the distance function estimators in (5.1)–(5.2) with the corresponding *true* distance function values. Under the null hypothesis H_0, it is clear that $\tau_1 = \tau_2 = 0$, whereas under the alternative hypothesis H_1, $\tau_1 > 0$ and $\tau_2 > 0$. Hence under H_0, both $\widehat{\tau}_1(\mathcal{S}_n)$ and $\widehat{\tau}_2(\mathcal{S}_n)$ are expected to be "small," whereas under H_1, $\widehat{\tau}_1(\mathcal{S}_n)$ and $\widehat{\tau}_2(\mathcal{S}_n)$ are expected to be "large," and the question is whether the statistics defined in (5.1)–(5.2) are large enough to reject the null hypothesis H_0. The sub-sampling bootstrap described in Section 4 can be used to determine the necessary critical values.

We simulate DGPs for the two-dimensional case (i.e., $p = q = 1$) by drawing (efficient) input values $\widetilde{x}$ from the uniform distribution on the interval $[0,\ 1]$, and then setting

$$y = \widetilde{x}^{\delta} \tag{5.3}$$

for some $\delta > 0$ to obtain the corresponding efficient output levels. Next, we set $x = \widetilde{x}e^{u}$, where $u \sim \text{Exp}(1/3)$ (i.e., u is exponentially distributed with parameter equal to 3, so that $E(u) = 1/3$) to obtain simulated observations (x, y). For the four-dimensional case (i.e., $p = 3$, $q = 1$), we first draw a triplet of efficient input quantities $\widetilde{x}_1,\ \widetilde{x}_2,\ \widetilde{x}_3$ from the uniform distribution on $[0, 1]$, and then set

$$y = \left[x_1^{0.33} x_2^{0.33} x_3^{0.34}\right]^{\delta} \tag{5.4}$$

where again $\delta > 0$. Next, we draw $u \sim \text{Exp}(1/3)$ and set $x_j = \widetilde{x}_j e^{u}$ for each $j \in \{1,\ 2,\ 3\}$ to obtain a simulated observations (x_1, x_2, x_3, y).

In our experiments, we simulate the DGPs described above using values $\delta \in \{0.5,\ 1.0,\ 1.1,\ 1.2,\ 1.3,\ 1.4,\ 1.6,\ 1.8,\ 2.4,\ 3.0\}$. When $\delta = 0.5$, the production set is strictly convex, while it is weakly convex when $\delta = 1.0$. For $\delta > 1$, the production set is not convex, with increasing departures from the null hypothesis of convexity as δ increases above one. Experiments were conducted using resampling without replacement as well as resampling with replacement. To conserve space, we report here only results from experiments using resampling without replacement.[4]

[4]Results from the experiments using resampling with replacement are available in a separate appendix,

In each experiment, we estimate rejection rates corresponding to each of 49 values $m_j \in \mathbb{M}_n$ by counting, for each m_j, the number of trials where the null hypothesis is rejected and then dividing these counts by the number of Monte Carlo trials (1,024). Overall, tests based on the statistic $\widehat{\tau}_{2n}(\mathcal{S}_n)$ out-performed those based on $\widehat{\tau}_{1n}(\mathcal{S}_n)$ in terms of both achieved size and power; consequently we focus the discussion that follows on the tests using $\widehat{\tau}_{2n}(\mathcal{S}_n)$.

Figure 1 shows the results of this analysis using the test statistic $\widehat{\tau}_2(\mathcal{S}_n)$ defined in (5.2) for the two sample sizes and the two dimensionalities that we considered in our experiments. Each panel of Figure 1 shows 10 curves corresponding to the 10 different values of δ plotted as alternating solid and dashed lines. Each curve represents rejection rate as a function of bootstrap sub-sample size. In each panel, starting from the southwest corner and moving toward the northeast corner, we encounter the first solid curve which corresponds to $\delta = 0.5$ and the first dashed curve, corresponding to $\delta = 1$. We next encounter alternating solid and dashed curves corresponding to $\delta = 1.1$, $\delta = 1.2$, and so on. Also in each panel, a horizontal line is plotted at height 0.05 on the vertical axis which measures rejection rates.

The two panels in the top half of Figure 1 show rejection rates for the two-dimensional case where $p = q = 1$. Comparing the two lowest curves (where H_0 is true) in these panels confirms that as n increases from 100 to 1,000, the range of values of m that yield rejection rates "close" to five percent becomes wider; i.e., the cures depicting rejection rates for various values of m become flatter and closer to the horizontal line at 0.05 when the sample size is increased. The two panels also illustrate that the test has good power for a wide range of values of m, and that power increases over all but the larges values of m as the sample size increases.

The two panels in the bottom half of Figure 1 show rejection rates for the four-dimensional case where $p = 3$ and $q = 1$. The story here is similar, but there is a cost of increasing dimensionality. With the same sample size (either $n = 100$ or $n = 1000$), the range of values of m that yield rejection rates near five percent when H_0 is true is narrower in this case than in the two-dimensional case. Nonetheless, the two panels in the bottom half of Figure 1 confirm again that as n increases, the curves corresponding to $\delta = 0.5$ and $\delta = 1$ become flatter and closer to the horizontal line drawn at 0.05 on the vertical axis.[5]

available on-line at `http://www.stat.ucl.ac.be/ISpub/ISdp.html/`.

[5]Figure A.1 in the separate appendix mentioned in footnote 4 shows plots of rejection rates versus sub-sample sizes analogous to those in Figure 1, except that the rejection rates depicted in Figure A.1 are those

Overall, the results shown in Figures 1 confirm the theoretical results in Section 4. However, the results shown in the Figures give *average* rejection rates for *fixed* values of sub-sample sizes m. The applied researcher, by contrast, must choose a single value of m using the procedure described in Section 4. In order to assess expected rejection rates when m is chosen by minimizing volatility of critical values, we employed the method described in Section 5.1 using $J_{\text{lo}} = 15$, $J_{\text{hi}} = 45$, and $k \in \{1,\ 2,\ 3\}$ on each of 1,024 Monte Carlo trials in each experiment. For each experiment using resampling without replacement, we report in Tables 1–2 the proportion of Monte Carlo trials where H_0 was rejected using the test statistic $\widehat{\tau}_2(\mathcal{S}_n)$ defined in (5.2) with critical values chosen by minimizing volatility; Table 1 gives results for the two-dimensional case (with $p = q = 1$), while Table 2 gives results for the four-dimensional case (with $(p = 3,\ q = 1)$. On a given Monte Carlo trial, either $k = 1$, 2, or 3 was used to optimize the choice of sub-sample size m at .90, .95, and .99 significance levels. The optimization was done independently for each significance level and each value of k to produce the nine columns of results in Tables 1–2.[6]

The results reported in Tables 1–2 have clear implications for implementing tests of convexity based on sub-sampling. First, setting $k = 1$ typically results in greater test power than $k = 2$ or 3. Second, power increases rapidly as the sample size increases. Even in the four-dimensional case, when $k = 1$ and resampling without replacement is used, the probability of rejecting the null at .95 significance rises from 0.01 when $\delta = 1.0$ to 0.36 when $\delta = 1.1$ with $n = 1,000$ as shown in Table 2. Third, regardless of the value of k, the tests are conservative; i.e., when $\delta = 0.5$ or 1.0 (and H_0 is true), the average rejection rates are less than the nominal size or one minus the significance level. Nonetheless, power typically increases rapidly as δ is increased above one; i.e., power increases rapidly with departures from the null.[7] Finally, comparing results in Tables 1–2 with similar results from experiments using resampling with replacement (reported in Tables A.1–A.2 in the separate appendix

obtained using resampling *with* replacement. Comparing the results shown in the two figures, it is apparent that for given dimensionality $(p + q)$ and sample size n, the optimal sub-sample size m is smaller when resampling is done with replacement as opposed to without replacement. In addition, holding dimensionality and sample size constant, resampling with replacement typically results in less test power than resampling without replacement for given subsample sizes and departures from the null.

[6]Results for tests of convexity based on the statistic $\widehat{\tau}_1(\mathcal{S}_n)$ defined in (5.1) are available in the separate appendix mentioned in footnote 4.

[7]Of course, the null hypothesis in our test is a composite hypothesis. We have considered only two values of δ where the null is true while the size of the test is equal to the supremum of rejection rates for each value of $\delta \in (0, 1]$.

mentioned earlier), it is apparent that for given dimensionality and given k, resampling without replacement yields greater test power than resampling with replacement when m is optimized for each Monte Carlo trial.

Delving further into the results of our experiments, Figure 2 shows, for $p = q = 1$ and $n = 1,000$, results from six individual Monte Carlo trials chosen at random.[8] Figure 2 contains six panels; those in the first column show results from individual trials in the experiment where $\delta = 1.0$ (where $\mathcal{P}$ is weakly convex), while those in the second column give results from trials in the experiment where $\delta = 1.1$ (where $\mathcal{P}$ is not convex). In each panel, estimated 95-percent critical values are plotted (using small crosses) as a function of the various sub-sample sizes $m_j \in \mathbb{M}_n$. The solid horizontal line in each panel intersects the vertical axis at the value of the test statistic $\widehat{\tau}_2(\mathcal{S}_n)$, while the vertical dotted line represents the value $\widehat{m}$ (and hence the corresponding critical value) chosen by minimizing volatility using $k = 1$ as described in Section 5.1.

In the three panels in the left column of Figure 2, where the null is true, most of the estimated critical values lie above the horizontal line showing the values of the test statistic; in these trials, the null is not rejected. Meanwhile, in the right-hand column, most of the estimated critical values lie below the horizontal line showing the values of the test statistic; in each of these trials, the null is rejected. In the Monte Carlo trial represented in the lower right-hand panel, the test statistic is 2.360, while the critical value chosen by minimizing volatility is equal to 2.357; hence the null is rejected.

In an applied setting, the researcher could examine plots of estimated critical values versus sub-sample sizes as we have done in Figure 2. With dimensionality larger than we have considered here, or with smaller sample sizes, the results may be less clear-cut than those shown in Figure 2. Nonetheless, visual examination of the results is likely to give useful information in addition to information obtained using the mechanism described in Section 5.1 to minimize volatility.

We also considered tests of convexity based on statistics similar to those defined in (5.1) and (5.2), but where the linearly interpolated FDH (LFDH) estimator proposed by Jeong and Simar (2006) replaces the FDH estimator used to define $\widehat{\tau}_{1n}(\mathcal{S}_n)$ and $\widehat{\tau}_{2n}(\mathcal{S}_n)$. The per-

[8] Trials represented in Figure 2 were chosen by generating uniform random integers between 1 and 1,024 (inclusive).

formance of these modified tests were similar to those of the original statistics; consequently, there seems to be no reason to incur the extra computational burden involved when the LFDH estimator is used.[9]

5.3 Testing Returns to Scale

To examine rejection rates of tests of returns to scale, we modified the DGPs described by (5.3) and (5.4) to allow for variable returns to scale under departures from the null hypothesis of constant returns to scale. In the case of one input and one output ($p = q = 1$), we draw (efficient) input values $\widetilde{x}$ from the uniform distribution on the interval $[1-\delta,\ 2-\delta]$ and then set

$$y = (\widetilde{x} - (1-\delta))^{\delta} \tag{5.5}$$

to obtain the corresponding efficient output levels. Next, we set $x = \widetilde{x}e^{u}$, where $u \sim \text{Exp}(1/3)$ to obtain a simulated observation (x, y). For the case of three inputs and one output, (i.e., $p = 3,\ q = 1$), we first draw a triplet of efficient output quantities $\widetilde{x}_1,\ \widetilde{x}_2,\ \widetilde{x}_3$ from the uniform distribution on $[1-\delta,\ 2-\delta]$, and then set

$$\widetilde{y} = \left[(x_1 - (1-\delta))^{0.33}\,(x_2 - (1-\delta))^{0.33}\,(x_3 - (1-\delta))^{0.34}\right]^{\delta} \tag{5.6}$$

to obtain the corresponding output quantity. Next, we draw $u \sim \text{Exp}(1/3)$ and set $x_j = \widetilde{x}_j e^{u}$ for each $j \in \{1,\ 2,\ 3\}$ to obtain a simulated observation (x_1, x_2, x_3, y). In both the two- and four-dimensional cases, we consider values of $\delta \in \{1.0,\ 0.95,\ 0.90,\ \ldots,\ 0.60,\ 0.50\}$. When $\delta = 1$, the technologies in (5.5) and (5.6) exhibit constant returns to scale; as δ decreases from unity, the technologies are characterized by (increasingly) variable returns to scale and greater departures from the null hypothesis of constant returns to scale.

We examined two statistics for testing the null hypothesis of constant returns to scale versus the alternative hypothesis of variable returns to scale (i.e., testing $H_0' :\ \mathcal{P} = \mathcal{V}(\mathcal{P})$ versus $H_1' :\ \mathcal{P} = \mathcal{C}(\mathcal{P}) \subset \mathcal{V}(\mathcal{P})$), namely

$$\widehat{\tau}_3(\mathcal{S}_n) = n^{-1} \sum_{i=1}^{n} \left(\frac{\widehat{\theta}_{\text{CRS}}(\boldsymbol{x}_i, \boldsymbol{y}_i \mid \mathcal{S}_n)}{\widehat{\theta}_{\text{VRS}}(\boldsymbol{x}_i, \boldsymbol{y}_i \mid \mathcal{S}_n)} - 1 \right) \geq 0 \tag{5.7}$$

[9] Results for convexity tests based on the LFDH estimator are available in the separate appendix mentioned in footnote 4.

and

$$\widehat{\tau}_4(\mathcal{S}_n) = n^{-1} \sum_{i=1}^{n} \boldsymbol{D}_{4i}' \boldsymbol{D}_{4i} \geq 0, \tag{5.8}$$

where $\boldsymbol{D}_{4i} = \left(\boldsymbol{x}_i \widehat{\theta}_{\text{CRS}}(\boldsymbol{x}_i, \boldsymbol{y}_i \mid \mathcal{S}_n)^{-1} - \boldsymbol{x}_i \widehat{\theta}_{\text{VRS}}(\boldsymbol{x}_i, \boldsymbol{y}_i \mid \mathcal{S}_n)^{-1} \right)$ is a $(p \times 1)$ vector. These statistics are similar to those defined in (5.1) and (5.2) for testing convexity (versus non-convexity) of $\mathcal{P}$. In addition, the statistics defined in (5.7)–(5.8) estimate the corresponding population quantities τ_3 and τ_4 obtained by replacing the distance function estimators in (5.7)–(5.8) with the corresponding *true* values. Under the null, $\tau_3 = \tau_4 = 0$, whereas under the alternative, $\tau_3 > 0$ and $\tau_4 > 0$.

As in the experiments involving the convexity tests, we conducted experiments using resampling without replacement as well as resampling with replacement. In all of our experiments analyzing tests of returns to scale, the statistic $\widehat{\tau}_3(\mathcal{S}_n)$ dominated the statistic $\widehat{\tau}_4(\mathcal{S}_n)$ in terms of achieved size and power. This result contrasts with our experience with the tests of convexity described in Section 5.2, where the statistic $\widehat{\tau}_2(\mathcal{S}_n)$ based on mean integrated difference dominated the statistic $\widehat{\tau}_1(\mathcal{S}_n)$, which is analogous to $\widehat{\tau}_3(\mathcal{S}_n)$; both $\widehat{\tau}_1(\mathcal{S}_n)$ and $\widehat{\tau}_3(\mathcal{S}_n)$ are based on ratios of distance function estimators. In order to save space, we report only results for tests using $\widehat{\tau}_3(\mathcal{S}_n)$; results for tests based on $\widehat{\tau}_4(\mathcal{S}_n)$ are available in the separate appendix mentioned in Section 5.2.

Tables 3 shows results for tests of returns to scale using $\widehat{\tau}_3(\mathcal{S}_n)$ and resampling without replacement in the two-dimensional case, while Table 4 shows similar results for the four-dimensional case.[10] The results from our experiments reveal some clear patterns. First, as was the case in Section 5.2, resampling without replacement produces tests with better size and power properties than resampling with replacement, holding sample size, k, and dimensionality constant. Second, setting $k = 1$ again dominates $k = 2$ or 3 in terms of size and power. Third, with $k = 1$ and resampling with replacement, the test has good power. Moreover, power increases rapidly with sample size as well as with departures from the null.

5.4 Confidence Interval Estimation

In order to examine the performance of the sub-sampling bootstrap for inference regarding technical efficiency of a given point or firm, we simulate DGPs using the technologies defined

[10]Similar results from experiments using resampling with replacement appear in Tables A.7–A.8 of the separate appendix.

by (5.3) for the two-dimensional case and by (5.4) for the four-dimensional case, setting $\delta = 0.8$ and drawing observations as described in Section 5.2. In our experiments, we consider the coverage of estimated confidence intervals for the technical efficiency of a single fixed point $(\boldsymbol{x}_0, \boldsymbol{y}_0)$. For $p = q = 1$, we set $\boldsymbol{y}_0 = 0.5745$ and $\boldsymbol{x}_0 = \theta_0 \left(0.5745^{1/\delta}\right)$, where $\theta_0 = 2$ is the "true" value of the Shephard input distance function defined in (2.3). For the four-dimensional case with $p = 3$, $q = 1$, the fixed point of interest is given by $\boldsymbol{y}_0 = 0.4902$ and $\boldsymbol{x}_0 = \theta_0 \left[0.4902^{1/\delta} \quad 0.4902^{1/\delta} \quad 0.4902^{1/\delta}\right]$; again, $\theta_0 = 2$ is the "true" value of the Shephard input distance function defined in (2.3).

We consider resampling both with and without replacement. In addition, we consider balanced as well as unbalanced sampling. Unbalanced sampling amounts to drawing m observations from $\mathcal{S}_n$, without regard to the position of the fixed point of interest. Balanced sampling, by contrast, involves dividing the observations in $\mathcal{S}_n$ into the set $\mathcal{S}_{1n}$ of n_1 observations where $\boldsymbol{y}_i \not\geq \boldsymbol{y}_0$ and the set $\mathcal{S}_{2n}$ of $n_2 = n - n_1$ observations where $\boldsymbol{y}_i \geq \boldsymbol{y}_0$. Define the operator $\text{Nint}(\cdot)$ as returning the whole number nearest its argument, with fractional portions equal to 0.5 rounded to the nearest whole number that is larger in magnitude than the argument of the function. Then for a sub-sample of size m, $m_2 = \text{Nint}\left(\frac{m\, n_2}{n}\right)$ observations are drawn from $\mathcal{S}_{2n}$, and $m_1 = m - m_2$ observations are drawn from $\mathcal{S}_{1n}$ (in both cases, either with or without replacement). Balanced (sub)-sampling avoids situations where, on a given bootstrap replication, the bootstrap efficiency estimate is infeasible. This will occur whenever the sub-sample contains only observations where $\boldsymbol{y}_i < \boldsymbol{y}_0 \; \forall \; i = 1, \; \ldots, \; m$ (or, in the output orientation, where $\boldsymbol{x}_i > \boldsymbol{x}_0 \; \forall \; i = 1, \; \ldots, \; m$). When using unbalanced resampling, in bootstrap replications where the bootstrap efficiency estimator is infeasible, we set the estimate equal to one.[11]

Table 5 gives estimated coverages of confidence intervals obtained using the input-oriented DEA efficiency estimator $\widehat{\theta}_{\text{VRS}}(\boldsymbol{x}_0, \boldsymbol{y}_0 \mid \mathcal{S}_n)$ defined in (3.3). Comparing rows 1–4 with rows 5–8, and rows 9–12 with rows 13–16, it is apparent that resampling *with* replacement yields coverages closer to nominal levels than resampling *without* replacement. This is in contrast to the results for testing convexity and returns to scale. The choice of balanced versus

[11] In our experiments with unbalanced resampling, setting the bootstrap efficiency estimator equal to one when the estimate is infeasible amounts to recognizing that the particular bootstrap replication has no useful information for inference, and avoids imposing conditioning in the bootstrap world that is not present in the real world. This does not alter the asymptotic properties of our bootstrap. A similar device was used by Jeong and Simar (2006).

unbalanced resampling seems to make little difference in the coverages that are achieved. In addition, the choice of value for k used to construct the volatility indices seems less critical than in the tests of convexity and returns to scale examined previously in Sections 5.2 and 5.3. Here, setting $k = 1$ seems to give slightly better results than $k = 2$ or 3, but the differences are small and insignificant when resampling is done with replacement.

Overall, the coverages achieved using resampling with replacement and $k = 1$ in the volatility minimization are typically farther from the nominal coverages than obtained using the bootstrap proposed by Kneip et al. (2009, Table 2). For example, with $\alpha = 0.05$ and two dimensions, the Kneip et al. (2009) bootstrap yields coverages of 0.929 and 0.941 with $n = 100$ and 1,000 (respectively), whereas in Table 5 the best coverages with $\alpha = 0.05$ and two dimensions are 0.902 and 0.883 with $n = 100$ and 1,000 (respectively). Note also that the results obtained with the sub-sampling bootstrap in Table 5 show little or no improvement as sample size increases from $n = 100$ to $n = 1000$, while the results reported in Kneip et al. (2009) show clear improvement as sample size increases. Clearly, there is a price to pay for throwing away data when inferences are made by subsampling.

Table 6 gives results similar to those displayed in Table 5, but the results in Table 6 show achieved coverages of confidence intervals estimated using the FDH efficiency estimator defined in (3.5). The pattern of results obtained with the FDH estimator are similar to those obtained with the DEA estimator. In particular, resampling with replacement gives better coverages than resampling without replacement, while using balanced or unbalanced sampling makes little difference. Also, as was the case with the DEA estimator, using $k = 1$ to minimize volatility typically gives better coverages than $k = 2$ or 3, but the differences are neither large nor significant. Overall, the coverages obtained with the FDH estimator are worse than those obtained with the DEA estimator. This is perhaps not surprising, given the slower convergence rate of the FDH estimator.

6 Conclusions

Sub-sampling is an attractive option for inference in situations where DEA efficiency estimators are used. The substantial computational burden of the double-smooth procedure proposed by Kneip et al. (2008) is avoided, and the method is much simpler than the computationally-efficient method of Kneip et al. (2009). However, lunch is not free. For

purposes of estimating confidence intervals, our simulation results discussed above indicate that the achieved coverages of confidence intervals estimated by sub-sampling are farther from the corresponding nominal coverages than the coverages of intervals estimated using the method of Kneip et al. (2009). This is not surprising—the sub-sampling gives up data, and hence information, in order to avoid the inconsistency problems discussed by Simar and Wilson (1999a, 1999b). The method proposed by Kneip et al. (2009) also trades information to avoid inconsistency, but only in a small neighborhood near the frontier; since less information is sacrificed, coverages are better with the Kneip et al. (2009) method.

For purposes of testing hypotheses about the nature of the production set where the test statistic involves a function of DEA and perhaps FDH or other estimators, there seems to be no viable alternative to sub-sampling. The double-smooth bootstrap proposed by Kneip et al. (2008) as well as the computationally-efficiency method proposed by Kneip et al. (2009) are specific to a single point. These methods give the sampling distribution of a DEA efficiency estimator for a specific point, but cannot give the sampling distribution of a function of DEA estimators corresponding to different points. Hence sub-sampling is required if the goal is to test hypotheses about the structure of the production set.

Our simulation results indicate that when the sub-sample size m is chosen using the methods we employed in our Monte Carlo experiments, tests of convexity and returns to scale yield reasonable size properties and good power. To the extend that the realized sizes of our tests differs from nominal sizes, evidence from our experiments indicate that the tests are conservative; i.e., when testing a null hypothesis at nominal size α, the probability of rejecting the null may be less than α. At the same time, the probability of rejection increases rapidly with even small departures from the null, and hence the tests have good power properties.

Although choosing the sub-sample size m requires performing bootstraps for an array of sub-sample sizes, the computational burden remains much smaller than methods that involve smoothing which require cross-validation to choose bandwidths. Moreover, we have discussed in Section 5.2 how one might use graphical methods to choose the sub-sample size and to determine whether the null hypothesis should be rejected. These methods would be easy for the applied researcher to implement using the *FEAR* software library developed by Wilson (2008) or perhaps other software, and therefore should be useful.

References

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

—— (1996), "Hypothesis tests using data envelopment analysis," *Journal of Productivity Analysis,* 7, 139–159.

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

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

—— (1979), "Measuring the efficiency of decision making units," *European Journal of Operational Research,* 3, 339.

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

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

Färe, R. (1988), *Fundamentals of Production Theory,* Berlin: Springer-Verlag.

Färe, R., Grosskopf, S., and 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, 253–281.

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

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

Härdle, W., and Mammen, E. (1993), "Comparing nonparametric versus parametric regression fits," *Annals of Statistics,* 21, 1926–1947.

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

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

Kittelsen, S. A. C. (1999), "Monte Carlo simulations of DEA efficiency measures and hypothesis tests," Unpublished working paper, memorandum no. 09/99, Department of Economics, University of Oslo, Norway.

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

Kneip, A., Simar, L., and Wilson, P. W. (2008), "Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models," *Econometric Theory,* 24, 1663–1697.

——— (2009), "A computationally efficient, consistent bootstrap for inference with non-parametric dea estimators," Discussion paper #0903, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Korostelev, A., Simar, L., and Tsybakov, A. B. (1995), "On estimation of monotone and convex boundaries," *Publications de l'Institut de Statistique de l'Université de Paris XXXIX,* 1, 3–18.

Park, B. U., Jeong, S.-O., and Simar, L. (2009), "Asymptotic distribution of conical-hull estimators of directional edges," *Annals of Statistics.* forthcoming.

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

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

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics,* New York: John Wiley & Sons, Inc.

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

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

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

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

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

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

——— (2000b), "Statistical inference in nonparametric frontier models: The state of the art," *Journal of Productivity Analysis,* 13, 49–78.

——— (2001a), "Nonparametric tests of returns to scale," *European Journal of Operational Research,* 139, 115–132.

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

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

Table 1: Rejection Rates of Convexity Test using $\widehat{\tau}_2(\mathcal{S}_n)$ ($p = q = 1$, resampling without replacement)

		$k=1$			$k=2$			$k=3$		
			$(1-\alpha)$			$(1-\alpha)$			$(1-\alpha)$	
n	δ	.90	.95	.99	.90	.95	.99	.90	.95	.99
100	0.5	0.042	0.013	0.000	0.026	0.008	0.001	0.023	0.004	0.000
	1.0	0.032	0.009	0.000	0.019	0.008	0.000	0.018	0.005	0.000
	1.1	0.137	0.044	0.007	0.047	0.012	0.001	0.033	0.008	0.000
	1.2	0.263	0.118	0.021	0.088	0.032	0.006	0.046	0.015	0.001
	1.3	0.425	0.229	0.080	0.188	0.049	0.021	0.109	0.019	0.003
	1.4	0.497	0.263	0.147	0.279	0.062	0.033	0.202	0.014	0.002
	1.6	0.688	0.383	0.253	0.495	0.136	0.049	0.435	0.054	0.006
	1.8	0.763	0.477	0.302	0.632	0.197	0.041	0.580	0.114	0.005
	2.4	0.914	0.665	0.397	0.857	0.419	0.080	0.838	0.319	0.010
	3.0	0.926	0.714	0.444	0.890	0.534	0.117	0.880	0.447	0.013
1000	0.5	0.035	0.009	0.001	0.021	0.004	0.000	0.020	0.006	0.000
	1.0	0.019	0.002	0.000	0.009	0.001	0.000	0.004	0.000	0.000
	1.1	0.925	0.818	0.568	0.911	0.727	0.321	0.900	0.678	0.182
	1.2	0.995	0.974	0.938	0.996	0.965	0.899	0.991	0.960	0.873
	1.3	0.999	0.993	0.979	0.998	0.991	0.966	0.997	0.990	0.964
	1.4	0.998	0.993	0.978	0.998	0.985	0.971	0.999	0.984	0.960
	1.6	0.999	0.995	0.987	0.999	0.990	0.981	0.999	0.990	0.977
	1.8	0.998	0.997	0.992	0.998	0.997	0.987	0.998	0.996	0.985
	2.4	1.000	0.999	0.998	1.000	0.998	0.995	1.000	0.998	0.992
	3.0	0.998	0.995	0.998	0.999	0.994	0.988	0.998	0.993	0.987

Table 2: Rejection Rates of Convexity Test using $\widehat{\tau}_2(\mathcal{S}_n)$ ($p = 3,\ q = 1$, resampling without replacement)

		$k=1$			$k=2$			$k=3$		
		$(1-\alpha)$			$(1-\alpha)$			$(1-\alpha)$		
n	δ	.90	.95	.99	.90	.95	.99	.90	.95	.99
100	0.5	0.061	0.028	0.006	0.021	0.002	0.003	0.014	0.001	0.000
	1.0	0.038	0.023	0.005	0.007	0.003	0.000	0.003	0.002	0.000
	1.1	0.113	0.063	0.012	0.021	0.009	0.003	0.010	0.004	0.000
	1.2	0.159	0.095	0.032	0.035	0.012	0.006	0.020	0.002	0.000
	1.3	0.242	0.122	0.054	0.079	0.021	0.007	0.057	0.009	0.001
	1.4	0.312	0.171	0.107	0.150	0.032	0.019	0.114	0.016	0.004
	1.6	0.463	0.246	0.158	0.284	0.056	0.013	0.241	0.017	0.001
	1.8	0.581	0.281	0.199	0.412	0.107	0.035	0.361	0.064	0.007
	2.4	0.776	0.461	0.290	0.690	0.250	0.046	0.666	0.192	0.009
	3.0	0.812	0.560	0.290	0.737	0.341	0.061	0.723	0.285	0.009
1000	0.5	0.032	0.014	0.000	0.010	0.002	0.000	0.002	0.001	0.000
	1.0	0.026	0.010	0.001	0.007	0.000	0.000	0.000	0.000	0.000
	1.1	0.619	0.360	0.186	0.486	0.136	0.039	0.448	0.093	0.013
	1.2	0.974	0.887	0.660	0.970	0.853	0.367	0.968	0.837	0.231
	1.3	0.992	0.970	0.875	0.989	0.955	0.782	0.989	0.950	0.731
	1.4	0.995	0.979	0.947	0.997	0.972	0.917	0.996	0.971	0.896
	1.6	0.999	0.995	0.984	0.998	0.989	0.971	0.998	0.988	0.959
	1.8	1.000	0.993	0.983	0.998	0.989	0.976	0.999	0.988	0.965
	2.4	0.999	0.995	0.990	0.999	0.994	0.979	0.998	0.992	0.974
	3.0	1.000	0.999	0.994	1.000	0.995	0.983	1.000	0.994	0.982

Table 3: Rejection Rates of Returns to Scale Test using $\widehat{\tau}_3(\mathcal{S}_n)$ ($p = q = 1$, resampling without replacement)

		$k=1$			$k=2$			$k=3$		
			$(1-\alpha)$			$(1-\alpha)$			$(1-\alpha)$	
n	δ	.90	.95	.99	.90	.95	.99	.90	.95	.99
100	1.00	0.147	0.071	0.020	0.115	0.064	0.007	0.103	0.047	0.008
	0.95	0.940	0.840	0.698	0.917	0.789	0.561	0.911	0.778	0.531
	0.90	0.971	0.889	0.778	0.957	0.871	0.696	0.957	0.859	0.658
	0.85	0.966	0.906	0.822	0.965	0.882	0.739	0.964	0.877	0.715
	0.80	0.983	0.929	0.848	0.977	0.904	0.758	0.977	0.900	0.715
	0.75	0.984	0.927	0.835	0.982	0.906	0.727	0.981	0.898	0.698
	0.70	0.983	0.933	0.847	0.979	0.898	0.739	0.980	0.893	0.692
	0.65	0.977	0.914	0.794	0.975	0.884	0.652	0.975	0.874	0.610
	0.60	0.987	0.927	0.793	0.983	0.894	0.642	0.980	0.880	0.580
	0.50	0.970	0.883	0.728	0.961	0.831	0.527	0.960	0.815	0.476
1000	1.00	0.096	0.043	0.009	0.071	0.035	0.003	0.062	0.027	0.006
	0.95	0.996	0.983	0.979	0.997	0.979	0.963	0.997	0.977	0.959
	0.90	0.995	0.987	0.982	0.995	0.987	0.977	0.997	0.983	0.971
	0.85	0.995	0.993	0.994	0.997	0.991	0.987	0.997	0.991	0.984
	0.80	0.999	0.995	0.995	1.000	0.993	0.991	0.999	0.991	0.988
	0.75	1.000	1.000	0.998	1.000	0.999	0.995	1.000	0.999	0.995
	0.70	1.000	0.999	0.999	1.000	0.999	0.997	1.000	0.999	0.997
	0.65	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	0.999
	0.60	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
	0.50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4: Rejection Rates of Returns to Scale Test using $\widehat{\tau}_3(\mathcal{S}_n)$ ($p = 3,\ q = 1$, resampling without replacement)

		$k=1$			$k=2$			$k=3$		
		$(1-\alpha)$			$(1-\alpha)$			$(1-\alpha)$		
n	δ	.90	.95	.99	.90	.95	.99	.90	.95	.99
100	1.00	0.072	0.044	0.009	0.024	0.009	0.001	0.018	0.004	0.000
	0.95	0.446	0.254	0.118	0.379	0.133	0.018	0.368	0.117	0.007
	0.90	0.520	0.288	0.147	0.462	0.181	0.024	0.456	0.165	0.007
	0.85	0.639	0.370	0.161	0.578	0.243	0.038	0.575	0.229	0.017
	0.80	0.660	0.382	0.197	0.608	0.288	0.036	0.606	0.273	0.017
	0.75	0.681	0.396	0.207	0.634	0.275	0.038	0.626	0.264	0.012
	0.70	0.720	0.390	0.225	0.661	0.278	0.040	0.659	0.260	0.019
	0.65	0.691	0.370	0.198	0.637	0.241	0.031	0.634	0.229	0.010
	0.60	0.698	0.388	0.192	0.637	0.244	0.030	0.633	0.230	0.009
	0.50	0.651	0.349	0.228	0.569	0.207	0.039	0.562	0.189	0.012
1000	1.00	0.056	0.031	0.010	0.008	0.006	0.004	0.006	0.003	0.000
	0.95	0.930	0.900	0.837	0.923	0.877	0.782	0.923	0.875	0.768
	0.90	0.957	0.930	0.890	0.950	0.918	0.857	0.946	0.915	0.838
	0.85	0.969	0.949	0.931	0.965	0.939	0.899	0.965	0.938	0.888
	0.80	0.983	0.972	0.952	0.980	0.967	0.938	0.980	0.966	0.934
	0.75	0.988	0.981	0.975	0.987	0.980	0.960	0.987	0.980	0.950
	0.70	0.991	0.989	0.979	0.991	0.986	0.972	0.991	0.985	0.966
	0.65	0.998	0.995	0.992	0.998	0.993	0.984	0.998	0.993	0.984
	0.60	0.999	0.997	0.994	0.998	0.998	0.993	0.998	0.997	0.992
	0.50	0.999	1.000	0.998	0.999	1.000	0.998	0.999	1.000	0.995

Table 5: Coverages of Estimated Confidence Intervals using DEA Efficiency Estimator

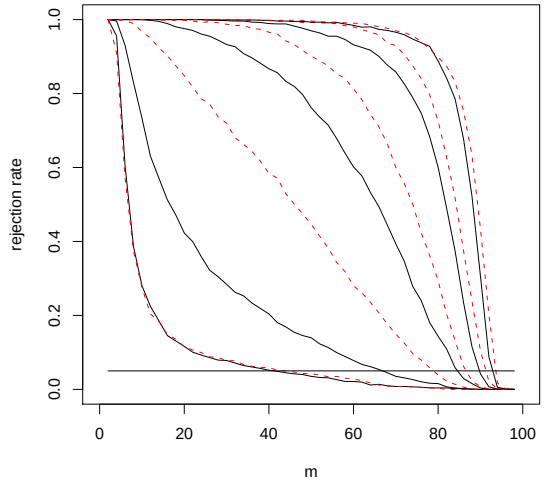
					$k=1$ $(1-\alpha)$			$k=2$ $(1-\alpha)$			$k=3$ $(1-\alpha)$		
p	q	n	balanced?	with repl.?	.90	.95	.99	.90	.95	.99	.90	.95	.99
1	1	100	N	N	0.773	0.846	0.936	0.762	0.828	0.926	0.753	0.827	0.922
1	1	1000	N	N	0.749	0.823	0.917	0.735	0.818	0.920	0.741	0.818	0.910
1	1	100	Y	N	0.769	0.835	0.927	0.757	0.827	0.922	0.750	0.812	0.920
1	1	1000	Y	N	0.751	0.822	0.915	0.737	0.824	0.906	0.746	0.815	0.910
1	1	100	N	Y	0.844	0.902	0.977	0.847	0.896	0.970	0.838	0.898	0.962
1	1	1000	N	Y	0.817	0.883	0.955	0.811	0.878	0.951	0.812	0.875	0.954
1	1	100	Y	Y	0.840	0.900	0.969	0.835	0.896	0.964	0.834	0.885	0.962
1	1	1000	Y	Y	0.819	0.881	0.948	0.813	0.874	0.947	0.810	0.875	0.948
3	1	100	N	N	0.832	0.901	0.979	0.827	0.894	0.969	0.811	0.886	0.967
3	1	1000	N	N	0.832	0.906	0.979	0.816	0.891	0.971	0.803	0.884	0.962
3	1	100	Y	N	0.821	0.905	0.973	0.814	0.891	0.969	0.803	0.877	0.961
3	1	1000	Y	N	0.836	0.906	0.975	0.814	0.887	0.969	0.807	0.880	0.960
3	1	100	N	Y	0.932	0.969	0.997	0.926	0.964	0.997	0.920	0.961	0.995
3	1	1000	N	Y	0.919	0.965	0.990	0.912	0.959	0.987	0.906	0.950	0.985
3	1	100	Y	Y	0.930	0.966	0.995	0.919	0.959	0.995	0.912	0.958	0.995
3	1	1000	Y	Y	0.917	0.955	0.991	0.914	0.957	0.988	0.910	0.947	0.983

Table 6: Coverages of Estimated Confidence Intervals using FDH Efficiency Estimator

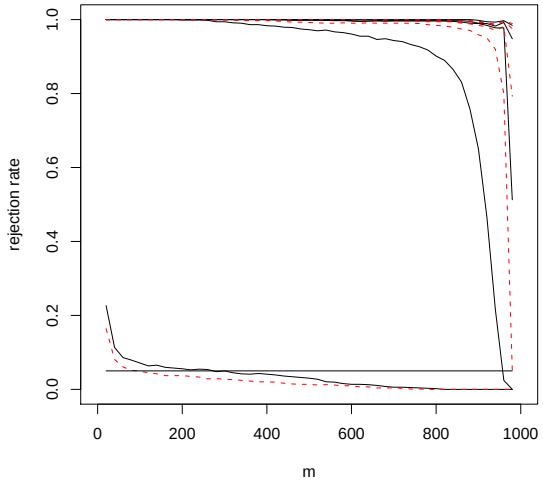
					$k=1$ $(1-\alpha)$			$k=2$ $(1-\alpha)$			$k=3$ $(1-\alpha)$		
p	q	n	balanced?	with repl.?	.90	.95	.99	.90	.95	.99	.90	.95	.99
1	1	100	N	N	0.706	0.786	0.892	0.708	0.777	0.883	0.714	0.787	0.874
1	1	1000	N	N	0.699	0.785	0.884	0.688	0.773	0.890	0.701	0.776	0.891
1	1	100	Y	N	0.685	0.772	0.864	0.695	0.773	0.867	0.694	0.783	0.867
1	1	1000	Y	N	0.700	0.780	0.889	0.691	0.771	0.888	0.705	0.787	0.881
1	1	100	N	Y	0.791	0.844	0.935	0.783	0.835	0.934	0.783	0.834	0.928
1	1	1000	N	Y	0.779	0.844	0.938	0.765	0.838	0.931	0.774	0.838	0.933
1	1	100	Y	Y	0.771	0.831	0.917	0.759	0.821	0.914	0.766	0.824	0.918
1	1	1000	Y	Y	0.773	0.849	0.936	0.773	0.835	0.929	0.769	0.843	0.930
3	1	100	N	N	0.616	0.694	0.782	0.618	0.684	0.776	0.623	0.682	0.768
3	1	1000	N	N	0.731	0.807	0.896	0.731	0.809	0.897	0.732	0.802	0.891
3	1	100	Y	N	0.601	0.673	0.769	0.621	0.680	0.764	0.626	0.688	0.770
3	1	1000	Y	N	0.736	0.808	0.899	0.729	0.803	0.896	0.740	0.801	0.891
3	1	100	N	Y	0.681	0.746	0.812	0.682	0.738	0.817	0.681	0.734	0.813
3	1	1000	N	Y	0.800	0.858	0.936	0.798	0.855	0.926	0.798	0.855	0.929
3	1	100	Y	Y	0.679	0.741	0.813	0.673	0.738	0.809	0.677	0.727	0.808
3	1	1000	Y	Y	0.797	0.857	0.929	0.802	0.848	0.930	0.797	0.855	0.928

Figure 1: Rejection Rates for Convexity Tests using $\widehat{\tau}_2(\mathcal{S}_n)$ and Resampling without Replacement

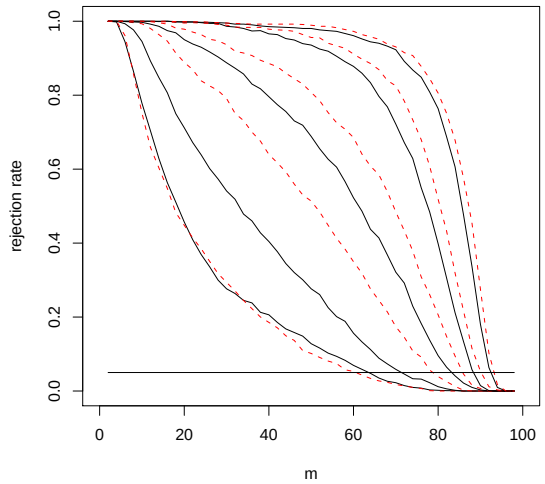
$p = q = 1,\ n = 100$



$p = q = 1,\ n = 1000$



$p = 3,\ q = 1, ; n = 100$



$p = 3,\ q = 1, ; n = 1000$

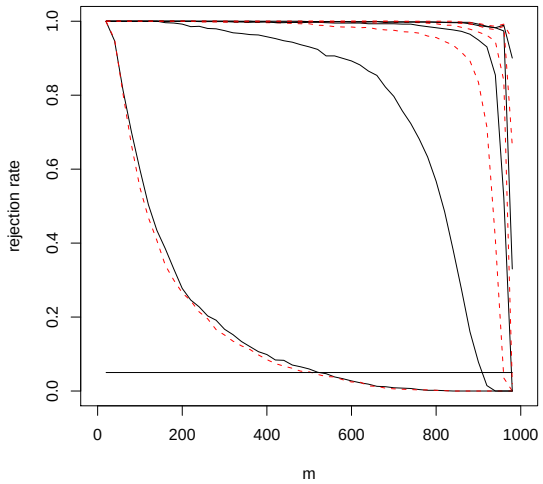


Figure 2: Estimated Critical Values versus Sub-Sample Size m for Convexity Tests using $\widehat{\tau}_2(\mathcal{S}_n)$ using Resampling without Replacement ($p = q = 1$, $n = 1000$)