

Testing whether Two-Stage Estimation is Meaningful in Non-Parametric Models of Production

CINZIA DARAIIO

LÉOPOLD SIMAR

PAUL W. WILSON*

June 2011

Abstract

Simar and Wilson (*J. Econometrics*, 2007) provided a statistical model that can rationalize two-stage estimation of technical efficiency in non-parametric settings. Two-stage estimation has been widely used, but requires a strong assumption: the second-stage environmental variables cannot affect the support of the input and output variables in the first stage. In this paper, we provide a fully non-parametric test of this assumption; in addition, we provide a theoretical link to results obtained by Politis et al. (*Statistica Sinica*, 2001), allowing us to estimate critical values for our test statistics using bootstrap sub-sampling while optimizing the choice of sub-sample size by minimizing a measure of volatility. Our simulation results indicate that our tests perform well both in terms of size and power. We present a real-world empirical example by updating the analysis performed by Aly et al. (*R. E. Stat.*, 1990) on U.S. commercial banks; our tests easily reject the assumption required for two-stage estimation, calling into question results that appear in *hundreds* of papers that have been published in recent years.

Keywords: technical efficiency, two-stage estimation, bootstrap, sub-sampling, data envelopment analysis (DEA).

*Daraio: Dipartimento di Scienze Aziendali, Università di Bologna, Bologna, Italy; email cinzia.daraio@unibo.it. Simar: Institut de Statistique, Université Catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium; email leopold.simar@uclouvain.be. Wilson: The John E. Walker Department of Economics, 222 Sarrine Hall, Clemson University, Clemson, South Carolina 29634-1309, USA; email wilson@clemson.edu. Financial support from the “Inter-university Attraction Pole”, Phase VI (No. P6/03) from the Belgian Government (Belgian Science Policy) and from l’Institut National de la Recherche Agronomique (INRA) and Le Groupe de Recherche en Economie Mathématique et Quantitative (GREMAQ), Toulouse School of Economics, Toulouse, France are gratefully acknowledged. Part of this research was done while Simar and Wilson were visiting professors at GREMAQ, and while Wilson was a visiting professor at the Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium. 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. Any remaining errors are solely our responsibility.

1 Introduction

Two-stage estimation procedures wherein technical efficiency is estimated by data envelopment analysis (DEA) or free disposal hull (FDH) estimators in the first stage, and the resulting efficiency estimates are regressed on some environmental variables in a second stage, are very popular in the literature.¹ Simar and Wilson (2007) cited 48 published papers that employed this approach and commented that “as far as we have been able to determine, none of the studies that employ this two-stage approach have described the underlying data-generating process.” Simar and Wilson went on to (i) define a statistical model where truncated (but not censored, i.e., tobit, nor ordinary least squares) regression yields consistent estimation of model features, (ii) demonstrated that conventional, likelihood-based approaches to inference are invalid, (iii) and developed a bootstrap approach that yields valid inference in the second-stage regression.

A number of papers have appeared in recent years using the approach suggested by Simar and Wilson (2007). However, papers that estimate technical efficiency in the first stage and then regress these estimates on some environmental variables in a second-stage tobit model continue to appear; a search on Google Scholar on 14 May 2010 using the keywords “dea,” “efficiency,” “tobit,” and “two stage” returned 362 papers with dates between 2007 and 2010. As far as we know, none of these papers present a statistical model in which second-stage tobit estimation would consistently estimate features of the model; the approach is ad hoc in each case.

Recently, Daraio and Simar (2005) have developed *conditional* measures of efficiency, which allow nonparametric estimation of technical efficiency conditional on some explanatory variables in a single stage. This raises some important questions for practitioners, such as the question of precisely how environmental variables might affect the production process. In the model presented by Simar and Wilson (2007), environmental variables affect the shape (i.e., mean, variance, etc.) of the distribution of inefficiencies, but not the support of input or output variables. Conceivably, however, environmental variables might have other effects; in particular, they might affect the production possibilities themselves. The statistical model in Simar and Wilson rationalizes second-stage regression of efficiency estimates on some

¹ The Google Scholar search engine returned about 4,120 articles after a search on “efficiency,” “two-stage,” and “dea” on 14 May 2010.

environmental variables, but does not allow for the possibility that environmental variables might affect the production possibilities. If they do, then a different model is needed, and second-stage regression may not be appropriate.

In this paper, we present a carefully-developed framework—i.e., a statistical model—in order to make clear how environmental variables might be relevant, and how to test whether two-stage approaches might be meaningful (i.e., whether assumptions given by Simar and Wilson, 2007 and required by most studies that have used the two-stage approach are satisfied). In addition, we describe a bootstrap method that can be used to assess the significance of test statistics without incurring a large computational burden. In cases where two-stage approaches are found to be inappropriate, one can estimate efficiency conditionally on environmental variables.

In the next section, we develop the statistical model. Estimators and test statistics are discussed in Section 3, and the bootstrap procedure necessary for inference is given in Section 4. Section 5 describes Monte Carlo experiments used to assess the size and power of our tests as well as results. In Section 6 we provide a real-world example by revisiting the work of Aly et al. (1990) and testing whether the assumptions given by Simar and Wilson (2007) that are required for the two-stage approach used by Aly et al. to be meaningful are satisfied. Conclusions are given in the final section.

2 Production in the Presence of Environmental Factors

Let $\mathbf{X} \in \mathbb{R}_+^p$ denote a vector of p input quantities, and let $\mathbf{Y} \in \mathbb{R}_+^q$ denote a vector of q output quantities. In addition, let $\mathbf{Z} \in \mathcal{Z} \subseteq \mathbb{R}^r$ denote a vector of r environmental variables with domain \mathcal{Z} . Firms transform quantities of inputs into various quantities of outputs, but environmental variables may affect how well firms do this on average, or possibly the range of possibilities for production. Given \mathbf{Z} , a firm becomes more technically efficient if it increases at least some of its output levels without increasing its input levels (output orientation), or alternatively if it reduces its use of at least some inputs without decreasing output levels (input orientation). In the real world, the analyst observes a set of observations $\mathcal{S}_n = \{(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i)\}_{i=1}^n$.

Assumption 2.1. *The sample observations $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)$ in \mathcal{S}_n are realizations of identically, independently distributed random variables $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ with probability density function $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ which has support over a compact set $\mathcal{P} \subset \mathbb{R}_+^{p+q} \times \mathbb{R}^r$ with level sets $\mathcal{P}(\mathbf{z})$ defined by*

$$\mathcal{P}(\mathbf{z}) = \{(\mathbf{X}, \mathbf{Y}) \mid \mathbf{Z} = \mathbf{z}, \mathbf{X} \text{ can produce } \mathbf{Y}\}. \quad (2.1)$$

Let

$$\Psi = \bigcup_{\mathbf{z} \in \mathcal{Z}} \mathcal{P}(\mathbf{z}) \subset \mathbb{R}_+^{p+q}. \quad (2.2)$$

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

Assumption 2.2. $\mathcal{P}(\mathbf{z}) = \Psi \forall \mathbf{z} \in \mathcal{Z}$.

Under Assumption 2.2, $\mathcal{P} = \Psi \times \mathcal{Z}$; this is the “separability” condition described by Simar and Wilson (2007), who also noted that this condition should be tested.

In the context of Assumptions 2.1–2.2, the standard, unconditional Farrell (1957) measure of output technical efficiency corresponding to an arbitrary point $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}_+^{p+q}$ is given by

$$\lambda = \lambda(\mathbf{x}, \mathbf{y}) \equiv \sup\{\lambda \mid (\mathbf{x}, \lambda \mathbf{y}) \in \Psi, \lambda > 0\} \quad (2.3)$$

and is the reciprocal of the Shephard (1970) output distance function. For $(\mathbf{x}, \mathbf{y}) \in \Psi$, $\lambda(\mathbf{x}, \mathbf{y}) \geq 1$. Note that λ provides a measure of Euclidean distance from the point $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}_+^{p+q}$ to the boundary of Ψ in a direction parallel to the output axes and orthogonal to the input axes.²

Note that $\mathbf{y} \in \mathbb{R}_+^q$ can be described in terms of polar coordinates; Simar and Wilson (2007) demonstrate that the modulus is related to the Farrell output efficiency measure in (2.3), and so \mathbf{y} can be described by $(\boldsymbol{\eta}, \lambda)$, where $\boldsymbol{\eta}$ is a vector of $(q - 1)$ angles and λ is the Farrell output efficiency corresponding to \mathbf{Y} . Then the joint density $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ can be described by a series of conditional densities and in terms of cylindrical coordinates:

$$f(\mathbf{x}, \boldsymbol{\eta}, \lambda, \mathbf{z}) = f(\mathbf{x}, \boldsymbol{\eta} \mid \lambda, \mathbf{z}) f(\lambda \mid \mathbf{z}) f(\mathbf{z}). \quad (2.4)$$

² Of course, one could also work in the input direction using the input-oriented analog of (2.7). In the remainder of the paper, we work only in the output direction to save space; analogous results for the input-orientation follow with suitable changes in notation.

The order of the conditioning on the right-hand side of (2.4) reflects the sequential nature of the DGP. Firm i is faced with environmental variables \mathbf{Z}_i drawn from $f(\mathbf{z})$. Given this \mathbf{Z}_i , an efficiency level λ_i is drawn from $f(\lambda \mid \mathbf{Z}_i)$, and then \mathbf{X}_i and $\boldsymbol{\eta}_i$ are drawn from $f(\mathbf{x}, \boldsymbol{\eta} \mid \lambda_i, \mathbf{Z}_i)$, resulting in a realization $(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i)$ from the joint density $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ after transforming the polar coordinates $(\boldsymbol{\eta}_i, \lambda_i)$ to Cartesian coordinates \mathbf{Y}_i .

Assumptions 2.1–2.2, together with the next assumption, comprise the nucleus of the model developed by Simar and Wilson (2007).

Assumption 2.3. *The conditioning in $f(\lambda \mid \mathbf{z})$ in (2.4) operates through the following mechanism:*

$$\lambda_i = \psi(\mathbf{Z}_i) + \varepsilon_i \geq 1, \tag{2.5}$$

where ψ is a smooth, continuous, function and ε_i is a continuous iid random variable, independent of \mathbf{Z}_i .

Assumptions 2.1–2.3 impose a “separability” condition on the model of Simar and Wilson (2007). In this model, $\mathcal{P}(\mathbf{z})$ is unaffected by \mathbf{z} , but the distribution of technical efficiency depends on \mathbf{z} through Assumption 2.3.³ As Simar and Wilson (2007, p. 36) noted, however, “one might reasonably wonder whether the implied separability condition is supported by the data.”

Alternatively, one might consider one of the formulations discussed by Coelli et al. (1997, pp. 166–171) where the environmental variables take the role of either inputs or outputs and may be either discretionary or non-discretionary. Simar and Wilson (2007) remark that the testing methods described by Simar and Wilson (2001) can be used to distinguish whether the covariates in \mathbf{Z} act as inputs or outputs. However, treating environmental variables as inputs or outputs incurs some problems. First, some environmental variables may be discrete; this may be problematic due to the assumptions required to obtain consistency of DEA and FDH estimators (see Simar and Wilson, 2008 for details). Second, incorporating elements of \mathbf{Z} as inputs or outputs requires knowing whether each element behaves as an input or as an output, which may be difficult to determine a priori. Third, this approach

³ In the empirical literature, researchers have typically assumed $\psi(\mathbf{Z}_i) = \mathbf{Z}_i\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is an $(r \times 1)$ vector of parameters. In addition to Assumptions 2.1–2.3, Simar and Wilson (2007) assume the error ε in (2.5) is distributed (truncated) normal in order to reflect the empirical literature. Alternatively, $\psi(\mathbf{Z}_i)$ and the distribution of ε can be assumed to be nonparametric; see Park et al. (2008) for details.

requires that the effects of the environmental variables must be monotonic; when DEA (as opposed to FDH) estimators are being used, their effects must also satisfy convexity of the production set. Perhaps most important, environmental variables describing regulatory regimes, climatic factors, etc. are likely to operate not as inputs or outputs, but as constraints on feasible combinations of inputs and outputs.

Alternatively, Assumption 2.2 may not hold; consequently, Assumptions 2.2–2.3 may be replaced with the following assumption.

Assumption 2.4. $\mathcal{P}(z) \neq \Psi$ for some $z \in \mathcal{Z}$; i.e., $\mathcal{P}(z) \neq \mathcal{P}(\tilde{z})$ for some $z \neq \tilde{z}$, $z, \tilde{z} \in \mathcal{Z}$.

If the sets $\mathcal{P}(z)$ are convex for all $z \in \mathcal{Z}$, then the set Ψ is convex under Assumption 2.2. However, if Assumption 2.4 holds, Ψ is not in general convex, even if the sets $\mathcal{P}(z)$ are convex for all $z \in \mathcal{Z}$. Moreover, under Assumption 2.4, some of the input-output combinations in Ψ are not attainable. Consequently, $\lambda(\mathbf{x}, \mathbf{y})$ measures distance from the point (\mathbf{x}, \mathbf{y}) to the frontier of an *unattainable* set when Assumption 2.4 holds, and thus has no meaning in familiar economic terms; moreover, as discussed below in Section 3, if Assumption 2.4 holds, the usual DEA estimators of Ψ and $\lambda(\mathbf{x}, \mathbf{y})$ are statistically inconsistent. This further implies that Assumption 2.3 is meaningless under Assumption 2.4; hence, regressing estimated efficiencies on environmental variables is also meaningless under Assumption 2.4. In cases where Assumption 2.4 holds, *conditional* measures of technical efficiency developed by Daraio and Simar (2005) are needed.

The joint density introduced in Assumption 2.1 implies a conditional distribution function

$$\begin{aligned} H(\mathbf{x}, \mathbf{y} \mid z) &= \Pr(\mathbf{X} \leq \mathbf{x}, \mathbf{Y} \geq \mathbf{y} \mid \mathbf{Z} = z) \\ &= \underbrace{\Pr(\mathbf{Y} \geq \mathbf{y} \mid \mathbf{X} \leq \mathbf{x}, \mathbf{Z} = z)}_{=S(\mathbf{y} \mid \mathbf{x}, z)} \underbrace{\Pr(\mathbf{X} \leq \mathbf{x} \mid \mathbf{Z} = z)}_{=F(\mathbf{x} \mid z)} \end{aligned} \quad (2.6)$$

where $S(\mathbf{y} \mid \mathbf{x}, z)$ is the conditional survivor function of \mathbf{Y} . Note that the conditioning on $\mathbf{X} \leq \mathbf{x}$ and $\mathbf{Z} = z$ is non-standard; however, the distribution function and its components are well-defined. A conditional version the Farrell (1957) output measure technical efficiency given in (2.3) can be defined by writing

$$\begin{aligned} \lambda(\mathbf{x}, \mathbf{y} \mid z) &\equiv \sup\{\lambda \mid (\mathbf{x}, \lambda \mathbf{y}) \in \mathcal{P}(z), \lambda > 0\} \\ &= \sup\{\lambda \mid S(\lambda \mathbf{y} \mid \mathbf{x}, z) > 0\}. \end{aligned} \quad (2.7)$$

One can adopt either Assumptions 2.1–2.3, or alternatively Assumptions 2.1 and 2.4. In either case, some additional assumptions are needed. The following assumptions on \mathcal{P} are standard in microeconomics; here, we adapt those of Shephard (1970) and Färe (1988) to account for the environmental variables:

Assumption 2.5. *For any $\mathbf{z} \in \mathcal{Z}$, $(\mathbf{x}, \mathbf{y}) \notin \mathcal{P}(\mathbf{z})$ if $\mathbf{x} = 0$, $\mathbf{y} \not\geq 0$; i.e., all production requires use of some inputs.*

Assumption 2.6. *For any $\mathbf{z} \in \mathcal{Z}$, $\tilde{\mathbf{x}} \geq \mathbf{x}$, and $\tilde{\mathbf{y}} \leq \mathbf{y}$, if $(\mathbf{x}, \mathbf{y}) \in \mathcal{P}(\mathbf{z})$ then $(\tilde{\mathbf{x}}, \mathbf{y}) \in \mathcal{P}(\mathbf{z})$ and $(\mathbf{x}, \tilde{\mathbf{y}}) \in \mathcal{P}(\mathbf{z})$, i.e., both inputs and outputs are strongly disposable.*

Throughout, inequalities involving vectors are assumed to hold element by element; e.g., $\mathbf{a} \leq \mathbf{b}$ denotes $a_j \leq b_j$ for each $j = 1, \dots, k$, where k is the length of \mathbf{a} and \mathbf{b} . Assumption 2.6 is equivalent to an assumption of monotonicity of the technology.

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

Assumption 2.7. *The n observations in \mathcal{S}_n are identically, independently distributed (iid) random variables on the attainable set \mathcal{P} .*

Assumption 2.8. *(a) Conditional on $\mathbf{Z} = \mathbf{z}$, the (\mathbf{X}, \mathbf{Y}) possess a joint density $f_{XY|Z}$ with support $\mathcal{P}(\mathbf{z})$; (b) $f_{XY|Z}$ is continuous on $\mathcal{P}(\mathbf{z})$; and (c) $f_{XY|Z}(\mathbf{x}, \lambda(\mathbf{x}, \mathbf{y})\mathbf{y}) > 0 \forall (\mathbf{x}, \mathbf{y})$ in the interior of $\mathcal{P}(\mathbf{z})$.*

Assumption 2.8(c) imposes a discontinuity in f at points on the boundary of $\mathcal{P}(\mathbf{z})$, ensuring a strictly positive, non-negligible probability of observing production units close to the production frontier. For points lying outside $\mathcal{P}(\mathbf{z})$, $f \equiv 0$.

Assumption 2.9. *The function $\lambda(\mathbf{x}, \mathbf{y} | \mathbf{z})$ is twice continuously differentiable for all $\mathbf{z} \in \mathcal{Z}$ and $(\mathbf{x}, \mathbf{y}) \in \mathcal{P}(\mathbf{z})$.*

Assumption 2.9 imposes some smoothness on the boundary of \mathcal{P} . This assumption is slightly stronger, but simpler, than a corresponding assumption needed by Kneip et al.

(1998) to establish consistency of the DEA estimators. We have adopted Assumption 2.9 from Kneip et al. (2008) and Jeong et al. (2010), where additional discussion is given.

Recall that the empirical researcher only observes the n observations in the sample \mathcal{S}_n . In order to decide between the model of Simar and Wilson (2007) defined by Assumptions 2.1–2.3 and 2.5–2.9, or alternatively the model defined by Assumptions 2.1 and 2.4–2.9, the researcher must test Assumption 2.2 versus Assumption 2.4. In other words, the empiricist must test $H_0: \mathcal{P}(\mathbf{z}) = \Psi \forall \mathbf{z} \in \mathcal{Z}$ versus $H_1: \mathcal{P}(\mathbf{z}) \neq \Psi$ for some $\mathbf{z} \in \mathcal{Z}$. In order to implement a test of this null hypothesis, estimators of $\lambda(\mathbf{x}, \mathbf{y})$ and $\lambda(\mathbf{x}, \mathbf{y} \mid \mathbf{z})$ are needed; these are introduced in the next section.

3 Non-parametric Efficiency Estimators

The distance functions in Section 2 are defined in terms of the *unknown, true* production set \mathcal{P} , and must be *estimated* from the set \mathcal{S}_n of sample observations. In cases where Assumption 2.2 holds, traditional non-parametric approaches used in analyses of efficiency and production typically assume $\Pr((\mathbf{x}_i, \mathbf{y}_i) \in \Psi) = 1 \forall i = 1, \dots, n$ and replace Ψ in (2.3) with an estimator of the unobserved production set Ψ to obtain estimators of the Farrell (1957) input- and output-oriented efficiency measures. Several possibilities exist.

Deprins et al. (1984) proposed estimating Ψ by the free-disposal hull (FDH) of the observed pairs $(\mathbf{X}_i, \mathbf{Y}_i)$ in \mathcal{S}_n , i.e.,

$$\widehat{\Psi}_{\text{FDH}}(\mathcal{S}_n) = \bigcup_{(\mathbf{X}_i, \mathbf{Y}_i) \in \mathcal{S}_n} \{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}_+^{p+q} \mid \mathbf{y} \leq \mathbf{Y}_i, \mathbf{x} \geq \mathbf{X}_i\}. \quad (3.1)$$

This estimator is consistent under Assumptions 2.1–2.3 and 2.5–2.9, but is not consistent if Assumption 2.4 replaces Assumption 2.2. If, in addition to Assumptions 2.1–2.3 and 2.5–2.9, one is willing to assume that Ψ is convex, then the convex hull of $\widehat{\Psi}_{\text{FDH}}$,

$$\widehat{\Psi}_{\text{DEA}}(\mathcal{S}_n) = \left\{ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}_+^{p+q} \mid \mathbf{y} \leq \sum_{i=1}^n \omega_i \mathbf{Y}_i, \mathbf{x} \geq \sum_{i=1}^n \omega_i \mathbf{X}_i, \sum_{i=1}^n \omega_i = 1, \omega_i \geq 0 \forall i = 1, \dots, n \right\}, \quad (3.2)$$

can be used to consistently estimate Ψ . As with the FDH estimator, however, consistency is lost if Assumption 2.4 replaces Assumption 2.2.

As a practical matter, DEA estimates of input or output distance functions are obtained by solving familiar linear programs; in the case of the output-oriented measure defined in (2.3), one would compute

$$\widehat{\lambda}_{\text{DEA}}(\mathbf{x}, \mathbf{y} \mid \mathcal{S}_n) = \max_{\lambda, \omega_1, \dots, \omega_n} \left\{ \lambda > 0 \mid \lambda \mathbf{y} \leq \sum_{i=1}^n \omega_i \mathbf{Y}_i, \mathbf{x} \geq \sum_{i=1}^n \omega_i \mathbf{X}_i, \sum_{i=1}^n \omega_i = 1, \omega_i \geq 0 \forall i = 1, \dots, n \right\}. \quad (3.3)$$

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 output orientation, one can compute

$$\widehat{\lambda}_{\text{FDH}}(\mathbf{x}, \mathbf{y} \mid \mathcal{S}_n) = \max_{i=1, \dots, n \mid \mathbf{X}_i \leq \mathbf{x}} \left(\min_{j=1, \dots, p} \left(\frac{\mathbf{Y}_i^j}{\mathbf{y}^j} \right) \right), \quad (3.4)$$

where $\mathbf{y}^j, \mathbf{Y}_i^j$ denote the j th elements of \mathbf{y} (i.e., the input vector corresponding to the fixed point of interest) and \mathbf{Y}_i (i.e., the output vector corresponding to the i th observation in \mathcal{S}_n).

Asymptotic properties of the estimators in (3.3)–(3.4) based on $\widehat{\Psi}_{\text{DEA}}(\mathcal{S}_n)$ and $\widehat{\Psi}_{\text{FDH}}(\mathcal{S}_n)$, as well as the assumptions needed to establish consistency of the estimators, are summarized in Simar and Wilson (2000b). In particular, under the assumption of convexity and Assumptions 2.1–2.3, 2.5–2.9, $\widehat{\lambda}_{\text{DEA}}(\mathbf{x}, \mathbf{y} \mid \mathcal{S}_n)$ is a consistent estimator of $\lambda(\mathbf{x}, \mathbf{y})$, with convergence rate $n^{-2/(p+q+1)}$ (Kneip et al., 1998).⁴ Irrespective of whether Ψ is convex, under Assumptions 2.1–2.3 and 2.5–2.9, $\widehat{\lambda}_{\text{FDH}}(\mathbf{x}, \mathbf{y} \mid \mathcal{S}_n)$ is a consistent estimator of $\lambda(\mathbf{x}, \mathbf{y})$, but with convergence rate $n^{-1/(p+q)}$ (Park et al., 2000).

In cases where Assumption 2.4 replaces Assumption 2.2, one can use the conditional estimators introduced by Daraio and Simar (2005, 2007a, 2007b). In particular, the conditional FDH estimator of $\lambda(\mathbf{x}, \mathbf{y} \mid \mathbf{z})$ defined in (2.7) is given by

$$\widehat{\lambda}_{\text{FDH}}(\mathbf{x}, \mathbf{y} \mid \mathbf{z}, \mathcal{S}_n) = \max_{\substack{i=1, \dots, n \\ \mathbf{X}_i \leq \mathbf{x} \cap \|\mathbf{Z}_i - \mathbf{z}\| \leq h}} \left(\min_{j=1, \dots, p} \left(\frac{\mathbf{Y}_i^j}{\mathbf{y}^j} \right) \right), \quad (3.5)$$

⁴ If, in addition, one assumes globally constant returns to scale and drops the constraint $\sum_{i=1}^n \omega_i = 1$ from the right-hand side of (3.6), the convergence rate becomes $n^{-2/(p+q)}$ (Park et al., 2010).

where \mathbf{h} is an r -vector of bandwidth parameters. Alternatively, Daraio and Simar (2007b) define a conditional DEA estimator of $\lambda(\mathbf{x}, \mathbf{y} \mid \mathbf{z})$, namely

$$\widehat{\lambda}_{\text{DEA}}(\mathbf{x}, \mathbf{y} \mid \mathbf{z}, \mathcal{S}_n) = \max_{\lambda, \omega_1, \dots, \omega_n} \left\{ \lambda > 0 \mid \lambda \mathbf{y} \leq \sum_{\substack{i=1, \dots, n \\ |\mathbf{Z}_i - \mathbf{z}| \leq \mathbf{h}}} \omega_i \mathbf{Y}_i, \mathbf{x} \geq \sum_{\substack{i=1, \dots, n \\ |\mathbf{Z}_i - \mathbf{z}| \leq \mathbf{h}}} \omega_i \mathbf{X}_i, \right. \\ \left. \sum_{\substack{i=1, \dots, n \\ |\mathbf{Z}_i - \mathbf{z}| \leq \mathbf{h}}} \omega_i = 1, n \omega_i \geq 0 \forall i \text{ such that } |\mathbf{Z}_i - \mathbf{z}| \leq \mathbf{h} \right\} \quad (3.6)$$

where \mathbf{h} is again an r -vector of bandwidth parameters.

Asymptotic results for both of the conditional estimators in (3.5)–(3.6) are given by Jeong et al. (2010); with appropriate size of the bandwidth, the convergence rates of the conditional estimators are slower than their unconditional counterparts by a factor $n^{-4/(4+r)}$. For either estimator in (3.5)–(3.6), the bandwidth parameter h can be optimized using the least-squares cross-validation technique discussed by Bădin et al. (2010), provided the elements of \mathbf{Z} are continuous. If \mathbf{Z} contains qualitative variables, then the sample observations must first be divided into groups defined by the qualitative variables. For example, if \mathbf{Z} includes r_c continuous variables and r_d binary dummy variables, we can partition \mathbf{Z} by writing $\mathbf{Z} = [\mathbf{Z}^c \ \mathbf{Z}^d]$. There are potentially 2^{r_d} groups of observations where the r_d binary variables have the same values within each group (note some cells may be empty). For each group, either $\widehat{\lambda}_{\text{FDH}}(\mathbf{x}, \mathbf{y} \mid \mathbf{z}, \mathcal{S}_n)$ or $\widehat{\lambda}_{\text{DEA}}(\mathbf{x}, \mathbf{y} \mid \mathbf{z}, \mathcal{S}_n)$ can be computed using only the continuous elements \mathbf{Z}^c of \mathbf{Z} , optimizing the bandwidth for continuous components of \mathbf{Z} using only observations within the given group. By computing estimates separately for each group, one necessarily conditions on $\mathbf{Z}^d = \mathbf{z}^d$. Dividing into groups is necessary since the bandwidths in (3.5)–(3.6) merely determine which observations fall into the reference set that defines the estimators. This contrasts with nonparametric regression estimators such as the Nadaraya-Watson estimator, where bandwidths determine kernel weights; here, however, no smoothing is performed since only the support of $H(\mathbf{x}, \mathbf{y} \mid \mathbf{z})$ is estimated.

In order to test the null hypothesis given at the end of Section 2, consider the test statistics

$$\widehat{\tau}_{\text{FDH},n}(\mathcal{S}_n) = n^{-1} \sum_{i=1}^n \widehat{\mathbf{D}}'_{\text{FDH},i} \widehat{\mathbf{D}}_{\text{FDH},i} \geq 0 \quad (3.7)$$

and

$$\widehat{\tau}_{\text{DEA},n}(\mathcal{S}_n) = n^{-1} \sum_{i=1}^n \widehat{\mathbf{D}}'_{\text{DEA},i} \widehat{\mathbf{D}}_{\text{DEA},i} \geq 0 \quad (3.8)$$

where $\widehat{\mathbf{D}}_{\text{FDH},i} = \left(\mathbf{Y}_i \widehat{\lambda}_{\text{FDH}}(\mathbf{X}_i, \mathbf{Y}_i \mid \mathcal{S}_n) - \mathbf{Y}_i \widehat{\lambda}_{\text{FDH}}(\mathbf{X}_i, \mathbf{Y}_i \mid \mathbf{Z}_i, \mathcal{S}_n) \right)$ and $\widehat{\mathbf{D}}_{\text{DEA},i} = \left(\mathbf{Y}_i \widehat{\lambda}_{\text{DEA}}(\mathbf{X}_i, \mathbf{Y}_i \mid \mathcal{S}_n) - \mathbf{Y}_i \widehat{\lambda}_{\text{DEA}}(\mathbf{X}_i, \mathbf{Y}_i \mid \mathbf{Z}_i, \mathcal{S}_n) \right)$ are $(q \times 1)$ vectors. The statistics defined by (3.7)–(3.8) give estimates of the mean integrated square difference between \mathcal{P} and $\Psi \times \mathcal{Z}$. If Assumption 2.2 holds, we should expect these statistics to be “close” to zero; otherwise, we should expect them to lie “far” from zero.

The sub-sampling method described by Simar and Wilson (2009) can be used to determine critical values in order to implement tests using the statistics defined above. We give an overview of this method in the next section.

4 Bootstrap Inference using Sub-Sampling

4.1 A Probabilistic Framework for Testing

In order to test the “separability” condition in Assumption 2.2 (versus the alternative given by Assumption 2.4), we must first define a probabilistic framework within which the model characteristic to be tested can be described. This is necessary in order to find critical values for the statistics defined at the end of Section 3. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be the probability space on which the random variables \mathbf{X} , \mathbf{Y} , and \mathbf{Z} are defined; by Assumption 2.1, \mathcal{P} is the support of the joint distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. 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

$$\tau(P) = E(\mathbf{D}'\mathbf{D}) \tag{4.1}$$

where $\mathbf{D} = (\mathbf{Y}\lambda(\mathbf{X}, \mathbf{Y}) - \mathbf{Y}\lambda(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}))$, a $(q \times 1)$ vector. Since $\mathbf{D}'\mathbf{D}$ is a Borel (measurable) function, the expectation $\tau(P)$ is well-defined. Assume that the variance of $\mathbf{D}'\mathbf{D}$, denoted by $\sigma^2(P)$, is finite. By construction, $\tau(P) \geq 0 \forall P \in \mathbb{P}$, with $\tau(P) = 0$ if $P \in \mathbb{P}_0$ and $\tau(P) > 0$ if $P \in \mathbb{P}_1$. Hence testing the null amounts to testing $H_0: \tau(P) = 0$ versus $H_1: \tau(P) > 0$.

Consistent estimators of $\tau(P)$ are given by (3.7) and, under convexity of $\mathcal{P}(\mathbf{z})$, by (3.8). To simplify notation, let $\mathbf{W} = (\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ denote a generic observation. Define

$$T(\mathbf{W}) = \mathbf{D}'\mathbf{D} \tag{4.2}$$

and

$$\widehat{T}(\mathbf{W} \mid \mathcal{S}_n) = \widehat{\mathbf{D}}' \widehat{\mathbf{D}}, \quad (4.3)$$

where $\mathcal{S}_n = \{\mathbf{W}_i\}_{i=1}^n$; $\widehat{\mathbf{D}}$ can represent either $\widehat{\mathbf{D}}_{\text{FDH}}$ or $\widehat{\mathbf{D}}_{\text{DEA}}$ defined near the end of Section 3. Then

$$\tau(P) = E(T(\mathbf{W})) \quad (4.4)$$

and

$$\tau_n(\mathcal{S}_n) = n^{-1} \sum_{i=1}^n \widehat{T}(\mathbf{W}_i \mid \mathcal{S}_n), \quad (4.5)$$

with τ_n representing either of the statistics defined in (3.7)–(3.8), depending on whether $\widehat{\mathbf{D}}$ in (4.3) represents either $\widehat{\mathbf{D}}_{\text{FDH}}$ or $\widehat{\mathbf{D}}_{\text{DEA}}$.

4.2 Asymptotic Behavior of $\widehat{T}(\mathbf{W} \mid \mathcal{S}_n)$

The framework introduced above ensures that for all P , $T(\mathbf{W}) \stackrel{a.s.}{\geq} 0$ and if $P \in \mathbb{P}_0$, then $T(\mathbf{W}) \stackrel{a.s.}{=} 0$. In addition, either under Assumptions 2.1–2.3 and 2.5–2.9, or under Assumptions 2.1 and 2.4–2.9, for all fixed $\mathbf{w} \in \mathcal{P}$,

$$n^\kappa \left(\widehat{T}(\mathbf{w} \mid \mathcal{S}_n) - T(\mathbf{w}) \right) \xrightarrow{\mathcal{L}} G_P(\cdot \mid \mathbf{w}), \quad (4.6)$$

where $G_P(\cdot \mid \mathbf{w})$ is a nondegenerate distribution whose characteristics depends on \mathbf{w} . The value of κ is known and depends on which of the statistics in (3.7)–(3.8) is being used. This rate is governed by the smallest rate of convergence of the DEA or FDH estimators used to define $T(\mathbf{W})$; for example, $\kappa = \left(\frac{4}{4+r}\right) \left(\frac{1}{p+q}\right)$ when the statistic in (3.7) based on the conditional FDH estimator is used, or $\kappa = \left(\frac{4}{4+r}\right) \left(\frac{2}{p+q+1}\right)$ when the statistic in (3.8) based on the conditional DEA estimator is used. This implies that

$$\lim_{n \rightarrow \infty} \Pr \left[n^\kappa \left(\widehat{T}(\mathbf{w} \mid \mathcal{S}_n) - T(\mathbf{w}) \right) \leq a \right] = G_P(a \mid \mathbf{w}). \quad (4.7)$$

Since $\widehat{T}(\mathbf{W} \mid \mathcal{S}_n)$ and $T(\mathbf{W})$ are well-defined random variables on (Ω, \mathcal{A}) , (4.7) can be considered as a conditional statement, with conditioning on $\mathbf{W} = \mathbf{w}$; hence

$$\lim_{n \rightarrow \infty} \Pr \left[n^\kappa \left(\widehat{T}(\mathbf{W} \mid \mathcal{S}_n) - T(\mathbf{W}) \right) \leq a \mid \mathbf{W} = \mathbf{w} \right] = G_P(a \mid \mathbf{w}). \quad (4.8)$$

By marginalizing on \mathbf{W} , we have

$$\lim_{n \rightarrow \infty} \Pr \left[n^\kappa \left(\widehat{T}(\mathbf{W} \mid \mathcal{S}_n) - T(\mathbf{W}) \right) \leq a \right] = \int_{\mathcal{P}} G_P(a \mid \mathbf{w}) f_W(\mathbf{w}) d\mathbf{w} = Q_P(a). \quad (4.9)$$

Note that the density introduced in Assumption 2.1 has been re-written here as $f_W(\cdot)$. Since $G_P(\cdot | \mathbf{w})$ and $f_W(\cdot)$ are nondegenerate, $Q_P(\cdot)$ is a nondegenerate distribution. It follows that

$$n^\kappa(\widehat{T}(\mathbf{W} | \mathcal{S}_n) - T(\mathbf{W})) \xrightarrow{\mathcal{L}} Q_P(\cdot). \quad (4.10)$$

Now let μ_{Q_P} and $\sigma_{Q_P}^2$ denote the finite mean and strictly positive variance of $Q_P(\cdot)$. Since $\widehat{T}(\mathbf{W} | \mathcal{S}_n) = T(\mathbf{W}) + n^{-\kappa}\xi(\mathbf{W})$, $\xi(\mathbf{W})$ must have limiting distribution $Q_P(\cdot)$ as $n \rightarrow \infty$. Combining this result with (4.4), we have

$$E(\widehat{T}(\mathbf{W} | \mathcal{S}_n)) = \tau(P) + \mu_{Q_P}/n^\kappa \quad (4.11)$$

and

$$\text{VAR}(\widehat{T}(\mathbf{W} | \mathcal{S}_n)) = \sigma^2(P) + \sigma(P)O(n^{-\kappa}) + \sigma_{Q_P}^2/n^{2\kappa}, \quad (4.12)$$

where the second term in (4.12) accounts for the covariance between $T(\mathbf{W})$ and $n^{-\kappa}\xi(\mathbf{W})$ (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 formulas (4.11) and (4.12) simplify accordingly.

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

From the results in (4.11) and (4.12), 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 $\widehat{T}(\mathbf{W}_j; \mathcal{S}_n)$ and $\widehat{T}(\mathbf{W}_k; \mathcal{S}_n)$, 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. (2011). 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. (2011), consider a point $\mathbf{w} \in \mathcal{P}$ where the DEA score (i.e., efficiency estimate) is evaluated and let $C_z(\zeta)$ be a neighborhood of the frontier point \mathbf{w}^∂ determined by the projection of the point \mathbf{w} on the true frontier; ζ is a bandwidth that controls the size of this neighborhood. If $\zeta^2 = O(n^{-2/(p+q+1)})$, $C_z(\zeta)$ will contain the DEA estimate of the frontier at \mathbf{w} with probability 1. Since $\zeta \rightarrow 0$ as $n \rightarrow \infty$, the probability of an observation \mathbf{W}_i falling in $C_z(\zeta)$ is approximated by

$$\pi_n = \Pr(\mathbf{W} \in C_z(\zeta)) \approx f_W(\mathbf{w}^\partial)(2\zeta)^{p+q-1}\zeta^2 = O(n^{-1}). \quad (4.13)$$

For large n , the distribution of the number of points \mathbf{W}_i falling in $C_z(\zeta)$ follows approximately a Poisson distribution with parameter $n\pi_n = O(1)$. As shown in Kneip et al. (2011), when $n \rightarrow \infty$, only points falling in this neighborhood influence the distribution of the DEA estimator at the point \mathbf{w} . The number of such points is $O(1)$. Consequently, the covariances between the DEA estimator at \mathbf{W}_j and the $(n - 1)$ DEA estimators at the other points \mathbf{W}_k is nonzero for at most $O(1)$ of these $(n - 1)$ estimators.⁵ Moreover, each of the nonzero covariances is bounded by the product of the standard deviations derived from (4.12); therefore, the n covariance terms sum to $nO(1)[\sigma^2(P) + \sigma(P)O(n^{-\kappa}) + \sigma_{Q_P}^2/n^{2\kappa}]$. Combining these results, we obtain

$$E(\tau_n(\mathcal{S}_n)) = \tau(P) + \frac{\mu_{Q_P}}{n^\kappa} \quad (4.14)$$

and

$$\text{VAR}(\tau_n(\mathcal{S}_n)) = \frac{1}{n^2} \{n \times [\sigma_{Q_P}^2/n^{2\kappa} + O(n^{-\kappa})\sigma(P) + \sigma^2(P)]\} = O(n^{-1}). \quad (4.15)$$

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 (4.14) we also see that μ_{Q_P}/n^κ acts as a bias term that disappears asymptotically. Under the null, since $\tau(P) = \sigma(P) = 0$, we obtain for all $P \in \mathbb{P}_0$,

$$E(\tau_n(\mathcal{S}_n)) = \mu_{Q_P}/n^\kappa \quad (4.16)$$

and

$$\text{VAR}(\tau_n(\mathcal{S}_n)) = \sigma_{Q_P}^2/n^{1+2\kappa} = O(n^{-(1+2\kappa)}), \quad (4.17)$$

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 sub-sampling approximation to the distribution of our test statistics 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 $f_W(\cdot)$ in order to obtain a normal limiting distribution.

Proposition 4.1. *If the joint density $f_W(\mathbf{w})$ of \mathbf{W} is such that the moments of $Q_P(\cdot)$ exist up to the fourth order, then under the null hypothesis $H_0: P \in \mathbb{P}_0$,*

$$n^\kappa \sqrt{n} (\tau_n(\mathcal{S}_n) - \mu_{Q_P}/n^\kappa) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_{Q_P}^2). \quad (4.18)$$

⁵ In the case of the FDH estimators, value of the estimators is determined by only one point, and so $O(1)$ becomes simply 1 for the FDH case (see Jeong and Simar, 2006 for additional discussion).

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

$$\begin{array}{cccc} \widehat{T}(\mathbf{W}_1; \mathcal{S}_1); & & & \\ \widehat{T}(\mathbf{W}_1; \mathcal{S}_2) & \widehat{T}(\mathbf{W}_2; \mathcal{S}_2); & & \\ \vdots & & & \\ \widehat{T}(\mathbf{W}_1; \mathcal{S}_n) & \widehat{T}(\mathbf{W}_2; \mathcal{S}_n) & \dots & \widehat{T}(\mathbf{W}_n; \mathcal{S}_n); \\ \vdots & & & \end{array}$$

The mean and the variance of the sums were derived above. As explained by Simar and Wilson (2009), Proposition 4.1 follows from the Lyapunov condition with $\nu = 3$ (see the corollary in Section 1.9.3 of Serfling), i.e.,

$$\frac{nE|\widehat{T}(\mathbf{W}_j; \mathcal{S}_n) - \mu_{Q_P}/n^\kappa|^3}{(n\sigma_{Q_P}^2/n^{2\kappa})^{3/2}} = o(1), \quad (4.19)$$

which holds provided moments of $Q_P(\cdot)$ exist up to fourth order.⁶

4.4 Testing by Subsampling

In principle, one could use the asymptotic result in Proposition 4.1 for inference. However, this would require estimating both μ_{Q_P} and $\sigma_{Q_P}^2$, and one should not expect the asymptotic normal approximation to give good results except in perhaps extraordinarily large samples, perhaps larger than those typically encountered by practitioners. Hence, it is safer for the practitioner to estimate critical values using the bootstrap sub-sampling methods described below.

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

⁶ 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 (4.14) and (4.15). Of course, to satisfy the Lyapunov condition an additional technical regularity condition on the random variable $T(\mathbf{W})$ is needed. In particular, $T(\mathbf{W})$ must have finite moments up to order 4 (when H_0 is true, $P \in \mathbb{P}_0$ and $T(\mathbf{W})$ is a degenerate random variable equal to zero). Hence, for $P \in \mathbb{P}$,

$$\sqrt{n}(\tau_n(\mathcal{S}_n) - (\tau(P) + \mu_{Q_P}/n^\kappa)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_{Q_P}^2/n^{2\kappa} + O(n^{-\kappa})\sigma(P) + \sigma^2(P)).$$

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

$$\hat{G}_{m,n}(a) = \frac{1}{B} \sum_{b=1}^B \mathbb{I}(m^\kappa \sqrt{m} \tau_m(\mathcal{S}_m^{*,b}) \leq a), \quad (4.20)$$

where $\mathbb{I}()$ denotes the indicator function, $B \leq \binom{n}{m}$ is the number of bootstrap replications, and $\{\tau_m(\mathcal{S}_m^{*,b})\}_{b=1}^B$ is the set of bootstrap estimates, each computed from different random sub-samples 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 $m, n \rightarrow \infty$ with $m/n \rightarrow 0$.

Note that in the procedure proposed here, we neglect the bias term μ_{Q_P}/n^κ appearing in (4.18). This bias term could be estimated while performing the bootstrap computations, but Monte-Carlo results from Simar and Wilson (2009) 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.

In the context of DEA estimates of technical efficiency, Kneip et al. (2008) proved that a sub-sampling approximation for the distribution of a DEA efficiency estimator is consistent for any choice $m = n^\gamma$ of the sub-sample size with $\gamma \in (0, 1)$. However, their Monte Carlo results reveal that the quality of the approximation in finite samples depends crucially on γ , i.e., the choice of the sub-sample size m . Using extensive Monte Carlo experiments, Simar and Wilson (2009) showed that ideas discussed by Politis et al. (2001) and Bickel and Sakov (2008) for choosing m (or equivalently, choosing γ) yield confidence interval estimates with reasonable coverage properties as well as critical values resulting in tests with good size and power properties.

Both Politis et al. (2001) and Bickel and Sakov (2008) proposed computing the object of interest (e.g., the 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; we use the approach of Politis et al.. One can compute the bootstrap approximation in (4.20) for various sub-sample sizes $m_1 < m_2 < \dots < m_J$, and obtain critical values c_j corresponding to each value of m_j for tests of (nominal) size α . Then, volatility corresponding to m_j can be

measured by computing the standard deviations of the critical values $c_{j-k}, \dots, c_j, \dots, c_{j+k}$ corresponding to sub-sample sizes $m_{j-k}, \dots, m_j, \dots, m_{j+k}$ for some small integer k , where $j = (k + 1), \dots, (J - k)$. 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 (4.20) will become too concentrated as $m \rightarrow n$; in fact, if $m = n$, the bootstrap distribution collapses to a single probability mass. On the other hand, as $m \rightarrow 0$, the resulting critical values will not be informative 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 critical values since as explained by Politis et al. (2001), the bootstrap approximation is valid for a wide range of values of sub-sample sizes m .

5 Monte Carlo Experiments

In order to examine the performance of our tests in terms of size and power, we conducted a series of Monte Carlo experiments for sample sizes $n \in \{50, 100, 200, 400\}$. We consider four DGPs in our experiments. For $p = q = r = 1$, we consider a DGP where the null hypothesis is false (i.e., where Assumption 2.4 holds) and another DGP where the null is true (i.e., where Assumption 2.2 holds). For $p = q = r = 2$, we consider another two DGPs, again one where the null is false, and one where it is true.

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 $m \in \mathbb{M}_n = \{\frac{n}{50}, \frac{2n}{50}, \frac{3n}{50}, \dots, \frac{49n}{50}\}$. 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 4.4. All experiments are conducted using resampling without replacement.⁷

In each experiment 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

⁷ The “optimal” sub-sample size is chosen as in Simar and Wilson (2009). Simar and Wilson found that in their Monte Carlo experiments examining tests of convexity of Ψ , the optimal sub-sample size m was smaller when resampling was done with replacement as opposed to without replacement. In addition, holding dimensionality and sample size constant, resampling with replacement typically resulted in less test power than resampling without replacement for given sub-sample sizes and departures from the null.

replications and compute corresponding critical values $\{c_1, c_2, \dots, c_{49}\}$ 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 $j \in \{J_{lo}, \dots, J_{hi}\}$ and for a small integer value k , compute volatility indices given by the standard deviations \widehat{s}_j of the critical values $\{c_{j-k}, \dots, c_{j+k}\}$.
- [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 $\widehat{m} = m_{\widehat{j}}$.

In all of our experiments, differences in results across the three different values of k were small; since $k = 1$ typically yielded the best results in terms of achieved size and power, we report only results for $k = 1$ below in order to conserve space.

In order to generate efficient input-output combinations $(\mathbf{x}, \mathbf{y}^+)$, let \mathbf{v} be a draw from the uniform distribution on the surface of the unit $(p + q - 1)$ -sphere centered at the origin.⁸ The vector \mathbf{v} has length $(p + q)$, and can be partitioned by writing $\mathbf{v} = [\mathbf{v}_x \ \mathbf{v}_y]$ where \mathbf{v}_x has length p and \mathbf{v}_y has length q . Now define

$$\mathbf{x} = 1 - |\mathbf{v}_x| \tag{5.1}$$

and

$$\mathbf{y}^+ = |\mathbf{v}_y|. \tag{5.2}$$

Then $(\mathbf{x} - 1, \mathbf{y}^+)$ represents a draw from the uniform distribution on the surface defined by the intersection of the unit $(p + q - 1)$ -sphere centered at the origin and the closed orthant in \mathbb{R}^{p+q} defined by the Cartesian product $\mathbb{R}_-^p \times \mathbb{R}_+^q$. The addition of 1 to $-|\mathbf{v}_x|$ on the right-hand side of (5.1) amounts to shifting this surface one unit in the positive direction along the axes in \mathbb{R}^p -space. In the case where $p = q = 1$, $(\mathbf{x}, \mathbf{y}^+)$ represents a draw from the uniform distribution on the northwest quarter of a circle of radius 1 centered at $(1, 0)$ in \mathbb{R}^2 .

For $p = q = r = 1$, we considered two DGPs as noted above. In the first DGP, we define

$$\text{DGP \#1:} \quad y = y^+ e^{-\delta(z-2)^2} e^{-|u|} \tag{5.3}$$

⁸ Recall that for any natural number d , a unit d -sphere is the set of points in $(d+1)$ -dimensional Euclidean space lying at distance one from a central point; the set of points comprises a d -dimensional manifold in Euclidean $(d+1)$ -space. A draw \mathbf{v} from the uniform distribution on the surface of the unit $(p + q - 1)$ -sphere can be simulated using the method of Muller (1959) and Marsaglia (1972).

where $u \sim N(0, \sigma_u^2)$ with $\sigma_u^2 = 0.17$, $z \sim \text{uniform on } [0, 4]$, and $\delta \geq 0$. In the second DGP, we define

$$\text{DGP \#2: } \quad y = y^+ e^{-|uz^\delta|} \quad (5.4)$$

where $z \sim \text{uniform on } [0.8, 1.2]$ and u and δ are defined as before. We also considered two DGPs for the case $p = q = r = 2$, namely

$$\text{DGP \#3: } \quad \mathbf{y} = \mathbf{y}^+ e^{-\delta(z_1 - z_2)^2} e^{-|u|} \quad (5.5)$$

where z_1 and z_2 are distributed uniform on $[0, 4]$, and

$$\text{DGP \#4: } \quad \mathbf{y} = \mathbf{y}^+ e^{-|u((z_1 - z_2)^2)^\delta|} \quad (5.6)$$

where z_1 and z_2 are distributed uniform on $[0, 0.4]$; in both cases, u and δ are defined as above.

In DGPs #2 and #4, the environmental variables \mathbf{z} do not affect the support of \mathbf{y} , irrespective of the value of δ . Since the environmental variables multiply the random inefficiency u , the environmental variables in these DGPs can be viewed as affecting the mean and variance of the one-sided inefficiency process. Hence these DGPs, are consistent with Assumption 2.2; in other words, the null hypothesis to be tested is true within the context of DGPs #2 and #4. This stands in contrast to DGPs #1 and #3, where the environmental variables affect the support of \mathbf{y} whenever $\delta > 0$. For example, in DGP #1 given in (5.3), for $\delta > 0$, $y \leq y^+$ even if $u = 0$. If $\delta = 0$, then Assumption 2.2 is satisfied and the null hypothesis to be tested is true, but as δ increases from zero, we have increasing departures from the null; hence δ controls the degree of departure from the null hypothesis.

In our experiments, we consider values $\delta \in \{0, 0.1, 0.2, \dots, 0.8, 1.0\}$. In addition, for each experiment, we consider both the FDH-based statistic defined in 3.7 as well as the DEA-based statistic given in 3.8, choosing the optimal m independently for the two different statistics.

For each value of m , we record whether the null hypothesis is rejected for nominal test sizes $\alpha \in \{0.1, 0.05, 0.01\}$ on each Monte Carlo trial. Dividing these counts by the number of Monte Carlo trials gives an estimate of the rejection rate achieved for each alternative sub-sample size. Figures 1–2 show plots of rejection rates for tests of nominal size $\alpha = 0.05$ as a function of the sub-sample size m for experiments with DGP #1 defined by (5.3). In the

figures, each panel corresponds to one of the four samples sizes we considered. In each panel, rejection rates corresponding to $\delta = 0, 0.1, \dots$ are plotted as alternating solid and dashed curves. The lowest solid curve in each panel corresponds to $\delta = 0$; moving upward, the next curve is dashed, corresponding to $\delta = 0.1$, and so on. Figure 1 shows results obtained using $\widehat{\tau}_{\text{FDH},n}(\mathcal{S}_n)$, while Figure 2 shows results obtained using $\widehat{\tau}_{\text{DEA},n}(\mathcal{S}_n)$.

The results shown in Figures 1–2 indicate that even with only $n = 50$, the tests have good power. The optimal sub-sample sizes in each panel are indicated by the intersections of the horizontal line drawn at 0.05 on the vertical axes and the bottom, solid curve corresponding to $\delta = 0$ (recall that the null is true when $\delta = 0$). The optimal sub-sample sizes in Figure 1 vary between about 50 and 75 percent of the sample size n . Comparing the panels within each of the two figures, it is evident that power increases rapidly with increasing departures from the null.

Figures 3–4 show similar results from experiments with DGP #2, where the null is true regardless of the value of δ . The differences between these figures and the previous pair are striking; here, if the optimal sub-sample size is used, the rejection rate is about five percent regardless of the value of δ . Since the null is true in DGP #2 for all values of δ , this is as expected.

Comparing Figures 1 and 3 with Figures 2 and 4 reveals few real differences; with either the FDH-based or the DEA-based statistic, size and power properties are good for an appropriate choice of the sub-sample size. The optimal sub-sample size appears to be smaller when the DEA-based statistic is used, as opposed to the FDH-based statistic.

Figures 5–6 show estimated rejection rates obtained for DGP #3, where $p = q = r = 2$ and the null is again false whenever $\delta > 0$. Comparing these results with the results in Figures 1–2 indicates that power suffers when dimensionality is increased, which is to be expected given dependence on $p + q$ and r of the convergence rate in (4.20). Nonetheless, even with moderate sample sizes n , power increases rapidly with increasing departures from the null. Also, as before, there are few differences in the estimated rejection rates across the two statistics.

Figures 7–8, show estimated rejection rates for DGP #4, where $p = q = r = 2$ and the null is true for all values of δ . In this case, some differences between the FDH-based and DEA-based statistics become apparent. In Figure 8, the curves in each panel lie close

together, indicating that the rejection rates will be close to five percent for each value of δ . But in Figure 7 where results for the FDH-based statistic are shown, we see that at the optimal sub-sample size (defined by the intersection of the horizontal line at 0.05 on the vertical axis and the lowest of the solid curves corresponding to $\delta = 0$), the test will reject the null with increasing frequency as δ increases from zero, although far less so than in Figure 5 with DGP #3. The convergence rates in the FDH estimators are slower than in the DEA estimators; the curse of dimensionality is more acute for the FDH estimators than for the DEA estimators. Hence with increasing dimensionality, the DEA estimators should be expected to out-perform their FDH counterparts for a fixed sample size. The results suggest that while there is a price to pay for increasing dimensionality, the price is lower if one can safely assume convexity of the level sets $\mathcal{P}(z)$; here, the price involves Type-I errors.

The results in Figures 1–8 illustrate rejection rates for a range of sub-sample sizes. In an application, however, the researcher must choose the sub-sample size, and furthermore, has only one set of observed data. In each of our experiments, on each Monte Carlo trial, we selected an “optimal” sub-sample size m using the method described in Section 4.4 and the beginning of this section. We counted the number of rejections obtained for each nominal size $\alpha \in \{0.1, 0.05, 0.01\}$ over the Monte Carlo trials, and then divided these counts by the number of trials. Results obtained with DGP #1, using both our statistics, are shown in Table 1.

The results in Table 1 confirm the overall impression given by Figures 1–2; more importantly, however, the results in Table 1 give an idea of how well the applied researcher with only one sample can be expected to do when optimizing m as described earlier. With sample size $n = 100$ or more, the achieved sizes of the tests (corresponding to the rejection rates when $\delta = 0$) are slightly too small, resulting in conservative tests. In addition, with the DEA-based statistic, the realized size decreases with n in Table 1. This is apparently due to the fact that in our simulations, we use the grid of sub-sample values given by \mathbb{M}_n as discussed above; consequently, the grid over which we search becomes more coarse as we increase the sample size. We used a grid of 49 sub-sample sizes to avoid excessive computational burdens in our Monte Carlo experiments, but in an application with a single dataset, a researcher could use a finer grid, or perhaps first use a coarse grid, and then refine the choice of sub-sample size by subsequently searching using a fine grid over sub-sample sizes near the

one initially selected. Nonetheless, our results indicate the tests work well; it is certainly better to have a conservative test than one where the realized size exceeds the nominal size. Moreover, the power of the tests increases quickly with increasing departures from the null.

Table 2 shows similar results obtained with DGP #3. As expected, the increased dimensionality over DGP #1 degrades the power of the tests; for $n = 50$ and $n = 100$, the power with either statistic is quite poor. However, power improves when n is increased to 200, and becomes quite good when n reaches 400. In addition, the achieved sizes of the tests are close to the nominal sizes with $n = 200$ or 400.

Tables 3–4 show similar results obtained with DGPs #2 and #4, where the null is true regardless of the value of δ . The results in these tables confirm the earlier observations made in conjunction with Figures 3–4 and 7–8; i.e., the rejection rates are low for all values of δ and all sample sizes that were considered.

While use of a rule such as the one described in Section 4.4 and at the beginning of this section is necessary for choosing the “optimal” sub-sample size in a Monte Carlo setting, the applied, empirical researcher can use graphical methods as discussed by Simar and Wilson (2009). In our experiments, on a given Monte Carlo trial, we have one sample of size n ; for each of 49 values of m , we compute the bootstrap approximation in (4.20) and obtain a critical value. Figure 9 shows results from randomly chosen Monte Carlo trials in our experiments using DGP #1 and the DEA-based statistic defined in (3.7) with $n = 400$ and nominal test size $\alpha = 0.05$. The three panels in the first column of Figure 9 correspond to particular trials where $\delta = 0$ and the null hypothesis is true; the three panels in the second column correspond to particular trials where $\delta = 0.3$ (and so the null is false). In each panel, the value of the test statistic is given by the height on the vertical axis of the horizontal solid line (the scales on the vertical axes are different in order to make the plots legible). Values of the critical values corresponding to each value of the sub-sample size m (measured on the horizontal axis) are represented by “+” symbols. The vertical dashed line in each panel indicates the sub-sample size optimized using the deterministic rule described above.

In the three Monte Carlo trials represented in the first column of Figure 9, use of the deterministic rule leads to failure to reject the null; in each case a sub-sample size yielding a critical value greater than the value of the test statistic is chosen. Recalling the results in Table 1, in the experiment with DGP #1 and the DEA-based statistic defined in (3.8)

with $n = 400$ and nominal test size $\alpha = 0.05$, the null was rejected in only one percent of all Monte Carlo trials, a sensible outcome given that the null is true in these cases. If an empirical researcher working with a single dataset produced a plot along the lines of those in the first column of Figure 9, he would have little reason to doubt that the null should not be rejected.

On the other hand, the three panels in the second column of Figure 9 show quite different results. In each of these cases, $\delta = 0.3$ and so the null is false. In each of the three Monte Carlo trials that are illustrated, the null is rejected using the deterministic rule, although just so in trials #408 and #721. However, an empirical researcher looking at a plot similar to any of the three in the second column of Figure 9 would see that the value of his test statistic is larger than almost every critical value, and hence could feel safe in rejecting the null hypothesis.

6 Empirical Example

Simar and Wilson (2007) included an empirical example based on Aly et al. (1990) using data on 6,955 U.S. Commercial Banks observed at the end of the fourth quarter, 2002. The specification in the example included three inputs ($p = 3$), four outputs ($q = 4$), and four environmental variables ($r = 4$), two of which are continuous and two of which are binary ($r_c = r_d = 2$). Simar and Wilson (2007) computed DEA estimates of efficiency for each of the 6,955 banks, and then estimated a truncated regression where these estimates appeared as the dependent variable and the four environmental variables (as well as interaction and squared terms) were treated as exogenous, right-hand side variables. Inference was made using the bootstrap methods discussed by Simar and Wilson. Specific definitions for the variables, etc. are given in Section 7 of Simar and Wilson (2007).

Using the same data, we tested the “separability” condition in Assumption 2.2 versus the alternative in Assumption 2.4 using both the FDH- and DEA-based statistics defined above in (3.7)–(3.8) that were examined in Section 5 using Monte Carlo methods. As in our Monte Carlo experiments discussed in Section 5, we searched over the grid of sub-sample sizes $m \in \mathbb{M}_n = \left\{ \left\lfloor \frac{n}{50} \right\rfloor, 2 \left\lfloor \frac{n}{50} \right\rfloor, 3 \left\lfloor \frac{n}{50} \right\rfloor, \dots, 49 \left\lfloor \frac{n}{50} \right\rfloor \right\}$, where $[a]$ denotes the integer part of a real number a ; with $n = 6,955$ our grid of sub-sample sizes includes values 139, 278, 417, \dots , 6,811. We used $k = 1$ to define the volatility measure as discussed above. For each of the

49 sub-sample sizes, we performed 2,048 bootstrap replications, using resampling without replacement.

Figure 10 illustrates the results of our two tests; the left panel corresponds to the FDH-based test, while the right panel corresponds to the DEA-based test. In either panel, estimated 95-percent critical values corresponding to each of 49 sub-sample sizes are represented by “+” symbols, and the value of the test statistic (measured on the vertical axes) is represented by the solid horizontal line. The vertical dashed lines indicate the sub-sample size chosen using the rule described above in Section 5. From the evidence shown in Figure 10, it is clear that the null hypothesis given by Assumption 2.2 should be rejected. In fact, the results are quite strong—with the FDH-based test, only one very small sub-sample size yields a critical value larger than the value of our statistic (i.e., $m = 973$), while with the DEA-based test, *no* critical values fail to reject the null.

Our tests indicate that the empirical illustration in Simar and Wilson (2007), while providing an example of how the bootstrap methods developed in Simar and Wilson (2007) can be applied in a real-world setting, gives results that are meaningless. Although we have not explicitly tested the “separability” assumption in Assumption 2.2 using the data used by Aly et al. (1990) (here, we have used more recent data and a larger sample size, but with the same variable definitions used by Aly et al.), our results certainly call into question whether the results obtained by Aly et al. are meaningful. Assuming we would similarly reject Assumption 2.2 in favor of the alternative Assumption 2.4 using the same data that Aly et al. used, our results suggest that not only is the second-stage regression estimated by Aly et al. meaningless, but the first-stage efficiency estimates are also meaningless since the DEA efficiency estimators used by Aly et al. estimate distance to the frontier of a set Ψ that is unattainable due to the violation of Assumption 2.2.

7 Conclusions

As discussed in the Introduction, numerous articles where technical efficiency is estimated using DEA or FDH estimators and then the estimates are regressed on some environmental variables in a second stage continue to appear in a variety of academic journals. Apart from the fact that few, if any, of these articles specify a coherent statistical model, and also the fact that many estimate second-stage models that cannot be justified in the context of

the model presented by Simar and Wilson (2007), as far as we know none have tested the “separability” assumption necessary for either the first-stage or the second-stage results to be sensible.

We have developed a fully non-parametric test employing sub-sampling methods in order to test the separability condition needed for two-stage estimation in production models. Our simulation results indicate that the test works quite well, with both good power and good size properties. Applying our test using data and variables from the empirical example in Simar and Wilson (2007), we easily reject separability, calling into question the results presented by Aly et al. (1990). The results of our empirical example also raise the question of whether the numerous articles cited by Simar and Wilson (2007) as examples of where two-stage efficiency estimation has been employed would be meaningful even if the second-stage estimation in those papers had been done sensibly.

It is important to note that in Simar and Wilson (2007), we did not recommend the two-stage approach for efficiency estimation. Rather, our purpose there was to rationalize what had frequently appeared in the literature (and, as noted in Section 1, continues to appear). Simar and Wilson (2007) remarked that the separability condition introduced in our earlier paper should be tested, and in this paper we have introduced tools that will enable applied researchers to test whether separability holds. If it does not hold, one should certainly not use two-stage estimation, but instead should use the conditional efficiency estimators that we have used to construct our test statistics here.

References

- Aly, H. Y., C. P. R. G. Grabowski, and N. Rangan (1990), Technical, scale, and allocative efficiencies in U.S. banking: an empirical investigation, *Review of Economics and Statistics* 72, 211–218.
- Bickel, P. J. and A. Sakov (2008), On the choice of m in the m out of n bootstrap and confidence bounds for extrema, *Statistica Sinica* 18, 967–985.
- Bădin, L., C. Daraio, and L. Simar (2010), Optimal bandwidth selection for conditional efficiency measures: A data-driven approach, *European Journal of Operational Research* 201, 633–664.
- Coelli, T., D. S. P. Rao, and G. E. Battese (1997), *An Introduction to Efficiency and Productivity Analysis*, Boston: Kluwer Academic Publishers.
- Daraio, C. and L. Simar (2005), Introducing environmental variables in nonparametric frontier models: A probabilistic approach, *Journal of Productivity Analysis* 24, 93–121.
- (2007a), *Advanced Robust and Nonparametric Methods in Efficiency Analysis*, New York: Springer Science+Business Media, LLC.
- (2007b), Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach, *Journal of Productivity Analysis* 28, 13–32.
- Deprins, D., L. Simar, and H. Tulkens (1984), Measuring labor inefficiency in post offices, in M. M. P. Pestieau and H. Tulkens, eds., *The Performance of Public Enterprises: Concepts and Measurements*, Amsterdam: North-Holland, pp. 243–267.
- Färe, R. (1988), *Fundamentals of Production Theory*, Berlin: Springer-Verlag.
- Farrell, M. J. (1957), The measurement of productive efficiency, *Journal of the Royal Statistical Society A* 120, 253–281.
- Jeong, S. O., B. U. Park, and L. Simar (2010), Nonparametric conditional efficiency measures: asymptotic properties, *Annals of Operational Research* 173, 105–122.
- Jeong, S. O. and L. Simar (2006), Linearly interpolated FDH efficiency score for nonconvex frontiers, *Journal of Multivariate Analysis* 97, 2141–2161.
- Kneip, A., B. Park, and L. Simar (1998), A note on the convergence of nonparametric DEA efficiency measures, *Econometric Theory* 14, 783–793.
- Kneip, A., L. Simar, and P. W. Wilson (2008), Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models, *Econometric Theory* 24, 1663–1697.
- (2011), A computationally efficient, consistent bootstrap for inference with non-parametric DEA estimators, *Computational Economics* Forthcoming.
- Marsaglia, G. (1972), Choosing a point from the surface of a sphere, *Annals of Mathematical Statistics* 43, 645–646.
- Muller, M. E. (1959), A note on a method for generating points uniformly on n -dimensional spheres, *Communications of the Association for Computing Machinery* 2, 19–20.

- Park, B. U., S.-O. Jeong, and L. Simar (2010), Asymptotic distribution of conical-hull estimators of directional edges, *Annals of Statistics* 38, 1320–1340.
- Park, B. U., L. Simar, and C. Weiner (2000), FDH efficiency scores from a stochastic point of view, *Econometric Theory* 16, 855–877.
- Park, B. U., L. Simar, and V. Zelenyuk (2008), Local likelihood estimation of truncated regression and its partial derivative: Theory and application, *Journal of Econometrics* 146, 185–2008.
- Politis, D. N., J. P. Romano, and M. Wolf (2001), On the asymptotic theory of subsampling, *Statistica Sinica* 11, 1105–1124.
- 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 P. W. Wilson (1998), Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models, *Management Science* 44, 49–61.
- (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.
- (2001), Testing restrictions in nonparametric efficiency models, *Communications in Statistics* 30, 159–184.
- (2007), Estimation and inference in two-stage, semi-parametric models of productive efficiency, *Journal of Econometrics* 136, 31–64.
- (2008), Statistical inference in nonparametric frontier models: Recent developments and perspectives, in H. O. Fried, C. A. K. Lovell, and S. S. Schmidt, eds., *The Measurement of Productive Efficiency*, chapter 4, Oxford: Oxford University Press, 2nd edition, pp. 421–521.
- (2009), Inference by subsampling in nonparametric frontier models. Discussion paper #0933, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Table 1: Rejection Rates for Separability Test (DGP #1, $p = q = r = 1$)

FDH												
δ	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
	$(1 - \alpha)$			$(1 - \alpha)$			$(1 - \alpha)$			$(1 - \alpha)$		
	.90	.95	.99	.90	.95	.99	.90	.95	.99	.90	.95	.99
0.0	0.096	0.022	0.007	0.060	0.021	0.009	0.058	0.018	0.007	0.037	0.021	0.009
0.1	0.070	0.025	0.010	0.060	0.032	0.018	0.104	0.045	0.029	0.231	0.133	0.061
0.2	0.065	0.031	0.013	0.139	0.071	0.053	0.378	0.219	0.127	0.756	0.540	0.353
0.3	0.091	0.042	0.034	0.293	0.132	0.101	0.736	0.498	0.268	0.992	0.940	0.733
0.4	0.130	0.055	0.050	0.467	0.244	0.146	0.931	0.741	0.471	1.000	0.999	0.948
0.5	0.162	0.072	0.064	0.606	0.363	0.175	0.992	0.931	0.647	1.000	1.000	1.000
0.6	0.228	0.101	0.085	0.782	0.485	0.248	1.000	0.992	0.818	1.000	1.000	1.000
0.7	0.299	0.136	0.098	0.872	0.604	0.322	1.000	0.997	0.936	1.000	1.000	1.000
0.8	0.352	0.156	0.103	0.944	0.780	0.385	1.000	1.000	0.986	1.000	1.000	1.000
1.0	0.452	0.211	0.126	0.998	0.927	0.521	1.000	1.000	0.999	1.000	1.000	1.000
DEA												
0.0	0.146	0.091	0.063	0.064	0.035	0.023	0.040	0.020	0.016	0.022	0.010	0.007
0.1	0.113	0.069	0.045	0.086	0.039	0.026	0.137	0.074	0.040	0.298	0.189	0.087
0.2	0.097	0.054	0.029	0.185	0.101	0.065	0.499	0.296	0.188	0.858	0.684	0.468
0.3	0.137	0.062	0.042	0.383	0.219	0.115	0.834	0.606	0.392	0.996	0.975	0.873
0.4	0.211	0.096	0.049	0.596	0.382	0.208	0.975	0.884	0.641	1.000	1.000	0.990
0.5	0.307	0.159	0.087	0.747	0.518	0.303	0.998	0.987	0.830	1.000	1.000	1.000
0.6	0.360	0.184	0.099	0.904	0.723	0.385	1.000	1.000	0.956	1.000	1.000	1.000
0.7	0.458	0.253	0.150	0.967	0.813	0.480	1.000	1.000	0.998	1.000	1.000	1.000
0.8	0.538	0.311	0.137	0.986	0.926	0.648	1.000	1.000	1.000	1.000	1.000	1.000
1.0	0.697	0.424	0.207	1.000	0.991	0.799	1.000	1.000	1.000	1.000	1.000	1.000

Table 2: Rejection Rates for Separability Test (DGP #3, $p = q = r = 2$)

FDH												
δ	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
	$(1 - \alpha)$			$(1 - \alpha)$			$(1 - \alpha)$			$(1 - \alpha)$		
	.90	.95	.99	.90	.95	.99	.90	.95	.99	.90	.95	.99
0.0	0.222	0.043	0.009	0.138	0.060	0.010	0.091	0.049	0.013	0.090	0.048	0.027
0.1	0.166	0.067	0.016	0.083	0.053	0.021	0.122	0.078	0.037	0.213	0.107	0.075
0.2	0.111	0.062	0.010	0.075	0.068	0.025	0.127	0.088	0.059	0.354	0.159	0.120
0.3	0.093	0.051	0.021	0.091	0.053	0.032	0.174	0.092	0.078	0.705	0.330	0.183
0.4	0.093	0.060	0.020	0.110	0.079	0.046	0.303	0.125	0.128	0.957	0.681	0.278
0.5	0.088	0.059	0.025	0.136	0.063	0.054	0.467	0.190	0.134	0.996	0.893	0.420
0.6	0.086	0.073	0.021	0.150	0.062	0.084	0.634	0.245	0.158	1.000	0.972	0.534
0.7	0.082	0.061	0.030	0.177	0.077	0.082	0.715	0.312	0.172	1.000	0.994	0.667
0.8	0.082	0.061	0.029	0.180	0.067	0.083	0.787	0.393	0.169	1.000	0.998	0.747
1.0	0.097	0.066	0.035	0.232	0.089	0.111	0.874	0.504	0.230	1.000	1.000	0.876
DEA												
0.0	0.070	0.034	0.022	0.071	0.029	0.016	0.092	0.046	0.012	0.094	0.045	0.013
0.1	0.085	0.042	0.027	0.082	0.068	0.028	0.104	0.056	0.022	0.153	0.070	0.032
0.2	0.090	0.048	0.027	0.101	0.068	0.024	0.137	0.086	0.048	0.259	0.127	0.106
0.3	0.086	0.047	0.028	0.110	0.075	0.045	0.202	0.110	0.064	0.654	0.289	0.191
0.4	0.093	0.058	0.038	0.151	0.069	0.062	0.298	0.124	0.107	0.952	0.633	0.274
0.5	0.098	0.066	0.044	0.163	0.076	0.060	0.474	0.160	0.122	0.998	0.891	0.376
0.6	0.105	0.069	0.056	0.175	0.081	0.075	0.610	0.235	0.137	0.999	0.963	0.515
0.7	0.112	0.055	0.053	0.225	0.088	0.081	0.685	0.338	0.164	1.000	0.995	0.621
0.8	0.110	0.050	0.042	0.212	0.076	0.091	0.743	0.343	0.175	1.000	0.996	0.675
1.0	0.114	0.049	0.041	0.214	0.078	0.075	0.721	0.348	0.195	1.000	0.994	0.707

Table 3: Rejection Rates for Separability Test (DGP #2, $p = q = r = 1$)

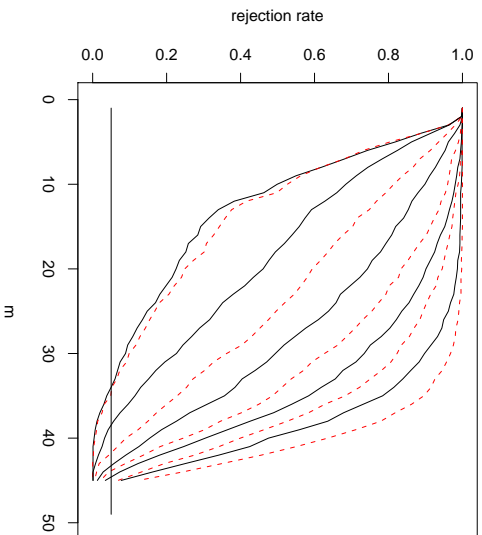
FDH												
δ	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
	$(1 - \alpha)$			$(1 - \alpha)$			$(1 - \alpha)$			$(1 - \alpha)$		
	.90	.95	.99	.90	.95	.99	.90	.95	.99	.90	.95	.99
0.0	0.092	0.016	0.007	0.063	0.021	0.009	0.062	0.021	0.013	0.041	0.014	0.005
0.1	0.111	0.020	0.006	0.075	0.022	0.009	0.065	0.024	0.011	0.073	0.031	0.005
0.2	0.080	0.021	0.004	0.061	0.014	0.011	0.052	0.013	0.010	0.055	0.021	0.003
0.3	0.078	0.014	0.004	0.084	0.026	0.014	0.056	0.015	0.006	0.048	0.020	0.010
0.4	0.085	0.016	0.008	0.062	0.029	0.010	0.044	0.022	0.008	0.051	0.021	0.009
0.5	0.075	0.019	0.006	0.070	0.021	0.011	0.059	0.023	0.009	0.061	0.022	0.010
0.6	0.091	0.030	0.008	0.067	0.023	0.007	0.056	0.021	0.008	0.043	0.014	0.005
0.7	0.085	0.017	0.007	0.062	0.019	0.014	0.041	0.019	0.010	0.044	0.018	0.005
0.8	0.077	0.021	0.007	0.073	0.020	0.008	0.061	0.022	0.010	0.046	0.014	0.005
1.0	0.097	0.034	0.011	0.066	0.023	0.009	0.049	0.021	0.009	0.047	0.018	0.009
DEA												
0.0	0.138	0.089	0.067	0.073	0.050	0.039	0.046	0.032	0.020	0.023	0.012	0.005
0.1	0.140	0.098	0.062	0.074	0.037	0.029	0.049	0.027	0.020	0.029	0.017	0.008
0.2	0.156	0.099	0.067	0.068	0.045	0.028	0.041	0.027	0.017	0.027	0.020	0.011
0.3	0.134	0.085	0.061	0.071	0.042	0.029	0.042	0.025	0.015	0.030	0.012	0.008
0.4	0.110	0.073	0.056	0.081	0.043	0.033	0.040	0.021	0.015	0.030	0.016	0.008
0.5	0.131	0.075	0.058	0.073	0.039	0.035	0.040	0.028	0.019	0.024	0.012	0.007
0.6	0.119	0.079	0.058	0.071	0.042	0.029	0.046	0.030	0.016	0.026	0.017	0.010
0.7	0.127	0.073	0.062	0.073	0.043	0.030	0.055	0.029	0.011	0.018	0.011	0.006
0.8	0.120	0.082	0.062	0.070	0.050	0.029	0.043	0.024	0.021	0.033	0.018	0.006
1.0	0.130	0.086	0.062	0.081	0.046	0.035	0.040	0.023	0.017	0.019	0.009	0.007

Table 4: Rejection Rates for Separability Test (DGP #4, $p = q = r = 2$)

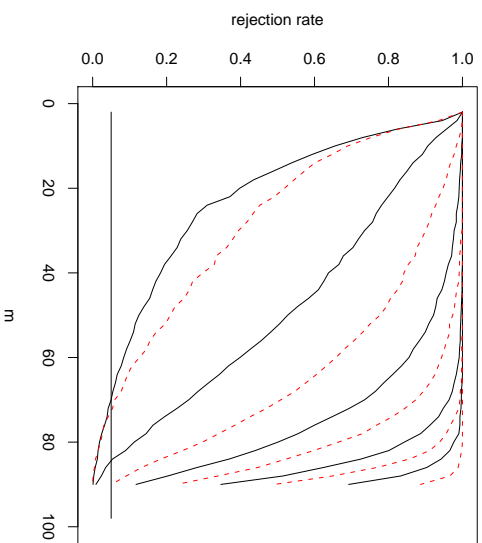
FDH												
δ	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
	$(1 - \alpha)$			$(1 - \alpha)$			$(1 - \alpha)$			$(1 - \alpha)$		
	.90	.95	.99	.90	.95	.99	.90	.95	.99	.90	.95	.99
0.0	0.221	0.045	0.011	0.133	0.057	0.012	0.090	0.051	0.013	0.085	0.052	0.024
0.1	0.268	0.018	0.009	0.170	0.060	0.008	0.104	0.071	0.014	0.115	0.062	0.022
0.2	0.297	0.010	0.005	0.198	0.033	0.007	0.121	0.036	0.011	0.105	0.049	0.009
0.3	0.251	0.007	0.004	0.209	0.021	0.007	0.173	0.054	0.008	0.114	0.046	0.010
0.4	0.184	0.004	0.003	0.212	0.010	0.006	0.203	0.027	0.004	0.131	0.059	0.018
0.5	0.113	0.001	0.001	0.208	0.007	0.004	0.215	0.016	0.006	0.154	0.034	0.013
0.6	0.071	0.001	0.001	0.170	0.001	0.001	0.210	0.011	0.002	0.201	0.041	0.011
0.7	0.033	0.000	0.000	0.112	0.001	0.001	0.198	0.009	0.003	0.190	0.020	0.003
0.8	0.031	0.000	0.000	0.070	0.001	0.001	0.127	0.001	0.001	0.212	0.010	0.006
1.0	0.012	0.000	0.000	0.020	0.000	0.000	0.059	0.000	0.000	0.128	0.002	0.001
DEA												
0.0	0.065	0.032	0.023	0.066	0.040	0.023	0.081	0.040	0.011	0.103	0.044	0.016
0.1	0.082	0.042	0.020	0.068	0.047	0.025	0.075	0.042	0.015	0.073	0.062	0.016
0.2	0.102	0.062	0.027	0.073	0.044	0.021	0.071	0.039	0.010	0.081	0.033	0.010
0.3	0.091	0.070	0.042	0.068	0.044	0.021	0.071	0.029	0.012	0.073	0.039	0.011
0.4	0.143	0.107	0.036	0.080	0.040	0.016	0.065	0.033	0.017	0.071	0.032	0.008
0.5	0.246	0.139	0.052	0.069	0.053	0.023	0.071	0.031	0.013	0.062	0.032	0.006
0.6	0.335	0.211	0.077	0.083	0.055	0.028	0.062	0.031	0.018	0.073	0.027	0.007
0.7	0.478	0.294	0.089	0.110	0.090	0.039	0.065	0.038	0.021	0.065	0.032	0.012
0.8	0.564	0.353	0.108	0.170	0.120	0.043	0.071	0.053	0.034	0.063	0.037	0.011
1.0	0.639	0.431	0.121	0.359	0.220	0.078	0.119	0.064	0.031	0.058	0.044	0.022

Figure 1: Rejection Rates for Separability Test (DGP #1, $\hat{\tau}_{\text{FDH},n}(\mathcal{S}_n)$, $p = q = r = 1$, resampling without replacement, $\alpha = 0.05$)

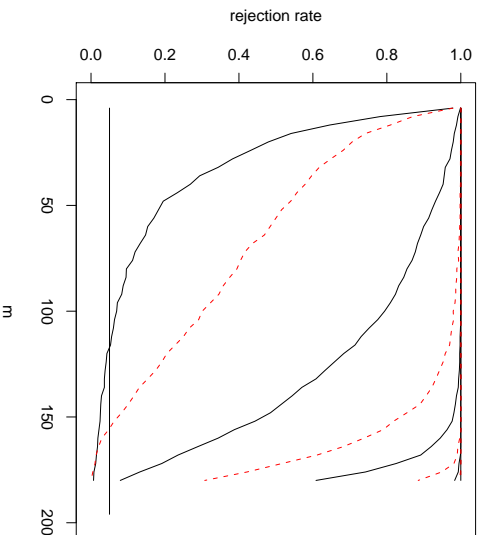
$n = 50$



$n = 100$



$n = 200$



$n = 400$

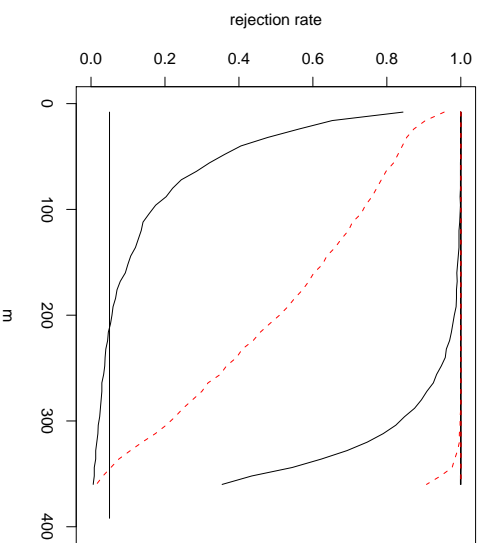
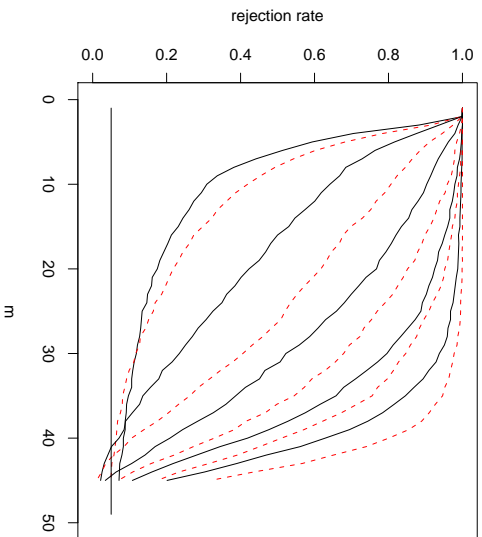
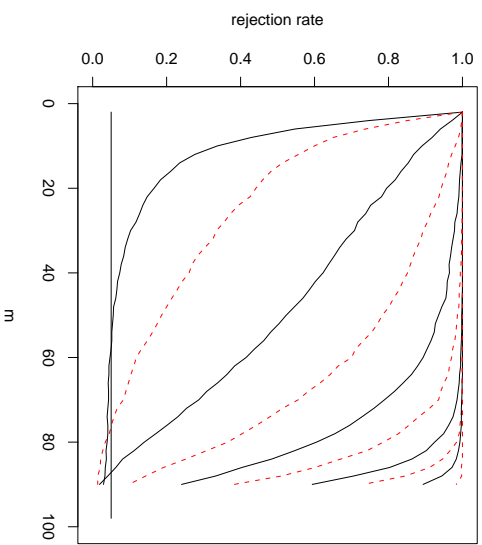


Figure 2: Rejection Rates for Separability Test (DGP #1, $\hat{\tau}_{DEA,n}(\mathcal{S}_n)$, $p = q = r = 1$, resampling without replacement, $\alpha = 0.05$)

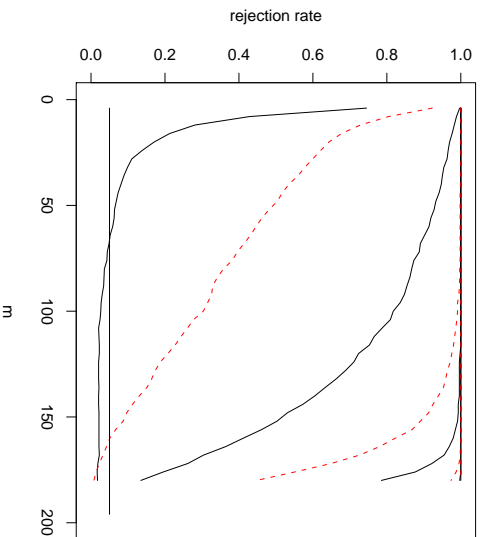
$n = 50$



$n = 100$



$n = 200$



$n = 400$

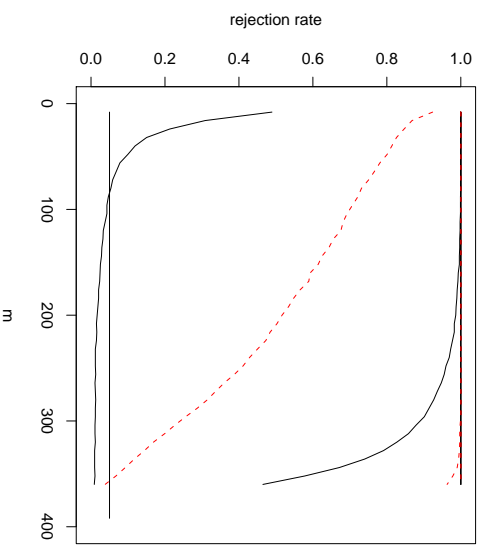
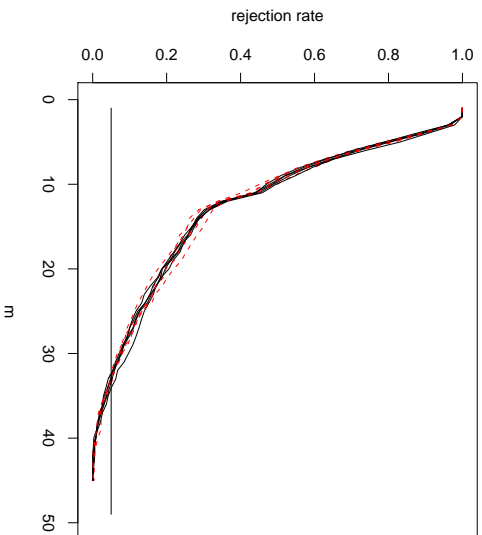
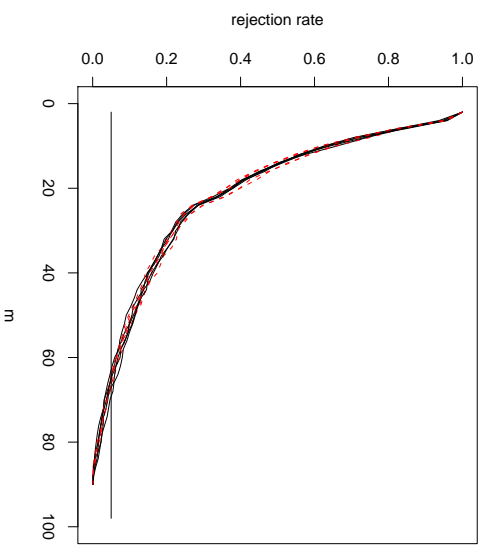


Figure 3: Rejection Rates for Separability Test (DGP #2, $\hat{\tau}_{\text{FDH},n}(\mathcal{S}_n)$, $p = q = r = 1$, resampling without replacement, $\alpha = 0.05$)

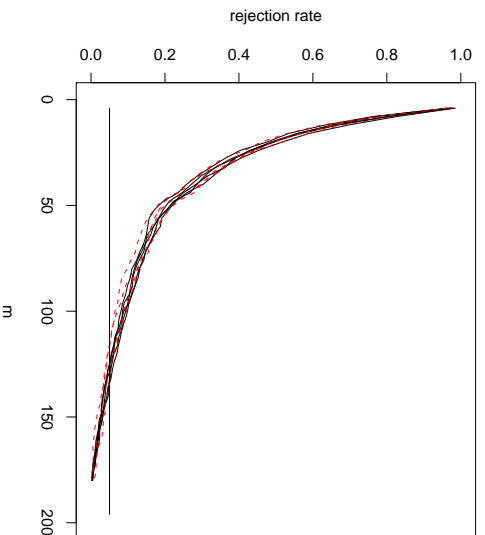
$n = 50$



$n = 100$



$n = 200$



$n = 400$

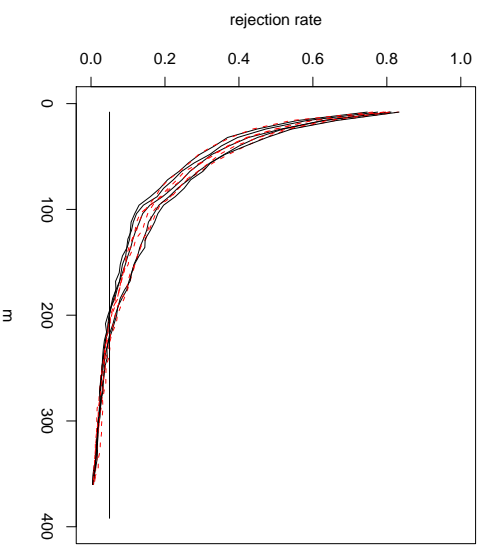
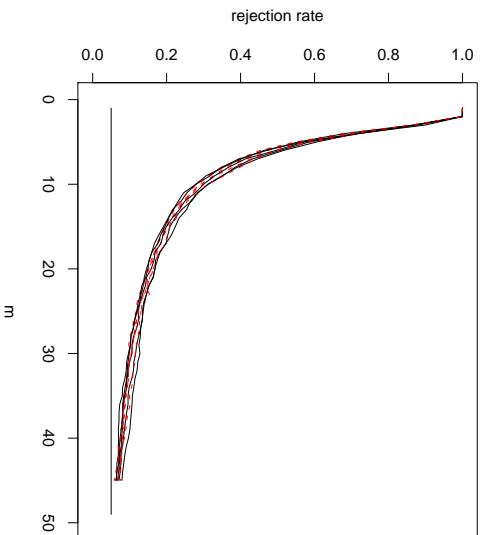
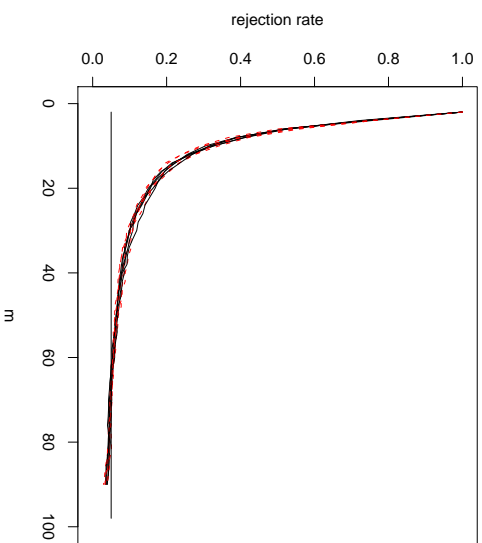


Figure 4: Rejection Rates for Separability Test (DGP #2, $\hat{\tau}_{DEA,n}(\mathcal{S}_n)$, $p = q = r = 1$, resampling without replacement, $\alpha = 0.05$)

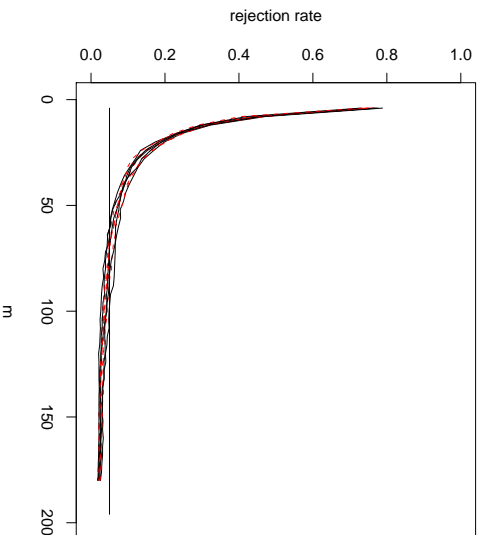
$n = 50$



$n = 100$



$n = 200$



$n = 400$

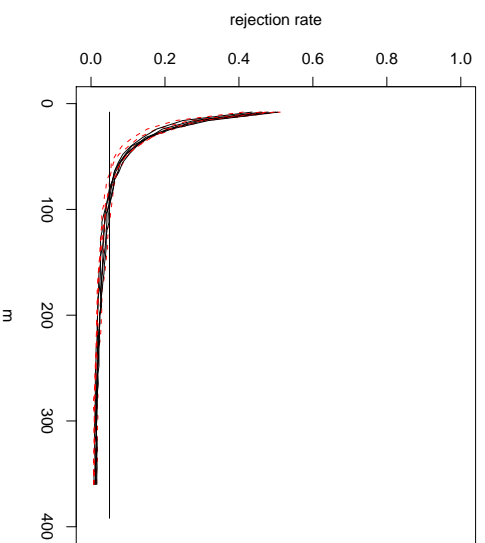
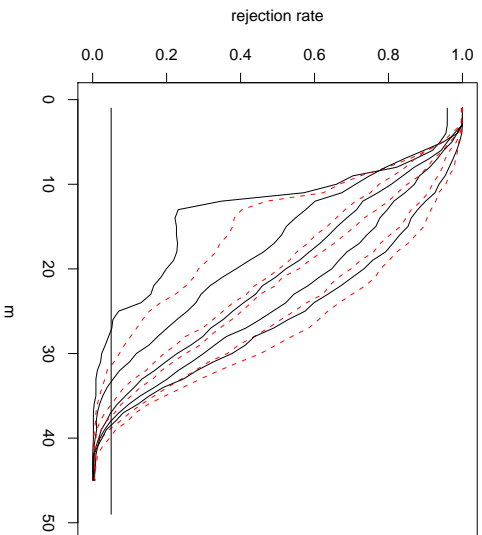
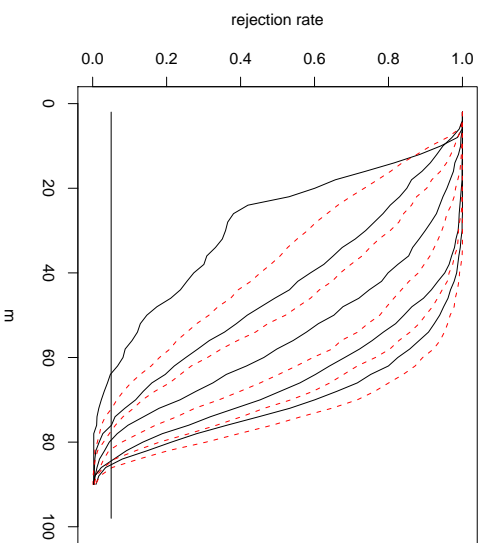


Figure 5: Rejection Rates for Separability Test (DGP #3, $\hat{\tau}_{FDH,n}(\mathcal{S}_n)$, $p = q = r = 2$, resampling without replacement, $\alpha = 0.05$)

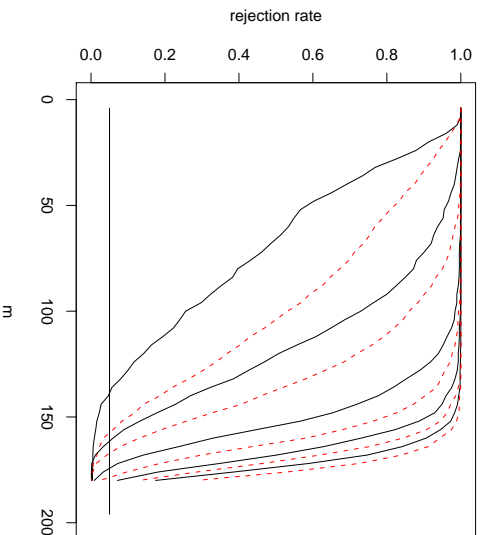
$n = 50$



$n = 100$



$n = 200$



$n = 400$

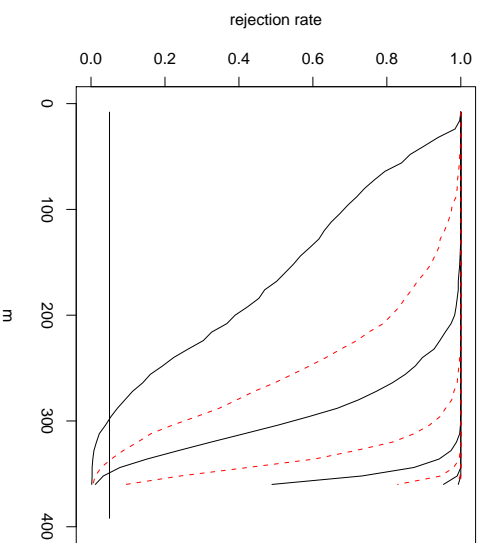
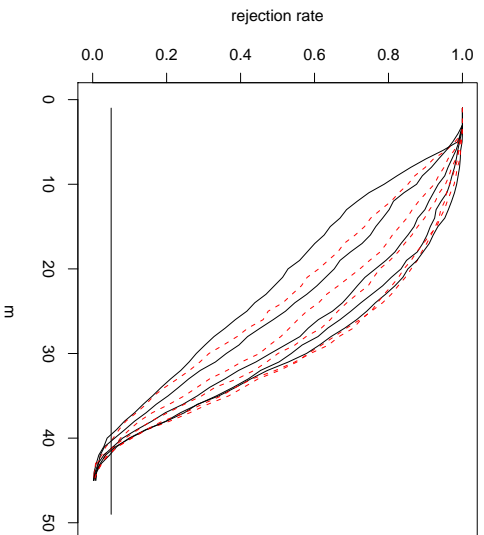
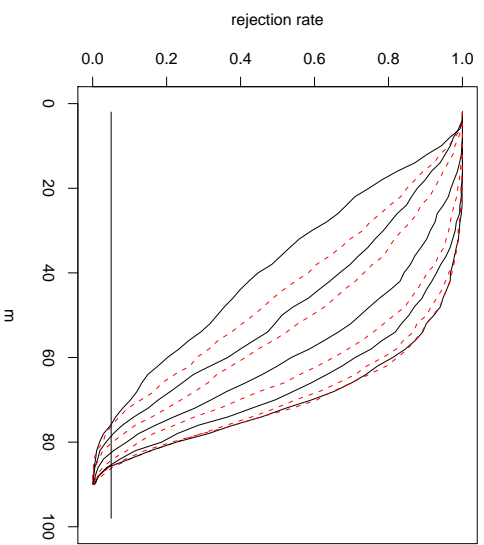


Figure 6: Rejection Rates for Separability Test (DGP #3, $\hat{\tau}_{\text{DEA},n}(\mathcal{S}_n)$, $p = q = r = 2$, resampling without replacement, $\alpha = 0.05$)

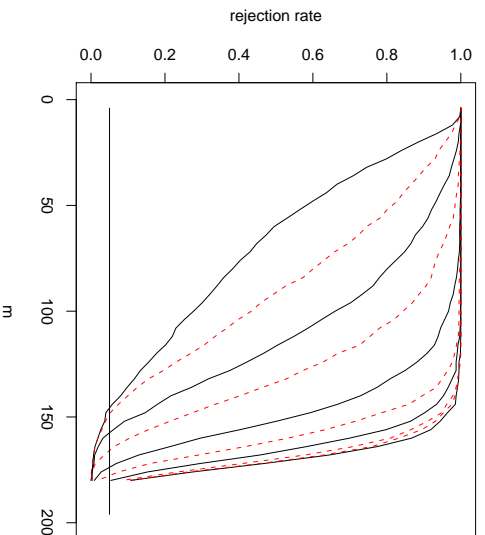
$n = 50$



$n = 100$



$n = 200$



$n = 400$

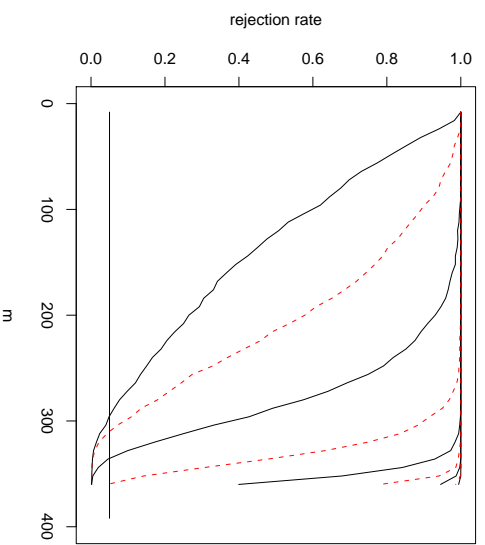


Figure 7: Rejection Rates for Separability Test (DGP #4, $\hat{\tau}_{\text{FDH},n}(\mathcal{S}_n)$, $p = q = r = 2$, resampling without replacement, $\alpha = 0.05$)

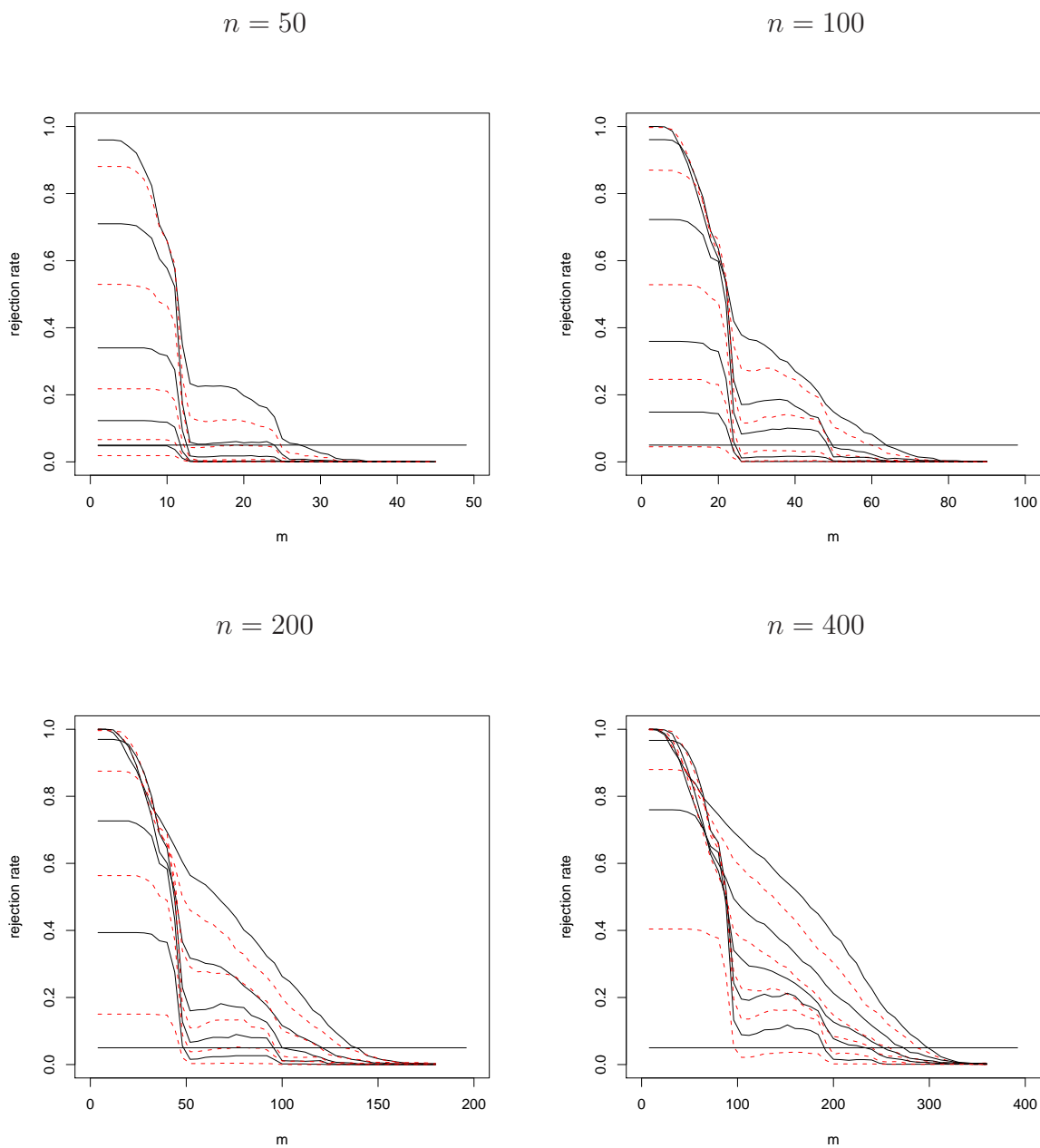


Figure 8: Rejection Rates for Separability Test (DGP #4, $\hat{\tau}_{\text{DEA},n}(\mathcal{S}_n)$, $p = q = r = 2$, resampling without replacement, $\alpha = 0.05$)

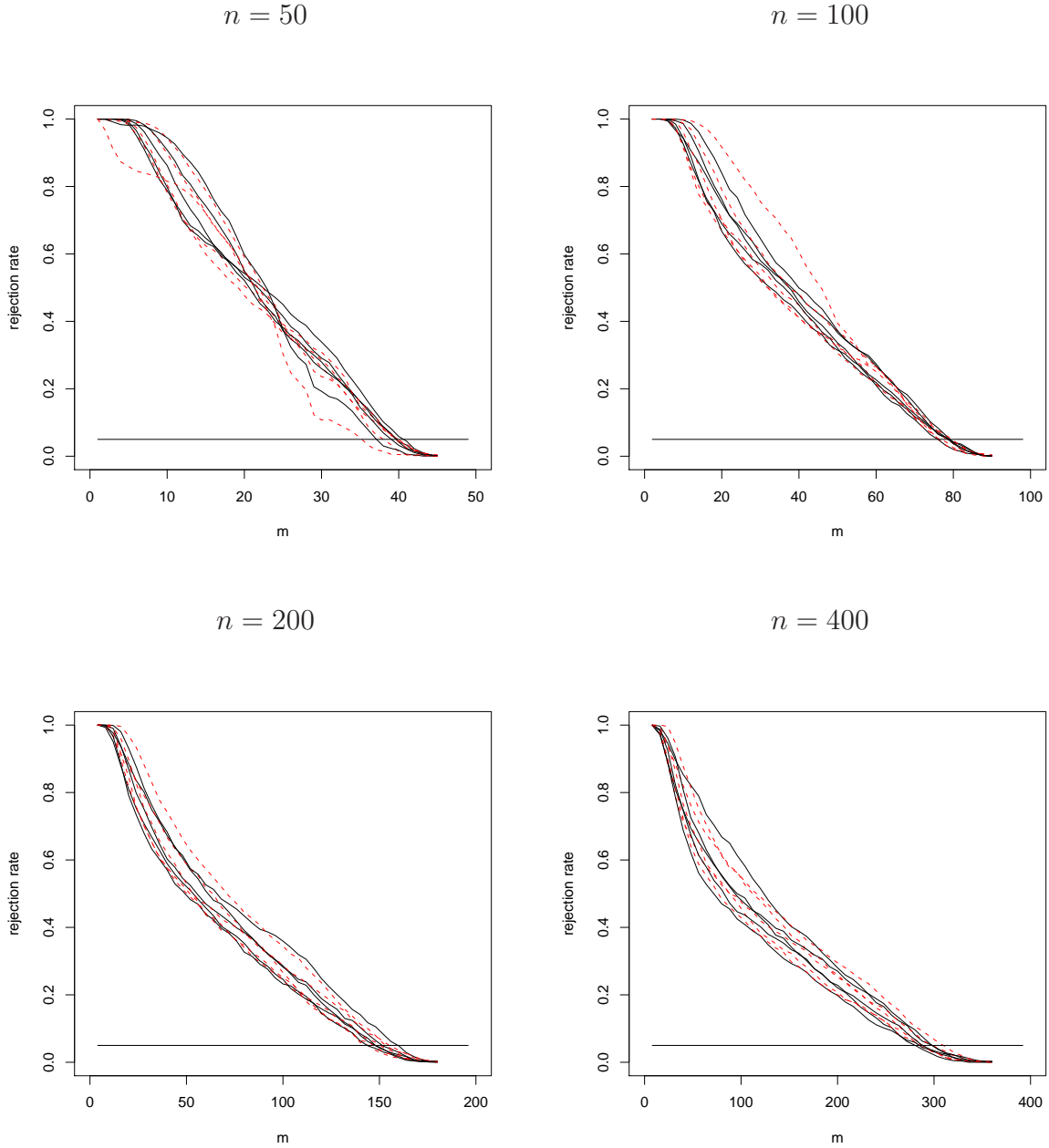


Figure 9: Estimated Critical Values versus Sub-Sample Size m for Separability Tests (DGP #1, $\hat{\tau}_{DEA,n}(\mathcal{S}_n)$, $p = q = r = 1$, resampling without replacement, $\alpha = 0.05$)

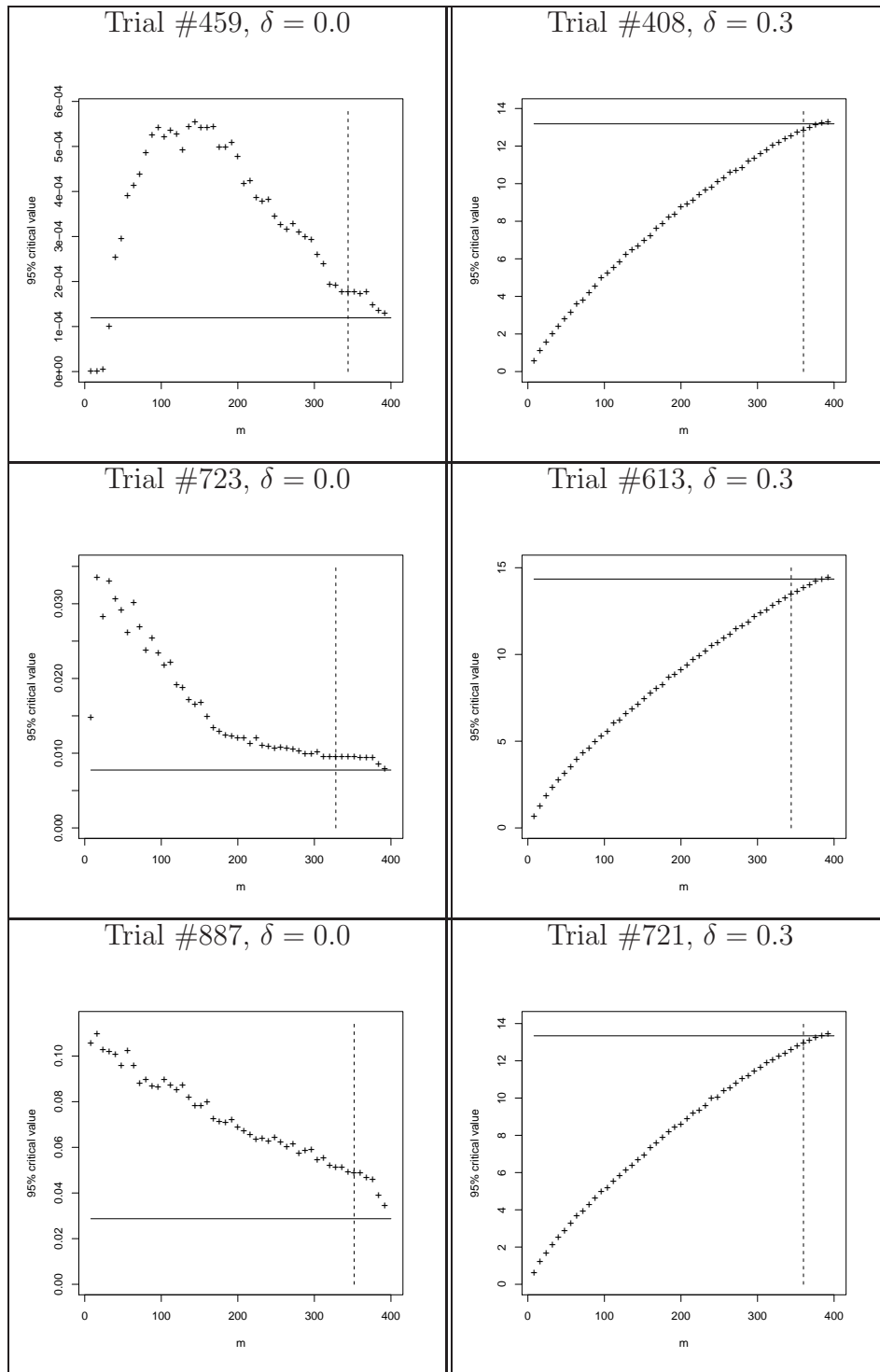


Figure 10: Tests of Separability for Banking Data used in Example by Simar and Wilson (2007), $n = 6,955$

$$\hat{\tau}_{\text{FDH},n}(\mathcal{S}_n)$$

$$\hat{\tau}_{\text{DEA},n}(\mathcal{S}_n)$$

