

MODERN TOOLS FOR EVALUATING THE PERFORMANCE OF HEALTH- CARE PROVIDERS

Léopold Simar, Paul Wilson

LIDAM Discussion Paper ISBA
2022 / 06

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

<https://uclouvain.be/en/research-institutes/lidam/isba/publication.html>

Modern Tools for Evaluating the Performance of Health-Care Providers

LÉOPOLD SIMAR

PAUL W. WILSON*

February 2022

Abstract

The hospital industry in many countries is characterized by right-skewed distributions of hospitals' sizes and varied ownership types, raising numerous questions about the performance of hospitals of different sizes and ownership types. In an era of aging populations and increasing health-care costs, evaluating and understanding the consumption of resources to produce health-care outcomes is increasingly important for policy discussions. This paper discusses recent developments in the statistical and econometric literature on DEA and FDH estimators that can be used to examine hospitals' technical efficiency and productivity. Use of these new results and methods is illustrated by revisiting the Burgess and Wilson hospital studies of the 1990s to estimate and make inference about the technical efficiency of U.S. hospitals, to make inference about returns to scale and other model features, and to test for differences among U.S. hospitals across ownership types and size groups in the context of a rigorous, statistical paradigm that was unavailable to researchers until recently.

*Simar: Institut de Statistique, Biostatistique, et Sciences Actuarielles, Université Catholique de Louvain-la-Neuve, Louvain-la-Neuve, Belgium; ORCID ID: 0000-0003-0791-8490; email: leopold.simar@uclouvain.be.

Wilson: Department of Economics and School of Computing, Clemson University, Clemson, South Carolina, USA; ORCID ID 0000-0002-9865-033X; email: pww@clemson.edu.

We regret the loss of James F. Burgess, Jr. (Jim, or “Chet” to his grad-school friends) who died at age 59 in 2017. He was a friend and a scholar. *Requiescat in Pace.*

We are grateful for (i) the Palmetto cluster operated by the Cyber Infrastructure Technology Integration group at Clemson University, and (ii) high-performance computing equipment purchased with funds from U.S. National Science Foundation MRI award no. 1725573. Any remaining errors are solely our responsibility.

1 Introduction

Data envelopment analysis (DEA) and free-disposal hull (FDH) estimators are widely used to examine the performance of health-care providers. A Google Scholar search on 4 February 2022 using the keywords “health,” “care” and “efficiency,” and either “DEA” or “FDH” yielded approximately 52,300 results. Limiting the search to items dated 2017 or later resulted in approximately 16,100 results. The studies revealed by these searches examine provision of care by physicians, hospitals and other organizations, and make comparisons across regions, states, countries, or other categories.

Both DEA and FDH estimators were used in many studies before their statistical properties, and perhaps their limitations, were known. As discussed below, papers developing the properties of the estimators began to appear in the late 1990s. By now, a large number of results (that are discussed below) are available permitting inference about efficiency of individual units, comparisons of mean efficiencies across groups or time, and other model features. Sadly, the available statistical tools are often ignored, even in recent papers.

This chapter provides some discussion of the available statistical results that can be used in conjunction with DEA and FDH estimators to allow rigorous statistical inference. We revisit Burgess and Wilson (1995, 1996 and 1998) and use their original data to re-analyze U.S. hospitals, showing what is possible using modern methods and thereby providing a useful template or road map for applied researchers.

The series of studies by Burgess and Wilson examine various aspects of the performance of U.S. hospitals operating in the 1980s using non-parametric, DEA efficiency estimators. By today’s standards, the analysis in these papers appears primitive, consisting of little more than descriptive statistics regarding the efficiency estimates and some ad hoc regressions of efficiency estimates on some environmental variables. The studies are typical of efficiency studies from the 1990s and earlier. Unfortunately, they are also typical of many efficiency studies that continue to appear in the literature. In the mid-1990s, no statistical results were available to enable inference about efficiency or productivity estimated using DEA estimators. Kneip et al. (1998) established consistency and the rate of convergence of the input-oriented variable-returns-to-scale (VRS) DEA estimator, but the other results needed for inference arrived intermittently over the next 20-plus years.

Each of the Burgess and Wilson studies use input- and output-oriented, VRS-DEA estimators to estimate efficiency of hospitals operating in the U.S. during 1985–1988. Each study considers four different ownership types. Each study estimates efficiency both in the input and the output directions. Each presents sample means of efficiency estimates and corresponding standard deviations. Burgess and Wilson (1995) estimate efficiency in each of the four years 1985 through 1988, and regress the efficiency estimates on a series of environmental variables in a panel-data framework. Burgess and Wilson (1996) focus just on 1988, and report mean efficiency estimates for each of four ownership types and discuss scale efficiency. Burgess and Wilson (1998) estimate changes in productivity and report estimated confidence intervals for geometric means of estimated Malmquist indices, but the inference is now known to be invalid. None of the studies describe a coherent statistical model from which data might have been drawn, and all three make implicit assumptions that are rejected by the data as shown below. Much has been learned in the years since the work in the Burgess and Wilson studies was done in the early-to-mid 1990s.

In the next section, we describe a non-parametric model of production and discuss various features of the model that are important for deciding how to estimate efficiency. Section 3 summarizes recent Central Limit Theorem (CLT) results for means of efficiency scores. In Section 4, we discuss a strategy for testing hypotheses about model structure that inform the choice between different estimators that might be used to analyze performance. The tests discussed in Section 4 rely on the CLT results discussed in Section 3. Section 5 provides a description of the data used by Burgess and Wilson (1995, 1996 and 1998). We use the same data, but we avoid omitting observations that were deleted unnecessarily in the Burgess and Wilson studies. Section 6 discusses results obtained by applying the tools discussed in Sections 2–4 to the data described in Section 5. Summary and Conclusions are given in the final section.

2 A Non-parametric Model of Production

2.1 Basic Framework

To establish notation, let $X \in \mathbb{R}_+^p$ and $Y \in \mathbb{R}_+^q$ denote random vectors of input and output quantities, and similarly let $x \in \mathbb{R}_+^p$ and $y \in \mathbb{R}_+^q$ denote corresponding fixed, nonstochastic

vectors of input and output quantities. In addition, let $Z \in \mathbb{R}^r$ denote a random vector of “environmental” variables that may affect how hospitals convert inputs X into outputs Y , and let $z \in \mathbb{R}^r$ denote a fixed, nonstochastic vector of values of the environmental variables in Z . Along the lines of Daraio et al. (2018), let $f_{XYZ}(x, y, z)$ denote the joint density of (X, Y, Z) with support $\mathcal{P} \subset \mathbb{R}_+^p \times \mathbb{R}_+^q \times \mathbb{R}^r$. This joint density can be decomposed into the joint density of X and Y conditional on Z times the marginal density of Z , i.e.,

$$f_{XYZ}(x, y, z) = f_{XY|Z}(x, y | z) f_Z(z). \quad (2.1)$$

Denote the conditional support of $f_{XY|Z}$ by Ψ^z , and denote the support of $f_Z(z)$ by $\mathcal{Z}$. Then Ψ^z is the set of feasible combinations of inputs and outputs for a firm facing the environmental conditions $Z = z$; i.e.,

$$\Psi^z = \{(x, y) \mid x \text{ can produce } y \text{ when } Z = z\}. \quad (2.2)$$

As discussed by Simar and Wilson (2007, 2011b) and Daraio et al. (2018), the environmental variables in Z can affect the production process either (i) only through Ψ^z , the support of (X, Y) , or (ii) only through the density $f_{XY|Z}(x, y | z)$, thereby affecting the probability for a firm to be near its optimal boundary, or (iii) through both Ψ^z and $f_{XY|Z}(x, y | z)$. Define

$$\Psi = \bigcup_{z \in \mathcal{Z}} \Psi^z. \quad (2.3)$$

Necessarily, $\Psi^z \subseteq \Psi \ \forall \ z \in \mathcal{Z}$ by construction, and clearly $\Psi \subset \mathbb{R}_+^{p+q}$. But whether Ψ is a feature of the underlying data-generating process (DGP), or whether whether Ψ is useful for benchmarking the performance of a firm producing output levels y from input levels x while facing levels z of the environmental variables, depends on whether the “separability” condition first described by Simar and Wilson (2007) is satisfied. In terms of the notation introduced here, this condition can be stated explicitly as follows.

Assumption 2.1. (*Separability Condition*): $\Psi^z = \Psi$ for all $z \in \mathcal{Z}$.

Obviously, this condition requires that Z affect production *only* through the conditional density $f_{XY|Z}(x, y | z)$ without affecting its support Ψ^z .

Clearly, if the separability condition given by Assumption 2.1 holds, then the joint support of (X, Y, Z) can be factorized as

$$\mathcal{P} = \Psi \times \mathcal{Z}, \quad (2.4)$$

and Ψ defined in (2.3) can be interpreted as the unconditional attainable set

$$\Psi = \{(x, y) \mid x \text{ can produce } y\}. \quad (2.5)$$

But, the interpretation of Ψ in (2.5) is valid if and only if Assumption 2.1 holds. Assumption 2.1 is very strong and restrictive, and is likely to *not* hold in many situations. Under Assumption 2.1, the environmental factors influence neither the *shape* nor the *level* of the boundary of the attainable set, and the potential effect of Z on the production process is only through the distribution of the inefficiencies. If the separability condition holds, it is meaningful to measure the efficiency of a particular production plan (x, y) by its distance to the boundary of Ψ .

Alternatively, if the separability condition in Assumption 2.1 *does not* hold, then Z may influence the level and the shape of the boundary of the attainable sets Ψ^z . In this case, the environmental variables in Z may also influence the conditional density $f_{XY|Z}(x, y \mid z)$, but the more important point is that Z affects the sets Ψ^z . This situation is characterized explicitly by the following assumption.

Assumption 2.2. (*Non Separability Assumption*): $\Psi^z \neq \Psi$ for some $z \in \mathcal{Z}$.

It is obvious that Assumptions 2.1 and 2.2 are mutually exclusive, i.e., only one can hold in a given scenario. Simar and Wilson (2011b) illustrate the difference between situations where separability holds (i.e., Assumption 2.1 holds) versus situations where separability does *not* hold (i.e., Assumption 2.2 holds and Assumption 2.1 does not hold) by considering two DGPs given by

$$Y^* = g(X)e^{-(Z-2)^2U} \quad (2.6)$$

and

$$Y^{**} = g(X)e^{-(Z-2)^2}e^{-U} \quad (2.7)$$

where $g(X) = (1 - (X - 1)^2)^{1/2}$, $X \in [0, 1]$, $Z \in [0, 4]$, and $U \geq 0$ is a one-sided inefficiency process. Setting $U = 0$ in (2.6)–(2.7) gives the frontiers for the two DGPs, as illustrated in Figure 1, where the frontier corresponding to (2.6) is shown in panel (a), and the frontier corresponding to (2.7) is shown in panel (b). Clearly, the frontiers are very different; it is clear that for a given level of the input variable X , the maximal output level Y^* in (2.6)

does not vary with Z , but the maximal output level Y^{**} in (2.7) does vary with Z . The separability condition is satisfied by the DGP in (2.6), but not by the DGP in (2.7).

If Assumption 2.1 holds, then for purposes of estimating efficiency, the environmental variables can be ignored. But if Assumption 2.2 holds, conditional estimators of efficiency introduced by Cazals et al. (2002) and Daraio and Simar (2005) must be used. Moreover, if Assumption 2.2 holds, then the set Ψ can still be defined as in (2.3), but this set and its boundary are not features of the DGP and hence are neither interesting nor relevant. Daraio et al. (2018) provide a test of Assumption 2.1 versus Assumption 2.2, and it is important to use this test whenever environmental variables are present.

From the viewpoint of the applied researcher, all that is observed is a single realization of an independent, identically distributed (iid) sample $\mathcal{S}_n = \{(X_i, Y_i, Z_i)\}_{i=1}^n$ drawn from $f_{XYZ}(x, y, z)$. Neither Ψ , $\mathcal{P}$, nor the Ψ^z are observed, and hence these must be estimated from data. Consequently, the researcher cannot know whether Assumption 2.1 or Assumption 2.2 holds. We will return to this point below, but for now we consider further the two cases where either (i) the separability condition holds, or (ii) it does not.

2.2 Case I: Separability Holds

In this case, we can focus on the unconditional production set given in (2.5). We adopt standard assumptions from microeconomic theory of the firm (e.g., see Shephard, 1970 or Färe, 1988). We assume (i) Ψ is closed; (ii) any non-zero production requires the use of some inputs (i.e., $(x, y) \notin \Psi$ if $x = 0, y \geq 0, y \neq 0$); and (iii) both inputs and outputs are freely disposable so that (i) $\tilde{x} \geq x \Rightarrow (\tilde{x}, y) \in \Psi$ and (ii) $\tilde{y} \leq y \Rightarrow (x, \tilde{y}) \in \Psi, \forall (x, y) \in \Psi$.

The Assumption that Ψ is closed ensures that the *efficient frontier* or *technology*

$$\Psi^\partial := \{(x, y) \mid (x, y) \in \Psi, (\gamma x, \gamma^{-1} y) \notin \Psi \text{ for any } \gamma \in (0, 1)\} \quad (2.8)$$

consisting of the set of extreme points of Ψ is contained in Ψ .¹ The requirement that some strictly positive amount of input must be used to produce any output greater than zero rules out the existence of free lunches. Free disposability of inputs and outputs implies weak monotonicity of the frontier. When Assumption 2.1 holds, the environmental variables Z have no effect on the set of feasible input-output combinations nor on the frontier in (2.8).

¹Inequalities involving vectors are defined on an element-by-element basis.

The frontier Ψ^∂ provides a benchmark against which production units' performance can be measured. Units operating on the frontier are said to be *technically efficiency*, while those operating under the frontier, in the interior of Ψ , are said to be *technically inefficient*. Several measures of technical efficiency are employed in the literature. The Farrell (1957) input efficiency measure

$$\theta(x, y) := \inf \{ \theta \mid (\theta x, y) \in \Psi \} \quad (2.9)$$

gives the proportion by which input levels can be feasibly reduced without reducing output levels. Alternatively, the Farrell (1957) output efficiency measure

$$\lambda(x, y) := \sup \{ \lambda \mid (x, \lambda y) \in \Psi \} \quad (2.10)$$

gives the feasible proportion by which output levels can be increased without increasing input quantities. The hyperbolic measure

$$\gamma(x, y) := \inf \{ \gamma > 0 \mid (\gamma x, \gamma^{-1} y) \in \Psi \} \quad (2.11)$$

proposed by Färe et al. (1985) gives the proportion by which output quantities can be increased while simultaneously reducing input quantities by the same proportion. Clearly, $\theta(x, y) \leq 1$, $\lambda(x, y) \geq 1$, and $\gamma(x, y) \leq 1$ for all $(x, y) \in \Psi$, with strict equality holding for any $(x, y) \in \Psi^\partial$.²

It is important to note that Ψ , and hence Ψ^∂ as well as the measures $\theta(x, y)$, $\lambda(x, y)$ and $\gamma(x, y)$ are unobserved. Consequently, they must be estimated from a sample $\mathcal{S}_n = \{(X_i, Y_i)\}_{i=1}^n$ of observed input-output pairs. There are several possibilities. First, Deprins et al. (1984) estimate Ψ by the free disposal hull of the sample observations in $\mathcal{S}_n$, i.e.,

$$\widehat{\Psi}_{\text{FDH}} := \bigcup_{(X_i, Y_i) \in \mathcal{S}_n} \{ (x, y) \in \mathbb{R}_+^{p+q} \mid x \geq X_i, y \leq Y_i \}. \quad (2.12)$$

This estimator does not impose convexity on Ψ . Substituting $\widehat{\Psi}_{\text{FDH}}$ for Ψ in (2.9)–(2.11) provides FDH estimators of the input, output and hyperbolic efficiency measures.

²In principle, for some $(x, y) \in \Psi^\partial$ one might have $\theta(x, y) < 1$ or $\lambda(x, y) > 1$ if the frontier is parallel to either all of the input axes or all of the output axes in some regions. However, this is ruled out by additional assumptions required to define a statistical model in which the efficiency measure defined in (2.9)–(2.11) are statistically consistent; see Kneip et al. (1998) and Kneip et al. (2008) for details.

The resulting FDH efficiency estimators can be computed easily by letting $\mathcal{D}_{x,y}$ denote the set indices of points in $\mathcal{S}_n$ dominating (x, y) , i.e., $\mathcal{D}_{x,y} = \{i \mid (X_i, Y_i) \in \mathcal{S}_n, X_i \leq x, Y_i \geq y\}$. Then the input-oriented FDH efficiency estimator is computed by solving

$$\hat{\theta}_{\text{FDH}}(x, y \mid \mathcal{S}_n) = \min_{i \in \mathcal{D}_{x,y}} \max_{j=1, \dots, p} \left(\frac{X_i^j}{x^j} \right), \quad (2.13)$$

where for a vector a , a^j denotes its j -th component, and the output-oriented FDH estimator can be computed by solving

$$\hat{\lambda}_{\text{FDH}}(x, y \mid \mathcal{S}_n) = \max_{i \in \mathcal{D}(x,y)} \min_{j=1, \dots, q} \left(\frac{Y_i^j}{y^j} \right), \quad (2.14)$$

The hyperbolic FDH estimator can be computed by solving

$$\hat{\gamma}_{\text{FDH}}(x, y \mid \mathcal{S}_n) = \min_{i=1, \dots, n} \left(\max_{\substack{j=1, \dots, p \\ k=1, \dots, q}} \left(\frac{X_i^j}{x^j}, \frac{y^k}{Y_i^k} \right) \right) \quad (2.15)$$

as shown by Wilson (2011).

Second, the variable returns to scale (VRS) version of the DEA estimator $\hat{\Psi}_{\text{VRS}}$ of Ψ is given by the convex hull of $\hat{\Psi}_{\text{FDH}}$, i.e.,

$$\hat{\Psi}_{\text{VRS}} = \{(x, y) \in \mathbb{R}_+^{p+q} \mid y \leq \mathbf{Y}\omega, x \geq \mathbf{X}\omega, \mathbf{i}'_n \omega = 1, \omega \in \mathbb{R}_+^n\} \quad (2.16)$$

where $\mathbf{X} = [X_1 \ \dots \ X_n]$ is a $(p \times n)$ matrix of input vectors, $\mathbf{Y} = [Y_1 \ \dots \ Y_n]$ is a $(q \times n)$ matrix of output vectors, $\mathbf{i}_n$ is an $(n \times 1)$ vector of ones and ω is an $(n \times 1)$ vector of weights. Then the VRS-DEA estimators of $\theta(x, y)$, $\lambda(x, y)$ and $\gamma(x, y)$ are obtained by substituting $\hat{\Psi}_{\text{VRS}}$ for Ψ in (2.9)–(2.11). The resulting estimators can be computed by solving

$$\hat{\theta}_{\text{VRS}}(x, y \mid \mathcal{S}_n) = \min_{\theta, \omega} \{\theta > 0 \mid y \leq \mathbf{Y}\omega, \theta x \geq \mathbf{X}\omega, \mathbf{i}'_n \omega = 1, \omega \in \mathbb{R}_+^n\}, \quad (2.17)$$

$$\hat{\lambda}_{\text{VRS}}(x, y \mid \mathcal{S}_n) = \max_{\lambda, \omega} \{\lambda > 0 \mid \lambda y \leq \mathbf{Y}\omega, x \geq \mathbf{X}\omega, \mathbf{i}'_n \omega = 1, \omega \in \mathbb{R}_+^n\} \quad (2.18)$$

and

$$\hat{\gamma}_{\text{VRS}}(x, y \mid \mathcal{S}_n) = \max_{\gamma, \omega} \{\gamma > 0 \mid \gamma^{-1}y \leq \mathbf{Y}\omega, \gamma x \geq \mathbf{X}\omega, \mathbf{i}'_n \omega = 1, \omega \in \mathbb{R}_+^n\}. \quad (2.19)$$

Estimates $\hat{\theta}_{\text{VRS}}(x, y \mid \mathcal{S}_n)$ or $\hat{\lambda}_{\text{VRS}}(x, y \mid \mathcal{S}_n)$ can be computed using linear programming methods, and $\hat{\gamma}_{\text{VRS}}(x, y \mid \mathcal{S}_n)$ can be computed using the numerical method given by Wilson

(2011) and implemented in the freely-available *FEAR* software package described by Wilson (2008).³

Finally, the constant returns to scale (CRS) version of the DEA estimator $\hat{\Psi}_{\text{CRS}}$ of Ψ is obtained by dropping the constraint $\mathbf{z}'_n \omega$ in (2.16). Then the CRS-DEA estimators of $\theta(x, y)$, $\lambda(x, y)$ and $\gamma(x, y)$ are obtained by substituting $\hat{\Psi}_{\text{CRS}}$ for Ψ in (2.9)–(2.11). The resulting estimators $\hat{\theta}_{\text{CRS}}(x, y \mid \mathcal{S}_n)$ or $\hat{\lambda}_{\text{CRS}}(x, y \mid \mathcal{S}_n)$ can be computed by solving (2.17) or (2.18) (respectively), and the estimator $\hat{\gamma}_{\text{CRS}}(x, y \mid \mathcal{S}_n)$ can be computed as the square root of either $\hat{\theta}_{\text{CRS}}(x, y \mid \mathcal{S}_n)$ or the reciprocal of $\hat{\lambda}_{\text{CRS}}(x, y \mid \mathcal{S}_n)$. See Simar and Wilson (2013, 2015) for discussion and details.

Thus, when Assumption 2.1 holds, three categories of nonparametric efficiency estimators (FDH, VRS-DEA and CRS-DEA) are available. For each of these categories, there are estimators for efficiency measures in input, output and hyperbolic directions. Of course, estimation reveals nothing; learning from data can occur only when statistical inference is made, and inference requires a well-defined statistical model. The appropriate statistical model consists of the assumptions on Ψ described above as well as a number of assumptions on the statistical process (i.e., the data-generating process) that yields observed data. Among these assumptions, we require that observed data on inputs and outputs consist of identically, independently distributed (iid) realizations of random variables $(X, Y) \in \mathbb{R}_+^{p+q}$ with bounded support on Ψ . We require also that the joint density of (X, Y) must be strictly positive along the frontier Ψ^∂ and continuously differentiable in a neighborhood of Ψ^∂ . We also require that the frontier Ψ^∂ be sufficiently smooth, depending on the class of estimators—FDH, VRS-DEA or CRS-DEA—being used. Technical details are given by Kneip et al. (2008,

³Färe et al. (2016) suggest an alternative method for computing $\hat{\gamma}_{\text{VRS}}(x, y \mid \mathcal{S}_n)$ and remark (page 313) that their “algorithm is very fast.” Curiously, Färe et al. do not cite Wilson (2011), nor do they give any evidence on how fast their algorithm is. Simple experiments using simulated data indicate that with $n = 1,000$, $p = q = 2$, the *FEAR* package for *R* by Wilson (2008) using the algorithm of Wilson (2011) is faster than the Färe et al. algorithm implemented in the *R* package *DJL* by Lim (2021) by a factor of about 0.01157. In other words, with $p = q = 2$, 1,000 VRS-DEA efficiency estimates can be computed using *FEAR* in about 1.157 percent of the time required using the Färe et al. algorithm implemented by Lim (2021). *Fast* is a relative term, and the Färe et al. algorithm (as implemented by Lim, 2021) is far slower than the Wilson (2011) algorithm.

Färe et al. also state (page 313) that “we develop an LP-based computational algorithm to generate the exact value of” the hyperbolic distance function. Presumably they mean its estimator, as (2.11) is not observable. More importantly, many base-10 numbers do not have exact representations in base-2, and so almost all computations on digital computers are approximate. The algorithm of Wilson (2011) can be made accurate to an arbitrary number of decimal digits. See Wilson (2011) for details.

2015, 2021).

The statistical properties of the nonparametric efficiency estimators are developed in a number of papers. In all cases, the estimators converge at rate n^κ and have limiting distributions. For the case of FDH estimators, Park et al. (2000) and Daouia et al. (2017) establish that $\kappa = 1/(p+q)$, and the limiting distributions are Weibull, but with parameters that depend on features of the model being estimated (e.g., curvature of Ψ^θ) and which are difficult to estimate. These results hold regardless of whether Ψ is convex. For the case of VRS-DEA estimators, under VRS and convexity of Ψ , Kneip et al. (1998) prove that $\kappa = 2/(p+q+1)$ and Kneip et al. (2008) establish existence of and characterize limiting distributions. However, the limiting distributions do not have explicit, analytical expressions. Park et al. (2010) establish corresponding results for CRS-DEA estimators under CRS and convexity of Ψ , where $\kappa = 2/(p+q)$. Kneip et al. (2016) prove that under CRS and convexity of Ψ , the VRS-DEA estimators converge at the faster rate achieved by CRS-DEA estimators under the same conditions. The existence of limiting distributions is needed to ensure validity of bootstrap methods developed by Jeong and Simar (2006), Kneip et al. (2008, 2011) and Simar and Wilson (2011a) for estimating confidence intervals for the efficiency of individual producers.

2.3 Case II: Separability does not Hold

2.3.1 Conditional Efficiency Estimation

When Assumption 2.2 holds, Ψ remains as defined in (2.3) but the interpretation in (2.5) is not valid. Nor can the joint support of (X, Y, Z) be factored as in (2.4). In addition, the frontier defined in (2.8) has no operational meaning. To illustrate this point, return to the example discussed above in Section 2.1 in (2.6)–(2.7) and $(0.5, 0.2, 1.75)$, $(0.5, 0.2, 2.0)$, and $(0.5, 0.2, 2.5)$ for (X, Y, Z) . If the true DGP is as given by (2.6), the same algebra reveals that the *true* Farrell output efficiencies defined by (2.10) are (approximately) $0.8660/0.2 = 4.33$ for each of these three observations. On the other hand, if the true DGP is given by (2.7), then the Farrell output efficiencies corresponding to these three observations are (approximately) $0.8136/0.2 = 4.068$, $0.8660/0.2 = 4.33$, and $0.6745/0.2 = 3.375$ (respectively). As noted in Section 2.1, the frontier corresponding to (2.6) (where Assumption 2.1 is satisfied) is invariant with respect to Z , while the frontier corresponding to (2.7) (where Assumption 2.2

holds) is not. It is clear that whether the “separability” condition holds has an impact on the underlying, true efficiency levels, and this impact may be large. In the example considered here, the output efficiency level for the third observation is about 28.3 percent larger if the DGP is given by (2.6), where the “separability” condition is satisfied, as opposed to the case where the DGP is given by (2.7), where the “separability” condition is not satisfied.

The example illustrates the fact that efficiency must be measured against the frontier of Ψ^z given by

$$\Psi^{z\partial} := \{(x, y) \mid (x, y) \in \Psi^z, (\gamma x, \gamma^{-1}y) \notin \Psi^z \text{ for any } \gamma \in (0, 1)\} \quad (2.20)$$

when Assumption 2.2 holds. This requires *conditional* efficiency measures and estimators. The conditional analogs of the efficiency measures defined in (2.9)–(2.11) are

$$\theta(x, y \mid z) := \inf \{\theta \mid (\theta x, y) \in \Psi^z\} \quad (2.21)$$

$$\lambda(x, y \mid z) := \sup \{\lambda \mid (x, \lambda y) \in \Psi^z\} \quad (2.22)$$

and

$$\gamma(x, y \mid z) := \inf \{\gamma > 0 \mid (\gamma x, \gamma^{-1}y) \in \Psi^z\}, \quad (2.23)$$

respectively.

How one might estimate the conditional efficiency measures in (2.21)–(2.23) depends on whether Z is discrete or continuous. If all of the r environmental variables in Z are discrete, then the sample observations can be split into groups according to the values of the discrete variables in Z . For example, if all of the environmental variables are discrete, and the j th variable takes K_j different values, then there are $G = \prod_{j=1}^r K_j$ possible groups. Within each group, the conditional efficiency of each observation can be estimated using the unconditional efficiency estimators in (2.9)–(2.11) and only the observations in the given group. The conditional efficiency estimates for each observation in a given group are then independent of the observations in all *other* groups.

Alternatively, if all of the r environmental variables in Z are continuous, then localization is needed. The conditional analog of the input-oriented FDH estimator in (2.13) for the case where Z is continuous is

$$\widehat{\theta}_{\text{FDH}}(x, y \mid z, \mathcal{S}_n) = \min_{i \in \mathcal{D}_{x, y, z}} \max_{j=1, \dots, p} \left(\frac{X_i^j}{x^j} \right), \quad (2.24)$$

where $\mathcal{D}_{x,y,z} = \{i \mid (X_i, Y_i) \in \mathcal{S}_n, X_i \leq x, Y_i \geq y, |Z_i - z| \leq h\}$ and h is a vector of r bandwidths. Similarly, the conditional analogs of the output-oriented and hyperbolic FDH estimators in (2.14)–(2.15) are given by

$$\hat{\lambda}_{\text{FDH}}(x, y \mid z, \mathcal{S}_n) = \max_{i \in \mathcal{D}(x,y,z)} \min_{j=1, \dots, q} \left(\frac{Y_i^j}{y^j} \right) \quad (2.25)$$

and

$$\hat{\gamma}_{\text{FDH}}(x, y \mid z, \mathcal{S}_n) = \min_{i \mid (|Z_i - z| \leq h)} \left(\max_{\substack{j=1, \dots, p \\ k=1, \dots, q}} \left(\frac{X_i^j}{x^j}, \frac{y^k}{Y_i^k} \right) \right). \quad (2.26)$$

Similarly, corresponding, conditional (on continuous Z) VRS-DEA estimators are given by

$$\hat{\theta}_{\text{VRS}}(x, y \mid z, \mathcal{S}_n) = \min_{\theta, \omega} \{ \theta > 0 \mid y \leq \mathbf{Y}\zeta(z), \theta x \geq \mathbf{X}\zeta(z), \mathbf{i}'_n \zeta(z) = 1, \omega \in \mathbb{R}_+^n \}, \quad (2.27)$$

$$\hat{\lambda}_{\text{VRS}}(x, y \mid z, \mathcal{S}_n) = \max_{\lambda, \omega} \{ \lambda > 0 \mid \lambda y \leq \mathbf{Y}\zeta(z), x \geq \mathbf{X}\zeta(z), \mathbf{i}'_n \zeta(z) = 1, \omega \in \mathbb{R}_+^n \} \quad (2.28)$$

and

$$\hat{\gamma}_{\text{VRS}}(x, y \mid z, \mathcal{S}_n) = \max_{\gamma, \omega} \{ \gamma > 0 \mid \gamma^{-1} y \leq \mathbf{Y}\zeta(z), \gamma x \geq \mathbf{X}\zeta(z), \mathbf{i}'_n \zeta(z) = 1, \omega \in \mathbb{R}_+^n \} \quad (2.29)$$

where $\zeta(z) = \omega \circ d(z)$, $d(z)$ an $(n \times 1)$ Bernoulli vector whose i th element equals 1 if $|Z_i - z| \leq h$ and 0 otherwise, $\mathbf{X}$, $\mathbf{Y}$, ω and $\mathbf{i}$ are defined as before, and “ $\circ$ ” denotes the Hadamard product.⁴ The conditional estimators $\hat{\theta}_{\text{VRS}}(x, y \mid z, \mathcal{S}_n)$ and $\hat{\lambda}_{\text{VRS}}(x, y \mid z, \mathcal{S}_n)$ can be computed using standard linear-programming methods, and $\hat{\gamma}_{\text{VRS}}(x, y \mid z, \mathcal{S}_n)$ is easily computed after a simple modification of the numerical algorithm given by Wilson (2011). Conditional (on continuous Z) CRS-DEA estimators are obtained by dropping the constraint $\mathbf{i}'_n \zeta(z) = 1$ in (2.27)–(2.29).

In situations where Z contains *both* discrete and continuous variables, one could split the observations in $\mathcal{S}_n$ into G groups defined by values of the discrete values as discussed above, and then independently for each group, apply the conditional estimators discussed above. The resulting estimates are then conditional on both the discrete as well as the continuous environmental variables in Z .

⁴Recall that the Hadamard product of two vectors $A = [a_i]$ and $B = [b_i]$ of length n , denoted by $A \circ B$, is the vector $C = [c_i]$ where $c_i = a_i b_i \forall i = 1, \dots, n$. Hence the i th element of the vector of weights $\zeta(z)$ in (2.27)–(2.29) is active (i.e., non-zero) if and only if Z_i is in an h -neighborhood of z (i.e., if $|Z_i - z| \leq h$).

2.3.2 Bandwidth Selection Issues

The conditional efficiency estimators discussed above for continuous $Z \in \mathbb{R}^r$ depends on a vector of r bandwidths h that controls the size of the neighborhood around z for defining local versions of the various envelopment estimators. These bandwidths must be small enough to compare firms facing similar environmental conditions (to avoid too much bias), but at the same time, large enough to have enough observations in the neighborhood of z , i.e., $n_h = \text{card}\{Z_i \mid |Z_i - z| \leq h\}$ should be large to avoid excessive variance of the resulting estimators. This tradeoff between bias and variance is common in many nonparametric econometric problems, and optimal bandwidths are determined by minimizing mean integrated squared error in order to find a balance between bias and variance. Theoretical results are given by Jeong et al. (2010), Daraio et al. (2018) and Bădin et al. (2019). In particular, least-squares-cross-validation (LSCV) techniques can be used for nonparametric estimation of the conditional distribution of (X, Y) given $Z = z$ to obtain bandwidths of appropriate order. Use of kernel functions permits smoothing values of Z_i near z . As discussed by Daraio and Simar (2005), kernel functions with bounded support (e.g., Epanechnikov, quartic, ...) have to be used to ensure that the estimation for $Z = z$ is obtained by only using data points such that $|Z_i - z| \leq h$. As shown by Li et al. (2013), this leads to bandwidths $h_j = c_j n^{-1/(r+4)}$ for some constant $c_j > 0$, and various nonparametric software packages can be used to provide these values. Note that, as shown in Jeong et al. (2010), the effective number of points used for the local estimation when $Z = z$ should be of order $n_h \approx n \prod_{j=1}^r h_j$.⁵

2.3.3 Statistical Properties of Conditional Measures

The statistical properties of the conditional efficiency measures are similar to those of the unconditional measures described above for the “separable” case. But of course the “effective” sample size used to define the envelopment estimators will be drastically reduced as the dimension of Z increases.

To see this, first consider the case where Z is discrete. As discussed above, the values

⁵ Bădin et al. (2019) suggest a slight modification of the LSCV bandwidths to improve the local bias in the estimation. This adjustment produces bandwidths $h_j = d_j n^{-1/(r+\kappa^{-1})}$ for some constant $d_j > 0$ where κ defined above determines the rate of convergence and depends on the chosen estimator (i.e., FDH or DEA types). In practice one may chose $h_j = n^{1/(r+4)} n^{-1/(r+\kappa^{-1})} h_{j,\text{LSCV}}$, where $h_{j,\text{LSCV}}$ is produced by typical software packages using LSCV.

in Z serve to divide the data into G possible groups, with n_g observations in group g , $g = 1, \dots, G$ and $\sum_{g=1}^G n_g = n$. From a mathematical viewpoint, the “curse of dimensionality” is avoided since $n_g \approx n \Pr(Z \in \text{Group } g)$ is proportional to n . Consequently, the rate of convergence is unaffected when the data are divided into the G groups. In other words, when the unconditional efficiency estimators are used for observations in group g , the rate of convergence (as $n \rightarrow \infty$) remains the same as in cases where all the data fall into a single group of size n . But in practice, for a given sample size n , some of the n_g may be too small to obtain reasonable estimates. In such cases, some regrouping of data is recommended. Moreover, when $n_g < n$ are used for estimation one should expect variance of the estimates to be larger than when n observations are used.

For continuous variables $Z \in \mathbb{R}^r$, the situation is more subtle because the number of points used for the local estimation when $Z = z$ is of order $n_h \approx n \prod_{j=1}^r h_j$. If the all the bandwidths have the LSCV optimal order $n^{-1/(r+4)}$, then $n_h \approx n^{4/(r+4)}$ which determines the rates of convergence of the conditional estimators. Consequently, in this case, the dimension r of Z deteriorates the rate of convergence from n^κ to $n_h^\kappa = n^{4\kappa/(r+4)}$. If corrected LSCV bandwidths are used (see footnote 5), then $n_h^\kappa = n^{1/(\kappa r+1)}$, with similar consequences when r increases. Conditional efficiency estimation requires more data to achieve similar precision of the estimates, as for the discrete case.

3 Central Limit Theorems for Efficiency Scores

The results in this section hold for the input, output and hyperbolic efficiency measures $\theta(x, y)$, $\lambda(x, y)$ and $\gamma(x, y)$ defined in (2.9)–(2.11), their corresponding unconditional FDH, VRS-DEA and CRS-DEA estimators introduced in Section 2.2, and the corresponding conditional measures and estimators discussed in Section 2.3. Thus, in order to simplify the discussion in this section, let $\varphi(x, y)$ and $\varphi(x, y \mid z)$ represent any of the unconditional or conditional efficiency measures described above (input, output, or hyperbolic) and let $\hat{\varphi}(x, y \mid \mathcal{S}_n)$ and $\hat{\varphi}(x, y \mid z, \mathcal{S}_n)$ represent the appropriate, corresponding envelopment estimator (FDH, VRS-DEA or CRS-DEA, depending on the particular case at hand). The results of the preceding section allow one to make inference about individual efficiencies (by using appropriate bootstrap techniques as in Kneip et al. (2008, 2011) and Simar and Wilson (2011a) for DEA estimators and Jeong and Simar (2006) for FDH estimators) to make infer-

ence about individual efficiency scores. Specifically, the appropriate bootstrap methods can be used to estimate the bias and variance of the estimators of $\widehat{\varphi}(x, y \mid \mathcal{S}_n)$ and $\widehat{\varphi}(x, y \mid z, \mathcal{S}_n)$ and to estimate confidence intervals for $\varphi(x, y)$ or for $\varphi(x, y \mid z)$ for fixed points of interest.

However, for purposes of testing hypotheses about the mean efficiency of a group of producers or for testing hypotheses about model features, more is needed. For these purposes, test statistics based on the efficiency estimators evaluated at *random* points (X_i, Y_i) in the unconditional case or (X_i, Y_i, Z_i) in the conditional case are needed.

3.1 CLTs for Unconditional Efficiency Measures

For the unconditional case, consider the (population) average efficiency $\mu_\varphi = E_{XY}(\varphi(X, Y))$ over some group of firms of interest. Many researchers have naïvely supposed that standard, textbook Central Limit Theorems (CLT) can be applied along with sample averages such as $\widehat{\mu}_n = n^{-1} \sum_{i=1}^n \widehat{\varphi}(X_i, Y_i \mid \mathcal{S}_n)$ to obtain appropriate statistics and asymptotic normal theory enabling inference about μ_φ .⁶ However, due to the work of Kneip et al. (2015) it is clear that this is wrong and that the problem is more complicated due to the inherent bias of the envelopment estimators. The results of Kneip et al. (2015) reveal that the bias of each efficiency estimate is of order $n^{-\kappa}$, where κ depends on the chosen estimator and was defined above. When averaging in $\widehat{\mu}_n$, the bias remains of the same order whereas the variance, under appropriate regularity assumptions, is shown to be of order n^{-1} (see Kneip et al. (2015) for details).

One of the most important practical consequence of the moment properties derived in Theorems 3.1–3.3 in Kneip et al. (2015) is their Theorem 4.1 which can be summarized as follows. Let $\sigma_\varphi^2 = \text{Var}(\varphi(X, Y))$. Then under mild regularity conditions, we have the following important theorem.

Theorem 3.1. *As $n \rightarrow \infty$,*

- (i) $\widehat{\mu}_n$ converges in probability to μ_φ with a bias of order $O(n^{-\kappa})$,
- (ii) $\sqrt{n}(\widehat{\mu}_n - \mu_\varphi - Cn^{-\kappa} - R_{n,\kappa}) \xrightarrow{\mathcal{L}} N(0, \sigma_\varphi^2)$, for some $C > 0$, and $R_{n,\kappa} = o(n^{-\kappa})$,
- (iii) $\widehat{\sigma}_\varphi^2 = n^{-1} \sum_{i=1}^n (\widehat{\varphi}(X_i, Y_i \mid \mathcal{S}_n) - \widehat{\mu}_n)^2$ is a consistent estimator of σ_φ^2 .

⁶For examples of standard, textbook presentations of CLT results, see Pfeiffer (1978, Section 6.5), Zaman (1996, Section 13.3), Casella and Berger (2002, pp. 236–240) or Mittelhammer (2013, Section 5.8).

The main error when using naive, textbook versions of CLTs for means of efficiency scores involves ignoring part (ii) of Theorem 3.1. The bias term $Cn^{-\kappa} + R_{n,\kappa}$ vanishes when multiplied by $\sqrt{n}$ only if $\kappa > 1/2$. Since $p, q \geq 1$, this *never* happens for FDH estimators, and happens only when $p = q = 1$ for VRS-DEA estimators and for $p + q \leq 3$ for the CRS-DEA case. But the bias term does not vanish when $\kappa = 1/2$, i.e., when $p + q = 3$ with VRS-DEA estimators or $p + q = 4$ with CRS-DEA estimators. For larger values of $p + q$, the bias term explodes to infinity.

Kneip et al. (2015) establish CLTs to overcome these problems. The new CLTs involve estimating the leading term $Cn^{-\kappa}$ of the bias and, for cases where $\kappa < 1/2$, using a sample mean over a subsample of the observations to control the other term, $R_{n,\kappa}$ in part (ii) of Theorem 3.1. The bias estimate is based on a jackknife idea: the sample $\mathcal{S}_n$ is split randomly into two mutually exclusive, collectively exhaustive (and hence independent) parts $\mathcal{S}_{n/2}^{(j)}$, $j = 1, 2$ of size $n/2$ (here, n is assumed even to simplify notation). For $j = 1, 2$, sample means of the efficiency estimates using only the sample $\mathcal{S}_{n/2}^{(j)}$ are computed, leading to the means $\hat{\mu}_{n/2}^{(j)}$. Formally, this yields

$$\hat{\mu}_{n/2}^{(j)} = 2n^{-1} \sum_{i|(X_i, Y_i) \in \mathcal{S}_{n/2}^{(j)}} \hat{\varphi}(X_i, Y_i) \mid \mathcal{S}_{n/2}^{(j)} \quad (3.1)$$

for $j = 1, 2$. Intuitively, the sample means $\hat{\mu}_{n/2}^{(1)}$ and $\hat{\mu}_{n/2}^{(2)}$ given by (3.1) both provide consistent estimates of μ_φ based on $n/2$ observations, so that the leading bias term is now $C(n/2)^{-\kappa}$. Kneip et al. (2015) show that

$$\hat{B}_{n,\kappa} = (2^\kappa - 1)^{-1} \left[(\hat{\mu}_{n/2}^{(1)} + \hat{\mu}_{n/2}^{(2)})/2 - \hat{\mu}_n \right] \quad (3.2)$$

is a consistent estimator of the leading term of the bias, $Cn^{-\kappa}$, with an error of the same order as $R_{n,\kappa}$, i.e., $o(n^{-\kappa})$. In practice, for reducing the variance of this bias estimate, Kneip et al. (2016) suggest computing the estimate in (3.2) by randomly splitting the sample a large number of times, computing the bias estimate for each random split, and then averaging these estimates at the end to obtain a final bias estimate.

This resolves the leading part of the bias, but the remainder $R_{n,\kappa} = o(n^{-\kappa})$ still creates problems when $\kappa < 1/2$. The idea here is to limit the mean of efficiency scores appearing in the CLT above to a smaller number of observations in order to increase the variance of the

mean. When $\kappa < 1/2$, randomly draw a subsample $\mathcal{S}_{n_\kappa}^* \subset \mathcal{S}_n$ of size $n_\kappa = \lfloor n^{2\kappa} \rfloor < n$ and compute

$$\hat{\mu}_{n_\kappa} = n_\kappa^{-1} \sum_{\{j | (X_j, Y_j) \in \mathcal{S}_{n_\kappa}^*\}} \hat{\varphi}(X_j, Y_j | \mathcal{S}_n), \quad (3.3)$$

where the notation makes clear that each individual $\hat{\varphi}(X_j, Y_j | \mathcal{S}_n)$ is estimated with the full sample $\mathcal{S}_n$.⁷ Using this subsampling for cases where $\kappa < 1/2$, Kneip et al. (2015) derive the following CLT.⁸

Theorem 3.2. *For $p + q \leq 3$ if the FDH estimator is used, for $p + q \leq 4$ and if Ψ is convex and the VRS-DEA estimator is used, for $p + q \leq 5$ and if Ψ exhibits CRS and the CRS-DEA or the VRS-DEA estimator is used,*

$$\sqrt{n} \left(\hat{\mu}_n - \mu_\varphi - \hat{B}_{n,\kappa} \right) \xrightarrow{\mathcal{L}} N(0, \sigma_\varphi^2) \quad (3.4)$$

as $n \rightarrow \infty$. When $\kappa < 1/2$,

$$\sqrt{n_\kappa} \left(\hat{\mu}_{n_\kappa} - \mu_\varphi - \hat{B}_{n,\kappa} \right) \xrightarrow{\mathcal{L}} N(0, \sigma_\varphi^2) \quad (3.5)$$

as $n \rightarrow \infty$.

As noted by Kneip et al. (2015), either version of CLT can be used for the FDH estimator when $p + q = 3$ (or for the VRS-DEA estimator when $p + q = 4$ or for the CRS-DEA estimator when $p + q = 5$) with the second approximation providing better properties in finite samples in such cases.

Using the fact that part (iii) of Theorem 3.1 provides a consistent estimator of the variance σ_φ^2 , Kneip et al. (2015) show how the above results can be used to provide asymptotic confidence intervals by using the quantiles of a standard normal. Similarly these results can also be used, as established by Kneip et al. (2016), to compare the mean of two groups of firms for which we have two independent samples.

⁷For $a \in \mathbb{R}$, $\lfloor a \rfloor$ denotes the largest integer that is less than or equal to a .

⁸Kneip et al. (2015) derive their CLT in terms of unconditional, input-oriented efficiency measures and estimates, but the results extend to the output and hyperbolic directions. Kneip et al. (2015) produce particular and different expressions of $R_{n,\kappa}$ as a power of the dimension $p + q$, according the chosen estimator, this implies the various conditions on $p + q$ in the theorem.

3.2 CLTs for Conditional Efficiency Measures

For discrete Z , in each group $g = 1, \dots, G$ with sample of size n_g (recall the discussion in Section 2.3.1), $n_g \rightarrow \infty$ as $n \rightarrow \infty$. Hence for the case where Z is discrete, the CLTs in Theorem 3.2 hold, separately and independently, for each group using the sample size n_g for each group. But in the case where Z is continuous, things are more complicated. Nonetheless, Daraio et al. (2018) show that the conditional efficiency scores share properties similar to those obtained in the unconditional case, with n replaced by n_h , the order of the effective number of observations used to estimate conditional efficiencies. As noted by Jeong et al. (2010), for a given h , the conditional estimators do not target $\varphi(x, y | z)$, but instead estimate

$$\varphi^h(x, y | z) = \sup\{\varphi > 0 \mid (x, \varphi y) \in \Psi^{z,h}\}, \quad (3.6)$$

with the *conditional* attainable set given by

$$\Psi^{z,h} = \{(X, Y) \mid X \text{ can produce } Y, \text{ when } |Z - z| \leq h\} \quad (3.7)$$

Moreover, it is clear that $\Psi^{z,h} = \bigcup_{|\tilde{z}-z| \leq h} \Psi^{\tilde{z}}$. As in Jeong et al. (2010) and Daraio et al. (2018), $\varphi(x, y | z)$ is assumed to be continuous in z so that $\varphi^h(x, y | z) - \varphi(x, y | z) = O(h)$ as $h \rightarrow 0$.

Consequently, CLTs for conditional measures are derived in terms of the true moments defined by

$$\mu_c^h = E(\varphi^h(X, Y | Z)) = \int_{\mathcal{P}} \varphi^h(x, y | z) f_{XYZ}(x, y, z) dx dy dz \quad (3.8)$$

and

$$\sigma_c^{2,h} = \text{VAR}(\varphi^h(X, Y | Z)) = \int_{\mathcal{P}} (\varphi^h(x, y | z) - \mu_c^h)^2 f_{XYZ}(x, y, z) dx dy dz. \quad (3.9)$$

These are the conditional analogs of μ and σ^2 , and conditional estimators are given by

$$\hat{\mu}_{c,n} = n^{-1} \sum_{i=1}^n \hat{\varphi}(X_i, Y_i | Z_i, \mathcal{S}_n) \quad (3.10)$$

and

$$\hat{\sigma}_{c,n}^{2,h} = n^{-1} \sum_{i=1}^n [\hat{\varphi}(X_i, Y_i | Z_i, \mathcal{S}_n) - \hat{\mu}_{c,n}]^2. \quad (3.11)$$

Unfortunately, as shown by Daraio et al. (2018), a CLT for μ_c^h cannot be developed using $\hat{\mu}_{c,n}$, the mean computed over the full sample $\mathcal{S}_n$. This results from the fact that as before, the inherent bias of $\hat{\mu}_{c,n}$ is determined by the number of effective observations used in computing the individual $\hat{\varphi}(X_i, Y_i | Z_i, \mathcal{S}_n)$, and this bias is of order $O(n_h^{-\kappa})$ (see Lemma 4.1 in Daraio et al. (2018)), whereas its variance is of order n^{-1} . Daraio et al. show however that $\hat{\sigma}_{c,n}^{2,h}$ is a consistent estimator of $\sigma_c^{2,h}$.

Daraio et al. (2018) use a device similar to the one above to avoid this problem. In particular, instead of using the sample mean in (3.10), Daraio et al. use an estimate of μ_c^h based on a random subsample $\mathcal{S}_{n_h}^*$ of size n_h given by

$$\hat{\mu}_{c,n_h} = n_h^{-1} \sum_{\{(X_i, Y_i, Z_i) \in \mathcal{S}_{n_h}^*\}} \hat{\varphi}(X_i, Y_i | Z_i, \mathcal{S}_n). \quad (3.12)$$

Even so, as for the unconditional case above, standard, textbook CLTs based on $\hat{\mu}_{c,n_h}$ are not valid when $\kappa \leq 1/2$ due to the presence of the bias term which here is of order $O(n_h^{-\kappa}) + o(n_h^{-\kappa})$. Thus, Daraio et al. (2018) provide a way to estimate the leading term in the bias by adapting the jackknife algorithm of Kneip et al. (2015) to conditional measures. This provides $\hat{B}_{n_h, \kappa}^c$ in the same way as $\hat{B}_{n, \kappa}$ defined above. Due to the particular expressions available for the part $o(n_h^{-\kappa})$ of the bias, depending on the chosen estimator, if $p + q \leq 3$ for the FDH case, and if $p + q \leq 4$ for the VRS-DEA case, this correction is sufficient to derive a valid CLT. However, when $\kappa < 1/2$, this is not enough due to the remainder term $o(n_h^{-\kappa})$ and, as above for the unconditional case, one must compute the average on a even smaller random subsample $\mathcal{S}_{n_h, \kappa}^{**}$ of $\mathcal{S}_n$ of size $n_{h, \kappa} = \lfloor n_h^{2\kappa} \rfloor < n_h$ to control its variance. This leads to the estimator

$$\hat{\mu}_{c,n_{h,\kappa}} = \frac{1}{n_{h,\kappa}} \sum_{(X_i, Y_i, Z_i) \in \mathcal{S}_{n_{h,\kappa}}^{**}} \hat{\varphi}(X_i, Y_i | Z_i, \mathcal{S}_n). \quad (3.13)$$

Combining these results, Daraio et al. establish the following CLTs (appearing as Theorems 4.3 and 4.4 in Daraio et al., 2018) All together, this provides the following CLTs for conditional efficiency scores.

Theorem 3.3. *For $p + q \leq 3$ when the FDH estimator is used, for $p + q \leq 4$ and if Ψ is convex when the VRS-DEA estimator is used, or for $p + q \leq 5$ and if Ψ exhibits CRS and*

either the CRS-DEA or VRS-DEA estimator is used,

$$\frac{\sqrt{n_h} \left(\hat{\mu}_{c,n_h} - \mu_c^h - \hat{B}_{n_h,\kappa}^c \right)}{\sqrt{\hat{\sigma}_{c,n}^{2,h}}} \xrightarrow{\mathcal{L}} N(0, 1) \quad (3.14)$$

as $n \rightarrow \infty$ When $\kappa < 1/2$,

$$\frac{\sqrt{n_{h,\kappa}} \left(\hat{\mu}_{c,n_{h,\kappa}} - \mu_c^h - \hat{B}_{n_{h,\kappa}}^c \right)}{\sqrt{\hat{\sigma}_{c,n}^{2,h}}} \xrightarrow{\mathcal{L}} N(0, 1) \quad (3.15)$$

as $n \rightarrow \infty$.

The next section shows that the CLTs provided by Theorems 3.2 and 3.3 offer ways to test various hypotheses about model structure in various situations.

4 Testing Hypotheses about Model Structure

4.1 Why Model Structure is Important

The discussion in Section 2 reveals that there are three broad types of nonparametric estimators that one might consider for estimation of efficiency. The FDH estimators do not require convexity of the production set. The VRS-DEA estimators require that the production set be convex, and the CRS-DEA estimators require both convexity of the production set and constant returns to scale of the frontier. In terms of flexibility, the FDH estimators are more flexible than the VRS-DEA estimators in the sense that the FDH estimators remain consistent regardless of whether the production set is convex, whereas the VRS-DEA estimators are inconsistent if the production set is not convex. Similarly, the VRS-DEA estimators are consistent regardless of whether the frontier exhibits CRS, but the CRS-DEA estimators are inconsistent if the frontier does not exhibit CRS.

In addition, for each of the three types of estimators, both unconditional and conditional versions of the estimators are available. Where environmental variables are present, the unconditional estimators provide consistent estimates only if the separability condition in Assumption 2.1 holds. The conditional estimators provide consistent estimates regardless of whether Assumption 2.1 or Assumption 2.2 holds, but if Assumption 2.1 holds, the conditional estimators will have slower convergence rates—and hence likely will have greater estimation error—than the unconditional estimators.

Since there are three broad types of efficiency estimators (i.e., FDH, VRS-DEA and CRS-DEA) and both unconditional and conditional versions of each type, there are altogether six classes of nonparametric efficiency estimators that one might use. Moreover, within each class, one can choose the direction in which to measure inefficiency. The choice of direction is not particularly problematic, given that observations are assumed to be drawn independently from the joint density $f_{XYZ}(x, y, z)$. But the choice between the six classes of estimators is certainly important, and to a large extent determines whether the resulting efficiency estimates are consistent and meaningful.

Most studies that employ nonparametric efficiency estimation make ad hoc choices between the six classes. The VRS-DEA and CRS-DEA estimators are used far more often than FDH estimators, and CRS-DEA estimators are often chosen over VRS-DEA estimators to avoid having large numbers of estimates equal to one. The issue of separability versus non-separability is often ignored—papers that do so continue to appear despite the warnings of Simar and Wilson (2007, 2011b). In the analysis of hospital data that appears below, we first test for separability with respect to both time and ownership type. We do this using FDH estimators, which do not require convexity of the production set. We also test for separability with respect to the continuous environmental variables considered by Burgess and Wilson (1995). After testing for separability, we then test for convexity to see whether the VRS-DEA estimators might be used. Moreover, when examining the hospital data, we work in the hyperbolic direction in order to avoid sensitivity to the curvature of the frontier as discussed by Wilson (2011). Consequently, in the discussion that follows below, we describe our tests in terms of hyperbolic measures and estimators, but the discussion in Section 3 makes clear that all of the tests we use can be carried out in the input or output directions.

4.2 Testing of Separability

4.2.1 The Discrete Case

We first test the separability condition in Assumption 2.1 before testing convexity of the production set. Consequently, we use FDH estimators to test separability. As discussed below in Section 5, observations in the data used by Burgess and Wilson (1995, 1996 and 1998) can be grouped by year, by ownership type, or by both year and type. Conceivably, the production sets may differ across groups. Suppose there are G groups, indexed by

$g = 1, \dots, G$. Then let $Z_i = g$, reflecting that the i th observation falls in group g . Since Z_i is discrete, the conditional, hyperbolic FDH efficiency estimator amounts to using the unconditional, hyperbolic FDH efficiency estimator independently on separate groups.

In order to implement the test developed by Daraio et al. (2018), we must first split the sample $\mathcal{S}_n$ randomly into two sub-samples $\mathcal{S}_{1,n_1}$ and $\mathcal{S}_{2,n_2}$ of sizes n_1 and n_2 such that $n_1 = \lfloor n \rfloor$ and $n_2 = n - n_1$. Necessarily, $\mathcal{S}_{1,n_1} \cup \mathcal{S}_{2,n_2} = \mathcal{S}_n$ and $\mathcal{S}_{1,n_1} \cap \mathcal{S}_{2,n_2} = \emptyset$. This ensures that efficiency estimates obtained using $\mathcal{S}_{1,n_1}$ are independent of those obtained using $\mathcal{S}_{2,n_2}$. This independence is needed to obtain non-degenerate limiting distributions for the test statistics that will be used, which is essential for making inference about separability versus non-separability.

In order to test separability across G groups, divide the observations in the second sub-sample into their respective groups so that $\mathcal{S}_{2,g,n_{2,g}}$ is the set of $n_{2,g}$ observations for group g , with $\sum_{g=1}^G n_{2,g} = n_2$, $\mathcal{S}_{2,g,n_{2,1}} \cup \dots \cup \mathcal{S}_{2,g,n_{2,G}} = \mathcal{S}_{2,n_2}$ and $\mathcal{S}_{2,g,n_{2,1}} \cap \dots \cap \mathcal{S}_{2,g,n_{2,G}} = \emptyset$. Then consider the sample means

$$\hat{\mu}_{1,n_1} = n_1^{-1} \sum_{(X_i, Y_i, Z_i) \in \mathcal{S}_{1,n_1}} \hat{\gamma}_{\text{FDH}}(X_i, Y_i \mid \mathcal{S}_{1,n_1}) \quad (4.1)$$

and

$$\hat{\mu}_{2,n_2} = n_2^{-1} \sum_{g=1}^G \left(\sum_{(X_i, Y_i, Z_i) \in \mathcal{S}_{2,g,n_{2,g}}} \hat{\gamma}_{\text{FDH}}(X_i, Y_i \mid Z_i = g, \mathcal{S}_{2,g,n_{2,g}}) \right). \quad (4.2)$$

The estimates under the summation sign in (4.1) are unconditional in the sense that observations in $\mathcal{S}_{1,n_1}$ may be from any of the G groups. The estimates under the second summation sign in (4.2) are conditional on membership in group g , i.e., the summation is only over observations in group g , and only the observations in group g are used to compute efficiency estimates corresponding to observations in group g . But, the estimator $\hat{\gamma}_{\text{FDH}}(X_i, Y_i \mid Z_i = g, \mathcal{S}_{2,g,n_{2,g}})$ is identical to the unconditional estimator defined in (2.15) since the conditioning variable Z_i is on a discrete random variable.

Under mild assumptions, the sample means $\hat{\mu}_{1,n_1}$ and $\hat{\mu}_{2,n_2}$ defined by (4.1) and (4.2) are consistent estimators of $E_{X,Y}(\gamma(X, Y))$ and $E_{X,Y,Z}(\gamma(X, Y \mid Z))$, respectively, where $E_{X,Y}(\cdot)$ denotes expectation with respect to X and Y and $E_{X,Y,Z}(\cdot)$ denotes expectation with respect to X , Y and Z .

Now define $\xi = E_{X,Y,Z}(\gamma(X,Y | Z)) - E_{X,Y}(\gamma(X,Y))$. Since for all (x,y,z) , $\Psi^z \subseteq \Psi$, we have $\gamma(x,y | z) \geq \gamma(x,y)$, by construction. So that $\xi \geq 0$ with equality to 0 if and only if the (separability) Assumption 2.1 holds. Hence, a natural test statistic for testing separability can be defined by looking to the difference $\hat{\mu}_{2,n_2} - \hat{\mu}_{1,n_1}$, a consistent estimator of ξ , and rejecting the null hypothesis of separability if this difference is too “large” (the distribution of the difference is needed in order to know what constitutes “large.”)

However we now that naïve application of the normal machinery is wrong and that we must use the CLTs developed above in Theorem 3.2, to construct a test statistics that has a known distribution under the null. This is rather easy because the two samples $\mathcal{S}_{1,n_1}$ and $\mathcal{S}_{2,n_2}$ provide independent estimators. The same is true for the G groups inside the sample $\mathcal{S}_{2,n_2}$. The first sample provides the estimates $\hat{\mu}_{1,n_1}$, $\hat{B}_{1,n_1,\kappa}$ and $\hat{\sigma}_{1,n_1}^2$ as described above.

For the conditional measures in sample $\mathcal{S}_{2,n_2}$, we have to compute the bias correction estimator within each groups $\mathcal{S}_{2,g,n_{2,g}}$, $g = 1 \dots, G$. To be specific, we first notice that $\hat{\mu}_{2,n_2}$ in (4.2) can be written as a weighted average of the mean within each group

$$\hat{\mu}_{2,n_2} = n_2^{-1} \sum_{g=1}^G n_{2,g} \hat{\mu}_{2,g,n_{2,g}}, \quad (4.3)$$

where $\hat{\mu}_{2,g,n_{2,g}} = n_{2,g}^{-1} \sum_{(X_i,Y_i,Z_i) \in \mathcal{S}_{2,g,n_{2,g}}} \hat{\gamma}_{\text{FDH}}(X_i, Y_i | Z_i = g, \mathcal{S}_{2,g,n_{2,g}})$. For each group g the jackknife procedure described above produces an estimate of the bias correction $\hat{B}_{2,g,n_{2,g},\kappa}$, so that the global bias correction for the conditional mean in the sample $\mathcal{S}_{2,n_2}$ is given by

$$\hat{B}_{2,n_2,\kappa} = n_2^{-1} \sum_{g=1}^G n_{2,g} \hat{B}_{2,g,n_{2,g},\kappa}. \quad (4.4)$$

Similarly within each group g , the consistent estimator of the variance is given by

$$\hat{\sigma}_{2,g,n_{2,g}}^2 = n_{2,g}^{-1} \sum_{(X_i,Y_i,Z_i) \in \mathcal{S}_{2,g,n_{2,g}}} (\hat{\gamma}_{\text{FDH}}(X_i, Y_i | Z_i = g, \mathcal{S}_{2,g,n_{2,g}}) - \hat{\mu}_{2,g,n_{2,g}})^2, \quad (4.5)$$

so that (each sample $\mathcal{S}_{2,g,n_{2,g}}$ being independent of each other), a consistent estimator of the variance of the mean defined in (4.3) is given by

$$n_2^{-1} \hat{\sigma}_{2,n_2}^2 = n_2^{-1} \sum_{g=1}^G n_{2,g} \hat{\sigma}_{2,g,n_{2,g}}^2. \quad (4.6)$$

Now we can adapt the procedure described in Kneip et al. (2016) for testing the equality of two means from two independent samples. Since we use here the FDH estimator, applying

Theorem 3.2 for both means, we have the following test statistics for testing the separability for discrete Z . If $p + q \leq 3$, then under the null hypothesis

$$T_{1,n} = \frac{\hat{\mu}_{2,n_2} - \hat{\mu}_{1,n_1} - (\hat{B}_{2,n_2,\kappa} - \hat{B}_{1,n_1,\kappa})}{\sqrt{\frac{\hat{\sigma}_{1,n_1}^2}{n_1} + \frac{\hat{\sigma}_{2,n_2}^2}{n_2}}} \xrightarrow{\mathcal{L}} N(0, 1) \quad (4.7)$$

as $n \rightarrow \infty$. Similarly, if $\kappa < 1/2$ we should rather use the result

$$T_{2,n} = \frac{\hat{\mu}_{2,n_{2,\kappa}} - \hat{\mu}_{1,n_{1,\kappa}} - (\hat{B}_{2,n_{2,\kappa}} - \hat{B}_{1,n_{1,\kappa}})}{\sqrt{\frac{\hat{\sigma}_{1,n_{1,\kappa}}^2}{n_{1,\kappa}} + \frac{\hat{\sigma}_{2,n_{2,\kappa}}^2}{n_{2,\kappa}}}} \xrightarrow{\mathcal{L}} N(0, 1), \quad (4.8)$$

as $n \rightarrow \infty$, where $n_{1,\kappa} = \lfloor n_1^{2\kappa} \rfloor$ and $n_{2,\kappa} = \sum_{g=1}^G n_{2,g,\kappa}$ with $n_{2,g,\kappa} = \lfloor n_{2,g}^{2\kappa} \rfloor$ for $g = 1, \dots, G$. In both cases, we reject the null hypothesis of separability if the test statistic is too large. So, for a test of level α , we reject whenever $1 - \Phi(T_{1,n})$, or $1 - \Phi(T_{2,n})$ depending on the value of $p + q$, is less than α .

4.2.2 The Continuous Case

When Z is continuous, the test proposed by Daraio et al. (2018) follows lines similar to those for the case when Z is discrete. But here, the test exploits the results of Theorem 3.3 given above for the mean of conditional efficiency scores. The splitting of the sample $\mathcal{S}_n$ into two random sub-samples $\mathcal{S}_{1,n_1}$ and $\mathcal{S}_{2,n_2}$ of sizes n_1 and n_2 such that $n_1 = \lfloor n \rfloor$ and $n_2 = n - n_1$ is the same as above. The important point is that the two subsamples are independent. As above, we can define $\xi = E_{X,Y,Z}(\gamma(X, Y \mid Z)) - E_{X,Y}(\gamma(X, Y))$. Then $\xi \geq 0$, with strict equality holding if and only if the (separability) Assumption 2.1 holds. Hence, a natural test statistic for testing separability is based on the difference $\hat{\mu}_{2,c,n_2} - \hat{\mu}_{1,n_1}$, a consistent estimator of ξ , implying rejection of the null hypothesis of separability if this difference is too large.

As above using $\hat{\mu}_{1,n_1}$ or $\hat{\mu}_{1,n_{1,\kappa}}$ according the value of κ , we can apply Theorem 3.2 with the subsample $\mathcal{S}_{1,n_1}$. For the independent subsample $\mathcal{S}_{2,n_2}$ we can use $\hat{\mu}_{2,c,n_{2,h}}$ or $\hat{\mu}_{2,c,n_{2,h,\kappa}}$ adapted from the definition in (3.12) or (3.13), according to the value of κ , and apply Theorem 3.3. After some derivations, accounting for bias corrections and variance estimation, Daraio et al. (2018) develop the following test statistics. Using FDH estimators, if $p + q \leq 3$, then under

the null hypothesis of separability

$$T_{3,n} = \frac{\hat{\mu}_{2,c,n_{2,h}} - \hat{\mu}_{1,n_1} - (\hat{B}_{2,n_{2,\kappa}}^c - \hat{B}_{1,n_1,\kappa})}{\sqrt{\frac{\hat{\sigma}_{1,n_1}^2}{n_1} + \frac{\hat{\sigma}_{2,c,n_2}^{2,h}}{n_{2,h}}}} \xrightarrow{\mathcal{L}} N(0,1) \quad (4.9)$$

as $n \rightarrow \infty$. Alternatively, if $\kappa < 1/2$ then

$$T_{4,n} = \frac{\hat{\mu}_{2,c,n_{2,h,\kappa}} - \hat{\mu}_{1,n_{1,\kappa}} - (\hat{B}_{2,n_{2,\kappa}} - \hat{B}_{1,n_{1,\kappa}})}{\sqrt{\frac{\hat{\sigma}_{1,n_1}^2}{n_{1,\kappa}} + \frac{\hat{\sigma}_{2,c,n_2}^{2,h}}{n_{2,h,\kappa}}}} \xrightarrow{\mathcal{L}} N(0,1), \quad (4.10)$$

$n \rightarrow \infty$ under the null hypothesis of separability, where $n_{1,\kappa} = \lfloor n_1^{2\kappa} \rfloor$ and $n_{2,h,\kappa} = \lfloor n_{2,h}^{2\kappa} \rfloor$. In either case we reject the null hypothesis of separability if the test statistic is too large. So, for a test of size α , the null hypothesis is rejected whenever $1 - \Phi(T_{2,n})$ or $1 - \Phi(T_{3,n})$ (depending on the value of $p + q$) is less than α .

4.3 Testing Convexity

As discussed above, the FDH estimators do not require convexity of the production set Ψ and are consistent estimator under the milder assumption of free disposability. By contrast, the VRS-DEA estimators require both convexity of the production set as well as free disposability. So the idea for testing convexity versus non-convexity of Ψ is to compare the average of the efficiency scores estimated by FDH (say $\hat{\mu}_{\text{FDH}}$) with those estimated by VRS-DEA (say $\hat{\mu}_{\text{DEA},n}$). By construction, $\hat{\gamma}_{\text{DEA}}(X_i, Y_i \mid \mathcal{S}_n) \leq \hat{\gamma}_{\text{FDH}}(X_i, Y_i \mid \mathcal{S}_n)$ and hence $\hat{\mu}_{\text{DEA},n} \leq \hat{\mu}_{\text{FDH}}$. Under the null hypothesis that Ψ is convex, the two averages estimate the same quantity, so we expect this difference to be “small” if Ψ is convex. Alternatively, if Ψ is not convex, we should expect this difference to be “large,” suggesting that the null hypothesis of convexity should be rejected.

Recall that Theorem 3.2 establishes CLT results for both of the means $\hat{\mu}_{\text{DEA},n}$ and $\hat{\mu}_{\text{FDH},n}$. But, as discussed by Kneip et al. (2016), Theorem 3.2 cannot be used directly to provide the limiting distribution of the difference between the two means. This is due to the fact that the same observations in $\mathcal{S}_n$ are used to compute the efficiency estimates in both means, and the scaled difference $n^\nu (\hat{\mu}_{\text{FDH}} - \hat{\mu}_{\text{DEA},n})$ would converge to a degenerate distribution for any $\nu \leq 1/2$ due to the dependence between the two estimators.

So, as with the separability test described above, the solution proposed by Kneip et al. (2016) is to randomly split the sample $\mathcal{S}_n$ into two independent, mutually exclusive and collectively exhaustive parts $\mathcal{S}_{n_1}$ and $\mathcal{S}_{n_2}$, with $n_1 + n_2 = n$, and then compute the efficiency estimators and their average using VRS-DEA for $\mathcal{S}_{n_1}$ and by using the FDH estimators for $\mathcal{S}_{n_2}$.⁹ With the hyperbolic distances, this will lead to two independent means

$$\hat{\mu}_{\text{DEA},n_1} = n_1^{-1} \sum_{(X_i, Y_i) \in \mathcal{S}_{n_1}} \hat{\gamma}_{\text{DEA}}(X_i, Y_i \mid \mathcal{S}_{n_1}), \quad (4.11)$$

and

$$\hat{\mu}_{\text{FDH},n_2} = n_2^{-1} \sum_{(X_i, Y_i) \in \mathcal{S}_{n_2}} \hat{\gamma}_{\text{FDH}}(X_i, Y_i \mid \mathcal{S}_{n_2}). \quad (4.12)$$

Under the null hypothesis that Ψ is convex, these two averages are consistent estimators of $E_{XY}(\gamma(X, Y))$ so the difference $\hat{\mu}_{\text{FDH},n_2} - \hat{\mu}_{\text{DEA},n_1}$ is expected to be, small but under the alternative it will be large. So a natural test statistics is based on this difference, and we reject the null hypothesis of convexity if it is sufficiently large.

Applying the CLTs in Theorem 3.2, and the bias estimates and the variance estimates from the two independent subsamples, Kneip et al. (2016) obtain the following test statistics depending on the dimension $p + q$ of the problem. Under the null hypothesis of convexity, provided $p + q \leq 3$,

$$T_{5,n} = \frac{\hat{\mu}_{\text{FDH},n_2} - \hat{\mu}_{\text{DEA},n_1} - (\hat{B}_{\text{FDH},n_2,\kappa_2} - \hat{B}_{\text{DEA},n_1,\kappa_1})}{\sqrt{\frac{\hat{\sigma}_{\text{DEA},n_1}^2}{n_1} + \frac{\hat{\sigma}_{\text{FDH},n_2}^2}{n_2}}} \xrightarrow{\mathcal{L}} N(0, 1). \quad (4.13)$$

as $n \rightarrow \infty$. Alternatively, if $p + q \geq 3$ and letting $\kappa = \kappa_2$,

$$T_{6,n} = \frac{\hat{\mu}_{\text{FDH},n_2,\kappa} - \hat{\mu}_{\text{DEA},n_1,\kappa} - (\hat{B}_{\text{FDH},n_2,\kappa_2} - \hat{B}_{\text{DEA},n_1,\kappa_1})}{\sqrt{\frac{\hat{\sigma}_{\text{DEA},n_1}^2}{n_{1,\kappa}} + \frac{\hat{\sigma}_{\text{FDH},n_2}^2}{n_{2,\kappa}}}} \xrightarrow{\mathcal{L}} N(0, 1), \quad (4.14)$$

as $n \rightarrow \infty$, where $n_{\ell,\kappa} = \lfloor n_\ell^{2\kappa} \rfloor$ and for $\ell = 1, 2$ and as in Theorem 3.2, the means are computed over a random subsample of size $n_{\ell,\kappa}$, $\ell = 1, 2$. As for the separability test, for a test of size α , we reject the null hypothesis of convexity of Ψ whenever $1 - \Phi(T_{5,n})$, or $1 - \Phi(T_{6,n})$ (depending on the value of $p + q$) is less than α .

⁹Note that here the rates of convergence of the two estimators are different, with we have $\kappa_1 = 2/(p+q+1)$ for the VRS-DEA and $\kappa_2 = 1/(p+q)$ for the FDH. Kneip et al. (2016) suggest splitting the sample $\mathcal{S}_n$ unevenly into parts of size n_1 and $n_2 = n - n_1$ such that $n_1^{\kappa_1} = n_2^{\kappa_2}$, so that the achieved precision of the estimators in the two samples will be of the same order.

4.4 Multiple Splittings

The test of separability for continuous Z described in Section 4.2.2 and the test for convexity described in Section 4.3 both require randomly splitting the sample $\mathcal{S}_n$ into two independent parts $\mathcal{S}_{n_1}$ and $\mathcal{S}_{n_2}$. The independence so created is crucial for obtaining non-degenerate limiting (normal) distributions of the test statistics. Some researchers may feel uncomfortable with the ambiguity or the arbitrariness of such a single random split, and different results may be obtained by different random splits of the sample. But, the results obtained from one split are not independent of results obtained from another split since the same data are involved. Nonetheless, Simar and Wilson (2020) propose a way to remove the ambiguity by in effect “integrating out” the uncertainty surrounding a single random split. The idea is to repeat many times the random splitting, say J times, and to exploit the information obtained in each of these splits while dealing with the dependence across splits.

For purposes of discussion, suppose we use a statistic T_n (T_n can represent any of the test statistics described above). For each split $j = 1, \dots, J$ we have an observed value $T_{j,n}$ and a p -value $p_j = 1 - \Phi(T_{j,n})$. As explained by Simar and Wilson (2020), if the J splits were independent, the average of the values of T_n

$$\bar{T}_n = J^{-1} \sum_{j=1}^J T_{n,j}, \quad (4.15)$$

would give a stabilized and less arbitrary indication of the credibility of the null hypothesis, with a large value of $\bar{T}_n$ indicating that the null should be rejected. Similarly, if the null is true and if all the splits were independent, the distribution of the p -values p_j , $j = 1, \dots, J$ would have, by the probability integral transform, a uniform distribution on $[0, 1]$ and so the under the null, the Kolmogorov distance between the empirical distribution of these p -values and a uniform distribution would be small. This distance is given by

$$K_n = \sup_{u \in [0,1]} \{ | \hat{F}_{J,p}(u) - u | \}, \quad (4.16)$$

where $\hat{F}_{J,p}(u)$ is the empirical distribution function of the values p_j , $j = 1, \dots, J$. If the null is true and if observed p -values were independent, then K_n would behave as a one-sample Kolmogorov-Smirnov statistics, with rejection of the null if it is too large.

However the J splits are *not* independent and so we cannot easily use the information contained in either $\bar{T}_n$ or K_n . Hence Simar and Wilson (2020) suggest using the bootstrap

to approximate the unknown distribution of $\bar{T}_n$ and of K_n under the null hypothesis. The details of how to generate bootstrap samples are given in Simar and Wilson (2020); the important point is that the bootstrap samples must be generated under the conditions of the null hypothesis that is to be tested. Splitting Each bootstrap sample $b = 1, \dots, B$ produces, by applying the J splits, J times yields sets of values $\{\bar{T}_n^{*,(b)}\}_{b=1}^B$ and $\{K_n^{*,(b)}\}_{b=1}^B$ of the statistics of interest in the bootstrap world. Hence, the bootstrap estimates of the p -values of the observed $\bar{T}_n$ and K_n are given by

$$\hat{p}_T = \frac{\#\{\bar{T}_n^{*,(b)} \geq \bar{T}_n\}}{B} \quad (4.17)$$

and

$$\hat{p}_K = \frac{\#\{K_n^{*,(b)} \geq K_n\}}{B}, \quad (4.18)$$

respectively. For a test of size α , we reject the null if these bootstrap estimates less than α . Simar and Wilson (2020) show under various Monte-Carlo scenarios that the two bootstrapped p -values are often very similar and that this approach gives even better results than applying an arbitrary single split, at a moderate cost of numerical burden when analyzing a given sample of size n . In practice, from their Monte-Carlo experiments, it appears that $J = 100$ and $B = 1,000$ give stable results.

5 Data and Variable Definitions

5.1 Variable Specification and Construction

As noted in Section 1, we revisit the studies by Burgess and Wilson (1995, 1996 and 1998) to show how recently-developed statistical tools can be used to analyze the performance of health-care providers (in this case, hospitals). For this purpose, we use a superset of the data used by Burgess and Wilson (1996, 1998). The data used here consist of an unbalanced panel of annual observations on U.S. hospitals for 1985–1988.¹⁰

The hospitals represented in our data have four different ownership types. These include “VA” hospitals operated by the U.S. Department of Veterans Affairs (formerly, the Veterans

¹⁰Burgess and Wilson (1995, 1998) omit observations for hospitals that are not observed in each year. Since this is an unnecessary loss of information, we use all available observations in each year. Burgess and Wilson (1996) use only data for 1988.

Administration, which became a cabinet-level department with enactment of the Department of Veterans Affairs Act of 1988), hospitals operated by state and local governments (LGOVT), hospitals operated by religious, charitable or other non-profit (NONPROFIT) organizations, and hospitals operated by for-profit (FORPROFIT) firms. Among these four types, NONPROFIT hospitals form the largest group in each year. Table 1 gives the number of hospitals in our data for each year by ownership type.

The data are taken from several sources, including VA financial accounting systems, the VA national patient discharge and outpatient data files, VA nursing home discharge files, the annual survey of hospitals by the American Hospital Association (AHA), and the Medicare Cost Reports submitted to the federal government by all hospitals receiving Medicare reimbursements. Data from the various VA files correspond to federal fiscal years 1985–88 which run from October 1 through September 30 in each year. Unfortunately, for some non-VA hospitals, data from the AHA survey and the Medicare Cost Reports do not correspond to federal fiscal years; the beginning (and end) of reporting periods may vary by as much as nine months across hospitals, although the most common starting date is October 1, followed by January 1 and July 1. For purposes of analysis in cases where hospitals’ reporting periods do not correspond to the federal fiscal years, data are taken from the reporting period beginning between April 1, 1984 and March 31, 1985 and continuing in turn for each of the other study years. Consequently, any temporal misalignment in the data does not exceed six months in either direction.

As discussed by Burgess and Wilson (1995), the data are initially limited to 137 non-psychiatric hospitals in the VA system during 1985–1988, excluding independent outpatient clinics, and non-VA hospitals with an average of at least 100 beds over the period 1985–1988. All VA hospitals have more than 100 beds during the period covered by the data. Moreover, the technology and scope of services offered in small hospitals is generally very different from that of larger hospitals. For example, non-VA hospitals with fewer than 100 beds are allowed to have swing beds between acute and long-term care, leading to different technology constraints. Since all non-psychiatric VA hospitals have at least 100 beds during 1985–1988, Burgess and Wilson (1995, 1996, 1998) omit non-VA hospitals with fewer than 100 beds in order to compare VA hospitals with those which have alternative ownership structures. In general, the operation of VA hospitals is characterized by a significantly higher proportion of

outpatient to in-patient care, a higher psychiatric workload, and more long-term care than typically found in non-VA hospitals.¹¹ After omitting observations with missing values or implausible values, we have for this study 7,327 observations, with 1,588, 2,058, 1,921 and 1,960 observations for each year 1985–1988, respectively.¹²

Burgess and Wilson (1995, 1996, 1998) specify $p = 7$ inputs and $q = 6$ outputs. The inputs include the number of acute-care hospital beds, weighted by a scope-of-services index (*BDH*); the number of long-term hospital beds (*BDLT*); registered nurses (RNs) measured in full time equivalents (*RNFTE*); licensed practical nurses (LPNs) measured in full time equivalents (*LPFTE*); other clinical labor (excluding RNs and LPNs) measured in full time equivalents (*XCFTE*); non-clinical labor measured in full time equivalents (*NCFTE*); and long-term care labor measured in full-time equivalents (*LTfTE*). The labor inputs are taken from VA records for VA hospitals and from the AHA survey for all other hospitals. Full time equivalents are computed by counting full time employees as 1 full time equivalent and part time employees as 0.5 full time equivalent.

The six distinct, direct hospital outputs include acute care in-patient days (*HIND*); case-mix weighted acute care in-patient discharges (*HDIS*); long-term care in-patient days (*LTIND*); number of outpatient visits (*OPVIS*); ambulatory surgical procedures (*SAM*); and in-patient surgical procedures (*SIP*). Sample variances indicate that each of the variables has a good deal of variation within each year, reflecting the dispersion of sizes of hospitals in the sample. The surgical procedure variables (*SAM* and *SIP*) and the long-term care variable (*LTIND*) have minimum values of zero for some hospitals in each of the four years. Some of the smaller hospitals in the sample perform no surgical procedures and some of the hospitals do not offer long-term care.

Burgess and Wilson (1995) note that variation in the intensity of care given to patients may reflect differences in quality across hospitals. To control for this, the number of acute care in-patient discharges is multiplied by the Medicare case-mix index to obtain the variable

¹¹Table 2 gives the proportion of hospitals providing long-term care by type and by year. The results in Table 2 reveal that more than half of VA hospitals provide long-term care, while between 10 and 20 percent of LGOVT and NONPROFIT hospitals and less than 7 percent of FORPROFIT hospitals provide long-term care in each year.

¹²See Burgess and Wilson (1995, 1996, 1998) for additional details on the data. Burgess and Wilson (1995, footnote no. 14 discuss use of the methodology in Wilson (1995) to delete potential outliers from their final sample, but this is not used here.

HDIS. Burgess and Wilson (1995) further discuss how case-mix severity is widely known to contribute to hospital costs, and failure to account for differences in case mix across hospitals could lead to biased estimates of technical efficiency. The Medicare case-mix index is an index for the average resource intensiveness of the average Medicare discharge in a given hospital. For the non-VA data, this information is obtained from the Medicare Cost Reports. VA facilities are not reimbursed under Medicare rules; hence, the case-mix index for VA hospitals is calculated using VA data on patients aged 65 years or older (a detailed description of these calculations appears in U.S. Department of Veterans Affairs, 1993).

In addition, Burgess and Wilson (1995) observe that physicians and dentists in VA hospitals are salaried employees of the hospital, which is uncommon among non-VA hospitals. The input of physicians in the hospital production process can be evaluated and compared to the medical staff measures used in existing private sector studies. Burgess and Wilson (1995) note that this important input is often ignored in studies of non-VA hospitals because no reliable data exist and cite *e.g.*, see Conrad and Strauss (1983) and Sherman (1984) as examples. Others such as Grosskopf and Valdmanis (1987) include a variable for the number of physicians, but this number adds active and associate physicians to the number of house staff. Many of those included in this heterogeneous total may spend little time performing work directly related to outputs produced by the hospital. Hence, using available data on physicians to construct an additional hospital input variable would bias the measured level of efficiency in favor of VA hospitals by overstating the actual physician input for non-VA hospitals. Consequently, physicians are omitted from the set of inputs. This issue is revisited below in Section 5.2.

As is common in studies of hospital efficiency, *BDH* and *BDLT* are included as proxies for capital used by the hospitals. Direct measurement of capital in the hospital industry is problematic; in addition to the usual problems with using book value of plant and equipment, accounting practices vary widely across the four ownership types. Data on beds are obtained from the AHA survey for all hospitals using the AHA definition of beds that are set up, staffed, and ready for use at the end of the reporting period. The intensity of acute care in hospitals can vary enormously and is related to capital differences among hospitals (by comparison, the intensity of long-term care is much less variable across hospitals). Furthermore, technological capabilities might serve as a proxy for quality, or at least the perception

of quality on the part of patients. Thus, we use an index of the scope of services offered by hospitals to weight the number of acute care beds in computing BDH . The scope of services index was constructed separately for each year of the analysis using the methodology described by Klastorin and Watts (1982), who present evidence that the order in which hospitals acquire services and facilities tends to be systematic and hierarchical. For example, applying their methodology and using the service availability data from the 1988 AHA survey yielded 24 well-defined groups with a Guttman scale coefficient of reproducibility (McIver and Caremines, 1981) of 0.89 as compared to the ideal 1.00 of a perfect hierarchy.

In addition to the input and output variables described above, Burgess and Wilson (1995) consider several environmental variables. These include the discrete variables VA , equal to 1 for VA hospitals and 0 otherwise; $LGOVT$, equal to 1 for state and local-government hospitals, and 0 otherwise; $NONPROFIT$, equal to 1 for non-profit hospitals and 0 otherwise; and $FORPROFIT$, equal to 1 for for-profit hospitals and 0 otherwise. In addition, Burgess and Wilson (1995) consider six continuous environmental variables, including administrative cost per bed-day of care ($AMCDAY$); the ratio of clinical staff FTE to non-clinical FTE ($CNCFTE$); Herfindahl index of competition using the county as market ($HERF$); the average length of stay, computed as in-patient days divided by discharges (LOS); the percentage of nurses who are registered nurses ($PCTR_N$); and the Health Care Financing Administration regional medical wage index (WI).

Summary statistics for the input, output, and continuous environmental variables are shown in the first 19 lines of Table 3.

5.2 Reducing Dimensionality

Each of the Burgess and Wilson (1995, 1996 and 1998) studies use the same variable specification, with $p + q = 13$. As noted earlier, Burgess and Wilson examine only data for one year, with 2,246 observations, and Burgess and Wilson (1995, 1998) estimate efficiency separately for each year and use balanced panels; Burgess and Wilson (1998) use 1,545 observations in each year, while Burgess and Wilson (1995) use 1,480 observations in each year. As discussed earlier, little was known about the properties of nonparametric efficiency estimators at the time of the Burgess and Wilson studies. However, today it is well known that DEA and FDH estimators suffer from the “curse of dimensionality.” This means that for

a given sample size, estimation error should be expected to increase as dimensionality (i.e., the sum of the numbers of inputs and outputs) increases. Even with 2,246 observations in Burgess and Wilson (1996), the effective parametric sample size defined by Wilson (2018) is approximately 9 for VRS-DEA estimators, and only approximately 3 if FDH estimators are used. With only 1,480 observations in Burgess and Wilson (1995), the effective parametric sample sizes are approximately 4 and 2 for VRS-DEA and FDH estimators (respectively). In other words, the nonparametric VRS-DEA estimators in the Burgess and Wilson studies yield estimation error *of the same order* as parametric estimators (in correctly specified models) converging at the usual root- n rate with only 4 to 9 observations.

Further evidence on the problem of large dimensionality is found by comparing efficiency estimates obtained by FDH, VRS-DEA and CRS-DEA estimators. Table 4 shows the proportions of FDH, VRS-DEA and CRS-DEA hyperbolic efficiency estimates that are equal to one by year and by type when using the original data with $p + q = 13$ dimensions. The FDH estimator yields estimates equal to one for more than 90 percent of the observations in each case. As expected, the VRS-DEA estimator yields fewer estimates equal to one in each case, and the CRS-DEA estimator yields even fewer estimates equal to one. When the data are pooled across types and years, the FDH estimator yields ones for more than 92 percent of the observations, while both the VRS-DEA and CRS-DEA estimators produce estimates equal to one for fewer than 10 percent of the observations. This casts doubt on convexity of the production set Ψ and perhaps also on separability with respect to years and types since the proportion of estimates equal to one obtained with the VRS-DEA estimator is much higher, 24.9 to 75.0 percent, when the data are separated into groups by year and type.

As discussed by Daraio and Simar (2007) and Wilson (2018), production data are often closely related, and this is the case with the hospital data used here. However, U.S. hospitals provide both acute care and long-term care, and the inputs used by Burgess and Wilson include two inputs ($BDLT$, $LTFTE$) and an output ($LTIND$) related to long-term care, while the remaining inputs and outputs are related to hospitals' production of acute care. Moreover, hospitals' provision of acute care is rather different from their provision of long-term care in the sense that different resources are used. We first divide each input variable by its sample standard deviation. Next, we arrange the standardized observations on the input quantities related to acute care (i.e., BDH , $RNFTE$, $LPFTE$, $XCFTFTE$ and $NCFTE$)

into the $7,527 \times 5$ matrix $\mathbf{X}$ and then compute the (raw) moment matrix $\mathbf{X}'\mathbf{X}$. We then compute eigenvalues of this moment matrix; computing the ratio of the largest eigenvalue to the sum of the eigenvalues yields 0.8925. Similar computations using the two long-term care inputs *BDLT* and *LTFTE* yield a ratio of 0.9275. These values are well within the guidelines given by the simulation results of Wilson (2018), and hence we reduce the number of inputs to $p = 2$ by using the first principle components based on moment matrices for the five acute-care inputs and the two long-term care inputs. These principle components are denoted by X_1^* and X_2^* in Table 3 where their summary statistics are displayed.¹³

We similarly standardize the six output variables. We then arrange the five that are related to acute-care provision, namely *HIND*, *HDIS*, *OPVIS*, *SAM* and *SIP* into a $7,527 \times 5$ matrix $\mathbf{Y}$ and compute the moment matrix $\mathbf{Y}'\mathbf{Y}$. Computing the eigenvalues of this moment matrix, and then computing the ratio of the largest eigenvalue to the sum of the eigenvalues yields a value 0.8778. Again following the guidelines of Wilson (2018), we retain the standardized long-term output *LTIND* (Y_1^*) and replace the acute-care outputs *HIND*, *HDIS*, *OPVIS*, *SAM* and *SIP* with the first principle component (Y_2^*) based on the corresponding 5×5 moment matrix described above resulting in a reduced-dimensional output space with $q = 2$. Summary statistics for these are given in Table 3.¹⁴

Recall from the discussion in Section 5.1 that the input variables defined in the Burgess and Wilson studies and used here do not include the input of physicians since this is not observed for non-VA hospitals in the U.S. However, it is reasonable to believe that physician input is likely closely related with the acute-care input variables that are included. Consequently, it seems plausible that little information is excluded by the absence of physicians' input, particularly since we use reduced-dimensional data for estimation.

For purposes of testing separability, we standardize each of the continuous environmental variables defined in Section 5.1 by dividing the variables by their sample standard deviations. Arranging the standardized variables into the $7,527 \times 6$ matrix $\mathbf{Z}$ and then computing the

¹³Computing eigenvalues of the 7×7 raw moment matrix obtained using all of the standardized input variables yields 0.7718 for the ratio of the largest eigenvalue to the sum of the eigenvalues. In view of the simulation results obtained by Wilson (2018), one should hesitate to replace all of the inputs and outputs by their first principle components in order to reduce dimensionality. Moreover, this confirms that hospitals' provision of acute care is rather different from their provision of long-term care.

¹⁴Computing the 6×6 moment matrix based on all six of the output variables, and then computing eigenvalues, yields a value of 0.7718 for the ratio of the largest eigenvalue to the sum of the eigenvalues. This confirms again that long-term care is rather different from acute-care provision.

ratio of the largest eigenvalue to the sum of the eigenvalues of the moment matrix $\mathbf{Z}'\mathbf{Z}$ yields 0.9578, indicating that the six continuous environmental variables are highly collinear. We compute their first principle component based on $\mathbf{Z}'\mathbf{Z}$ and denote this by Z^* . Summary statistics for this variable appear in the last row of Table 3.

To re-cap, we reduce the dimensionality of the input-output space from 13 to 4 by exploiting the collinearity among subsets of inputs and outputs. The simulation results provided by Wilson (2018) suggest that using the reduced-dimensional data is likely to reduce estimation error from the level that would result from using the original, 13-dimensional input-output data. Of course, some information is lost by reducing dimensionality, but the eigen-system analysis described above suggests that the loss is minimal compared to what is gained by reducing dimensionality from 13 to 4 dimensions. In particular, with 13 dimensions, the FDH and VRS-DEA estimators converge at rates $n^{1/13}$ and $n^{1/7}$ (respectively), but with only 4 dimensions, the rates are $n^{1/4}$ and $n^{2/5}$. With the reduced-dimensional data, the rate of the VRS-DEA estimator is not far from the root- n parametric rate, but whether it should be used depends on whether the production set Ψ is convex.

All of our estimation results discussed below in Section 6 are obtained using the reduced-dimensional data. Before turning to the results, note that Table 5 gives the number of observations with efficiency estimates equal to one using hyperbolic FDH, VRS-DEA and CRS-DEA estimators, analogous to Table 4. In every case, the proportion of observations with efficiency estimates equal to one is reduced substantially by using dimension-reduction. This is important for two reasons. First, the hyperbolic estimators are biased upward, and dimension reduction reduces some of the bias, especially at the frontier. Second, under the assumptions of the model, there is no probability mass along the frontier, and efficiency equal to one is a negligible, measure-zero event. Although a non-zero proportion of the observations in each case have efficiency estimates equal to one even with dimension reduction, the reduction from 13 to 4 dimensions moves us in the right direction. The remaining bias is accounted for when making inference below in Section 6.

6 Estimation Results

6.1 Tests of Separability with Respect to Year and Type

As noted above in Section 5.2, we use the reduced-dimensional data with $p = q = 2$ for all of our estimation and inference. A goal of the Burgess and Wilson studies was to compare the efficiencies of hospitals across types and years, but as discussed in Section 1, not statistical tools were available at the time of the original studies. Here, we take a methodical approach and employ various statistical results that have been developed since the time of the original Burgess and Wilson papers. In the tests of separability and convexity described below, we use 1,000 sample-splits and 1,000 bootstrap replications. We refer to tests using averaged statistics across splits as in (4.15) as “Test #1,” and to test using Kolmogorov-Smirnov statistics along the lines of (4.16) as “Test #2” (recall the discussion in Section 4.4 for details).

Given that hospitals are observed across four different years and four different ownership types, a basic question is whether data can be pooled across years or types. Burgess and Wilson pooled data across types, but not across years, when estimating efficiency. But if the frontiers faced by different types are different, then separability (i.e., Assumption 2.1) does not hold. With four years and four types, we have $G = 16$ groups of hospitals. Since we do not know whether the production set for any year-type is convex, we use hyperbolic FDH efficiency estimators to test separability with respect to both year and type.¹⁵ The results of this test (see Section 4.2.1 for details) are given in the first row of Table 6, where the estimated p -values for both Tests #1 and #2 are 0.000. Hence separability with respect to year *and* type is soundly rejected.

In order to investigate why separability is rejected with respect to year and type, we next test for separability with respect to year, for each of the four ownership types. The results of these tests are shown in the second part of Table 6, and again, the estimated p -values are 0.000 in every case. We then test separability with respect to type for each year. The results of these test are given in the third part of Table 6, and as in the other cases, the estimated p -values are 0.000 in each case. These tests confirm that rejection of the null when we test for separability with respect to both year and type occurs not because of non-separability with respect to only years, or because of non-separability with respect to only type, but is

¹⁵We test for separability with respect to year and type before testing for convexity in order to determine whether we might reasonably pool data across years or types when testing convexity of the production set(s).

due to non-separability with respect to *both* year and type. Consequently, for subsequent estimation, data should not be pooled across types nor across years. Instead, efficiency must be estimated separately and independently for each year-type.

6.2 Tests of Separability with Respect to Continuous Environmental Variables

Burgess and Wilson (1995) regress efficiency estimates for hospitals on the six continuous environmental variables *AMCDAY*, *CNCFTE*, *HERF*, *LOS*, *PCTR**N* and *WI* described earlier in Section 5.1, and include fixed effects for time as well as dummy variables. Burgess and Wilson (1995) estimate efficiency by pooling data across types, but not across years. However, given the results in Table 6, efficiency estimates should be computed independently for each year-type. In addition, Burgess and Wilson use VRS-DEA estimators, and we have not yet tested convexity versus non-convexity of the production set for each year-type.

We test separability (Assumption 2.1) versus non-separability (Assumption 2.2) with respect to each of the continuous environmental variables, as well as their first principle component Z^* constructed as described above in Section 5.2. We perform the tests separately and independently for each year-type, and we use hyperbolic FDH estimators. As in the previous separability tests, we use 1,000 sample-splits and 1,000 bootstrap replications. For each year-type, we obtain results for Test #1 and Test #2 as described above. The results of these tests appear in Table 7. All together, there are $2 \times 4 \times 4 \times 7 = 224$ tests, and none reject the null hypothesis of separability. While there are many other problems with the regression analysis in Burgess and Wilson (1995), we find no evidence that lack of separability with respect to the continuous environmental variables is among these problems. However, this is learned only after testing.

We do not pursue the environmental variables farther here. However, if one is curious about the effects of such variables, one should use conditional (as opposed to unconditional) efficiency estimators to gauge their effect, even if separability with respect to the environmental variables is not rejected. Even if separability holds, the results of Kneip et al. (2015) present serious and difficult problems for inference about parameters in any second-stage regression of estimated efficiencies (either FDH or DEA) on environmental variables. At the time when the Burgess and Wilson (1995) was written, none of this was known, and

conditional efficiency estimators would not be developed until later. But today we know much more, and there is no excuse for continuing to do things as they were done many years ago.

6.3 Tests of Convexity

We now turn to the issue of convexity versus non-convexity. The results of the tests of separability with respect to year and type in Table 6 dictate that we must test convexity separately and independently for each year-type. We implement the test described above in Section 4.3, again using 1,000 sample-splits and 1,000 bootstrap replications and Tests #1 and #2 as described above. Results are shown in Table 8. Here, the two tests give conflicting results in a number of cases. Test #1 rejects the null hypothesis of convexity in only 3 cases—for VA hospitals in 1985, 1986 and 1987. But Test #2 rejects the null in each year for VA and LGOVT hospitals, and in 1985, 1986 and 1987 for NONPROFIT and for FORPROFIT hospitals.

Simulation results reported in Simar and Wilson (2020, Tables 5–6) indicate that Tests #1 and #2 have similar size and power. But, tame (i.e., simulated) data typically are better-behaved than wild (i.e., real-world) data. Recall from Table 5 that even with the reduced-dimensional data, many of the FDH efficiency estimates are equal to one. Moreover, when the sample $\mathcal{S}_n$ is split into the two parts $\mathcal{S}_{1,n_1}$ and $\mathcal{S}_{2,n_2}$, a larger proportion of observations in each of the two parts yield efficiency estimates equal to one than in the full sample $\mathcal{S}_n$. Some investigation indicates that this can sometimes cause surprising results in the bootstrap used to estimate p -values for Tests #1 and #2. The theory underlying the tests is valid asymptotically, but finite-sample results may vary as they do here.

Consideration of some basic facts make clear what conclusions should be drawn from the results in Table 8. First, both Test #1 and #2 reject convexity for VA hospitals in three out of four years. This provides compelling evidence against convexity for VA hospitals for 1985–1987. Second, failure to reject the null hypothesis in any statistical test never provides evidence that the null is true. Failure to reject the null can occur for many reasons, including too-few observations, skewness, etc. Test #2 rejects convexity in 14 out of 16 cases, and this also provides compelling evidence against convexity. Third, FDH estimators do not require convexity, but DEA estimators do. DEA estimators are inconsistent if convexity does not

hold (this means that even with an arbitrarily large number of observations, one should have no expectation of obtaining “reasonable” or meaningful estimates if convexity does not hold when using DEA estimators). Consequently, to be conservative in our approach, and to facilitate comparison across year-types, FDH estimators are used for all of the empirical analysis that follows.

6.4 Differences in Efficiency Across Years and Types

Table 9 displays summary statistics for hyperbolic FDH efficiency estimates for each year-type, with efficiency estimated separately and independently for each year-type. Within each year, there is some variation in mean estimates of efficiency across types, but this is influenced in part by differences in sample sizes across types within each year (see Table 1 for the number of observations for each year-type). More interestingly, the results in Table 9 show a substantial decline in mean efficiency from 1985 to 1986 for each of the four types, and a further decline from 1986 to 1987 for each type. From 1987 to 1988, mean efficiency is shown to increase for each type, but the mean levels in 1988 are well-below the corresponding levels in 1985.

To further examine the variations in mean efficiency seen in Table 9, we test for differences in mean efficiency across pairs of years for each hospital type. We use the results in Theorem 3.2 to implement the hyperbolic analog of the test developed by Kneip et al. (2016, Section 3.1.1). The test statistic includes jackknife bias estimates and is asymptotically distributed $N(0, 1)$. Table 10 shows the results of these tests.

Significant and negative (positive) test statistics in Table 10 indicate declines (increases) in mean efficiency from the first year to the second year. For VA hospitals, we find a significant decrease in mean efficiency from 1985 to 1986, and significant increases from 1986 to 1987 and 1987 to 1988. Moreover, among VA hospitals, mean efficiency is significantly higher in 1988 than in either 1985 or 1986. The significant increases between 1986 and 1987 and between 1985 and 1988 is contrary to what is seen in Table 9, but this merely indicates the danger of relying on point estimates without inference. From the discussion in Section 3.1, we know that standard CLTs do not hold here. Moreover, the CLT results in Theorem 3.2 that are used here correct for bias, and this is the reason for the apparent discrepancy between the patterns seen when looking at sample averages alone versus making statistical

inference. It is always important to remember that nothing is learned from point estimates; learning can only occur when inference is made.

Among LGOVT hospitals, we see in Table 10 significant declines in mean efficiency from 1986 to 1987 and between 1986 and 1988. For NONPROFIT hospitals, there are significant reductions in mean efficiency between 1986 and 1988 and from 1987 to 1988. FORPROFIT hospitals show significant declines in mean efficiency from 1985 to 1986 and between 1985 and 1987 as well as between 1985 and 1988, consistent with what is seen in Table 10.

Our results in Tables 9 and 10 are rather different from the results of Burgess and Wilson (1995). In particular, means for the input and output estimates reported in Tables 2–3 of Burgess and Wilson (1995) show different patterns and less variation from year to year than what we find here in Table 9. In addition, Burgess and Wilson (1995, Tables 2–3) indicate that mean efficiency *improved* (i.e., moved close to one) for VA hospitals from 1985 to 1986, and remark (on page 94) that “over this period, the VA was implementing a new budgeting system for its hospitals” and attribute their finding of improved efficiency to this change. While it is true that this change in budgeting system occurred, its effects, at least initially, appear to be opposite of what Burgess and Wilson (1995) found. As discussed elsewhere, the efficiency estimates of Burgess and Wilson (1995) were obtained using VRS-DEA estimators and thus are likely inconsistent. Moreover, data were pooled across types, and the evidence we obtain in Table 6 indicates that FDH estimates would also be inconsistent when data are pooled. Consequently, one should not be surprised that the careful, rigorous statistical analysis employed here overturns some of the conclusions of the earlier study.

Table 11 shows results from similar tests for differences in mean efficiency across hospital types for each year 1985–1988, with positive (negative) values of test statistics indicating greater (lesser) mean efficiency for the second hospital type than the first. The results indicate that on average, VA hospitals have significantly lower efficiency than either LGOVT, NONPROFIT or FORPROFIT hospitals in each of the four years. This result is perhaps not surprising, as VA hospitals are part of a system administered by the U.S. federal government. VA hospitals tend to be larger than other types, and many are teaching hospitals. By contrast, control of the other hospitals types is far more decentralized. Among the other results in Table 11, NONPROFIT hospitals are found to be significantly more efficient on average than LGOVT hospitals in 1987, and FORPROFIT hospitals are found to be significantly

more efficient on average than LGOVT hospitals in 1985. FORPROFIT hospitals have significantly higher mean efficiency than NONPROFIT hospitals in 1985, but the reverse is true in 1986 and 1987.

7 Summary and Conclusions

We describe modern tools now available for evaluation of the performances of firms or productive units, with a focus on Health-Care providers. Recent developments the last 20 years in the statistical and econometric literature on popular nonparametric envelopment estimators (FDH, and DEA) permit one to conduct valid inference about efficiency measures, and also about assumptions on the underlying model that generate observed data. In particular, we describe the set of assumptions (both economic and statistical) needed to justify the use of the various estimators (conditional, unconditional, FDH, VRS-DEA, CRS-DEA, etc.). We describe also the statistical properties of these estimators and indicate how these underlying assumptions can be tested to obtain statistically valid results.

Various procedures are implemented with data on U.S. hospitals used in a series of papers by Burgess and Wilson in appearing in the 1990s. We show how to proceed in practice using these modern tools and provide a useful road map for applied researchers. We indicate also how conclusions regarding the performance of U.S. hospitals change when the modern tools (that were unavailable to Burgess and Wilson in the 1990s) are used. All of the estimates and tests that we compute can be computed using the *FEAR* software package described by Wilson (2008). As such, there is little excuse for not using coherent, valid statistical methods for examining the productive performance of hospitals or other producers.

Burgess and Wilson (1998) estimate changes in productivity, and decompose productivity into changes in efficiency and technology. Unfortunately, their index of productivity change is badly defined, and all estimates in the paper are obtained with input-oriented VRS-DEA estimators which have been shown above in Section 6 to be inappropriate for the data used here and by Burgess and Wilson (1998).

Kneip et al. (2021) provide CLT results for geometric means of Malmquist indices constructed from hyperbolic DEA efficiency estimators that can be used to make inference about productivity change between two points in time. Simar and Wilson (2019) show that the hyperbolic Malmquist index considered by Kneip et al. (2021) can be decomposed into hy-

perbolic measures of change in efficiency, change in technology, etc., provide additional CLT results for such components of productivity change. Both Simar and Wilson (2019) and Kneip et al. (2021) are based on DEA estimators and require convexity of production sets, but work is underway (Kneip et al., 2022) to extend these results to cases where production sets are not convex.

References

- Bădin, L., C. Daraio, and L. Simar (2019), A bootstrap approach for bandwidth selection in estimating conditional efficiency measures, *European Journal of Operational Research* 277, 784–797.
- Burgess, J. F., Jr. and P. W. Wilson (1995), Decomposing hospital productivity changes 1985–1988: A nonparametric Malmquist approach, *Journal of Productivity Analysis* 6, 343–363.
- (1996), Hospital ownership and technical inefficiency, *Management Science* 42, 110–123.
- (1998), Variation in inefficiency among US hospitals, *Canadian Journal of Operational Research and Information Processing (INFOR)* 36, 84–102.
- Casella, G. and R. L. Berger (2002), *Statistical Inference*, Pacific Grove, California: The Wadsworth Group, Inc.
- Cazals, C., J. P. Florens, and L. Simar (2002), Nonparametric frontier estimation: A robust approach, *Journal of Econometrics* 106, 1–25.
- Conrad, R. F. and R. P. Strauss (1983), A multiple-output multiple-input model of the hospital industry in North Carolina, *Applied Economics* 15, 341–352.
- Daouia, A., L. Simar, and P. W. Wilson (2017), Measuring firm performance using nonparametric quantile-type distances, *Econometric Reviews* 36, 156–181.
- Daraio, C. and L. Simar (2005), Introducing environmental variables in nonparametric frontier models: A probabilistic approach, *Journal of Productivity Analysis* 24, 93–121.
- (2007), *Advanced Robust and Nonparametric Methods in Efficiency Analysis*, New York: Springer Science+Business Media, LLC.
- Daraio, C., L. Simar, and P. W. Wilson (2018), Central limit theorems for conditional efficiency measures and tests of the ‘separability condition’ in non-parametric, two-stage models of production, *The Econometrics Journal* 21, 170–191.
- 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.

- Färe, R., S. Grosskopf, and C. A. K. Lovell (1985), *The Measurement of Efficiency of Production*, Boston: Kluwer-Nijhoff Publishing.
- Färe, R., D. Margaritis, P. Rouse, and I. Roshdi (2016), Estimating the hyperbolic distance function: A directional distance function approach, *European Journal of Operational Research* 254, 312–319.
- Farrell, M. J. (1957), The measurement of productive efficiency, *Journal of the Royal Statistical Society A* 120, 253–281.
- Grosskopf, S. and V. Valdmanis (1987), Measuring hospital performance: A nonparametric approach, *Journal of Health Economics* 6, 89–107.
- Jeong, S. O., B. U. Park, and L. Simar (2010), Nonparametric conditional efficiency measures: asymptotic properties, *Annals of Operations 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.
- Klasterin, T. D. and C. A. Watts (1982), A current reappraisal of Berry’s hospital typology, *Medical Care* 20, 441–449.
- 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* 38, 483–515.
- (2015), When bias kills the variance: Central limit theorems for DEA and FDH efficiency scores, *Econometric Theory* 31, 394–422.
- (2016), Testing hypotheses in nonparametric models of production, *Journal of Business and Economic Statistics* 34, 435–456.
- (2021), Inference in dynamic, nonparametric models of production: Central limit theorems for Malmquist indices, *Econometric Theory* 37, 537–572.
- (2022), Conical hull estimators of general technologies, with applications to returns to scale and Malmquist productivity indices. In progress.
- Li, Q., J. Lin, and J. S. Racine (2013), Optimal bandwidth selection for nonparametric conditional distribution and quantile functions, *Journal of Business and Economic Statistics* 31, 57–65.
- Lim, D.-J. (2021), *DJL: Distance Measure Based Judgment and Learning*. R package version 3.7, available at <https://CRAN.R-project.org/package=DJL>.
- McIver, J. P. and E. G. Caremines (1981), *Unidimensional Scaling*, Berverly Hills, California: Sage Publications. Sage University Paper series on Quantitative Applications in the Social Sciences.

- Mittelhammer, R. C. (2013), *Mathematical Statistics for Economics and Business*, New York: Springer-Verlag, Inc.
- 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.
- Pfeiffer, P. E. (1978), *Concepts of Probability Theory*, New York: Dover Publications, Inc.
- Shephard, R. W. (1970), *Theory of Cost and Production Functions*, Princeton: Princeton University Press.
- Sherman, H. D. (1984), Hospital efficiency measurement and evaluation: Empirical test of a new technique, *Medical Care* 22, 922–938.
- Simar, L. and P. W. Wilson (2007), Estimation and inference in two-stage, semi-parametric models of productive efficiency, *Journal of Econometrics* 136, 31–64.
- (2011a), Inference by the m out of n bootstrap in nonparametric frontier models, *Journal of Productivity Analysis* 36, 33–53.
- (2011b), Two-Stage DEA: Caveat emptor, *Journal of Productivity Analysis* 36, 205–218.
- (2013), Estimation and inference in nonparametric frontier models: Recent developments and perspectives, *Foundations and Trends in Econometrics* 5, 183–337.
- (2015), Statistical approaches for non-parametric frontier models: A guided tour, *International Statistical Review* 83, 77–110.
- (2019), Central limit theorems and inference for sources of productivity change measured by nonparametric Malmquist indices, *European Journal of Operational Research* 277, 756–769.
- (2020), Hypothesis testing in nonparametric models of production using multiple sample splits, *Journal of Productivity Analysis* 53, 287–303.
- U.S. Department of Veterans Affairs (1993), Comparison of the operating costs and performance of VA and nonfederal hospitals: 1984–1992, Technical report, Management Science Group, Bedford, Massachusetts.
- Wilson, P. W. (1995), Detecting influential observations in data envelopment analysis, *Journal of Productivity Analysis* 6, 27–45.
- (2008), FEAR: A software package for frontier efficiency analysis with R, *Socio-Economic Planning Sciences* 42, 247–254.
- (2011), Asymptotic properties of some non-parametric hyperbolic efficiency estimators, in I. Van Keilegom and P. W. Wilson, eds., *Exploring Research Frontiers in Contemporary Statistics and Econometrics*, Berlin: Springer-Verlag, pp. 115–150.
- (2018), Dimension reduction in nonparametric models of production, *European Journal of Operational Research* 267, 349–367.
- Zaman, A. (1996), *Statistical Foundations for Econometric Techniques*, San Diego: Academic Press, Inc.

Table 1: Numbers of Observations by Year and Type

Year	VA	LGOVT	NONPROFIT	FORPROFIT	Total
1985	128	205	1092	163	1588
1986	129	268	1430	231	2058
1987	129	247	1322	223	1921
1988	130	247	1365	218	1960
Total	516	967	5209	835	7527

Table 2: Proportion of Hospitals Providing Long-Term Care, by Year and Type

Type	1985	1986	1987	1988
VA	0.6641	0.6977	0.6822	0.7077
LGOVT	0.1073	0.1007	0.1215	0.1538
NONPROFIT	0.1364	0.1280	0.1520	0.1729
FORPROFIT	0.0675	0.0649	0.0448	0.0550

Table 3: Summary Statistics for Variables

Variable	Min	Quartile 1	Median	Mean	Quartile 3	Max
<i>BDH</i>	1.8700(2)	1.0640(3)	2.5850(3)	3.8287(3)	5.1160(3)	4.6508(4)
<i>BDLT</i>	0.0000	0.0000	0.0000	1.3970(1)	0.0000	9.5900(2)
<i>RNFTE</i>	1.4000(1)	1.0400(2)	1.8000(2)	2.3447(2)	2.9700(2)	1.5390(3)
<i>LPFTE</i>	1.0000	2.5000(1)	4.3000(1)	5.4639(1)	7.1000(1)	4.4300(2)
<i>XCFTE</i>	3.0000	1.4600(2)	2.3900(2)	3.2467(2)	4.0900(2)	4.1500(3)
<i>NCFTE</i>	6.0000	1.5100(2)	2.6800(2)	3.6788(2)	4.6600(2)	4.5590(3)
<i>LTFTE</i>	0.0000	0.0000	0.0000	1.2377(1)	0.0000	5.4400(2)
<i>HIND</i>	3.7385(3)	3.6449(4)	6.4145(4)	8.4559(4)	1.0774(5)	6.1281(5)
<i>HDIS</i>	3.0900(2)	5.1880(3)	8.1960(3)	9.9071(3)	1.2802(4)	9.4516(4)
<i>LTIND</i>	0.0000	0.0000	1.6830(3)	8.6476(3)	1.0492(4)	2.4065(5)
<i>OPVIS</i>	0.0000	3.0596(4)	5.2176(4)	7.5573(4)	9.2203(4)	1.9457(6)
<i>SAM</i>	0.0000	1.2555(3)	2.1620(3)	2.7620(3)	3.6120(3)	2.3611(4)
<i>SIP</i>	0.0000	1.8430(3)	3.0650(3)	3.8678(3)	4.9600(3)	3.2673(4)
<i>AMCDAY</i>	1.5667	1.8671(2)	2.3651(2)	2.4696(2)	2.9488(2)	8.9397(2)
<i>CNCFTE</i>	4.4543(−1)	1.4267	1.7479	2.0002	2.2118	9.5143(1)
<i>HERF</i>	1.4800(−2)	9.2600(−2)	2.3990(−1)	3.3929(−1)	5.0050(−1)	1.0000
<i>LOS</i>	1.0000	6.0000	7.0000	7.7799	8.0000	8.2000(1)
<i>PCTRN</i>	2.6168(1)	7.1795(1)	8.0211(1)	7.8670(1)	8.7097(1)	9.9893(1)
<i>WI</i>	5.3870(3)	9.0280(3)	1.0031(4)	1.0154(4)	1.1097(4)	1.6387(4)
X_1^*	3.5029(−1)	1.1994	1.9607	2.5523	3.2377	1.6539(1)
X_2^*	0.0000	0.0000	0.0000	4.3396(−1)	0.0000	1.9442(1)
Y_1^*	0.0000	0.0000	9.8795(−2)	5.0763(−1)	6.1593(−1)	1.4127(1)
Y_2^*	1.0488(−1)	1.4042	2.3061	2.7497	3.5163	1.7794(1)
Z^*	5.3577	9.3317	1.0103(1)	1.0085(1)	1.0915(1)	1.5810(1)

NOTE: Values are represented in scientific notation with $a(b)$ denoting $a \times 10^b$.

Table 4: Proportion of Observations Yielding FDH and VRS-DEA Hyperbolic Efficiency Estimates Equal to One, by Year and Type (Original Data, $p = 7$ and $q = 6$)

Type	1985	1986	1987	1988	All Years
FDH Estimates:					
VA	1.0000	1.0000	1.0000	1.0000	1.0000
LGOVT	0.9854	0.9925	0.9919	0.9960	0.9855
NONPROFIT	0.9744	0.9685	0.9758	0.9839	0.9420
FORPROFIT	0.9939	0.9784	0.9731	0.9541	0.9533
All Types	0.9533	0.9533	0.9533	0.9533	0.9272
VRS-DEA Estimates:					
VA	0.7500	0.8062	0.8062	0.8462	0.5930
LGOVT	0.6537	0.5187	0.5547	0.6154	0.3475
NONPROFIT	0.2747	0.2490	0.2640	0.2703	0.1302
FORPROFIT	0.5767	0.5281	0.4843	0.5275	0.2766
All Types	0.2766	0.2766	0.2766	0.2766	0.0975
CRS-DEA Estimates:					
VA	0.5312	0.6434	0.6124	0.6385	0.3605
LGOVT	0.4927	0.2948	0.3441	0.4291	0.1944
NONPROFIT	0.1355	0.1392	0.1475	0.1648	0.0660
FORPROFIT	0.4110	0.3203	0.3812	0.3349	0.1605
All Types	0.1605	0.1605	0.1605	0.1605	0.0473

Table 5: Proportion of Observations Yielding FDH, VRS-DEA and CRS-DEA Hyperbolic Efficiency Estimates Equal to One, by Year and Type (Reduced Data, $p = q = 2$)

Type	1985	1986	1987	1988	All Years
FDH Estimates:					
VA	0.6406	0.8217	0.6744	0.6538	0.4089
LGOVT	0.3220	0.2537	0.3117	0.2874	0.1706
NONPROFIT	0.1310	0.1196	0.1233	0.1385	0.0572
FORPROFIT	0.2147	0.2338	0.2332	0.2294	0.1162
All Types	0.1162	0.1162	0.1162	0.1162	0.0493
VRS-DEA Estimates:					
VA	0.1562	0.1705	0.1860	0.1462	0.0484
LGOVT	0.0634	0.0448	0.0648	0.0648	0.0155
NONPROFIT	0.0192	0.0133	0.0121	0.0117	0.0044
FORPROFIT	0.0675	0.0433	0.0448	0.0321	0.0144
All Types	0.0144	0.0144	0.0144	0.0144	0.0029
CRS-DEA Estimates:					
VA	0.0391	0.0620	0.0698	0.0385	0.0097
LGOVT	0.0195	0.0112	0.0121	0.0202	0.0041
NONPROFIT	0.0046	0.0028	0.0030	0.0029	0.0008
FORPROFIT	0.0184	0.0173	0.0224	0.0183	0.0048
All Types	0.0048	0.0048	0.0048	0.0048	0.0007

Table 6: Tests of Separability with Respect to Year and Type

Year	Type	— Test #1 —		— Test #2 —	
		Statistic	<i>p</i> -value	Statistic	<i>p</i> -value
Year and Type:					
—	—	7.321	0.000	0.985	0.000
Year by Type:					
—	VA	3.293	0.000	0.726	0.000
—	LGOVT	1.661	0.000	0.535	0.000
—	NONPROFIT	3.232	0.000	0.682	0.000
—	FORPROFIT	1.981	0.000	0.599	0.000
Type by Year:					
1985	—	3.272	0.000	0.738	0.000
1986	—	2.608	0.000	0.564	0.000
1987	—	2.196	0.000	0.608	0.000
1988	—	2.454	0.000	0.732	0.000

NOTE: Test #1 is based on averaging test statistics across sample-splits, and Test #2 is the Kolmogorov-Smirnov test for uniform distribution of *p*-values obtained on each sample-split as described in Section 4.4. All tests are based on 1,000 sample-splits, and *p*-values are estimated from 1,000 bootstrap replications. Estimated *p*-values equal to 0.000 indicate that bootstrapped test statistics were less than the original test statistic for each bootstrap replication. Hence *p*-values shown as 0.000 should be construed to mean “less than 0.001.”

Table 7: Tests of Separability with Respect to Continuous Environmental Variables, by Year and Type

Type	— 1985 —		— 1986 —		— 1987 —		— 1988 —	
	Test #1	Test #2	Test #1	Test #2	Test #1	Test #2	Test #1	Test #2
—Tests of Separability: <i>AMCDAY</i> —								
VA	3.524 (0.999)	0.841 (0.998)	2.427 (0.999)	0.769 (0.999)	3.339 (1.000)	0.837 (0.999)	3.906 (0.999)	0.858 (0.998)
LGOVT	4.829 (0.999)	0.770 (0.998)	5.247 (0.999)	0.724 (0.999)	5.139 (0.999)	0.763 (0.999)	4.781 (0.999)	0.724 (1.000)
NONPROFIT	7.443 (0.999)	0.760 (0.999)	8.414 (0.999)	0.757 (0.999)	8.048 (0.999)	0.757 (0.999)	7.985 (0.999)	0.812 (0.999)
FORPROFIT	4.778 (0.999)	0.690 (0.999)	4.122 (0.999)	0.623 (0.999)	4.513 (0.999)	0.741 (0.999)	4.408 (0.999)	0.684 (0.999)
—Tests of Separability: <i>CNCFTE</i> —								
VA	3.461 (0.999)	0.817 (0.998)	2.396 (0.999)	0.772 (0.999)	3.464 (0.999)	0.849 (0.999)	3.793 (0.999)	0.843 (0.998)
LGOVT	5.043 (0.999)	0.800 (0.999)	5.130 (0.999)	0.709 (1.000)	5.142 (0.999)	0.752 (0.999)	4.739 (0.999)	0.719 (1.000)
NONPROFIT	7.689 (0.999)	0.758 (1.000)	8.305 (0.999)	0.770 (0.999)	7.631 (0.999)	0.749 (1.000)	7.818 (0.999)	0.790 (1.000)
FORPROFIT	4.560 (0.999)	0.676 (0.999)	4.169 (0.999)	0.607 (1.000)	4.318 (0.999)	0.717 (1.000)	4.751 (0.999)	0.724 (0.999)

NOTE: Numbers in parentheses give p -values. Test #1 is based on averaging test statistics across sample-splits, and Test #2 is the Kolmogorov-Smirnov test for uniform distribution of p -values obtained on each sample-split as described in Section 4.4. All tests are based on 1,000 sample-splits, and p -values are estimated from 1,000 bootstrap replications.

Table 7: Tests of Separability with Respect to Continuous Environmental Variables, by Year and Type (continued)

Type	— 1985 —		— 1986 —		— 1987 —		— 1988 —	
	Test #1	Test #2	Test #1	Test #2	Test #1	Test #2	Test #1	Test #2
—Tests of Separability: <i>HERF</i> —								
VA	3.490 (0.999)	0.816 (0.999)	2.393 (0.999)	0.777 (0.998)	3.333 (0.999)	0.841 (0.998)	3.822 (0.999)	0.854 (0.999)
LGOVT	4.793 (0.999)	0.767 (0.999)	5.413 (0.999)	0.761 (0.999)	5.039 (0.999)	0.748 (0.998)	4.841 (0.999)	0.740 (1.000)
NONPROFIT	7.824 (0.999)	0.796 (0.999)	8.763 (0.999)	0.811 (0.999)	8.098 (0.999)	0.807 (0.999)	8.181 (0.999)	0.817 (1.000)
FORPROFIT	4.568 (0.999)	0.679 (0.999)	4.353 (0.999)	0.638 (0.999)	4.331 (0.999)	0.699 (1.000)	4.194 (0.999)	0.677 (1.000)
—Tests of Separability: <i>LOS</i> —								
VA	3.438 (0.999)	0.821 (0.998)	2.433 (0.999)	0.759 (0.998)	3.433 (0.999)	0.845 (0.999)	3.872 (0.999)	0.866 (0.999)
LGOVT	4.882 (0.999)	0.790 (0.999)	5.487 (0.999)	0.753 (0.999)	5.268 (0.999)	0.781 (0.998)	4.854 (0.999)	0.732 (0.999)
NONPROFIT	8.330 (0.999)	0.879 (0.999)	9.256 (0.999)	0.887 (0.999)	8.277 (0.999)	0.818 (0.999)	8.200 (0.999)	0.840 (0.999)
FORPROFIT	4.789 (0.999)	0.699 (0.999)	4.412 (0.999)	0.645 (0.999)	4.201 (0.999)	0.726 (0.999)	4.756 (0.999)	0.725 (0.999)

NOTE: Numbers in parentheses give p -values. Test #1 is based on averaging test statistics across sample-splits, and Test #2 is the Kolmogorov-Smirnov test for uniform distribution of p -values obtained on each sample-split as described in Section 4.4. All tests are based on 1,000 sample-splits, and p -values are estimated from 1,000 bootstrap replications.

Table 7: Tests of Separability with Respect to Continuous Environmental Variables, by Year and Type (continued)

Type	— 1985 —		— 1986 —		— 1987 —		— 1988 —	
	Test #1	Test #2	Test #1	Test #2	Test #1	Test #2	Test #1	Test #2
	—Tests of Separability: <i>PCTRN</i> —							
VA	3.405 (0.999)	0.819 (0.999)	2.334 (0.999)	0.755 (0.999)	3.380 (0.999)	0.836 (0.999)	3.725 (0.999)	0.846 (0.999)
LGOVT	4.979 (0.999)	0.802 (0.999)	5.546 (0.999)	0.765 (0.999)	5.326 (0.999)	0.774 (0.999)	4.832 (0.999)	0.749 (0.999)
NONPROFIT	8.374 (0.999)	0.836 (0.999)	9.272 (0.999)	0.820 (0.999)	8.564 (0.999)	0.838 (0.999)	8.677 (0.999)	0.869 (0.999)
FORPROFIT	4.639 (0.999)	0.693 (0.999)	4.454 (0.999)	0.665 (0.999)	4.390 (0.999)	0.720 (0.999)	4.437 (0.999)	0.696 (0.999)
	—Tests of Separability: <i>WI</i> —							
VA	3.505 (0.999)	0.842 (0.999)	2.446 (0.999)	0.779 (0.999)	3.416 (0.999)	0.846 (0.999)	3.931 (0.999)	0.862 (0.999)
LGOVT	4.788 (0.999)	0.780 (0.999)	5.381 (0.999)	0.747 (0.999)	5.233 (0.999)	0.784 (0.999)	4.981 (0.999)	0.738 (1.000)
NONPROFIT	8.202 (0.999)	0.812 (0.999)	8.413 (0.999)	0.787 (0.999)	8.192 (0.999)	0.803 (0.999)	8.696 (0.999)	0.859 (0.999)
FORPROFIT	4.587 (0.999)	0.680 (0.999)	4.262 (0.999)	0.627 (0.999)	4.423 (0.999)	0.732 (0.999)	4.463 (0.999)	0.696 (0.999)

NOTE: Numbers in parentheses give p -values. Test #1 is based on averaging test statistics across sample-splits, and Test #2 is the Kolmogorov-Smirnov test for uniform distribution of p -values obtained on each sample-split as described in Section 4.4. All tests are based on 1,000 sample-splits, and p -values are estimated from 1,000 bootstrap replications.

Table 7: Tests of Separability with Respect to Continuous Environmental Variables, by Year and Type (continued)

Type	— 1985 —		— 1986 —		— 1987 —		— 1988 —	
	Test #1	Test #2	Test #1	Test #2	Test #1	Test #2	Test #1	Test #2
—Tests of Separability: Z^* —								
VA	3.450 (0.999)	0.824 (0.998)	2.360 (0.999)	0.760 (0.999)	3.428 (0.999)	0.847 (0.998)	3.939 (0.999)	0.865 (0.999)
LGOVT	4.875 (0.999)	0.775 (0.999)	5.456 (0.999)	0.754 (0.999)	5.353 (0.999)	0.789 (0.999)	4.844 (0.999)	0.734 (1.000)
NONPROFIT	7.692 (0.999)	0.778 (0.999)	8.642 (0.999)	0.783 (0.999)	8.313 (0.999)	0.810 (0.999)	8.287 (0.999)	0.840 (0.998)
FORPROFIT	4.779 (0.999)	0.709 (0.999)	4.076 (0.999)	0.627 (0.999)	4.338 (0.999)	0.709 (1.000)	4.590 (0.999)	0.703 (0.999)

NOTE: Numbers in parentheses give p -values. Test #1 is based on averaging test statistics across sample-splits, and Test #2 is the Kolmogorov-Smirnov test for uniform distribution of p -values obtained on each sample-split as described in Section 4.4. All tests are based on 1,000 sample-splits, and p -values are estimated from 1,000 bootstrap replications.

Table 8: Tests of Convexity by Year and Type

Type	1985	1986	1987	1988
— Test #1 —				
VA	1.200 (0.004)	2.301 (0.018)	1.529 (0.000)	−0.351 (1.000)
LGOVT	−0.656 (1.000)	−1.180 (1.000)	−0.401 (1.000)	−0.322 (1.000)
NONPROFIT	−1.576 (1.000)	−1.771 (1.000)	−1.239 (1.000)	0.030 (0.430)
FORPROFIT	−1.953 (1.000)	−1.236 (1.000)	−0.342 (1.000)	−0.359 (1.000)
— Test #2 —				
VA	0.411 (0.000)	0.708 (0.000)	0.509 (0.000)	0.260 (0.000)
LGOVT	0.343 (0.000)	0.441 (0.000)	0.206 (0.071)	0.207 (0.003)
NONPROFIT	0.530 (0.000)	0.578 (0.000)	0.411 (0.000)	0.147 (0.520)
FORPROFIT	0.587 (0.000)	0.427 (0.000)	0.168 (1.000)	0.197 (1.000)

NOTE: Numbers in parentheses give p -values estimated from 1,000 bootstrap replications. Values 0.000 indicate that bootstrapped test statistics were less than the original test statistic for each bootstrap replication. Hence p -values shown as 0.000 should be construed to mean “less than 0.001.”

Table 9: Summary Statistics for Efficiency Estimates by Year and Type

Year	Type	Min	Quartile 1	Median	Mean	Quartile 3	Max
1985	VA	0.8132	0.9664	1.0000	0.9733	1.0000	1.0000
	LGOVT	0.8340	1.0000	1.0000	0.9878	1.0000	1.0000
	NONPROFIT	0.7266	0.9733	1.0000	0.9780	1.0000	1.0000
	FORPROFIT	0.5904	0.9347	1.0000	0.9545	1.0000	1.0000
1986	VA	0.5189	0.8379	0.9343	0.9034	1.0000	1.0000
	LGOVT	0.5090	0.7995	0.8875	0.8790	1.0000	1.0000
	NONPROFIT	0.6295	0.8177	0.9059	0.8952	1.0000	1.0000
	FORPROFIT	0.6519	0.8193	0.9151	0.8981	1.0000	1.0000
1987	VA	0.4339	0.7659	0.8444	0.8430	0.9338	1.0000
	LGOVT	0.4860	0.7578	0.8323	0.8322	0.9210	1.0000
	NONPROFIT	0.4367	0.7679	0.8346	0.8407	0.9208	1.0000
	FORPROFIT	0.5371	0.7907	0.8671	0.8634	0.9458	1.0000
1988	VA	0.4637	0.7685	0.8626	0.8546	0.9771	1.0000
	LGOVT	0.4794	0.7783	0.8683	0.8615	0.9810	1.0000
	NONPROFIT	0.5979	0.8203	0.9038	0.8872	0.9861	1.0000
	FORPROFIT	0.5703	0.7982	0.9020	0.8829	0.9951	1.0000

Table 10: Tests of Differences in Mean Efficiency across Years, by Type

Years	VA	LGOVT	NONPROFIT	FORPROFIT
1985 v. 1986	−1.9800 (0.0477)	0.9967 (0.3189)	0.5655 (0.5718)	−2.4849 (0.0130)
1985 v. 1987	−0.1320 (0.8949)	−1.0410 (0.2979)	0.4545 (0.6495)	−3.3643 (0.0008)
1985 v. 1988	1.7980 (0.0722)	−0.9609 (0.3366)	−1.2825 (0.1997)	−3.3785 (0.0007)
1986 v. 1987	1.9026 (0.0571)	−2.2092 (0.0272)	−0.2601 (0.7948)	−0.8220 (0.4111)
1986 v. 1988	3.1500 (0.0016)	−2.0355 (0.0418)	−2.0144 (0.0440)	−0.9059 (0.3650)
1987 v. 1988	2.0098 (0.0445)	0.2492 (0.8032)	−1.7876 (0.0738)	−0.2458 (0.8058)

NOTE: Each cell in the table contains two values; the top value is the value of the test statistic, and the bottom value is the corresponding p -value, with $a \times 10^{-b}$ represented as $a(-b)$. Positive (negative) values of test statistics indicate larger (smaller) mean efficiency for the second year than the first year.

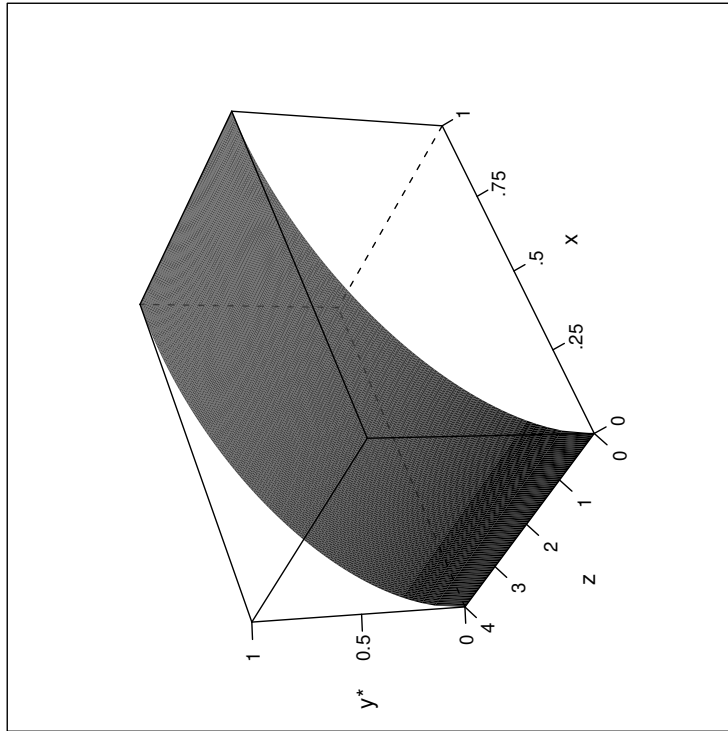
Table 11: Tests of Differences in Mean Efficiency across Types, by Year

Types	1985	1986	1987	1988
VA v. LGOVT	6.7963 (0.0000)	9.8275 (0.0000)	5.9734 (0.0000)	3.5201 (0.0004)
VA v. NONPROFIT	0.0759 (0.0000)	3.7812 (0.0000)	1.8058 (0.0000)	5.4503 (0.0000)
VA v. FORPROFIT	8.3528 (0.0000)	7.3312 (0.0000)	5.8254 (0.0000)	2.8736 (0.0041)
LGOVT v. NONPROFIT	1.2045 (0.2284)	0.2964 (0.7669)	2.8651 (0.0042)	1.2262 (0.2201)
LGOVT v. FORPROFIT	2.5376 (0.0112)	-1.1263 (0.2600)	0.1008 (0.9197)	-0.2110 (0.8329)
NONPROFIT v. FORPROFIT	1.7607 (0.0783)	-1.7326 (0.0832)	-2.5301 (0.0114)	-1.3143 (0.1888)

NOTE: Each cell in the table contains two values; the top value is the value of the test statistic, and the bottom value is the corresponding p -value, with $a \times 10^{-b}$ represented as $a(-b)$. Positive (negative) values of test statistics indicate larger (smaller) mean efficiency for the second type than the first type.

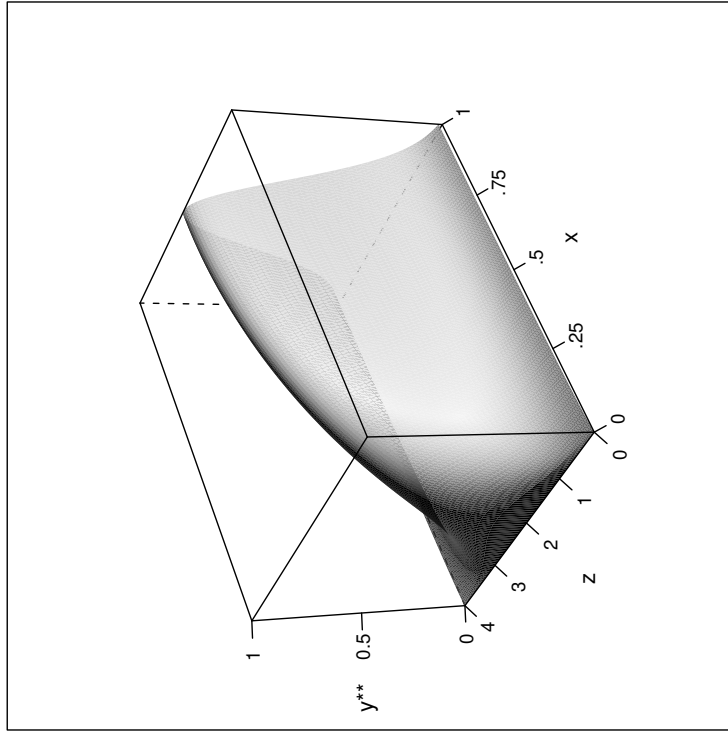
Figure 1: Illustration of Separability versus Non-Separability from Sinar and Wilson (2011b)

$$Y^* = g(X)e^{-(Z-2)^2}U$$



(a)

$$Y^{**} = g(X)e^{-(Z-2)^2}e^{-U}$$



(b)