HYPOTHESIS TESTING IN NONPARAMETRIC MODELS OF PRODUCTION USING MULTIPLE SAMPLE SPLITS

Simar, L. and Paul Wilson

DISCUSSION PAPER | 2019 / 19

Hypothesis Testing in Nonparametric Models of Production using Multiple Sample Splits

LÉOPOLD SIMAR

PAUL W. WILSON*

September 2019

Abstract

Several tests of model structure developed by Kneip et al. (2016) and Daraio et al. (2018) rely on comparing sample means of two different efficiency estimators, one appropriate under the conditions of the null hypothesis and the other appropriate under the conditions of the alternative hypothesis. These tests rely on central limit theorems developed by Kneip et al. (2015) and Daraio et al., but require that the original sample be split randomly into two independent subsamples. This introduces some ambiguity surrounding the sample-split, which may be determined by choice of a seed for a random number generator. We develop a method that eliminates much of this ambiguity by repeating the random splits a large number of times. We use a bootstrap algorithm to exploit the information from the multiple sample-splits. Our simulation results show that in many cases, eliminating this ambiguity results in tests with better size and power than tests that employ a single sample-split.

Keywords: DEA, FDH, bootstrap, inference, hypothesis testing

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

Wilson: Department of Economics and School of Computing, Division of Computer Science, Clemson University, Clemson, South Carolina, USA; email: pww@clemson.edu.

We are grateful to the Cyber Infrastructure Technology Integration group at Clemson University for operating the Palmetto Cluster used for simulations in this paper..

1 Introduction

The statistical properties (under appropriate assumptions) such as consistency, rates of convergence, and limiting distributions of either the conditional or unconditional versions of data envelopment analysis (DEA) or free-disposal hull (FDH) estimators of technical efficiency are by now well-known due to the work of a number of authors (see Simar and Wilson, 2013 and 2015 for recent surveys). DEA estimators can be divided into two classes, namely those that impose constant returns to scale (CRS) which we refer to as CRS-DEA estimators, and those that permit variable returns to scale (VRS) which we refer to as VRS-DEA estimators. Recently, Kneip et al. (2015) derive properties on the moments of unconditional, input-oriented FDH and VRS-DEA as well as CRS-DEA estimators. These results extend trivially to the output-orientation, and are extended to hyperbolic versions of the FDH and DEA estimators by Kneip et al. (2018). The results of Kneip et al. (2015) are used by Kneip et al. (2016) to develop statistical tests of (i) differences in mean efficiency across two groups of producers, (ii) CRS versus VRS and (iii) convexity versus non-convexity of the production set. Alternatively, Daraio et al. (2018) derive properties on the moments of conditional, output-oriented FDH and VRS-DEA efficiency estimators, and use these results to establish statistical tests of the infamous “separability condition” described by Simar and Wilson (2007, 2011).

The test developed by Kneip et al. (2015) for differences in mean efficiency across two groups of producers is straightforward provided one deals with the bias problems discovered by Kneip et al. (2015). In particular, all of the nonparametric efficiency estimators mentioned above have bias that disappears asymptotically as sample size increases, but only at the same rate at which the particular estimator converges. As explained by Kneip et al. (2015), this means that unless the number of inputs and outputs is very small, standard central limit theorem (CLT) results such as the Lindeberg-Feller CLT do not hold for mean efficiency estimated by the nonparametric estimators. This is explained in more detail below; see also Kneip et al. (2015) for details and proofs. Kneip et al. (2015) develop new CLTs for mean efficiencies estimated by unconditional FDH or DEA estimators, with bias estimated by a generalized jackknife estimator. Similarly, Daraio et al. (2018) provide new CLTs for mean efficiencies estimated by conditional FDH or DEA estimators, with bias again estimated by

a generalized jackknife estimator. These new CLTs also require subsampling, with means computed over a random subsample of observations when the number of inputs and outputs exceeds a small bound that depends on the particular estimator being used.

To test for differences in mean efficiency across two groups, Kneip et al. (2016) use their CLT results to establish asymptotic normality of a test statistic involving the difference in bias-corrected sample means of technical efficiency estimates for each of the two groups. Kneip et al. (2016) discuss how independence between observations across the two groups is crucial for their results. Without independence, it is not clear how one might estimate unknown, non-zero correlations in order to establish a limiting distribution for a test statistic. For the case of testing for differences in mean efficiency across two groups, the applied researcher presumably has in hand two sets of observations, one from each group. But when testing CRS versus VRS, or convexity versus non-convexity of the production set or separability versus non-separability in the sense of Simar and Wilson (2007), there is typically only one set of observations, from one group of producers. Each of these tests involves comparing a sample mean of efficiency estimates that impose the conditions of the null hypothesis against a sample mean of estimates that do not. To ensure independence between the two means under comparison, the test Kneip et al. (2016) and Daraio et al. (2018) randomly split the original sample into two independent sub-samples. Both Kneip et al. (2016) and Daraio et al. (2018) establish asymptotic normality of their test statistics under the null, enabling testing using critical values from the standard normal distribution. Both papers provide extensive Monte Carlo evidence showing that the tests work well in finite samples of the size often encountered by applied researchers.

While the results of Kneip et al. (2016) and Daraio et al. (2018) hold for a single random split of the original sample, some have noticed (and complained) that p -values resulting from the tests vary across different random splits of the original sample. In fact, one can obtain almost any result (i.e., almost any p -value) by repeatedly splitting the original sample. As will be seen below, this is neither surprising nor evidence that something is wrong. Nor is this the only instance where randomization is introduced into a testing situation; e.g., see Hoel et al. (1971, pp. 63–67). Nonetheless, if two researchers working with the same data do not split the data the same way, they will obtain different results. Daraio et al. (2018) present a pseudo-random, computer science approach for splitting data that guarantees that two

researchers will obtain the same splits, even if one shuffles the observations before splitting. But of course, once the data have been split into two sub-samples, different results will be obtained depending on which sub-sample is used for estimation under conditions of the null hypothesis and which is not. Again, this is neither surprising or an indication of something amiss in the proofs of Kneip et al. (2016) and Daraio et al. (2018). Nonetheless, the ambiguity introduced by the requirement to split samples is perhaps annoying.

We propose a method that eliminates much of the ambiguity of a single split of the sample by splitting the sample many times. This results in dependence across the multiple sample splits. We deal with this dependence using a bootstrap algorithm to exploit information from multiple splits and to provide valid inference. To our knowledge, this is the only method that can be used to eliminate the seeming arbitrariness of a single random split and to deal with the dependence that arises when a single sample is split multiple times.

In the next section we present the nonparametric efficiency estimators used in the tests of convexity and returns to scale (RTS) developed by Kneip et al. (2016), as well as in the separability test of Daraio et al. (2018). In Section 3 we illustrate the need for randomly splitting a given sample, and discuss the implications of splitting the sample more than once. We use the RTS test of Kneip et al. (2016) to motivate the discussion, but the issues are the same for tests of convexity or separability. We provide in Section 4 a bootstrap procedure that can be used to combine information across multiple sample splits to give a valid test of the null hypothesis. Section 5 gives results from Monte Carlo experiments showing how well the new bootstrap method can be expected to perform in finite samples. Implications for applied researchers are discussed in Section 6, and Conclusions are presented in Section 7.

2 Nonparametric Efficiency Estimators and Their Properties

To establish notation, let $x, X \in \mathbb{R}_+^p$ denote nonstochastic and random p -vectors of input quantities, and let $y, Y \in \mathbb{R}_+^q$ denote nonstochastic and random q -vectors of output quantities (respectively). The production set or set of feasible combinations of inputs and outputs is

given by

$$\Psi := \{(x, y) \mid x \text{ can produce } y\}. \quad (2.1)$$

We assume the usual characteristics of the production framework. In particular, we assume Ψ is closed, and production of any non-zero output quantity requires use of non-zero level of at least one output. In addition, both inputs and outputs are freely disposable so that for $\tilde{x} \geq x$, $\tilde{y} \leq y$, if $(x, y) \in \Psi$ then $(\tilde{x}, y) \in \Psi$ and $(x, \tilde{y}) \in \Psi$, with inequalities involving vectors are defined on an element-by-element basis, as is standard. The assumption of freely disposable inputs and outputs amounts to an assumption of weak monotonicity on the frontier.

The Farrell (1957) input efficiency measure

$$\theta(x, y \mid \Psi) := \inf \{\theta \mid (\theta x, y) \in \Psi\} \quad (2.2)$$

and measures efficiency in terms of the amount by which input levels can be scaled downward by the same factor without reducing output levels. Alternatively, the Farrell (1957) output efficiency measure

$$\lambda(x, y \mid \Psi) := \sup \{\lambda \mid (x, \lambda y) \in \Psi\} \quad (2.3)$$

gives the feasible proportionate expansion of output quantities, holding input quantities fixed. One can consider other measures as well, e.g., the hyperbolic measure introduced by Färe et al. (1985) or the directional measures proposed by Chambers et al. (1998), Färe and Grosskopf (2006) and Färe et al. (2008).

Unconditional estimators of the efficiency measures in (2.2) and (2.3) are obtained by replacing Ψ with suitable estimators. Let $\mathcal{X}_n = \{(X_i, Y_i)\}$ denote a random sample of n observations on input and output quantities. The FDH estimator of Ψ due to Deprins et al. (1984) is given by

$$\hat{\Psi}_{\text{FDH},n} := \bigcup_{(X_i, Y_i) \in \mathcal{X}_n} \{(x, y) \in \mathbb{R}_+^{p+q} \mid x \geq X_i, y \leq Y_i\}. \quad (2.4)$$

The VRS-DEA estimator $\hat{\Psi}_{\text{VRS},n}$ of Ψ proposed by Farrell (1957) and popularized by Charnes et al. (1978) is the convex hull of $\hat{\Psi}_{\text{FDH},n}$, i.e.,

$$\hat{\Psi}_{\text{VRS},n} := \{(x, y) \in \mathbb{R}^{p+q} \mid y \leq \mathbf{Y}\boldsymbol{\omega}, x \geq \mathbf{X}\boldsymbol{\omega}, \mathbf{i}'_n \boldsymbol{\omega} = 1, \boldsymbol{\omega} \in \mathbb{R}_+^n\}, \quad (2.5)$$

where $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_n)$ are $(p \times n)$ and $(q \times n)$ matrices of input and output vectors, respectively; $\mathbf{i}_n$ is an $(n \times 1)$ vector of ones, and $\boldsymbol{\omega}$ is a $(n \times 1)$ vector of weights. The CRS-DEA estimator $\widehat{\Psi}_{\text{CRS},n}$ of $\Psi_{\text{VRS},n}$ is given by the conical hull of $\widehat{\Psi}_{\text{DEA},n}$ (with vertex at the origin) obtained by dropping the constraint $\mathbf{i}'_n \boldsymbol{\omega} = 1$ in (2.5).

Substituting $\widehat{\Psi}_{\text{FDH},n}$ into (2.2) or (2.3) leads to integer programming problems, but the resulting estimators can be computed easily using

$$\widehat{\theta}_{\text{FDH}}(x, y \mid \mathcal{X}_n) = \min_{i \in D_{x,y}} \max_{j=1, \dots, p} \left(\frac{X_i^j}{x^j} \right) \quad (2.6)$$

or

$$\lambda_{\text{FDH}}(\mathbf{x}, \mathbf{y} \mid \mathcal{X}_n) = \max_{i \in D_{x,y}} \min_{j=1, \dots, q} \left(\frac{\mathbf{Y}_i^j}{\mathbf{y}^j} \right), \quad (2.7)$$

where $D_{x,y} = \{i \mid (X_i, Y_i) \in \mathcal{X}_n, X_i \leq x, Y_i \geq y\}$ is the set of indices of points in $\mathcal{X}_n$ dominating (x, y) and where for a vector a , a^j denotes its j -th component. Alternatively, substituting $\widehat{\Psi}_{\text{VRS},n}$ into (2.2) or (2.3) leads to

$$\widehat{\theta}_{\text{VRS}}(x, y \mid \mathcal{X}_n) = \min_{\theta, \boldsymbol{\omega}} \{ \theta \mid y \leq \mathbf{Y} \boldsymbol{\omega}, \theta x \geq \mathbf{X} \boldsymbol{\omega}, \mathbf{i}'_n \boldsymbol{\omega} = 1, \boldsymbol{\omega} \in \mathbb{R}_+^n \} \quad (2.8)$$

or

$$\widehat{\lambda}_{\text{VRS}}(x, y \mid \mathcal{X}_n) = \max_{\lambda, \boldsymbol{\omega}} \{ \lambda \mid \lambda y \leq \mathbf{Y} \boldsymbol{\omega}, x \geq \mathbf{X} \boldsymbol{\omega}, \mathbf{i}'_n \boldsymbol{\omega} = 1, \boldsymbol{\omega} \in \mathbb{R}_+^n \}, \quad (2.9)$$

which can be solved easily using standard optimization methods for linear programs. Substituting $\widehat{\Psi}_{\text{CRS},n}$ into (2.2) or (2.3) leads to the CRS-DEA estimators $\widehat{\theta}_{\text{CRS}}(x, y \mid \mathcal{X}_n)$ and $\widehat{\lambda}_{\text{CRS}}(x, y \mid \mathcal{X}_n)$ obtained by removing the constraint $\mathbf{i}'_n \boldsymbol{\omega} = 1$ in (2.8) and (2.9).

The *conditional* output-oriented FDH efficiency estimator due to Daraio and Simar (2007) is given by

$$\widehat{\lambda}_{\text{FDH}}(x, y \mid z, \mathcal{X}_n) = \max_{i \in \mathcal{I}_{\text{FDH}}(z, h)} \left(\min_{j=1, \dots, p} \left(\frac{Y_i^j}{y^j} \right) \right), \quad (2.10)$$

where $\mathcal{I}_{\text{FDH}}(z, h) = \{i \mid z - h \leq Z_i \leq z + h \cap X_i \leq x\}$, and the conditional output-oriented VRS-DEA efficiency estimator is given by

$$\begin{aligned} \widehat{\lambda}_{\text{VRS}}(x, y \mid z, \mathcal{X}_n) &= \max_{\lambda, \boldsymbol{\omega}} \{ \lambda \mid \lambda y \leq \mathbf{Y}(\boldsymbol{\beta} \circ \boldsymbol{\omega}), x \geq \mathbf{X}(\boldsymbol{\beta} \circ \boldsymbol{\omega}), \\ &\quad \mathbf{i}'_n(\boldsymbol{\beta} \circ \boldsymbol{\omega}) = 1, \boldsymbol{\omega} \in \mathbb{R}_+^n, \boldsymbol{\beta} \in \{0, 1\}^n \}, \end{aligned} \quad (2.11)$$

where “ $\circ$ ” denotes the Hadamard product β is a Bernoulli vector of length n with i th element β_i such that $\beta_i = 1$ if $(z - h) \leq Z_i \leq (z + h)$ or $\beta_i = 0$ otherwise.¹ The conditional, output-oriented CRS-DEA efficiency estimator $\hat{\lambda}_{\text{CRS}}(x, y \mid z, \mathcal{X}_n)$ is obtained by dropping the constraint $i'_n(\beta \circ \omega) = 1$ in 2.11. Note that the conditional estimators in (2.10) and (2.11) are localized versions of the corresponding unconditional efficiency estimators, with the degree of localization controlled by the bandwidth h . See Daraio et al. (2018) for practical issues regarding the choice of bandwidths. Conditional, input-oriented efficiency estimators can be constructed using similar localization techniques.

Statistical properties of the estimators discussed above, under assumptions appropriate for each estimator, have been derived in a number of papers. Let n^κ denote the rate of convergence, with the value of κ depending on the particular estimator.² Kneip et al. (1998) establish consistency of and the rate of the unconditional VRS-DEA estimator (under VRS) with $\kappa = 2/(p + q + 1)$, Kneip et al. (2008) establish the estimator’s limiting distribution. Park et al. (2010) establish consistency, limiting distribution and the rate of the unconditional CRS-DEA estimator (under CRS) with $\kappa = 2/(p + q)$. Kneip et al. (2016) prove that under CRS, the unconditional VRS-DEA estimator attains the faster rate of the unconditional CRS-DEA estimator. Park et al. (2000) and Daouia et al. (2017) establish analogous properties for the unconditional FDH estimator, with $\kappa = 1/(p + q)$. All of these results are for input-oriented estimators, but the results extend trivially to the output-orientation.³ Jeong et al. (2010) establish similar properties for the conditional FDH and VRS-DEA estimators, with convergence at rate $\kappa/(\kappa r + 1)$ when the bandwidths are the same across the r variables conditioned upon and have the optimal rate.

In addition, Kneip et al. (2015) prove that the bias of each of the unconditional estimators is of order $O(n^{-\kappa})$, while the variance and covariance are of orders $o(n^{-\kappa})$ and $o(n^{-1})$

¹ For two vectors a and b of length n with i th elements a_i and b_i , $c = a \circ b$ is a vector of length n with i th element $c_i = a_i b_i$.

² For an estimator $\hat{\theta}(x, y)$ of $\theta(x, y)$ converging at rate n^κ , $\hat{\theta}(x, y) - \theta(x, y) = O_p(n^{-\kappa})$. In other words, the estimation error of $\hat{\theta}(x, y)$ is of order in probability $n^{-\kappa}$. In such cases, estimation error becomes less in probabilistic terms as the sample size n increases, but how fast this happens depends on the magnitude of κ .

³ In addition, the results on consistency, limiting distributions and rates of convergence have been extended to hyperbolic versions of the FDH and VRS-DEA estimators by Wheelock and Wilson (2008) and Wilson (2011), and to directional-distance versions by Simar and Vanhems (2012) and Simar et al. (2012). For each type of estimator—FDH, VRS-DEA or CRS-DEA—the value of κ remains the same across the different orientations.

(respectively). Kneip et al. show that because the bias is of order larger than the order of the variance, standard CLTs do not hold for means of unconditional, input-oriented FDH or DEA estimators unless $\kappa > 1/2$. For the VRS-DEA and CRS-DEA estimators under CRS, this requires $(p+q) \leq 3$. For the VRS-DEA estimators under VRS, this requires $(p+q) \leq 2$, and for the FDH estimators it requires $(p+q) < 2$.⁴ Similarly, Daraio et al. (2018) prove that the bias of the output-oriented FDH and VRS-DEA estimators is of order $O(n^{-\kappa/(\kappa r+1)})$ when bandwidths are of the optimal order, while the variance and covariance are of orders $o(n^{-\kappa/(\kappa r+1)})$ and $o(n^{-1/(\kappa r+1)})$, respectively. Standard CLTs for means of the conditional estimators are never valid, but Daraio et al. (2018) provide new CLTs involving bias corrections using generalized jackknife estimators of bias and requiring means over subsamples of observations when $(p+q)$ is sufficiently large.

As noted above, Kneip et al. (2016) and Daraio et al. (2018) use these results to develop statistical tests of various model features. Each test involves comparing a sample mean of efficiency estimates that impose the conditions of the null hypothesis against a sample mean of efficiency estimates where the null is not imposed. All of the tests require that the two sample means under comparison be independent of each other, and this requires randomly splitting the original sample. In the next section, we show how the ambiguity resulting from the random split can be largely removed.

3 Why Random Splitting is Needed

In this section we illustrate the issues arising from random splitting as well as solutions to the resulting issues in terms of the test of CRS versus VRS. The ideas developed here extend easily and naturally to the Kneip et al. (2016) test of convexity versus non-convexity of Ψ or the test of separability developed by Daraio et al. (2018).

Under the null hypothesis of CRS, both the VRS-DEA and CRS-DEA estimators converge at rate n^κ with $\kappa = 2/(p+q)$. Given the random sample $\mathcal{X}_n$ introduced in Section 2, the test of CRS versus VRS developed by Kneip et al. (2016) requires splitting the original sample $\mathcal{X}_n$ into two sub-samples $\mathcal{X}_{1,n_1}$, $\mathcal{X}_{2,n_2}$ such that $\mathcal{X}_{1,n_1} \cup \mathcal{X}_{2,n_2} = \mathcal{X}_n$ and $\mathcal{X}_{1,n_1} \cap \mathcal{X}_{2,n_2} = \emptyset$, where $n_1 = \lfloor n/2 \rfloor$ and $n_2 = n - n_1$ and $\lfloor a \rfloor$ denotes the largest integer less than or equal to

⁴ These results extend trivially to the output-oriented estimators. Wilson (2019) extends the results to the hyperbolic orientation.

a. Apply the input-oriented VRS-DEA estimator $\hat{\theta}_{\text{VRS}}(X_i, Y_i \mid \mathcal{X}_{1,n_1})$ to each observation in $\mathcal{X}_{1,n_1}$, and apply the CRS-DEA estimator $\hat{\theta}_{\text{CRS}}(X_i, Y_i \mid \mathcal{X}_{2,n_2})$ to each observation in $\mathcal{X}_{2,n_2}$. Let

$$\hat{\mu}_{\text{VRS},n_1} = n_1^{-1} \sum_{(X_i, Y_i) \in \mathcal{X}_{1,n_1}} \hat{\theta}_{\text{VRS}}(X_i, Y_i \mid \mathcal{X}_{1,n_1}) \quad (3.1)$$

and

$$\hat{\mu}_{\text{CRS},n_2} = n_2^{-1} \sum_{(X_i, Y_i) \in \mathcal{X}_{2,n_2}} \hat{\theta}_{\text{CRS}}(X_i, Y_i \mid \mathcal{X}_{2,n_2}). \quad (3.2)$$

Let $\hat{B}_{\text{VRS},\kappa,n_1}$ and $\hat{B}_{\text{CRS},\kappa,n_1}$ denote the corresponding generalized jackknife estimates of bias appearing in equations (35)–(36) of Kneip et al. (2016), and let

$$\hat{\sigma}_{\text{VRS},n_1}^2 = n_1^{-1} \sum_{(X_i, Y_i) \in \mathcal{X}_{1,n_1}} \left[\hat{\theta}_{\text{VRS}}(X_i, Y_i \mid \mathcal{X}_{1,n_1}) - \hat{\mu}_{\text{VRS},n_1} \right]^2 \quad (3.3)$$

and

$$\hat{\sigma}_{\text{CRS},n_2}^2 = n_2^{-1} \sum_{(X_i, Y_i) \in \mathcal{X}_{2,n_2}} \left[\hat{\theta}_{\text{CRS}}(X_i, Y_i \mid \mathcal{X}_{2,n_2}) - \hat{\mu}_{\text{CRS},n_2} \right]^2. \quad (3.4)$$

Kneip et al. (2016) show that under the conditions of Kneip et al. ((2015), Lemma 4.1), Theorem 4.2 of Kneip et al. (2015) and Theorem 3.1 of Kneip et al. (2016) are sufficient to establish that under the null hypothesis of CRS,

$$\hat{T}_{1,n} = \frac{(\hat{\mu}_{\text{VRS},n_1} - \hat{\mu}_{\text{CRS},n_2}) - (\hat{B}_{\text{VRS},\kappa,n_1} - \hat{B}_{\text{CRS},\kappa,n_2})}{\sqrt{\frac{\hat{\sigma}_{\text{VRS},n_1}^2}{n_1} + \frac{\hat{\sigma}_{\text{CRS},n_2}^2}{n_2}}} \xrightarrow{\mathcal{L}} N(0, 1) \quad (3.5)$$

provided $(p + q) \leq 5$.

Alternatively, if $\kappa < 1/2$, one can compute the sample means using subsets of the available observations. For $\ell \in \{1, 2\}$, let $n_{\ell,\kappa} = \lfloor n_\ell^{2\kappa} \rfloor$; then $n_{\ell,\kappa} < n_\ell$ for $\kappa < 1/2$. Let $\mathcal{X}_{\ell,n_{\ell,\kappa}}^*$ be a random subset of $n_{\ell,\kappa}$ input-output pairs from $\mathcal{X}_{\ell,n_\ell}$. Let

$$\hat{\mu}_{\text{VRS},n_{1,\kappa}} = n_{1,\kappa}^{-1} \sum_{(X_i, Y_i) \in \mathcal{X}_{1,n_{1,\kappa}}^*} \hat{\theta}(X_i, Y_i \mid \mathcal{X}_{1,n_1}) \quad (3.6)$$

and

$$\hat{\mu}_{\text{CRS},n_{2,\kappa}} = n_{2,\kappa}^{-1} \sum_{(X_i, Y_i) \in \mathcal{X}_{2,n_{2,\kappa}}^*} \hat{\theta}(X_i, Y_i \mid \mathcal{X}_{2,n_2}). \quad (3.7)$$

Note that the full subsets $\mathcal{X}_{1,n_1}$ and $\mathcal{X}_{2,n_2}$ are used to compute the efficiency estimates in (3.6)–(3.7), but the summations are over subsets of the observations used to compute the efficiency estimates. Then Theorem 4.4 of Kneip et al. (2015) and Theorem 3.1 of Kneip et al. (2015) establish that

$$\hat{T}_{2,n} = \frac{(\hat{\mu}_{\text{VRS},n_1,\kappa} - \hat{\mu}_{\text{CRS},n_2,\kappa}) - (\hat{B}_{\text{VRS},\kappa,n_1} - \hat{B}_{\text{CRS},\kappa,n_2})}{\sqrt{\frac{\hat{\sigma}_{\text{VRS},n_1}^2}{n_{1,\kappa}} + \frac{\hat{\sigma}_{\text{CRS},n_2}^2}{n_{2,\kappa}}}} \xrightarrow{\mathcal{L}} N(0,1) \quad (3.8)$$

for $(p+q) \geq 5$.

Depending on the value of $(p+q)$, either $\hat{T}_{1,n}$ or $\hat{T}_{2,n}$ (but not both, except when $p+q = 5$) can be used to test the null hypothesis of CRS versus VRS, with critical values obtained from the standard normal distribution. In particular, for $j \in \{1, 2\}$, the null hypothesis of CRS is rejected if $\hat{p} = 1 - \Phi(\hat{T}_{j,n})$ is less than, say, .1, .05, or .01. If $(p+q) = 5$, either of the statistics $\hat{T}_{1,n}$ or $\hat{T}_{2,n}$ can be used. In this case, one should expect $\hat{T}_{2,n}$ to perform better in finite samples for reasons similar to those given in Kneip et al. (2018).⁵

Splitting the original sample $\mathcal{X}_n$ into two independent subsamples $\mathcal{X}_{1,n_1}$ and $\mathcal{X}_{2,n_2}$ is essential for obtaining non-degenerate limiting distributions for the test statistics in (3.5) and (3.8). Kneip et al. (2016) show that if one does not split the original sample and instead applies the VRS-DEA and CRS-DEA estimators to all of the n observations in $\mathcal{X}_n$, and then builds a statistic similar to (3.5) or (3.8) but using sample means over all n observations, the resulting covariances wipe out the variances, resulting in a test statistic with a degenerate distribution for any value of $(p+q)$. The same is true for the test of convexity versus non-convexity of Ψ developed by Kneip et al. (2016) as well as for the test of separability developed by Daraio et al. (2018).⁶

The test of CRS versus VRS, as well as the tests of convexity and separability, are valid

⁵ A typo appears in Kneip et al. (2016) after equation (42), which holds for $(p+q) \geq 4$ instead of $(p+q) > 5$. A similar typo follows equation (52), which holds for $(p+q) \geq 2$ instead of $(p+q) > 3$.

⁶ In the test of convexity versus non-convexity, one applies the FDH estimator to the observations in $\mathcal{X}_{1,n_1}$ and the VRS-DEA estimator to the observations in $\mathcal{X}_{2,n_2}$. In the test of separability, one applies the conditional VRS-DEA (or conditional FDH) estimator to observations in $\mathcal{X}_{1,n_1}$ and the unconditional VRS-DEA (or unconditional FDH) estimator to the observations in $\mathcal{X}_{2,n_2}$. In the test of CRS to VRS outlined here, both estimators used to construct the test statistic have the same rate of convergence, and so $\kappa = 2/(p+q)$. In the tests of convexity and separability, the two estimators used for each test have different rates of convergence under the null, and so the variances in the denominators of the statistics for these tests are divided by different powers of n , reflecting the different convergence rates of the estimators. See Kneip et al. (2016) and Daraio et al. (2018) for details.

for a single, random split of the data as made clear by the results of Kneip et al. (2016) and Daraio et al. (2018). Moreover, the Monte Carlo results provided by Kneip et al. (2016) and Daraio et al. (2018) indicate that the tests have reasonable size and power properties in finite samples. However, given n observations, there are $m = \binom{n}{n/2}$ possible splits. For $n = 50$, $m \approx 1.246 \times 10^{14}$; for $n = 100$, $m \approx 1.009 \times 10^{29}$; and for $n = 1,000$, $m \approx 2.703 \times 10^{299}$. Moreover, one should expect to obtain a different p -value with different splits of the same sample. Even if one uses the splitting algorithm proposed by Daraio et al. (2018), the results will differ if one changes the labels on the two sub-samples. The temptation might be to split the sample different ways until one obtains his or her desired result, but of course this is anti-science and is dishonest. At a minimum, it is unnerving to some to realize that test results depend on an arbitrary, random split of the data into two subsamples, perhaps determined by the choice of a seed for a random number generator if one does not use the splitting algorithm of Daraio et al. (2018).

Almost every first-year, graduate-level statistics textbook discusses the probability integral transform. Given a continuous random variable W with distribution function $F_W(\cdot)$, the probability integral transform ensures that a random variable $U = F(W)$ will be uniformly distributed on the $(0,1)$ interval. Consequently for a test statistic with some asymptotic, non-degenerate distribution (not necessarily normal) under the null hypothesis, the corresponding p -values obtained from multiple, random samples of size n where the null is true will be asymptotically uniformly distributed on the $(0,1)$ interval. This is illustrated below in Figure 1 for one of our Monte Carlo experiments.

Now Suppose an applied researcher wishes to test CRS versus VRS, and that $(p + q)$ is small enough so that the statistic in (3.5) can be used. Given a random sample $\mathcal{X}_n$, one can randomly split the sample once to obtain a value of the statistic in (3.5). Shuffling the observations and splitting again yields a second realization of the statistic, almost surely different from the first value that was obtained. Suppose the researcher shuffles and then splits the sample $s \ll m$ times, obtaining a set $\mathcal{T} = \{\hat{T}_{1,n,j}\}_{j=1}^s$, of s different values of the test statistic given in (3.5). Corresponding to each $\hat{T}_{1,n,j}$ is a realized p -value $\hat{p}_j$, so that the researcher also has a set $\mathcal{P} = \{\hat{p}_j\}_{j=1}^s$ of p -values. If the statistics in $\mathcal{T}$ were independent, then by the probability integral transform, the p -values in $\mathcal{P}$ would be independent and uniformly distributed on $(0,1)$.

However, the realized values of the test statistic over s splits contained in $\mathcal{T}$ cannot be independent. To illustrate the reason for this, suppose n is even. The first split of the sample $\mathcal{X}_n$ results in two sub-samples A and B of size $n/2$. Now suppose the original sample is shuffled and split again, resulting in two sub-samples C and D of size $n/2$. Then the value of the test statistic computed using A and B cannot not be independent of the value of test statistic computed using C and D since on average, about half the observations in sub-sample C will be observations also appearing in sub-sample A, and the other half will be identical to observations in sub-sample B. Similar observations hold for sub-sample D. This induces dependence of a complicated form. Moreover, while each of the p -values in $\mathcal{P}$ corresponding to the test statistics in $\mathcal{T}$ are uniformly distributed on $(0, 1)$ due to the probability integral transform, they cannot be independent since the values in $\mathcal{T}$ are not independent. Consequently, any attempt to combine the information in $\mathcal{T}$ or $\mathcal{P}$ across multiple splits of the sample $\mathcal{X}_n$ that ignores this dependence cannot give a valid test of the null hypothesis.

4 Using Information from Multiple Sample Splits

If test statistics using multiple splits of the same sample were independent, one could combine the information from different splits in any of several ways. For example, since the statistics are distributed $N(0, 1)$, the s values in the set $\mathcal{T}$ described at the end of Section 3 could be averaged; multiplying their sample mean by $\sqrt{s}$ would yield a new statistic with asymptotic distribution $N(0, 1)$. Alternatively, one might use the one-sample Kolmogorov-Smirnov test to test the null hypothesis that the p -values in the set $\mathcal{P}$ are uniformly distributed on $(0, 1)$, and then reject CRS in favor of VRS if the Kolmogorov-Smirnov test rejects uniformity of the p -values. But neither the test statistics in $\mathcal{T}$ nor the p -values in $\mathcal{P}$ are independent, and hence the previous two approaches, as well as any others that require independence, cannot provide valid inference.

The bootstrap presented below provides a simple way to deal with the dependence across multiple splits. Continuing to use the RTS test outlined in Section 3 for illustrative purposes, suppose again that $(p + q)$ is small enough so that the statistic $\hat{T}_{1,n}$ in (3.5) can be used. For

$s \ll m$ splits of the original sample $\mathcal{X}_n$, define

$$\bar{T}_n := s^{-1} \sum_{j=1}^s \hat{T}_{\ell,n,j} \quad (4.1)$$

for $\ell = 1$ or 2 as appropriate depending on the value of $(p + q)$, and define

$$\hat{K}_n := \sup_{u \in [0,1]} \left(|\hat{F}_{s,\hat{p},n}(u) - u| \right), \quad (4.2)$$

where $\hat{F}_{s,\hat{p},n}(u)$ is the empirical distribution function of the s values in $\mathcal{P}$, i.e.,

$$\hat{F}_{s,\hat{p},n}(u) = s^{-1} \sum_{j=1}^s I(\hat{p}_j \leq u) \quad (4.3)$$

where $I(\cdot)$ denotes the indicator function. The statistic $\hat{K}_n$ is a one-sample Kolmogorov-Smirnov statistic, but its distribution is unknown due to the dependence induced by splitting repeatedly the same sample. Similarly, the distribution of $\bar{T}_n$ is also unknown, for the same reason.

The following bootstrap algorithm provides an estimate of the sampling distributions of either $\bar{T}_n$ or $\hat{K}_n$ and enables inference-making from multiple sample-splits.

Bootstrap Algorithm:

- [1] Randomly shuffle the observations in $\mathcal{X}_n$ and then split into two subsamples $\mathcal{X}_{1,n_1}$, $\mathcal{X}_{2,n_2}$ as described in Section 3, and compute the test statistic $\hat{T}_{\ell,n}$ for $\ell = 1$ or 2 as appropriate.
- [2] Repeat step [1] $(s - 1)$ times to obtain $\mathcal{T} = \{\hat{T}_{\ell,n,j}\}_{j=1}^s$. Compute $\bar{T}_n$ using (4.1), the set $\mathcal{P}$ of p -values corresponding to the elements of $\mathcal{T}$, and K_n using (4.2).
- [3] Compute $\hat{\theta}_i = \hat{\theta}_{\text{CRS}}(X_i, Y_i \mid \mathcal{X}_n)$ for each observation $i = 1, \dots, n$ in $\mathcal{X}_s$. Set $b = 0$.
- [4] Increment b by 1. Draw k_i , $i = 1, \dots, n$ independently and with replacement from the set of integers 1 through n , such that each integer has probability n^{-1} of being selected on a particular draw, and set $\theta_i^* = \hat{\theta}_{k_i}$.
- [5] Create a bootstrap sample $\mathcal{X}_n^* = \{(X_i^*, Y_i^*)\}_{i=1}^n$ where $X_i^* = X_i \hat{\theta}_i / \theta_i^*$ and $Y_i^* = Y_i$.

- [6] Analogous to step [1], randomly shuffle the observations in $\mathcal{X}_n^*$ and then split into two subsamples $\mathcal{X}_{1,n_1}^*$, $\mathcal{X}_{2,n_2}^*$, and compute the test statistic $\widehat{T}_{\ell,n}^*$ for the value of ℓ used in step [1] using (3.5) if $\ell = 1$ or (3.8) if $\ell = 2$.
- [7] Repeat step [6] s times to obtain $\mathcal{T}_b^* = \{\widehat{T}_{\ell,n,j}^*\}_{j=1}^s$. Compute $\overline{T}_{n,b}^*$ using (4.1), the set $\mathcal{P}_b^*$ of p -values corresponding to the elements of $\mathcal{T}_b^*$, and $K_{n,b}^*$ analogous (4.2) using the values in $\mathcal{P}_b^*$.
- [8] Repeat steps [6]–[7] $(B - 1)$ times to obtain $\{\overline{T}_{n,b}^*\}_{b=1}^B$ and $\{\widehat{K}_{n,b}^*\}_{b=1}^B$.
- [9] Compute

$$\widehat{p}_T = \frac{\#\{\overline{T}_{n,b}^* \geq \overline{T}_n\}}{B} \quad (4.4)$$

and

$$\widehat{p}_K = \frac{\#\{\widehat{K}_{n,b}^* \geq \widehat{K}_n\}}{B}. \quad (4.5)$$

For a test of size α (e.g., $\alpha = 0.10, 0.05$ or 0.01), reject the null hypothesis of CRS if $\widehat{p}_T$ and $\widehat{p}_K$ are less than α .

Note that in step [3] of the algorithm, CRS-DEA estimates are computed for the entire, original sample. These are then used in step [5] to first project observations onto the estimated frontier in the input direction, and then to project away from the frontier by a random amount. These two steps are crucial and ensure that the bootstrap data are generated under the restrictions of the null hypothesis. This amounts to a “naive” bootstrap, with resampling from an empirical distribution. The naive bootstrap fails when making inference about efficiencies of individual producers because the nonparametric efficiency estimators discussed in Section 2 measure distance from a fixed point to an estimated support boundary. The naive bootstrap is known to fail in such situations; see Simar and Wilson (1998, 1999a and 1999b) for additional discussion. The problem here is very different. By contrast, we are making inference about a mean or the distribution of estimated p -values rather than distance to a support boundary, and there are no such problems. Valid inference is ensured by the CLT results of Kneip et al. (2015) and Daraio et al. (2018), and Mammen (1992, Theorem 1).

To implement the algorithm, one must choose values for B and s . Our experience with various simulations suggests that $B = 1,000$ bootstrap replications and $s = 10$ splits provide

a good compromise between performance of the tests and computational burden. Since one is using the bootstrap to estimate p -values, which necessarily involve features of tails of sampling distributions, one should use at least 1,000 bootstrap replications. We performed several experiments with $s = 100$ splits (which increases the computational burden by a factor of about 10 over the case where $s = 10$), and found little difference in terms of the achieved sizes or power of the tests.

One should expect the two p -values computed in step [9] to be close, as is the case in our simulations discussed below in Section 5. However, for a particular nominal size α chosen a priori, it is conceivable that one might reject the null with one test but not the other. This possibility arises whenever one uses two different tests for the same null hypothesis (e.g., in fully-parametric work, one might employ a likelihood-ratio test as well as a Wald test or perhaps a Lagrange multiplier test). In all such cases, one should be hesitant to draw strong conclusions, as the tests provide conflicting information. One might avoid this possibility by using only one of the tests; we have included both to economize on our presentation. But of course one should also be reluctant to conclude strongly when a single test barely rejects (or fails to reject) the null. One cannot learn truth from data, but the estimated p -values give an idea of how much (or how little) evidence there is in the data to reject the null hypothesis.

We have illustrated our bootstrap method in terms of the RTS test discussed in Section 3. It is trivial to adapt the algorithm to the convexity test developed by Kneip et al. (2016). To do so, one would use the FDH and VRS-DEA estimator to construct the appropriate test statistic given in Kneip et al. (2016) instead of the VRS-DEA and CRS-DEA estimators used in the RTS test, adjusting the value of κ accordingly as described in Kneip et al. (2016). In step [3] of the algorithm, one would compute VRS-DEA estimates instead of CRS-DEA estimates in order to impose convexity under the null hypothesis. The algorithm can be similarly adapted for the separability test introduced by Daraio et al. (2018) by using conditional and unconditional estimators, with appropriate values for κ as described in Daraio et al. (2018). For the separability test, one would use the unconditional estimator in steps [3] and [5] to impose separability, i.e., the restriction implied by the null hypothesis.

In the next section, we present evidence from Monte Carlo experiments for each of the RTS, convexity, and separability tests showing how well one might expect the tests to perform in finite samples.

5 Monte Carlo Evidence

5.1 RTS Test

We first consider the RTS test. We generate data from the data-generating process (DGP) described in Kneip et al. (2016, Section 5.1, equations (64)–(65) and Figure 1). The parameter $\delta \geq 0$ controls departures from the null, with the CRS holding when $\delta = 0$, and departure from the null increasing with δ . The only difference between the DGP here and the one in Kneip et al. (2016) is that here, the value of $s_{\max}$ (in the notation of Kneip et al.) has been reduced to 70-percent of the value used in Kneip et al. to avoid generating observations with output levels very close to zero which can cause numerical problems. See Kneip et al. for additional details and discussion.

We consider values of $(p+q)$ ranging from 2 to 6, and $\delta \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.6, 0.8, 1.4\}$. We use the statistic in (3.5) based on full sample means when $(p+q) \leq 4$ (case I), and we use the statistic in (3.8) based on sub-sample means when $(p+q) > 4$ (case II). Figure 1 of Kneip et al. (2016) gives a visual illustration of the degree of departure from CRS provided by each of the non-zero values of δ . In all of our experiments, we use $s = 10$ sample-splits (except in a few experiments as noted earlier where we used $s = 100$) and $B = 1,000$ bootstrap replications. In each experiment, we perform 1,000 Monte Carlo trials, and report the proportion among these 1,000 trials where we reject the null hypothesis (i.e., the rate at which the null is rejected) using nominal test sizes of .10, .05 and .01. We conduct experiments involving the RTS test with sample sizes $n \in \{100, 1000\}$. With 1,000 trials, the confidence intervals on a reported rejection rate $\tilde{p}$ for test of size α is $\tilde{p} \pm z_{1-\frac{\alpha}{2}} (\alpha(1-\alpha)/1000)^{1/2}$ where $z_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$ quantile of the standard normal distribution function. For $\alpha \in \{0.1, 0.05, 0.01\}$ this corresponds to $\tilde{p} \pm 0.0156, 0.0135, 0.0081$ (respectively).

For comparison purposes, Table 1 reports rejection rates for tests similar to those in Kneip et al. (2016), i.e., with only one split and with critical values taken from the $N(0, 1)$ distribution instead of being estimated by bootstrapping. The results are broadly similar to those shown in Kneip et al. (2016) for the RTS test, with rejection rates approaching nominal sizes as n increases, and with power increasing with n and with increasing departures from the null. For the case where $n = 1,000$, $p = 5$ and $q = 1$, Figure 1 shows the empirical distributions of the p -values obtained with one sample-split on each of the 1,000 Monte

Carlo trials for $\delta = 0.0$ (where the null hypothesis of CRS holds) and for $\delta = 0.2, 0.4$ and 1.4 . The red line in each plot in Figure 1 shows the uniform $(0,1)$ distribution function, and the vertical dashed line is drawn at 0.1 on the horizontal axis to make it easy to see the proportion of p -values that are greater than or less than 0.1 in each plot. In the upper-left plot, where the null is true, it is clear that the p -values across 1,000 independent samples are uniformly distributed, exactly as predicted by the probability integral transform discussed above in Section 3. With $\delta = 0.2$ in the upper-right plot of Figure 1, there is some divergence between the empirical distribution of the p -values and the uniform $(0,1)$ distribution function. This divergence increases as δ increases to 0.4 and then 1.4 . When the null is true, with nominal test size $\alpha = 0.1$ the null hypothesis is rejected in 10.7 percent of the 1,000 Monte Carlo trials as seen from the upper-left plot in Figure 1 and from Table 1, reflecting the rate of type-I errors, which is seen to be close to the nominal size of the test.

Tables 2 and 3 show rejection rates obtained with the test statistics in (3.5) and (3.8) (respectively) using the bootstrap algorithm in Section 4. The rejection rates for either of the statistics defined in (3.5) or (3.8) are similar over the various scenarios. However, either of the statistics based on multiple splits in many cases yield realized size closer to nominal values than the single-split test relying on asymptotic normality in Table 1. Moreover, the power of the multiple-split tests is in many cases superior to that of the single-split test. In addition, much of the ambiguity created by a single split of a given sample has been removed.

5.2 Convexity Test

The test of convexity versus non-convexity of the production set Ψ developed by Kneip et al. (2016) involves comparing means of FDH efficiency estimates (where convexity is not imposed) against means of VRS-DEA efficiency estimates (where convexity is imposed). As discussed above in Section 2, FDH estimators converge at rate $n^{1/(p+q)}$, while the VRS-DEA estimator converges at rate $n^{2/(p+q+1)}$ under strict convexity or at rate $n^{2/(p+q)}$ under (weak) convexity and CRS.

To simulate data to examine performance of the convexity test, we generate data from the DGP described in Kneip et al. (2016, Section 5.1, “third set of experiments”), again reducing the value of the parameter $s_{\max}$ (in the notation of Kneip et al., 2016) to 70 percent of the value used in Kneip et al. (2016) to avoid possible numerical problems. We consider

$q = 1$ and $p \in \{1, 2, 3, 4, 5\}$. The parameter δ controls departures from the null of convexity for Ψ . As in Kneip et al. (2016), we first simulate a case using same DGP used in the RTS tests described in Section 5.1 with $\delta = 1.4$ so that Ψ is strictly convex and the null hypothesis holds. We then use the DGP described in Kneip et al. (2016) where $\delta = 0.0$ corresponds to a convex but CRS technology, and increasing values of δ result in increasing degrees of non-convexity for the production set. Using this DGP, we simulate cases for $\delta \in \{0.1, 0.2, 0.3, 0.4, 0.6, 1.4\}$ where departure from the null increases with δ . Figure 2 of Kneip et al. (2016) provides an illustration of the degree of departure from the null for each value of δ . For each experiment, we again compute 1,000 Monte Carlo trials, each with $s = 10$ sample-splits and $B = 1,000$ bootstrap replications. As in Kneip et al. (2016), we split sample unevenly, with more observations to the subsample used by the FDH estimator than in the subsample used with the VRS-DEA estimator due to the slower convergence rate of the FDH estimator. In particular, we set $n_1^{2/(p+q+1)} = n_2^{1/(p+q)}$ and $n_1 + n_2 = n$ and then solve for n_1 and n_2 as described in Kneip et al. (2016, Section 3.3).

Table 4 reports rejection rates using a single sample split with critical values taken from the standard normal distribution, similar to the test in Kneip et al. (2016). The results in Table 4 are broadly similar to the corresponding results in Table 5 of Kneip et al. (2016), with size and power improving with sample size n , although not as quickly for larger as opposed to smaller dimensions. Tables 5–6 show rejection rates for the tests with multiple splits. The rejection rates in both 5 and 6 where the production set is strictly convex are similar to the rejection rates in Table 4 where only a single sample-split is used. However, the power of the tests with multiple splits increases more rapidly with increasing departures from the null than is the case in Table 4. For example, with 6 dimensions and $n = 1,000$, the rejection rate in Table 4 at the 10-percent level is 0.231 when $\delta = 0.4$, while the tests based on multiple splits reject at rates 0.430 and 0.354. When δ is increased to 0.6, the difference is even more dramatic; with a single split in Table 4 is 0.350 at the 10-percent level, but the tests that rely on multiple splits yield rejection rates of 0.816 and 0.721 in Tables 5 and 6 (respectively). As in Kneip et al. (2016), rejection rates are smaller when the production set is strictly convex than when it is weakly convex with CRS.

5.3 Separability Test

Daraio et al. (2018) develop a test of the separability condition described by Simar and Wilson (2007). In a nutshell, this condition requires that any environmental variables one might regress estimated efficiencies on in a second-stage analysis cannot influence the frontier, but only the distribution of efficiency. The test of Daraio et al. (2018) compares means of conditional efficiency estimates against means of unconditional efficiency estimates. If the difference is “large,” then the null hypothesis of separability is rejected. As with the tests already examined, this test also requires splitting the original sample in order to maintain independence between the two sample means under comparison.

We revisit the first set of experiments described in Daraio et al. (2018, Separate Appendix E). We simulate cases with $n \in \{100, 200\}$ and $p = q = 1$, $p = 2$ and $q = 1$, $p = q = 2$, $p = 3$ and $q = 2$, and with $p = q = 3$. The parameter δ controls the degree of departure from the null hypothesis of separability, with separability holding when $\delta = 0.0$. Departures from the null increase with increasing values of δ . In each of our experiments we computed 1,000 Monte Carlo trials, and on each trial we use $s = 10$ sample splits and $B = 1,000$ bootstrap replications.⁷

For purposes of comparison, we again report in Table 7 rejection rates obtained with a single sample-split and with critical values chosen from the standard normal distribution. The results are similar to those reported by Daraio et al.; any differences are due only to the fact that the sequence of generated random numbers differs here from the sequence used in Daraio et al. due to the addition of additional code to handle multiple splits and bootstrapping.

Tables 8–9 give rejection rates for the test statistics using multiple splits. The achieved test size with the Kolmogorov-Smirnov statistic based on (3.8) in many—but not all—cases has size closer to the nominal test size than the averaged statistic based on (3.5).⁸ Both of the new statistics typically have better size properties at the .10 level than the statistic

⁷ Daraio et al. report results from experiments with $n = 1,000$ in addition to $n = 100$ and 200 . However, with 10 sample splits, and 1,000 bootstrap replications, the computational burden for each experiment here is 10,010 times that of the experiments in Daraio et al.. Moreover, with the separability test, a bandwidth parameter must be selected by cross-validation, which requires time of order $O(n^2)$, and this must be done 10,010 times. Consequently, we consider only $n = 100, 200$ for the separability test.

⁸ Note that while the statistic in (3.8) is a Kolmogorov-Smirnov statistic, the usual tables cannot be used to assess significance due to the dependence problem here.

that uses only one sample split. In addition, in many cases, power is better with the new statistics using multiple splits as opposed to the original statistic based on a single split. With 5 or 6 dimensions, the power is low even for large departures from the null with only 100 or 200 observations. On the other hand, power increases rapidly when $p = q = 1$, and so one might consider dimension reduction methods along the lines of Wilson (2018), which are shown to work well (in terms of reducing mean-square error of the efficiency estimates) in many cases.

6 Practical Issues for the Applied Researcher

The tests proposed by Kneip et al. (2016) and Daraio et al. (2018), where the sample is split only once, do not require bootstrapping and involve asymptotic standard normal limiting distributions. This makes the tests attractive by avoiding the computational burden of the bootstrap. For these tests, the computational burden depends on the sample size n , the number of dimensions $(p + q)$, and the number K of replications of the generalized jackknife used to compute bias corrections as described by Kneip et al. (2016). For a single sample-split and n even, the RTS test requires solution of n linear programs (LPs) each with $(1 + \frac{n}{2})$ weights, and Kn LPs each with $(1 + \frac{n}{4})$ weights. But with s sample-splits and B bootstrap replications, the test based on either of the statistics in (3.5) or (3.8) require solution of n LPs with $(1 + n)$ weights, $ns(B + 1)$ LPs with $(1 + \frac{n}{2})$ weights, and $Kns(B + 1)$ LPs with $(1 + \frac{n}{4})$ weights.

In our Monte Carlo experiments discussed above in Section 5, this computational burden is multiplied by the number of Monte Carlo trials, i.e., by a factor of 1,000. While this is not an issue for the applied researcher who faces only a single sample of data, the computational burden forced us to limit our simulations to only those with $s = 10$ sample-splits on each Monte Carlo trial. Nonetheless, the applied researcher, who has a single sample, must decide how many sample splits to use, and this choice must be balanced against the required computational burden which increases with the number of splits. For the applied researcher, n is given by the size of his sample. Moreover, the number B of bootstrap samples should be at least 1,000, and our experience suggests the number K of generalized jackknife replications should be at least 100 to minimize noise (there is apparently little benefit to increasing K beyond 100). With $B = 1,000$, $s = K = 100$ and $n = 1,000$, this results in 1,000 LPs with

1,000 weights, 1.001×10^8 LPs with 501 weights and 1.001×10^{10} LPs with 251 weights for a single sample.

To gain some insight into the role of the number s of sample splits, we conducted some additional experiments using the RTS test based on averaging statistics over the sample splits as in (3.5). In each experiment, we set with $n = 100$, $p = 5$ and $q = 1$ and generate a single sample of size n with δ equal to either 0.0 (so the null is true) or 1.4 (where the null is false) as described above in Section 5.1 and in Kneip et al. (2016). We then split the sample s times to compute the statistic in (3.5) and its corresponding p -value (which requires the bootstrap discussed above in Section 4), and we repeat this 100 times, resulting in 100 p -values for a single sample. Of course, these p -values are not independent, and we know of no way to combine them into a single p -value that would provide valid inference, but their distribution is illuminating.

Figure 2 shows the distributions of p -values over 100 repetitions with $\delta = 0.0$ in the upper panel and $\delta = 1.4$ in the lower panel. Each panel shows a red, 45-degree line representing the uniform distribution function, and three empirical distribution functions in green, blue and black corresponding to the distributions of p -values over the 100 repetitions with $s = 2$, 10 and 100 splits, respectively. Each panel also shows a vertical dotted line drawn at 0.1 on the horizontal axis, and a horizontal dotted line drawn at 0.1 on the vertical axis.

In the top panel of Figure 2, the null hypothesis is true and so one should hope to avoid rejecting the null. The figure shows that almost any p -value can be obtained with a single set of s splits, just as is the case for the original test proposed by Kneip et al. (2016) involving only a single sample-split. With $s = 2$ or 10 splits, the null is rejected in about 10-percent of the 100 repetitions at the .10 level, but with $s = 100$ splits, the null is rejected in only 2 of 100 repetitions, suggesting that there is some benefit to increasing the number of splits.

This effect becomes more dramatic in the bottom panel of Figure 2, where the null hypothesis is false. As the number of splits increases, the range of p -values decreases, and the distribution of p -values becomes increasingly concentrated on the left. With only 2 splits, the null is rejected in about 60-percent of the 100 repetitions, but this increases to about 95-percent with 10 splits and 100-percent with 100 splits. Of course the two cases shown in Figure 2 with $\delta \in \{0.0, 1.4\}$ are extreme in the sense that the null is either true or grossly false, but as δ increases starting from 0.0, one should expect the distributions to move toward

those shown in the lower panel of Figure 2.

The message from Figure 2 is clear—more splits are preferred to fewer splits. Moreover, one should *not* do what was done in the experiments leading to Figure 2, i.e., only one set of s splits should be done to compute the test statistic, as there is not way to aggregate 100 dependent p -values from a single sample to give a valid test. So, s should be made as large as possible, but this must be balanced against the resulting computational burden. As to computational burden, it is important to note that modern desktop (and even laptop) computers have become increasingly fast and inexpensive. Moreover, modern desktop and laptop computers typically have 4 or more *cores*, allowing as many computational threads as there are cores. The bootstrap discussed in Section 4 is trivially parallel, since each of the B replications are independent.

To give additional evidence, our Monte Carlo experiments were performed on the massively parallel Palmetto Cluster at Clemson University. Each experiment was run on a mix of Intel XEON E5345, E5410, L5420 and E5640 quad-core chips with clock speeds ranging from 2.33 to 2.67GHz. On each Monte Carlo trial, parallelization was used for the initial $s = 10$ sample-splits, as well as for the $B = 1,000$ bootstrap replications. Each experiment used 128 cores, with 1 core used as a counter since shared memory is not available on clustered super-computing systems such as the Palmetto Cluster, and 127 cores were used for computation. Our code records the elapsed time for each trial in seconds. Each trial uses 127 cores for computation, hence multiplying this time by 127/60 gives time in core-hours, or the time one would need for a single sample while performing computations on a single processing core. The times given below are upper bounds due to latency of some cores near the end of each Monte Carlo trial. In addition, the cores we used on the Palmetto Cluster are now several years old; new desktop and laptop machines often have cores with clock speeds of 3.0GHz or more.

With $n = 100$ and $p = q = 1$, each Monte Carlo trial for the RTS test required 0.1567 to 0.1857 core-hours, and with $n = 100$, $p = 5$ and $q = 1$, each trial required 0.4175 to 0.6680 core-hours. With $n = 1,000$ and $p = q = 1$, times range from 6.7330 to 10.4126 core-hours, and for $n = 1,000$, $p = 5$ and $q = 1$ times range from 24.0894 to 58.3517 core-hours, with a mean of 35.0822 core-hours. Recall that our Monte Carlo experiments used $s = 10$ splits. With 100 observations and 6 dimensions, increasing the number of splits to 100 would require

about 4.175 hours to 6.680 hours running on a single core; one could use 1,000 splits and still have results likely in less than 66.80 hours (i.e., 2.78 days) using a single core. With 1,000 observations, 10 splits should be feasible for anyone. While some might not want to wait long enough to use 100 splits, the fact that modern desktop and laptop computers have clock speeds greater than the chips used for our experiments, and the fact that modern desktops and laptops have 4 or more cores, one should be able easily use 20–30 splits and perhaps more depending on one's patience. One could also use 2 or more computers or a grid-computing system such as Condor (e.g., see Thain et al., 2005) since the computational problem is trivially parallel.⁹

7 Conclusions

Kneip et al. (2016) and Daraio et al. (2018) provide tests of CRS versus VRS, convexity versus non-convexity, and separability that require a single split of the original sample and which involve statistics with asymptotic $N(0, 1)$ distributions. Here, we provide a bootstrap method to remove some of the dependence of test results on the particular random split, which may be determined by choice of a seed for a random number generator if the randomization algorithm of Daraio et al. (2018) is not used. A key element of our bootstrap method involves generating the bootstrap data under the conditions of the null hypothesis, which is not unlike other bootstrap tests. Our simulation results show that removing this randomness results in many cases where size and power of the tests are improved over what is obtained with a single split. The tests, together with the new bootstrap method introduced here, will soon be added to a future version of the FEAR software package described by Wilson (2008).

Theoretical results in Kneip et al. (2018, 2019) on estimators of changes in productivity (either under convexity or non-convexity of the production set) can be used to construct tests similar to the ones discussed in this paper, and which also require randomly splitting samples to maintain independence. The methods developed here are useful in these situations, as well as in other future developments where testing might require randomly splitting a sample.

⁹ Computation times for the convexity test are faster than those given here due to the fact that the FDH estimator involves lower computational burden than the VRS-DEA estimator. On the other hand, times for the separability test are slower than those for the RTS test due to the necessity of cross-validation to optimize bandwidths used by the conditional efficiency estimators.

References

- Chambers, R. G., Y. Chung, and R. Färe (1998), Profit, directional distance functions, and nerlovian efficiency, *Journal of Optimization Theory and Applications* 98, 351–364.
- Charnes, A., W. W. Cooper, and E. Rhodes (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research* 2, 429–444.
- 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 (2007), Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach, *Journal of Productivity Analysis* 28, 13–32.
- 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. and S. Grosskopf (2006), *New Directions: Efficiency and Productivity*, New York: Springer Science & Business Media.
- Färe, R., S. Grosskopf, and C. A. K. Lovell (1985), *The Measurement of Efficiency of Production*, Boston: Kluwer-Nijhoff Publishing.
- Färe, R., S. Grosskopf, and D. Margaritis (2008), Productivity and efficiency: Malmquist and more, in H. Fried, C. A. K. Lovell, and S. Schmidt, eds., *The Measurement of Productive Efficiency*, chapter 5, Oxford: Oxford University Press, 2nd edition, pp. 522–621.
- Farrell, M. J. (1957), The measurement of productive efficiency, *Journal of the Royal Statistical Society A* 120, 253–281.
- Hoel, P. G., S. C. Port, and C. J. Stone (1971), *Introduction to Statistical Theory*, Boston: Houghton Mifflin Company.
- Jeong, S. O., B. U. Park, and L. Simar (2010), Nonparametric conditional efficiency measures: asymptotic properties, *Annals of Operations Research* 173, 105–122.
- 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.
- (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.

- (2018), Infernece in dynamic, nonparametric models of production: Central limit theorems for Malmquist indices. Discussion paper #2018/10, Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.
- (2019), Infernece in dynamic, nonparametric models of production with constant returns to scale and non-convex production sets. In progress.
- Mammen, E. (1992), *When Does Bootstrap Work? Asymptotic Results and Simulations*, Berlin: Springer-Verlag.
- 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.
- Simar, L. and A. Vanhems (2012), Probabilistic characterization of directional distances and their robust versions, *Journal of Econometrics* 166, 342–354.
- Simar, L., A. Vanhems, and P. W. Wilson (2012), Statistical inference for DEA estimators of directional distances, *European Journal of Operational Research* 220, 853–864.
- Simar, L. and P. W. Wilson (1998), Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models, *Management Science* 44, 49–61.
- (1999a), Some problems with the Ferrier/Hirschberg bootstrap idea, *Journal of Productivity Analysis* 11, 67–80.
- (1999b), Of course we can bootstrap DEA scores! But does it mean anything? Logic trumps wishful thinking, *Journal of Productivity Analysis* 11, 93–97.
- (2007), Estimation and inference in two-stage, semi-parametric models of productive efficiency, *Journal of Econometrics* 136, 31–64.
- (2011), 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.
- Thain, D., T. Tannenbaum, and M. Livny (2005), Distributed computing in practice: the Condor experience, *Concurrency and Computation: Practice and Experience* 17, 323–356.
- Wheelock, D. C. and P. W. Wilson (2008), Non-parametric, unconditional quantile estimation for efficiency analysis with an application to Federal Reserve check processing operations, *Journal of Econometrics* 145, 209–225.
- Wilson, P. W. (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.
- (2019), U.S. banking in the post-crisis era: New results from new methods, in C. Parmeter and R. Sickles, eds., *Methodological Contributions to the Advancement of Productivity and Efficiency Analysis*, Cham, Switzerland: Springer International Publishing AG. Forthcoming.

Table 1: Rejection Rates for RTS Test, Asymptotic Normality, 1 Split

n	δ	Case I						Case II					
		$p = 1, q = 1$		$p = 2, q = 1$		$p = 3, q = 1$		$p = 4, q = 1$		$p = 5, q = 1$			
		.10	.05	.10	.05	.10	.05	.10	.05	.10	.05	.10	.05
100	0.0	0.181	0.111	0.031	0.217	0.145	0.057	0.262	0.175	0.054	0.190	0.127	0.044
	0.1	0.196	0.114	0.038	0.223	0.151	0.058	0.264	0.182	0.061	0.190	0.132	0.044
	0.2	0.235	0.155	0.055	0.256	0.166	0.068	0.279	0.202	0.071	0.198	0.135	0.046
	0.3	0.336	0.244	0.100	0.331	0.213	0.096	0.323	0.230	0.091	0.230	0.144	0.052
	0.4	0.497	0.368	0.174	0.426	0.309	0.139	0.393	0.293	0.132	0.275	0.161	0.070
	0.6	0.710	0.598	0.342	0.594	0.471	0.259	0.548	0.416	0.245	0.370	0.257	0.106
	0.8	0.826	0.752	0.548	0.723	0.618	0.401	0.675	0.569	0.352	0.484	0.353	0.173
	1.4	0.977	0.949	0.835	0.925	0.878	0.720	0.901	0.851	0.686	0.699	0.601	0.377
1000	0.0	0.115	0.064	0.014	0.154	0.077	0.016	0.143	0.082	0.029	0.110	0.061	0.008
	0.1	0.179	0.093	0.022	0.176	0.098	0.025	0.161	0.089	0.036	0.111	0.060	0.009
	0.2	0.397	0.268	0.102	0.315	0.198	0.063	0.247	0.158	0.055	0.137	0.075	0.018
	0.3	0.789	0.684	0.412	0.591	0.442	0.226	0.473	0.332	0.149	0.212	0.114	0.039
	0.4	0.974	0.946	0.845	0.856	0.772	0.552	0.737	0.633	0.390	0.346	0.220	0.073
	0.6	0.999	0.999	0.996	0.994	0.982	0.946	0.958	0.942	0.828	0.612	0.471	0.230
	0.8	1.000	1.000	1.000	1.000	1.000	0.997	0.997	0.992	0.977	0.818	0.701	0.457
	1.4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.986	0.978	0.918

Table 2: Rejection Rates for RTS Test, Bootstrap, Average over 10 Splits

n	δ	Case I						Case II					
		$p = 1, q = 1$		$p = 2, q = 1$		$p = 3, q = 1$		$p = 4, q = 1$		$p = 5, q = 1$		$p = 5, q = 1$	
		.10	.05	.10	.05	.10	.05	.10	.05	.10	.05	.10	.05
100	0.0	0.075	0.034	0.010	0.118	0.057	0.011	0.145	0.079	0.014	0.145	0.082	0.019
	0.1	0.090	0.041	0.009	0.128	0.065	0.011	0.151	0.081	0.015	0.145	0.084	0.017
	0.2	0.175	0.082	0.016	0.174	0.087	0.016	0.194	0.102	0.022	0.172	0.091	0.017
	0.3	0.390	0.249	0.058	0.289	0.181	0.043	0.290	0.172	0.041	0.226	0.122	0.030
	0.4	0.708	0.521	0.213	0.508	0.344	0.126	0.451	0.302	0.100	0.335	0.199	0.063
	0.6	0.957	0.882	0.651	0.828	0.715	0.403	0.768	0.626	0.329	0.625	0.457	0.205
	0.8	0.997	0.985	0.879	0.966	0.917	0.717	0.944	0.865	0.616	0.838	0.720	0.432
	1.4	1.000	1.000	0.995	0.999	0.999	0.986	0.998	0.997	0.976	0.997	0.985	0.892
1000	0.0	0.098	0.047	0.009	0.135	0.061	0.014	0.138	0.083	0.015	0.152	0.075	0.026
	0.1	0.229	0.132	0.037	0.214	0.123	0.026	0.186	0.097	0.025	0.175	0.093	0.028
	0.2	0.859	0.731	0.484	0.625	0.496	0.246	0.466	0.329	0.130	0.257	0.161	0.053
	0.3	1.000	1.000	0.995	0.984	0.959	0.858	0.898	0.825	0.632	0.519	0.389	0.186
	0.4	1.000	1.000	1.000	1.000	1.000	0.999	0.988	0.988	0.974	0.851	0.767	0.555
	0.6	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.996	0.996	0.981	0.975	0.958
	0.8	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.996	0.996	0.991	0.989	0.989
	1.4	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.997	0.997	0.994	0.994	0.994

Table 3: Rejection Rates for RTS Test, Bootstrap, KS-statistic, 10 Splits

n	δ	Case I						Case II					
		$p = 1, q = 1$		$p = 2, q = 1$		$p = 3, q = 1$		$p = 4, q = 1$		$p = 5, q = 1$		$p = 5, q = 1$	
		.10	.05	.01	.10	.05	.01	.10	.05	.01	.10	.05	.01
100	0.0	0.080	0.031	0.007	0.119	0.057	0.007	0.137	0.069	0.015	0.124	0.061	0.017
	0.1	0.094	0.042	0.008	0.130	0.067	0.009	0.142	0.072	0.013	0.125	0.061	0.013
	0.2	0.170	0.082	0.019	0.167	0.091	0.015	0.173	0.094	0.022	0.136	0.070	0.014
	0.3	0.379	0.225	0.059	0.263	0.163	0.040	0.255	0.147	0.030	0.181	0.090	0.021
	0.4	0.630	0.456	0.194	0.447	0.310	0.098	0.401	0.244	0.084	0.276	0.155	0.041
	0.6	0.923	0.832	0.530	0.766	0.624	0.331	0.694	0.542	0.258	0.504	0.369	0.149
	0.8	0.987	0.962	0.814	0.930	0.858	0.627	0.893	0.793	0.493	0.747	0.615	0.337
1000	1.4	1.000	0.997	0.982	0.999	0.994	0.950	0.999	0.993	0.944	0.977	0.942	0.801
	0.0	0.101	0.055	0.010	0.109	0.055	0.013	0.105	0.064	0.017	0.123	0.068	0.016
	0.1	0.184	0.117	0.034	0.166	0.096	0.027	0.144	0.081	0.026	0.144	0.078	0.017
	0.2	0.775	0.638	0.355	0.542	0.404	0.157	0.381	0.257	0.095	0.201	0.130	0.025
	0.3	0.998	0.992	0.972	0.953	0.907	0.727	0.841	0.727	0.484	0.421	0.294	0.124
	0.4	1.000	1.000	1.000	0.999	0.999	0.989	0.989	0.979	0.925	0.740	0.630	0.379
	0.6	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.972	0.954	0.900
1000	0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.994	0.993	0.984
	1.4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998

Table 4: Rejection Rates for Convexity Test, Asymptotic Normality, 1 Split

n	δ	Case I						Case II					
		$p = 1, q = 1$			$p = 2, q = 1$			$p = 3, q = 1$			$p = 4, q = 1$		
		.10	.05	.01	.10	.05	.01	.10	.05	.01	.10	.05	.01
100	1.4*	0.112	0.073	0.024	0.094	0.054	0.018	0.100	0.042	0.010	0.106	0.064	0.024
	0.0	0.271	0.187	0.079	0.217	0.125	0.052	0.164	0.111	0.048	0.158	0.109	0.034
	0.1	0.281	0.195	0.088	0.218	0.133	0.053	0.165	0.110	0.049	0.158	0.108	0.034
	0.2	0.311	0.225	0.097	0.225	0.140	0.060	0.171	0.117	0.053	0.161	0.111	0.036
	0.3	0.376	0.281	0.149	0.256	0.168	0.068	0.187	0.125	0.057	0.164	0.117	0.039
	0.4	0.466	0.371	0.221	0.289	0.201	0.091	0.207	0.136	0.066	0.177	0.132	0.047
	0.6	0.628	0.557	0.404	0.379	0.284	0.148	0.262	0.177	0.081	0.206	0.150	0.064
	1.4	0.943	0.916	0.856	0.717	0.637	0.475	0.567	0.463	0.300	0.460	0.341	0.189
1000	1.4*	0.116	0.064	0.011	0.108	0.056	0.013	0.106	0.056	0.015	0.119	0.083	0.027
	0.0	0.184	0.111	0.034	0.185	0.112	0.029	0.169	0.097	0.031	0.176	0.109	0.036
	0.1	0.209	0.137	0.044	0.186	0.117	0.030	0.174	0.100	0.032	0.180	0.112	0.038
	0.2	0.361	0.264	0.123	0.211	0.135	0.037	0.192	0.115	0.036	0.191	0.118	0.042
	0.3	0.696	0.597	0.390	0.288	0.191	0.065	0.223	0.142	0.045	0.219	0.131	0.050
	0.4	0.931	0.884	0.774	0.436	0.315	0.148	0.302	0.201	0.080	0.263	0.173	0.069
	0.6	1.000	0.996	0.988	0.719	0.610	0.402	0.494	0.364	0.172	0.412	0.296	0.138
	1.4	1.000	1.000	1.000	0.997	0.983	0.956	0.916	0.864	0.708	0.851	0.773	0.583

*Data generated by the DGP used for RTS tests; the production set is strictly convex.

Table 5: Rejection Rates for Convexity Test, Bootstrap, Average over 10 Splits

n	δ	Case I						Case II					
		$p = 1, q = 1$			$p = 2, q = 1$			$p = 3, q = 1$			$p = 4, q = 1$		
		.10	.05	.01	.10	.05	.01	.10	.05	.01	.10	.05	.01
100	1.4*	0.069	0.035	0.005	0.073	0.038	0.010	0.074	0.039	0.007	0.082	0.045	0.008
	0.0	0.209	0.116	0.031	0.185	0.110	0.033	0.193	0.126	0.029	0.166	0.097	0.026
	0.1	0.216	0.124	0.033	0.190	0.112	0.034	0.197	0.130	0.031	0.167	0.102	0.027
	0.2	0.251	0.162	0.041	0.221	0.126	0.041	0.212	0.142	0.042	0.176	0.110	0.030
	0.3	0.351	0.244	0.088	0.280	0.175	0.063	0.248	0.159	0.056	0.202	0.125	0.039
	0.4	0.554	0.424	0.217	0.380	0.270	0.115	0.318	0.214	0.086	0.236	0.154	0.052
	0.6	0.885	0.807	0.619	0.686	0.571	0.342	0.566	0.420	0.213	0.388	0.272	0.115
	1.4	1.000	1.000	1.000	1.000	1.000	0.991	0.991	0.983	0.931	0.949	0.907	0.774
1000	1.4*	0.069	0.028	0.004	0.089	0.033	0.003	0.091	0.045	0.010	0.106	0.049	0.012
	0.0	0.158	0.075	0.022	0.169	0.088	0.021	0.178	0.098	0.028	0.210	0.119	0.035
	0.1	0.185	0.100	0.033	0.182	0.097	0.023	0.187	0.103	0.034	0.218	0.126	0.038
	0.2	0.433	0.327	0.141	0.238	0.135	0.043	0.228	0.137	0.044	0.245	0.136	0.042
	0.3	0.951	0.908	0.745	0.500	0.348	0.159	0.363	0.237	0.096	0.338	0.228	0.078
	0.4	1.000	1.000	0.997	0.901	0.819	0.612	0.683	0.541	0.268	0.565	0.425	0.203
	0.6	1.000	1.000	1.000	1.000	1.000	0.998	0.987	0.973	0.902	0.941	0.901	0.741
	1.4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

*Data generated by the DGP used for RTS tests; the production set is strictly convex.

Table 6: Rejection Rates for Convexity Test, Bootstrap, KS-statistic, 10 Splits

n	δ	Case I						Case II					
		$p = 1, q = 1$			$p = 2, q = 1$			$p = 3, q = 1$			$p = 4, q = 1$		
		.10	.05	.01	.10	.05	.01	.10	.05	.01	.10	.05	.01
100	1.4*	0.114	0.066	0.009	0.143	0.078	0.021	0.113	0.055	0.014	0.136	0.079	0.022
	0.0	0.172	0.095	0.028	0.154	0.093	0.026	0.150	0.096	0.022	0.135	0.068	0.016
	0.1	0.177	0.103	0.030	0.160	0.096	0.027	0.152	0.096	0.023	0.140	0.069	0.018
	0.2	0.208	0.126	0.042	0.178	0.112	0.033	0.164	0.098	0.026	0.144	0.072	0.020
	0.3	0.296	0.190	0.067	0.221	0.143	0.046	0.195	0.115	0.035	0.151	0.081	0.024
	0.4	0.476	0.338	0.141	0.321	0.205	0.080	0.248	0.163	0.056	0.172	0.098	0.030
	0.6	0.806	0.703	0.459	0.568	0.460	0.216	0.393	0.276	0.119	0.230	0.146	0.054
	1.4	0.998	0.998	0.989	0.968	0.928	0.805	0.853	0.752	0.527	0.659	0.541	0.286
1000	1.4*	0.085	0.043	0.009	0.084	0.039	0.009	0.099	0.050	0.013	0.094	0.050	0.011
	0.0	0.150	0.076	0.014	0.161	0.076	0.014	0.155	0.089	0.022	0.186	0.105	0.022
	0.1	0.169	0.093	0.019	0.165	0.080	0.016	0.158	0.091	0.022	0.191	0.106	0.023
	0.2	0.375	0.258	0.087	0.212	0.117	0.030	0.191	0.111	0.030	0.217	0.121	0.026
	0.3	0.896	0.818	0.549	0.410	0.279	0.109	0.297	0.191	0.062	0.292	0.193	0.060
	0.4	1.000	0.998	0.978	0.801	0.684	0.415	0.575	0.427	0.185	0.476	0.326	0.141
	0.6	1.000	1.000	1.000	0.999	0.996	0.977	0.958	0.914	0.761	0.860	0.772	0.553
	1.4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999

*Data generated by the DGP used for RTS tests; the production set is strictly convex.

Table 7: Rejection Rates for Separability Test ($r = 1$), Asymptotic Normality, 1 Split

n	δ	Case I						Case II					
		$p = 1, q = 1$			$p = 2, q = 1$			$p = 2, q = 2$			$p = 3, q = 2$		
		.10	.05	.01	.10	.05	.01	.10	.05	.01	.10	.05	.01
100	0.0	0.153	0.087	0.024	0.190	0.120	0.049	0.160	0.101	0.043	0.181	0.121	0.048
	0.1	0.180	0.120	0.048	0.201	0.133	0.053	0.168	0.105	0.050	0.192	0.129	0.057
	0.2	0.267	0.204	0.084	0.224	0.155	0.069	0.188	0.127	0.054	0.202	0.141	0.053
	0.3	0.386	0.275	0.149	0.256	0.185	0.088	0.210	0.143	0.065	0.203	0.149	0.057
	0.4	0.506	0.393	0.229	0.293	0.230	0.118	0.239	0.163	0.067	0.209	0.153	0.060
	0.6	0.696	0.598	0.399	0.364	0.292	0.186	0.276	0.208	0.099	0.203	0.155	0.055
	0.8	0.795	0.725	0.553	0.459	0.372	0.241	0.320	0.238	0.128	0.200	0.152	0.063
	1.0	0.862	0.804	0.655	0.527	0.440	0.307	0.364	0.275	0.152	0.204	0.144	0.058
	2.0	0.956	0.939	0.859	0.718	0.637	0.503	0.485	0.407	0.248	0.209	0.144	0.065
200	0.0	0.147	0.069	0.020	0.151	0.099	0.031	0.144	0.080	0.031	0.141	0.080	0.023
	0.1	0.192	0.126	0.045	0.170	0.118	0.043	0.155	0.093	0.030	0.154	0.088	0.030
	0.2	0.380	0.266	0.128	0.231	0.173	0.086	0.193	0.121	0.044	0.163	0.091	0.029
	0.3	0.546	0.462	0.278	0.353	0.275	0.156	0.249	0.167	0.077	0.162	0.097	0.032
	0.4	0.729	0.620	0.440	0.482	0.390	0.249	0.324	0.237	0.110	0.169	0.110	0.034
	0.6	0.896	0.831	0.712	0.670	0.599	0.448	0.496	0.390	0.210	0.191	0.128	0.038
	0.8	0.961	0.937	0.843	0.786	0.720	0.598	0.617	0.526	0.324	0.224	0.147	0.049
	1.0	0.986	0.975	0.926	0.873	0.816	0.719	0.717	0.617	0.434	0.275	0.181	0.073
	2.0	0.997	0.996	0.991	0.980	0.964	0.914	0.858	0.801	0.643	0.403	0.299	0.145

Table 8: Rejection Rates for Separability Test ($r = 1$), Bootstrap, Average over 10 Splits[illegible]

Table 9: Rejection Rates for Separability Test ($r = 1$), Bootstrap, KS-statistic, 10 Splits

n	δ	Case I						Case II					
		$p = 1, q = 1$			$p = 2, q = 1$			$p = 2, q = 2$			$p = 3, q = 2$		
		.10	.05	.01	.10	.05	.01	.10	.05	.01	.10	.05	.01
100	0.0	0.118	0.068	0.017	0.124	0.071	0.016	0.119	0.058	0.015	0.110	0.059	0.013
	0.1	0.168	0.091	0.030	0.148	0.079	0.020	0.139	0.077	0.016	0.112	0.054	0.016
	0.2	0.323	0.235	0.093	0.195	0.123	0.042	0.165	0.086	0.022	0.112	0.055	0.017
	0.3	0.565	0.459	0.260	0.271	0.162	0.066	0.184	0.117	0.024	0.119	0.062	0.014
	0.4	0.769	0.682	0.469	0.362	0.259	0.111	0.231	0.142	0.050	0.126	0.063	0.019
	0.6	0.949	0.909	0.788	0.551	0.448	0.264	0.340	0.243	0.101	0.129	0.078	0.018
	0.8	0.982	0.975	0.923	0.674	0.581	0.434	0.451	0.349	0.176	0.155	0.090	0.028
	1.0	0.995	0.991	0.975	0.781	0.685	0.528	0.523	0.438	0.266	0.172	0.097	0.033
	2.0	1.000	1.000	1.000	0.943	0.909	0.814	0.775	0.700	0.520	0.217	0.138	0.047
200	0.0	0.105	0.048	0.011	0.118	0.065	0.014	0.108	0.051	0.015	0.124	0.061	0.017
	0.1	0.221	0.141	0.053	0.153	0.093	0.030	0.122	0.064	0.018	0.112	0.057	0.015
	0.2	0.627	0.507	0.298	0.350	0.225	0.119	0.197	0.123	0.034	0.116	0.061	0.018
	0.3	0.901	0.838	0.691	0.598	0.489	0.294	0.348	0.250	0.111	0.131	0.069	0.019
	0.4	0.983	0.971	0.908	0.796	0.715	0.548	0.542	0.417	0.216	0.137	0.080	0.028
	0.6	1.000	1.000	0.993	0.957	0.942	0.868	0.811	0.731	0.558	0.207	0.121	0.040
	0.8	1.000	1.000	1.000	0.992	0.989	0.969	0.929	0.899	0.787	0.288	0.190	0.077
	1.0	1.000	1.000	1.000	0.998	0.998	0.994	0.970	0.950	0.898	0.362	0.262	0.128
	2.0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.989	0.640	0.535	0.363

Figure 1: Distribution of p -Values Obtained with Single Sample-Split, RTS Test, $n = 1000$, $p = 5$, $q = 1$

Figure 2: Distribution of p -Values across $s \in \{2, 10, 100\}$ Splits in a Single Sample, RTS Test, $n = 100$, $p = 5$, $q = 1$

$\delta = 0.0$

$\delta = 1.4$

NOTE: In both frames, a single sample is split s times to produce a p -value, and this is repeated 100 times to obtain 100 p -values.