Rønn-Nielsen, Anders; Kronborg, Dorte; Asmild, Mette

Working Paper

Exact tests on returns to scale and comparisons of production frontiers in nonparametric models

IFRO Working Paper, No. 2019/04

Provided in Cooperation with:
Department of Food and Resource Economics (IFRO), University of Copenhagen

Suggested Citation: Rønn-Nielsen, Anders; Kronborg, Dorte; Asmild, Mette (2019) : Exact tests on returns to scale and comparisons of production frontiers in nonparametric models, IFRO Working Paper, No. 2019/04, University of Copenhagen, Department of Food and Resource Economics (IFRO), Copenhagen

This Version is available at:
https://hdl.handle.net/10419/204433

DEPARTMENT OF FOOD AND RESOURCE ECONOMICS
UNIVERSITY OF COPENHAGEN

IFRO Working Paper

Exact tests on returns to scale and comparisons of production frontiers in nonparametric models

Anders Rønn-Nielsen
Dorte Kronborg
Mette Asmild

2019 / 04

IFRO Working Paper 2019 / 04

Exact tests on returns to scale and comparisons of production frontiers in nonparametric models

Authors: Anders Rønn-Nielsen, Dorte Kronborg, Mette Asmild

JEL-classification: C12, C14, C44, C46, C61, D 24

Published: May 2019

See the full series IFRO Working Paper here:
www.ifro.ku.dk/english/publications/ifro_series/working_papers/

Department of Food and Resource Economics (IFRO)
University of Copenhagen
Rolighedsvej 25
DK 1958 Frederiksberg DENMARK
www.ifro.ku.dk/english/

Exact tests on returns to scale and comparisons of production frontiers in nonparametric models

Anders Rønn-Nielsen & Dorte Kronborg
Center for Statistics, Department of Finance
Copenhagen Business School

Mette Asmild
Institute of Food and Resource Economics
University of Copenhagen

Abstract

When benchmarking production units by non-parametric methods like data envelopment analysis (DEA), an assumption has to be made about the returns to scale of the underlying technology. Moreover, it is often also relevant to compare the frontiers across samples of producers. Until now, no exact tests for examining returns to scale assumptions in DEA, or for test of equality of frontiers, have been available. The few existing tests are based on asymptotic theory relying on large sample sizes, whereas situations with relatively small samples are often encountered in practical applications.

In this paper we propose three novel tests based on permutations. The tests are easily implementable from the algorithms provided, and give exact significance probabilities as they are not based on asymptotic properties. The first of the proposed tests is a test for the hypothesis of constant returns to scale in DEA. The others are tests for general frontier differences and whether the production possibility sets are, in fact, nested. The theoretical advantages of permutation tests are that they are appropriate for small samples and have the correct size. Simulation studies show that the proposed tests do, indeed, have the

correct size and furthermore higher power than the existing alternative tests based on asymptotic theory.

Keywords: Data Envelopment Analysis (DEA), returns to scale, equality of production frontiers, exact tests, permutations.

Correspondence: Anders Rønn-Nielsen, Center for Statistics, Department of Finance, Copenhagen Business School, Solbjerg Plads 3, 2000 Frederiksberg, Denmark, email: aro.fi@cbs.dk

1 Introduction

A widely used method for benchmarking of a set of production units is the non-parametric Data Envelopment Analysis (DEA) approach, which estimates the production possibility set as a convex envelopment of the observed set of input and output quantities (Farrell, 1957; Charnes *et al.*, 1978). It is well known that the DEA method results in a biased estimate of the production frontier and also that the estimated efficiencies are correlated, introducing the need for extra caution when these are used for further statistical analysis. The statistical properties of the estimated efficiencies have been the subject of numerous studies, and in a recent paper Kneip *et al.* (2015) developed asymptotic results usable for inference for mean efficiencies.

Most of the developed theory is concentrated on inference for individual or mean efficiencies, whereas the theory on another important issue - comparison of production frontiers - is sparse. Methods for comparison of mean efficiencies for two independent samples, based on asymptotic normal approximations, are developed in Kneip *et al.* (2016), focusing on testing equality of means across two independent groups, both when the efficiencies for the two samples are measured relative to a common frontier or to different frontiers. The former test is formally a test for a composite hypothesis, namely whether the groups are facing the same frontier *and* the mean efficiencies are the same in the two groups and therefore it is not possible to determine whether rejection is due to different frontiers or different mean efficiencies.

Here we consider the situation where the production frontiers across independent samples are to be compared. Equality of production frontiers across samples is an implicit (but often not tested) assumption when applying DEA, and the meaningfulness of many two-stage analyses rely on the assumption of equal support for the production possibility sets, also known as the 'separability' condition c.f. e.g. Simar and Wilson (2007), Simar and Wilson (2015) and recently Daraio *et al.* (2018). Daraio *et al.* (2018) develop central limit theorems for means of conditional efficiencies and propose an asymptotic test for the 'separability' condition when conditioning on a continuous environmental variable. Further, when dealing with a discrete dichotomous environmental variable the proposed test is basically the same as in the continuous case expect for the bias-correction method. To avoid a degenerate test statistic, the method relies on comparison of means of efficiencies calculated on random splits of the sample: one sample used for estimating efficiencies assuming separability and the other sample used for estimating conditional efficiencies. When splitting the sample one does not utilize all the available

information which, especially when dealing with small samples, can be critical. Further, if for example the two frontiers intersect, the proposed test will not necessarily reject the separability hypothesis.

Another important issue when studying productivity is the assumption about returns to scale made by the researcher. Methods for deciding the appropriate technology assumptions have been the subject of several papers, lately Kneip *et al.* (2016), but among others Simar and Wilson (2002) and Banker (1996) also adress the importance of imposing the correct assumption. Kneip *et al.* (2016) propose tests for returns to scale, with the hypothesis being constant returns to scale (CRS) within a model assuming variable returns to scale (VRS), based on the asymptotic distribution of the difference of sample means of efficiencies calculated assuming CRS and VRS respectively. Again, to avoid a degenerate test statistic under the CRS hypothesis the sample is randomly split into two samples, one used for estimating the VRS efficiencies and one for the CRS efficiencies. This procedure, however, requires a substantial amount of data which may not necessarily be available in practice. Moreover, the significance probabilities derived from the above mentioned method rely heavily on the suggested initial random split. The impact of this effect is larger the smaller the sample. Even when applying the suggested bootstrap confidence intervals based on an asymptotic pivotal test statistic giving an asymptotic refinement of the test, a considerable amount of data is still needed to obtain a powerful test with correct size.

In this paper we introduce permutation tests for inference in nonparametric production frontier models. The use of permutation tests for exact inference was originally proposed by Fisher (1935), and subsequently mathematically formalized (see e.g. Lehmann and Romano (2005)). Recently such methods have gained popularity due to increased computational possibilities.

Here we will describe how permutation tests can be formulated when testing hypotheses regarding returns to scale, with the hypothesis being constant returns to scale. The power and size of the permutation test will be investigated through Monte-Carlo simulations and the performance of the test will be compared to that of the test proposed in Kneip *et al.* (2016).

Further, we introduce and examine two test statistics which can be used for comparison of frontiers across separate (independent) groups of production units. The result of the first test simply indicates whether the frontiers are likely to be different (including intersecting frontiers) and the second test supplements the first as it is designed to detect whether one group overall has better production possibilities

than the other (nested frontiers). The tests are based on the relative locations of the frontiers and account for differences in bias obtained in DEA estimated frontiers by the use of jackknife methods.

The structure of the paper is as follows: First we introduce our notation and the production frontier methodology in Section 2. The method of inference on returns to scale is described in Section 3 and Section 4 describes the methods for comparison of frontiers for independent groups. In Section 5 the results of a series of Monte Carlo experiments for evaluating the performance of the described tests are presented and compared with an existing asymptotic test. Section 6 concludes the paper.

2 The non-parametric frontier model

Let the vector of input (x) and output (y) quantities be denoted by $(x, y) \in \mathbb{R}_+^{p+q}$. Using standard notation the feasible set of input-output combinations, i.e the production possibility set, is

$$\Psi = \{(x, y) \in \mathbb{R}_+^{p+q} \,|\, x \text{ can produce } \ y\}.$$

The production possibility set is assumed to be closed, convex and satisfying strong disposablity in both inputs and outputs. The efficient frontier of Ψ is given by

$$\Psi^\delta = \{(x, y) \in \Psi \,|\, (\gamma^{-1}x, \gamma y) \notin \Psi, \ \forall \gamma > 1\}.$$

Efficiency of a given production unit is often measured by either the Farrell input index or the corresponding output index, with the input index given by

$$\theta(x, y) = \inf\{\theta > 0 | \, (\theta x, y) \in \Psi\},$$

and the output index given by

$$\vartheta(x, y) = \sup\{\vartheta > 0 | \, (x, \vartheta y) \in \Psi\}.$$

If $\theta = 1$ the firm is said to be technically efficient in the input direction while if $\vartheta = 1$ the firm is technically efficient in the output direction. Otherwise, the firm is referred to as technically inefficient in either the input and/or the output direction. Technical efficiency can alternatively be measured in hyperbolic distance

$$\gamma(x, y) = \inf\{\gamma > 0 | \, (\gamma x, \gamma^{-1} y) \in \Psi\}.$$

Various assumptions about returns to scale are possible. Here we concentrate on the two most commonly used assumptions: Constant returns to scale, i.e the production can be scaled arbitrarily up and down

$$(x, y) \in \Psi \text{ then } \lambda(x, y) \in \Psi, \forall \lambda > 0, \tag{1}$$

and variable returns to scale where rescaling of all points in the production possibility set is not necessarily possible ($\lambda = 1$). Constant returns to scale implies that the efficiency is invariant under simultaneous scaling of both inputs and outputs versus the variable returns to scale scenario where the efficiency varies when scaling the units. We will denote the input efficiencies when assuming CRS or VRS as θ_{CRS} and θ_{VRS} respectively.

2.1 Statistical model and estimation

In practice, the production possibility set Ψ and the corresponding efficiencies are unobserved and estimated from a set of n observations, $i = 1, \ldots, n$. The observations (X_i, Y_i) are assumed to be independent and identically distributed on Ψ, such that (X_i, Y_i) has distribution F with density on Ψ for all $i = 1, \ldots, n$.

Along the lines of Farrell (1957) and Charnes *et al.* (1978) we use the DEA approach and estimate the production possibility set assuming constant return to scale by

$$\hat{\Psi} = \{(x, y) \in \mathbb{R}_+^{p+q} \,|\exists \omega \in \mathbb{R}_+^n : x \geq \mathbf{X}\omega, y \leq \mathbf{Y}\omega\},$$

where $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$. The input efficiency index for (x, y) is estimated by

$$\hat{\theta}_{CRS}(x, y) = \min_{\theta, \omega} \{\theta \,|\, \theta x \geq \mathbf{X}\omega, y \leq \mathbf{Y}\omega, \, \omega \in \mathbb{R}_+^n\}.$$

Analogously, the production possibility set assuming VRS can be estimated and the equivalent DEA estimator for the efficiencies in the input direction are

$$\hat{\theta}_{VRS}(x, y) = \min_{\theta, \omega} \{\theta \,|\, \theta x \geq \mathbf{X}\omega, y \leq \mathbf{Y}\omega, \, \sum \omega_i = 1, \omega \in \mathbb{R}_+^n\}.$$

In Section 4 we consider frontiers from two different production technologies. Let the efficient frontier be indexed by a subscript $g \in \{1, 2\}$, such that Ψ_g^δ denotes the frontier for technology g. Assume that there exists a distribution F_g with density

on Ψ_g. Let (X^g, Y^g) denote random variables with distribution F_g and assume that the random variables, (X^1, Y^1), and (X^2, Y^2) are independent. Note that if the distributions F_g are equal then the frontiers Ψ_g^δ are equal.

Similarly to the notation for one production plan, the vector $(X_1^g, \ldots, X_{n_g}^g)$ is denoted $\mathbf{X}^g$ and the vector $(Y_1^g, \ldots, Y_{n_g}^g)$ is denoted $\mathbf{Y}^g$ for $g \in \{1, 2\}$.

3 Inference on returns to scale

Here we explain how permutation tests can be used to analyze whether a production possibility set can be assumed to fulfil an assumption of constant returns to scale, within a model assuming variable returns to scale. Assume that n independent and identically distributed observations $(X_i, Y_i)_{i=1\ldots n}$, with a common production possibility set Ψ, are given. For the test method below to work we need a few additional assumptions. Therefore, let $Z_i = \|Y_i\|$ and $V_i = \frac{Y_i}{Z_i}$. Furthermore, let $W_i = \frac{X_i}{\|X_i\|}$, let X_i^δ be such that (X_i^δ, Y_i) is a point belonging to the frontier, and let Θ_i be the efficiency of (X_i, Y_i). Thus $(X_i^\delta, Y_i) \in \Psi^\delta$ and $X_i = \frac{X_i^\delta}{\Theta_i}$. Note that X_i^δ is deterministically known from Z_i, V_i and W_i. Assume now for each $i = 1, \ldots, n$ that Z_i is independent of (V_i, W_i, Θ_i). This means, that we assume that the length of the output vector is independent of the output direction, the input direction, as well as the efficiency.

For an individual observation (x, y)

$$F_{rts}(x, y) = \frac{\hat{\theta}_{CRS}(x, y)}{\hat{\theta}_{VRS}(x, y)},$$

is a measure of the difference between the estimated frontiers in the input direction, x, when assuming CRS and VRS respectively. The overall difference between the estimated frontiers can be measured by the geometric mean of these n ratios and is calculated as

$$T_{rts} = \prod_{i=1}^{n} F_{rts}(X_i, Y_i)^{\frac{1}{n}}. \tag{2}$$

This statistic can be used to test the hypothesis of constant returns to scale within a model assuming variable returns to scale. If a value of T_{rts} significantly below one is observed, the hypothesis of CRS is rejected.

The distribution of the test statistic T_{rts} is unknown under the hypothesis of constant returns to scale. However the significance of the hypothesis of CRS can be evaluated using a permutation test. To implement the permutation test, for each $j = 1, \ldots, N$ we use the following procedure to construct T^j_{rts} :

1. For each observation (X_i, Y_i), $i = 1, \ldots, n$, calculate $Z_i = \|Y_i\|$. Define $U_i = \frac{X_i}{Z_i}$ and $V_i = \frac{Y_i}{Z_i}$.
2. Let $(\tilde{Z}_1, \ldots, \tilde{Z}_n)$ be a random permutation of $(Z_1, \ldots, Z_n)$.
3. Define $\tilde{X}_i = \tilde{Z}_i \cdot U_i$ and $\tilde{Y}_i = \tilde{Z}_i \cdot V_i$.
4. Calculate T^j_{rts} by applying (2) to the new dataset $(\tilde{X}_i, \tilde{Y}_i)_{i=1\ldots n}$.

The significance probability is obtained by comparing the observed value of T_{rts} with the empirical distribution of T^j_{rts}, $j = 1, \ldots, N$. Since small values provide evidence against the hypothesis of CRS, the significance probability is more precisely calculated as

$$\hat{p} = \frac{1}{N} \sum_{j=1}^{N} 1_{\{T^j_{rts} \leq T_{rts}\}}, \tag{3}$$

i.e., the proportion of T^j_{rts}'s smaller than T_{rts}.

Under the assumption of CRS, the frontier point X^δ_i in fact has the form

$$X^\delta_i = Z_i \cdot H(V_i, W_i),$$

for some appropriate function H. We now have that X_i and Y_i are given as

$$X_i = \frac{Z_i \cdot H(V_i, W_i)}{\Theta_i} \qquad \text{and} \qquad Y_i = Z_i \cdot V_i,$$

which together with the assumptions of independence between Z_i and (V_i, W_i, θ_i), and all $(X_i, Y_i)_{i=1\ldots n}$ being independent and identically distributed, leads to the conclusion that the observations $(\tilde{X}_i, \tilde{Y}_i)_{i=1,\ldots,n}$ are independent and distributed with the same distribution F as the original observations (X_i, Y_i).

According to Lehmann and Romano (2005), the significance probability p, as calculated in (3), satisfies that

$$P(\hat{p} \leq u) \leq u,$$

for all $u \in (0, 1)$. That implies that the rejection rate is controlled, i.e. rejecting the hypothesis when $\hat{p} < \alpha$ will give an actual rejection rate that is not higher

than α when the CRS hypothesis is true.

If (1) is satisfied for all $0 < \lambda \leq 1$, the technology is called non-increasing returns to scale while $\lambda \geq 1$ is referred to as non-decreasing returns to scale. If relevant, both non-increasing and non-decreasing returns to scale can by used as models when testing CRS by applying (2) with the relevant efficiencies in the denominator for the $F_{rts}(x, y)$'s.

4 Inference on equality of frontiers

In this section we consider two production technologies, with corresponding production possibility sets Ψ_1 and Ψ_2 and independent observation vectors $(\mathbf{X}^1, \mathbf{Y}^1)$ and $(\mathbf{X}^2, \mathbf{Y}^2)$. Without loss of generality assume that $n_1 \leq n_2$, where n_g is the number of observations in $(\mathbf{X}^g, \mathbf{Y}^g)$ for $g = 1, 2$. We shall assume that the two production technologies both are CRS. For simplicity of notation $\hat{\theta}_g$, $g = 1, 2$, is used for the estimated efficiency regardless of the relevant technology. The method outlined below can, with only a slight change in notation, be adapted to the case of output oriented efficiencies. If VRS is the appropriate technology the hyperbolic efficiency measure can be used to ensure the test statistics below being well defined.

We formulate two tests for the hypothesis about equality of the production frontiers: $\Psi_1^\delta = \Psi_2^\delta$, or more precisely if the two distributions F_1 and F_2 are equal. However, the test statistics in the following are only designed to detect differences of the distributions F_1 and F_2 close to the frontiers, or simply differences between the frontiers Ψ_1^δ and Ψ_2^δ. Confer Asmild *et al.* (2018) for a similar discussion.

The frontier difference test statistic (GT_{diff}) defined in (e) below is the general test designed to detect differences between the frontiers, whereas the second test statistic (GT_{nest}) is designed to detect whether one of the groups has better production possibilities than the other i.e whether one of the technologies is nested within the other.

For the first test we define the difference between the two frontiers in the direction of an observation (x, y) as the proportion of the larger of the distances from the observation to one of the estimated frontiers to the smaller distance,

$$F_{diff}(x, y) = \frac{\max_{g\in\{1,2\}}\hat{\theta}_g(x, y)}{\min_{g\in\{1,2\}}\hat{\theta}_g(x, y)}\,. \tag{4}$$

The second test considers which frontier corresponds to the best production pos-

sibilities and thus we simply use the ratio of the efficiencies for (x, y) relative to each of the two frontiers

$$F_{nest}(x, y) = \frac{\hat{\theta}_1(x, y)}{\hat{\theta}_2(x, y)}\,. \tag{5}$$

The test statistics are calculated using the following jackknife procedure: For $k \in 1 \ldots m$

a. Draw randomly without replacement n_1 observations from the sample $(\mathbf{X}^2, \mathbf{Y}^2)$. Denote these $(\bar{\mathbf{X}}^2, \bar{\mathbf{Y}}^2)$.

b. Estimate the DEA frontiers for each of the datasets $(\mathbf{X}^1, \mathbf{Y}^1)$ and $(\bar{\mathbf{X}}^2, \bar{\mathbf{Y}}^2)$.

c. Calculate $F_{diff}(x, y)$ and $F_{nest}(x, y)$ for each of the $2n_1$ observations (X_i^1, Y_i^1) and $(\bar{X}_i^2, \bar{Y}_i^2)$ for $i = 1 \ldots n_1$.

d. Calculate the geometric mean of the geometric means of the frontier differences.

$$T_{diff}^k = \prod_{i=1}^{n_1} F_{diff}(X_i^1, Y_i^1))^{\frac{1}{2n_1}} \times \prod_{i=1}^{n_1} F_{diff}(\bar{X}_i^2, \bar{Y}_i^2))^{\frac{1}{2n_1}}, \tag{6}$$

and similarly for the nested test calculate the geometric mean

$$T_{nest}^k = \prod_{i=1}^{n_1} F_{nest}(X_i^1, Y_i^1))^{\frac{1}{2n_1}} \times \prod_{i=1}^{n_1} F_{nest}(\bar{X}_i^2, \bar{Y}_i^2))^{\frac{1}{2n_1}}. \tag{7}$$

e. Repeat (a)-(d) m times and calculate the geometric means GT_{diff} of the T_{diff}^k's and GT_{nest} of the T_{nest}^k's respectively.

Both statistics are positive and close to one if the two frontiers are equal. The distributions of the test statistics GT_{diff} and GT_{nest} under the hypotheses are unknown, but whether the observed value of each test statistic are extreme can be evaluated using permutation tests. The permutation test based on N permutations is performed as follows: For each $j \in 1, \ldots, N$

1. Permute all $n = n_1 + n_2$ observations and divide the n observations randomly into two groups of size n_1 and n_2 respectively. Let $(\tilde{\mathbf{X}}_j^1, \tilde{\mathbf{Y}}_j^1)$ and $(\tilde{\mathbf{X}}_j^2, \tilde{\mathbf{Y}}_j^2)$ be the two new sets of independent observations of sizes n_1 and n_2.

2. Calculate GT_{diff}^j and GT_{nest}^j as described in (a)-(e) above.

Under the hypothesis of equal frontiers the sets of observations $(\tilde{\mathbf{X}}_j^1, \tilde{\mathbf{Y}}_j^1)$ and $(\tilde{\mathbf{X}}_j^2, \tilde{\mathbf{Y}}_j^2)$ are independent and identically distributed and follow the same distribution as the observations in the dataset. The significance probability for the

hypothesis can therefore be calculated by looking up GT_{diff} in the empirical distribution of the GT^j_{diff}, regarding values of GT_{diff} much larger than one as critical for the hypothesis. Similarly, the significance probability for the test for nestedness is calculated by comparing the observed value of GT_{nest} with the empirical distribution of the GT^j_{nest}. Note that the use of GT_{diff} always leads to a one-sided test, whereas GT_{nest} can be used for both one-sided and two-sided tests. Again, according to Lehmann and Romano (2005) the rejection rate is controlled as described in Section 3.

Note that the purpose of the jackknife procedure described above is to make the bias that arises when estimating the frontier in each of the two groups of the same magnitude. This has no effect on the rejection rate under the hypothesis of equal distributions in the two groups – the permutation argument above works both with and without jackknifing. However, the use of jackknifing is of substantial importance for the rejection rate when the hypothesis of equal frontiers (or distributions) is false, i.e. for the power of the test. Without jackknifing, different group sizes will lead to unequal magnitudes of the biases for the two frontiers, and this may neutralize the real difference between them. This effect will be illustrated in the simulation study below.

4.1 Test under additional assumptions

The hypothesis that is tested can be made clearer by making an additional assumption:

Under a VRS assumption assume furthermore that the distributions of $(Z^1, V^1, W^1, \Theta^1)$ and $(Z^2, V^2, W^2, \Theta^2)$ are equal, where $(Z^g, V^g, W^g, \Theta^g)$ denotes the input length, input direction, output direction and efficiency, respectively, of an observation using technology $g = 1, 2$. The observation index has been suppressed.

Under this additional assumptions, the test proposed in the present section will exclusively be a test of the hypothesis of the two frontiers being equal: $\Psi^\delta_1 = \Psi^\delta_2$. This is simply due to the fact that Ψ^δ_g together with the distribution of $(Z^g, V^g, W^g, \Theta^g)$ determines the distribution F_g uniquely.

If instead constant returns to scale can be assumed, it suffices to assume that (V^1, W^1, Θ^1) and (V^2, W^2, Θ^2) have equal distributions in order for the test to be exclusively about the hypothesis of equal frontiers. This is due to (X^g, Y^g) having

the form,

$$X^g = \frac{Z^g \cdot H_g(V^g, W^g)}{\Theta^g} \qquad \text{and} \qquad Y^g = Z^g \cdot V^g,$$

for an appropriate function H_g that only depends on Ψ_g^δ (see also Section 3). Thus both X^g and Y^g are scaled by the same factor Z^g. Since $\hat{\theta}_{CRS}(x, y)$ is invariant under rescaling X^g and Y^g by the same factor, also both GT_{diff} and GT_{nest} do not depend on the length of the input vectors.

5 Monte Carlo procedure

5.1 Test for returns to scale

Simulation procedure 1

In the following simulations we let $p = 2$ and $q = 1$. The frontier is defined by a Cobb–Douglas function

$$f(x_1, x_2) = x_1^\alpha x_2^\alpha \,,$$

where $\alpha = \frac{\gamma}{2}$ and $0 < \gamma \leq 1$. A point $((x_1, x_2), y)$ is placed on the frontier, if

$$y = f(x_1, x_2) \,.$$

Note that the case $\gamma = 1$ corresponds to a CRS situation, since then with the frontier Ψ^δ defined by f, it always holds that $\|(x_1, x_2)\| = |y|$, when $((x_1, x_2), y) \in \Psi^\delta$. If, on the other hand, $\gamma < 1$ and $((x_1, x_2), y)$ satisfies $f(x_1, x_2) = y$, then for $a > 0$

$$f(a \cdot x_1, a \cdot x_2) = a^\gamma \cdot y \ ,$$

demonstrating that CRS cannot be assumed when $\gamma < 1$. Thus f is homogeneous of order 1 under the hypothesis of CRS while $\gamma < 1$ corresponds to the alternative hypothesis of VRS. When γ decreases from one to zero then the 'distance' to the CRS hypothesis becomes 'larger'.

We generate each of the points $(X_i, Y_i)_{i=1,\ldots,n}$ in the following way, where we suppress i in the notation:

1. Generate U_1 and U_2 independently from a Beta$(3, 3)$–distribution[1].

[1]We use the Beta–distribution to make directions close to the axes less likely than directions "in the middle"

2. Define the unit vector (W_1, W_2) by normalizing (U_1, U_2). That is

$$(W_1, W_2) = \frac{(U_1, U_2)}{\|(U_1, U_2)\|}\,.$$

3. Generate A from a $\Gamma(3,3)$–distribution, and calculate (X_1^δ, X_2^δ) as

$$(X_1^\delta, X_2^\delta) = A \cdot \frac{(W_1, W_2)}{f(W_1, W_2)^{1/\gamma}}\,.$$

4. Calculate Y as

$$Y = f(X_1^\delta, X_2^\delta)$$

5. Generate Θ from a Beta$(3, 1.5)$–distribution[2] and calculate (X_1, X_2) as

$$(X_1, X_2) = \frac{(X_1^\delta, X_2^\delta)}{\Theta}.$$

It should be noted that Y could also be calculated directly from A by using that $Y = A^\gamma$. This means that the simulation procedure satisfies the independence property stated in Section 3 requiring that the length of Y is independent of the joint distribution of V, W and Θ.

Results from simulation procedure 1

For different combinations of n, the number of observations, and γ, the degree of departure from the CRS hypothesis, we have simulated 1000 sets of observations. For each set we have used the permutation procedure proposed in Section 3 to calculate a significance probability. From this we have derived the proportion of rejected hypotheses on a 5% significance level across the 1000 simulations. The upper part of Table 1 shows the resulting rejection rates.

The first column in the upper part of Table 1 shows the rejection rates for $\gamma = 1$. This corresponds to the situation where the CRS hypothesis is actually true. As expected, the simulated rejection rate here is close to 5% for all the demonstrated values of n, meaning that the test has the correct size. In the next columns, the γ–parameter is decreased, which means that the departure from the CRS hypothesis increases. Here we see that rejection rates, i.e. the power of the test, increases rather fast, when γ decreases. It is also clear that the test procedure is more

[2]We use the Beta–distribution to obtain a distribution on $(0, 1)$. The parameter 1.5 is chosen to limit the probability of observations close to the frontier

	Permutation test (Section 3)										
	γ										
n	1.00	0.98	0.96	0.94	0.92	0.90	0.88	0.86	0.84	0.82	0.80
50	0.043	0.061	0.081	0.105	0.184	0.286	0.424	0.542	0.648	0.753	0.830
100	0.056	0.075	0.159	0.289	0.472	0.699	0.873	0.944	0.987	0.995	0.998
200	0.054	0.102	0.346	0.698	0.921	0.988	0.998	1.000	1.000	1.000	1.000
300	0.045	0.167	0.594	0.911	0.996	1.000	1.000	1.000	1.000	1.000	1.000

	Asymptotic test (Kneip *et al.* (2016))										
	γ										
n	1.00	0.98	0.96	0.94	0.92	0.90	0.88	0.86	0.84	0.82	0.80
50	0.204	0.186	0.243	0.244	0.309	0.354	0.396	0.432	0.540	0.537	0.585
100	0.150	0.190	0.225	0.281	0.367	0.433	0.503	0.630	0.673	0.749	0.817
200	0.143	0.177	0.273	0.376	0.475	0.618	0.734	0.829	0.880	0.910	0.940
300	0.128	0.163	0.262	0.432	0.586	0.744	0.841	0.916	0.930	0.941	0.945

Table 1: Proportions of rejected hypotheses when testing returns to scale on a significance level of $\alpha = 5\%$ with observations simulated according to procedure 1. Using varying values of n and γ.

powerful for larger sets of observations.

For comparison, we have included an application of the test procedure[3] proposed in Section 3.2 in Kneip *et al.* (2016). We have used the test for each of the 1000 sets of observations simulated according to procedure 1 for the same combinations of values of n and γ. As described in Section 1 this test relies on asymptotic arguments. The resulting rejection rates from this application of the test are seen in the lower part of Table 1. Here we see that under the CRS–hypothesis, i.e. where $\gamma = 1$, the rejection rate is much higher than the desired 5% for the chosen values of n – in agreement with the findings in Kneip *et al.* (2016), who describe their test for returns to scale as being conservative. After all, the rejection rate, when the CRS hypothesis is true, does decrease as n increases. This is consistent with the fact that the test only relies on asymptotic arguments as n goes to infinity.

The subsequent columns demonstrate that the test also has increasing power for an increasing departure from the CRS–hypothesis. However the rejection rates do not increase nearly as fast as the ones from the permutation test seen in the upper part, when γ decreases. This illustrates that the test proposed in Section 3 has a substantially higher power than the similar test suggested in Kneip *et al.* (2016).

Simulation procedure 2

To make the Monte Carlo evaluation of the test procedure for returns to scale as thorough and diverse as possible, we have also included a simulation study, where the sets of observations are simulated according to the procedure described in Section 5.1 in Kneip *et al.* (2016). Hereby it is possible to compare rejection rates from the permutation test with both rejection rates using the asymptotic test from Section 3.2 in Kneip *et al.* (2016) and using the bootstrapped confidence intervals suggested in their Section 4. In the simulations, the parameter δ supposedly measures the departure from the hypothesis of CRS: When $\delta = 0$, the CRS hypothesis is true, while values different from zero will make it false. However, increasing the value of δ is not identical to increasing the deviation from the hypothesis of CRS. The reason is that for large values of δ the simulated points at the upper right part of the frontiers (up to half of the points) have almost identical benchmarks near to or at the CRS frontier, and consequently, the corresponding inefficiencies assuming CRS and VRS do not differ substantially. This is also reflected in the simulation study performed in Kneip *et al.* (2016). In their Table 3 and Table 4 it is seen that the power generally decreases again for large δ and for a large number of observations, probably due to the simulation procedure.

[3]The number of subsample splits in the procedure is chosen to be $K = 1000$

δ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Rejection rate	0.053	0.075	0.529	0.985	0.997	0.996	1.000	0.997
δ	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.6
Rejection rate	0.999	0.997	0.995	0.998	0.994	0.997	0.996	0.891

Table 2: Proportions of rejected hypotheses when testing returns to scale on a significance level of $\alpha = 5\%$ with observations simulated according to procedure 2. Using $n = 100$ observations, $p = 2$ inputs and varying values of δ.

Results from simulation procedure 2

We have simulated according to procedure 2 with $n = 100$ observations in each dataset, $p = 2$ input variables, and varying values of δ. For each combination of the parameters we have simulated 1000 sets of observations, and the rejection rates for the permutation test are shown in Table 2.

Also in this simulation study we see that the test introduced in Section 3 is very powerful, and again it has the correct size under the CRS hypothesis. For comparison, corresponding rejection rates, using the asymptotic test of Kneip *et al.* (2016) can be found in their Table 3 while results from the bootstrapped method are found in their Table 4. It is noticeable that the rejection rates in Table 2 represent a test that is much more powerful and with more accurate size than the methods proposed by Kneip *et al.* (2016). However, the simulation procedure for generating the sets of observations suffers from the obvious drawback pointed out above: Increasing values of δ cannot be interpreted directly as the CRS–hypothesis becoming more incorrect. In fact in all cases, the rejection rate starts to decrease for very large values of δ.

5.2 Test for equality of frontiers

Simulation procedure 3

In this procedure for generating observations from two independent samples, let $p = 2$, $q = 1$ and assume CRS. Thus we can without loss of generality let $Y_i^g = 1$ for $g = 1, 2$ and all $i = 1, \ldots, n_g$ and focus on generating the points X_i^g. For each sample $g = 1, 2$, let the frontier be defined by the Cobb-Douglas function

$$f_g(x_1, x_2) = \beta_g x_1^{\alpha_g} x_2^{1-\alpha_g},$$

such that a point $((x_1, x_2), 1)$ is placed on the frontier, if

$$1 = f_g(x_1, x_2)\,.$$

In each group $g = 1, 2$ we generate $X_i^g \in \mathbb{R}^2$ as follows, where i is suppressed in the notation.

1. Generate U_1 and U_2 independently from a Beta$(3, 3)$–distribution.
2. Define the unit vector (W_1, W_2) by normalizing (U_1, U_2). That is

$$(W_1, W_2) = \frac{(U_1, U_2)}{\|(U_1, U_2)\|}\,.$$

3. Generate Θ from a Beta$(3, 1.5)$–distribution and calculate (X_1, X_2) as

$$(X_1, X_2) = \frac{(W_1, W_2)}{f_g(W_1, W_2)\Theta}\,.$$

Results from simulation procedure 3

In this section we investigate the performance of the test procedures proposed in Section 4. The two independent groups of observations are simulated according to procedure 3. In all simulation studies we have generated 500 datasets each consisting of two groups of observations with varying group sizes n_1 and n_2. When calculating the test statistics, the number of jackknife replications, m, is chosen to be 50, and for the test procedures we have used $N = 1000$ permutations. For each combination of the used parameters we have derived the proportion of rejected hypotheses on a 5% significance level across the 500 simulations.

First we have investigated the performance of the tests in a situation, where the two production possibility sets are nested. Thus we have chosen β_1 and β_2 unequal, while α_1 and α_2 are both chosen to be 0.5. We let $\beta_1 = 1$ and let β_2 vary between 1 and 1.1. Therefore, the production possibility set for group 1 is nested within the production possibility set for group 2. Furthermore, the sample sizes vary such that one of them is 50 and the other is either 100 or 200.

In the left part of Table 3 the rejection rates are seen for the cases, where sample 1 is smaller than sample 2, i.e. $n_1 = 50$ and $n_2 \in \{100, 200\}$, and in the right part group 2 is smaller than group 1. In the first row $\beta_2 = 1$, which means that the two frontiers are equal. Here the rejection rates are approximately 5% for both tests – as expected.

	$n_1 = 50$					$n_2 = 50$				
β_2	$n_2 = 100$		$n_2 = 200$		β_2	$n_1 = 100$		$n_1 = 200$	
	diff	*nest*	*diff*	*nest*		*diff*	*nest*	*diff*	*nest*
1.00	0.062	0.048	0.050	0.052	1.00	0.060	0.060	0.058	0.046
1.02	0.106	0.092	0.110	0.128	1.02	0.058	0.086	0.088	0.138
1.04	0.148	0.278	0.246	0.278	1.04	0.134	0.268	0.164	0.314
1.06	0.350	0.382	0.448	0.514	1.06	0.310	0.422	0.446	0.578
1.08	0.596	0.650	0.706	0.762	1.08	0.596	0.702	0.680	0.752
1.10	0.806	0.824	0.892	0.918	1.10	0.816	0.846	0.892	0.910

Table 3: Proportions of rejected hypotheses on a 5% significance level, when testing equality of frontiers using both the general difference (denoted *diff*) and the nested (denoted *nest*) test. Observations are simulated according to procedure 3 with $\beta_1 = 1$ and varying values of β_2 such that the two production possibility sets are in fact nested.

In the following rows β_2 is increased, which corresponds to the two frontiers becoming more and more different, such that the production possibility set for group 1 is nested within the production possibility set for group 2. Here the rejection rates are seen to increase substantially for both tests – no matter which of the groups is larger than the other. However, the rejection rate generally seems to be slightly higher for the nested test based on (5) than for the general difference test. This is not surprising, since the nested test is, in fact, designed to detect exactly the kind of difference between the two frontiers that have been used to produce the two samples.

To illustrate the importance of the use of the jackknife method in the test procedures, we have included a simulation study similar to the one in Table 3, but without jackknifing, both in the calculation of GT_{diff} and GT_{nest} and when finding GT^j_{diff} and GT^j_{nest} for all permutations, $j = 1, \ldots, N$. Here the findings are remarkably different from those of Table 3: When group 1 is smaller than group 2, the rejection rate increases faster than before. On the other hand, when group 1 is larger than group 2, both of the tests seem to be unable to reject the false hypothesis of no difference for almost all of the simulated datasets. This is due to the different magnitude of bias when estimating the two frontiers. In the table to the right, the production possibility set for group 1 is nested within the production possibility set for group 2, but at the same time, the frontier of group 1

β_2	$n_1 = 50$				β_2	$n_2 = 50$			
	$n_2 = 100$		$n_2 = 200$			$n_1 = 100$		$n_1 = 200$	
	diff	*nest*	*diff*	*nest*		*diff*	*nest*	*diff*	*nest*
1.00	0.047	0.041	0.051	0.051	1.00	0.052	0.049	0.059	0.055
1.02	0.139	0.155	0.144	0.150	1.02	0.029	0.020	0.015	0.013
1.04	0.311	0.349	0.320	0.329	1.04	0.009	0.004	0.005	0.005
1.06	0.489	0.509	0.599	0.603	1.06	0.009	0.002	0.000	0.000
1.08	0.746	0.766	0.832	0.831	1.08	0.025	0.002	0.001	0.000
1.10	0.900	0.904	0.971	0.973	1.10	0.069	0.014	0.000	0.000

Table 4: Simulations similar to Table 3 but without jackknifing.

is estimated with a smaller bias. These two effects counteract, such that the estimated frontiers are so close that the tests are unable to distinguish between them.

Another part of the evaluation of the two tests is to consider a situation, where the frontiers are different without one production possibility set being nested within the other. Here we have chosen $\beta_1 = \beta_2 = 1$, $\alpha_1 = 0.5$ and varying values of α_2. When α_1 and α_2 are different, the two frontiers will be different, and since they intersect the corresponding production possibility sets are not nested. The rejection rates from this simulation study are seen in Table 5. The middle row with $\alpha_2 = 0.5$ is identical to the first row in the left part of Table 3. Thus, the rejection rate is approximately 5 %.

In the other rows the frontiers are different, with a larger difference as α_2 becomes more different from 0.5. Here we see that only the general difference test, in the table denoted as *diff*, detects the difference with an increasing rejection rate, as the two frontiers become more and more different. On the other hand, the nested test is unable to distinguish between the two frontiers: This test keeps track of, which frontier corresponds to the best production possibilities for each observation, and when averaging over all observations these differences tend to cancel out when the frontiers intersect. Thus, in this situation, the two tests jointly detect the difference between the two frontiers with high power and furthermore correctly concludes that the production possibility sets are not nested.

	$n_1 = 50$			
α_2	$n_2 = 100$		$n_2 = 200$	
	diff	*nest*	*diff*	*nest*
0.1	0.988	0.022	1.000	0.016
0.2	0.948	0.020	0.992	0.030
0.3	0.646	0.024	0.782	0.036
0.4	0.184	0.034	0.262	0.054
0.5	0.062	0.048	0.050	0.052
0.6	0.160	0.038	0.230	0.054
0.7	0.654	0.020	0.778	0.040
0.8	0.952	0.018	0.998	0.012
0.9	0.952	0.014	0.996	0.028

Table 5: Proportions of rejected hypotheses on a 5% significance level, when testing equality of frontiers using both the general difference (denoted *diff*) and the nested (denoted *nest*) test. Observations are simulated according to procedure 3 with $\beta_1 = \beta_2 = 1$, $\alpha_1 = 0.5$ and varying values of α_2.

Simulation procedure 4

For completeness of the evaluation of our test for equality of frontiers, we have included a simulation study using the following procedure that is inspired by a procedure described in Daraio *et al.* (2018). However, the two procedures are still somewhat different: While the observations generated in our procedure are divided into two groups with a distinct frontier in each group, the datasets generated in Daraio *et al.* (2018) all have separate frontiers, determined parametrically by the value of a numerical covariate.

We let $p = q = 2$ and assume CRS. In each of the two groups $g = 1, 2$ we generate each of the observations (X_i, Y_i) for $i = 1, \ldots, n_g$ in the following way:

1. Generate U and V independently and each following the uniform distribution on the part of the unit circle, where both coordinates are positive.

2. Generate Z from a standard normal distribution.

3. Calculate $X, Y \in \mathbb{R}^2$ as

$$X = \left(1.01 - \frac{U}{\|U\|}\right) \cdot (1 + |Z|) \cdot (1 + \gamma_g) \qquad \text{and} \qquad Y = \frac{V}{\|V\|} + 0.01\,.$$

	$n_1 = 50$			
γ_2	$n_2 = 100$		$n_2 = 200$	
	F_{diff}	F_{nest}	F_{diff}	F_{nest}
0.0	0.040	0.062	0.050	0.050
0.2	0.634	0.902	0.756	0.938
0.4	0.998	1.000	1.000	1.000
0.6	1.000	1.000	1.000	1.000

Table 6: Simulation results for test for equality of frontiers, when data is simulated according to procedure 4. Proportions of rejected hypotheses for varying values of γ_2.

Choosing γ differently for the two groups corresponds to the two groups having different frontiers in such a way that one of the production possibility sets will be nested within the other.

Results from simulation procedure 4

The purpose of Table 6 is to investigate the performance of the two tests from Section 4, when the two groups of observations are simulated according to procedure 4. For each combination of parameters we have generated 500 datasets. For the test procedures, we have again used $m = 50$ jackknife replications and $N = 1000$ permutations. Furthermore, the proportions of rejected hypotheses on a 5% significance level are derived across the 500 simulations.

For different combinations of group sizes, the first row of Table 6 shows the rejection rates when the two frontiers are equal, i.e. when $\gamma_1 = \gamma_2 = 0$. As expected theoretically, all rejection rates in this row are close to 5%. In the next rows the parameter $\gamma_1 = 0$ is fixed, while γ_2 increases. This corresponds to the two frontiers becoming more different. Here the two tests with a very high power correctly identifies both the difference and the fact that the production possibility sets are nested.

6 Conclusion

In the paper we have proposed three exact tests based on permutations: One test for returns to scale and the others testing equality of frontiers, one for detecting general differences and the other specifically whether the production possibility

sets are nested. Simulation studies revealed that all three tests are of correct size and have high power. Specifically, both size and power are found to be better than those of the existing tests for similar hypotheses based on asymptotic theory. Therefore the proposed permutation tests are more appropriate for practical applications.

We propose that the tests are used as follows: For any dataset, in order to examine an underlying assumption of constant returns to scale in DEA, the returns to scale test can be used. After a decision on returns to scale has been made, if one wishes to compare the frontiers for two independent groups of observations, the test for general differences can be applied. If the hypothesis of no difference between the frontiers is rejected, the analysis can be supplemented with the test examining nestedness of the production possibility sets.

References

Asmild, M., Kronborg, D., and Rønn-Nielsen, A. (2018). Testing productivity change, frontier shift, and efficiency change. Institute of Food and Resource Economics, University of Copenhagen, Frederiksberg, Denmark. IFRO Working Paper 2018/07.

Banker, R. D. (1996). Hypothesis tests using data envelopment analysis. *Journal of Productivity Analysis*, **7**, 139-159.

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

Daraio, C., Simar, L. and Wilson P. W. (2018). Central limit theorems for conditional efficiency measures and tests of the 'separability' condition in non-parametric, two-stage models of production. *Econometrics Journal*, **21**, 170-191.

Farrell, M.J. (1957). The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society*, **120**, 253-281.

Fisher, R. A. (1935). **The Design of Experiments**, Oliver and Boyd, Edinburgh.

Kneip, A., Simar, L., and Wilson, P.W. (2015). When bias kills the variance: central limit theorems for DEA and FDH efficiency scores. *Econometric Theory*, **31**, 294-422.

Kneip, A., Simar, L., and Wilson, P.W. (2016). Testing Hyphoteses in Nonparametric Models of Production. *Journal of Business & Economic Statistics*, **34**, 435-456.

Lehmann, E.L. and Romano, J.P. (2005). **Testing Statistical Hypotheses**, Third edition, Springer texts in statistics.

Simar, L. and Wilson, P.W. (2002). Non-parametric tests of returns to scale, *European Journal of Operational Research*, **139**, 115-132.

Simar, L. and Wilson, P.W. (2007). Estimation and inference in two-stage, semiparametric models of production processes, *Journal of Econometrics*, **136**, 31-64.

Simar, L. and Wilson, P.W. (2015). Statistical approaches for Non-parametric frontier models: A guided tour. *International Statistical Review*, **83**, 77-110.