Notes on Technical efficiency Estimation with Multiple Inputs and Outputs

Mike G. Tsionas[*]

September 3, 2015

Abstract

Collier, Johnson and Ruggiero (2011) deal with the problem of estimating technical efficiency using regression analysis that allows multiple inputs and outputs. This revives an old problem in the analysis of production. In this note we provide an alternative maximum likelihood estimator that addresses the concerns. A Monte Carlo experiment shows that the technique works well in practice. A test for homotheticity, a critical assumption in Collier, Johnson and Ruggiero (2011) is constructed and its behavior is examined using Monte Carlo simulation and an empirical application to European banking.

[*]Lancaster University Management School, LA1 4YX, UK, tsionas@otenet.gr.

1 Introduction

To make things concrete, suppose we have multiple outputs in vector $\mathbf{y}_{it} \in \Re^M$, and there exists an aggregator function of the constant elasticity of substitution (CES) type:

$$g(\mathbf{y}_{it}) = (\alpha_1 y_{1,it}^\rho + \alpha_2 y_{2,it}^\rho)^{1/\rho}, \ i = 1, ..., N, \ t = 1, ..., T, \quad (1)$$

although other aggregator can serve just as well. The model is:

$$g(\mathbf{y}_{it}) = f(\mathbf{x}_{it}) + v_{it} - u_{it}, \quad (2)$$

where $f(\mathbf{x}_{it})$ is a functional form that shows how inputs $\mathbf{x}_{it} = [x_{1,it}, ..., x_{K,it}]' \in \Re^K$ contribute to the production of aggregate output, v_{it} is a two sided error term and $u_{it} \geq 0$ represents technical inefficiency. For example, with a Cobb-Douglas functional form and K inputs, equation (2) is:

$$\ln g(\mathbf{y}_{it}) = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_{k,it} + v_{it} - u_{it}, \ i = 1, ..., N, \ t = 1, ..., T. \quad (3)$$

There are certain important issues in technical efficiency analysis with multiple outputs. Since outputs are jointly produced given the inputs, equations (2) or (3) do *not* specify a joint system of equations for $\mathbf{y}_{it}$. In fact, with M outputs, $M - 1$ equations are missing, as noted by Fernandez et al. (2002). Therefore, the aggregator function approach is not enough for a complete analysis of the problem.

It is a, relatively, common practice in the literature to assume, say, [1] $g(\mathbf{y}_{it}) = y_{1,it}^{\alpha_1} y_{2,it}^{\alpha_2}$ and use OLS in the following equation[2]:

$$\alpha_1 \ln y_{1,it} + \alpha_2 \ln y_{2,it} = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_{k,it} + v_{it} - u_{it}, \quad (4)$$

or, alternatively, in obvious notation:

$$\ln y_{1,it} = \beta_0 + \gamma_0 \ln y_{2,it} + \sum_{k=1}^{K} \gamma_k \ln x_{k,it} + \tfrac{1}{\alpha_1}(v_{it} - u_{it}). \quad (5)$$

This equation shows plainly that we have an *endogenous* variable in the right-hand-side of the equation and, therefore, OLS *cannot* be used because of correlation with the error term, $v_{it} - u_{it}$. This is very often forgotten in applied

---

[1] To the author's knowledge the only reason why a Cobb-Douglas aggregator may be undesirable is because it does not satisfy the second order conditions of profit maximization. A Cobb-Douglas aggregator is, however, consistent with cost minimization. Of course it is also undesirable in the sense that it is not flexible enough like the CES or translog functional forms. Here we use it for simplicity in presentation of the main points.

[2] This, and similar practices, are nicely reviewed in Kumbhakar and Lovell (2000, pp. 93-95) and the cited references.

research. This point applies not only to regressions involving aggregator functions but also to distance functions. Input- or output-oriented distance functions are homogeneous of degree one with respect to inputs and outputs, respectively. Therefore, imposition of homogeneity results in a form similar to (5).

The proper method of estimation is limited information maximum likelihood (LIML) *if* the inputs are exogenous and prices are not available[3]. However, this is also a problematic assumption.

Under specific behavioral assumptions, inputs are endogenous. For example under profit maximization or cost minimization, inputs are endogenously selected and outputs are, respectively, endogenous or predetermined. In effect, OLS is not the proper method of estimation. Using LIML requires predetermined variables like prices (which, by assumption, are not available.) However, lagged values of $\mathbf{x}_{it}$ and $\mathbf{y}_{it}$ can be used as they are predetermined from the point of view of period t, provided v_{it} and u_{it} are not autocorrelated. An alternative estimation technique is the generalized method of moments (GMM) with a one-sided error term and explicit correlation allowed between $\mathbf{x}_{it}$ and the error components (Tran and Tsionas, 2014).

Collier, Johnson and Ruggiero (2011), CJR henceforth, deal with the problem of estimating technical efficiency using OLS regression analysis that allows multiple inputs and outputs. Specifically, technical efficiency can be estimated using regression models with multiple inputs and outputs *without* input price data. CJR propose to use DEA analysis in a first stage -without the input constraints since they assume separability from inputs- to derive an aggregate output measure, say S_{it}. The DEA problem solved in CJR is the following, for each observation "o":

$$\begin{aligned}
\max : \ & \Theta_o, \\
\sum_{j=1}^{NT} \Lambda_j y_{jm} \geq & \ \Theta_o y_{om}, \ m = 1, ..., M, \\
\sum_{j=1}^{NT} & \Lambda_j = 1, \\
\Lambda_j \geq 0, \ & j = 1, ..., NT.
\end{aligned} \tag{6}$$

Aggregate output is defined as:

$$S_o = \frac{1}{\Theta_o}, \ o = 1, ..., NT. \tag{7}$$

Then, instead of (2) they propose regression analysis in the following model:

$$\log S_{it} = f(\mathbf{x}_{it}) + v_{it} - u_{it}, \ i = 1, ..., N, \ t = 1, ..., T. \tag{8}$$

As $\mathbf{x}_{it}$ is often used in logs in (8) a Cobb-Douglas function has the following form:

$$\log S_{it} = \beta_0 + \mathbf{x}'_{it}\boldsymbol{\beta} + v_{it} - u_{it} \ i = 1, ..., N, \ t = 1, ..., T. \tag{9}$$

---

[3]If prices are available the system can be completed using the first order conditions from cost minimization or profit maximization. An alternative has been introduced by Atkinson and Tsionas (2015) where the first order conditions are used, and unobserved prices are treated as latent variables in the context of a Bayesian hierarchical model.

This methodology has been applied by Collier, Mamula, and Ruggiero (2014). Under homotheticity, this bypasses the problem of having an endogenous variable in the right-hand-side, and falls in between DEA and stochastic frontier analysis: It allows measurement error in the aggregate output, and statistical assumptions about the error terms v_{it} and u_{it}. However, it does not addresses concerns about the potential endogeneity of inputs, in which case $\mathbf{x}_{it}$ and v_{it} are correlated.

The main problem, however, is not to seek an output aggregator, since homotheticity cannot be taken for granted. In Fernandez, Koop, and (2002) the problem in (3) is recognized for *what it truly is*, viz. several inputs $\mathbf{x}_{it}$ are used to *jointly produce* outputs $\mathbf{y}_{it}$. The problem is not which aggregator function must be used (a CES would be just fine for most purposes) or how to aggregate the outputs, but how to account for the *endogenous character* of $\mathbf{y}_{it}$ in this context. From the econometric point of view, the problem is that with M outputs there are M endogenous variables but only one equation, viz. (2) or (3). Therefore, there are $M-1$ missing equations to complete the system.

2   An alternative estimation technique

The question is how to deal with the problem when *both* outputs and inputs are endogenous and prices are *not* observed. To complete the model provided by (3) we consider the reduced form:

$$\begin{bmatrix} \tilde{\mathbf{y}}_{it} \\ \mathbf{x}_{it} \end{bmatrix} = \mathbf{\Pi}\mathbf{z}_{it} + \boldsymbol{\varepsilon}_{it}, \tag{10}$$

where $\tilde{\mathbf{y}}_{it} = [y_{2,it}, \ldots, y_{m,it}]'$, $\mathbf{z}_{it}$ is a vector of predetermined variables, $\boldsymbol{\varepsilon}_{it}$ is a vector of error terms and $\mathbf{\Pi}$ is a matrix of unknown parameters. Under the assumption that $u_{it} = u_i$, $\forall t = 1, ..., T$ are fixed parameters and:

$$[v_{it}, \boldsymbol{\varepsilon}'_{it}]' \sim \mathcal{N}_M(\mathbf{0}, \mathbf{\Sigma}), \tag{11}$$

the system of (3) and (10) can be estimated using limited information maximum likelihood (LIML) along the lines suggested by Pagan (1979). The variables in $\mathbf{z}_{it}$ include *only* firm and time dummies so that it is not necessary to think about the possibility of other instruments, which may not be available at all in practice. For example, the use of lagged values of $\tilde{\mathbf{y}}_{it}$ and $\mathbf{x}_{it}$ is often problematic if they are only weakly correlated with the endogenous variables.

As in CJR, we assume time-invariant technical inefficiency, that is $u_{it} = u_i, \forall t = 1, ..., T$. As the inclusion of firm-specific effects in (10) prevents the identification of u_is we use the *nonlinear transformation*:

$$u_i = \exp(-\varphi_i^2), \ \forall i = 1, ..., n, \tag{12}$$

where the φ_is are unrestricted. With this transformation, the u_is are *always positive and less than one*, thus making it unnecessary to apply a corrected

ordinary least squares (COLS) transformation, viz. $\hat{u}_i = u_i - \max_{i=1,\dots,n} u_i$, to obtain technical inefficiencies. To the author's knowledge, this transformation has not been used before although it has considerable merit, at the cost of requiring nonlinear estimation techniques.

3 Monte Carlo evidence

To see specifically how the problems mentioned in the previous section can be addressed, we consider a model using (1) with $\rho = \frac{1}{2}$. The first output is are generated from a standard lognormal distribution. Next, we generate $v_{it} \sim \mathcal{N}(0, \sigma_v^2)$. The three inputs (in log terms) are generated as follows:

$$
\begin{aligned}
x_{1,it} &= \alpha v_{it} + \mathbf{D}_{it}\boldsymbol{\gamma}_1 + \xi_{1,it}, \\
x_{2,it} &= x_{1,it} + \mathbf{D}_{it}\boldsymbol{\gamma}_2 + \xi_{2,it}, \\
x_{3,it} &= x_{1,it} + \mathbf{D}_{it}\boldsymbol{\gamma}_3 + \xi_{3,it},
\end{aligned}
\tag{13}
$$

where $\xi_{j,it} \sim N(0,1)$. Therefore, inputs are mutually correlated as well as correlated with v_{it}. The correlation coefficient between $x_{1,it}$ and v_{it} is $\varrho = \frac{\alpha\sigma_v^2+1}{\sigma_v\sqrt{\alpha^2\sigma_v^2+1}}$. Given σ_v we can vary α so that we obtain different values of ϱ. Moreover, $\mathbf{D}_{it}$ is a vector of firm and time dummies and $\boldsymbol{\gamma}_j$ are respective coefficients, which are generated from a uniform distribution in $[1,1]$. We have $N + T - 1$ coefficients in each input equation.

The model is the same as in (3) with constant term -1 and slope coefficients equal to $\frac{1}{3}$. Finally, the second output is generated from (2) where $v_{it} \sim \mathcal{N}(0, \sigma_v^2)$ and independently $u_{it} \sim \mathcal{N}_+(0, \sigma_u^2)$. Denote $\sigma = \sqrt{\sigma_v^2 + \sigma_u^2}$ and $\lambda = \frac{\sigma_u}{\sigma_v}$. For practical reasons, we can set $\sigma = 0.3$ and $\lambda = 1$ which is a typical case in empirical studies and examine the rank correlation between true and estimated inefficiencies for various values of ϱ. The number of firms is N, the number of time periods is T and we assume $u_{it} = u_i, \forall t = 1, \dots, T$.

LIML is implemented using a standard conjugate-gradients algorithm without analytic derivatives and we consider 10,000 alternative data sets. Our evidence is summarized in Table 1. Clearly, the behavior of OLS is disappointing, as expected, and gets worse in larger sample sizes as ϱ increases, contrary to LIML which performs well across the board. Another problem is how the two-stage method of CJR performs when homotheticity is violated. For simplicity in presentation, we proceed with the following specification:

$$
g(\mathbf{y}_{it}) = (1-\psi)(\alpha_1 y_{1,it}^\rho + \alpha_2 y_{2,it}^\rho)^{1/\rho} + \psi(\alpha_1 x_{1,it}^\tau + \alpha_2 x_{2,it}^\tau + \alpha_3 x_{3,it}^\tau)^{1/\tau}. \tag{14}
$$

We set $\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{3}$ and $\tau = \frac{1}{2}$. The parameter $\psi \in (0,1)$ measures the degree of departure from homotheticity (which corresponds to $\psi = 0$). As this does not affect the behavior of LIML, we focus on how different values of ψ affect the CJR estimator. A fair comparison with CJR must be based on taking account of the endogeneity between $\mathbf{x}_{it}$ and v_{it}. Therefore, the regressors

are replaced by the fitted values of regressions in (13) as is commonly done in two-stage least squares. [4]

From the evidence in Table 1 it turns out that even minor departures from homotheticity have an important impact on efficiency estimates obtained through CJR. As the rank correlations tend to stabilize by increasing N and T it becomes clear that there is (a significant) asymptotic bias when homotheticity is violated.

Table 1. Rank correlation coefficients

	OLS	LIML	CJR			
$\varrho = 0.3$			$\psi = 0.01$	$\psi = 0.3$	$\psi = 0.5$	$\psi = 0.9$
$n = 50, T = 5$	0.117	0.515	0.313	0.117	-0.224	-0.487
$n = 50, T = 10$	0.121	0.589	0.308	0.032	-0.335	-0.515
$n = 200, T = 5$	0.125	0.771	0.298	-0.115	-0.361	-0.528
$n = 200, T = 10$	0.129	0.814	0.214	-0.117	-0.369	-0.557
$n = 500, T = 5$	0.131	0.921	0.213	-0.116	-0.370	-0.562
$n = 500, T = 10$	0.135	0.975	0.214	-0.117	-0.370	-0.562
$\varrho = 0.5$						
$n = 50, T = 5$	0.022	0.622	0.119	-0.026	-0.381	-0.555
$n = 50, T = 10$	0.031	0.791	0.120	-0.031	-0.394	-0.585
$n = 200, T = 5$	0.017	0.873	0.022	-0.035	-0.317	-0.592
$n = 200, T = 10$	0.019	0.913	0.017	-0.041	-0.321	-0.590
$n = 500, T = 5$	0.015	0.925	0.016	-0.041	-0.320	-0.606
$n = 500, T = 10$	0.017	0.981	0.017	-0.040	-0.320	-0.606
$\varrho = 0.8$						
$n = 50, T = 5$	0.012	0.635	0.051	-0.037	-0.357	-0.552
$n = 50, T = 10$	0.015	0.802	0.047	-0.043	-0.371	-0.589
$n = 200, T = 5$	-0.007	0.899	0.012	-0.046	-0.394	-0.617
$n = 200, T = 10$	-0.009	0.978	0.007	-0.047	-0.317	-0.622
$n = 500, T = 5$	-0.004	0.992	0.007	-0.048	-0.322	-0.620
$n = 500, T = 10$	-0.005	0.998	0.007	-0.048	-0.322	-0.620

The important question in relation to the CJR technique is whether we can construct a test that has power against violations of homotheticity. The test can be used to decide whether homotheticity is satisfied so that CJR can be applied in practice. Suppose S_{it} and $\tilde{S}_{it}$ are the output indicators obtained using, respectively, CJR (see program (6)) and an "output indicator" that one obtains by including the input constraints in their DEA problem. This problem is the following:

---

[4]All computations are performed in Fortran 77 using g77. The envelopment problems in (6) and (15) are solved using DLPRS or DSLPRS. The conjugate gradients technique is implemented using DUMCGG in the IMSL library, version 5.

$$\max : \tilde{\Theta}_o,$$
$$\sum_{j=1}^{NT} \Lambda_j y_{jm} \geq \tilde{\Theta}_o y_{om}, \ m = 1, ..., M,$$
$$\sum_{j=1}^{NT} \Lambda_j x_{jk} \leq x_{ok}, \ k = 1, ..., K, \tag{15}$$
$$\sum_{j=1}^{NT} \Lambda_j = 1,$$
$$\Lambda_j \geq 0, \ j = 1, ..., NT.$$

Aggregate output is constructed similarly to (7). Consider the test:

$$D = (NT)^{-1} \sum_{i,t} \frac{|S_{it} - \tilde{S}_{it}|}{\frac{1}{2}(S_{it} + \tilde{S}_{it})}. \tag{16}$$

The test measures relative deviations between the two aggregate output indicators relative to their simple average. The average is used so that we do not have to choose which case is a benchmark, viz. (6) or (15). We use the same data generating process as before. We consider only the typical case $\varrho = 0.3$, which is more favorable to regressions using (6). We report size and power results in Table 2. The test is, approximately, correctly sized for $NT = 50$, approximately, and it is almost 5% for $NT = 100$. The power of the test in (16) is reasonable even when the sample size is small ($NT = 20$ for example) and deviations for homotheticity are minor ($\psi = 0.01$). These results are encouraging and show that the D-statistic can be used in the CJR approach to detect whether violations of homotheticity in the data are statistically significant.

Table 2. Size and power of D-statistic

		Power		
$\varrho = 0.3$	$\psi = 0$, Size	$\psi = 0.01$	$\psi = 0.1$	$\psi = 0.3$
$nT = 10$	0.0717	0.451	0.781	0.892
$nT = 20$	0.0623	0.711	0.915	0.938
$nT = 50$	0.0517	0.825	0.947	0.951
$nT = 100$	0.0504	0.877	0.953	0.953

Notes: Size is measured using the proportion of cases where the hypothesis $\psi = 0$ is rejected when, in fact, it holds true. Power is measured using the proportion of cases where the hypothesis $\psi = 0$ is rejected when, in fact, $\psi = 0.01, 0.1, 0.3$.

4 Empirical application

In this section we present an empirical application to European banking along with some important clarifications for the implementation of LIML. Our data set includes prices of labor, capital, deposits. The two outputs funds are loans and other earning assets. We have also obtained the quantity values for loans, deposits and other earning assets as well as other funds. Labor quantity is obtained as personnel expenses divided by the price of labor. Capital is defined

as total fixed assets divided by the price of capital. Quantities for loans, deposits and other earning assets are obtained from total values divided by respective prices. Total assets and equity (in logs) are included as environmental variables in all models. The data is for 27 European countries over the period 2006-2011 and were retrieved from Datastream. The total number of observations is 27,187. To minimize computational costs in the implementation of CJR we consider only large banks, viz. those who have total assets in excess of the ninth decile.

First, we present our translog specification under the assumption of output homotheticity. This is given by the following:

$$
\begin{aligned}
\sum_{m=1}^{M} \alpha_m y_{m,it} + \tfrac{1}{2} \sum_{m=1}^{M} \sum_{m' \geq m}^{M} \alpha_{mm'} y_{m,it} y_{m',it} = \\
\beta_0 + \sum_{k=1}^{K} \beta_k x_{k,it} + \tfrac{1}{2} \sum_{k=1}^{K} \sum_{k' \geq k}^{K} \beta_{kk'} x_{k,it} x_{k',it} + v_{it} - u_{it}.
\end{aligned}
\tag{17}
$$

Without the assumption of output homotheticity, the specification is as follows:

$$
\begin{aligned}
\sum_{m=1}^{M} \alpha_m y_{m,it} + \tfrac{1}{2} \sum_{m=1}^{M} \sum_{m' \geq m}^{M} \alpha_{mm'} y_{m,it} y_{m',it} = \\
\beta_0 + \sum_{k=1}^{K} \beta_k x_{k,it} + \tfrac{1}{2} \sum_{k=1}^{K} \sum_{k' \geq k}^{K} \beta_{kk'} x_{k,it} x_{k',it} + \\
\sum_{k=1}^{K} \sum_{m=1}^{M} \gamma_{km} x_{k,it} y_{m,it} + v_{it} - u_{it}.
\end{aligned}
\tag{18}
$$

The constant term of the output "aggregator" in the left-hand-side cannot be identified separately and it is merged with an overall intercept, β_0. Under the KM restrictions $\gamma_{km} = 0, k = 1, ..., K, m = 1, ..., M$, (18) reduces to (17). LIML is implemented to estimate the more general form in (18). The instruments $\mathbf{z}_{it}$ in (10) now include time-specific and bank-specific dummy variables, log of total assets and log of equity. It is important to mention that LIML does not depend on assigning *one* output as *the* "left hand side" variable in (18), see Schmidt (1976, p. 171). Let us write (18) in compact form as follows:

$$
\boldsymbol{\alpha}' \mathbf{y}_{it} + \tfrac{1}{2} \mathbf{y}'_{it} \boldsymbol{A} \mathbf{y}_{it} = \boldsymbol{\beta}' \mathbf{x}_{it} + \tfrac{1}{2} \mathbf{x}'_{it} \boldsymbol{B} \mathbf{x}_{it} + \mathbf{x}'_{it} \boldsymbol{\Gamma} \mathbf{y}_{it} + v_{it} - u_{it}.
\tag{19}
$$

The definition of the various vectors and matrices is as follows:

$$
\begin{aligned}
\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_m]', \ \boldsymbol{\beta} = [\beta_1, ..., \beta_K]', \\
\boldsymbol{A} = [\alpha_{mm'}, \ m, m' = 1, ..., M], \ \boldsymbol{B} = [\beta_{kk'}, \ k, k' = 1, ..., K], \\
\boldsymbol{\Gamma} = [\gamma_{km}, \ k = 1, ..., K, \ m = 1, ..., M].
\end{aligned}
$$

The matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric. Suppose $\epsilon_{it} = v_{it} - u_{it}$. Taking derivatives, $\frac{\partial \epsilon_{it}}{\partial \mathbf{y}_{it}} = \boldsymbol{\alpha} + \boldsymbol{A} \mathbf{y}_{it} + \boldsymbol{\Gamma} \mathbf{x}_{it}$, an $M \times 1$ vector. In the system consisting of (19) and (10) the Jacobian of transformation from $[\epsilon_{it}, \boldsymbol{\varepsilon}'_{it}]'$ to $\mathbf{y}_{it}$ is:

$$
J_{it} = | \prod_{j=1}^{M} [\boldsymbol{\alpha} + \boldsymbol{A} \mathbf{y}_{it} + \boldsymbol{\Gamma} \mathbf{x}_{it}]_j |, \ i = 1, ..., N, \ t = 1, ..., T,
\tag{20}
$$

where $[\mathbf{a}]_j$ denotes the jth element of vector $\mathbf{a}$. Under normality of v_{it} the likelihood function can be formulated easily taking also into account the Jacobian term in (20).

From LIML estimation of (18) the F-statistic to test homothetic production, viz. $\gamma_{km} = 0, k = 1, ..., K, m = 1, ..., M$ has a p-value very close to zero (0.00072) suggesting that, statistically, homotheticity is rejected through this parametric form. The D-statistic yields an empirical p-value of 0.00014. The empirical p-value is obtained through a bootstrap procedure (using 200 replications to minimize computational costs) where the data are re-sampled in blocks corresponding to the same bank over time. We follow the bootstrap implementation in Simar and Wilson (2010, 2011). This implementation is rather computationally intensive. Since homotheticity is rejected, the CJR techniques are not reliable as it has been stated in the previous section. Nevertheless, a comparison is interesting. Another problem is that the assumption of time-invariant inefficiency, in the second stage of CJR, is not likely to hold. Therefore, both CJR and LIML are implemented using the assumption that technical inefficiency is constant for the sub-periods 2006-2007, 2008-2009 and 2010-2011. The p-values of the F-statistic and the D-statistic are practically zero when the models are re-estimated in the three sub-periods. For a fair comparison between CJR and LIML, a translog functional form is estimated in the second stage of CJR where trend, and logs of total assets and equity are included also as regressors.

The empirical results for technical efficiency scores are provided in Table 3. Apart from CJR and LIML we also consider another estimator, CJR/2SLS (2SLS stands for two-stage least squares). This is similar to CJR but instead of OLS, we use 2SLS in the second stage of CJR. The instruments are fitted values from OLS regressions of inputs on the instruments $\mathbf{z}_{it}$ in (10) used in LIML. This may provide a more fair comparison between CJR and LIML.

Table 3. Technical efficiency results

	2006-2007	2008-2009	2010-2011
LIML	0.919	0.734	0.884
	(0.177)	(0.144)	(0.211)
CJR	0.872	0.810	0.913
	(0.202)	(0.187)	(0.109)
CJR/2SLS	0.880	0.832	0.910
	(0.165)	(0.157)	(0.122)
p-value of D-statistic	0.0027	0.00015	0.00011
Rank Correlation, LIML-CJR	0.202	0.0144	0.181
Rank Correlation, LIML-CJR/ 2SLS	0.245	0.0261	0.203

Notes: Sample standard deviations of efficiency scores appear in parentheses. The p-values of D-statistics are obtained from a block-bootstrap applied to the CJR linear programs. "RAnk correlation" refers to the rank correlation coefficients between efficiency scores from LIML and CJR. CJR-2SLS (2SLS stands for two-stage least squares) is similar to CJR but instead of OLS, we use 2SLS in the second stage of CJR. The instruments are fitted values from OLS regressions of inputs on the instruments $\mathbf{z}_{it}$ in (10) used in LIML.

Technical efficiency, according to LIML, is 91.9% for 2006-2007, 73.4% for 2008-2009 and 88.4% for 2010-2011. CJR estimates are 87.2%, 81% and 91.3% respectively. The (rank) correlation coefficients between efficiency scores are very low at 0.202, 0.014 and 0.181 for the three sub-periods. The low rank correlations and the differences in estimated scores (as well as their sample standard deviations) are, probably, attributed to the violation of homotheticity in this data set. Notably, efficiency scores from LIML drop during 2008-2009, when the sub-prime crisis occurred, to recover in the sub-period 2010-2011. CJR scores also drop but not as much as those derived from LIML. The use of 2SLS in CJR yields similar results with respect to rank correlations with LIML although the first two moments of CJR and CJR/2SLS seem to be different.

5   Conclusions

This study addresses concerns raised by CJR. The main concern is how to use econometric tools in the presence of multiple inputs and multiple outputs when prices are not available. CJR assume homotheticity and propose a two-stage approach to the problem. In the first stage, a DEA problem is formulated that drops the input constraints, and a composite output aggregator is obtained. In the second stage, OLS is used to regress the composite output aggregator on inputs and obtain efficiency measures. This also allows for noise in the second stage. We show that a more promising approach to the same problem is to couple an explicit multiple-output specification with a reduced form for the endogenous inputs and outputs. LIML is used to estimate the system and obtain estimates of technical inefficiency. Moreover, a test is proposed to examine whether deviations from the critical homotheticity assumption are statistically significant. The approach is shown to perform well in Monte Carlo experiments and an empirical application using European banking data.

References

Atkinson, S., E.G. Tsionas (2015). Directional distance functions: Optimal endogenous directions. Journal of Econometrics, in press, doi:10.1016/j.jeconom.2015.06.006.

Collier, T., A.L. Johnson, J. Ruggiero (2011). Technical efficiency estimation with multiple inputs and multiple outputs using regression analysis, European Journal of Operational Research 208, 153-160.

Collier, T., A. Mamula, J. Ruggiero (2014). Estimation of multi-output production functions in commercial fisheries, Omega 42, 157-165.

Fernandez, C., G. Koop, M.F.J. Steel (2002). Multiple-Output production with undesirable outputs. An application to nitrogen surplus in agriculture, Journal of the American Statistical Association 97, 432-442.

Kumbhakar, S.C., C.A.K. Lovell (2000). Stochastic frontier analysis. Cambridge University Press.

Pagan, A. (1979). Some consequences of viewing LIML as an iterated Aitken estimator, Economics Letters 3, 369-372.

Schmidt, P. (1976). Econometrics. Marcel Dekker, New York.

Simar, L., P. Wilson (2000). Statistical inference in nonparametric frontier models: The State of the Art, Journal of Productivity Analysis 13 (1), 49-78.

Simar, L., P. Wilson (2011). Two stage DEA: *Caveat Emptor*. Journal of Productivity Analysis 36, 205-218.

Tran, K.C., E.G. Tsionas (2013). GMM estimation of stochastic frontier model with endogenous regressors, Economics Letters 118, 233-236.