Further Results on Estimating Inefficiency Effects in Stochastic Frontier Models

Mike G. Tsionas* Emmanuel Mamatzakis†

October 29, 2018

Abstract

Paul and Shankar [Satya Paul, Sriram Shankar, On Estimating Efficiency Effects in a Stochastic Frontier Model, European Journal of Operational Research (2018)] proposed an inefficiency effects stochastic frontier model which is easy to implement and avoids some of the difficulties of existing models. Unfortunately, the model has the restrictive feature that the ratio of the inefficiency effects of any two environmental variables remains fixed independently of the values of the variables. We modify the model so that this restriction can be avoided. Moreover, we provide a substantive extension of the model under quite general endogeneity assumptions. In turn, the model can be estimated using the Generalized Method of Moments technique allowing identification of efficiency estimates and inefficiency effects.

Keywords: Decision Processes; Stochastic frontier; Technical efficiency; Artificial Neural Networks; Generalized Method of Moments.

Acknowledgments: The author wishes to thank three anonymous reviewers for providing constructive comments on an earlier version.

*Lancaster University Management School, LA1 4YX, U.K., m.tsionas@lancaster.ac.uk Department
†of Accounting and Finance, Sussex Business School, University of Sussex, Brighton, U.K.
e.mamatzakis@sussex.ac.uk

1 Introduction

Paul and Shankar (2018) proposed the following model to incorporate inefficiency effects:

$$Y_{it} = \exp(x'_{it}\beta + v_{it})\Phi(z'_{it}\gamma), i = 1, ..., N, t = 1, ..., T, \tag{1}$$

where $x_{it} \in \Re^K$ is the input vector, Y_{it} is output, $z_{it} \in \Re^M$ is the vector of observations on inefficiency effects, $\beta \in \Re^K, \gamma \in \Re^M$ are parameters, v_{it} is a two-sided error and Φ is any cumulative distribution function (cdf), for example the normal. In terms of logs we have:

$$y_{it} = x'_{it}\beta + \log \Phi(z'_{it}\gamma) + v_{it}, i = 1, ..., N, t = 1, ..., T, \tag{2}$$

where $y_{it} = \log Y_{it}$ and the model can be estimated using nonlinear Least Squares (LS). The model has been proposed in Deprins and Simar (1989) although they used $\exp(z'_{it}\gamma)$ instead of the log of a cdf. See also Kumbhakar and Lovell (2000, p. 265).

The efficiency effects are simply:

$$\frac{\partial E(y_{it}|x_{it}, z_{it})}{\partial z_{it}} = \frac{\varphi(z'_{it}\gamma)}{\Phi(z'_{it}\gamma)}\gamma, \tag{3}$$

and they have always the same sign as γ. The ratio of any two effects is constant:

$$\frac{\partial E(y_{it}|x_{it}, z_{it})/\partial z_{it,m}}{\partial E(y_{it}|x_{it}, z_{it})/\partial z_{it,m'}} = \frac{\gamma_m}{\gamma_{m'}}, m \neq m'. \tag{4}$$

This feature is clearly a drawback of the model. Another drawback is that the shape of any efficiency effect depends only on the ratio $\frac{\varphi(z'_{it}\gamma)}{\Phi(z'_{it}\gamma)}$ which (as a function of any variable conditional on the others) is a function decreasing at increasing rate. This suggests that despite its simplicity, the proposed formulation is not sufficiently flexible for empirical work.

A more flexible formulation is:

$$y_{it} = x'_{it}\beta + \sum_{g=1}^{G} \delta_g \log \Phi(z'_{it}\gamma_g) + v_{it}, i = 1, ..., N, t = 1, ..., T, \tag{5}$$

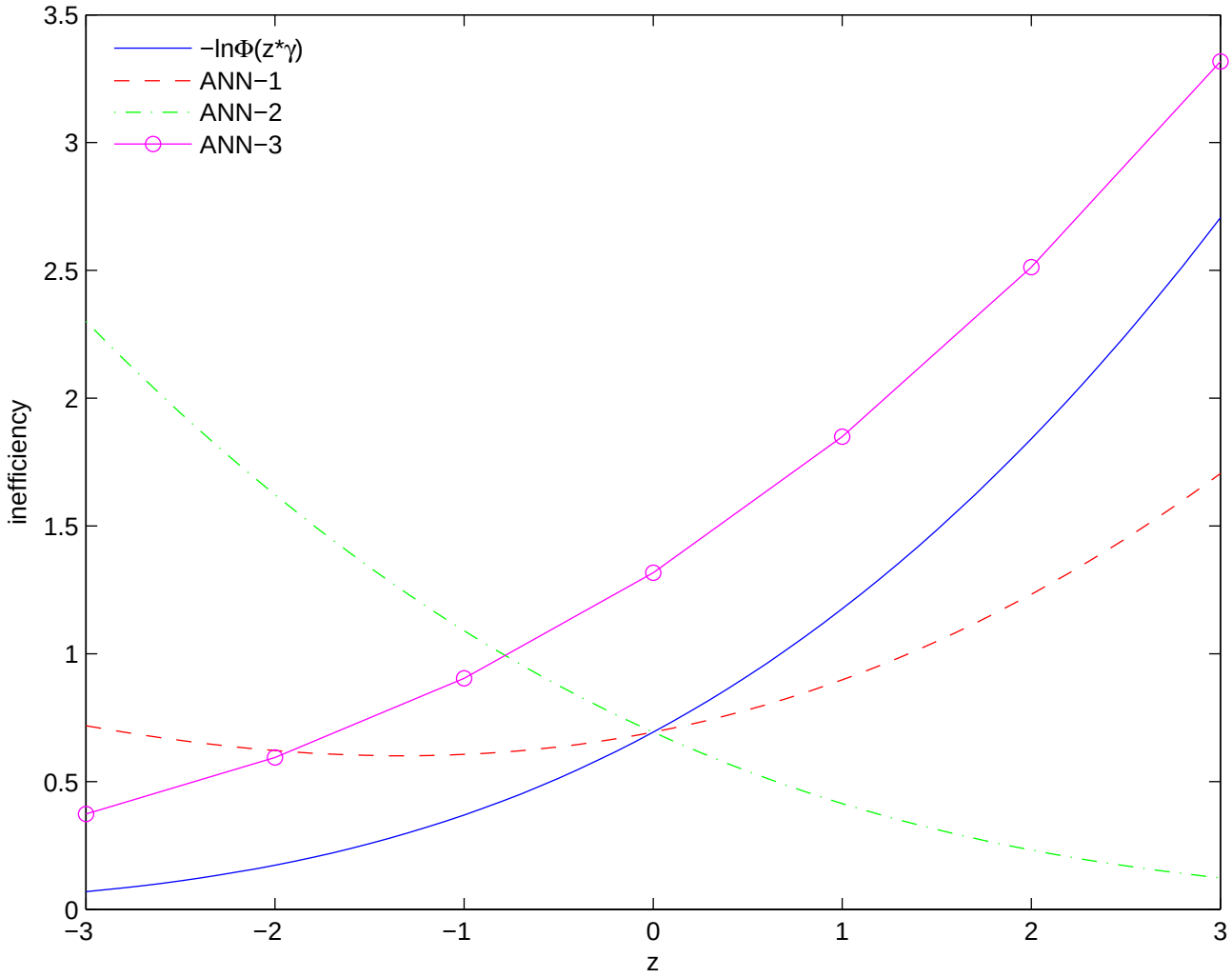
where $\delta_g \geq 0$ $(g = 1, ..., G \geq 2)$. This formulation is essentially an artificial neural network (ANN) approximation to the unknown inefficiency effects function. The universal approximation properties of neural networks are too familiar to present here in detail (see Hornik et al., 1989). The inefficiency effects are now

$$me_{it} = \sum_{g=1}^{G} \delta_g \frac{\varphi(z'_{it}\gamma_g)}{\Phi(z'_{it}\gamma_g)}\gamma_g, \tag{6}$$

and the ratio of any two inefficiency effects varies with z_{it}. It varies, however, in a *flexible* way as it inherits the universal approximation properties of neural networks.

In Figure 1 we present (2) three parametrizations of (5). Variable z runs from -3 to 3 with step 0.1. In ANN-1 we have $\gamma_1 = -0.5$, $\gamma_2 = 0.3$, $\delta_1 = 0.3$ and $\delta_2 = 0.6$. In ANN-2 we have $\gamma_1 = 0.5$, $\gamma_2 = 0.3$, $\delta_1 = 0.3$ and $\delta_2 = 0.6$. Finally, the parametrization in ANN-3 is $\gamma_1 = -0.5$, $\gamma_2 = -0.3$, $\delta_1 = 0.1$ and $\delta_2 = 1.8$.

Figure 1: Comparison of (2) and alternative parametrizations of (5)



As we can see from Figure 1, the solid blue line represents how inefficiency varies with the underlying variable z. Inefficiency in this case always increases with z, a fact that may or may not be the case in practice. However, there are parametrizations of the neural network that yield inefficiency decreasing with z (dotted green line), inefficiency that increases with z (line with circles) or even inefficiency that is non-monotonic in terms of z (dotted red line). In fact, due to the approximation properties of neural networks, arbitrary patterns of the dependence between inefficiency and z can be accommodated as G increases. In fact, neural networks are *universal* approximations to arbitrary functions, see Hornik et al. (1980).

Given the formulation, the model can accommodate a wide range of assumptions about endogeneity. Although nonlinear least squares is consistent under the restrictive assumption that x_{it} and z_{it} are orthogonal to v_{it}, more general assumptions can be used *without* affecting the identification of inefficiency. In stochastic frontier models this is not the case and explicit distributional assumptions have to be made. In this paper, we propose the use of Generalized Method of Moments (GMM) which can deal with endogeneity concerns. Instead of estimating (5) using nonlinear least squares, we can use certain instrumental variables to deal with the endogeneity problem. Toi explain the technique, let us write (5) in the following form:

$$y_{it} = f(\mathbf{q}_{it}; \theta) + v_{it}, i = 1, ..., n, t = 1, ..., T, \tag{7}$$

where $\mathbf{q}_{it}$ is a vector that contains both x_{it} and z_{it}, and θ contains the parameters β and γ. When q_{it} and v_{it} are correlated but there is a vector of instruments, say $\mathbf{w}_{it} \in \Re^d$ such that they are correlated with q_{it} but not the error term, then we have: $E\left[y_{it} - f(\mathbf{q}_{it}; \theta)\mathbf{w}_{it} | \mathbf{w}_{it}\right] = \mathbf{0}$. GMM relies on the "principle of analogy" which implements these conditions in the sample by invoking a law of large numbers:

$$(NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - f(\mathbf{q}_{it}; \theta)\mathbf{w}_{it}\right) \mathbf{w}_{it} = \mathbf{0}. \tag{8}$$

This is a system of d nonlinear equations in parameters θ and provided $d \geq dim(\theta)$ the equations in (8) identify the parameters. For example, if we have a linear model[1]:

$$y_t = x_t'\beta + v_t, t = 1, ..., T, \tag{9}$$

where $\beta \in \Re^k$ the explanatory variables x_t and the error are correlated, but there is a vector of instruments $\mathbf{w}_t$ correlated with x_t but not with the error term, the method of moments provided the following estimating equations:

$$T^{-1} \sum_{t=1}^{T} \left(y_t - x_t'\beta\right) \mathbf{w}_t = \mathbf{0}. \tag{10}$$

If we have as many moment conditions as the number of elements in β, viz. $d = k$, the system can be solved in closed

[1] We abstract from the panel structure in this discussion, in the interest of simplicity in presentation.

form to provide the Instrumental Variables (IV) estimator:

$$\hat{\beta}_{IV} = (\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\mathbf{y}, \tag{11}$$

where $\mathbf{W}$ is the matrix containing all observations for $\mathbf{w}_t$, $\mathbf{X} = [x_t', t = 1, ..., T]$, $\mathbf{y} = [y_t, t = 1, ..., T]$. When $d > k$ (the so called over-identified case) the matrix $\mathbf{W}'\mathbf{X}$ is no longer square and, therefore, it cannot be inverted. Given the linear model in matrix notation: $\mathbf{y} = \mathbf{X}\beta + \mathbf{v}$, where $\mathbf{v} = [v_t, t = 1, ..., T]$, pre-multiplying both sides by $\mathbf{W}'$, we obtain: $\mathbf{W}'\mathbf{y} = \mathbf{W}'\mathbf{X}\beta + \mathbf{W}'\mathbf{v}$. Under the assumption that the distribution of $\mathbf{v}$ is proportional to an identity matrix (with unknown constant of proportionality), application of Generalized Least Squares (GLS) yields the following estimator:

$$\tilde{\beta} = [\mathbf{X}'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}, \tag{12}$$

which is known as Generalized Instrumental Variables Estimator (GIVE) or GMM. A similar approach is applied when the model is nonlinear as in (5). In this case, given a set of instruments, $\mathbf{w}_{it} \in \Re^d$, we obtain the following moment conditions:

$$(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'\beta - \sum_{g=1}^{G}\delta_g \log \Phi(z_{it}'\gamma_g))\mathbf{w}_{it} \equiv (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{g}(\theta, \mathcal{Y}_{it}) = 0, \tag{13}$$

where $\theta \in \Theta \subseteq \Re^p$ is the parameter vector containing β, γ_gs and δ_gs, and $\mathcal{Y}_{it} = (y_{it}, x_{it}, z_{it}, w_{it})$. Moreover,

$$\mathbf{g}(\theta, \mathcal{Y}_{it}) = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'\beta - \sum_{g=1}^{G}\delta_g \log \Phi(z_{it}'\gamma_g))\mathbf{w}_{it}. \tag{14}$$

The GMM estimator, $\hat{\theta}$, solves the following program:

$$\min_{\theta\in\Theta}[(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{g}(\theta, \mathcal{Y}_t)]'\mathbb{M}[(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{g}(\theta, \mathcal{Y}_{it})], \tag{15}$$

where $\mathbb{M}$ is a weighting matrix. Let $\bigtriangledown_\theta$ denote gradient taken with respect to θ. If we define $\mathbf{\Gamma} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \mathbf{g}(\theta, \mathcal{Y}_{it})$ and $\mathbf{\Omega} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{g}(\theta, \mathcal{Y}_{it})\mathbf{g}(\theta, \mathcal{Y}_{it})'$, then the GMM estimator has an asymptotic normal distribution centered at the true value of the parameter and covariance matrix: $cov(\hat{\theta}) = (\mathbf{\Gamma}'\mathbf{\Omega}^{-1}\mathbf{\Gamma})^{-1}$ provided we select $\mathbb{M} \propto \mathbf{\Omega}^{-1}$. This is the optimal choice of the weighting matrix that yields asymptotic efficiency. Apparently, it depends on the parameter θ. In this paper we use the so-called Continuously-Updated-Estimator (CUE) version of GMM in which we substitute $\mathbb{M}$ as a function of θ directly in (15). An alternative is the two-step estimator: First, we use $\mathbb{M} = I$, obtain a consistent estimator from (15), say, $\hat{\theta}_o$. We obtain $\hat{\mathbf{\Omega}}_o = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{g}(\hat{\theta}_o, \mathcal{Y}_{it})\mathbf{g}(\hat{\theta}_o, \mathcal{Y}_{it})'$ and, in the second stage, we plug in this estimate in (15) to provide the GMM estimator. Both versions are consistent and asympotically efficiency, although there is some evidence that the CUE-GMM estimator has better finite sample performance.

Of course, there is a parallel literature in econometrics that deals with endogeneity in panel data, see for example Ahn

Table 1: Marginal inefficiency effects from (5)

	nonlinear LS estimation of ANN		GMM estimation of ANN	
	sample mean	sample s.d.	sample mean	sample s.d.
age	-6.08	2.21	-2.38	9.19
education	16.71	6.08	1.69	6.54
household size	-16.17	5.88	-2.28	8.81
percentage of upland	-10.79	3.93	-1.45	5.58
time trend	8.52	3.10	0.76	2.92

Notes: This table shows summary statistics for the marginal effects of each variable. These marginal effects depend on the data.

et al. (2013), Kneip et al. (2012) as well as an important paper by Bai (2009). The problem with GMM estimators in panel data is that inefficiency estimates are not automatically non-negative and other methods are needed to transform residuals into inefficiency scores. Specifically, the application of the CSS estimator (Cornwell, Schmidt and Sickles, 1990) in Kneip et al. (2012) requires to determine efficiency as $v_i(t) - \max_{j=1,...,n} v_j(t)$ where $v_i(t)$ is interpreted as a time-varying firm efficiency factor. This paper and Paul and Shankar (2018) focus on models where inefficiency is non-negative by construction. Although we agree that Ahn et al. (2013) and Kneip et al. (2012) can handle endogeneity easily, this is not so when one is explicitly interested in inefficiency effects of exogenous variables and / or when one is not content with the assumption that one firm is always 100% efficient, which is a feature of CSS.[2]

2 Application

We apply the new model to the Philippines rice data of Paul and Shankar (2018) but we use different environmental variables. The data set contains annual data collected from 43 smallholder rice producers in Tarlac, Philippines between 1990 and 1997. The dependent variable is log of output (tonnes of freshly threshed rice). Inputs are logs of area planted (hectares), labor used (man-days of family and hired labor), fertilizer (kg of active ingredients) and other inputs used (Laspeyres index = 100 for Firm 17 in 1991). We use a Cobb-Douglas production function as in previous applications.[3] The environmental variables are: Age of the household head (years), education of the household head (years), household size, number of adults in the household and percentage of area classified as *bantog* (upland) fields. This data set is published as supplement to Coelli et al. (2005).

The Bayes Information Criterion criterion[4] in (2) was 262.2. With $G = 2$ the BIC was 60.038, and 319.9 with $G = 3$ so using $G = 2$ is optimal in this instance. Using the ANN formulation in (5) we obtain the marginal effects reported in Table 1. This table shows summary statistics for the marginal effects of each variable. These marginal effects depend on the data.

From these results it turns out that age, household size and percentage of upland, decrease inefficiency while inefficiency increases with education. Adding a squared term in education did not change this result. Using GMM[5] produces estimates that have the same signs but are numerically very different from nonlinear LS. The sample standard deviations are also

[2]CSS estimators provide relative inefficiency scores, that is one firm is always fully efficient.

[3]We use a standard conjugate gradients algorithm preceded by a random search to locate good initial conditions. WinGauss programs to perform the computations are available on request.

[4]The BIC is computed as BIC=$NT \log \sigma^2 + p \log(NT)$ where p is the number of parameters and σ^2 is the nonlinear least squares objective function divided by NT.

[5]The instruments include variables in z_{it} plus the logs of output price and the four input prices. GMM is implemented with a conjugate gradients algorithm staring from the final nonlinear LS estimates.

Table 2: Neural network coefficients

	$g=1$	$g=2$
δ_g	0.322	0.678
	(0.177)	(0.025)
γ_g constant	3.12	-0.52
	(0.74)	(0.11)
γ_g age	0.44	-0.31
	(0.32)	(0.44)
γ_g education	0.12	-0.43
	(0.06)	(0.21)
γ_g household size	-0.33	0.071
	(0.14)	(0.12)
γ_g percentage of upland	0.17	0.21
	(0.06)	(0.030)
γ_g time trend	0.002	0.23
	(0.001)	(0.017)

Notes: Standard errors appear in parentheses. Here, g denotes the particular component of the neural network in (5).

quite large. The effects of variables on inefficiency can be summarizing as in Table 1 but it is best to present figures similar to Figure 1 to examine whether the marginal effects are in effect constant. The marginal effects as a function of a specific variable (holding other variables fixed at their medians) are reported in Figure 2. All variables are scaled to be in the interval [0, 1] using the transformation[6] $z := \frac{z-min(z)}{max(z)-min(z)}$ so that they correspond, approximately, to percentiles.

In this instance, all marginal effects are statistically insignificant. In effect, variables such as upland, education etc. do not seem to exercise important effects on technical inefficiency, Having that said, the effects themselves are clearly non-constant and non-monotonic in the case of age and education. In Table 2, we report estimates of δs and γs along with their standard errors following the request of an anonymous reviewer. However, these coefficients lack any structural (economic) interpretation and are related only to the neural network approximation. In Table 2, we report estimates of δs and γs along with their standard errors following the request of an anonymous reviewer. However, these coefficients lack any structural (economic) interpretation and are related only to the neural network approximation.

To investigate further the issue of large standard errors we consider, in turn, another empirical application of the techniques.

3 Another empirical application

We use a global banking sample to provide comprehensive measures of bank efficiency scores across different economies. Our sample consists of 17,399 observations for 31 advanced countries, 7,130 observations for 35 emerging economies, and 2,471 observations for 40 developing countries. All bank-specific variables were obtained from Bankscope database. We use a cost function with two outputs (net loans and other earning assets), three inputs (financial capital -deposits and short-term funding-, labor, and physical capital -fixed assets). We include equity as a quasi-fixed input. We also include nonperforming loans (NPL) as a negative quasi-fixed input. As determinants of inefficiency we choose the Z-score, and the ratio of liquid assets over total assets. In addition, we use GDP per capita and inflation to proxy macroeconomic stability. Also, we include population density and market size to capture size effects of the banking industry. For detailed description of the data and

[6]The model is estimated using the original data. The transformation is applied only when inefficiency effects are plotted.

Figure 2: Marginal effects

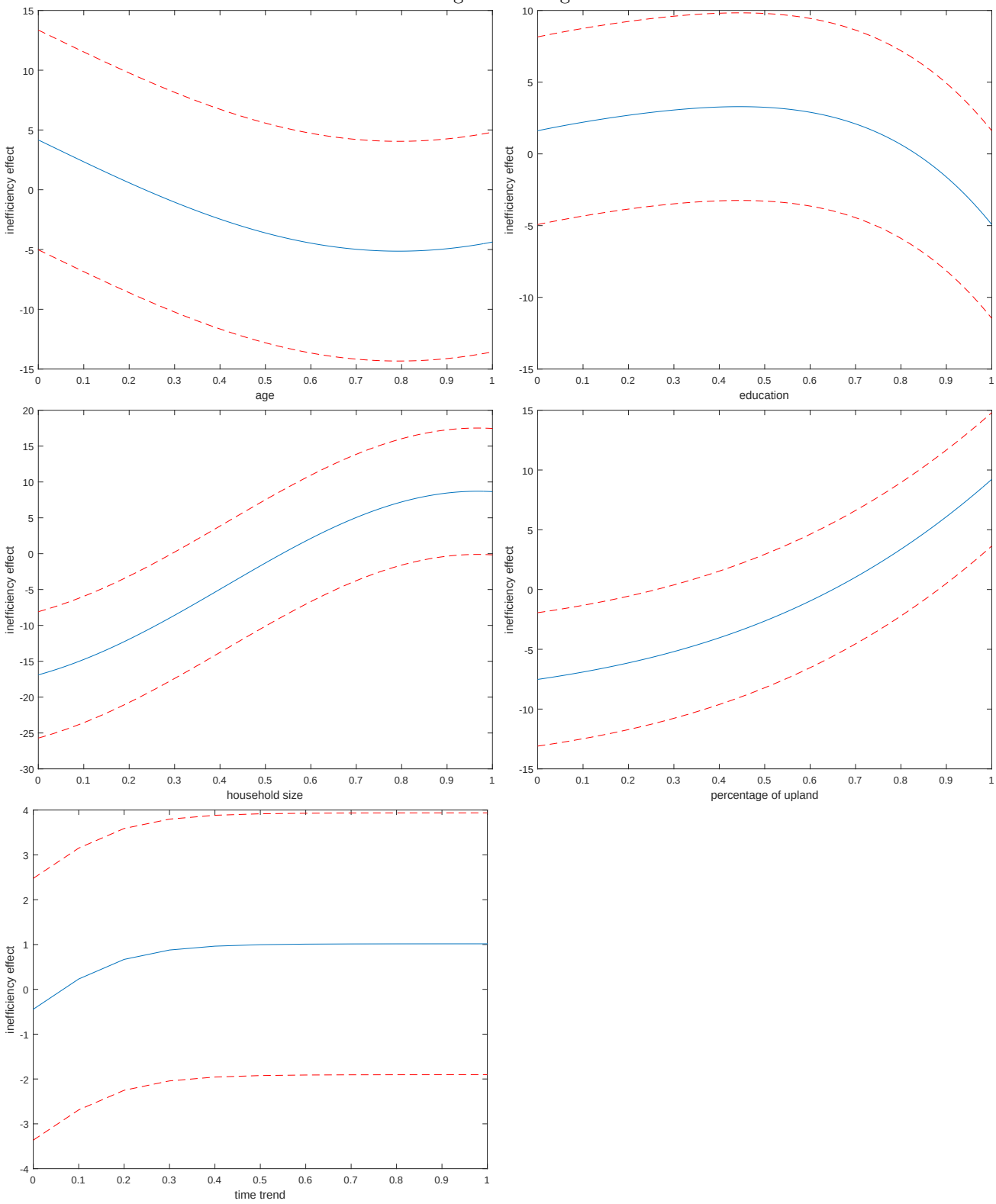


Table 3: Neural network parameters (global banking data)

	$g = 1$	$g = 2$	$g = 3$
δ_g	0.13	0.25	0.34
	(0.01)	(0.05)	(0.06)
δ_g Z-score	0.44	-0.23	0.32
	(0.05)	(0.06)	(0.11)
δ_g liquid assets / total assets	-0.13	0.14	0.25
	(0.06)	(0.01)	(0.01)
δ_g GDP per capita	0.33	0.12	-0.22
	(0.05)	(0.09)	(0.14)
δ_g inflation	-0.14	-0.22	0.13
	(0.05)	(0.03)	(0.01)
δ_g population density	0.33	0.25	-0.33
	(0.06)	(0.04)	(0.12)
δ_g market size	-0.33	-0.28	0.14
	(0.04)	(0.02)	(0.06)

Notes: Standard errors appear in parentheses, Here, g denotes the particular node / component of the neural network. Method of estimation was GMM to account for endogeneity.

the rationale for using these variables, see Tran et al. (2016). We include bank-specific fixed effects in both the cost function and the inefficiency determinants function to deal with the presence of heterogeneity in the sample, expecting a priori that banking systems differ widely across the world.

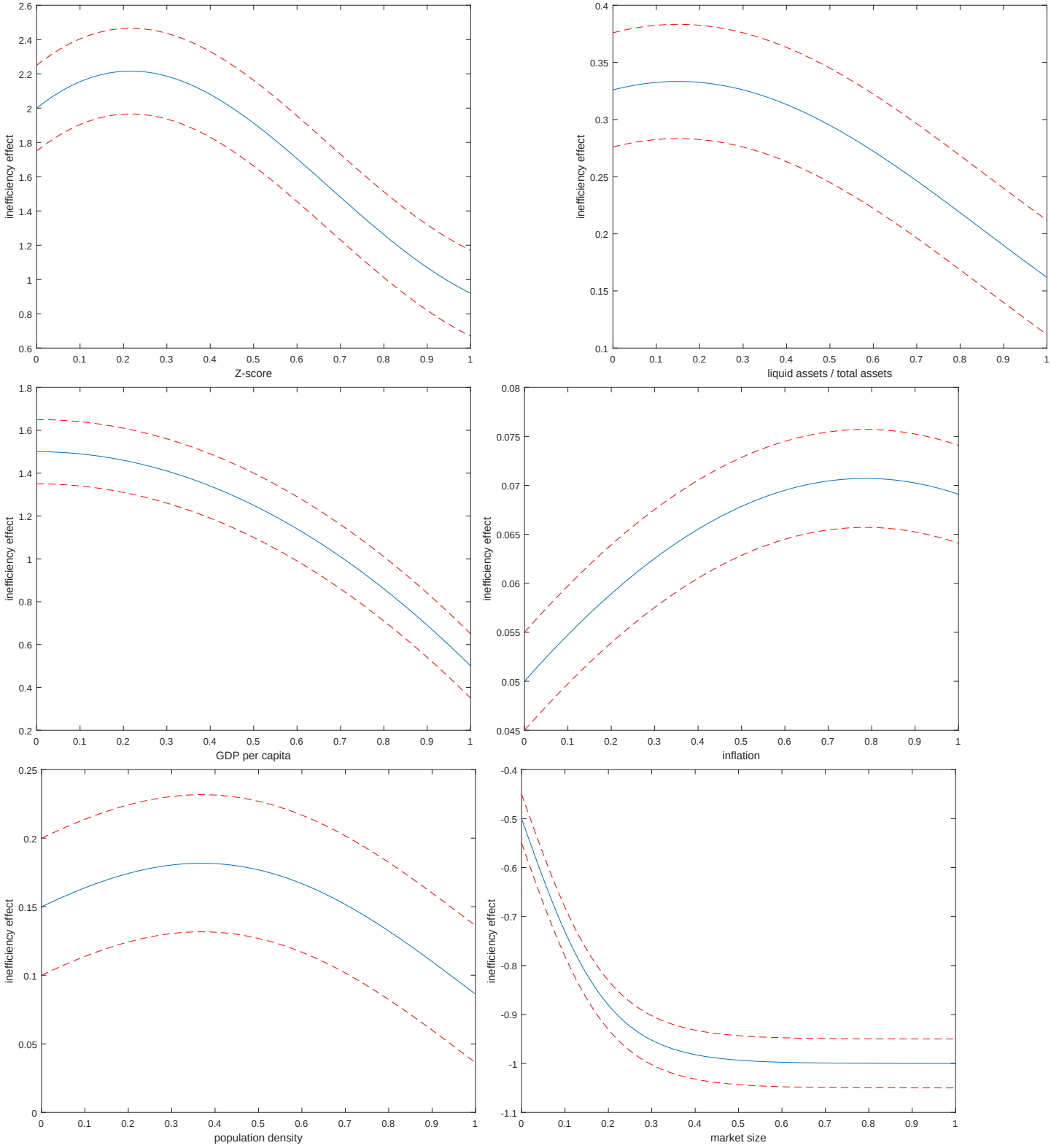
Our interest focuses on marginal effects of the different variables and their confidence intervals. For this application, the optimal number of nodes was $G = 3$ based on the BIC criterion. Again, we normalize all variables to be in the interval [0, 1]. The neural network parameters along with standard errors are reported in Table 3.

From the results in Figure 3, we can see that confidence bands are much tighter compared to the previous application. The method of estimation was CUE-GMM and we provide some details in the next paragraph. The narrow confidence bands owe much to the large sample size and, presumably, the fact that the sample is quite informative with respect to these quantities of interest.

Interestingly, Z-score increases inefficiency when it is lower than (approximately) the 25% percentile but it decreases inefficiency at higher levels. The result shows that considerable improvement in Z-score is necessary to increase efficiency. The same is true for the ratio of liquid assets to total assets and the cutoff point is near the 20% percentile for this variable, implying that liquid assets provide more flexibility to the banking sector as it reduces the risks that it is exposed to. GDP per capita and market size reduce inefficiency across the board, inflation increases inefficiency up to some level, and population density has a non-monotonic effect: It increases inefficiency up to some point (nearly the mean or median) but then it has a beneficial effect. As population density increases, servicing customers is increasingly problematic but after some point experience reduces overall cost inefficiency.

To implement CUE-GMM we remove the assumption that outputs and input prices in the cost function are exogenous. As instruments we use the lagged values of these variables (lagged once). In the inefficiency equation, Z-score and liquidity ratio are likely to be endogenous so we use their lagged values as instruments. The resulting orthogonality conditions are validated by Hansen's J-statistic (p-value was 0.30). The selection of the number of nodes, G, for the neural network is problematic as GMM does not deliver a BIC statistic. In the application of section 2 we took G directly from NLS estimation which may or may not be always correct. Although it is a reasonable procedure, we propose here a modification.

Figure 3: Marginal effects (Global banking application)



To determine the optimal value of G we follow Chernozhukov and Hong (2003). The criterion in (15) may be written as:

$$L_{NT}(\theta) = -\tfrac{1}{2}\left[\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{g}(\theta,\mathcal{Y}_t)\right]' \mathbb{M}\left[\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{g}(\theta,\mathcal{Y}_t)\right], \tag{16}$$

which we need to maximize (for selection of $\mathbb{M}$ we use the CUE approach). Although L_{NT} is not, in general, a log-likelihood function, one can define the following quasi-posterior:

$$p_{NT}(\theta) = \frac{e^{L_{NT}(\theta)}\tilde{\pi}(\theta)d\theta}{\int_{\Theta} e^{L_{NT}(\theta)}\pi(\theta)d\theta}, \tag{17}$$

where $\tilde{\pi}(\theta)$ is a "prior" which we take to be flat: $\tilde{\pi}(\theta) \propto$ const., $\forall\theta \in \Theta$. What we are interested in is the denominator of (17) which is known as marginal or integrated likelihood. The BIC criterion is, in fact, an asymptotic approximation to the marginal likelihood. The quantity $\mathscr{M} \equiv \int_{\Theta} e^{L_{NT}(\theta)}\tilde{\pi}(\theta)d\theta$ depends only on the data and can be estimated as follows. First, notice that

$$\mathscr{M} = \frac{\mathscr{L}_{NT}(\theta)\tilde{\pi}(\theta)}{p_{NT}(\theta)}, \ \forall\theta \in \Theta, \tag{18}$$

where $\mathscr{L}_{NT}(\theta) = e^{L_{NT}(\theta)}$. This is the well-known marginal likelihood identity (Chib, 1995). Given GMM parameter estimates $\hat{\theta}$ the numerator can be computed easily. The denominator is not available but can be estimated using the Laplace approximation (Lewis and Raftery, 1997) which assumes a normal (large-sample) approximation to the posterior, viz. $p_{NT}(\theta) \cong (2\pi)^{-d/2}|\Sigma|^{-1/2}e^{-\frac{1}{2}(\theta-\theta_o)'\Sigma^{-1}(\theta-\theta_o)}$, where d is the dimensionality of θ, θ_o is the mean which we can take it to be $\theta_o = \hat{\theta}$, and Σ is the asymptotic covariance matrix of the GMM estimator. Under these assumptions, the denominator of (18) becomes: $p_{NT}(\theta) \cong (2\pi)^{-d/2}|\Sigma|^{-1/2}$. With a flat prior, we have:

$$\log\mathscr{M} \cong L_{NT}(\hat{\theta}) + \tfrac{d}{2}\log(2\pi) + \tfrac{1}{2}\log|\Sigma|. \tag{19}$$

We choose G by computing this expression for different values of G and choose the one which maximizes the value of log marginal likelihood ($G = 3$ for this application).

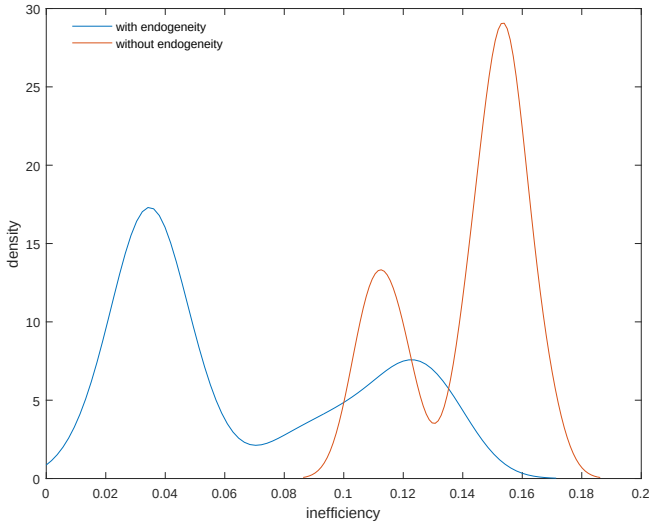
Finally, in Figure 4 we provide the density of inefficiency across the global banking industry.

We present densities under two assumptions, viz. when endogeneity of the arguments of the cost function is taken into account (blue line) and when this endogeneity is ignored (orange line). Although both densities are clearly bimodal, ignoring endogeneity produces inefficiency scores averaging (approximately) 15% and ranging from about 9% to 19%. With endogeneity taken into account, the density is supported over 9-19% but the dominant mode allows much smaller inefficiency estimates (from zero to almost 7%). Therefore, the two densities are quite different indicating that accounting for endogeneity may alter the results in substantive ways. In fact, the rank correlation between the two sets of inefficiency estimates is close to zero.

We omit a more detailed discussion of these effects as it would take us astray from the main point of this section. The main point is that confidence bands need not be as wide as in the application of the previous section, and the model is capable of capturing non-constant and non-monotonic effects on inefficiency. This is due to both the large sample size as well as the information contained in the sample as regards the functions of interest in this application. Finally, we should mention

Figure 4: Sampling densities of inefficiency (global banking data set)

Notes: Blue line: Density of inefficiency when endogeneity of the arguments of the cost function is taken into account. Orange line: Density of inefficiency when endogeneity of the arguments of the cost function is ignored.



that standard errors of parameter estimates can be made robust to autocorrelation by using a HAC (Heteroskedasticity-Autocorrelation-Consistent) correction. This correction involves simple modification of the covariance matrix of the GMM estimator and is readily available in commonly available software.

Concluding remarks

We have proposed a neural-network formulation of the inefficiency effects function which is general enough to accommodate arbitrary inefficiency effects functions. It avoids the unfortunate drawback of the model in Paul and Shankar (2018) that the ratio of any two marginal effects is constant across all values of the variables. We show that the model can be estimated using GMM and accommodate arbitrary patterns of endogeneity. Moreover, unlike other GMM techniques for panel data, the model provides absolute, not relative efficiency scores. Therefore, it is not necessary to assume that a firm is always fully efficient. The model is, therefore, along with the one proposed by Paul and Shankar (2018), the only one that it delivers absolute efficiency measures without distributional assumptions on the two components of the error term. In addition, this paper shows how to provide such absolute efficiency scores under endogeneity. This opens up the possibility to implement stochastic frontier analysis with inefficiency effects in much more general settings that was thought before. In terms of future applications, the new model opens up the possibility of extensions in several directions, including dynamic panel data -a novel class of models in stochastic frontier analysis that has not been analyzed before- as well as many other areas such as models with heteroskedasticity and autocorrelation.

References

Ahn, S.C., Lee, Y. H., & Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. Journal of Econometrics 174 (1), 1–14.

Bai, J. (2009). Panel Data Models With Interactive Fixed Effects, Econometrica 77 (4), 1229–1279.

Chernozhukov, V., & Hong, H. (2003). An MCMC approach to classical estimation. Journal of Econometrics, 115, 293– 346

Chib, S. (1995). Marginal likelihood from the Gibbs output. Journal of the American Statistical Association, 90: 1313–1321.

Coelli, T. J., Rao, D. S. P., O'Donnell, C. J. , & Battese, G. E. (2005). *An Introduction to Efficiency and Productivity Analysis*. Springer, New York.

Cornwell, C., Schmidt, P., & Sickles, R. C. (1990). Production frontiers with cross-sectional and time-series variation in efficiency levels. Journal of Econometrics 46: 185–200.

Deprins, D. & Simar, L. (1989). Estimation de frontières déterministes avec facteurs exogènes d'inefficacité. Annales d'Economie et de Statistique 14, pp. 117–150.

Hornik, K., Stinchcombe, M. & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks 2 (5), 359–366.

Kneip, A., Sickles, R., & Song, W. (2012). A new panel data treatment for heterogeneity in time trends. Econometric Theory, 28(3), 590–628.

Kumbhakar, S. C., & Lovell, C. A. K. (2000). *Stochastic Frontier Analysis*. Cambridge, Cambridge University Press.

Lewis, S. M. and Raftery, A. E. (1997). "Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator." Journal of the American Statistical Association, 92: 648–655.

Paul, S. & Shankar, S. (2018). On estimating efficiency effects in a stochastic frontier model. European Journal of Operational Research, 271, 769–774.

Tran, K., Mamatzakis, E., & Tsionas, M. (2016). Why Fully Efficient Banks Matter? A Nonparametric Stochastic Frontier Approach in the Presence of Fully Efficient Banks. Text available at : https://www.researchgate.net/publication/301219515