A Novel Model of Costly Technical Efficiency

Mike G. Tsionas, Marwan Izzeldin

PII: S0377-2217(18)30034-1
DOI: 10.1016/j.ejor.2018.01.016
Reference: EOR 14917

To appear in: *European Journal of Operational Research*

Received date: 17 March 2017
Revised date: 13 December 2017
Accepted date: 6 January 2018

Please cite this article as: Mike G. Tsionas, Marwan Izzeldin, A Novel Model of Costly Technical Efficiency, *European Journal of Operational Research* (2018), doi: 10.1016/j.ejor.2018.01.016

Highlights

- A novel model of technical inefficiency based where improvements are costly.
- Equivalent to model where inefficiency is arbitrary function in input-output space.
- We determine optimal directions in input-output space.
- Guide for policy actions to reduce inefficiency.
- Bayesian techniques are used to statistical inferences.

A Novel Model of Costly Technical Efficiency

Mike G. Tsionas*and Marwan Izzeldin†

February 8, 2018

Abstract

This paper presents a novel model of measuring technical inefficiency based on the notion that higher efficiency requires a certain cost. First, we apply the "rational inefficiency hypothesis" of Bogetoft and Hougaard (2003) but we fail to find that it rationalizes our data set of large U.S banks with multiple inputs and outputs. In consequence, we adopt a novel model of profit maximization which explicitly incorporates the cost of technical inefficiency. The cost of inefficiency is treated as unknown and is parametrized as a function of inputs, outputs and decision-making-unit specific fixed effects. More importantly, by showing the model to be equivalent to one in which inefficiency is an arbitrary function of inputs, outputs and the inefficiency cost, we are able to determine optimal directions in the input-output space that would reduce inefficiency. Bayesian techniques organized around Markov Chain Monte Carlo are used to perform the computations and provide statistical inferences as well as useful policy measures to reduce inefficiencies in the U.S banking sector through an examination of different realistic scenarios.

Key Words: Production; Technical Inefficiency; Profit Maximization; Distance Functions; Bayesian Methods.

*Lancaster University Management School, LA1 4YX, U.K. & Athens University of Economics and Business, Greece, m.tsionas@lancaster.ac.uk

†Lancaster University Management School, LA1 4YX, U.K.

1 Introduction

Economic and operational research models that address technical inefficiency, abstract from a fundamental concern, that improvements in efficiency require certain costs in terms of re-allocating resources and better management of the production process etc.[1] However, it is clear that technical efficiency cannot be improved unless real resources are used and certain costs are incurred. As efficiency is under the control of the firm, it must be the case that observed efficiencies are the result of an optimization process that takes explicitly into account the real costs of improvements in efficiency. Although Bogetoft and Hougaard (2003) make a similar point that observed efficiencies are optimal, their techniques are quite distinct to those proposed in this paper.

The "rational inefficiency hypothesis" of Bogetoft and Hougaard (2003) has empirical implications which are not supported by our data set. Therefore, we adopt a profit maximization model with an output distance function to handle the multiple-output, multiple-input nature of production. We incorporate the cost of technical inefficiency in profit maximization and we use duality of the profit function to derive, not only input demand functions and output supply functions, but also inefficiency as a function of input and output prices and the cost of technical inefficiency. The resulting system of equations, which includes the unobserved cost of inefficiency, is estimated using Bayesian techniques organized around Markov Chain Monte Carlo (MCMC), specifically the Girolami and Calderhead (2011) Riemannian Manifold Hamiltonian technique.

Another important implication of our model is that it can be viewed as a model where inefficiency is an arbitrary function of inputs and outputs. In this sense, these results, of course, give us directions in the input-output space that can be reasonably used in other studies or banking policy makers with a purpose towards reducing inefficiency. Usually, such concerns are not taken into account although they clearly impose an overhead in empirical analysis, see for example Asmild and Matthews (2012), Balezentis and DeWitte (2015), Biener et al. (2016), Reyes et al. (2016), Saranga (2009), Tecles and Tabak (2010), Sena (2016), Tsionas and Mamatzakis (2017), Annaert et al. (2003), Chen et al. (2016), Badunenko and Kumbhakar (2017) and Sun et al. (2015). Simar et.al (2016) provide a probabilistic formulation of production to give an original characterization of the directional distances. They define robust versions of directional distance functions, introducing order-m and order-α quantile versions of distance functions. They also allow exogenous factors in the production process. Bădin et.al (2012) examine the impact of environmental factors on production process using a new two-stage type approach. The authors use conditional measures to overcome the inconsistencies associated with the traditional two-stage analysis. Daraio and Simar (2014) compute conditional and unconditional directional distances and outline how the approach of Bădin et al. (2012) can be modified in the context of directional distance to develop two-stage analysis of efficiency scores. In this literature, inefficiency is taken as a "given" in the sense that it is not the outcome of an optimization problem by the firm. In this paper we take a different approach and assume that inefficiency is under the control of the decision making unit, albeit at a cost. Input demands and output supply functions are derived formally from the profit maximization problem along with an optimal inefficiency function which depends on input and output prices as well as the cost of inefficiency. The challenging part in the empirical implementation of the model is that the inefficiency cost is unobserved.

Since we estimate specific directions[2] in the input-output space, it is critical for policy makers to reduce inefficiencies at the level of the firm and the sector as a whole. Thus far, many studies estimate technical inefficiency but few focus on what types of measures can be taken to reduce it. Other studies focused on incorporating *exogenous* influences on inefficiency (see previous references). However, one can argue that it is not only exogenous or environmental factors that matter. We can imagine, for example, that regulation is an important exogenous or environmental factor but that the input-output choices of the firm should also determine its overall level of technical

[1] See for example Aigner, Lovell and Schmidt (1977), Cornwell, Schmidt and Sickles (1990), Sickles (2005), Kutlu and Sickles (2012), Kumbhakar and Tsionas (2006) and chapter 3 in the excellent monograph of Kumbhakar and Lovell (2000).

[2] When inefficiency is measured we use a radial measure, that is all inputs are expanded or contracted by the same %. with different directions this is no longer necessary so we can for example increase capital by 5% and decrease labor by 7%

inefficiency, conditional on the environment. In our work we do not use the "separability" assumption of Simar and Wilson (2011) which can be important in their data generating process but not in ours. The reason is that because technical inefficiency is the outcome of profit maximization, it is impossible to argue that "separability" holds. In fact inefficiency depends on all inputs and outputs.

Our work is related to Bogetoft and Hougaard (2003) model that explains how inefficiency, on some occasions, may be a rational choice. For example, the gains that may accrue from observed inefficiency might derive from meeting other preferences or from unusual market conditions that permit, say, rent seeking. Where there is generally ignorance of such considerations, the argument that inefficiency should be eliminated is undermined. The alternative approach taken by Bogetoft and Hougaard (2003) is to model the slack selection process and to draw inferences about the relative value of different types of slack. As those authors explain: where the conventional approach is to attribute X-inefficiency to sub-optimal behavior, 'if what appears to be technical inefficiency at first sight actually is the result of rational and optimizing behavior of the firm such undertakings seem fruitless indeed, if not directly harmful over time' (Bogetoft and Hougaard, 2003, p. 246). Although we agree that inefficiency is, certainly, at least to a great extent, the outcome of an optimization process, we highlight that it is not fruitless at all to examine specific policy measures to reduce inefficiency. This is because inefficiency may also depend on the decision variables of the optimization process which can be manipulated to move firms closer to the frontier. For example, if inefficiency itself depends on inputs and / or outputs, which is quite likely in practice, then the Bogetoft and Hougaard (2003) framework may not be appropriate. In their model specification, input slacks are decision variables, which is hard to disagree with, but they are not allowed to vary systematically with inputs or outputs. Therefore, it is not feasible to explore the implications of alternative input - output combinations that would yield more efficiency at the cost of non-detrimental decreases in the objective function of the firm -which is a non-decreasing function of profit and input slacks.

The core message is that, where the elimination of inefficiency requires costly resources, observed inefficiencies may be a proxy 'for the omitted aspects or values of slack and that this proxy may be used to predict more precisely the results of changes in the control instruments as well as to define other types of measurement, e.g., measures of allocative efficiency' (Bogetoft and Hougaard, 2003, p. 246). However, contrary to that view, we do not regard slack as desirable. Rather, while we hold to the Leibensteinian view that inefficiency is undesirable, we complement that analysis by showing how inefficiency is difficult to eliminate. Bogetoft and Hougaard (2003) adopt the view that inefficiency represents resources that produce unobserved desirable outputs. Although valid in principle, this view cannot always hold. For example, in our application, slacks are not positive and, therefore, the Bogetoft and Hougaard (2003) view is not definitive (see Asmild et.al, 2013). Despite that this does not preclude model mis-specifications, it is not compatible with the 'benevolent' approach to inefficiency. Therefore, we look for an alternative in a model is equivalent to a model where inefficiency is an *arbitrary function of inputs, outputs and the inefficiency cost*, so that we can determine optimal directions in the input-output space. In turn, we can use these directions to suggest policy actions that must be taken to reduce inefficiency. We do this using alternative plausible scenarios.

The paper is organized as follows. In Section 2 we present the data for this study (U.S banks during the period 2000-2010) and outline the "rational inefficiency hypothesis" of Bogetoft and Hougaard (2003). The proposed model specification is presented in section 3. In section 4, we provide an alternative interpretation of the model that is compatible with optimal directions in the input - output space. The empirical application is presented in section 5, where policy implications are also derived.

2 Data and the "rational inefficiency hypothesis"

2.1 Data

A decade of quarterly data, 2001-2010, are taken from the 'Call Reports' of the Federal Reserve Bank of Chicago. These relate to FDIC insured commercial banks, where the data vary widely by size, capitalization, regulatory environment, and so on. To ameliorate the potential for heterogeneity in production, Malikov, Kumbhakar and Tsionas (2016) select a sub-sample of larger banks, so giving an unbalanced panel of 2,397 bank-year observations for 285 banks. Metrics for the outputs of a bank's production process are: consumer loans (y_1); real estate loans (y_2); commercial and industrial loans (y_3); securities (y_4) and off-balance-sheet income (y_5). Financial equity is also included as a quasi-fixed input. For variable inputs, we use: full-time equivalent employees (x_1); physical capital (x_2); purchased funds (x_3); interest-bearing transaction accounts (x_4) and non-transaction accounts (x_5). All nominal stock variables are deflated to 2005 U.S. dollars. Summary statistics for the data are listed in Table 1 (see Malikov, Kumbhakar and Tsionas, 2016, pp. 1420-21).

2.2 An examination of the "rational inefficiency hypothesis"

Let $x \in \Re^K$ represent a vector of inputs whose prices $w \in \Re^K$, and $y \in \Re^M$ represent a vector of outputs.We denote by x_i and y_i the input and output vector for a decision-making-unit $i \in \{1, ..., n\}$. In this section we set out to explore the empirical implications of the Bogetoft and Hougaard (2003) approach. First, the technology is estimated using the DEA approach by solving the following problem:

$$\begin{array}{c} \min \vartheta : \\ (\vartheta x_i, y_i) \in \mathcal{T}^*, \end{array} \tag{1}$$

where

$$\begin{array}{c} \mathcal{T}^* = \{(x, y) \in \Re^K \times \Re^M : \sum_{j=1}^n \lambda_j x_{jk} \leq x_k,\ k = 1, ..., K, \\ \sum_{j=1}^n \lambda_j y_{jk} \geq y_m,\ m = 1, ..., M,\ \lambda_j \geq 0,\ j = 1, ..., n\}. \end{array} \tag{2}$$

We can also determine the cost-minimizing combinations by solving:

$$\begin{array}{c} \min_{x,\lambda} :\ w'x, \\ \sum_{j=1}^n \lambda_j x_{jk} \leq x_k,\ k = 1, ..., K, \\ \sum_{j=1}^n \lambda_j y_{jk} \geq y_m,\ m = 1, ..., M, \\ \lambda_j \geq 0,\ j = 1, ..., n. \end{array} \tag{3}$$

Suppose the optimal input mix from (3) is z_i. The "rational inefficiency hypothesis" implies that the following slacks:

$$s_{ik} = \frac{x_{ik} - z_{ik}}{z_{ik}},\ k = 1, ..., K, \tag{4}$$

must be non-negative[3], "*assuming that the estimated technology is a good estimate of the 'true' technology.*" (Asmild et al., 2013, p. 83). To estimate the technology as accurately as possible, taking account of (possible) heterogeneity, despite the fact that we use only large banks which are likely to operate under similar conditions, we estimate different DEA problems for each decile of banks (deciles defined by total assets). Therefore, we address ten different situations similar to (1) and (3). The results are reported in Table 1.

[3]In the Bogetoft and Hougaard (2003) approach the decision-making-unit is equipped with a utility function, $U_i(\Pi_i, s_i)$, where Π_i is profit and $s_i \in \Re^K$ is the vector of input slacks. The utility function is assumed non-decreasing in all its arguments.

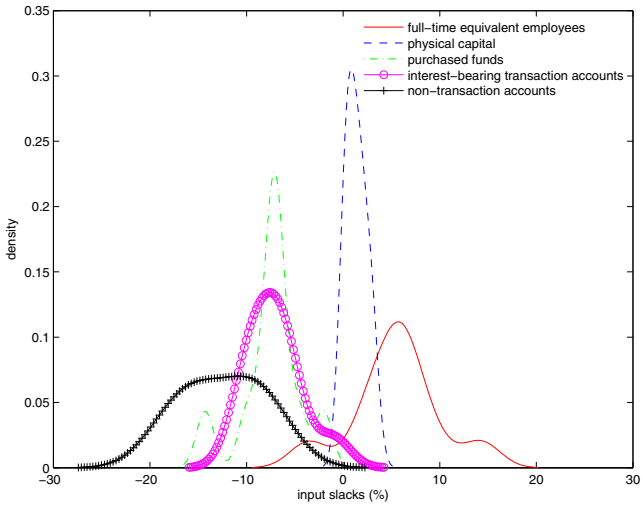
Table 1: Descriptive statistics of optimal input mix

input	slack on average (%)
x_1, full-time equivalent employees	1.37% (4.44%)
x_2, physical capital	1.42% (1.38%)
x_3, purchased funds	-6.17% (2.25%)
x_4, interest-bearing transaction accounts	-9.17% (2.50%)
x_5, non-transaction accounts	-12.25% (4.71%)

Notes: The table reports sample averages with sample standard deviations in parentheses.

The results in Table 1 give little support to the "rational inefficiency hypothesis" as most inputs are associated with sizable negative slacks. Although, on the average, the slacks of full-time equivalent employees and physical capital are positive, the standard deviations are quite large allowing for negative values. For purchased funds, interest-bearing transaction accounts and non-transaction accounts, all slacks are negative on the average and their sample distributions are not compatible with positive values (see Figure 1). Even for full-time equivalent employees there is a sizable proportion of negative slacks (slightly less than 50%) which cannot be reconciled with the "rational inefficiency hypothesis". As a consequence, we resort to a different way of rationalizing inefficiency.

Figure 1: Sample distributions of input slacks



Notes: For descriptive statistics, see Table 1.

In Asmild et al. (2013) the authors found some negative deviations but these were quite small. Therefore, they decided to ignore small negative deviations "small" being defined as less than 0.5 full-time-equivalent employment. In our case, in Figure 1, even if we truncate small negative deviations, most remain quite large, i.e., greater than -10% for full-time-equivalent employment and purchased funds and -20% for non-transaction accounts.

3 A New Model

3.1 Theoretical Framework

Let $x \in \Re^K$ again, represent a vector of inputs whose prices$w \in \Re^K$ and $y \in \Re^M$ represent a vector of outputs whose prices are $p \in \Re^M$, and $0 < u \leq 1$ is technical inefficiency. Production possibilities are represented by the technology set $\mathcal{T} = \left\{(x, y) \in \Re^K \times \Re^M : \ x \text{ can produce } y\right\}$ which can be described by an output distance function $D(x, y) = 1$, where $D : \Re^K \times \Re^M \to \Re$. We can write the output distance function in the form: $y_1 = f(y_2, ..., y_M, x)$. Introducing technical inefficiency we have the representation:[4]

$$y_1 = f(y_2, ..., y_M, x)\, e^{-u}. \tag{5}$$

Since technical inefficiency is commonly assumed to be under the control of the firm, it should be possible to reduce it at the cost of q per unit, where $q > 0$. The profit maximization problem of the firm then becomes:

$$\Pi(p, w, q) = \max_{x \in \Re^K, y \in \Re^M, u \in (0,1]} : \ \sum_{m=1}^{M} p_m y_m - \sum_{k=1}^{K} w_k x_k - qu, \tag{6}$$

where$\Pi(p, w, q)$ is the profit function. Using (5), we can write the problem as:

$$\Pi(p, w, q) = \max_{x \in \Re^K, y \in \Re^M, u \in (0,1]} : \ p_1 \left\{f(y_2, ..., y_M, x)\, e^{-u}\right\} + \sum_{m=2}^{M} p_m y_m - \sum_{k=1}^{K} w_k x_k - qu. \tag{7}$$

From standard duality, we have the following results:

$$\frac{\partial \Pi(p, w, q)}{\partial p_1} = y_1^*(p, w, q) = f(y_2^*, ..., y_M^*, x)\, e^{-u^*}, \tag{8}$$

$$\frac{\partial \Pi(p, w, q)}{\partial p_m} = y_m^*(p, w, q), \ m = 2, ..., M, \tag{9}$$

$$-\frac{\partial \Pi(p, w, q)}{\partial w} = x^*(p, w, q), \tag{10}$$

$$-\frac{\partial \Pi(p, w, q)}{\partial q} = u^*(p, w, q), \tag{11}$$

where $y^*(p, w, q) \in \mathbb{R}^M$ represents optimal output supplies, $x^*(p, w, q) \in \mathbb{R}^K$ represents optimal input demands and $u^*(p, w, q)$ represents optimal inefficiency which is a function of all prices as well as q. Therefore, we can derive technical inefficiency endogenously with multiple inputs and outputs.

Given the duality results, we can specify a profit function $\Pi(p, w, q)$ and use the duality results in (8)-(11). The advantage is clear. For example, in the classical profit maximization framework:

$$\max_{x,y} : \ p'y - w'x, \ \text{s.t. } F(x, y) = 1,$$

where $F(x, y)$ is a general production transformation function, the objective function is $\Pi(p, w)$. Duality then, ascertains that we can obtain input demand functions as $x(p, y) = -\frac{\partial \Pi(p,w)}{\partial w}$ and output supply functions as $y(p, y) = \frac{\partial \Pi(p,w)}{\partial p}$. Therefore, we do not have to specify the transformation function, $F(x, y)$, and solve the possibly

[4]From linear homogeneity with respect to outputs, we can write the output distance function in log terms as follows: $\ln y_1 = g(\ln \frac{y_2}{y_1}, ..., \ln \frac{y_M}{y_1}, \ln x) - u$.

cumbersome profit maximization problem. Instead we can specify directly a profit function, $\Pi(p, w)$ and derive input demand and output supply functions by direct differentiation.

The problem in our particular context is, of course, that we rarely if ever we have information on q, which can be thought of as a subset of prices. The problem can be solved if we parametrize q. Specifically suppose $\{p_{it}, w_{it}, q_{it}\} \equiv \{z_{it}\}$ and we specify a flexible quadratic functional form for the profit function:

$$\Pi_{it} = \beta_o + \beta' z_{it} + \tfrac{1}{2} z'_{it} \Gamma z_{it}, \tag{12}$$

where β_o is a scalar, $\beta \in \mathbb{R}^d$ and $\Gamma$is a $d \times d$ matrix of parameters, where $d = M + K + 1$. This function is flexible in the sense that it is a second-order approximation to an arbitrary profit function. Such second-order approximations are well established in the relevant literature and, therefore, we use them here as well with relative safety.

Let us rewrite (12) in expanded form as follows[5]:

$$\begin{aligned} \Pi_{it} = \beta_o + \beta'_p p_{it} + \beta'_w w_{it} + \beta_q q_{it} + \tfrac{1}{2} p'_{it} \Gamma_{pp} p_{it} + + \tfrac{1}{2} w'_{it} \Gamma_{ww} w_{it} + \tfrac{1}{2} \gamma_{qq} q^2_{it} + \\ p'_{it} \Gamma_{pw} w_{it} + p'_{it} \tilde{\gamma}_{pq} q_{it} + w'_{it} \tilde{\gamma}_{qw} q_{it} + \\ \delta_t t + \tfrac{1}{2} \delta_{tt} t^2 + p'_{it} \tilde{\gamma}_{pt} t + w'_{it} \tilde{\gamma}_{wt} t + \delta_{qt} q_{it} t, \end{aligned} \tag{13}$$

where we have also introduced a time trend to capture technical progress and a tilde indicates a vector. As can be seen from the above expression, the profit function depends on levels of prices and a trend, their squares as well as their interactions to provide a full second-order approximation to an arbitrary profit function.

By differentiating (13) and using the duality results in (8)-(11) we have the following equations:

$$\begin{aligned} \tfrac{\partial \Pi_{it}}{\partial p_{it}} &= \beta_p + \Gamma_{pp} p_{it} + \Gamma_{pw} w_{it} + \tilde{\gamma}_{pq} q_{it} + \tilde{\gamma}_{pt} t = y_{it}, \\ \tfrac{\partial \Pi_{it}}{\partial w_{it}} &= \beta_w + \Gamma_{pw} w_{it} + \Gamma_{ww} w_{it} + \tilde{\gamma}_{qw} q_{it} + \tilde{\gamma}_{wt} t = -x_{it} \\ \tfrac{\partial \Pi_{it}}{\partial q_{it}} &= \beta_q + \gamma_{qq} q_{it} + \tilde{\gamma}_{pq} p_{it} + \tilde{\gamma}_{qw} w_{it} + \delta_{qt} t = u_{it}. \end{aligned} \tag{14}$$

It remains necessary to specify a model for q_{it} as data are not available. We assume that inefficiency cost, q_{it}, is given by the following equation which depends on all inputs, outputs and the time trend.

$$q_{it} = \mu_i + \lambda_t + \alpha_e e_{it}, \tag{15}$$

where μ_i represents fixed effects, λ_t represents time effects and e_{it} is equity which is considered as a quasi-fixed input [6]. In this specification, the inefficiency cost is decomposed into a firm-specific component, a time-specific component and a component related to equity which is, possibly, an important determinant of the inefficiency cost. The coefficient α_e captures the effect of equity on the cost of inefficiency.

After introducing error terms (denoted by v) which are necessary for econometric inferences, we can write (14) and (15) in the following form:

$$\begin{aligned} y_{it} &= \beta_p + \Gamma_{pp} p_{it} + \Gamma_{pw} w_{it} + \tilde{\gamma}_{pq} q_{it} + \tilde{\gamma}_{pt} t + v_{y,it}, \\ -x_{it} = \tfrac{\partial \Pi_{it}}{\partial w_{it}} &= \beta_w + \Gamma_{pw} w_{it} + \Gamma_{ww} w_{it} + \tilde{\gamma}_{qw} q_{it} + \tilde{\gamma}_{wt} t + v_{x,it}, \\ q_{it} &= \mu_i + \lambda_t + \alpha_e e_{it} + v_{q,it}. \end{aligned} \tag{16}$$

For the error vector $v_{it} = [v_{y,it}, v_{x,it}, v_{q,it}]$ we assume $v_{it} \sim \mathcal{N}_{M+K+1}(O, \Sigma)$, a multivariate normal distribution. Notice that we omit the third equation in (14) as it contains the unobserved inefficiency u_{it}. Inefficiency can

[5] Linear homogeneity of the profit function in all outputs is imposed by dividing all prices by the first input price.
[6] Apparently a vector of quasi-fixed inputs can be included in the specification.

be recovered using the duality result:

$$u_{it} = \beta_q + \gamma_{qq} q_{it} + \tilde{\gamma}_{pq} p_{it} + \tilde{\gamma}_{qw} w_{it} + \delta_{qt} t. \tag{17}$$

To obtain meaningful results we have to impose the standard restrictions in the profit function, viz. that it is non-increasing in input prices and inefficiency cost, and non-decreasing in output prices. This implies that profits cannot go up, *ceteris paribus*, when input prices or inefficiency costs increase, and profits cannot decrease when output prices increase, *ceteris paribus*. These intuitive results have also been proven mathematically in the relevant literature.

Most importantly, perhaps, we have also to impose the restriction that the RHS of (17) must also be positive, otherwise we would end up with negative inefficiency estimates. Suppose $\boldsymbol{\theta} \in \mathbb{R}^D$ denotes the vector of all parameter estimates in (13) and (15).D is the dimensionality of the parameter vector. The monotonicity restrictions adopted have the form:

$$\mathcal{M}(\boldsymbol{\theta}, \mathcal{Y}) \leq O, \tag{18}$$

where $\mathcal{Y} = \{y_{it}, x_{it}, p_{it}, w_{it}\}$ denotes the available data and $\mathcal{M} : \mathbb{R}^D \times \mathbb{R}^{2(M+K)} \to \mathbb{R}^{(M+K+2)N'}$ denotes a vector field describing the restrictions, where N' denotes the number of available observations. Specifically, we have D elements in the parameter vector $\boldsymbol{\theta}$, $2(M+K)$ elements in the data for each observation and $M+K+1$ restrictions in total from (16) and one restriction from (17) to ensure inefficiency is non-negative. The set of restrictions in (18) simply enforces the requirements that the profit function is non-increasing in input prices and inefficiency cost, and non-decreasing in output prices. Due to the quadratic approximation, these restrictions are not simple restrictions on the parameter vector $\boldsymbol{\theta}$. Instead they involve the data as well, thus the notation $\mathcal{M}(\boldsymbol{\theta}, \mathcal{Y})$.

The system of (16) and (15) can be estimated by maximum likelihood (ML) if we ignore (18). As these constraints cannot be ignored we have to use special techniques which are reported in the Technical Appendix. These techniques are Bayesian in nature and rely on Markov Chain Monte Carlo (MCMC).

3.2 Returns to scale and inefficiency

For simplicity, let us assume we have a single output but multiple inputs, and production possibilities are described by

$$y = f(x)e^{-u},\ 0 < u \leq 1.$$

The profit function is generalized to be:

$$\Pi(p, w) = \max_{x,u} :\ pf(x)e^{-u} - w'x - q(u, x), \tag{19}$$

where p is the price of output and $q(u, x)$ represents a *general* function of technical inefficiency *and* inputs themselves, with $\frac{\partial q(u)}{\partial u} < 0,\ \forall u \in (0, 1]$. The first-order conditions for profit maximization are:

$$\begin{aligned} pf_k(x)e^{-u} &= w_j + \tfrac{\partial q(u,x)}{\partial x_k},\ \forall k = 1, ..., K, \\ pf(x)e^{-u} &= -\tfrac{\partial q(u,x)}{\partial u}. \end{aligned} \tag{20}$$

From the first K conditions we obtain:

$$\frac{f_k(x)}{f_1(x)} = \frac{w_k + \frac{\partial q(u,x)}{\partial x_k}}{w_1 + \frac{\partial q(u,x)}{\partial x_1}},\ \forall k = 2, ..., K, \tag{21}$$

which implies that the first-order conditions for cost minimization are *not* satisfied and, therefore, *allocative inefficiency is induced as well.*

By dividing the two conditions in (20) we obtain:

$$\frac{f_k(x)}{f(x)} = -\frac{w_k + \frac{\partial q(u,x)}{\partial x_k}}{\frac{\partial q(u,x)}{\partial u}}. \tag{22}$$

Multiplying both sides by x_j and summing up, we have:

$$\sum_{k=1}^{K} \frac{f_k(x)x_j}{f(x)} = -\frac{\sum_{k=1}^{K} w_k x_k + \sum_{k=1}^{K} \frac{\partial \log q(u,x)}{\partial \log x_k} q(u,x)}{\frac{\partial \log q(u,x)}{\partial \log u} \cdot \frac{q(u,x)}{u}}. \tag{23}$$

In this expression, $RTS(x) \equiv \sum_{k=1}^{K} \frac{f_k(x)x_k}{f(x)}$ is the returns-to-scale and $TC_x = \sum_{k=1}^{K} w_k x_k$ represents total input cost. Moreover, $\varepsilon_k(u,x) \equiv \frac{\partial \log q(u,x)}{\partial \log x_k}$ represents the elasticity of the inefficiency costs with respect to input x_k and $\varepsilon_u(u,x) \equiv \frac{\partial \log q(u,x)}{\partial \log u}$ is the elasticity of the inefficiency costs with respect to technical inefficiency. The expression can be simplified as follows:

$$RTS(x) = -\frac{TC_x + \sum_{k=1}^{K} \varepsilon_k(u,x)}{\varepsilon_u(u,x)} \cdot u. \tag{24}$$

This provides a way to obtain *returns-to-scale* from the profit maximization model. Apparently $x = x^*(p,w)$, $u = u^*(p,w)$, the profit-maximizing input demand vector field. In the special case where inefficiency costs do not depend on the inputs, that is $\frac{\partial q(u,x)}{\partial x_j} = 0$, $j = 1, ..., K$ we obtain:

$$RTS(x) = -\frac{TC_x \cdot u}{\varepsilon_u(u)}, \tag{25}$$

where $\varepsilon_u(u) = \frac{d \log q(u)}{d \log u}$. A similar analysis can be carried out in the case of multiple outputs but we do not pursue it further in the interest of space.

4 Model generalization and profit maximization

In this section we generalize the model so that it can provide directions in the input-output space that are compatible with profit maximization. Despite the fact that we do not specify a directional distance function (Atkinson and Tsionas, 2016 and Kumbhakar and Tsionas, 2016), we can determine directions from the data and indicate, both qualitatively and quantitatively, policy actions that must be taken by the banks to reduce waste of resources. The way to do this, without destroying the nice duality properties of the profit function, is to assume that technical inefficiency is a function of all inputs and outputs, viz.

$$u = u(x,y),\ u : \Re^K \times \Re^M \to (0,1]. \tag{26}$$

The profit maximization problem of the firm can then be expressed as follows:

$$\begin{aligned} \Pi(p,w,q) = \max_{x,y} :\ & p'y - w'x - q \cdot u(x,y) \\ \text{s.t. } & D(x,y) \leq 1, \end{aligned} \tag{27}$$

where $D(x,y)$ is the (output) distance function and $u(x,y)$ is an arbitrary inefficiency function depending on inputs and outputs. We wish to empahasize that there is a certain strand of the literature which has been heavily occupied with parametric specifications of inefficiency on exogenous (or "environmental factors" as they are called).

See Battese and Coelli (1995), Kumbhakar et.al (1991), Deprins and Simar (1989a,b), Reifschneider and Stevenson (1991), Simar et.al (1994), Sickles et.al (1986).

From duality we have the conditions:

$$\begin{array}{l} \frac{\partial \Pi(p,w,q)}{\partial p} = y^*(p,w,q), \\ -\frac{\partial \Pi(p,w,q)}{\partial w} = x^*(p,w,q), \\ -\frac{\partial \Pi(p,w,q)}{\partial q} = u^*(p,w,q). \end{array} \tag{28}$$

It is essential to observe that the new approach does not deliver results that are different from *(8)-(11)*. We obtain the same duality relationships without optimizing with respect to inefficiency. This is due to its dependence on inputs and outputs through (26). The great advantage of the new formulation in (28), that we are permitted to utilize these expressions to solve for u^* as a function of x^*, y^* and q. Specifically, we can solve (28) to obtain

$$\begin{array}{c} p = p(w,q,y^*) \\ w = w(p(w,q,y^*),q,x^*) \Rightarrow w = w(q,y^*,x^*) \end{array} . \tag{29}$$

From these relationships, the third equation in (28) can be written as

$$u^* = u^*(q,y^*,x^*). \tag{30}$$

We label **Model B** the model with (26) and **Model A** the model in (7) and (8)-(11). Estimation details, organized around Bayesian techniques using Markov Chain Monte Carlo (MCMC), are provided in the Technical Appendix.

5 Empirical application

5.1 Empirical results

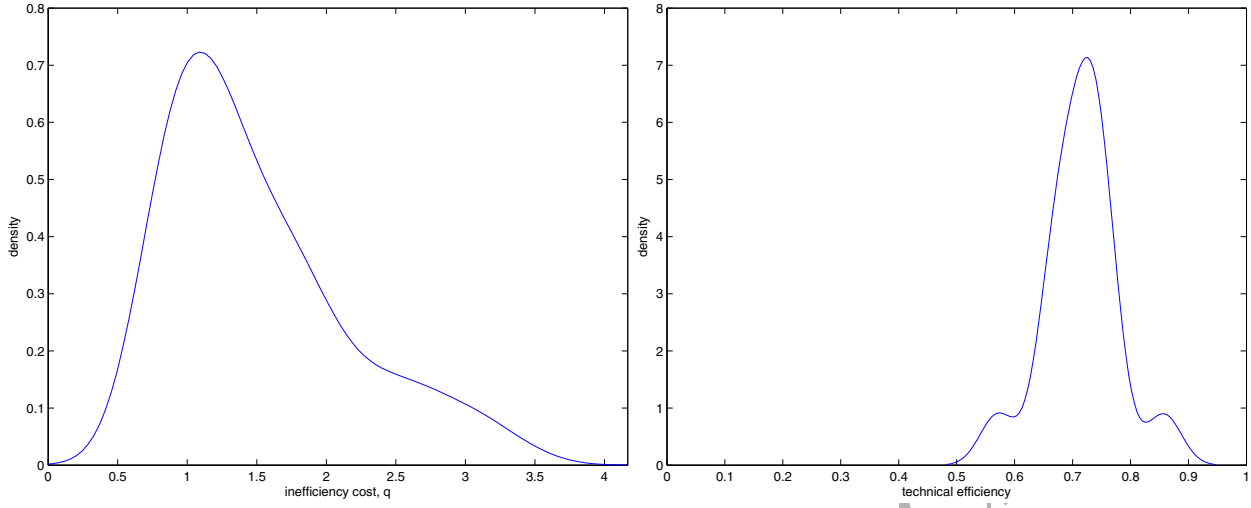
Our empirical results are reported in Table 2. Specifically, we report posterior means and posterior standard deviations produced from our MCMC procedure. To implement MCMC we use 15,000 draws, the first 5,000 of which are discarded to mitigate the possible impact of start up effects.

Table 2: Estimates of equation (26)

	posterior mean	posterior s.d.
y_1, consumer loans	-0.202	0.235
y_2, real estate loans	0.174	0.155
y_3, commercial & industrial loans	-0.223	0.135
y_4, securities	-0.128	0.103
y_5, off-balance-sheet income	-0.092	0.075
x_1, full-time equivalent employees	-0.171	0.122
x_2, physical capital	-0.104	0.098
x_3, purchased funds	0.081	0.072
x_4, interest-bearing transaction accounts	0.065	0.045
x_5, non-transaction accounts	0.033	0.029
inefficiency costs, q	0.832	0.014
Equity, e	-0.044	0.012

Our results show that purchased funds, interest-bearing transaction accounts and non-transaction accounts contribute negatively to optimal inefficiency. The most notable being the interest-bearing purchased funds (0.081),

Figure 2: Sample densities of inefficiency cost, q and technical efficiency, r_{it}



followed by transaction accounts (0.065), and non-transaction accounts (0.033). In terms of outputs, the cost of inefficiency and inefficiency itself can be reduced primarily by expanding commercial & industrial loans (-0.223), followed by consumer loans (-0.202), real estate loans (-0.174), securities (-0.128), and off-balance-sheet income (-0.092). These findings provide us with directions in the input-output space that can be reasonably adopted by policy makers to reduce inefficiency.

Sample densities of q and efficiency are reported in Figure 2. The sample distribution of q is skewed to the right, suggesting that for most banks the inefficiency cost is large. Technical efficiency ranges from 50% to 95%, averaging, approximately, 75%.

Figure 3 presents the relationship between efficiency and q, which is positive, as expected, and approximately linear. The results in Table 2 imply that relatively large reductions of inefficiency costs are required to make it cost-effective in order to improve efficiency, further implying that inefficiency is likely to be quite persistent. Specifically, a 10% reduction in inefficiency requires, *ceteris paribus*, a reduction of 8.32% in its cost.

Figure 3: Sample relationship between q and efficiency

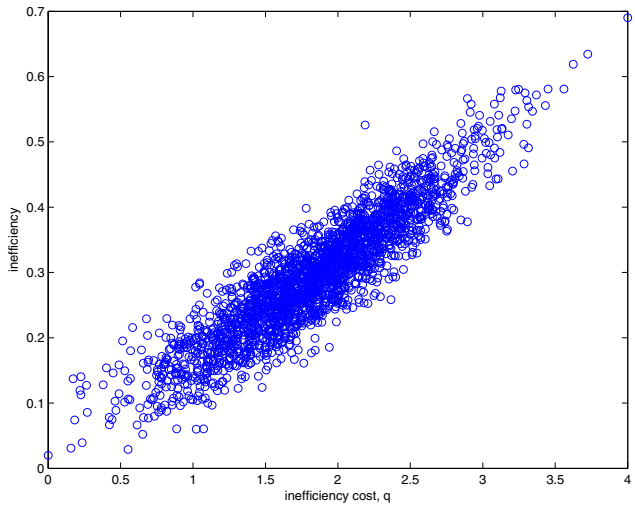
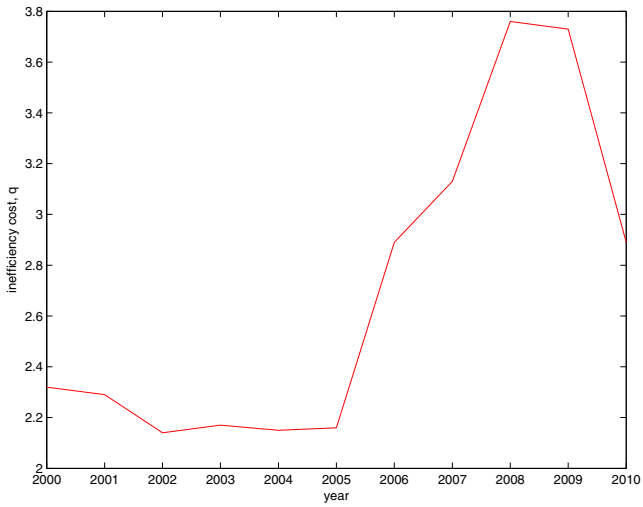


Figure 4: Evolution of average inefficiency cost over time



As an additional implication of the model we wish to examine how average inefficiency cost evolves over time. This is presented in Figure 4. Interestingly enough, q increases sharply after 2006 and continues to rise through to 2009; that is, well before and also during the sub-prime crisis. The explanation is that q depends on all inputs and outputs which have been misallocated in the banking sector as the result of the forces that lead to the financial crisis of 2008-2009. This can also be explained by the positive contribution of real estate loans to q providing further support to the misallocation hypothesis.

5.2 Policy measures

We now turn to policy suggestions for reducing inefficiency in the banking sector. We consider several scenarios, where our analysis is performed at the medians of the variables. The results are reported in Table 3 and are obtained as follows. Suppose we examine the first scenario where all outputs are simultaneously expanded by 10%. Since we have an MCMC sample $\left\{\boldsymbol{\theta}^{(s)} s = 1, ..., S\right\}$ (where S =20,000).

Let the median vector of outputs be given by $\tilde{\mathbf{y}}$. The vector is disturbed by 10% and becomes $\tilde{\mathbf{y}}_{new} = \tilde{\mathbf{y}} + \triangle\tilde{\mathbf{y}}$, where $\triangle\tilde{\mathbf{y}} = h \cdot \tilde{\mathbf{y}}$, where $h = 0.1$. With inefficiency evaluated at $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{y}}_{new}$ using equation (17) we respectively obtain, say $\tilde{u}$ and $\tilde{u}_{new}$ respectively. The effect on inefficiency, in percentage terms, is: $\frac{\tilde{u}_{new} - \tilde{u}}{\tilde{u}}$. As this computation is performed for each MCMC, we can easily compute the sample average and sample posterior standard deviations of the inefficiency effect.

Table 3: Policy actions to affect inefficiency

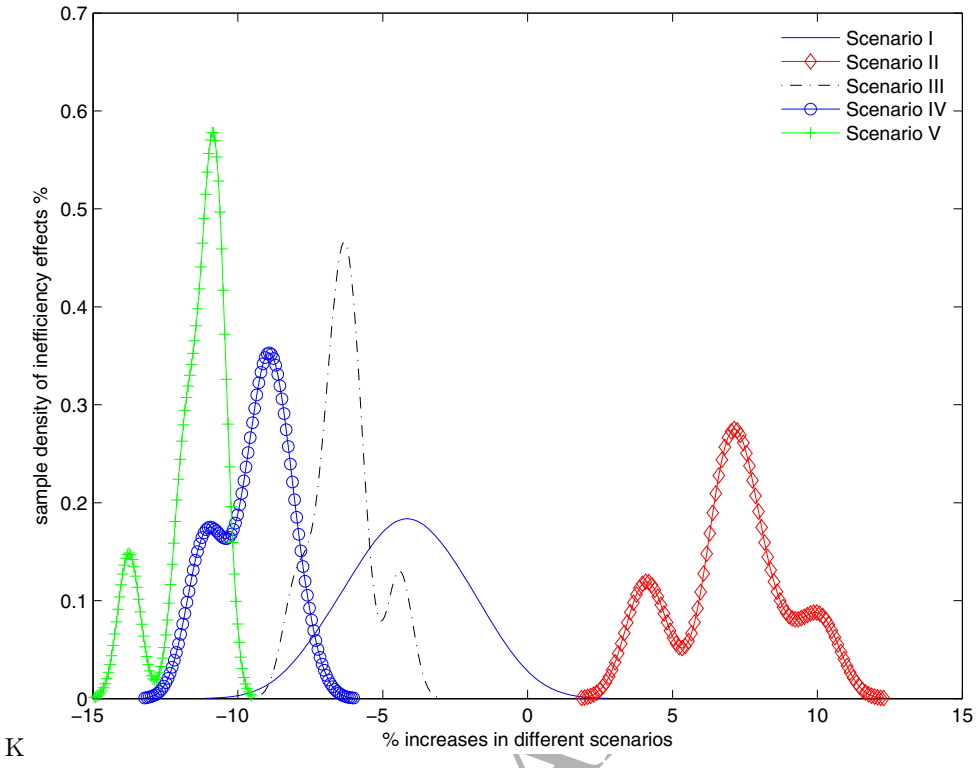
Scenario		sample post. mean inefficiency	change in optimal profit
I.	Expand all outputs by 10%	-4.55% (1.32%)	-0.017% (0.0044%)
II.	Expand all inputs by 10%	7.45% (2.24%)	-0.023% (0.0032%)
III.	Reduce q and increase e by 10%	-5.52% (0.48%)	7.155% (1.47%)
IV.	Expand all outputs except real estate by 10%	-9.18% (0.76%)	-0.0082% (0.0012%)
V.	Expand all inputs except labor and physical capital by 10%	-11.71% (1.25%)	-0.012% (0.0037%)

Notes: Sample standard deviations are reported in parentheses

The results in Table 3 indicate that the banking sector can substantially reduce inefficiencies. For example the most effective measure would be the expansion of all inputs (with the exception of labor and physical capital) by 10% (effect -11.71%). Close behind would be the expansion of all outputs (except real estate) by 10% (effect -9.18%). However, the expansion of all inputs by 10% would be deleterious, since inefficiency would rise by 7.45%.

The relevance of these results is not limited to the "median" banking firm: for example, when we reduce q and increase e by 10%, the median effect is -5.52%. Taking into account a sample posterior standard deviation of 0.48%, we would expect banks in general to experience reductions in inefficiency of between -4.56% and -6.48%. In emphasising that we are no longer taking the median bank as our reference, an expansion of all inputs by 10% would increase inefficiency from between 2.97% and 11.93%. These results, of course, rely on asymptotic normality of the effects. A clearer picture emerges if we present the full marginal posterior distributions of the five effects corresponding to the different scenarios in Table 3. These are reported in Figure 5, for all banks across all parameter draws. With the one exception of Scenario III (where profits increase by 7.1%) the third column of Table 3 indicates how profit falls with a change in the optimal decision of the banks. The significance of a 0.01% profit loss of (*i.e.*, \$100,000 per \$1 billion) is difficult to judge given that changes in efficiency range from 4.5% to 11.7%. Our 'success' is in being able to attribute gains and losses against scenarios that policy-makers and decision-makers can examine.

Figure 5: Sample marginal posteriors for all observations and parameter draws



K

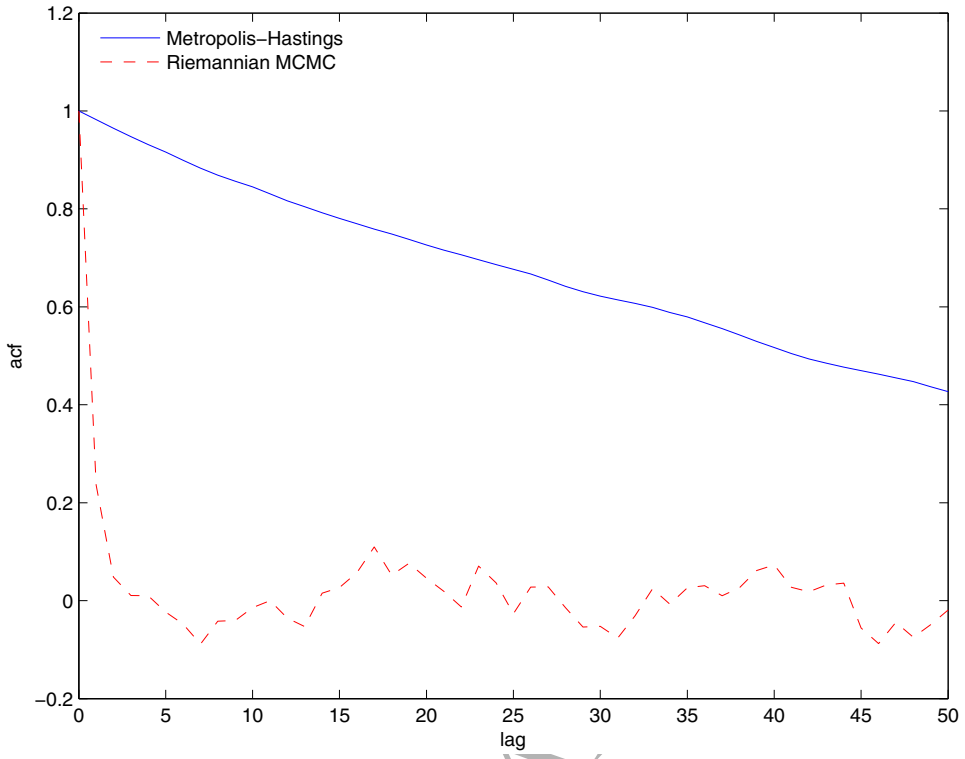
It also turns out that exact marginal posteriors across all bank observations and MCMC parameter draws are far from normality with the exception of scenario I. This picture provides a more accurate description of the inefficiency effects arising from the different scenarios, although the normality-based "rule of thumb" (mean plus and minus two standard deviations) is simple enough to provide a first-approximation advice to policy makers and banking authorities.

5.3 Sensitivity analysis and other technical remarks

Given the use of the MCMC, it is essential to assess convergence and examine autocorrelation of MCMC draws. Convergence is assessed using the Geweke (1992) convergence diagnostic. Specifically, for each parameter, we test for equality of means in the first and last 25% of the draws. Geweke's (1992) statistic converges asymptotically (in the number of draws) to a standard normal distribution. In our application, these statistics ranged from 0.220 to 1.344 in absolute value, suggesting that MCMC has converged. To examine autocorrelation, we compute autocorrelation functions (acf) for each parameter based on the MCMC draws. In Figure 6, we report the *maximal* values of autocorrelation coefficients at each lag from 1 to 50 (in absolute value but retaining the sign for plotting). For comparison, we also plot the *maximal* values of autocorrelation coefficients at each lag from a Metropolis-Hastings MCMC scheme: The Metropolis-Hastings MCMC scheme generates a candidate draw as $\boldsymbol{\theta}^c \sim \mathcal{N}\left(\boldsymbol{\theta}^{(s-1)}, hI\right)$, where $\boldsymbol{\theta}^{(s-1)}$ is the previous draw and $h > 0$ is a smoothing constant. The candidate is accepted with probability $\min\left\{1, \frac{p(\boldsymbol{\theta}^c|\mathcal{Y})}{p(\boldsymbol{\theta}^{(s-1)}|\mathcal{Y})}\right\}$ and we set $\theta^{(s)} = \theta^c$, otherwise we repeat the previous draw and we set $\boldsymbol{\theta}^{(s)} = \boldsymbol{\theta}^{(s-1)}$, $\forall s = 1, ..., S$. We use the same number of draws, S, and we select h by trial-and-error so that approximately 25% of all candidates are, eventually, accepted.

The autocorrelation functions in Figure 6 shows that the performance of Riemannian MCMC is much better than the Metropolis-Hastings MCMC. The autocorrelations from the Riemannian MCMC are, practically, zero after

Figure 6: Autocorrelation functions



about lag 5 whereas autocorrelations from Metropolis-Hastings MCMC remains close to 0.5 at lag 50.

Another important technical question is the sensitivity of the results to the number of draws. To address this, we increase the number of draws to 25,000, 35,000,...,125,000. We omit the first 5,000 draws to mitigate possible start up effects and re-compute the posterior means and posterior standard deviations of the parameters. Figure 7 shows the percentage deviations relative to the case with 15,000 draws and it can be seen that increasing the number of draws leaves the results unchanged.

The prior is flat over the region described by the constraints in (18), see also (A.3). As such, we do not impose any particular prior information other than the imposition of economic constraints. It would be interesting, however, to adopt an informative prior of the form:

$$\boldsymbol{\theta} \sim \mathcal{N}\left(\underline{\boldsymbol{\theta}}, \underline{\mathbf{V}}\right), \tag{31}$$

where $\underline{\boldsymbol{\theta}}$, $\underline{\mathbf{V}}$ are, respectively, the prior mean and prior covariance matrix. This prior is truncated to the set in (18) to account for the economic restrictions. To examine sensitivity with respect to the prior, we set $\underline{\boldsymbol{\theta}} = \varphi I$ and $\underline{\mathbf{V}} = \omega^2 I$. We generate 10,000 different φ and ω values from uniform distributions in the interval $\left(-10^6, 10^6\right)$ and $\left(1, 10^6\right)$ respectively. In turn, we re-run our Riemannian MCMC scheme with 15,000 draws the first 5,000 are discarded to mitigate start up effects. We compute the percentage deviations of posterior means and posterior standard deviations relative to the baseline case which corresponds to a flat prior over (18). The densities are computed using all parameter draws and all the 10,000 different φ and ω values relative to the baseline case.

Figure 7: Sensitivity to number of draws

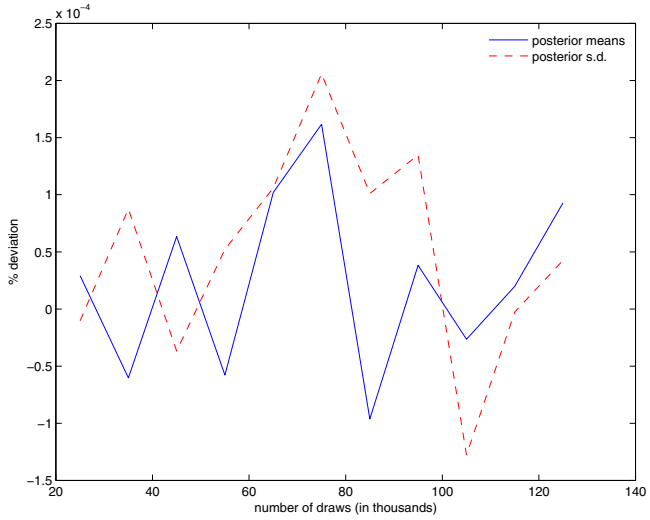
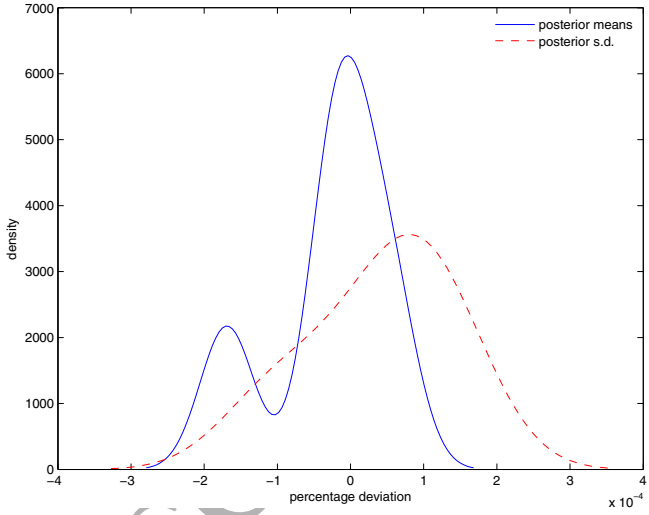


Figure 8: Senstivity to prior



The densities displayed in Figure 8 show that the results to be robust with respect to large changes in the prior. This provides some assurance that the results corresponding to the different scenarios are not artificial as they do not depend on a specific prior.

6 Concluding remarks

In this paper we have proposed to incorporate explicitly the cost of technical inefficiency in a maximization profit maximization model recognizing that improvements in efficiency require real resources and, therefore, certain costs. This point has been ignored in the literature. Bogetoft and Hougaard (2003) have proposed the "rational inefficiency hypothesis" but the empirical implications of this hypothesis, unfortunately, are rejected in our U.S banking application making it essential to provide an alternative hypothesis.

We have used an output distance function with multiple outputs and multiple inputs and duality of the profit function to derive input demand functions, output supply functions but more importantly inefficiency as a function of input and output prices and the cost of technical inefficiency. The resulting system of equations, which include the unobserved cost of inefficiency is estimated using Bayesian techniques based on MCMC, in specific the Girolami and Calderhead (2011) Riemannian Manifold Hamiltonian technique. We show that the model is equivalent to a model where inefficiency is an arbitrary function of inputs, outputs and the inefficiency cost, so that we can determine optimal directions in the input-output space to take policy actions to reduce inefficiency.We have applied the new techniques to an unbalanced panel of U.S banks for the period 2000-2010. The empirical results suggest that a 10% decrease of inefficiency requires a 9.59% decrease of its cost. This implies that relatively large reductions of inefficiency costs are required in order to improve efficiency, implying further that inefficiency is likely to be quite persistent. Additionally, inefficiency cost increases sharply after 2006 and increases through to 2009, that is well before and also during the sub-prime crisis. The explanation is, apparently, the positive contribution of real estate loans to inefficiency cost, providing further support to the misallocation hypothesis. Our results also provide us with directions in the input-output space that can be reasonably used in other studies or banking policy makers with a purpose towards reducing inefficiency.

As the model provides directions in the input-output space, it can be used to address more complex problems where competing approaches have been suggested as, for example, in Atkinson and Tsionas (2016) where such

directions are estimated using Bayesian techniques rather than derived directly from the model. The two approaches can be compared and contrasted to decide which model is best in the light of the data, using the concepts of marginal likelihood and Bayes factors. Simar et. al (2016) use nonparametric conditional efficiencies and propose a model where the heterogeneity variable is linked to a particular input or output. In our context we allow for a full set of environmental variables that determine the heterogeneity. Apparently, there is room for improvement in this context by introducing semiparametric components into the model and examining further its robustness, particularly when it comes to the examination of results corresponding to different policy scenarios.

References

Aigner, D., C.A.K. Lovell and P. Schmidt (1977). Formulation and estimation of stochastic frontier production function models. Journal of Econometrics 6 (1), 21-37.

Annaert, J., J. van den Broeck, and R.V. Vennet (2003). Determinants of mutual fund underperformance: A Bayesian stochastic frontier approach. European Journal of Operational Research 151 (3), 617-632.

Asmild, M., P. Bogetoft, J. L. Hougaard (2013). Rationalising inefficiency: Staff utilisation in branches of a large Canadian bank, Omega 41, 80-87.

Asmild, M., and K. Matthews (2012). Multi-directional efficiency analysis of efficiency patterns in Chinese banks 1997–2008. European Journal of Operational Research 219 (2), 434-441.

Atkinson, S.E., and M.G. Tsionas (2016). Directional distance functions: Optimal endogenous directions. Journal of Econometrics 190 (2), 301-314.

Badin L, Daraio C, Simar L (2012) How to measure the impact of environmental factors in a nonparametric production model. European Journal of Operational Research 223(3):818-833.

Badunenko, O., and S.C. Kumbhakar (2017). Economies of scale, technical change and persistent and time-varying cost efficiency in Indian banking: Do ownership, regulation and heterogeneity matter? European Journal of Operational Research, in press.

Balezentis, T., and C. DeWitte (2015). One- and multi-directional conditional efficiency measurement – Efficiency in Lithuanian family farms. European Journal of Operational Research 245 (2), 612-622.

Battese, G. and T. Coelli (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. Empirical Economics 20, 325-332.

Berger, A. N. and Mester, L. J. (1997). Inside the black box: What explains differences in the efficiencies of financial institutions? Journal of Banking & Finance, 21(7):895–947.

Berger, A. N. and Mester, L. J. (2003). Explaining the dramatic changes in performance of US banks: Technological change, deregulation, and dynamic changes in competition. Journal of Financial Intermediation 12(1):57–95.

Biener, C., M. Eling, and J.K. Wifs (2016). The determinants of efficiency and productivity in the Swiss insurance industry. European Journal of Operational Research 248 (2), 703-714.

Bogetoft, P. and J.L. Hougaard (2003). Rational inefficiencies, Journal of Productivity Analysis 20, 243-271.

Chen, Y., Y. Li, L. Liang, A. Salo, and H. Wu (2016). Frontier projection and efficiency decomposition in two-stage processes with slacks-based measures. European Journal of Operational Research 250 (2), 543-554.

Cornwell, C., P. Schmidt and R.C. Sickles (1990). Production frontiers with cross-sectional and time-series variation in efficiency levels. Journal of Econometrics 46 (1-2), 185-200.

Daraio C, Simar L (2014) Directional distances and their robust versions: Computational and testing issues. Eur J Oper Res, 237(1), 358-369.

Deprins, S., and L. Simar (1989a). Estimation de frontieres deterministes avec facteurs exogenes d' Inefficite. Annales d' Ecomie et de Statistique 14, 117-150.

Deprins, S., and L. Simar (1989b). Estimating technical inefficiencies with corrections for environmental conditions with an application to railway companies. Annals of Public and Cooperative Economics 60 (1), 81-102.

Feng, G. and Serletis, A. (2009). Efficiency and productivity of the US banking industry, 1998–2005: Evidence from the Fourier cost function satisfying global regularity conditions. Journal of Applied Econometrics, 24(1):105–138.

Girolami, M., and B. Calderhead. (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73(2), 123–214.

Geweke, J. (1992), Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In Bayesian Statistics 4 (eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith), Oxford: Oxford University Press, 169-193.

Hughes, J. P. and Mester, L. J. (2010). Efficiency in banking: Theory and evidence. In Berger, A., Molyneux, P., and Wilson, J., editors, Oxford Handbook of Banking. Oxford University Press, Oxford, 1st edition.

Hughes, J. P. and Mester, L. J. (2013). Who said large banks don't experience scale economies? Evidence from a risk-return-driven cost function. Journal of Financial Intermediation, 22(4):559–585.

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

Kumbhakar, S.C. and M.G. Tsionas (2006). Estimation of stochastic frontier production functions with input-oriented technical efficiency. Journal of Econometrics 133 (1), 71-96.

Kutlu, L., and R.C. Sickles (2012). Estimation of market power in the presence of firm level inefficiencies. Journal of Econometrics 168 (1), 141-155.

Kumbhakar, S.C., S, Ghosh and J.T.McGuckin (1991).A generalized production frontier approach for estimating determinants of inefficiency in US dairy farms. Journal of Business and Economic Statistics 9 (3), 279-286.

Kumbhakar, S.C. and M.G. Tsionas (2016). The good, the bad and the technology: Endogeneity in environmental production models. Journal of Econometrics 190 (2), 315-327.

Malikov, E., S.C. Kumbhakar, M.G. Tsionas (2016). A Cost System Approach to the Stochastic Directional Technology Distance Function with Undesirable Outputs: The Case of us Banks in 2001–2010. Journal of Applied Econometrics 31 (7), 1407-1429.

Reifschneider, D., and R.S. Stevenson (1991). Systematic departures from the frontier: A framework for the analysis of firm inefficiency. International Economic Review 32(3), 715-723.

Reyes, P.M., S. Li., and J.K. Visich (2016). Determinants of RFID adoption stage and perceived benefits. European Journal of Operational Research 254 (3), 801-812.

Saranga, S. (2009). The Indian auto component industry – Estimation of operational efficiency and its determinants using DEA. European Journal of Operational Research 196 (2), 707-718.

Sealey, C. W. and Lindley, J. T. (1977). Inputs, outputs, and a theory of production and cost at depository financial institutions. Journal of Finance, 32(4):1251–1266.

Sena, V. (2006). The determinants of firms' performance: Can finance constraints improve technical efficiency? European Journal of Operational Research 172 (1), 311-325.

Sickles, R.C. (2005). Panel estimators and the identification of firm-specific efficiency levels in parametric, semi-parametric and nonparametric settings. Journal of Econometrics 126 (2), 305-334.

Sickles, R.C., D. Good, and R.L Johnson (1986). Allocative distortions and the regulatory transition of the U.S. Airlines industry. Journal of Econometrics 33, 143-163.

Simar, L., C.A.K. Lovell and P. vanden Eeckhaut (1994). Stochastic frontiers incorporating exogenous influences on efficiency. Discussion Paper 9403, Institute de Statistique, Universite Catholique de Louvain, Louvain-la-Neuve, Belgium.

Simar, L., P.W. Wilson (2011). Two-stage DEA: caveat emptor. Journal of Productivity Analysis 36, 205-218.

Simar, L., A. Vanhems and I. Van Keilegom (2016). Unobserved heterogeneity and endogeneity in nonparametric frontier estimation. Journal of Econometrics 190 (2), 360-373.

Sun, K., S.C. Kumbhakar, and R. Tveteras (2015). Productivity and efficiency estimation: A semiparametric stochastic cost frontier approach. European Journal of Operational Research 245 (1), 194-202.

Tecles, P.L., and B.M. Tabak (2010). Determinants of bank efficiency: The case of Brazil. European Journal of Operational Research 207 (3), 1587-1598.

Tsionas, M.G. (2012). Maximum likelihood estimation of stochastic frontier models by the Fourier transform. Journal of Econometrics 170 (1), 234-248.

Tsionas, M.G., and E. Mamatzakis (2017). Adjustment costs in the technical efficiency: An application to global banking. European Journal of Operational Research 256 (2), 640-649.

TECHNICAL APPENDIX. Inference and Markov Chain Monte Carlo (MCMC)

The likelihood function is the following:

$$\mathcal{L}(\boldsymbol{\theta}, \Sigma; \mathcal{Y}) = (2\pi)^{-NT(M+K+1)/2} |\Sigma|^{-NT/2} e^{\sum_{i=1}^{n} \sum_{t=1}^{T} V_{it}(\theta,\mathcal{Y})' \Sigma^{-1} V_{it}(\theta,\mathcal{Y})}, \tag{A.1}$$

where $V_{it}(\boldsymbol{\theta}, \mathcal{Y})$ denotes the errors in (16). It is possible to integrate out Σ and obtain:

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{Y}) \propto |\sum_{i=1}^{n} \sum_{t=1}^{T} V_{it}(\boldsymbol{\theta}, \mathcal{Y}) V(\boldsymbol{\theta}, \mathcal{Y})'|^{-NT/2} \equiv \mathcal{A}(\boldsymbol{\theta}, \mathcal{Y})^{-NT/2}, \tag{A.2}$$

where $\mathcal{A}(\boldsymbol{\theta}, \mathcal{Y}) = \sum_{i=1}^{n} \sum_{t=1}^{T} V_{it}(\boldsymbol{\theta}, \mathcal{Y}) V_{it}(\boldsymbol{\theta}, \mathcal{Y})'$.

The major problem in implementing ML is that we have to take account of the restrictions in (18). These are restrictions that we must impose for each observation in the sample, otherwise inefficiency estimates will not be meaningful as they are not produced from a proper profit function. Therefore, we opt for a Bayesian approach using a prior which is flat over the region describing the restrictions, viz.:

$$p(\boldsymbol{\theta}) \propto \mathbb{I}(\mathcal{M}(\boldsymbol{\theta}, \mathcal{Y}) \leq O), \tag{A.3}$$

where $\mathbb{I}()$ is the indicator function. Using Bayes' theorem we combine (A.2) and (A.3) and we have the posterior:

$$p(\boldsymbol{\theta}|\mathcal{Y}) \propto \mathcal{A}(\boldsymbol{\theta}, \mathcal{Y})^{-NT/2} \cdot \mathbb{I}(\mathcal{M}(\boldsymbol{\theta}, \mathcal{Y}) \leq O). \tag{A.4}$$

To explore the posterior we use Markov Chain Monte Carlo (MCMC) techniques. Specifically we use the Girolami and Calderhead (2011) Riemannian manifold Hamiltonian MCMC. The algorithm uses local information about both the gradient and the Hessian of the log-posterior conditional of $\boldsymbol{\theta}$ at the existing draw. Suppose $\mathscr{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta}|\boldsymbol{D})$ is used to denote for simplicity the log posterior of $\boldsymbol{\theta}$. Moreover, define:

$$\boldsymbol{G}(\boldsymbol{\theta}) = \text{est.cov} \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{D}|\boldsymbol{\theta}), \tag{A.5}$$

the empirical counterpart of

$$\boldsymbol{G}_o(\boldsymbol{\theta}) = -E_{\boldsymbol{D}|\boldsymbol{\theta}} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log p(\boldsymbol{D}|\boldsymbol{\theta}). \tag{A.6}$$

The Langevin diffusion is given by the following stochastic differential equation:

$$d\boldsymbol{\theta}(t) = \tfrac{1}{2} \tilde{\nabla}_{\boldsymbol{\theta}} \mathscr{L}\{\boldsymbol{\theta}(t)\} dt + d\mathbf{B}(t), \tag{A.7}$$

where

$$\tilde{\nabla}_{\boldsymbol{\theta}} \mathscr{L}\{\boldsymbol{\theta}(t)\} = -\boldsymbol{G}^{-1}\{\boldsymbol{\theta}(t)\} \cdot \tilde{\nabla}_{\boldsymbol{\theta}} \mathscr{L}\{\boldsymbol{\theta}(t)\}, \tag{A.8}$$

is the so called "natural gradient" of the Riemann manifold generated by the log posterior. The elements of the Brownian motion are

$$\begin{aligned}\boldsymbol{G}^{-1}\left\{\boldsymbol{\theta}\left(t\right)\right\}d\mathbf{B}_i\left(t\right) =&\left|\boldsymbol{G}\left\{\boldsymbol{\theta}\left(t\right)\right\}\right|^{-1/2}\sum_{j=1}^{K_\beta}\frac{\partial}{\partial\boldsymbol{\theta}}\left[\boldsymbol{G}^{-1}\left\{\boldsymbol{\theta}\left(t\right)\right\}_{ij}\left|\boldsymbol{G}\left\{\boldsymbol{\theta}\left(t\right)\right\}\right|^{1/2}\right]dt \\ &+\left[\sqrt{\boldsymbol{G}\left\{\boldsymbol{\theta}\left(t\right)\right\}}d\boldsymbol{B}\left(t\right)\right]_i\end{aligned} \tag{A.9}$$

The discrete form of the stochastic differential equation provides a proposal as follows:

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_i =&\boldsymbol{\theta}_i^o+\tfrac{\varepsilon^2}{2}\left\{\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^o\right)\nabla_{\boldsymbol{\theta}}\mathscr{L}\left(\boldsymbol{\theta}^o\right)\right\}_i-\varepsilon^2\sum_{j=1}^{K_\theta}\left\{\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^o\right)\frac{\partial\boldsymbol{G}\left(\boldsymbol{\theta}^o\right)}{\partial\boldsymbol{\theta}_j}\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^o\right)\right\}_{ij} \\ &+\tfrac{\varepsilon^2}{2}\sum_{j=1}^{K_\theta}\left\{\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^o\right)\right\}_{ij}\operatorname{tr}\left\{\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^o\right)\frac{\partial\boldsymbol{G}\left(\boldsymbol{\theta}^o\right)}{\partial\boldsymbol{\theta}_j}\right\}+\left\{\varepsilon\sqrt{\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^o\right)}\boldsymbol{\xi}^o\right\}_i \\ =&\boldsymbol{\mu}\left(\boldsymbol{\theta}^o,\varepsilon\right)_i+\left\{\varepsilon\sqrt{\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^o\right)}\boldsymbol{\xi}^o\right\}_i,\end{aligned} \tag{A.10}$$

where $\boldsymbol{\theta}^o$ is the current draw. The proposal density is:

$$q\left(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^o\right)=\mathcal{N}_{K_\theta}\left(\tilde{\boldsymbol{\theta}},\varepsilon^2\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^o\right)\right), \tag{A.11}$$

and convergence to the invariant distribution is ensured by using the standard form Metropolis-Hastings probability:

$$\min\left\{1,\frac{p\left(\tilde{\boldsymbol{\theta}}|\cdot,\boldsymbol{D}\right)q\left(\boldsymbol{\theta}^o|\tilde{\boldsymbol{\theta}}\right)}{p\left(\boldsymbol{\theta}^o|\cdot,\boldsymbol{D}\right)q\left(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^o\right)}\right\}. \tag{A.12}$$

The MCMC procedure is implemented using 15,000 iterations the first 5,000 are omitted to mitigate start-up effects. For all parameters the prior is uniform over the interval $(-M,M)^{\dim(\boldsymbol{\theta})}$ where $M=10^7$ and $\dim(\boldsymbol{\theta})$ is the dimensionality of the parameter vector. Any parameter draw that does not satisfy the monotonicity restrictions in (18) is rejected and another candidate is considered.

To facilitate imposition of the restrictions in (18) we proceed as follows: First, we impose the restrictions at the mean of the data in (16), and denote the mean by $\bar{\mathbf{z}}$. Second, the restrictions are imposed at 100 random points in the set $\bar{\mathbf{z}}+3\mathbf{s}$, where $\mathbf{s}$ is the standard deviation of the data. Finally, we check whether the restrictions hold at every other point in the data set. If the restrictions are not satisfied then we consider another candidate for the parameter vector $\boldsymbol{\theta}$. This procedure, improves vastly over a naive GC algorithm that simply accepts or rejects through the entire set in (18).