A Cost System Approach to the Stochastic Directional Technology Distance Function with Undesirable Outputs: The Case of U.S. Banks in 2001–2010*

Emir Malikov[1] Subal C. Kumbhakar[2] Efthymios G. Tsionas[3]

[1]Department of Economics, St. Lawrence University, Canton, NY, USA
[2]Department of Economics, State University of New York at Binghamton, Binghamton, NY, USA
[3]Department of Economics, Lancaster University Management School, Lancaster, UK

First Draft: September 30, 2014
This Draft: September 5, 2015

Abstract

This paper offers a methodology to address the endogeneity of inputs in the directional technology distance function (DTDF) based formulation of banking technology which explicitly accommodates the presence of *undesirable* nonperforming loans — an inherent characteristic of the bank's production due to its exposure to credit risk. Specifically, we model nonperforming loans as an undesirable output in the bank's production process. Since the stochastic DTDF describing banking technology is likely to suffer from the endogeneity of inputs, we propose addressing this problem by considering a system consisting of the DTDF and the first-order conditions from the bank's cost minimization problem. The first-order conditions also allow us to identify the "cost-optimal" directional vector for the banking DTDF, thus eliminating the uncertainty associated with an *ad hoc* choice of the direction. We apply our cost system approach to the data on large U.S. commercial banks for the 2001–2010 period, which we estimate via Bayesian MCMC methods subject to theoretical regularity conditions. We document dramatic distortions in banks' efficiency, productivity growth and scale elasticity estimates when the endogeneity of inputs is assumed away and/or the DTDF is fitted in an arbitrary direction.

Keywords: Bad Outputs, Commercial Banks, Directional Distance Function, Endogeneity, MCMC, Nonperforming Loans, Productivity, Technical Change

JEL Classification: C11, C33, D24, G21

*We would like to thank the editor, Edward Vytlacil, and anonymous referees for many insightful comments and suggestions that helped improve this article. Any remaining errors are our own.

Email: emalikov@stlawu.edu (Malikov), kkar@binghamton.edu (Kumbhakar), tsionas@aueb.gr (Tsionas).

1 Introduction

One can hardly overstate the role of the banking sector in the economy. Given the importance of financial intermediation in facilitating economic activity, in general, and the transformation that the banking industry has been undergoing in recent decades, a large number of studies have analyzed banks' performance in an attempt to quantify their productivity and efficiency (see Hughes and Mester, 2010, for an excellent review). The interest in quantifying banking technology has been particularly fueled by the wave of regulatory changes[1] as well as natural technological and financial innovations which have spurred the consolidation of the industry (e.g., see Berger et al., 1999; Berger and Mester, 2003; Berger, 2004). The recent financial crisis of 2007 has further focused researchers' attention on banks' production technology in the pursuit of the evidence of scale economies and high efficiency which may provide arguments against the "too-big-to-fail" criticism of large-sized banks (see Hughes and Mester, 2013; Restrepo-Tobón and Kumbhakar, 2015, and references therein).

In this paper, we address some econometric issues related to a consistent estimation of banking technology which explicitly accommodates the presence of *undesirable* nonperforming loans — an inherent characteristic of the bank's production due to its exposure to uncertainty. Textbooks explain that commercial banks are subject to the credit risk associated with the likelihood that a borrower will default on the debt by failing to make payments as obligated contractually (e.g., Freixas and Rochet, 2008). This credit risk materializes in the form of "nonperforming loans", i.e., loans that are not paid back duly. Researchers studying banks' productivity and efficiency have long ago acknowledged the importance of taking such nonperforming loans (NPL) into account when estimating banking technology. The two approaches to modeling bank's nonperforming loans prevailing in the literature have been to treat the former as either *(i)* a "control variable" capturing bank's risk and/or quality of loans (e.g., Berger and Mester, 1997, 2003; Altunbas et al., 2000; Koutsomanoli-Filippaki et al., 2009) or *(ii)* an "undesirable output" (e.g., Park and Weber, 2006; Assaf et al., 2013; Guarda et al., 2013).

Due to its apparent advantages, we focus on the second approach to modeling nonperforming loans. Treating NPL as an undesirable output is advantageous over modeling it as a mere banking technology shifter (i.e., a contextual control variable) because not only does this approach recognize that NPL is a *by*-product of producing desirable outputs, such as earning loans and securities, but it also accommodates the *undesirability* of NPL. That is, by acknowledging that nonperforming loans are undesirable, we can credit banks for the reduction in NPL along with the expansion in desirable outputs when computing their productivity and technological efficiency.

To estimate the banking technology in the presence of nonperforming loans, we use the directional technology distance function (DTDF) of Chambers et al. (1998) generalized to the case of undesirable outputs (also see Chung et al., 1997; Färe et al., 2005; Hudgins and Primont, 2007). The DTDF can be estimated parametrically in two ways: via a linear programming technique, if treating it as a deterministic function, or via a stochastic frontier, if allowing the DTDF to be subject to a random error. In this paper, we consider the latter approach due to its additional flexibility and an increasing popularity among productivity studies seeking to estimate directional distance functions (e.g., see Guarda et al., 2013; Atkinson et al., 2014; Feng and Serletis, 2014; Tsionas et al., 2014).

The estimation of the stochastic DTDF is however not trivial because of the potential endogeneity problem. While in many empirical applications (e.g., studies of service industries such as commercial banking) one can justify the exogeneity of outputs (both desirable and undesirable) with relative ease (e.g., Feng and Serletis, 2009; Hughes and Mester, 2010; Assaf et al., 2013; Malikov et al., 2015b), the exogeneity of inputs is however a much harder sell. Economists commonly agree

[1] E.g., permitting interstate branching or merging commercial banking with security trading and/or insurance, etc.

that *input* allocation is an outcome of *endogenous* choice made by bank managers.

In this paper, we propose addressing the endogeneity of inputs in the stochastic DTDF formulation of the banking technology from the perspective of economic theory. More specifically, we suggest invoking the assumption of the bank's cost minimizing behavior not only to justify the treatment of outputs as exogenous[2] but to also tackle the endogeneity of inputs in the DTDF. We do so by augmenting the stochastic DTDF with the set of independent (nonlinear) first-order conditions derived from the bank's cost minimization problem, which we then estimate as a system of simultaneous equations. We prefer the assumption of cost-minimizing behavior over profit-maximization primarily because it is consistent with the premise of exogenously determined outputs, which is common to the banking industry, and it does not require price information on undesirable outputs which is hard to measure and is rarely available in practice. Our identification strategy thus relies on competitively determined input prices as a source of exogenous variation.

Incidentally, Atkinson et al. (2014) and Atkinson and Tsionas (2015) have recently proposed tackling the endogeneity in the DTDF by formulating a system of equations based on the assumption of profit-maximizing behavior. While being a valid alternative to our estimator, the inconvenience of such an approach is that, due to the unobservability of prices of undesirable outputs, a researcher is forced to augment a system of *quasi*-profit-maximizing first-order conditions[3] with the reduced-form demand equations for undesirable outputs. Our *cost* system approach is however not subject to the above problem. Further, endogeneity in the estimation of the stochastic DTDF is also discussed, although in a more narrow sense, in Guarda et al. (2013). In their paper, the authors are primarily concerned with the endogeneity problem due to the appearance of the left-hand-side "dependent variable" on the right-hand side of the normalized DTDF regression equation, which they suggest to remedy by choosing the directional vector so that the dependent variable disappears from the right-hand side of the equation.[4] Guarda et al. (2013) however leave the endogeneity of inputs due to *simultaneity*, which is a primary focus of our paper, unaddressed.

We note that there are advantages to using our cost system approach even if the DTDF does not suffer from the endogeneity of inputs. Since additional equations (the first-order conditions) do not contain any extra parameters, the system-based parameter estimates are likely to be more precise. Furthermore, technological metrics obtained from the cost system of DTDF are likely to be more meaningful because the economic behavior is embedded into the system through the first-order conditions. The inclusion of the cost-minimizing first-order conditions in the system also permits us to estimate the "cost-optimal" directional vector for the banking DTDF. That is, in contrast to a common tradition in the literature, we do not pre-specify the (fixed) directional vector for the DTDF but rather employ Färe et al.'s (2013) idea and let the data help us determine the direction in which the bank's movement toward the stochastic frontier is to be estimated.[5] This approach eliminates the uncertainty associated with the fact that the DTDF is an implicit function of the direction, which implies that the DTDF-based estimates of banking technology change with the choice of the directional vector.

We apply our cost system approach to the data on large U.S. commercial banks for the 2001–2010 period. The nonlinear cost system of equations is estimated via Bayesian MCMC methods subject to monotonicity and curvature regularity conditions in order to ensure that our results are economically meaningful, as emphasized by Barnett et al. (1991) and Barnett (2002). The reported

[2] See Malikov et al. (2015b) and the references therein.

[3] The prefix "quasi" indicates that the employed profit-maximizing first-order conditions *omit* the terms containing the prices of undesirable outputs.

[4] Tsionas et al. (2014) propose a more general solution to the same problem, which is based on accounting for a proper Jacobian of the transformation during the estimation.

[5] Also, see Tsionas et al. (2014); Hampf and Krüger (2014); Atkinson and Tsionas (2015).

results on technological efficiency, technical change, productivity and (desirable) scale elasticity from our preferred system-based model are contrasted with those obtained from a *single*-equation DTDF model formulated under a rather strong assumption of exogenously determined inputs. We further assess the sensitivity of the results to the choice of the directional vector for the DTDF by re-estimating both models in the fixed Färe et al.'s (2005) "unit" direction specified prior to the estimation. We document dramatic distortions in banks' efficiency, productivity growth and scale elasticity estimates obtained from the DTDF estimated via a traditional single-equation approach and/or in an *ad hoc* pre-specified direction. We conclude that in studies of banking technology, where exogeneity of inputs is hardly plausible, a cost system approach, which we consider in this paper, is likely to provide a more robust estimation strategy.

The contribution of our paper is threefold. First, we offer a cost system approach to a consistent estimation of the DTDF based on the (behavioral) cost minimization assumption consistent with the economic theory. Second, our model of banking technology treats NPL as an undesirable output thus allowing us to credit banks for its reduction. To do so, we derive the DTDF-based Solow-type primal productivity index which is defined as the difference between the expansion rate in desirable outputs and contraction rates in inputs *and* undesirable outputs. The index thus explicitly recognizes the undesirability of NPL. Third, we estimate optimal directions for all inputs and outputs (both desirable and undesirable) conditional on banks' cost-minimizing behavior. The estimated DTDF direction captures the bank's movement to the point on a technological frontier where costs are minimized, thus eliminating the uncertainty associated with an *ad hoc* choice of the direction.

The rest of the paper unfolds as follows. Section 2 presents the DTDF framework for modeling banking production technology in the presence of undesirable outputs as well as discusses the associated productivity growth decomposition. In Section 3, we introduce a cost system approach to tackling endogeneity in the stochastic DTDF, followed by the discussion of the cost-optimal direction. Section 4 discusses Bayesian implementation of our estimator. Data are discussed in Section 5, and empirical results are reported in Section 6. We conclude in Section 7.

2 The Directional Technology Distance Function with Undesirable Outputs

We start by introducing the directional distance function formulation of the banking production process. Consider the production process in which J inputs $\mathbf{x} \in \mathbb{R}^J_+$ are being transformed into M desirable ("good") outputs $\mathbf{y} \in \mathbb{R}^M_+$ and P undesirable ("bad") outputs $\mathbf{b} \in \mathbb{R}^P_+$. The banking production technology is given by

$$\mathbb{T} \stackrel{\text{def}}{=} \{(\mathbf{x}, \mathbf{y}, \mathbf{b}) : \ \mathbf{x} \text{ can produce } (\mathbf{y}, \mathbf{b})\} \ , \tag{2.1}$$

subject to usual assumptions (see Chambers et al., 1998; Färe et al., 2005):

(1) closedness of $\mathbb{T}$;

(2) no free lunch: if $(\mathbf{x}, \mathbf{y}, \mathbf{b}) \in \mathbb{T}$ and $\mathbf{x} = \mathbf{0}$, then $(\mathbf{y}, \mathbf{b}) = (\mathbf{0}, \mathbf{0})$;

(3) null-jointness of the output production: if $(\mathbf{x}, \mathbf{y}, \mathbf{b}) \in \mathbb{T}$ and $\mathbf{b} = \mathbf{0}$, then $\mathbf{y} = \mathbf{0}$;

(4) free input disposability: if $(\mathbf{x}, \mathbf{y}, \mathbf{b}) \in \mathbb{T}$ and $\mathbf{x}' \geq \mathbf{x}$, then $(\mathbf{x}', \mathbf{y}, \mathbf{b}) \in \mathbb{T}$;

(5) weak joint disposability of desirable and undesirable outputs: if $(\mathbf{x}, \mathbf{y}, \mathbf{b}) \in \mathbb{T}$ and $0 \leq \kappa \leq 1$, then $(\mathbf{x}, \kappa\mathbf{y}, \kappa\mathbf{b}) \in \mathbb{T}$;

(6) free disposability of desirable outputs: if $(\mathbf{x}, \mathbf{y}, \mathbf{b}) \in \mathbb{T}$ and $\mathbf{y}' \leq \mathbf{y}$, then $(\mathbf{x}, \mathbf{y}', \mathbf{b}) \in \mathbb{T}$;

(7) feasibility of inaction: $(\mathbf{0}, \mathbf{0}, \mathbf{0}) \in \mathbb{T}$;

(8) convexity of $\mathbb{T}$.

The maximal distance between the observed $(\mathbf{x}, \mathbf{y}, \mathbf{b})$ and the boundary of banking technology $\mathbb{T}$ in a given direction $\mathbf{g} \equiv (-\mathbf{g}_x, \mathbf{g}_y, -\mathbf{g}_b) \in \mathbb{R}^J_- \times \mathbb{R}^M_+ \times \mathbb{R}^P_-$ is given by the value of the directional technology distance function (DTDF) defined as

$$\vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g}) \stackrel{\text{def}}{=} \sup_\beta \{\beta \in \mathbb{R}_+ : \ (\mathbf{x} - \beta \mathbf{g}_x, \mathbf{y} + \beta \mathbf{g}_y, \mathbf{b} - \beta \mathbf{g}_b) \in \mathbb{T}\} \ . \tag{2.2}$$

The DTDF in (2.2) seeks the simultaneous maximal expansion in desirable outputs and maximal reduction in inputs and undesirable outputs. It is advantageous over traditional (input- and/or output-oriented) distance functions because the distance that it measures is technology-oriented and can vary across individual inputs and outputs. However, unlike a traditional distance function, the DTDF constitutes an additive (not proportional) measure of inefficiency in a given direction $\mathbf{g}$, where the zero value of $\vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g})$ implies full technological efficiency. The function $\vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g})$ satisfies the following theoretical properties:[6]

(1) $\vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g}) \geq 0 \iff (\mathbf{x}, \mathbf{y}, \mathbf{b}) \in \mathbb{T}$;

(2) the translation property: for $\alpha \in \mathbb{R}$

$$\vec{D}_\tau(\mathbf{x} - \alpha \mathbf{g}_x, \mathbf{y} + \alpha \mathbf{g}_y, \mathbf{b} - \alpha \mathbf{g}_b; \mathbf{g}) = \vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g}) - \alpha \ ; \tag{2.3}$$

(3) concavity of $\vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g})$ in $(\mathbf{x}, \mathbf{y}, \mathbf{b}) \in \mathbb{T}$;

(4) positive monotonicity in inputs: if $(\mathbf{x}', \mathbf{y}, \mathbf{b}) \geq (\mathbf{x}, \mathbf{y}, \mathbf{b}) \in \mathbb{T}$, then $\vec{D}_\tau(\mathbf{x}', \mathbf{y}, \mathbf{b}; \mathbf{g}) \geq \vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g})$;

(5) negative monotonicity in desirable outputs: if $(\mathbf{x}, \mathbf{y}', \mathbf{b}) \leq (\mathbf{x}, \mathbf{y}, \mathbf{b}) \in \mathbb{T}$, then $\vec{D}_\tau(\mathbf{x}, \mathbf{y}', \mathbf{b}; \mathbf{g}) \geq \vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g})$;

(6) positive monotonicity in undesirable outputs: if $(\mathbf{x}, \mathbf{y}, \mathbf{b}') \geq (\mathbf{x}, \mathbf{y}, \mathbf{b}) \in \mathbb{T}$, then $\vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}'; \mathbf{g}) \geq \vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g})$;

(7) homogeneity of degree minus one in $\mathbf{g}$: $\vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \lambda \mathbf{g}) = \lambda^{-1} \vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g})$ for $\lambda > 0$.

Note that the DTDF in (2.2) nests several special cases of the directional distance functions. In the case of $\mathbf{b} = \mathbf{0}$ (and hence $\mathbf{g}_b = \mathbf{0}$), the DTDF in (2.2) collapses to the *standard* directional technology distance function (Chambers et al., 1998; Hudgins and Primont, 2007), which assumes *no* by-production of undesirable outputs, i.e.,

$$\vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{0}; (-\mathbf{g}_x, \mathbf{g}_y, \mathbf{0})) = \vec{D}_\top(\mathbf{x}, \mathbf{y}; (-\mathbf{g}_x, \mathbf{g}_y)) \ . \tag{2.4}$$

When $(\mathbf{g}_y, \mathbf{g}_b) = (\mathbf{0}, \mathbf{0})$, the DTDF in (2.2) becomes the directional *input* distance function similar to that of Chambers et al. (1996), i.e.,[7]

$$\vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; (-\mathbf{g}_x, \mathbf{0}, \mathbf{0})) = \vec{D}_i(\mathbf{x}, \mathbf{y}, \mathbf{b}; -\mathbf{g}_x) \ . \tag{2.5}$$

Lastly, if $\mathbf{g}_x = \mathbf{0}$, function (2.2) nests the directional *output* distance function with undesirable outputs (Chung et al., 1997; Färe et al., 2005), i.e.,

$$\vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; (\mathbf{0}, \mathbf{g}_y, -\mathbf{g}_b)) = \vec{D}_o(\mathbf{x}, \mathbf{y}, \mathbf{b}; (\mathbf{g}_y, -\mathbf{g}_b)) \ . \tag{2.6}$$

[6] The proofs are very similar to those in Chambers et al. (1996).

[7] To be precise, Chambers et al. (1996) consider the directional input distance function under the assumption of *no* undesirable outputs.

2.1 Productivity Growth Decomposition

Since the DTDF in (2.2) explicitly accommodates the by-production of undesirable outputs in the banking technology, we are able to credit banks for the reduction in undesirable outputs (nonperforming loans) along with the expansion in desirable outputs (earning assets such as loans, securities, etc.) when computing their productivity growth.

Specifically, letting the time enter the DTDF in (2.2) directly and recognizing that $(\mathbf{x}, \mathbf{y}, \mathbf{b})$ are all changing over time, the total time-differentiation of $\vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}, t; \mathbf{g})$ yields

$$\frac{\mathrm{d}\vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}, t; \mathbf{g})}{\mathrm{d}t} = \sum_j \frac{\partial \vec{D}_\tau(\cdot)}{\partial \log x_j} \dot{x_j} + \sum_m \frac{\partial \vec{D}_\tau(\cdot)}{\partial \log y_m} \dot{y_m} + \sum_p \frac{\partial \vec{D}_\tau(\cdot)}{\partial \log b_p} \dot{b_p} + \frac{\partial \vec{D}_\tau(\cdot)}{\partial t} \ , \tag{2.7}$$

where the "dot" above a variable denotes the time derivative of its log. Dividing both sides of (2.7) by (non-zero) $\left(1 + \vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}, t; \mathbf{g})\right)$ and some rearranging yields an equivalent expression:

$$\begin{aligned} \mathrm{PG} &\equiv -\sum_j \frac{\partial \log\left(1 + \vec{D}_\tau(\cdot)\right)}{\partial \log x_j} \dot{x_j} - \sum_m \frac{\partial \log\left(1 + \vec{D}_\tau(\cdot)\right)}{\partial \log y_m} \dot{y_m} - \sum_p \frac{\partial \log\left(1 + \vec{D}_\tau(\cdot)\right)}{\partial \log b_p} \dot{b_p} \\ &= \frac{\partial \log\left(1 + \vec{D}_\tau(\cdot)\right)}{\partial t} - \frac{\mathrm{d} \log\left(1 + \vec{D}_\tau(\cdot)\right)}{\mathrm{d}t} \ , \end{aligned} \tag{2.8}$$

where the left-hand side of the equality is a Solow-type Divisia index of productivity growth which we label PG, and the right-hand side of the equality yields a natural decomposition of PG into technical change $\mathrm{TC} \equiv \partial \log\left(1 + \vec{D}_\tau(\cdot)\right)/\partial t$ and efficiency change $\mathrm{EC} \equiv -\mathrm{d} \log\left(1 + \vec{D}_\tau(\cdot)\right)/\mathrm{d}t$. The PG index constitutes a weighted aggregate of percentage growth rates in inputs and desirable and undesirable outputs: $\dot{x_j}\ \forall\ j$, $\dot{y_m}\ \forall\ m$ and $\dot{b_p}\ \forall\ p$, respectively. The corresponding weights are

$$\begin{aligned} -\frac{\partial \log\left(1 + \vec{D}_\tau(\cdot)\right)}{\partial \log x_j} &\leq 0 \quad \forall \quad j = 1, \ldots, J \\ -\frac{\partial \log\left(1 + \vec{D}_\tau(\cdot)\right)}{\partial \log y_m} &\geq 0 \quad \forall \quad m = 1, \ldots, M \\ -\frac{\partial \log\left(1 + \vec{D}_\tau(\cdot)\right)}{\partial \log b_p} &\leq 0 \quad \forall \quad p = 1, \ldots, P \ , \end{aligned} \tag{2.9}$$

where the signs are dictated by the monotonicity properties of the DTDF in (2.2) and are in line with one's intuition.

3 Endogeneity in the Stochastic Directional Technology Distance Function: A Cost System Approach

In order to estimate the DTDF, one needs to specify its functional form. The quadratic form is a common choice for the directional distance functions, since it can be easily restricted to satisfy the translation property (e.g., Färe et al., 2005; Hudgins and Primont, 2007; Feng and Serletis, 2014). The function can then be estimated in two ways. Treating it as a deterministic function, one may

estimate it for a given direction $\mathbf{g}$ using the linear programming technique (subject to symmetry and theoretical regularity conditions).[8] Here, we employ an alternative approach where we treat the DTDF as a stochastic function in the spirit of Guarda et al. (2013) and Feng and Serletis (2014), who estimate the stochastic directional *input* and *output* distance functions.

3.1 Stochastic Formulation of the Directional Technology Distance Function

Before we proceed, note that the direction in which the DTDF is specified is uniquely defined by $(J+M+P-1)$ *ratios* of the elements in the $(J+M+P)$-dimensional vector $\mathbf{g}$. For example, in the case of one input ($J=1$), one desirable ($M=1$) and zero undesirable outputs ($P=0$), the directional vector is given by $(-g_x, g_y, 0)$ and is uniquely defined by the ratio of g_x to g_y.[9] Without the loss of generality, we can therefore normalize one of the elements in the directional vector $\mathbf{g}$ to be equal to one and control the direction of the DTDF using (magnitudes of) its remaining elements.

Following Guarda et al. (2013) and Feng and Serletis (2014), we normalize the stochastic DTDF in (2.2) in order to impose the translation property onto it for a given direction $\mathbf{g}$.[10] Specifically, setting α equal to the negative of one of the desirable outputs, say $\alpha = -y_k$, and normalizing the corresponding $g_{y_k} = 1$ (in the light of the argument above), the translation property in (2.3) can be rewritten as[11]

$$\vec{D}_\tau(\mathbf{x} + y_k\mathbf{g}_x, \widehat{\mathbf{y}} - y_k\widehat{\mathbf{g}}_y, \mathbf{b} + y_k\mathbf{g}_b; \mathbf{g}) = \vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g}) + y_k \ , \tag{3.1}$$

where we define $\widehat{\mathbf{y}} \equiv (y_1, \ldots, y_{k-1}, y_{k+1}, \ldots, y_M)$ and $\widehat{\mathbf{g}}_y \equiv (\mathbf{g}_{y_1}, ..., \mathbf{g}_{y_{k-1}}, \mathbf{g}_{y_{k+1}}, \ldots, \mathbf{g}_{y_M})$. Note that, due to the normalization $g_{y_k} = 1$, the desirable output y_k now enters the left-hand side of (3.1) in the capacity of "α" from (2.3) only. The "output" y_k has however disappeared from $\vec{D}_\tau(\mathbf{x} + y_k\mathbf{g}_x, \widehat{\mathbf{y}} - y_k\widehat{\mathbf{g}}_y, \mathbf{b} + y_k\mathbf{g}_b; \mathbf{g})$ because $y_k + \alpha g_{y_k} = y_k - y_k \cdot 1 = 0$. After rearranging, from (3.1) we get

$$\begin{aligned} y_k &= \vec{D}_\tau(\mathbf{x} + y_k\mathbf{g}_x, \widehat{\mathbf{y}} - y_k\widehat{\mathbf{g}}_y, \mathbf{b} + y_k\mathbf{g}_b; \mathbf{g}) - \vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g}) \\ &\stackrel{\text{def}}{=} \vec{D}_\tau(\mathbf{x} + y_k\mathbf{g}_x, \widehat{\mathbf{y}} - y_k\widehat{\mathbf{g}}_y, \mathbf{b} + y_k\mathbf{g}_b; \mathbf{g}) - u \ , \end{aligned} \tag{3.2}$$

where $u \stackrel{\text{def}}{=} \vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g}) \geq 0$ represents the (unobserved) composite technological inefficiency.[12] After adding the white noise term, the normalized DTDF in (3.2) constitutes a standard stochastic frontier model.

3.2 Endogeneity in the Stochastic Directional Technology Distance Function

The estimation of (3.2) is not trivial because of the potential endogeneity problem. While in many empirical applications (e.g., studies of service industries such as banking or electric power

[8] E.g., see Chung et al. (1997) and Färe et al. (2005) who estimate the deterministic directional *output* distance function in such a way.

[9] Intuitively, the direction is uniquely defined by $(J+M+P-1)$ angles between the vector $\mathbf{g}$ and its $(J+M+P-1)$ projections on all two-dimensional subspaces. The magnitudes of elements in $(\mathbf{g}_x, \mathbf{g}_y, \mathbf{g}_b)$ *per se* do not matter.

[10] We note that the translation property may alternatively be imposed using the set of parameter restrictions applied to the DTDF directly during the estimation (e.g., see Atkinson et al., 2014).

[11] Note that the normalization can be performed by setting α equal to any of the arguments of the DTDF. For instance, other options include setting α equal to some input or some undesirable output. However, opting for an input to normalize the DTDF produces highly nonlinear (although equivalent to those we employ) cost-minimizing first-order conditions, which we make use of later in the paper.

[12] Defined over inputs and both the desirable and undesirable outputs. For the discussion of how to disentangle technical inefficiency, conventionally defined over inputs and desirable outputs, from environmental inefficiency, defined over undesirable output, see Malikov et al. (2015a).

generation/distribution) one can justify the exogeneity of outputs $(\mathbf{y}, \mathbf{b})$ with relative ease, the exogeneity of inputs $\mathbf{x}$ is however a much harder sell. Economists commonly agree that input allocation is an outcome of endogenous choice made by firms. Specifically, the endogeneity problem arises due to the presence of J endogenous inputs $\mathbf{x}$ on the right-hand side of (3.2). To mitigate the problem, one either needs to instrument for $\mathbf{x}$ or, alternatively, to employ a system approach. In what follows, we opt for the system approach solution.

In order to meet the rank condition for the identification of the model, one needs the total of at least J independent equations in the system,[13] which equals the total number of endogenous variables (in our case, inputs $\mathbf{x} \in \mathbb{R}^J_+$). Also note that, despite that the desirable output y_k appears in the normalized DTDF (3.2) in the capacity of a left-hand-side "dependent variable", it is *not* an endogenous variable. Equation (3.2) can therefore be classified as an implicit function, which requires a proper Jacobian of the transformation be taken into account during the estimation (which we discuss in detail in Section 4).

We turn to economic theory in order to formulate additional independent equations for $\mathbf{x}$. Specifically, we augment the normalized DTDF in (3.2) with $(J-1)$ independent first-order conditions from banks' cost minimization problem. We prefer the assumption of the cost-minimizing behavior over profit-maximization primarily because *(i)* it is consistent with the premise of exogenously determined outputs, which is common to the banking industry, and *(ii)* the assumption of profit-maximization requires price information on undesirable outputs which is hard to measure and is rarely available in practice.[14] Incidentally, Atkinson et al. (2014) and Atkinson and Tsionas (2015) consider the estimation of the DTDF with undesirable outputs, where they propose tackling the endogeneity problem by formulating a profit-maximizing system of equations. While being a valid alternative to our estimator, the inconvenience of such an approach is that, due to unobservability of prices of undesirable outputs (say, $\mathbf{p}$), a researcher is forced to augment a system of *quasi*-profit-maximizing first-order conditions (which omit the terms containing $\mathbf{p}$) with the reduced-form demand equations for $\mathbf{b}$. Our *cost* system approach is not subject to the above problem.

Endogeneity in the estimation of the *normalized* DTDF is also discussed, although in a more narrow sense, in Guarda et al. (2013). In their paper, the authors are primarily concerned with the endogeneity problem due to the appearance of the left-hand-side "dependent variable" on the right-hand side of the normalized regression equation, which they suggest to remedy by setting $\widehat{\mathbf{g}} = \mathbf{0}$ (in our notation).[15] The endogeneity of inputs due to simultaneity, which is a primary focus of our paper, is however left unaddressed.

To formulate our cost system approach, we start with the bank's cost-minimizing objective defined as

$$\min_{\mathbf{x}} \quad \mathbf{w}'\mathbf{x} : \ (\mathbf{x}, \mathbf{y}, \mathbf{b}) \in \mathbb{T} \ , \tag{3.3}$$

which can be equivalently defined in terms of the DTDF in (2.2) as

$$\min_{\mathbf{x}} \quad \mathbf{w}'\mathbf{x} : \ \vec{D}_\tau(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g}) \geq 0 \ , \tag{3.4}$$

where $\mathbf{w} \in \mathbb{R}^J_+$ is the vector of J exogenous, competitively determined inputs prices.[16]

[13] Including the normalized DTDF (3.2).

[14] Furthermore, outputs in banking are services which are demand-determined and non-storable. Hence, cost minimization appears to be a more natural framework as opposed to profit maximization. This is one of the primary reasons why cost functions are routinely estimated for service industries.

[15] Tsionas et al. (2014) propose a more general solution to the same problem, which is based on accounting for a proper Jacobian of the transformation during the estimation.

[16] It may at first seem somewhat counterintuitive to speak of optimal behavior such as cost minimization while also

To link the optimization problem in (3.4) to the normalized DTDF in (3.2) which we seek to estimate, we substite (3.1) for $\vec{D}_\tau(\mathbf{x},\mathbf{y},\mathbf{b};\mathbf{g})$ [using the translation property for $\alpha=-y_k$] and let $g_{y_k}=1$. Equation (3.4) then transforms into

$$\min_{\mathbf{x}} \quad \mathbf{w}'\mathbf{x}: \ \vec{D}_\tau(\mathbf{x}+y_k\mathbf{g}_x, \widehat{\mathbf{y}}-y_k\widehat{\mathbf{g}}_y, \mathbf{b}+y_k\mathbf{g}_b;\mathbf{g})-y_k\geq 0\ . \tag{3.5}$$

The corresponding first-order conditions are

$$\frac{w_j}{w_q}=\left(\frac{\partial\vec{D}_\tau(\cdot)}{\partial\widetilde{x}_q}\right)^{-1}\frac{\partial\vec{D}_\tau(\cdot)}{\partial\widetilde{x}_j} \quad \forall \quad j(\neq q)=1,\ldots,J, \tag{3.6}$$

where, for convenience, we have defined $\widetilde{x}_j \stackrel{\text{def}}{=} x_j+y_kg_{x_j}\ \forall\ j=1,\ldots,J$.

Combining $(J-1)$ equations given in (3.6) with the normalized DTDF in (3.2) constitutes a complete, exactly identified (nonlinear) system of equations. Specifically, under the assumption of the quadratic functional form, the system consists of the following normalized DTDF (where we also introduce the time trend t)

$$\begin{aligned}
y_k=\theta_0&+\sum_j\alpha_j\widetilde{x}_j+\sum_{m(\neq k)}\beta_m\widetilde{y}_m+\sum_p\gamma_p\widetilde{b}_p+\theta_t t+\\
&\frac{1}{2}\sum_j\sum_{j'}\alpha_{jj'}\widetilde{x}_j\widetilde{x}_{j'}+\frac{1}{2}\sum_{m(\neq k)}\sum_{m'(\neq k)}\beta_{mm'}\widetilde{y}_m\widetilde{y}_{m'}+\frac{1}{2}\sum_p\sum_{p'}\gamma_{pp'}\widetilde{b}_p\widetilde{b}_{p'}+\frac{1}{2}\theta_{tt}t^2+\\
&\sum_j\sum_{m(\neq k)}\delta_{jm}\widetilde{x}_j\widetilde{y}_m+\sum_j\sum_p\phi_{jp}\widetilde{x}_j\widetilde{b}_p+\sum_{m(\neq k)}\sum_p\varphi_{mp}\widetilde{y}_m\widetilde{b}_p+\\
&\sum_j\theta_{x,j}t\widetilde{x}_j+\sum_{m(\neq k)}\theta_{y,m}t\widetilde{y}_m+\sum_p\theta_{b,p}t\widetilde{b}_p-u
\end{aligned} \tag{3.7}$$

and $(J-1)$ equations for $\widetilde{x}_j$, which we obtain from (3.6):[17]

$$\begin{aligned}
\widetilde{x}_j=\frac{2}{\alpha_{jj}w_q-\alpha_{qj}w_j}\Bigg[&w_j\left(\alpha_q+\frac{1}{2}\sum_{j'(\neq j)}\alpha_{qj'}\widetilde{x}_{j'}+\sum_{m(\neq k)}\delta_{qm}\widetilde{y}_m+\sum_p\phi_{qp}\widetilde{b}_p+\theta_{x,q}t\right)-\\
&w_q\left(\alpha_j+\frac{1}{2}\sum_{j'(\neq j)}\alpha_{jj'}\widetilde{x}_{j'}+\sum_{m(\neq k)}\delta_{jm}\widetilde{y}_m+\sum_p\phi_{jp}\widetilde{b}_p+\theta_{x,j}t\right)\Bigg] \quad \forall \quad j\neq q\ ,
\end{aligned} \tag{3.8}$$

where we have defined $\widetilde{y}_m \stackrel{\text{def}}{=} y_m-y_kg_{y_m}\ \forall\ m(\neq k)=1,\ldots,M$ and $\widetilde{b}_p \stackrel{\text{def}}{=} b_p+y_kg_{b_p}\ \forall\ p=1,\ldots,P$.

allowing for technological inefficiency in the banking production process, i.e., the instance when $\vec{D}_\tau(\cdot)$ takes a positive value. However, cost minimization (or even profit maximization) does not necessarily imply that the firm is fully efficient and operates on its technological frontier $[\vec{D}_\tau(\cdot)=0]$ (e.g., see Färe and Primont, 1995; Chambers et al., 1998). That is, banks may seek to minimize expenditures on inputs *given* the level of their technological inefficiency due to, say, imperfect organization of the production process, poor management and other (often unobserved) factors. In fact, this technological inefficiency can also be linked using duality to the cost inefficiency which measures the extra cost of producing below the technological frontier, as first shown by Schmidt and Lovell (1979). Also, see Chapters 4 and 6 of Kumbhakar et al. (2015).

[17] Note that, for the ease of it, we solve the first-order conditions for $\widetilde{x}_j$ thus treating them as endogenous variables (as opposed to x_j). The equivalence of working with $\widetilde{x}_j$ and working directly with x_j holds because $\partial\widetilde{x}_j/\partial x_j=1$.

We estimate the system (3.7)–(3.8) subject to symmetry and theoretical monotonicity and curvature restrictions.[18] Specifically, the symmetry conditions are

$$\alpha_{jj'} = \alpha_{j'j} \; ; \quad \beta_{mm'} = \beta_{m'm} \; ; \quad \gamma_{pp'} = \gamma_{p'p} \; . \tag{3.9}$$

Monotonicity of the (normalized) DTDF (3.2) in inputs and desirable and undesirable outputs, respectively, require that

$$\frac{\partial \vec{D}_\tau(\cdot)}{\partial x_j} = \alpha_j + \frac{1}{2}\sum_{j'} \alpha_{jj'}\widetilde{x}_{j'} + \sum_{m(\neq k)} \delta_{jm}\widetilde{y}_m + \sum_p \phi_{jp}\widetilde{b}_p + \theta_{x,j}t \geq 0 \quad \forall \quad j \tag{3.10a}$$

$$\frac{\partial \vec{D}_\tau(\cdot)}{\partial y_m} = \beta_m + \frac{1}{2}\sum_{m'(\neq k)} \beta_{mm'}\widetilde{y}_{m'} + \sum_j \delta_{jm}\widetilde{x}_j + \sum_p \varphi_{mp}\widetilde{b}_p + \theta_{y,m}t \leq 0 \quad \forall \quad m(\neq k) \tag{3.10b}$$

$$\frac{\partial \vec{D}_\tau(\cdot)}{\partial b_p} = \gamma_p + \frac{1}{2}\sum_{p'} \gamma_{pp'}\widetilde{b}_{p'} + \sum_j \phi_{jp}\widetilde{x}_j + \sum_{m(\neq k)} \varphi_{mp}\widetilde{y}_m + \theta_{b,p}t \geq 0 \quad \forall \quad p \; . \tag{3.10c}$$

Note that the above monotonicity restrictions are observation-specific and imposed at every data point. Further, the joint concavity of the DTDF in inputs and desirable and undesirable outputs is imposed by ensuring that all odd-numbered (even-numbered) principal minors of the Hessian matrix for (3.2) are non-positive (non-negative). The Hessian matrix is given by

$$\begin{bmatrix}
\alpha_{11} & \cdots & \alpha_{1J} & \delta_{11} & \cdots & \delta_{1(M-1)} & \phi_{11} & \cdots & \phi_{1P} \\
\vdots & & & & & & & & \vdots \\
\alpha_{J1} & \cdots & \alpha_{JJ} & \delta_{J1} & \cdots & \delta_{J(M-1)} & \phi_{J1} & \cdots & \phi_{JP} \\
\delta_{11} & \cdots & \delta_{J1} & \beta_{11} & \cdots & \beta_{1(M-1)} & \varphi_{11} & \cdots & \varphi_{1P} \\
\vdots & & & & & & & & \vdots \\
\delta_{1(M-1)} & \cdots & \delta_{J(M-1)} & \beta_{(M-1)1} & \cdots & \beta_{(M-1)(M-1)} & \varphi_{(M-1)1} & \cdots & \varphi_{(M-1)P} \\
\phi_{11} & \cdots & \phi_{J1} & \varphi_{11} & \cdots & \varphi_{(M-1)1} & \gamma_{11} & \cdots & \gamma_{1P} \\
\vdots & & & & & & & & \vdots \\
\phi_{1P} & \cdots & \phi_{JP} & \varphi_{1P} & \cdots & \varphi_{(M-1)P} & \gamma_{P1} & \cdots & \gamma_{PP}
\end{bmatrix} \tag{3.11}$$

where, for the ease of exposition, we have set $y_k = y_M$.

Since additional equations (the first-order conditions) do not contain any extra parameters, the system-based parameter estimates are likely to be more precise. Furthermore, technological metrics obtained from the cost system of DTDF are likely to be more meaningful because the economic behavior is embedded into the system through the first-order conditions. That is, if one believes in the economic behavior of bank managers, the first-order-conditions ought to be used in the estimation.

3.3 Optimal Direction

In our discussion above, following a common tradition in the literature, we have treated the directional vector $\mathbf{g}$, or its normalized counterpart $\widehat{\mathbf{g}}$, as fixed (pre-specified). However, it is important to recognize that the DTDF-based representation of the production process is an implicit function of the pre-specified direction. That is, there exists an infinite number of the DTDFs as defined by

[18] Recall that the translation property is already imposed by construction.

the unique values of $\widehat{\mathbf{g}}$. Hence, the estimates of the DTDF-based inefficiency, the productivity index PG as well as any other technological metric of the banking production process will change with the different choice of the directional vector.

To eliminate this uncertainty, we follow Tsionas et al. (2014) and let the data help us determine the direction in which the bank's movement toward the stochastic frontier is to be estimated. More specifically, we treat elements of the normalized directional vector $(\mathbf{g}_x, \widehat{\mathbf{g}}_y, \mathbf{g}_b)$ as unknown (non-negative) parameters which we estimate jointly with the remaining parameters in the cost system (3.7)–(3.8). The obtained "data-driven" estimates of the directional vector can then be interpreted as being "cost-optimal" due to the inclusion of the cost-minimizing first-order conditions in the system. That is, the estimated DTDF direction captures the bank's movement to the point on a technological frontier where costs are minimized. Atkinson et al. (2014) and Atkinson and Tsionas (2015) offer a similar "optimal" interpretation for the estimated directional vector. For the data-driven selection of the directional vector in a *deterministic* framework, also see Färe et al. (2013) and Hampf and Krüger (2014).

4 Bayesian Estimation

We start by rewriting the normalized DTDF in (3.7) for each cross-section $i = 1, \ldots, N$ and time period $t = 1, \ldots, T$ in the following form:

$$y_{k,it} = \widetilde{\mathbf{r}}_{it}(\widehat{\mathbf{g}})'\boldsymbol{\beta} - u_{it} + v_{0,it} \ , \tag{4.1}$$

where $\widetilde{\mathbf{r}}_{it}$ contains the quadratic regressors (including a unity for the intercept term) generated from $\widetilde{\mathbf{x}}_{it} \equiv (\widetilde{x}_{1,it}, \ldots, \widetilde{x}_{J,it}) \in \mathbb{R}^J$, $\widetilde{\mathbf{y}}_{it} \equiv (\widetilde{y}_{1,it}, \ldots, \widetilde{y}_{k-1,it}, \widetilde{y}_{k+1,it}, \ldots, \widetilde{y}_{M-1,it}) \in \mathbb{R}^{M-1}$, $\widetilde{\mathbf{b}}_{it} \equiv (\widetilde{b}_{1,it}, \ldots, \widetilde{b}_{P,it}) \in \mathbb{R}^P$ and the time trend t; $\boldsymbol{\beta}$ is a conformable vector of unknown parameters; and $\widehat{\mathbf{g}} \equiv (-\mathbf{g}_x, \widehat{\mathbf{g}}_y, -\mathbf{g}_b) \in \mathbb{R}^J_- \times \mathbb{R}^{M-1}_+ \times \mathbb{R}^P_-$ is the normalized vector of directional parameters. Here, we explicitly recognize that $\widetilde{\mathbf{r}}_{it}$ is a function of $\widetilde{\mathbf{g}}$ used in the construction of "tilde" regressors. Lastly, u_{it} is an (unobserved) one-sided composite technological inefficiency, and $v_{0,it}$ is a two-sided stochastic disturbance.

The $(J-1)$ first-order conditions in (3.8) can be written as

$$\widetilde{x}_{j,it}(g_{x_j}) = f_{j,it}(\boldsymbol{\alpha}_j)\widetilde{\mathbf{z}}_{j,it}(\widehat{\mathbf{g}})'\boldsymbol{\beta}_j + v_{j,it} \quad \forall \quad j = 1, \ldots, J-1 \ , \tag{4.2}$$

where $\widetilde{\mathbf{z}}_{j,it}$ denotes the vector of regressors in each one of the $(J-1)$ equations in (3.8), and $\boldsymbol{\beta}_j$ is a subset of $\boldsymbol{\beta}$ which can be obtained using a selection matrix $\mathbf{A}_j$, i.e., $\boldsymbol{\beta}_j = \mathbf{A}_j\boldsymbol{\beta}$ (selection matrices contain elements which are either 0 or 1). Further, $f_{j,it}(\boldsymbol{\alpha}_j) \overset{\text{def}}{=} 2/(\alpha_{jj}w_{q,it} - \alpha_{qj}w_{j,it})$ $(j \neq q)$ is a nonlinear function, where $\boldsymbol{\alpha}_j$ denotes the "α" coefficients from the DTDF in (3.7), a subset of $\boldsymbol{\beta}$. That is, $\boldsymbol{\alpha}_j \subset \boldsymbol{\beta}$ $\forall$ $j = 1, \ldots, J$. Lastly, $v_{j,it}$ is a two-sided stochastic disturbance.

We employ the following stochastic specification for system (4.1)–(4.2):

$$\mathbf{v}_{it} = (v_{0,it}, v_{1,it}, \ldots, v_{J-1,it})' \sim \text{i.i.d. } \mathbb{N}_J(\mathbf{0}, \boldsymbol{\Sigma}) \tag{4.3}$$

and independently of

$$u_{it} \sim \mathbb{N}_+\left(a_0 + a_1 t + \frac{1}{2}a_2 t^2, \sigma_u^2\right) \ , \tag{4.4}$$

where $\boldsymbol{a} \equiv (a_0, a_1, a_2)'$ are unknown parameters. Thus, we allow technological inefficiency to have a time-varying mean.

We next write the entire system (4.1)–(4.2) in the following form:

$$\begin{bmatrix} y_{it} \\ \widetilde{x}_{1,it} \\ \vdots \\ \widetilde{x}_{J-1,it} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{r}}_{it}(\widehat{\mathbf{g}})' & & & \\ & f_{1,it}(\boldsymbol{\alpha}_1)\widetilde{\mathbf{z}}_{1,it}(\widehat{\mathbf{g}})' & & \\ & & \ddots & \\ & & & f_{J-1,it}(\boldsymbol{\alpha}_{J-1})\widetilde{\mathbf{z}}_{J-1,it}(\widehat{\mathbf{g}})' \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_{J-1} \end{bmatrix} - \begin{bmatrix} u_{it} \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} v_{0,it} \\ v_{1,it} \\ \vdots \\ v_{J-1,it} \end{bmatrix} . \tag{4.5}$$

The (observation-specific) Jacobian of the transformation J_{it} from the vector of random disturbances $(v_{0,it} - u_{it}, v_{1,it}, \ldots, v_{J-1,it})'$ to the endogenous variables $\mathbf{x}_{it}$ (inputs) for system (4.5) is given by

$$\begin{aligned} J_{it} &= \left| \det \left[\frac{\partial (v_{0,it} - u_{it}, v_{1,it}, \ldots, v_{J-1,it})}{\partial \mathbf{x}_{it}'} \right] \right| \\ &= \left| \det \begin{bmatrix} -\partial y_{it} / \partial \widetilde{\mathbf{x}}_{it}' \\ \mathbf{C}_{it} \end{bmatrix} \right| , \end{aligned} \tag{4.6}$$

where we have made use of the fact that the gradient with respect to $x_{j,it}$ equals the gradient with respect to $\widetilde{x}_{j,it}$ for all j and $\mathbf{C}_{it} \stackrel{\text{def}}{=} [c_{jj'};\ c_{jj} = 1, c_{jj'} = \frac{\alpha_{jj'} w_{q,it} - \alpha_{qj} w_{j,it}}{\alpha_{jj} w_{q,it} - \alpha_{qj} w_{j,it}}\ \forall\ j \neq j']$. If it were not for the (observation-specific) nonlinear factors $f_{j,it}(\boldsymbol{\alpha}_j)$ in (4.5), then we would have a *linear* simultaneous equations model with technological inefficiency.

To perform Bayesian analysis and tackle the nonlinearity of our model, we estimate it in two stages. First, we estimate a single-equation DTDF in (4.1) subject to all constraints so that $\boldsymbol{\beta} \in \mathscr{B}$, where $\mathscr{B}$ denotes the set of acceptable parameters, using a Bayesian approach.[19] The resulting estimates of the parameter subset $\boldsymbol{\alpha} \equiv (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{J-1}) \subset \boldsymbol{\beta}$, say $\overline{\boldsymbol{\alpha}}$, can be used to produce an *approximately* linear system of the first-order conditions in (4.2) of the form

$$\widetilde{x}_{j,it}(g_{x_j}) = f_{j,it}(\overline{\boldsymbol{\alpha}}_j)\widetilde{\mathbf{z}}_{j,it}(\widehat{\mathbf{g}})'\boldsymbol{\beta}_j + v_{j,it} = \widetilde{\mathbf{Z}}_{j,it}(\widehat{\mathbf{g}})'\boldsymbol{\beta}_j + v_{j,it} \quad \forall \quad j = 1, \ldots, J-1 , \tag{4.7}$$

where $\widetilde{\mathbf{Z}}_{j,it}(\widehat{\mathbf{g}}) \stackrel{\text{def}}{=} f_{j,it}(\overline{\boldsymbol{\alpha}}_j)\widetilde{\mathbf{z}}_{j,it}(\widehat{\mathbf{g}})$.
Taking into account that $\boldsymbol{\beta}_j = \mathbf{A}_j\boldsymbol{\beta}$, we can rewrite the system in (4.5) as

$$\begin{bmatrix} y_{it} \\ \widetilde{x}_{1,it} \\ \vdots \\ \widetilde{x}_{J-1,it} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{r}}_{it}(\widehat{\mathbf{g}})' & & & \\ & \left(\mathbf{A}_1'\widetilde{\mathbf{Z}}_{1,it}(\widehat{\mathbf{g}})\right)' & & \\ & & \ddots & \\ & & & \left(\mathbf{A}_{J-1}'\widetilde{\mathbf{Z}}_{J-1,it}(\widehat{\mathbf{g}})\right)' \end{bmatrix} \boldsymbol{\beta} - \begin{bmatrix} u_{it} \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} v_{0,it} \\ v_{1,it} \\ \vdots \\ v_{J-1,it} \end{bmatrix} , \tag{4.8}$$

which can be written in a more compact notation as follows:

$$\mathbf{Y}_{it} = \mathbf{Z}_{it}(\widehat{\mathbf{g}})\boldsymbol{\beta} - u_{it}\boldsymbol{\iota} + \mathbf{v}_{it} , \tag{4.9}$$

where $\boldsymbol{\iota} = (1, 0, \ldots, 0)'$.

Although the system above is only approximate, we can use it to perform Bayesian inference through MCMC to obtain first-stage inferences for $\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{a}, \sigma_u^2\right)$. These approximate inferences can then be made *exact* by taking the nonlinear terms $f_{j,it}(\cdot)$, which have been previously fixed at $\boldsymbol{\alpha} = \overline{\boldsymbol{\alpha}}$, into account. Conditional on u_{it}, the implementation of MCMC for (4.9) is well known. Since

[19] We use 850,000 MCMC iterations, the first 350,000 of which are burned to mitigate start-up effects and obtain convergence.

first-stage inferences are only approximate, we impose the restrictions (monotonicity and curvature constraints) only at the means. Furthermore, while exact posterior inference requires taking the Jacobian of the transformation into account in (4.9), we however ignore it at this preliminary stage and thus treat the system as seemingly unrelated regressions.

In the second stage, we use a Metropolis-Hastings algorithm taking the exact Jacobian into account to impose regularity restrictions at each data point and to perform exact inferences for the entire parameter vector $\boldsymbol{\beta}$. Specifically, drawing from u_{it} relies on the following conditional posterior distribution:

$$\sigma_u^{-1} \exp\left\{-\frac{1}{2}\left(\mathbf{Q}_{it}+u_{it}\boldsymbol{\iota}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{Q}_{it}+u_{it}\boldsymbol{\iota}\right)\right\} \times \exp\left\{-\frac{1}{2\sigma_u^2}\left(u_{it}-\mu_{it}\right)^2\right\}\Phi\left(\frac{\mu_{it}}{\sigma_u}\right)^{-1} \tag{4.10}$$

which can be expressed more compactly as follows:

$$u_{it}|\mathbf{Y},\mathbf{Z},\boldsymbol{\beta},\widehat{\mathbf{g}},\sigma_u \sim \mathbb{N}_+\left(\widehat{u}_{it},v_{u,it}^2\right) \tag{4.11}$$

where $\mathbf{Q}_{it}=\mathbf{Y}_{it}-\mathbf{Z}_{it}(\widehat{\mathbf{g}})\boldsymbol{\beta}$, $\widehat{u}_{it}=\frac{-\boldsymbol{\Sigma}^{-1}\mathbf{Q}_{it}+\mu_{it}/\sigma_u^2}{\boldsymbol{\iota}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\iota}+\sigma_u^2}$, $\mu_{it}=a_0+a_1t+\frac{1}{2}a_2t^2$ and $v_{u,it}^2=\frac{1}{\boldsymbol{\iota}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\iota}+\sigma_u^2}$.

From (4.9), we obtain the approximate posterior mean $\overline{\boldsymbol{\beta}}$ and its posterior covariance matrix $\mathbf{V}$. We use a Student-t with 5 degrees of freedom, denoted by $t_5\left(\overline{\boldsymbol{\beta}},h\mathbf{V}\right)$, as a proposal distribution, where $h>0$ is a tuning parameter used to calibrate the acceptance rate of MCMC. We denote the corresponding density by $\phi\left(\boldsymbol{\beta};\overline{\boldsymbol{\beta}},h\mathbf{V}\right)$.

Suppose the MCMC is currently at draw $\boldsymbol{\beta}^{(s)}$ and we obtain a candidate draw $\boldsymbol{\beta}^{(c)}\sim t_5\left(\overline{\boldsymbol{\beta}},h\mathbf{V}\right)$ subject to regularity restrictions $\boldsymbol{\beta}^{(c)}\in\mathscr{B}$. Then, the candidate is accepted with probability

$$\min\left\{1,\ \frac{p\left(\boldsymbol{\beta}^{(c)}|\mathbf{Y},\mathbf{Z},\mathbf{u},\boldsymbol{\Sigma},\boldsymbol{a},\sigma_u\right)/\phi\left(\boldsymbol{\beta}^{(c)};\overline{\boldsymbol{\beta}},h\mathbf{V}\right)}{p\left(\boldsymbol{\beta}^{(s)}|\mathbf{Y},\mathbf{Z},\mathbf{u},\boldsymbol{\Sigma},\boldsymbol{a},\sigma_u\right)/\phi\left(\boldsymbol{\beta}^{(s)};\overline{\boldsymbol{\beta}},h\mathbf{V}\right)}\right\}\ , \tag{4.12}$$

where $\mathbf{u}=[u_{it}]$ denotes the vector of technological inefficiencies. The tuning parameter h is selected so that the acceptance rate is between 20% and 30%.

Drawings from the conditional posterior of $\boldsymbol{a}$ are facilitated by the re-parametrization $a_\kappa^*=a_\kappa/\sigma_u;\ \kappa=0,1,2$, the conditional posterior for which is given by

$$p\left(\boldsymbol{a}^*|\mathbf{u}\right)\propto\exp\left\{-\frac{1}{2}\sum_{i,t}\left(u_{it}-a_0^*-a_1^*t-a_2^*t^2\right)^2\right\}\prod_{i,t}\Phi\left(a_0^*+a_1^*t+a_2^*t^2\right)^{-1}\ . \tag{4.13}$$

Random drawings can be realized using the first term and accepting the draw with a Metropolis-Hastings probability determined by the second term. Specifically, from the first term, we have

$$\boldsymbol{a}\sim\mathbb{N}\left(\boldsymbol{a}^*,\mathbf{V}_a\right)\ , \tag{4.14}$$

where $\boldsymbol{a}^*=\left(\mathbf{D}'\mathbf{D}\right)^{-1}\mathbf{D}'\mathbf{u}$, $\mathbf{D}=(1,t,t^2)'$ and $\mathbf{V}_a=\left(\mathbf{D}'\mathbf{D}\right)^{-1}$.
Similarly, for σ_u^2, we have

$$\frac{\sum_{i,t}\left(u_{it}-a_0-a_1t-a_2t^2\right)^2}{\sigma_u^2}\Big|\mathbf{u}\sim\chi_{NT}^2\ . \tag{4.15}$$

The nonlinear term $\Phi\left(\frac{a_0+a_1t+a_2t^2}{\sigma_u}\right)^{-1}$ is accommodated by a Metropolis-Hastings step. Lastly, the system covariance matrix can be drawn easily using standard results for the Wishart distribution:

$$p\left(\boldsymbol{\Sigma}|\mathbf{Y},\mathbf{Z},\boldsymbol{\beta},\widehat{\mathbf{g}}\right) \propto |\boldsymbol{\Sigma}|^{-(NT+J+1)/2} \exp\left\{-\frac{1}{2}\text{tr}\boldsymbol{\Sigma}^{-1}\sum_{i,t}\mathbf{Q}_{it}(\boldsymbol{\beta},\widehat{\mathbf{g}})\mathbf{Q}_{it}(\boldsymbol{\beta},\widehat{\mathbf{g}})'\right\} . \tag{4.16}$$

To compute observation-specific posterior means of the vector functions of interest, such as returns to scale, technological change, productivity growth, etc., generically denoted by $\boldsymbol{h}_{it}\left(\boldsymbol{\beta}\right)$, we use an MCMC sample $\left\{\boldsymbol{\beta}^{(s)}, s=1,...,S\right\}$, i.e.,

$$\mathbb{E}\left[\boldsymbol{h}_{it}\left(\boldsymbol{\beta}\right)|\mathbf{Y},\mathbf{Z},\widehat{\mathbf{g}}\right] = S^{-1}\sum_{s=1}^{S}\boldsymbol{h}_{it}\left(\boldsymbol{\beta}^{(s)}\right) . \tag{4.17}$$

Priors. For the parameter vector $\boldsymbol{\beta}$, our prior is flat in the domain of restrictions, i.e., $p\left(\boldsymbol{\beta}\right) \propto I\left(\boldsymbol{\beta}\in\mathscr{B}\right)$. For σ_u^2 we assume a proper but relatively non-informative prior in the inverted-Gamma family: $\frac{1}{\sigma_u^2}\sim\chi^2(10)$. For parameters of the location of a truncated normal inefficiency distribution, we assume that $a_0\sim\mathbb{N}\left(-3,0.1^2\right)$, $a_1\sim\mathbb{N}(-0.1,0.01^2)$ and $a_2\sim\mathbb{N}(-0.001,0.01^2)$. Our choice of priors for $\boldsymbol{a}$ and σ_u is motivated by calibrating a proper prior for technological inefficiency $u_{it}\sim\mathbb{N}_+\left(a_0+a_1t+\frac{1}{2}a_2t^2,\sigma_u^2\right)$. For the covariance matrix $\boldsymbol{\Sigma}$, we choose a prior in the inverted Wishart family, i.e.,

$$p\left(\boldsymbol{\Sigma}\right) \propto |\boldsymbol{\Sigma}|^{-(\overline{\nu}+J+1)/2}\exp\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\overline{\mathbf{A}}\right)\right\}$$

for $\overline{\nu}=1$ and $\overline{\mathbf{A}}=0.01\times\mathbf{I}_J$, where $\mathbf{I}_J$ is the identity matrix of dimension J.

Optimal (data-driven) direction. The described MCMC scheme can be implemented for any given pre-specified (fixed) value of the normalized directional vector $\widehat{\mathbf{g}}$. However, when the directional vector is *un*known, given any normalization, we can redefine $\boldsymbol{\beta}$ as $(\boldsymbol{\beta}',\widehat{\mathbf{g}}')'$ so that both the parameter vector $\boldsymbol{\beta}$ and the directional vector $\widehat{\mathbf{g}}$ are merged into the same vector and then treat them together as unknown parameters of interest. We can then apply the Metropolis-Hastings MCMC scheme described above without any major changes. Our prior for $\widehat{\mathbf{g}}\equiv(-\mathbf{g}_x,\widehat{\mathbf{g}}_y,-\mathbf{g}_b)$ is non-informative and only imposes the restriction that each element $(\mathbf{g}_x,\widehat{\mathbf{g}}_y,\mathbf{g}_b)$ is positive. We do so by re-parameterizing $\widehat{\mathbf{g}}$ as $\widehat{\mathbf{g}}^*=\exp\left(\widehat{\mathbf{g}}\right)$ and treating $\widehat{\mathbf{g}}^*$ as the unrestricted parameters of interest.

5 Framework and Data

The data on commercial banks come from Call Reports available from the Federal Reserve Bank of Chicago and include all FDIC insured commercial banks with reported data for 2001:I-2010:IV. It is well-known that commercial banks may starkly differ from one another in terms of the size, capitalization, regulatory environment, etc., suggesting potential heterogeneity in production technologies across banks. To alleviate these complications, in this paper we focus on a selected subsample of relatively homogeneous large banks,[20] namely those with total assets in excess of one billion dollars (in 2005 U.S. dollars) in the first three years of observation. We further exclude internet banks, commercial banks conducting primarily credit card activities and banks chartered outside

[20]We thank an anonymous referee for this suggestion.

Table 1: Data Summary Statistics

Variable	Mean	5th Perc.	Median	95th Perc.
y_1	1,910,364.2	5,382.6	123,954.4	6,027,401.5
y_2	8,118,512.7	357,906.2	1,451,297.8	29,005,330.8
y_3	5,971,810.2	100,482.6	503,807.1	23,392,817.5
y_4	8,132,195.2	169,367.0	784,193.8	20,153,051.7
y_5	446,770.0	3,448.2	23,813.0	1,520,048.4
b	382,804.4	1,070.8	19,360.0	950,395.0
x_1	5,109.1	236.3	759.0	16,520.9
x_2	245,563.1	7,344.3	41,884.4	902,259.0
x_3	7,790,193.8	198,008.1	844,580.2	24,478,720.1
x_4	355,793.5	8,447.0	71,546.4	1,249,705.9
x_5	14,257,910.2	593,056.8	1,726,840.7	47,503,806.2
e	2,614,537.0	96,702.2	323,246.5	9,055,215.2
w_1	65.23	40.75	60.05	107.68
w_2	0.435	0.148	0.308	1.035
w_3	0.032	0.012	0.031	0.052
w_4	0.011	0.001	0.008	0.031
w_5	0.019	0.005	0.017	0.037
ta	27,222,337.8	1,207,845.8	3,267,751.7	85,622,664.2

NOTES: y_1 – consumer loans; y_2 – real estate loans; y_3 – commercial and industrial loans; y_4 – securities; y_5 – off-balance sheet income; b – total nonperforming loans; x_1 – labor; x_2 – physical capital; x_3 – purchased funds; x_4 – interest-bearing transaction accounts; x_5 – non-transaction accounts; e – financial (equity) capital; w_1–w_5 are the prices of x_1–x_5, respectively; ta – total assets. All variables but labor (x_1) and input prices (w_1–w_5) are in thousands of real 2005 U.S. dollars. x_1 is the number of full-time equivalent employees. All input prices but w_1 are interest rates and thus are unit-free. w_1 is measured in real 2005 U.S. dollars.

the continental U.S. We also omit observations for which negative values for assets, equity, outputs and input prices are reported. These are likely to be the result of erroneous data reporting. The remaining data sample is an unbalanced panel with 2,397 bank-year observations for 285 banks. We deflate all nominal stock variables to 2005 U.S. dollars using the Consumer Price Index (for all urban consumers).

We model the bank's production technology using the commonly used "intermediation approach" of Sealey and Lindley (1977), according to which a bank's balance sheet is assumed to capture the essential structure of its core business. Liabilities, together with physical capital and labor, are taken as inputs to the bank's production process, whereas assets (other than physical) are considered as outputs. Liabilities include core deposits and purchased funds; assets include loans and trading securities. In this paper, we generalize the standard framework of modeling banking technology by explicitly recognizing that the bank's production of desirable outputs, such as earning loans, is usually accompanied by the simultaneous by-production of an undesirable output that takes the form of non-performing loans.

We define the following *desirable* outputs of the bank's production process: consumer loans (y_1), real estate loans (y_2), commercial and industrial loans (y_3) and securities (y_4). These output categories are essentially the same as those in Berger and Mester (1997, 2003). Following Hughes and Mester (1998, 2013), we further include off-balance-sheet income (y_5) as an additional output.[21]

[21] In this paper, we measure off-balance-sheet income by the net non-interest income (less service charges on deposits). We acknowledge that this measure of off-balance-sheet income may be biased downward by losses. Ideally, one would want to use the *gross* non-interest income, which however is infeasible to construct based on the information

The bank's *undesirable* output is defined as the total non-performing loans (b). The variable inputs are labor, i.e., the number of 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). We also include financial (equity) capital (e) as an additional input to the production technology. However, due to the unavailability of the price of equity capital, we follow Berger and Mester (1997, 2003) and Feng and Serletis (2009) in modeling e as a quasi-fixed input. The treatment of equity capital as an input to banking production technology is consistent with Hughes and Mester's (1993, 1998) argument that banks may use the latter as a source of loanable funds and thus as a cushion against losses. Since equity capital is modeled as being quasi-fixed, it does not have a corresponding first-order condition and rather enters the system as a conditioning (contextual) variable. We compute the prices of variable inputs (w_1 through w_5) by dividing total expenses on each input by the corresponding input quantity. Table 1 presents summary statistics of the data we use.[22]

Lastly, we note that, because the DTDF yields an additive measure of the distance to the frontier, it is *not* invariant to the units of measurement of its arguments $(\mathbf{x}, \mathbf{y}, \mathbf{b})$. That is, the value of the DTDF in any given direction changes, should the variables be rescaled. To mitigate this problem, we follow Färe et al. (2005) and standardize all variables prior to the estimation by subtracting their respective sample means and dividing by their sample standard deviations.

6 Empirical Results

Instead of looking at the parameter estimates of the DTDF, we focus on more informative technological metrics of the banking production process: technological efficiency, productivity growth (PG) and its two components — technical change (TC) and efficiency change (EC), and scale elasticity. Observation-specific posterior estimates of technological efficiency are obtained from the posterior conditional mean of u, and the estimates of PG, TC and EC are computed as specified in (2.8). To compute posterior scale elasticity estimates, we generalize the results in Zelenyuk (2013) to the case of the DTDF with undesirable outputs. Specifically, for any given direction the *desirable* scale elasticity is given by

$$\mathrm{DSE}\left(\mathbf{x}, \mathbf{y}, \mathbf{b}; \mathbf{g}\right) \stackrel{\text{def}}{=} \left.\frac{\partial \log \lambda_y}{\partial \log \lambda_x}\right|_{\lambda_y = \lambda_x = 1} : \ \vec{D}_\tau(\lambda_x \mathbf{x}, \lambda_y \mathbf{y}, \mathbf{b}; \mathbf{g}) \geq 0 \ , \tag{6.1}$$

which, under our DTDF normalization, can be easily shown to equal

$$\mathrm{DSE}\left(\mathbf{x}, \widehat{\mathbf{y}}, \mathbf{b}; \widehat{\mathbf{g}}\right) = -\frac{\sum_j \frac{\partial \log \vec{D}_\tau(\cdot)}{\partial \log x_j}}{\sum_{m(\neq k)} \frac{\partial \log \vec{D}_\tau(\cdot)}{\partial \log y_m}} \ . \tag{6.2}$$

Here, we define DSE as the ratio of equiproportional percentage change in *desirable* outputs to equiproportional percentage change in inputs, while holding the levels of *undesirable* outputs unchanged. The latter can be accomplished by diverting some of the inputs into ensuring that no additional undesirable outputs are by-produced along with the production of desirable outputs. For instance, banks may engage labor and capital to more carefully examine new loan applications or more actively monitor the latter upon their issuance in order to ensure that all new loans are "performing", i.e., paid back duly. The above definition of scale elasticity is such that the bank is said to exhibit increasing/constant/ decreasing returns to (desirable) scale if DSE is greater than/equal to/less than one.

reported in the data.

[22] For more details on the construction of the variables, see the Appendix of Malikov et al. (2015b).

Table 2: Summary of Posterior Estimates over the 2001–2010 Period

	System Fixed Dir.	System Optimal Dir.	Single Eq. Fixed Dir.	Single Eq. Data-Driven Dir.
Efficiency				
Mean	0.9644	0.9431	0.4583	0.9459
Median	0.9640	0.9513	0.4580	0.9533
S.D.	0.0042	0.0403	0.0044	0.0390
95% Bayes Interval	(0.9560; 0.9740)	(0.8482; 0.9945)	(0.4490; 0.4680)	(0.8562; 0.9966)
Productivity Growth (PG)				
Mean	0.0177	0.0110	0.0119	0.0042
Median	0.0177	0.0110	0.0117	0.0040
S.D.	0.0020	0.0022	0.0105	0.0334
95% Bayes Interval	(0.0138; 0.0216)	(0.0063; 0.0158)	(0.0088; 0.0321)	(–0.0608; 0.0725)
Technical Change (TC)				
Mean	0.0177	0.0110	0.0120	0.0042
Median	0.0178	0.0111	0.0120	0.0042
S.D.	0.0017	0.0022	0.0032	0.0321
95% Bayes Interval	(0.0146; 0.0212)	(0.0057; 0.0184)	(0.0058; 0.0183)	(–0.0551; 0.0699)
Efficiency Change (EC)				
Mean	0.0000	–0.0000	–0.0000	0.0000
Median	0.0000	–0.0003	0.0000	–0.0001
S.D.	0.0010	0.0008	0.0100	0.0098
95% Bayes Interval	(–0.0019; 0.0019)	(–0.0015; 0.0016)	(–0.0192; 0.0195)	(–0.0190; 0.0197)
Desirable Scale Elasticity (DSE)				
Mean	1.0211	0.8979	1.2580	1.3323
Median	0.9782	0.8831	1.4657	1.2783
S.D.	0.2916	0.2170	0.2391	0.3214
95% Bayes Interval	(0.5863; 1.7018)	(0.5821; 1.2774)	(1.0066; 2.4332)	(0.8607; 2.1036)
Undesirable Output Elasticity (E_b)				
Mean	0.2796	0.2089	0.1702	0.2203
Median	0.2800	0.2100	0.1700	0.2200
S.D.	0.0901	0.0507	0.0394	0.0146
95% Bayes Interval	(0.1049; 0.4570)	(0.1090; 0.3060)	(0.0949; 0.2521)	(0.1910; 0.2500)

Table 3: Posterior Estimates of Directional Parameters

	System		Single Eq.	
	Mean	*S.D.*	*Mean*	*S.D.*
g_{x_1}	1.642	0.037	0.340	0.564
g_{x_2}	3.969	0.041	0.704	0.6413
g_{x_3}	36.563	0.018	8.550	0.868
g_{x_4}	5.858	0.018	1.854	0.511
g_{x_5}	1572.078	0.058	0.025	0.788
g_{y_1}	0.372	0.233	0.219	0.799
g_{y_2}	10.453	0.188	5.509	1.431
g_{y_3}	1.256	0.050	0.679	2.649
g_{y_4}	905.785	0.006	378.694	0.234
g_b	0.060	0.003	0.022	1.285

In order to comprehensively investigate how the results change if the endogeneity of inputs is taken for granted when estimating the DTDF for banks, we estimate several auxiliary models. Specifically, in addition to our preferred system-based model (3.7)–(3.8) developed in Sections 3–4, we also estimate a *single*-equation stochastic DTDF as commonly done in the literature (e.g., Koutsomanoli-Filippaki et al., 2009; Feng and Serletis, 2014). That is, we estimate (3.7) without additional first-order condition equations, which is equivalent to making a rather strong assumption of exogenously determined inputs. Furthermore, we estimate both system- and single-equation-based models twice: *(i)* in a given fixed direction specified prior to the estimation and *(ii)* in the data-driven direction as discussed in Section 3. However, note that, unlike in the case of the system-based model (3.7)–(3.8), the data-driven direction selected when estimating a single-equation DTDF cannot be interpreted as "optimal" since no economic behavior is imposed on the parameters of the equation. When fixing the direction, we use Färe et al.'s (2005) "unit" direction and set all elements of $(\mathbf{g}_x, \widehat{\mathbf{g}}_y, \mathbf{g}_b)$ equal to ones. In total, we estimate four models. In all four instances, we use y_M (off-balance-sheet income) to impose the translation property onto the DTDF (via the normalization $\alpha = -y_M$) and let $g_{y_M} = 1$ as described in Section 3.

Table 2 reports the summary of posterior estimates of the technological metrics computed based on the cost system of the DTDF and a single-equation DTDF for both fixed and estimated directional vectors. Comparing technological efficiency estimates across the models, we find that a single-equation approach, which assumes the endogeneity of inputs away, produces unreasonably low estimates of banks' efficiency when being fitted in the popular *ad hoc* fixed "unit" direction: the posterior mean and median are at 0.45. All other models, including our preferred system-based model estimated in the optimal direction, suggest mean efficiency at the 0.94–0.96 level, indicating a starkly low level of *technological* inefficiency exhibited by the large banks in the U.S. Perhaps unsurprisingly, our preferred estimates of banks' efficiency (column 2 of Table 2) are somewhat greater than comparable Bayesian estimates of primal[23] technological efficiency recently reported in the literature (e.g., Feng and Zhang, 2012), since our approach explicitly recognizes the undesirable nature of nonperforming loans and does not penalize banks for diverting some of their inputs in an attempt to lower NPL. Further, when estimating both the system and single-equation DTDF in the pre-specified direction, the computed posterior efficiency estimates hardly vary as indicated by the near-zero magnitudes of the respective posterior standard deviations (see columns 1 and 3). Letting the data select the direction produces estimates of technological efficiency with a consider-

[23]Note that our estimates of technological (in)efficiency cannot be meaningfully compared to those based on widely-used *dual* specifications of the banking production technology such as cost and/or profit functions.

ably greater variation across banks (also see Table 3 for the data-driven estimates of the directional parameters).

Regardless of the model used, we consistently fail to detect any time variability in the estimates of banks' efficiency levels. That is, the posterior estimates of parameters $\boldsymbol{a}$ inside the mean function of technological inefficiency u given in (4.4) are virtually zeros.[24] Consequently, we document no significant efficiency change across banks over the course of our sample period: the posterior means and medians of EC are zeros across all four models. As a result of this, the primary driving force behind the productivity growth in the banking industry appears to be technological change, as indeed confirmed in Table 2. Using Bayesian methods, Feng and Serletis (2010) and Feng and Zhang (2012) similarly document statistically insignificant EC for large U.S. banks and find technical change to be the dominant component in the productivity growth decomposition.

We next analyze the posterior estimates of technical change. Two remarks are warranted here. First, the TC estimates obtained based on a single-equation DTDF fitted in a data-driven direction appear to be of unreasonably low magnitude (see column 4 of Table 2). Specifically, the model indicates virtually *no* significant technological advancement by large banks over the course of 2001–2010, which seems quite difficult to believe given the favorable effects of recent advances in information technologies and regulatory changes on large financial institutions. In contrast, our preferred system-based model (3.7)–(3.8) produces TC estimates of rather more plausible magnitudes: a posterior mean of 1.1% per annum with the corresponding 95% posterior coverage region of (0.5%; 1.8%), when estimated in the optimal direction. Second, when compared to the estimates obtained in the data-driven direction, using the *ad hoc* fixed "unit" direction seems to overestimate TC in the case of both a cost system approach and a traditional single-equation approach. The latter exemplifies the sensitivity of empirical estimates of directional distance functions to the choice of directional vector. Researchers ought to exercise caution when choosing the direction for the DTDF. Following Atkinson et al. (2014), we advocate for the use of the data-driven optimal direction in conjunction with the imposition of economic behavior onto the DTDF.

Differences in the estimates of TC across models translate themselves into differences in the productivity growth estimates, which Figure 1 vividly illustrates. It plots the productivity indices that are normalized to 100 in the year 2001 and are constructed using the (total)-asset-weighted average annual productivity growth rates (over all banks in the sample). Figure 1, which plots such PG indices for all four models, graphically demonstrates the sensitivity of the productivity estimates to *(i)* potential endogeneity in the DTDF and *(ii)* the use of an *ad hoc* directional vector. Specifically, the implied cumulative wedge between the (asset-weighted) productivity index obtained from a system-based model and that from a single-equation DTDF is 540 percentage basis points when using the pre-specified "unit" direction and 1,120 percentage basis points when using data-driven directions. According to our preferred system-based model estimated in the optimal direction, over the course of 2001–2010, the productivity of large U.S. commercial banks grew at the weighted average rate of 1.1% per year with a rather timid but significantly positive cumulative ten-year growth of 10.2%. While our PG estimates may appear to be somewhat conservative in comparison to other estimates of the productivity growth exhibited by large U.S. banks during the recent decade reported in the literature,[25] one can plausibly argue that our productivity growth measure is expected to produce estimates on a lower level since it accounts for the presence of an undesirable by-product (NPL) of the banking production technology.[26]

[24] Modeling the mean of technological efficiency u as a function of time dummies (in place of the second-order polynomial of time trend) yields similar results.

[25] For instance, both Feng and Serletis (2009) and Feng and Zhang (2012) report an average annual PG of about 2% for large banks in 1998–2005.

[26] We thank an anonymous referee for bringing this point to our attention.

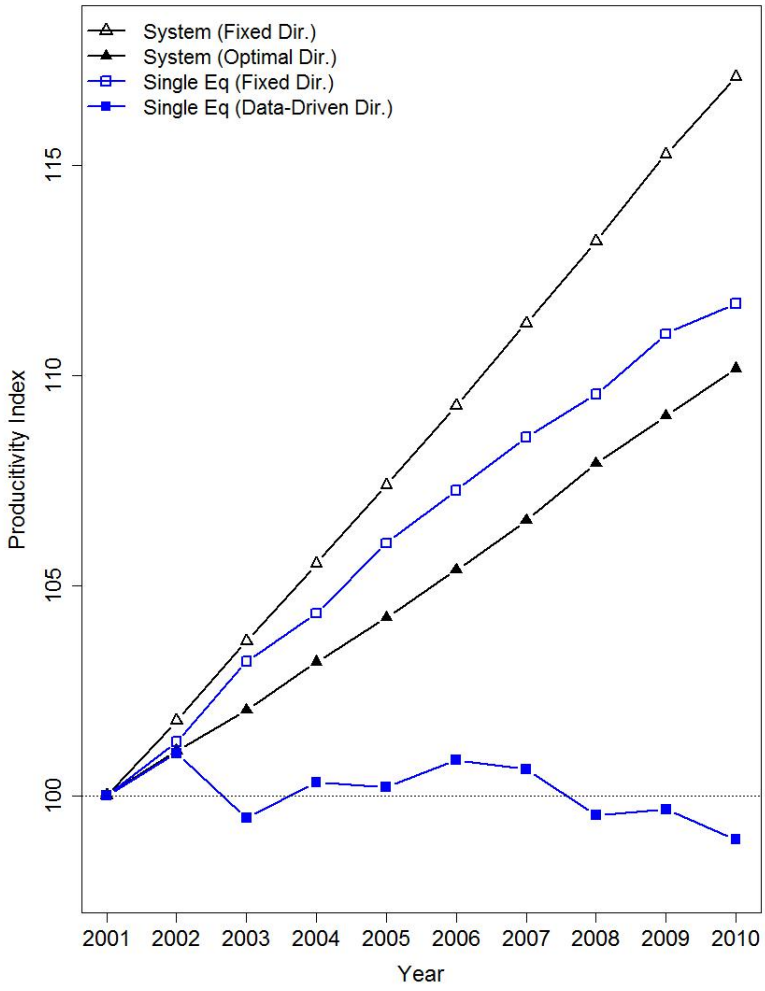

Figure 1: Posterior Estimates of Asset-Weighted Productivity Indices across All Models

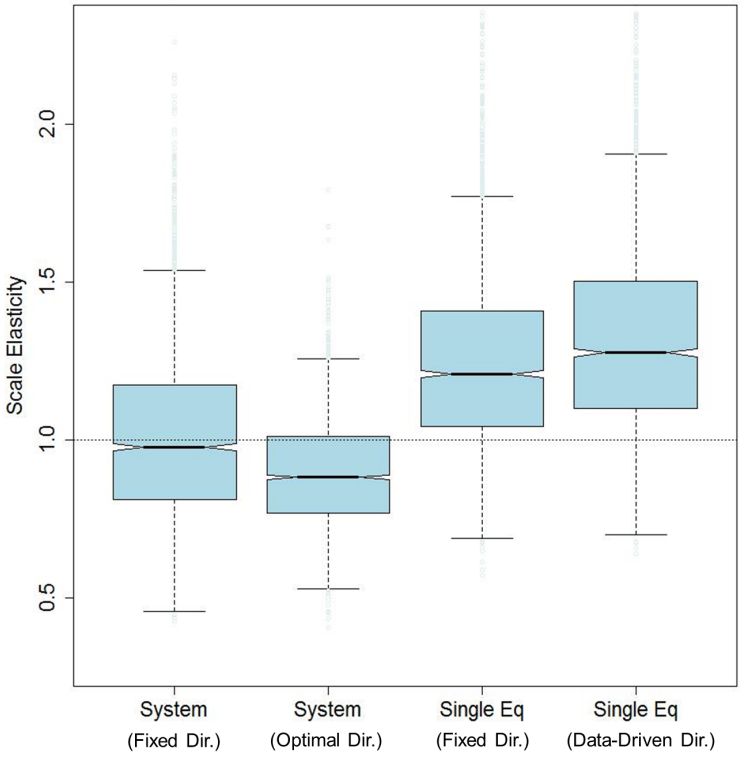

Figure 2: Posterior Scale Elasticity Estimates

We next consider the posterior estimates of bank's scale elasticity. Table 2 reports summary statistics for desirable scale elasticity measures across all models, whereas Figure 2 presents box-plots of their distributions. Regardless of the direction used, the single-equation-based model indicates that, on average, banks exhibit increasing returns to desirable scale with the mean and median posterior estimates of DSE being well above one, which is consistent with some recent estimates reported in the literature (e.g., Hughes and Mester, 2013; Malikov et al., 2015b). However, when controlling for the endogeneity of inputs, the scale elasticity estimates considerably decline in their magnitudes, as can clearly be seen in Figure 2. The mean posterior estimates of DSE from the system-based model are 1.02 and 0.90 when estimated in the pre-specified and cost-optimal directions, respectively. While the two models produce seemingly contradictory findings, respectively suggesting slightly increasing and decreasing returns to scale, the 95% credible intervals for DSE from both models (including our preferred one) do include unity which suggests that banks are generally *invariant* to desirable scale, i.e., they enjoy constant returns to scale. Feng and Zhang (2012) report similar Bayesian estimates of scale economies for large U.S. banks.

In addition to scale elasticities, we also report posterior estimates of the elasticity of the bank's DTDF with respect to an undesirable output (E_b), i.e., $\partial \log \vec{D}_\tau(\cdot)/\partial \log b \geq 0$. Due to duality (Chambers et al., 1996), the latter provides a measure of the cost elasticity with respect to nonperforming loans. Figure 3 plots kernel densities of the posterior distributions of E_b across all models. The kernel densities are constructed using the second-order Gaussian kernel with the optimal bandwidth selected via the data-driven least-squares cross-validation. When using the pre-specified directional vectors, we find that the single-equation-based estimates of E_b tend to be considerably lower than those obtained using a cost system approach: a posterior mean of 0.17 vs. 0.27 (also see

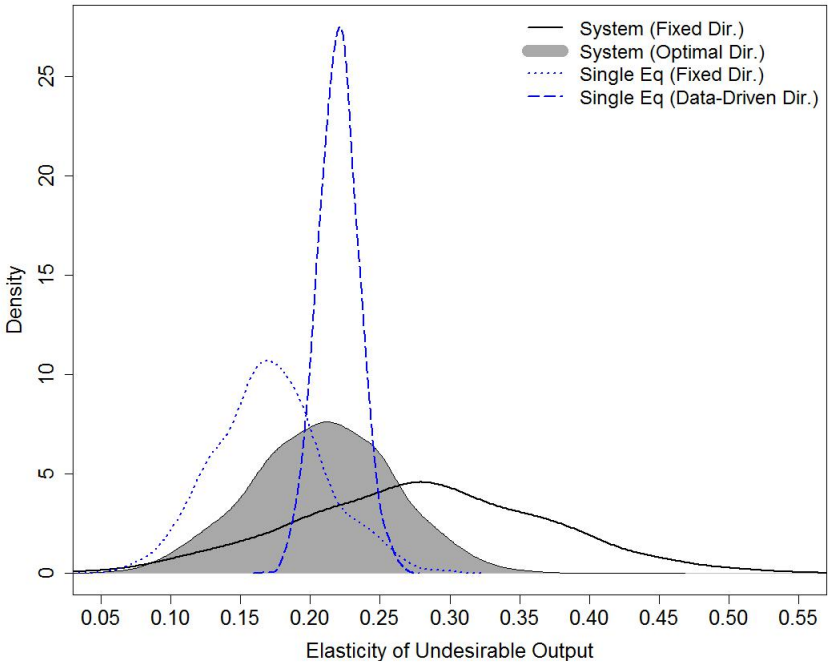

Figure 3: Posterior Estimates of Undesirable Output Elasticity

Table 2). However, letting the data select the direction closes the gap between the two models.

To conclude, we document dramatic distortions in banks' efficiency, productivity growth and scale elasticity estimates obtained from the DTDF estimated via a traditional single-equation approach, which leaves the endogeneity of inputs unaddressed. In studies where exogeneity of inputs is hardly plausible, such as studies of banking production and productivity, a cost system approach, which we offer in this paper, is likely to provide a more robust estimation strategy.

7 Conclusion

This paper addresses some econometric issues related to a consistent estimation of banking technology which explicitly accommodates the presence of *undesirable* nonperforming loans (NPL) — an inherent characteristic of the bank's production due to its exposure to credit risk. Specifically, we model NPL as an undesirable output in the bank's production process. This approach is advantageous over modeling NPL as a mere banking technology shifter (i.e., a contextual control variable) because it *(i)* recognizes that NPL is a *by*-product of producing desirable outputs, such as earning loans and securities, and *(ii)* accommodates the *undesirability* of NPL. By acknowledging that nonperforming loans are undesirable, we are able to credit banks for the reduction in NPL along with the expansion in desirable outputs when computing their productivity and technological efficiency.

We formulate banking technology using the directional technology distance function (DTDF) generalized to the case of undesirable outputs, which we treat as a *stochastic* function. The estimation of such a stochastic DTDF is however not trivial due to the potential simultaneity of inputs. We propose addressing this endogeneity of inputs from the perspective of economic theory. More specifically, we suggest invoking the assumption of the bank's cost minimizing behavior not only to justify the treatment of outputs as exogenous (as commonly done in the literature) but to also tackle

the endogeneity of inputs in the DTDF. We do so by augmenting the stochastic DTDF with the set of independent (nonlinear) first-order conditions from the bank's cost minimization problem, which we then estimate as a system of simultaneous equations. Our identification strategy thus relies on competitively determined input prices as a source of exogenous variation.

We note that there are advantages to using our cost system approach even if the DTDF does not suffer from the endogeneity of inputs. Since additional equations (the first-order conditions) do not contain any extra parameters, the system-based parameter estimates are likely to be more precise. Furthermore, technological metrics obtained from the cost system of DTDF are likely to be more meaningful because the economic behavior is embedded into the system through the first-order conditions. The inclusion of the cost-minimizing first-order conditions in the system permits us to also estimate the "cost-optimal" directional vector for the banking DTDF. That is, in contrast to a common tradition in the literature, we do not pre-specify the (fixed) directional vector for the DTDF but rather let the data help us determine the direction in which the bank's movement toward the stochastic frontier is to be estimated. The estimated optimal direction captures the bank's movement to the point on a technological frontier where costs are minimized, thus eliminating the uncertainty associated with an *ad hoc* choice of the direction.

We apply our cost system approach to the data on U.S. commercial banks in 2001–2010. The nonlinear cost system is estimated in two stages via Bayesian methods subject to theoretical regularity conditions. We document dramatic sensitivity of the estimates of banks' various technological metrics to controlling for the endogeneity of inputs and the choice of the directional vector. We conclude that in studies of banking technology, where exogeneity of inputs is hardly plausible, a cost system approach is likely to provide a more robust estimation strategy.

Lastly, we would like to note that our cost system approach to tackling the endogeneity of inputs in the stochastic DTDF is not limited to the banking application only. It can be applied to any other productivity study, where the endogeneity of inputs may be of concern while exogeneity of outputs is justified on the basis of market regulations and/or the nature of outputs.

References

Altunbas, Y., Liu, M-H, M. P., and Seth, R. (2000). Efficiency and risk in Japanese banking. *Journal of Banking & Finance*, 24:1605–1628.

Assaf, A. G., Matousek, R., and Tsionas, E. G. (2013). Turkish bank efficiency: Bayesian estimation with undesirable outputs. *Journal of Banking & Finance*, 37:506–517.

Atkinson, S. E., Primont, D., and Tsionas, E. G. (2014). Estimation of efficient production using optimal directions with multiple inputs and outputs. Working Paper, Southern Illinois University.

Atkinson, S. E. and Tsionas, E. G. (2015). Directional distance functions: Optimal endogenous directions. *Journal of Econometrics*. forthcoming.

Barnett, W. A. (2002). Tastes and technology: Curvature is not sufficient for regularity. *Journal of Econometrics*, 108:199–202.

Barnett, W. A., Geweke, J., and Wolfe, M. (1991). Seminonparametric Bayesian estimation of the Asymptotically Ideal Production Model. *Journal of Econometrics*, 49:5–50.

Berger, A. N. (2004). The economic effects of technological progress: Evidence from the banking industry. *Journal of Money, Credit, and Banking*, 35(2):141–176.

Berger, A. N., Demsetz, R. S., and Strahan, P. E. (1999). The consolidation of the financial services industry: Causes, consequences, and the implications for the future. *Journal of Banking & Finance*, 23(2–4):135–194.

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.

Chambers, R. G., Chung, Y., and Färe, R. (1996). Benefit and distance functions. *Journal of Economic Theory*, 70:407–419.

Chambers, R. G., Chung, Y., and Färe, R. (1998). Profit, directional distance functions, and Nerlovian efficiency. *Journal of Optimization Theory and Applications*, 98:351–364.

Chung, Y., Färe, R., and Grosskopf, S. (1997). Productivity and undesirable outputs: A directional distance function approach. *Journal of Environmental Management*, 51:229–240.

Färe, R., Grosskopf, S., Noh, D.-W., and Weber, W. (2005). Characteristics of a polluting technology: Theory and practice. *Journal of Econometrics*, 126:469–492.

Färe, R., Grosskopf, S., and Whittaker, G. (2013). Directional output distance functions: Endogenous directions based on exogenous normalization constraints. *Journal of Productivity Analysis*, 40:267–269.

Färe, R. and Primont, D. (1995). *Multi-Output Production and Duality: Theory and Applications*. Kluwer Academic Publishers, Norwell.

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.

Feng, G. and Serletis, A. (2010). Efficiency, technical change, and returns to scale in large US banks: Panel data evidence from an output distance function satisfying theoretical regularity. *Journal of Banking & Finance*, 34:127–138.

Feng, G. and Serletis, A. (2014). Undesirable outputs and a primal Divisia productivity index based on the directional output distance function. *Journal of Econometrics*. forthcoming.

Feng, G. and Zhang, X. (2012). Productivity and efficiency at large and community banks in the US: A Bayesian true random effects stochastic distance frontier analysis. *Journal of Banking & Finance*, 36(7):1883–1895.

Freixas, X. and Rochet, J. (2008). *Microeconomics of Banking*. The MIT Press, Cambridge.

Guarda, P., Rouabah, A., and Vardanyan, M. (2013). Identifying bank outputs and inputs with a directional technology distance function. *Journal of Productivity Analysis*, 40:185–195.

Hampf, B. and Krüger, J. J. (2014). Optimal directions for directional distance functions: An exploration of potential reductions of greenhouse gases. *American Journal of Agricultural Economics*. forthcoming.

Hudgins, L. B. and Primont, D. (2007). Derivative properties of directional technology distance functions. In Färe, R., Grosskopf, S., and Primont, D., editors, *Aggregation, Efficiency, and Measurement*. Springer.

Hughes, J. P. and Mester, L. J. (1993). A quality and risk-adjusted cost function for banks: Evidence on the "too-big-to-fail doctrine. *Journal of Productivity Analysis*, 4(3):293–315.

Hughes, J. P. and Mester, L. J. (1998). Bank capitalization and cost: Evidence of scale economies in risk management and signaling. *Review of Economics and Statistics*, 80(2):314–325.

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, 1 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.

Koutsomanoli-Filippaki, A., Margaritis, D., and Staikouras, C. (2009). Efficiency and productivity growth in the banking industry of Central and Eastern Europe. *Journal of Banking & Finance*, 33:557–567.

Kumbhakar, S. C., Wang, H.-J., and Horncastle, A. P. (2015). *A Practitioner's Guide to Stochastic Frontier Analysis using Stata*. Cambridge University Press, Cambridge.

Malikov, E., Kumbhakar, S. C., and Tsionas, E. G. (2015a). Bayesian approach to disentangling technical and environmental productivity. *Econometrics*, 3(2):443–465.

Malikov, E., Restrepo-Tobón, D., and Kumbhakar, S. C. (2015b). Estimation of banking technology under credit uncertainty. *Empirical Economics*, 49(1):185–211.

Park, K. H. and Weber, W. L. (2006). A note on efficiency and productivity growth in the Korean banking industry, 1992–2002. *Journal of Banking & Finance*, 30:2371–2386.

Restrepo-Tobón, D. A. and Kumbhakar, S. C. (2015). Nonparametric estimation of returns to scale using input distance functions: An application to large U.S. banks. *Empirical Economics*, 48(1):143–168.

Schmidt, P. and Lovell, C. A. K. (1979). Estimating technical and allocative inefficiency relative to stochastic production and cost frontiers. *Journal of Econometrics*, 9:343–366.

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.

Tsionas, E. G., Malikov, E., and Kumbhakar, S. C. (2014). Do direction, normalization and the Jacobian matter in the productivity measurements based on the directional distance function? Working Paper, Binghamton University.

Zelenyuk, V. (2013). A scale elasticity measure for directional distance function and its dual: Theory and DEA application. *European Journal of Operational Research*, 228:592–600.