

Efficiency, Technical Change, and Returns to Scale in Large U.S. Banks: Panel Data Evidence Based on Bayesian Estimation of the Output Distance Function

Guohua Feng and Apostolos Serletis

Department of Economics

University of Calgary

Calgary, Alberta, T2N 1N4

Dissertation Chapter Three

Abstract

This paper provides parametric estimates of technical change, efficiency change, economies of scale, and total factor productivity growth for large banks (those with assets in excess of \$1 billion) in the United States, over the period from 2000 to 2005. In doing so, we propose a distance function based primal total factor productivity growth index, which is valid under both perfect and imperfect competition, and estimate the output distance function, subject to theoretical regularity, within a Bayesian framework. The results show that total factor productivity of the large U.S. banks grew at an average rate of 1.98% over the sample period. However, our estimates also show a clear downward trend in the growth rate of total factor productivity and our decomposition of the primal Divisia total factor productivity growth index into its three components — technical change, efficiency change, and economies of scale — indicates that technical change is the driving force behind this decline.

JEL classification: G21; G28; D24; E58; C11.

Keywords: Bank; Productivity; Efficiency; Technical Change; Returns to Scale; Output distance function.

1 Introduction

In the last 25 years, fundamental regulatory changes together with technological and financial innovations have greatly transformed the commercial banking industry in the United States. Regarding regulation, major changes include the removal of geographic restrictions and the permission of combinations of banks, securities firms, and insurance companies — for a complete list of regulatory changes, see Jones and Critchfield (2005). On the other hand, the industry has widely adopted various innovations in technology and applied finance. These technological and financial innovations include (but are not limited to) information processing and telecommunication technologies, the securitization and sale of bank loans, and the development of derivatives markets — see Berger *et al.* (1995) and Berger (2004) for more details. One of the most important consequences of these regulatory changes and technological and financial innovations has been financial consolidation, leading to larger and more complex banking organizations — see, for example, Berger (2004) and Jones and Critchfield (2005). In fact, according to Jones and Critchfield (2005), the asset share of large banks in the United States (those with more than \$10 billion in assets) increased dramatically from 42 percent in 1984 to 73 percent in 2003.

This raises the issue of whether the recent transformation of the U.S. banking industry has made the industry more productive. In particular, has the adoption of technological and financial innovations caused any shift in the production frontier or more generally the best practice frontier of the banking industry (technical change)? Have legislative and regulatory changes increased the ability of banks to produce more output from a given set of inputs with existing technology (efficiency change)? Finally, has the increased concentration of industry assets among the very large banks brought these banks closer to their optimal output levels (economies of scale)? These interesting questions have been partially investigated in previous studies with data prior to 2000 — see, for example, Stiroh (2000), Alam (2001), and Berger and Mester (2003).

The purpose of this paper is to contribute to this literature (more generally the productivity analysis literature), by proposing a new productivity index, applying it to more recent data, and building on recent work by Feng and Serletis (2008) paying particular attention to the theoretical regularity conditions of neoclassical microeconomic theory. More specifically, we have three objectives in this paper. To propose a new distance function based primal productivity index which is suitable for both perfectly and imperfectly competitive markets and which can also be further decomposed into technical change, efficiency change, and economies of scale components. To slightly modify the O'Donnell and Coelli (2005) Bayesian method of imposing nonlinear constraints on the distance function to guarantee the economic meaningfulness of the new primal productivity index, and finally to apply the new productivity index to large U.S. banks using recent panel data over the period from 2000 to 2005.

The literature on productivity growth has been dominated by two methodologies —

the nonparametric Malmquist index approach [see Färe *et al.* (1994)] and the parametric stochastic frontier approach. For a comprehensive review of the different approaches to productivity measurement, see Feng and Serletis (2007). The nonparametric Malmquist index approach involves fitting distance functions to data on input and output quantities using the nonparametric, linear programming techniques of data envelopment analysis. This approach has two major advantages: it does not require behavioral assumptions and it does not require information on prices. The latter advantage is especially desirable for the study of productivity growth for sectors where price information is missing or distorted — for example, infrastructure, public sectors, regulated industries, and industries with pollutants. It is also very useful in the case where price information cannot be obtained as accurately as quantity information. Taking the studies on banking productivity and efficiency, for example, most of them assume that the input market is competitive and then take a cost function approach. Moreover, in the measurement of input prices, almost all existing studies use the actual prices paid by banks (i.e. dividing expenses by the stock of inputs) to proxy the prevailing market prices — one exception is Berger and Mester (2003) who use market average prices (i.e. the weighted average of the prices of the other banks in the market excluding the bank’s own price). However, actual input prices paid by banks vary greatly across banks, contradicting the assumption of a competitive input market. Thus, a primal measure of productivity growth, i.e. the nonparametric Malmquist index approach, becomes more appealing in this case.

However, the nonparametric Malmquist index approach suffers from several drawbacks. It assumes away any measurement error and so could potentially suffer from outliers. It cannot provide deep insights into important production structures (i.e. substitution elasticities), since it is nonparametric. More importantly, it has problems in measuring the contribution of scale economies, which has been proved to have important implications for market structure. Färe *et al.* (1994) imposed a constant returns to scale restriction on the frontier technology and used a variable returns to scale technology only when further decomposition of efficiency change was needed. Although Ray and Desli (1997) disagreed with Färe *et al.* (1994) on the roles of the constant returns to scale and variable returns to scale frontiers in the decomposition of productivity change indexes, the alternative method they proposed also necessitated the constant returns to scale assumption in computing the overall productivity change index — see Ray and Desli (1997) and Atkinson *et al.* (2003).

The stochastic frontier approach, based on the ideas of Aigner *et al.* (1977) and Meeusen and van den Broeck (1977), involves the estimation of parametric production, cost, or profit frontiers with a composite error term consisting of nonnegative inefficiency and noise components. With this approach, the contribution of scale economies can be easily identified. For example, Bauer (1990), using a production frontier and a cost frontier, successfully decomposed productivity growth into three components including a scale effect component. Kumbhakar and Lovell (2000) further analyzed the case of multiple outputs using cost and profit frontiers and decomposed productivity growth into even more components. However,

the decomposition of productivity growth using a production frontier suffers from the problem of not allowing for multiple output analysis. Thus, it is not suitable for the study of many industries (such as, for example, banking, agriculture, and telecommunications), where multiple outputs is a common feature of the production process. Moreover, the decomposition of productivity growth using cost and profit frontiers involves the use of prices, thus losing its appeal in many situations where information on prices is missing, distorted, or inaccurate.

In this paper, we extend and combine the best elements of the non-parametric approach and the parametric approach, and propose a distance-function based primal Divisia total factor productivity growth index. In particular, under the assumption of perfect competition, and by solving the problem of profit maximization subject to the output distance function being less than or equal to one, we replace in the conventional Divisia total factor productivity growth index the observed revenue shares by quantity-based shadow revenue shares and the observed cost shares by quantity-based shadow cost shares. We also show that this primal Divisia total factor productivity growth index obtained from the problem of profit maximization under perfect competition is also valid in the presence of imperfect competition. In this case, the obtained primal Divisia total factor productivity growth index is equal to a markup and markdown adjusted dual Divisia total factor productivity growth index, which reflects the firm's true marginal revenue and marginal costs. Then based on the primal Divisia total factor productivity growth index, we decompose the growth rate in total factor productivity into three components — technical change, efficiency change, and a scale effect. Due to its parametric nature, the proposed primal Divisia total factor productivity growth index does not suffer from the problem of not allowing for the scale effect or the problem of lacking deep insights into production structures, as the nonparametric Malmquist index approach does. At the same time, the primal Divisia total factor productivity growth index requires only information on input and output quantities, and thus can be widely used in sectors where information on prices is missing or distorted.

We also pay explicit attention to theoretical regularity. We show that for the primal multi-output Divisia total factor productivity growth index to be economically meaningful (that is, each of the shadow revenue shares and cost shares to be non-negative and the sum of the revenue/cost shares to be equal to unity), certain regularity conditions have to be imposed. In particular, we show that the non-negativity of the shadow revenue and cost shares can be guaranteed by the monotonicity conditions of the output distance function (i.e. the output distance function is non-decreasing in outputs and non-increasing in inputs), and that the unity sum of shadow revenue shares can be guaranteed by the linear homogeneity of the output distance function in outputs. As our empirical results show, the non-negativity of the shadow shares and unity sum of the shadow revenue shares cannot be automatically satisfied unless the regularity conditions are imposed on the output distance function. This suggests that an estimation method that is capable of imposing regularity conditions has to be employed.

In this regard, we use Bayesian methods to estimate a parametric translog (locally flexible) output distance function. The Bayesian approach has two major advantages that traditional econometric methods (such as the maximum likelihood method, the least squares dummy variables method, and the generalized least squares method) commonly used for productivity estimation do not possess. First, the Bayesian approach provides exact (small-sample) inference on the productivity components (i.e. firm efficiency, technical change, and returns to scale) whereas the traditional methods provide only point estimates of the productivity components without statistical inference. This is so, because there is no way to calculate the probability density function of those productivity components with traditional methods, since they are generally all nonlinear functions of the estimated parameters. Second, and even more importantly, the Bayesian approach allows us to incorporate the theoretical regularity restrictions of neoclassical microeconomic theory in the estimation. As discussed above, the imposition of regularity conditions is particularly important in this study to ensure that the shadow revenue and cost shares are economically meaningful. This can be done either by using the accept-reject algorithm — see Terrel (1996) — or the Metropolis-Hastings algorithm — see Griffiths *et al.* (2000) — within a Bayesian framework. It is to be noted that the imposition of theoretical regularity is beyond all of the simple traditional methods. Although we can reformulate those traditional methods within a constrained optimization method, as in Gallant and Golub (1984), in order to impose theoretical regularity, our experience shows that obtaining statistical inference is still a big problem with the constrained optimization method. Due to the complexities associated with the imposition of theoretical regularity, the vast majority of studies using traditional econometric methods in the productivity analysis literature have failed to incorporate theoretical regularity.

The rest of the paper is organized as follows. In Section 2, we derive a primal multiple output Divisia total factor productivity growth index, which is valid under both perfect competition and imperfect competition, and specify the conditions this index has to satisfy. We also present a decomposition of the growth rate of total factor productivity, isolating the separate contributions of scale economies, technical change, and technical efficiency change. In Section 3 we present the translog output distance function and specify the homogeneity, monotonicity, and curvature constraints required by the decomposition of the primal Divisia total factor productivity growth index. In Section 4 we discuss Bayesian estimation procedures for imposing theoretical regularity on the parameters of the translog output distance function. Section 5 deals with data issues. In Section 6 we apply our methodology to a panel data of 292 large banks in the United States, discuss the effects of incorporating monotonicity and curvature, and also report our estimates of total factor productivity growth and its components. The last section summarizes and concludes the paper.

2 Theoretical Framework

2.1 The Output Distance Function

For each input vector, $\mathbf{x}^t \in R_+^N$ at time t , let $P^t(\mathbf{x}^t)$ be the set of feasible outputs (or production possibilities set)

$$P^t(\mathbf{x}^t) = \{\mathbf{y}^t \in R_+^M : \mathbf{y} \text{ is producible from } \mathbf{x}\}.$$

Following Shephard (1970), we can define the output distance function relative to the output set as follows

$$D_o^t(\mathbf{y}^t, \mathbf{x}^t) = \inf_{\theta} \left\{ \theta > 0 : \frac{\mathbf{y}^t}{\theta} \in P^t(\mathbf{x}^t) \right\}. \quad (1)$$

Thus, for any output quantity vector, \mathbf{y}^t , at time t

$$\frac{\mathbf{y}^t}{D_o^t(\mathbf{y}^t, \mathbf{x}^t)},$$

is the largest output quantity vector on the ray from the origin through \mathbf{y}^t that can be produced by \mathbf{x}^t . In the case of a single output ($M = 1$),

$$F^t(\mathbf{x}^t) \equiv \frac{\mathbf{y}^t}{D_o^t(\mathbf{y}^t, \mathbf{x}^t)},$$

is the familiar production function, implying that $D_o^t(\mathbf{y}^t, \mathbf{x}^t)$ is just the ratio of the observed output \mathbf{y}^t to the maximal output $F^t(\mathbf{x}^t)$.

Output distance functions are non-decreasing, convex and linearly homogeneous in outputs, and non-increasing and quasi-convex in inputs — see Fare and Grosskopf (1994) for more details. From equation (1) it follows that

$$D_o^t(\mathbf{y}^t, \mathbf{x}^t) \leq 1. \quad (2)$$

In (2), the equality holds only if \mathbf{y}^t is on the output isoquants, which are given by

$$\text{Isoq}P^t(\mathbf{x}^t) \equiv \left\{ \mathbf{y}^t \mid D_o^t(\mathbf{y}^t, \mathbf{x}^t) = 1 \right\}, \quad (3)$$

where $\text{Isoq}P^t(\mathbf{x}^t)$ is the boundary of the output set or production ‘frontier.’

To intuitively motivate the output distance function, we can consider $P^t(\mathbf{x}^t)$ to be like a multi-input and multi-output production function. Then the output distance function represents the distance from the boundary of the output set or production frontier. If \mathbf{y} is on the boundary of the output set, the output distance function is equal to one, implying there is no ‘distance’ from the production frontier. If \mathbf{y} is within the boundary of the

output set, the output distance is less than one, indicating the deviation of the firm from the production frontier or technically ‘best-practice’ production. Hence, the output distance function coincides with the Farrell type output oriented measure of technical efficiency [see Kumbhakar and Lovell (2003)],

$$D_o^t(\mathbf{x}^t, \mathbf{y}^t) = TE_o^t(\mathbf{x}^t, \mathbf{y}^t).$$

A unity value of the output distance function indicates that the firm is operating at full technical efficiency level, and a value less than one indicates that the firm is operating with technical inefficiency.

To facilitate the calculation of technical change, we follow the common practice in the empirical literature and model the effect of time through an exogenous time variable, t . Thus, the output distance function defined in (1) can be rewritten as $D_o(\mathbf{x}, \mathbf{y}, t)$, which we will use throughout this paper. Deviation of the output distance function from one, due to technical inefficiency, can be accommodated as follows,

$$D_o(\mathbf{x}, \mathbf{y}, t)\psi(t) = 1, \quad (4)$$

where $\psi(t)$ is a function of a random variable, u , which will be discussed in more detail in Section 3. Equation (4) will be used below in the decomposition of productivity growth. Also, after specifying functional forms for $D_o(\mathbf{x}, \mathbf{y}, t)$ and $\psi(t)$, equation (4) will be econometrically estimated.

2.2 A Primal Divisia TFP Growth Index

2.2.1 Perfect Competition

We start by assuming that the markets for both outputs and inputs are perfectly competitive (that is, price-taking behavior in both markets). In this case, prices for outputs and inputs are exogenous. When all these prices are accurately available, total factor productivity growth for banks can be easily obtained from the conventional dual total factor productivity growth index [see Jorgenson and Griliches (1967)]:

$$\left. \frac{d \ln TFP}{dt} \right|_{\text{Dual}} \equiv \sum_{m=1}^M \tilde{s}_m \dot{y}_m - \sum_{n=1}^N s_n \dot{x}_n, \quad (5)$$

where x_n denotes input n , y_m denotes output m , and a dot over a variable indicates the growth (or change) rate of the variable — for example, $\dot{y} = d \ln y / dt$. Also, in equation (5), $\tilde{s}_m = p_m y_m / \sum_{m=1}^M p_m y_m$ denotes the observed revenue share of output y_m and $s_n = w_n x_n / \sum_{n=1}^N w_n x_n$ the observed cost share of input x_n . $\mathbf{w} = (w_1, \dots, w_n)$ and $\mathbf{p} = (p_1, \dots, p_n)$ are price vectors for inputs and outputs, respectively. In (5), the first term is a Divisia index

of real output growth and the second a Divisia index of real input growth. The Divisia total factor productivity growth index has been widely used in productivity research. In the special case of a single output, it is just the Solow (1957) residual.

However, there are many situations where information on prices is missing, distorted or inaccurate, as we noted above. In those cases, productivity growth has to be calculated by resorting to the primal approach — an approach that relies only on quantity information. This means that the price information required for the calculation of the dual Divisia total factor productivity growth index has to be replaced by quantity information. This can be done in many ways such as, for example, by exploiting the duality between the revenue function and the output distance function [see Shephard (1970)] and the duality between the output distance function and the indirect output distance function or cost function [see Färe and Primont (1990)]. But banks are generally assumed to be profit maximizing firms. To be consistent with this assumption, we replace the price information in (5) with quantity information by solving the following profit maximization problem in perfectly competitive markets.

$$\pi = \max_{\{y, x\}} \left\{ \sum_{m=1}^M p_m y_m - \sum_{n=1}^N w_n x_n : D_o(\mathbf{y}, \mathbf{x}, t) \psi(t) = 1 \right\}, \quad (6)$$

where the constraint is equivalent to $D_o(\mathbf{y}, \mathbf{x}, t) \leq 1$, which completely represents the firm's technology — see, for example, Färe and Primont (1990). The duality between the profit function and the output distance function under the assumption of perfect competition is discussed in Färe and Primont (1995, p. 129) and Kumbhakar and Lovell (2003, p. 206), and used in Brümmer *et al.* (2002) in the literature of agricultural economics.

The first-order conditions corresponding to output, y_m , are

$$p_m = \mu \frac{\partial D_o(\mathbf{y}, \mathbf{x}, t)}{\partial y_m} \psi(t), \quad m = 1, \dots, M, \quad (7)$$

where μ is the Lagrange multiplier. Multiplying both sides of (7) with $y_m/D_o(\mathbf{y}, \mathbf{x}, t)$ and rearranging yields

$$\frac{p_m \mathbf{y}_m}{D_o(\mathbf{y}, \mathbf{x}, t)} = \mu \psi(t) \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_m}, \quad m = 1, \dots, M, \quad (8)$$

Summing up the M equations in (8) yields

$$\sum_{i=1}^M \frac{p_i \mathbf{y}_i}{D_o(\mathbf{y}, \mathbf{x}, t)} = \mu \psi(t) \sum_{i=1}^M \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_i}, \quad (9)$$

since the output distance function is linearly homogeneous in \mathbf{y} and $D_o(y_m(\mathbf{p}, \mathbf{x}, t), \mathbf{x}, t) \psi(t) = 1$. Noting that $\sum_{i=1}^M \partial \ln D_o(\mathbf{y}, \mathbf{x}, t) / \partial \ln y_m = 1$ by linear homogeneity of the output function in outputs [see equation (31) below], we divide (8) by (9) to obtain

$$\frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_m} = \frac{p_m y_m}{R} = \tilde{s}_m, \quad m = 1, \dots, M, \quad (10)$$

according to which the observed revenue share for the m th output, \tilde{s}_m , is equivalent to the elasticity of the distance function with respect to the m th output, $\partial \ln D_o(\mathbf{y}, \mathbf{x}, t) / \partial \ln y_m$ under perfect competition and instantaneous adjustment when they are evaluated at the same point. In fact, in this case the elasticity of the distance function with respect to output is a shadow measure of the revenue share. However, the equivalency between the actual and shadow revenue shares will not hold with imperfect competition, as will be elaborated in the next subsection.

A similar procedure can be applied to the inputs. The first-order conditions corresponding to inputs are

$$w_n = \mu\psi(t) \frac{\partial D_o(\mathbf{y}, \mathbf{x}, t)}{\partial x_n}, \quad n = 1, \dots, N, \quad (11)$$

where μ is the Lagrange multiplier. Multiplying both sides of (11) with $x_n/D_o(\mathbf{y}, \mathbf{x}, t)$ and rearranging yields

$$\frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln x_n} = \frac{1}{\mu\psi(t)} \frac{w_n x_n}{D_o(\mathbf{y}, \mathbf{x}, t)}, \quad n = 1, \dots, N. \quad (12)$$

Summing up the N equations in (12) yields

$$\sum_{n=1}^N \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln x_n} = \frac{1}{\mu\psi(t)} \sum_{n=1}^N \frac{w_n x_n}{D_o(\mathbf{y}, \mathbf{x}, t)}. \quad (13)$$

Dividing (12) by (13) yields (for $n = 1, \dots, N$)

$$\frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln x_n} \frac{1}{\sum_{n=1}^N \partial \ln D_o(\mathbf{y}, \mathbf{x}, t) / \partial \ln x_n} = \frac{w_n x_n}{\sum_{n=1}^N w_n x_n} = s_n, \quad (14)$$

according to which the observed cost shares can be replaced by their corresponding normalized elasticities of the output distance function with respect to inputs. In fact, the left hand side of (13) is actually the shadow cost share.

Substituting (10) and (14) in (5) yields a primal measure of the Divisia total factor productivity growth index which needs only quantity information

$$\left. \frac{d \ln TFP}{dt} \right|_{\text{Primal}} = \sum_{m=1}^M \tilde{\omega}_m \dot{y}_m - \sum_{n=1}^N \omega_n \dot{x}_n, \quad (15)$$

where

$$\tilde{\omega}_m = \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_m}, \quad (16)$$

is the shadow revenue share for output m , and

$$\omega_n = \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t) / \partial \ln x_n}{\sum_{n=1}^N \partial \ln D_o(\mathbf{y}, \mathbf{x}, t) / \partial \ln x_n} \quad (17)$$

is the shadow cost share for input n . To further simplify the notation in (17), we define

$$\varepsilon_n = \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln x_n}, \quad (18)$$

and

$$\varepsilon = - \sum_{n=1}^N \varepsilon_n,$$

so that ω_n in equation (18) can thus be rewritten as

$$\omega_n = - \frac{\varepsilon_n}{\varepsilon}$$

where ε has been shown by Fare and Grosskopf (1994, p. 103) to be the returns to scale (RTS) in terms of the output distance function.

2.2.2 Imperfect Competition

While most studies on banking productivity and efficiency assume that the market for bank services (output market) is perfectly competitive, some empirical studies show that monopolistic competition is more appropriate for the banking industry in most countries — see, for example, Bikker and Haaf (2002) and Claessens and Laeven (2003). In particular, one widely used technique to empirically measure the degree of competitive behavior in the market is the H statistic, developed by Panzar and Rosse (1987). In particular, the H statistic is used to measure the elasticity of revenue with respect to input prices. $H = 1$ implies perfect competition, $H = 0$ indicates perfect collusion, and $0 < H < 1$ indicates monopolistic competition; values less than 0 are also consistent with perfect collusion. Both Bikker and Haaf (2002) and Claessens and Laeven (2003) have found the H statistic for the U.S. banking industry to be around 0.5, indicating that the U.S. market for bank services is characterized by monopolistic competition.

With this in mind, a natural question to ask is whether the primal Divisia total factor productivity growth index obtained under the assumption of perfect competition is the correct measure of productivity growth in the presence of imperfect competition. To address this question, in what follows we assume that market power is limited to output markets and that input markets are perfectly competitive (the assumption of competitive input markets

can be relaxed without affecting the validity of the primal Divisia total factor productivity growth index, as we shall show below). We assume that each firm (bank) solves the following profit maximization problem

$$\max_y \pi = \left\{ \sum_{m=1}^M p_m(y_m) y_m - C(\mathbf{y}, \mathbf{w}, t) \right\}, \quad (19)$$

where $p_m(y_m)$ is the inverse demand function, and $C(\mathbf{y}, \mathbf{w}, t)$ is obtained from the following first-stage cost minimization problem

$$C(\mathbf{y}, \mathbf{w}, t) = \min_x \{ \mathbf{w}'\mathbf{x} : D_o(\mathbf{y}, \mathbf{x}, t) \psi(t) = 1 \}. \quad (20)$$

The duality between the output distance function and cost function is discussed in Färe and Primont (1990) and Primont and Sawyer (1993).

The first-order conditions corresponding to (19) are

$$p_m(1 - m_m) = \lambda \frac{\partial C(\mathbf{y}, \mathbf{w}, t)}{\partial y_m}, \quad m = 1, \dots, M, \quad (21)$$

where λ is the Lagrange multiplier for the profit maximization problem in (19), and

$$m_m = -\frac{\partial p(y_m) y_m}{\partial y_m p_m} \geq 0,$$

is the nonnegative ad valorem monopolistic markup for the m th output. Applying the envelope theorem to equation (20) with respect to the m th output, we obtain

$$\frac{\partial C(\mathbf{y}, \mathbf{w}, t)}{\partial y_m} = -\tilde{\lambda} \psi(t) \frac{\partial D_o(\mathbf{y}, \mathbf{x}, t)}{\partial y_m}, \quad (22)$$

where $\tilde{\lambda}$ is the Lagrangian multiplier for the cost minimization problem in (20). Substituting (22) into (21) yields

$$p_m(1 - m_m) = -\lambda \tilde{\lambda} \psi(t) \frac{\partial D_o(\mathbf{y}, \mathbf{x}, t)}{\partial y_m}, \quad m = 1, \dots, M. \quad (23)$$

Multiplying both sides of (23) by $y_m/D_o(\mathbf{y}, \mathbf{x}, t)$, yields

$$\frac{p_m(1 - m_m) y_m}{D_o(\mathbf{y}, \mathbf{x}, t)} = -\lambda \tilde{\lambda} \psi(t) \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_m}, \quad m = 1, \dots, M. \quad (24)$$

Summing up the M equations in (24) yields

$$\sum_{i=1}^M \frac{p_m(1 - m_m) y_m}{D_o(\mathbf{y}, \mathbf{x}, t)} = -\lambda \tilde{\lambda} \psi(t) \sum_{i=1}^M \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_m}, \quad m = 1, \dots, M. \quad (25)$$

Noting that $\sum_{i=1}^M \partial \ln D_o(\mathbf{y}, \mathbf{x}, t) / \partial \ln \mathbf{y}_m = 1$, by linear homogeneity of the output function in outputs, and dividing (24) by (25) yields

$$\frac{p_m (1 - m_m) y_m}{\sum_{i=1}^M p_m (1 - m_m) y_m} = \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_m}, \quad (26)$$

according to which the elasticity of the output distance function with respect to the m th output is equivalent to a markup-adjusted revenue share of the m th output under imperfect competition and instantaneous adjustment when they are evaluated at the same point.

Combining (14) and (26) gives

$$\begin{aligned} \left. \frac{d \ln TFP}{dt} \right|_{\text{Primal}} &= \sum_{m=1}^M \tilde{\omega}_m \dot{y}_m - \sum_{n=1}^N \omega_n \dot{x}_n \\ &= \sum_{m=1}^M \frac{p_m (1 - m_m) y_m}{\sum_{i=1}^M p_m (1 - m_m) y_m} \dot{y}_m - \sum_{n=1}^N \frac{w_n x_n}{\sum_{n=1}^N w_n x_n} \dot{x}_n, \end{aligned} \quad (27)$$

where $\tilde{\omega}_m$ and ω_n are defined separately in (16) and (17). According to (27), in the presence of imperfect competition in the output market, the primal Divisia total factor productivity growth index is equal to a markup-adjusted Divisia real output index minus the Divisia real input index. In the special cases where the markups are zero (as in the case with perfect competition), or markups are constant across outputs, or there is only one output, the markup-adjusted dual Divisia total factor productivity growth index reduces to the conventional dual Divisia total factor productivity growth index without markup in (5).

It should be noted that (27) can be easily generalized to the case where market power is present in both output and input markets. In that case

$$\begin{aligned} \left. \frac{d \ln TFP}{dt} \right|_{\text{Primal}} &= \sum_{m=1}^M \tilde{\omega}_m \dot{y}_m - \sum_{n=1}^N \omega_n \dot{x}_n \\ &= \sum_{m=1}^M \frac{p_m (1 - m_m) y_m}{\sum_{i=1}^M p_m (1 - m_m) y_m} \dot{y}_m - \sum_{n=1}^N \frac{w_n (1 - n_n) x_n}{\sum_{i=1}^M w_n (1 - n_n) x_n} \dot{x}_n, \end{aligned} \quad (28)$$

where

$$n_n = -\frac{\partial w(x_n)}{\partial x_n} \frac{x_n}{w_n} \geq 0,$$

is the nonnegative ad valorem monopsony markdown for the n th input. According to (27), in the presence of imperfect competition in the output market, the primal Divisia total factor

productivity growth index is equal to a markup-adjusted Divisia real output index minus a markdown-adjusted Divisia real input index.

The primal Divisia total factor productivity growth index shown in (15) has several advantages. First, like the nonparametric Malmquist productivity index, it does not require price information and thus can be widely used in situations where price information is missing or distorted as, for example, in infrastructure, regulated industries, and industries with pollutants. Second, it is consistent with all types of returns to scale, (i.e. decreasing, constant, and increasing returns to scale) and does not require prior knowledge of the underlying market structure. This is a very desirable property since we don't have to impose returns to scale a priori. In this sense, the primal Divisia total factor productivity growth index is preferable to the nonparametric Malmquist productivity index proposed by Färe *et al.* (1994) where the assumption of returns to scale has to be imposed a priori. Finally, a parametric approach shares many desirable properties with the stochastic frontier approach — foreexample, it allows an easy calculation of the contribution of the scale effect and a deep insight into important production structures.

2.2.3 The Properties of the Primal TFP Growth Index

There is a general consensus among researchers that a total factor productivity growth index should satisfy four desirable properties: identity, separability, monotonicity, and proportionality [see Orea (2002)]. The identity property states that if outputs and inputs do not change, the productivity index should remain unchanged. Clearly, the primal Divisia total factor productivity growth index satisfies this property. The separability property implies that a total factor productivity index can be interpreted in the same way as in the single-output single-input case, for example, as a relationship between an (aggregated) output and an (aggregated) input. As Førsund (1997) pointed out, this property relies on a separability restriction on technology, instead of the formula chosen to construct the productivity index. Consequently, if technology is separable in outputs and inputs, the primal total factor productivity index has this desirable property.

The monotonicity property requires that the primal Divisia total factor productivity index be non-decreasing in the output vector and non-increasing in the input vector. An examination of (15) reveals that the monotonicity property can be satisfied is

$$\frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_m} \geq 0; \tag{29}$$

$$\frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln x_n} \leq 0,$$

which is equivalent to the monotonicity conditions of the output distance function (i.e. $\partial D_o(\mathbf{y}, \mathbf{x}, t) / \partial y_m \geq 0$ and $\partial D_o(\mathbf{y}, \mathbf{x}, t) / \partial x_n \leq 0$), since outputs, inputs, and distance are

all non-negative. Monotonicity violations will give rise to incorrectly signed elasticities, with the perverse implication that productivity can be improved by increasing inputs (decreasing outputs) while holding outputs (inputs) fixed.

The proportionality property means that whenever $(X_{t+1}, Y_{t+1}) = (\lambda X_t, \mu X_t)$, a total factor productivity index (i.e. $TFP = Y/X$ where Y and X are output and input quantity indexes, respectively) should be equal to μ/λ . The primal Divisia total factor productivity growth index will satisfy this property if and only if the shadow revenue/cost shares sum to unity, respectively. To see this, we take the exponential of both sides of the primal Divisia total factor productivity growth index to obtain its corresponding total factor productivity index

$$TFP = \frac{\exp \left[\sum_{m=1}^M \tilde{\omega}_m \ln (y_{m,t+1}/y_{m,t}) \right]}{\exp \left[\sum_{n=1}^N \omega_n \ln (x_{n,t+1}/x_{n,t}) \right]} = \frac{(y_{1,t+1}/y_{1,t})^{\tilde{\omega}_1} \times \dots \times (y_{M,t+1}/y_{M,t})^{\tilde{\omega}_M}}{(x_{1,t+1}/x_{1,t})^{\omega_1} \times \dots \times (x_{M,t+1}/x_{M,t})^{\omega_N}}.$$

From the above equation, it is clear that the proportionality property in our particular case requires

$$\sum_{m=1}^M \tilde{\omega}_m = 1 \text{ and } \sum_{m=1}^M \omega_n = 1. \quad (30)$$

(30) can actually be guaranteed by the linear homogeneity of the output distance function in outputs. Formally,

$$\sum_{m=1}^M \tilde{\omega}_m = \sum_{m=1}^M \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_m} = \sum_{m=1}^M \left(\frac{\partial D_o(\mathbf{y}, \mathbf{x}, t)}{\partial y_m} y_m \right) \frac{1}{D_o(\mathbf{y}, \mathbf{x}, t)} = 1. \quad (31)$$

Moreover, $\sum_{n=1}^N \omega_n = 1$ is also satisfied by definition.

It should be noted at this point that certain theoretical regularity conditions (i.e. non-decreasing, convexity and linearly homogeneity in outputs, and non-increasing and quasi-convexity in inputs) have to be imposed on the parameters of the output distance function. These theoretical regularity conditions are not only used for the validity of the output distance function to completely describe the technology, but also for guaranteeing the economic meaningfulness of the total factor productivity growth index, as shown in (29) and (30). This suggests that an estimation method that is capable of imposing the theoretical regularity conditions has to be employed.

2.3 Decomposition of the Primal Divisia TFP Growth Index

Equations (15), (29), and (30) provide a basic framework for further decomposing the total factor productivity growth index using the output distance function. In particular, totally

differentiating equation (4) with respect to time (after taking logs of both sides) and rearranging yields

$$\sum_{m=1}^M \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_m} \dot{y}_m = -\frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial t} - \frac{d \ln \psi(t)}{dt} - \sum_{n=1}^N \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln x_n} \dot{x}_n. \quad (32)$$

Substituting (32) into (15) yields

$$\left. \frac{d \ln TFP}{dt} \right|_{\text{Primal}} = TC + \Delta TE + SC, \quad (33)$$

where

$$TC = -\partial \ln D_o(\mathbf{y}, \mathbf{x}, t) / \partial t \quad (34)$$

$$\Delta TE = -\partial \ln \psi(t) / \partial t \quad (35)$$

$$SC = (\varepsilon - 1) \sum_{n=1}^N \left(-\frac{\varepsilon_n}{\varepsilon} \right) \dot{x}_n \quad (36)$$

The first term in (33) is a primal measure of the rate of technical change. In terms of the output distance function, it captures the change in the best practice distance frontier which is solely due to the passing of time. In fact, it is a continuous time version of the technical change term in the Malquist productivity index, which measures the shift in technology between the two periods evaluated at x_t and x_{t+1} . The second term is a primal measure of the change in technical efficiency. It represents the rate at which an observed firm is moving towards or away from the frontier. It is positive (negative) as technical efficiency increases (decreases) over time. It should be noted that what matters to productivity growth is not the level of technical efficiency, but its improvement over time. The third term captures the contribution of economies of scale. It is positive when increasing returns to scale prevails ($\varepsilon > 1$ in this case), negative when decreasing returns to scale prevails ($\varepsilon < 1$ in this case), and vanishes when constant returns to scale is present.

3 The Translog Output Distance Function

In order to implement our total factor productivity growth index decomposition, we need to parameterize and calculate the parameters of an output distance function. Here we choose to parameterize $D_o(\mathbf{y}, \mathbf{x}, t)$ as a translog function, which is the functional form often employed

to model bank technology. The translog output distance function, defined over M outputs and N inputs can be written as

$$\begin{aligned}
\ln D_o(\mathbf{y}, \mathbf{x}, t) &= a_0 + \sum_{m=1}^M a_m \ln y_m + \frac{1}{2} \sum_{m=1}^M \sum_{p=1}^M a_{mp} \ln y_m \ln y_p \\
&+ \sum_{n=1}^N b_n \ln x_n + \frac{1}{2} \sum_{n=1}^N \sum_{j=1}^N b_{nj} \ln x_n \ln x_j + \delta_t t + \frac{1}{2} \delta_{tt} t^2 \\
&+ \sum_{n=1}^N \sum_{m=1}^M g_{nm} \ln x_n \ln y_m + \sum_{m=1}^M \delta_{ym} t \ln y_m + \sum_{n=1}^N \delta_{xn} t \ln x_n, \quad (37)
\end{aligned}$$

where t denotes a time trend. Symmetry requires $a_{mp} = a_{pm}$ and $b_{nj} = b_{jn}$. The restrictions required for homogeneity of degree one in outputs are

$$\begin{aligned}
\sum_{m=1}^M a_m &= 1; \\
\sum_{p=1}^M a_{mp} &= 0 \quad \text{for all } m = 1, 2, \dots, M; \\
\sum_{m=1}^M g_{nm} &= 0 \quad \text{for all } n = 1, 2, \dots, N; \\
\sum_{m=1}^M \delta_{ym} &= 0.
\end{aligned}$$

One way of imposing these restrictions is to normalize the function by one of the outputs — see, for example, Lovell *et al.* (1994) and O'Donnell and Coelli (2005). This specific transformation through normalization has the advantage of converting equation (37), which is difficult to estimate directly, into an estimable regression model. We choose the M th output for normalization, which leads to the following expression

$$\ln D_o\left(\frac{\mathbf{y}}{y_M}, \mathbf{x}, t\right) = \ln \left[\frac{1}{y_M} D_o(\mathbf{y}, \mathbf{x}, t) \right].$$

Using the homogeneity restriction, replacing $-\ln D_o(\mathbf{y}, \mathbf{x}, t)$ with $u = \ln(\psi)$, and adding a

random error, v , yields the stochastic output distance function

$$\begin{aligned}
-\ln y_M &= a_0 + \sum_{m=1}^{M-1} a_m \ln \left(\frac{y_m}{y_M} \right) + \frac{1}{2} \sum_{m=1}^{M-1} \sum_{p=1}^{M-1} a_{mp} \ln \left(\frac{y_m}{y_M} \right) \ln \left(\frac{y_p}{y_M} \right) \\
&+ \sum_{n=1}^N b_p \ln x_p + \frac{1}{2} \sum_{n=1}^N \sum_{j=1}^N b_{nj} \ln x_n \ln x_j + \delta_t t + \frac{1}{2} \delta_{tt} t^2 \\
&+ \sum_{n=1}^N \sum_{m=1}^{M-1} g_{nm} \ln x_n \ln \left(\frac{y_m}{y_M} \right) + \sum_{m=1}^{M-1} \delta_{ym} t \ln \left(\frac{y_m}{y_M} \right) + \sum_{n=1}^N \delta_{xn} t \ln x_n + u + v, \quad (38)
\end{aligned}$$

where the v 's are assumed to be independently and identically distributed (iid) as $N(0, \sigma^2)$, intended to capture statistical noise; $u = -\ln D$ is a nonnegative random variable, intended to capture technical inefficiency. We assume that u follows an exponential distribution with scale parameter λ , which we will discuss in more detail in Section 4. Further, we assume that v and u are independent of each other, an assumption we maintain throughout this paper.

Technical efficiency, technical change, and returns to scale can thus be shown, respectively, to be

$$TE = \exp(-u) \quad (39)$$

$$TC = -\frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial t} = -\left(\delta_t + \delta_{tt} t + \sum_{m=1}^M \delta_{ym} \ln y_m + \sum_{n=1}^N \delta_{xn} \ln x_n \right) \quad (40)$$

$$RTS = -\sum_{n=1}^N \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln x_n}. \quad (41)$$

Equation (39) can then be used to obtain efficiency change, $\Delta TE = -du/dt$, and (41) can be used to obtain the scale effect,

$$(\varepsilon - 1) \sum_{n=1}^N \left(-\frac{\varepsilon_n}{\varepsilon} \right) \dot{x}_n.$$

3.1 Monotonicity Constraints

As required by microeconomic theory, the output distance function (37) has to satisfy the theoretical regularity conditions of monotonicity and curvature. Monotonicity requires that

$D_o(\mathbf{y}, \mathbf{x}, t)$ is non-increasing in \mathbf{x} and non-decreasing in \mathbf{y} . That is,

$$\frac{\partial D_o(\mathbf{y}, \mathbf{x}, t)}{\partial x_n} \leq 0 \quad \text{and} \quad \frac{\partial D_o(\mathbf{y}, \mathbf{x}, t)}{\partial y_m} \geq 0,$$

or, equivalently,

$$\frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln x_n} \leq 0 \quad \text{and} \quad \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_m} \geq 0, \quad (42)$$

since $x_n/D_o(\mathbf{y}, \mathbf{x}, t) > 0$ and $y_m/D_o(\mathbf{y}, \mathbf{x}, t) > 0$.

The monotonicity restrictions in (42) are critically important in ensuring that the shadow revenue and cost shares are economically meaningful when decomposing the primal total factor productivity growth index (15), as we discussed above.

We now explicitly produce the monotonicity conditions for the output distance function

$$\begin{aligned} k_n &= \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln x_n} \\ &= b_n + \sum_{j=1}^N b_{nj} \ln x_j + \sum_{m=1}^M g_{nm} \ln y_m + \delta_{xn} t \leq 0, \quad \text{for } n = 1, \dots, N; \end{aligned} \quad (43)$$

$$\begin{aligned} r_m &= \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_m} \\ &= a_m + \sum_{p=1}^M a_{mp} \ln y_p + \sum_{n=1}^N g_{nm} \ln x_n + \delta_{ym} t \geq 0, \quad \text{for } m = 1, \dots, M. \end{aligned} \quad (44)$$

Noting that [see equation (31)]

$$\sum_{m=1}^M \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_m} = 1,$$

the monotonicity condition for the M th output can be also rewritten as

$$1 - \sum_{m=1}^{M-1} \frac{\partial \ln D_o(\mathbf{y}, \mathbf{x}, t)}{\partial \ln y_m} \geq 0.$$

3.2 Curvature Constraints

Curvature requires that the output distance function $D_o(\mathbf{y}, \mathbf{x}, t)$ be quasi-convex in inputs and convex in outputs — see Färe and Grosskopf (1994, p.38). For $D_o(\mathbf{y}, \mathbf{x}, t)$ to be quasi-convex in \mathbf{x} it is sufficient that all the principal minors of the following bordered Hessian matrix

$$\mathbf{F} = \begin{bmatrix} 0 & f_1 & \cdots & f_N \\ f_1 & f_{21} & \cdots & f_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ f_N & f_{N1} & \cdots & f_{NN} \end{bmatrix}$$

are negative, where

$$f_n = \frac{\partial D_o(\mathbf{y}, \mathbf{x}, t)}{\partial x_n} = \frac{k_n D_o(\mathbf{y}, \mathbf{x}, t)}{x_n},$$

and

$$f_{nj} = \frac{\partial^2 D_o(\mathbf{y}, \mathbf{x}, t)}{\partial x_n \partial x_j} = (b_{nj} + k_n k_j - \phi_{nj} k_n) \frac{D_o(\mathbf{y}, \mathbf{x}, t)}{x_n x_j},$$

with $\phi_{nj} = 1$ if $n = j$ and 0 otherwise. Noting that factoring out $D_o(\mathbf{y}, \mathbf{x}, t)/x_n$ from the rows and $1/x_j$ from the columns of \mathbf{F} does not change the signs of its principal minors, we can consider the following matrix

$$\tilde{\mathbf{F}} = \begin{bmatrix} 0 & \tilde{f}_1 & \cdots & \tilde{f}_N \\ \tilde{f}_1 & \tilde{f}_{11} & \cdots & \tilde{f}_{1N} \\ \vdots & \vdots & \cdots & \vdots \\ \tilde{f}_N & \tilde{f}_{N1} & \cdots & \tilde{f}_{NN} \end{bmatrix}$$

where $\tilde{f}_n = k_n$, and $\tilde{f}_{nj} = b_{nj} + k_n k_j - \phi_{nj} k_n$. Thus, for $D_o(\mathbf{y}, \mathbf{x}, t)$ to be quasi-convex in \mathbf{x} it is sufficient that all the principal minors of $\tilde{\mathbf{F}}$ are negative.

Convexity in outputs will be ensured if and only if all the principal minors of the Hessian matrix,

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1M} \\ h_{21} & h_{22} & \cdots & h_{2M} \\ \vdots & \vdots & \cdots & \vdots \\ h_{1M} & h_{2M} & \cdots & h_{MM} \end{bmatrix},$$

are non-negative, where

$$h_{mp} \equiv \frac{\partial^2 D_o(\mathbf{y}, \mathbf{x}, t)}{\partial y_m \partial y_p} = (a_{mp} - r_m r_p - \phi_{mp} r_m) \frac{D_o(\mathbf{y}, \mathbf{x}, t)}{y_m y_p},$$

for $m, p = 1, \dots, M$ and $\phi_{mp} = 1$ if $m = p$ and 0 otherwise. Note that factoring out $D_o(\mathbf{y}, \mathbf{x}, t)/y_m$ from the rows and $1/y_p$ from the columns of \mathbf{H} does not change the signs of

its principal minors. Hence, we can simplify the problem by considering the following matrix

$$\widetilde{\mathbf{H}} = \begin{bmatrix} \tilde{h}_{11} & \tilde{h}_{12} & \cdots & \tilde{h}_{1M} \\ \tilde{h}_{21} & \tilde{h}_{22} & \cdots & \tilde{h}_{2M} \\ \vdots & \vdots & \cdots & \vdots \\ \tilde{h}_{1M} & \tilde{h}_{2M} & \cdots & \tilde{h}_{MM} \end{bmatrix}$$

where

$$\tilde{h}_{mp} = a_{mp} - r_m r_p - \phi_{mp} r_m. \quad (45)$$

Thus, the distance function will be convex in outputs if and only if $\widetilde{\mathbf{H}}$ is positive-semidefinite.

4 Bayesian Estimation

With the translog function for $D_o(\mathbf{y}, \mathbf{x}, t)$, the stochastic output distance function in (38) can be rewritten in a panel data framework as

$$q_{it} = \mathbf{z}'_{it} \boldsymbol{\beta} + u_{it} + v_{it}, \quad (46)$$

where $i = 1, \dots, K$ indicates firms, $t = 1, \dots, T$ indicates time, $q_{it} = -\ln y_{3,it}$, \mathbf{z}_{it} is a vector comprising all the variables which appear on the right hand side of (38), and $\boldsymbol{\beta}$ refers to the corresponding vector of coefficients of the translog function (including the intercept).

The formulation of our empirical model as a random effects model (46) is convenient for Bayesian analysis. Although equation (38) can also be formulated as a fixed effects model and then estimated using Bayesian procedures, we prefer a random effects model for the following two reasons. First, with a fixed effects model we have to specify the same number of intercepts as that of observational units, which makes the implementation of Bayesian estimation methods cumbersome, since we have 292 banks in this study. Second, most previous studies investigating U.S. bank efficiency adopted maximum likelihood models, a special case of random effect models. Hence, adopting a random effects model will enable us to compare our empirical results regarding bank efficiency to those from previous studies. It is also to be noted that fixed effects models generally give different results than maximum likelihood models, since in the fixed effects models technical efficiency is measured relative to the best performing bank in the sample, rather than using equation (39).

With this in mind, letting $h = 1/\sigma^2$, the Bayes theorem in our particular case can be restated as

$$f(\boldsymbol{\beta}, h, \mathbf{u}, \lambda^{-1} | \mathbf{q}) \propto L(\mathbf{q} | \boldsymbol{\beta}, h, \mathbf{u}, \lambda^{-1}) p(\boldsymbol{\beta}, h, \mathbf{u}, \lambda^{-1}), \quad (47)$$

where $f(\boldsymbol{\beta}, h, \mathbf{u}, \lambda^{-1} | \mathbf{q})$ is the posterior joint density function for all the parameters, $\boldsymbol{\beta}$, h , \mathbf{u} and λ^{-1} , given \mathbf{q} . The posterior density summarizes all the information about $\boldsymbol{\beta}$, h , \mathbf{u} and λ^{-1} after \mathbf{q} is observed. $L(\mathbf{q} | \boldsymbol{\beta}, h, \mathbf{u}, \lambda^{-1})$ is the likelihood function of the sample, which

summarizes all the sample information. $p(\boldsymbol{\beta}, h, \mathbf{u}, \lambda^{-1})$ is the joint prior density function for the parameters, $\boldsymbol{\beta}$, h , \mathbf{u} and λ^{-1} , summarizing the best initial guess of $\boldsymbol{\beta}$, h , \mathbf{u} and λ^{-1} .

Under the assumption that the v_{it} 's are iid normal, the likelihood function in (47) can be shown to be

$$\begin{aligned} L(\mathbf{q} | \boldsymbol{\beta}, h, \mathbf{u}, \lambda^{-1}) &= \prod_{i=1}^K \prod_{t=1}^T \left\{ \sqrt{\frac{h}{2\pi}} \exp \left[-\frac{h}{2} (q_{it} - \mathbf{z}'_{it} \boldsymbol{\beta} - u_{it})^2 \right] \right\} \\ &\propto h^{K \times T / 2} \exp \left[-\frac{h}{2} \mathbf{v}' \mathbf{v} \right], \end{aligned} \quad (48)$$

where $\mathbf{v} = (\mathbf{q} - \mathbf{z}' \boldsymbol{\beta} - \mathbf{I}_{KT} \mathbf{u})$, with \mathbf{I}_{KT} being the $KT \times KT$ identity matrix.

The Bayesian model in (47) also requires choosing prior parameter values. We choose a flat (constant) prior for $\boldsymbol{\beta}$ — it is to be noted that the sum or integral of the prior values may not even need to be finite to get sensible answers for the posterior probabilities

$$p(\boldsymbol{\beta}) \propto I(\boldsymbol{\beta} \in R_j), \quad (49)$$

where $I(\cdot)$ is an indicator function which takes the value 1 if the argument is true and 0 otherwise, and R_j is the set of permissible parameter values when no theoretical regularity constraints ($j = 0$) are imposed and when both the monotonicity and curvature constraints ($j = 1$) must be satisfied. Generally speaking, a flat (constant) prior is assumed when the researcher does not wish to impose prior constraints on model parameters, and thus renders the posterior proportional to the sampling density (likelihood function). With the constant equal to an indicator function, our particular flat prior for $\boldsymbol{\beta}$ allows us to slice away the portion of posterior density that violates monotonicity and curvature of the output distance function.

We adopt the following prior for h

$$p(h) \propto h^{-1}, \quad \text{where } h = \frac{1}{\sigma^2} > 0. \quad (50)$$

The main effect of such a prior is to downweigh excessively large values of the precision, h .

As stated above, we choose an exponential distribution for u_{it} . This is mainly because van den Broek *et al.* (1994) argue that this distribution for inefficiency u_{it} is more robust to prior assumptions about parameters than other distributions. Since the exponential distribution is a special case of the gamma distribution, the prior for u_{it} is

$$p(u_{it} | \lambda^{-1}) = f_{\text{Gamma}}(u_{it} | 1, \lambda^{-1}), \quad (51)$$

where f_{Gamma} is a gamma density function. According to Fernandez *et al.* (1997), in order to obtain a proper posterior we need a proper prior for the remaining parameter, λ . Accordingly, we use the proper prior

$$p(\lambda^{-1}) = f_{\text{Gamma}}(\lambda^{-1} | 1, -\ln \tau^*), \quad (52)$$

where τ^* is the prior median of the efficiency distribution.

With the priors (49)-(52), our joint prior probability density function is therefore

$$f(\boldsymbol{\beta}, h, \mathbf{u}, \lambda^{-1}) = p(\boldsymbol{\beta})p(h)p(\mathbf{u}|\lambda^{-1})p(\lambda^{-1}) \\ \propto h^{-1}I(\boldsymbol{\beta} \in R_j) f_{\text{Gamma}}(\lambda^{-1}|1, -\ln \tau^*) \prod_{i=1}^K \prod_{t=1}^T f_{\text{Gamma}}(u_{it}|1, \lambda^{-1}). \quad (53)$$

Finally, our best prior for the efficiency of large banks in the United States is the mean efficiency value of 0.899 reported by Tsionas (2006) who applied a Bayesian cost frontier (without constraints) to 128 large U.S. banks. In fact, after reviewing the results of 50 U.S. bank efficiency studies, Berger and Humphrey (1997) found that the annual average efficiency is 0.84 with a standard deviation of 0.07. So we are comfortable following Tsionas (2006), setting $\tau^* = 0.899$ in this study.

Combining the likelihood function in (48) and the joint prior distribution in (53) yields the posterior joint density function

$$f(\boldsymbol{\beta}, h, \mathbf{u}, \lambda^{-1} | \mathbf{q}) \propto h^{(KT/2-1)} \exp\left[-\frac{h}{2} \mathbf{v}' \mathbf{v}\right] I(\boldsymbol{\beta} \in R_j) \times \\ \times f_{\text{Gamma}}(\lambda^{-1}|1, -\ln \tau^*) \prod_{i=1}^K \prod_{t=1}^T f_{\text{Gamma}}(u_{it}|1, \lambda^{-1}). \quad (54)$$

Also, technical change (TC), elasticities (ε_n), returns to scale (RTS), and total factor productivity growth are all functions of $\boldsymbol{\beta}, h, \mathbf{u}$, and λ^{-1} . We are particularly interested in the posterior marginal densities of $\boldsymbol{\beta}, \mathbf{u}$, TE, ε_n , RTS, and TFP growth, and the means and standard deviations of these posterior densities.

Let $g(\boldsymbol{\beta}, h, \mathbf{u}, \lambda^{-1})$ represent these functions of interest. In theory, we could obtain the moments of $g(\boldsymbol{\beta}, h, \mathbf{u}, \lambda^{-1})$ from the posterior density through integration. Unfortunately, these integrals cannot be computed analytically. Therefore, we use the Gibbs sampling algorithm which draws from the joint posterior density by sampling from a series of conditional posteriors. Essentially, Gibbs sampling involves taking sequential random draws from full conditional posterior distributions. Under very mild assumptions [see, for example, Tierney (1994)], these draws then converge to draws from the joint posterior. Once draws from the joint distribution have been obtained, any posterior feature of interest can be calculated.

The full conditional posterior distributions for $\boldsymbol{\beta}$, h , \mathbf{u} , and λ^{-1} can be shown to be

$$p(\lambda^{-1} | \mathbf{q}, \boldsymbol{\beta}, h, \mathbf{u}) \propto f_{\text{Gamma}}(\lambda^{-1} | KT + 1, \mathbf{u}'\boldsymbol{\nu}_{KT} - \ln \tau^*), \quad (55)$$

$$p(h | \mathbf{q}, \boldsymbol{\beta}, \mathbf{u}, \lambda^{-1}) \propto f_{\text{Gamma}}\left(h \left| \frac{KT}{2}, \frac{1}{2} \mathbf{v}'\mathbf{v} \right.\right); \quad (56)$$

$$p(\boldsymbol{\beta} | \mathbf{q}, h, \mathbf{u}, \lambda^{-1}) \propto f_{\text{Normal}}\left[\boldsymbol{\beta} \left| \mathbf{b}, h^{-1} (\mathbf{z}'\mathbf{z})^{-1} \right.\right] I(\boldsymbol{\beta} \in R_j) \quad (57)$$

$$p(\mathbf{u} | \mathbf{q}, \boldsymbol{\beta}, h, \lambda^{-1}) = f_{\text{Normal}}(\mathbf{u} | \mathbf{q} - \mathbf{z}'\boldsymbol{\beta} - (h\lambda)^{-1}\boldsymbol{\nu}_{KT}, h^{-1}\mathbf{I}_{KT}) \prod_{i=1}^K \prod_{t=1}^T I(\mathbf{u}_{it} \geq 0) \quad (58)$$

where $\mathbf{b} = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'[\mathbf{q} - \mathbf{I}_{KT}\mathbf{u}]$, with $\boldsymbol{\nu}_{KT}$ being the KT vector of ones, and f_{Normal} is a normal density function.

The Gibbs sampler for Bayesian estimation without monotonicity and curvature constraints can be implemented by setting $I(\boldsymbol{\beta} \in R_0)$ in (57) equal to one and then drawing sequentially from the conditional posteriors in (55)–(58). Sampling from (55), (56), and (57) is straightforward. However, sampling from (58), a multivariate truncated normal distribution, is more complicated. Luckily, in our particular case, sampling from the multivariate truncated normal distribution (58) can be simplified as KT independent draws from the following univariate truncated normal distribution

$$p(\mathbf{u}_{it} | \mathbf{q}, \boldsymbol{\beta}, h, \lambda^{-1}) = f_{\text{Normal}}(\mathbf{q}_{it} - \mathbf{z}'_{it}\boldsymbol{\beta} - (h\lambda)^{-1}, h^{-1}) I(\mathbf{u}_{it} \geq 0), \quad (59)$$

by noting that the covariance matrix is a scalar times an identity matrix, and the truncations are independent. Sampling from univariate truncated normal distributions can be easily implemented, using procedures discussed in Griffiths (2004).

The Gibbs sampler for Bayesian estimation with monotonicity and curvature constraints also involves taking sequential random draws from the above full conditional posterior distributions. Sampling from (55), (56), and (58) is the same as in the case without monotonicity and curvature constraints. However, sampling from the multivariate normal distribution (57) is even more involved than sampling from the multivariate normal distribution (58) in that the region to which $\boldsymbol{\beta}$ is truncated cannot be explicitly specified. There are two approaches in this literature which can be used to handle the sampling from the truncated multivariate normal distribution like (57) — the accept-reject algorithm [see Terrell (1996)] and the Metropolis-Hastings (M-H) algorithm, proposed by Griffiths *et al.* (2000) and used by O'Donnell and Coelli (2005). The accept-reject algorithm has been criticized for its inefficiency in that it needs to generate an extremely large number of candidate draws before

finding one that is acceptable — see Griffiths *et al.* (2000). In this paper, we follow Griffiths *et al.* (2000) and sample the truncated multivariate normal distribution (57) using the Metropolis-Hastings algorithm, which in our case proceeds iteratively as follows:

- *Step 1:* Start with an initial value β^j satisfying both the monotonicity and curvature constraints. Let j denote the state of β , and set $j = 1$.
- *Step 2:* Using the current value β^j , sample a candidate point β^c from a symmetric proposal density $q(\beta^c, \beta^j)$, which is the probability of returning a value of β^c given a previous value of β^j .
- *Step 3:* Evaluate the monotonicity and curvature constraints at the specified data points using the candidate value, β^c . If any constraints are violated, set $\alpha(\beta^j, \beta^c) = 0$ (that is, the probability that the move from j to c is made) and go to Step 5.
- *Step 4:* Calculate $\alpha(\beta^j, \beta^c) = \min[a_1, 1]$ where a_1 is the ratio of the target density at the candidate (β^c) and current (β^j) points, and can be written (in our case) as

$$\frac{\exp [(\beta^c - \mathbf{b}) (h\mathbf{z}'\mathbf{z}) (\beta^c - \mathbf{b})]}{\exp [(\beta^j - \mathbf{b}) (h\mathbf{z}'\mathbf{z}) (\beta^j - \mathbf{b})]}$$

- *Step 5:* Generate independent uniform random variables, \mathbf{u} , from the interval $[0, 1]$.
- *Step 6:* Set $\beta^{j+1} = \beta^c$ if $\mathbf{u} < \alpha(\beta^j, \beta^c)$ and β^j otherwise.
- *Step 7:* Set $j = j + 1$ and return to Step 2.

The algorithm works best if the proposal density matches the shape of the target distribution. Therefore, the proposal density is chosen to be a multivariate normal with mean equal to the current state β^j and covariance matrix equal to the maximum likelihood estimate of the covariance matrix of the parameters, multiplied by a tuning parameter. The tuning parameter can be used to adjust the acceptance rate, which is the fraction of proposed samples that is accepted in a window of the last κ samples. The optimal acceptance rate (i.e., the one which minimizes the autocorrelations across the sample values) has been shown to lie within the range between 0.45 (in one-dimensional problems) and approximately 0.23 (as the number of dimensions becomes infinitely large) — see Roberts *et al.* (1997). In this paper, we choose the tuning parameter so that the acceptance rate lies within this range.

Compared with the conventional M-H algorithm, the above M-H algorithm is capable of imposing monotonicity and curvature constraints through manipulating $\alpha(\beta^j, \beta^c)$. More specifically, it sets $\alpha(\beta^j, \beta^c) = 0$ when any monotonicity and curvature constraints are violated, and equal to the expression in Step 4 (as in the conventional M-H algorithm) otherwise. And in the case where $\alpha(\beta^j, \beta^c) = 0$, the candidate draw will always be rejected, thus ensuring that monotonicity and curvature are satisfied.

5 The Data

The data used in this study are obtained from the Reports of Income and Condition (Call Reports) over the six-year period ($T = 6$) from 2000 to 2005. We examine only continuously operating banks to avoid the impact of entry and exit and to focus on the performance of a core of healthy, surviving institutions during the sample period. In this paper, we selected the subsample of large banks, namely those with total assets in excess of one billion dollars (in 2000 dollars) in the last three year in the sample. This gives a total of 292 banks ($K = 292$) observed over 6 years.

To select the relevant variables, we follow the commonly-accepted intermediation approach proposed by Sealey and Lindley (1977), which treats deposits as inputs and loans as outputs. On the input side, three inputs are included. The quantity of labor, x_1 ; the quantity of purchased funds and deposits, x_2 ; and the quantity of physical capital, x_3 , which includes premises and other fixed assets. On the output side, three outputs are specified. These are securities, y_1 , which includes all non-loan financial assets (i.e., all financial and physical assets minus the sum of consumer loans, non-consumer loans, securities, and equity); consumer loans, y_2 ; and non-consumer loans, y_3 , which is composed of industrial, commercial, and real estate loans. All the quantities are constructed by following the data construction method in Berger and Mester (2003). These quantities are also deflated by the CPI to the base year 2000, except for the quantity of labor.

While non-traditional activities are clearly increasing in importance, the wide range of activities and imperfect data make the measurement of non-traditional activities problematic — see Stiroh (2000) for a discussion of the different approaches to the measurement of non-traditional activities. To avoid the uncertainties associated with the introduction of non-traditional activities, we choose not to include it as an output. But we do run an alternative model where non-traditional activities are considered as an extra output to check the robustness of the estimates of technical change.

6 Empirical Results

6.1 Regularity Tests

We start with unconstrained parameter estimates and make 50,000 draws discarding the first 20,000 as a burn in. Table 1 presents the estimated parameters and also reports both standard deviations and 90% posterior density regions calculated as the 5th and 90th percentiles of the MCMC sample observations. We calculate 90% posterior density regions because it provides a better indication of likely values of the parameters when the marginal posterior distributions are asymmetric.

Regularity tests can be implemented by analyzing the estimated unconstrained marginal

posterior pdfs of k_n and r_m and the principal minors of $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{H}}$. We first evaluate the posterior means of k_n and r_m and the principal minors of $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{H}}$, at each of the 1752 ($= K \times T$) observations, and then calculate the proportions of regularity violations relative to the total number of observations. The results, presented in the first column of Table 2, indicate that only two (k_2 and r_1) of the six monotonicity conditions are satisfied at all the 1752 observations and that both curvature conditions are violated, with the quasi-convexity in outputs being violated at all observations. We then evaluate the posterior coverage regions of k_n and r_m and of the principal minors of $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{H}}$, again at each of the 1752 observations, and calculate the ratio of the number of observations, where posterior coverage regions span inadmissible values, to the total number of observations (1752). As can be seen in the second column of Table 2, all eight regularity conditions have a positive probability of being violated at some observations. In fact, both of the curvature conditions have a positive probability of being violated at all the 1752 observations.

These violations of monotonicity and curvature in the unconstrained model may lead to perverse conclusions concerning TFP growth. To see this, we also generate the marginal density plots for the shadow input cost shares, ω_n for $n = 1, 2, 3$ in (15), and the shadow output revenue shares, $\tilde{\omega}_m$ for $m = 1, 2, 3$ in (15), from the unconstrained model, evaluated at the mean value of all inputs and outputs in each year. As discussed above, both ω_n and $\tilde{\omega}_m$ are required to be positive and less than one. Due to space limitations, only the marginal densities in 2005 are plotted in Figure 1.1-1.6 — the marginal densities for other years are similar to those in 2005. Clearly, all the three shadow output shares are reasonable, containing no negative values or values larger than one. However, the plot of the shadow input shares shows that the labor share and the capital share may be negative. A negative input share implies that an increase in the use of that input (with all other inputs and outputs held constant) will increase the (measured) productivity of that bank, which is economically implausible. Moreover, Figure 1.2 shows that the shadow input share for funds may be greater than one, implying that an increase in the use of that input (with all other inputs and outputs held constant) will reduce the (measured) productivity of that bank by more than the growth rate of funds, which is again economically implausible.

Since monotonicity and curvature are not attained in the unconstrained model, we follow the procedures specified in Section 4 to impose those constraints on the translog output distance function. Again, we generated a total of 50,000 observations, and then discarded the first 20,000 as a burn-in. The associated estimates of parameters are reported in Table 3, the monotonicity and curvature violations reported in Table 4, and the marginal densities for the shadow input and output shares are plotted in Figure 2.1–2.6. Generally speaking, the constrained model has smaller posterior standard deviations and narrower confidence intervals in terms of posterior moments for the estimated parameters and shadow revenue and cost shares. This is consistent with Dorfman and McIntosh (2001) and O’Donnell and Coelli (2005) who find that incorporating inequality constraints into the estimation process has the effect of reducing the variances of the estimated marginal pdfs. In addition,

Figures 2.1–2.6 show that some densities are asymmetric — for example, those for the funds share, capital share, and non-consumer loans share. Kleit and Terrell (2001) found similar results and suggested that the asymmetry perhaps reflects the fact that the constrained posterior density slices away the portion of the unconstrained posterior density that violates monotonicity and curvature.

As we expected, monotonicity and curvature are satisfied by all measures after monotonicity and curvature are incorporated. In particular, k_n and r_m and the principal minors of $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{H}}$ are correctly signed at all 1752 observations whether they are evaluated by using posterior means or by using posterior coverages. Moreover, the shadow shares are all positive and less than one. In what follows, we will discuss technical efficiency, technological change, returns to scale, and the contributions of each of these components to TFP growth, based on the constrained translog output distance function.

6.2 Results from the Constrained Model

6.2.1 Technical Efficiency

Table 5.1 reports the estimates of average technical efficiency over the sample period, together with the 90% posterior density regions. The average technical efficiency for each year is evaluated at the mean value of all inputs and outputs in that year. As indicated by the standard deviations and 90% density regions, the estimates of the average technical efficiency are statistically significant for every year over the sample period. The scores of technical efficiency show a high level of efficiency, ranging from 92.43% to 93.41%. Thus, on average, a 7% to 8% proportional increase in outputs can be achieved by solely increasing efficiency, without altering production technology and input usage.

Our estimates of the technical efficiency are quite close to those from recent research; see, for example, Stiroh (2000) and Tsionas (2006) — both of these studies employed a translog cost frontier (dual method), rather than a distance frontier (primal method). Thus, one of the differences in efficiency estimates could be due to allocative efficiency. For example, Tsionas used the panel data on 128 large U.S. banks over the period from 1989 to 2000 and found that the average efficiency is 88.9% when a dynamic effect is not considered and 95.5% when a dynamic effect is considered. Further, our technical efficiency estimates show no specific pattern of temporal change. In particular, it starts at 93.41% in 2000, falls to 92.49% in 2001, rebounds slightly in the following two years, falls slightly again in 2004, and picks up again to 92.69% in 2005. This time pattern of technical efficiency means that the change in technical efficiency is not a consistent source of TFP growth.

To get a better understanding of the distribution of technical efficiency across banks, in Table 5.2 we report the minimum and maximum technical efficiency in each year, together with standard deviations, and the 5th and 95th percentile values. The results show that the scores of technical efficiency can differ greatly across banks in all the sample years. Taking

the technical efficiencies in 2005 as an example, the highest is 97.63% whereas the lowest is only 35.08%. Despite these extreme cases, the results on standard deviations and the 5th and 95th percentile values show that the vast majority of the banks fall within the range between 84% and 96%.

6.2.2 Returns to Scale

Table 6 summarizes the returns to scale (RTS) estimates, again evaluated at the mean value of all inputs and outputs each year. The standard deviations and 90% density regions indicate that the RTS estimates are statistically significant for every year over the sample period. Clearly, the point estimates of RTS in Table 6 are all greater than one, ranging from 1.037 to 1.056, suggesting that the large commercial banks in the sample exhibit moderate increasing returns to scale. This is consistent with recent research that found scale economies in the U.S. banking industry using data for the 1990s — see, for example, Berger and Mester (1997), Hughes and Mester (1998), and Stiroh (2000). Increasing returns to scale indicates the presence of imperfect competition in the U.S. banking industry, which is consistent with the findings of Bikker and Haaf (2002) and Claessens and Laeven (2003) that the U.S. banking industry is characterized by a relatively low level of competition.

The presence of moderate increasing returns to scale also has two implications for productivity growth. First, the presence of moderate increasing returns to scale implies that productivity growth will exhibit procyclical behavior to some extent. This is because the contribution of scale economies to productivity growth is positive when the share weighted input aggregate grows over time, but negative when the share weighted input aggregate declines over time, as can be seen from (36). Second, since the economies of scale is moderate in magnitude, the scale effect will not be a consistent significant source of TFP growth. In addition, the presence of moderate increasing returns to scale also implies that the large banks in the U.S. are expected to be engaged in more mergers and acquisitions until the returns to scale are exploited.

6.2.3 Technical Change

Table 7.1 reports technical change rate estimates, again evaluated at the mean value of all inputs and outputs each year. Clearly, the estimates are statistically significant in every year over the sample period. On average, the rate of technical change is 2.22% per year. Compared with the estimates of technical efficiency, which show no specific pattern of temporal change, the estimates of the rate of technical change show a declining trend. In particular, the rate of technical change falls consistently from 6.0% in 2000 to -1.79% in 2005. In terms of the output distance function, this decline in the rate of technical change means that the growth rate of the ratio of actual output to potential output declines as time passes (holding all other things constant). In terms of the revenue function, which is dual to the output

distance function, this decline in the rate of technical change means that the growth rate of the revenue generated from a fixed combination of inputs declines with the passing of time. Considering the small variation in technical efficiency and the small magnitude of the scale effect, technical change seems to be the dominant force driving the growth in total factor productivity.

Considering the importance of technical change, we also estimated three alternative models to check the robustness of our results regarding the time pattern of technical change. In the first alternative model (Model 1), we treat securities (instead of non-consumer loans) as the numeraire for normalizing the outputs, to see whether the choice of the numeraire has any effect on the time pattern of technical change. The second alternative model (Model 2) is just the unconstrained model, where all the outputs and inputs remain unchanged. This model, though having been discarded due to its violations of monotonicity and curvature, is used to see whether the imposition of constraints has greatly altered the time pattern of the rate of technical change. In the third alternative model (Model 3), we add an off-balance-sheet variable to see whether the exclusion of non-traditional activities affects the estimated time pattern of the rate of technical change. The estimates of the rate of technical change, together with 90% posterior density regions, from the three alternative models are reported in Table 7.2.

The estimates of the rate of technical change from the first alternative model (Model 1), reported in the first column of Table 7.2, are almost the same as those in Table 7.1 (our standard model), suggesting that the choice of the numeraire has almost no effect on the estimated time pattern of the rate of technical change. The estimates based on the second alternative model (Model 2), reported in the second column of Table 7.2, also follow the same pattern as in our standard model, although there is a slight difference in magnitude of the technical change rate estimates between the two models. This suggests that the imposition of constraints has little effect on the estimated time pattern of technical change. When the off-balance-sheet variable is added, the technical change rate estimates change on average by 0.45% in absolute terms. However, as can be clearly seen in the third column of Table 7.2, the time pattern of technical change is still almost the same. As we discussed above, the wide range and imperfect data of the non-traditional activities could introduce more uncertainty regarding the estimates of the rate of technical change. Thus, the third alternative model (Model 3) is not our preferred model.

In summary, the time pattern and (to a lesser degree) magnitude of the rate of technical change estimates are very robust to the different choice of the numeraire output, the imposition of monotonicity and curvature constraints, and the inclusion of off-balance-sheet variables.

6.2.4 TFP Growth and Its Components

We now turn to a decomposition of the growth rate of total factor productivity, as shown in Table 8 — it should be noted that the first year in the sample period is dropped because we have to difference the technical efficiencies in two consecutive years to obtain efficiency changes. Again, all the estimates are evaluated at the mean values of all inputs and outputs in each year. In addition to the estimates of the three TFP growth components, we also calculate the percentage contribution of each of the three productivity components to total factor productivity growth, shown in brackets in Table 8.

Overall, the results presented in the first column of Table 8 indicate that total factor productivity grew in all years, except the last, at an average annual rate of 1.98%. However, the estimates for total factor productivity growth also exhibit a clear downward trend. In particular, total factor productivity growth is quite impressive in the first three years, in all exceeding 2%. But, it falls almost to zero in 2004 and even turns negative in the last year in the sample. It should be noted that while TFP growth shows a downward trend, TFP level has been increased over the sample period except the last. In particular, if we normalize the productivity level in 2000 to 100, then the productivity level in the last year will be 109.91.

The decomposition of total factor productivity growth in Table 8 identifies the forces that drive its decline. In particular, the estimates for efficiency changes, $-du/dt$, in the second column of Table 8 are rather small in magnitude, averaging only 0.14% per year. Moreover, they fluctuate around zero, indicating that efficiency change has an unstable effect on total factor productivity growth. The small effects of efficiency changes on total factor productivity growth are also reflected in the percentage contribution to total factor productivity growth, reported in Column 3 of Table 8, averaging 7.27% per year. The estimates reported in the fourth column of Table 8 indicate that the scale effect has a moderate positive effect on total factor productivity growth, averaging 0.44% per year. In terms of average percentage contributions, the scale effect is the second largest factor contributing to growth in total factor productivity (22.30%). This is consistent with our estimates of returns to scale, which show moderate economies of scale in large commercial banks in the United States.

Without doubt, the last component, technical change, is the dominant force behind total factor productivity growth. This can be clearly seen from the average annual rate of technical change (of 1.39%) in column 6 of Table 8. The importance of technical change can also be seen from its percentage contribution — it contributes over 75% each year to productivity growth. Further, the technical change estimates show a clear downward trend, accounting for the decline in total factor productivity growth over the sample period.

6.3 Sensitivity Analysis

A possible problem with our estimation of the output distance function is endogeneity — that is, the regressors on the right hand side of equation (38) may not be exogenous. To investigate

the robustness of our results to alternative estimation procedures, in this subsection we use instrumental variables.

The variables on the right hand of (38) can be classified into two types of variables — the output ratio variables (i.e. y_m/y_M , $m = 1, \dots, M - 1$) and the input variables. According to Coelli and Perelman (1999), the output ratios are measures of the output mix which are more likely to be exogenous. Schmidt (1988) and Mundlak (1996) also find that, in the context of a production function, the input ratios do not suffer from the endogeneity problem; the basic argument also applies to the output ratios in the transformed output distance function. Thus, the only variables suspected of causing possible endogeneity problems are the input variables. To use instrumental variables for the input variables, we follow the assertion of Griliches (2000, p. 62) that “good instruments are hard to find without the supporting theory that give them a formal role in the model.” As we noted above, the U.S. banking industry is more likely to be characterized by monopolistic competition. Hence, in order to be consistent with the theoretical framework of profit maximization in the presence of imperfect competition, input prices and the time trend are chosen as instruments.

The empirical results are summarized in Table 9. A comparison of Tables 8 and 9 reveals that the major conclusions reached in the previous subsection are still valid, although we notice that there are some changes. First, total factor productivity growth still shows a clear downward trend, implying that productivity has been growing at a lower rate. In particular, it has consistently decreased from 0.0491 to 0.033 over the sample period. Second, technical change is still the driving force behind the decline in total factor productivity growth. From the contributions of the three productivity components, we see that technical change is still the dominant force, accounting for 70.32% of the productivity growth on average. With the contributions from the other productivity components being rather small, the consistent decline in technical change (see the last column of Table 9) results in the decline in productivity growth. Third, the estimates of efficiency change and the scale effect when instrumental variables are used are comparable to our earlier estimates. Finally, we also find that the estimates of the contributions of the three productivity components when instrumental variables are used are very similar to our earlier estimates as well. In particular, the average contributions of technical change, scale effect, and efficiency change when instrumental variables are used are 70.32%, 21.94%, and 7.74%, respectively, and they are 70.33%, 22.30%, and 7.27% when instrumental variables are not used. Therefore, our major conclusions in the previous subsection are quite robust to the use of instrumental variables.

7 Conclusion

In this paper, we extend and combine the best elements of the non-parametric approach and the parametric approach, and propose a distance-function based primal Divisia total factor

productivity growth index. In particular, we show that this Divisia total factor productivity growth index is equivalent to the conventional dual Divisia total factor productivity growth index under the assumption of a competitive market. We further show that, in the presence of imperfect competition, it is equivalent to a markup and markdown adjusted dual Divisia total factor productivity growth index, which reflects the firm’s true marginal revenue and marginal cost. Based on the primal Divisia total factor productivity growth index, we present a decomposition of productivity change, isolating the separate contributions of scale economies, technical change, and technical efficiency change. The primal Divisia total factor productivity growth index has several advantages, as it does not require price information (and thus can be widely used in situations where price information is missing), it does not require prior knowledge of the underlying market structure, and it allows an easy calculation of the contribution of scale effect and a deep insight into important production structures.

We also pay explicitly to the theoretical regularity conditions of the output distance function (i.e. non-decreasing, convex and linearly homogeneous in outputs, and non-increasing and quasi-convex in inputs). We show that these conditions are not only necessary for the validity of the output distance function as a means of completely describing the technology, but also some of the conditions are necessary for its validity as a productivity growth index. In order to impose these nonlinear theoretical regularity conditions, we need to adopt an estimation method which is capable of incorporating monotonicity and curvature conditions implied by neoclassical microeconomic theory. In this respect, we follow O’Donnell and Coelli (2005) and use the Bayesian approach to impose the theoretical regularity conditions on the parameters of a translog output distance function. Implementing the approach involves the use of a Gibbs sampler with data augmentation. A Metropolis-Hastings algorithm is also used within the Gibbs sampler to simulate observations from truncated pdfs.

We applied our methodology to the panel data on 292 large banks in the United States over the period from 2000 to 2005. Our results confirm that the monotonicity and concavity constrained model yields more accurate and favorable results than an unconstrained model. In particular, shadow revenue and cost shares are well behaved, and the standard deviations are largely reduced. Our results show that total factor productivity grew at an average rate of 1.98% for the large U.S. commercial banks over the sample period. However, the estimates of total factor productivity growth show a clear downward trend and our decomposition of the total factor productivity growth rate indicates that technical change is the driving force that leads to the decline in the total factor productivity growth rate. Our results indicate that returns to scale also have a positive effect on productivity growth, suggesting that the scale effect should be included when examining bank productivity growth.

In estimating technical change, returns to scale, and efficiency in large banks in the United States, we have used a translog output distance function. A locally flexible functional form, the translog is only suitable for samples composed of relatively homogenous firms — for example, only large banks with assets greater than \$1 billion are used in this study. In cases where the firms are of widely varying sizes, globally flexible functional forms which

can provide greater flexibility will be more appropriate. There are two globally flexible functional forms — the Asymptotically Ideal Model, introduced by Barnett *et al.* (1991), and the Fourier flexible functional form, introduced by Gallant (1982). However, due to the trigonometric terms which are not neoclassical, the Fourier functional forms has been criticized for its possibility of overfitting the data — see, for example, Barnett *et al.* (1988). In contrast, with the globally regular Müntz-Szatz series, the AIM model form fits only that part that is globally regular, thus eliminate the risk of overfitting. Therefore, using an AIM output distance function to estimate technical change, returns to scale, and efficiency is an area for potentially productive future research.

References

- [1] Aigner D.J., Lovell C.A.K., and Schmidt P. "Formulation and Estimation of Stochastic Frontier Production Function Models." *Journal of Econometrics* 6 (1977): 21-37.
- [2] Alam, I.M.S. "A Non-Parametric Approach for Assessing Productivity Dynamics of Large Banks." *Journal of Money, Credit, Banking* 33 (2001), 121-139.
- [3] Atkinson, S.E., Cornwell, C., and Honerkamp, O. "Measuring and Decomposing Productivity Change: Stochastic Distance Function Estimation Versus Data Envelopment Analysis." *Journal of Business and Economic Statistics* 21 (2003), 284-294.
- [4] Barnett, W.A., J. Geweke, and M. Wolfe. "Semi-nonparametric Bayesian Estimation of the Asymptotically Ideal Production Model." *Journal of Econometrics* 49 (1991), 5-50.
- [5] Barnett, W.A. and Yue, P. "Semi-Nonparametric Estimation of the Asymptotically Ideal Model: the AIM Demand System." In *Advances in Econometrics*, Vol VII, Rhodes, G and Fomby, T.B. (eds), Greenwich: CT: JAI Press (1988), 229-252.
- [6] Bauer, P. "Decomposing TFP Growth in the Presence of Cost Inefficiency, Non-Constant Returns to Scale and Technological Progress." *Journal of Productivity Analysis* 1 (1990), 287-301.
- [7] Berger, A.N. "The Economic Effects of Technological Progress: Evidence from the Banking Industry." *Journal of Money, Credit, and Banking* 35 (2004), 141-176.
- [8] Berger, A.N. and Humphrey, D.B. "Efficiency of Financial Institutions: International Survey and Directions for Future Research." *European Journal of Operational Research* 98 (1997), 175-212.
- [9] Berger, A.N., and Mester, L.J. "Inside the Black Box: What Explains Differences in the Efficiencies of Financial Institutions?" *Journal of Banking and Finance* 21 (1997), 895-947.
- [10] Berger, A.N. and Mester, L.J. "Explaining the Dramatic Changes in the Performance of U.S. Banks: Technological Change, Deregulation, and Dynamic Changes in Competition." *Journal of Financial Intermediation* 12 (2003), 57-95.
- [11] Berger, A.N., A.K. Kashyap, and J.M. Scalise. "The Transformation of the U.S. Banking Industry: What a Long, Strange Trip Its Been." *Brookings Papers on Economic Activity* 2 (1995) 54-219.

- [12] Bikker, J.A. and K. Haaf. “Competition, Concentration and Their Relationship: An Empirical Analysis of the Banking Industry.” *Journal of Banking and Finance* 26 (2002), 2191-2214.
- [13] Brummer, B., T. Glauben, and G. Thijssen. “Decomposition of Productivity Growth Using Distance Functions: the Case of Dairy Farms in Three European Countries.” *American Journal of Agricultural Economics* 84 (2002), 628–644.
- [14] Claessens S. and Laeven, L. “Financial Development, Property Rights, and Growth.” *Journal of Finance* 58 (2003), 2401-2436.
- [15] Coelli, T.J. and S. Perelman. “A Comparison of Parametric and Non-parametric Distance Functions: With Application to European Railways.” *European Journal of Operational Research* 117 (1999), 326–339.
- [16] Dorfman, J.H. and McIntosh, C.S. “Imposing Inequality Restrictions: Efficiency Gains from Economic Theory.” *Economics Letters* 71 (2001), 205-209.
- [17] Färe, R., Grosskopf, S., Norris, M., and Zhang, Z. “Productivity Growth, Technical Progress and Efficiency Change in Industrialized Countries.” *American Economic Review* 84 (1994), 66-83.
- [18] Färe, R. and Grosskopf, S. *Cost and Revenue Constrained Production*. Springer (1994).
- [19] Färe, R. and Primont, D. “A Distance Function Approach to Multi-Output Technologies.” *Southern Economic Journal* 56 (1990), 879-891.
- [20] Fare, R. and Primont, D. *Multi-Output Production and Duality: Theory and Applications*. Netherlands: Kluwer Academic Publishers (1995).
- [21] Feng, G. and A. Serletis. “Productivity Trends in U.S. Manufacturing: Evidence from the NQ and AIM Cost Functions.” *Journal of Econometrics* (2007, forthcoming), doi:10.1016/j.jeconom.2007.06.002.
- [22] Feng, G. and A. Serletis. “Efficiency and Productivity of the U.S. Banking Industry, 1998-2005: Evidence from the Fourier Cost Function Satisfying Global Regularity Conditions.” *Journal of Applied Econometrics* (2008, forthcoming)
- [23] Fernandez, C., J. Osiewalski, and M.F.J. Steel. “On the Use of Panel Data in Stochastic Frontier Models with Improper Priors.” *Journal of Econometrics* 79 (1997), 169-193.
- [24] Førsund, F.R. “The Malmquist Productivity Index, TFP and Scale.” Taipei International Conference on Efficiency and Productivity Growth, June 20–21, 1997.

- [25] Gallant, A.R. “Unbiased Determination of Production Technology.” *Journal of Econometrics* 20 (1982), 285-323.
- [26] Gallant AR, and Golub G. 1984. “Imposing Curvature Restrictions on Flexible Functional Forms.” *Journal of Econometrics* 26 (1984): 295-321.
- [27] Griffiths, W.E., C.J. O’Donnell, and A. Tan Cruz. “Imposing Regularity Conditions on A System of Cost and Cost-Share Equations: A Bayesian Approach.” *Australian Journal of Agricultural and Resource Economics* 44 (2000), 107-127.
- [28] Griffiths, W.E. “A Gibbs Sampler for the Parameters of a Truncated Multivariate Normal Distribution.” In *Contemporary Issues in Economics and Econometrics: Theory and Application*, R. Becker and S. Hurn (eds.), Cheltenham, U.K.: Edward Elgar (2004), 75-91.
- [29] Griliches, Z. *R & D, Education, and Productivity: A Retrospective*. Cambridge, Harvard University Press (2000).
- [30] Hughes, J.P. and L.J. Mester. “Bank Capitalization and Cost: Evidence of Scale Economies in Risk Management and Signaling.” *The Review of Economics and Statistics* 80 (1998), 314-325.
- [31] Jones, K.D. and T. Critchfield. “Consolidation in the U.S. Banking Industry: Is the Long, Strange Trip About to End?” *FDIC Banking Review* 17 (2005), 31-61.
- [32] Jorgenson, D.W. and Z. Griliches. “The Explanation of Productivity Change.” *Review of Economic Studies* 34 (1967), 249-280.
- [33] Kleit, A. and D. Terrell. “Measuring Potential Efficiency Gains from Deregulation of Electricity Generation: A Bayesian Approach.” *Review of Economics and Statistics* 83 (2001), 523-530.
- [34] Kumbhakar, S.C. and C.A.K. Lovell. *Stochastic Frontier Analysis*. Cambridge, UK.: Cambridge University Press (2003).
- [35] Mundlak, Y. “Production Function Estimation: Reviving the Primal.” *Econometrica* 64 (1996), 431-438.
- [36] Lovell, C.A.K., S. Richardson, P. Travers, and L.L. Wood. “Resources and Functionings: A New View of Inequality in Australia.” In *Models and Measurement of Welfare and Inequality*, W. Eichhorn (ed.), Berlin: Springer-Verlag Press (1994), 787-807.

- [37] Meeusen W. and J. van den Broeck. "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error." *International Economic Review* 18 (1977), 435-444.
- [38] O'Donnell, C.J. and T.J. Coelli. "A Bayesian Approach to Imposing Curvature on Distance Functions." *Journal of Econometrics* 126 (2005), 493-523.
- [39] Orea, L. "Parametric Decomposition of a Generalized Malmquist Productivity Index." *Journal of Productivity Analysis* 18 (2002), 5-22.
- [40] Panzar, J.C. and J.N. Rosse. "Testing for Monopoly Equilibrium." *Journal of Industrial Economics* 35 (1987), 443-456.
- [41] Primont, D. and C. Sawyer. "Recovering the Production Technology from the Cost Function." *Journal of Productivity Analysis* 4 (1993), 347-352.
- [42] Ray, S.C. and E. Desli. "Productivity Growth, Technical Progress and Efficiency Change in Industrialized Countries: Comment." *American Economic Review* 87 (1997), 1033-1039.
- [43] Roberts, G.O., A. Gelman, and W.R. Gilks. "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms." *Annals of Applied Probability* 7 (1997), 110-120.
- [44] Schmit, P. "Estimation of Fixed Effect Cobb-Douglas System Using Panel Data." *Journal of Econometrics* 37 (1988), 361-380.
- [45] Sealey C and J. Lindley. "Inputs, Outputs, and a Theory of Production and Cost at Depository Financial Institutions." *Journal of Finance* 32 (1977), 1251-1266.
- [46] Shephard, R.W. *Theory of Cost and Production Functions*. Princeton: Princeton University Press (1970).
- [47] Solow, R. "Technical Change and the Aggregate Production Function." *Review of Economics and Statistics* 39 (1957), 312-320.
- [48] Stiroh, K.J. "How Did Bank Holding Companies Prosper in the 1990s?" *Journal of Banking and Finance* 24 (2000), 1703-1745.
- [49] Terrell, D. "Incorporating Monotonicity and Concavity Conditions in Flexible Functional Forms." *Journal of Applied Econometrics* 11 (1996), 179-194.
- [50] Tierney, L. "Markov Chains for Exploring Posterior Distributions (with discussion)" *Annals of Statistics* 22 (1994), 1701-1762.

- [51] Tsionas, E.G. “Inference in Dynamic Stochastic Frontier Models.” *Journal of Applied Econometrics* 21 (2006), 669-676.
- [52] van den Broeck, J., G. Koop, J. Osiewalski, and M. Steel. “Stochastic Frontier Models: A Bayesian Perspective.” *Journal of Econometrics* 46 (1994), 39–56.

TABLE 1

PARAMETER ESTIMATES FROM THE UNCONSTRAINED MODEL

Variable	Parameter	Estimate	Standard deviation	90% posterior coverage regions
intercept	a_0	0.2060	0.0723	(0.0663, 0.2617)
$\ln x_1$	b_1	-0.0795	0.0418	(-0.1437, -0.0184)
$\ln x_2$	b_2	-0.9555	0.0357	(-1.0081, -0.8993)
$\ln x_3$	b_3	-0.0030	0.0276	(-0.0441, 0.0377)
$(\ln x_1)^2$	b_{11}	-0.3669	0.0618	(-0.4754, -0.2815)
$(\ln x_2)^2$	b_{22}	-0.0268	0.0372	(-0.0815, 0.0273)
$(\ln x_3)^2$	b_{33}	0.09353	0.0281	(0.0499, 0.1352)
$(\ln x_1)(\ln x_2)$	b_{12}	0.2467	0.0399	(0.1906, 0.3114)
$(\ln x_1)(\ln x_3)$	b_{13}	0.1210	0.0344	(0.0719, 0.1782)
$(\ln x_2)(\ln x_3)$	b_{23}	-0.0328	0.0053	(-0.0408, -0.0247)
$\ln y_1$	a_1	0.3956	0.0209	(0.3635, 0.4261)
$\ln y_2$	a_2	0.1094	0.0102	(0.0942, 0.1246)
$\ln y_3$	a_3	0.4951	0.0202	(0.4651, 0.5260)
$(\ln y_1)^2$	a_{11}	0.0987	0.0231	(0.0584, 0.1296)
$(\ln y_2)^2$	a_{22}	0.0268	0.0039	(0.0207, 0.0326)
$(\ln y_3)^2$	a_{33}	0.1376	0.0218	(0.0997, 0.1680)
$(\ln y_1)(\ln y_2)$	a_{12}	0.0061	0.0066	(-0.0040, 0.0163)
$(\ln y_1)(\ln y_3)$	a_{13}	-0.1047	0.0215	(-0.1342, -0.0672)
$(\ln y_2)(\ln y_3)$	a_{23}	-0.0328	0.0053	(-0.0408, -0.0247)
$(\ln x_1)(\ln y_1)$	g_{11}	-0.0211	0.0226	(-0.0565, 0.0125)
$(\ln x_1)(\ln y_2)$	g_{12}	0.0274	0.0132	(0.0080, 0.0479)
$(\ln x_1)(\ln y_3)$	g_{13}	-0.1047	0.0215	(-0.1342, -0.0672)
$(\ln x_2)(\ln y_1)$	g_{21}	0.0776	0.0196	(0.0483, 0.1083)
$(\ln x_2)(\ln y_2)$	g_{22}	-0.0090	0.0099	(-0.0241, 0.0062)
$(\ln x_2)(\ln y_3)$	g_{23}	-0.0328	0.0053	(-0.0408, -0.0247)
$(\ln x_3)(\ln y_1)$	g_{31}	-0.0543	0.0156	(-0.0785, -0.0313)
$(\ln x_3)(\ln y_2)$	g_{32}	-0.0127	0.0082	(-0.0250, -0.0006)
$(\ln x_3)(\ln y_3)$	g_{33}	0.0671	0.0148	(0.0452, 0.0897)
t	c_t	-0.0867	0.0138	(-0.1073, -0.0664)
t^2	c_{tt}	0.0163	0.0038	(0.0107, 0.0219)
$(\ln x_1)t$	g_{x1t}	-0.0088	0.0097	(-0.0230, 0.0056)
$(\ln x_2)t$	g_{x2t}	0.0028	0.0079	(-0.0094, 0.0145)
$(\ln x_3)t$	g_{x3t}	0.0053	0.0061	(-0.0036, 0.0145)
$(\ln y_1)t$	g_{y1t}	-0.0229	0.0047	(-0.0300, -0.0158)
$(\ln y_2)t$	g_{y2t}	0.0002	0.0022	(-0.0031, 0.0034)
$(\ln y_3)t$	g_{y3t}	0.02269	0.0046	(0.0158, 0.0297)

TABLE 2 REGULARITY VIOLATIONS

Regularity conditions	Regularity violations (at the posterior mean)	pdf > 0 (in inadmissible region)
<i>Monotonicity</i>		
$k_1 \leq 0$	11.59%	89.21%
$k_2 \leq 0$	0%	0.57%
$k_3 \leq 0$	69.29%	98.80%
$r_1 \geq 0$	0%	5.03%
$r_2 \geq 0$	6.51%	42.92%
$r_3 \geq 0$	0.34%	0.74%
<i>Curvature</i>		
All the principal minors of: $\widetilde{\mathbf{F}}$ are negative, and $\widetilde{\mathbf{H}}$ is positive semidefinite	100% 16.15%	100% 100%

Figure 1. Estimated Distributions of the Shadow Shares from Unconstrained Model Evaluated at Mean Prices in 2005

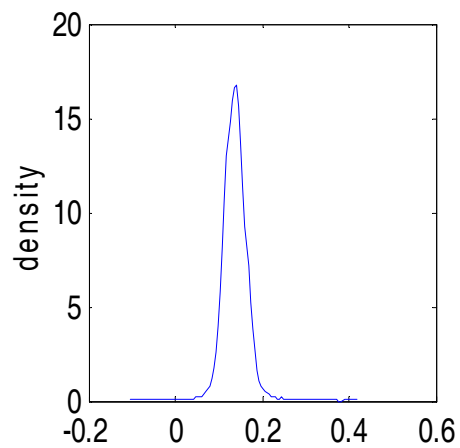


Figure 1.1: labor share

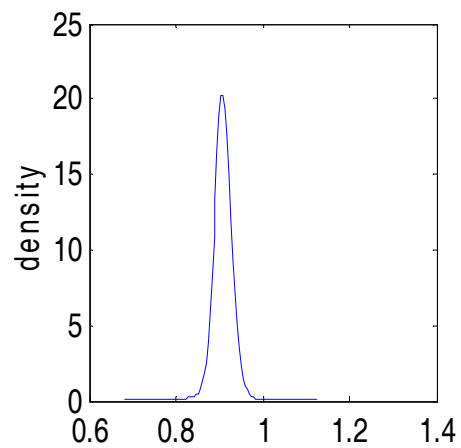


Figure 1.2: fund share

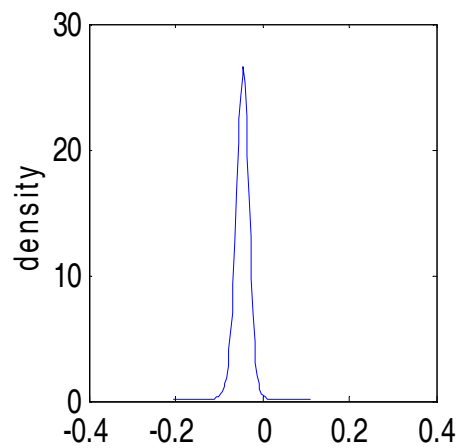


Figure 1.3: capital share

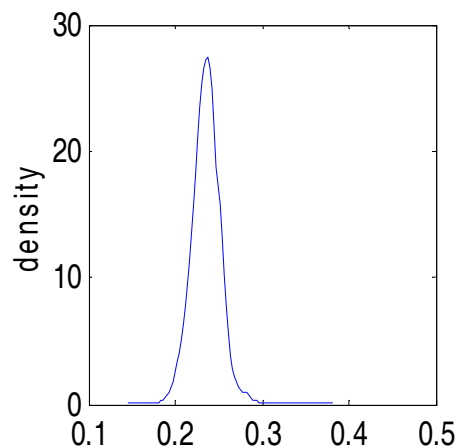


Figure 1.4: securities share

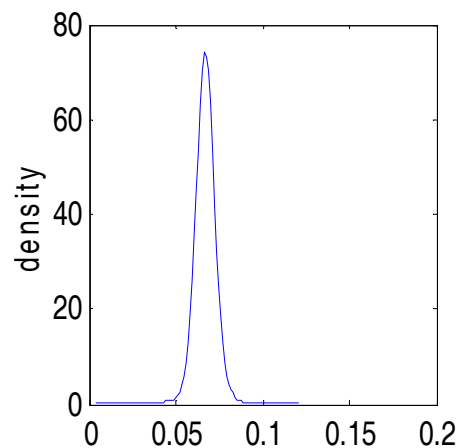


Figure 1.5: consumer loan share

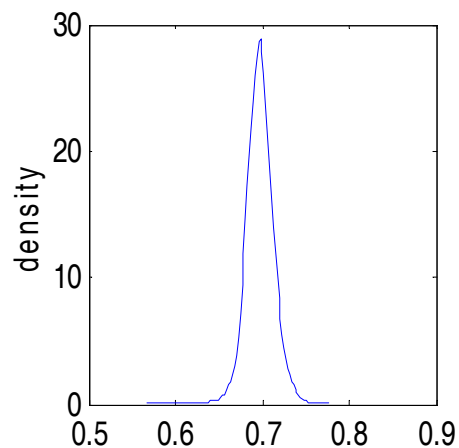


Figure 1.6: non-consumer loan share

TABLE 3

PARAMETER ESTIMATES FROM THE CONSTRAINED MODEL

Variable	Parameter	Estimate	Standard deviation	90% posterior coverage regions
intercept	a_0	0.2548	0.0194	(0.2222, 0.2873)
$\ln x_1$	b_1	-0.1166	0.0223	(-0.1580, -0.0825)
$\ln x_2$	b_2	-0.8705	0.0223	(-0.9094, -0.8363)
$\ln x_3$	b_3	-0.0521	0.0151	(-0.0763, -0.0283)
$(\ln x_1)^2$	b_{11}	-0.0288	0.0112	(-0.0465, -0.0092)
$(\ln x_2)^2$	b_{22}	0.0119	0.0223	(-0.0246, 0.0488)
$(\ln x_3)^2$	b_{33}	0.0076	0.0047	(0.0011, 0.0162)
$(\ln x_1)(\ln x_2)$	b_{12}	0.0140	0.0145	(-0.0105, 0.0370)
$(\ln x_1)(\ln x_3)$	b_{13}	0.0059	0.0042	(-0.0010, 0.0127)
$(\ln x_2)(\ln x_3)$	b_{23}	-0.0243	0.0084	(-0.0386, -0.0112)
$\ln y_1$	a_1	0.3996	0.0169	(0.3741, 0.4301)
$\ln y_2$	a_2	0.1171	0.0059	(0.1069, 0.1264)
$\ln y_3$	a_3	0.4834	0.0171	(0.4524, 0.5098)
$(\ln y_1)^2$	a_{11}	0.0720	0.0076	(0.0590, 0.0837)
$(\ln y_2)^2$	a_{22}	0.0099	0.0007	(0.0087, 0.0111)
$(\ln y_3)^2$	a_{33}	0.0865	0.0054	(0.0776, 0.0951)
$(\ln y_1)(\ln y_2)$	a_{12}	0.0023	0.0023	(-0.0014, 0.0061)
$(\ln y_1)(\ln y_3)$	a_{13}	-0.0743	0.0062138728	(-0.0842, -0.0639)
$(\ln y_2)(\ln y_3)$	a_{23}	-0.0122	0.0021915983	(-0.0158, -0.0086)
$(\ln x_1)(\ln y_1)$	g_{11}	-0.0264	0.010685836	(-0.0439, -0.0079)
$(\ln x_1)(\ln y_2)$	g_{12}	0.0123	0.0046178735	(0.0045, 0.0203)
$(\ln x_1)(\ln y_3)$	g_{13}	0.0141	0.010467973	(-0.0031, 0.0311)
$(\ln x_2)(\ln y_1)$	g_{21}	0.0582	0.012142793	(0.03789, 0.0790)
$(\ln x_2)(\ln y_2)$	g_{22}	0.0064	0.0056902557	(-0.0042, 0.0152)
$(\ln x_2)(\ln y_3)$	g_{23}	-0.0647	0.011890464	(-0.0848, -0.0443)
$(\ln x_3)(\ln y_1)$	g_{31}	-0.0075	0.0032	(-0.0130, -0.0027)
$(\ln x_3)(\ln y_2)$	g_{32}	-0.0023	0.0014	(-0.0046, -0.0002)
$(\ln x_3)(\ln y_3)$	g_{33}	0.0098	0.0036	(0.0041, 0.0158)
t	c_t	-0.0914	0.0120	(-0.1116, -0.0708)
t^2	c_{tt}	0.0183	0.0032	(0.0126, 0.0238)
$(\ln x_1)t$	g_{x1t}	-0.0056	0.0047	(-0.0133, 0.0020)
$(\ln x_2)t$	g_{x2t}	0.0040	0.0046	(-0.0029, 0.0116)
$(\ln x_3)t$	g_{x3t}	0.0010	0.0011	(-0.0009, 0.0028)
$(\ln y_1)t$	g_{y1t}	-0.0158	0.0036	(-0.0219, -0.0098)
$(\ln y_2)t$	g_{y2t}	-0.0021	0.0010	(-0.0037, -0.0003)
$(\ln y_3)t$	g_{y3t}	0.0179	0.0036	(0.0120, 0.0240)

TABLE 4 REGULARITY VIOLATIONS (CONSTRAINED MODEL)

Regularity conditions	Regularity violations (at the posterior mean)	pdf > 0 (in inadmissible region)
<i>Monotonicity</i>		
$k_1 \leq 0$	0%	0%
$k_2 \leq 0$	0%	0%
$k_3 \leq 0$	0%	0%
$r_1 \geq 0$	0%	0%
$r_2 \geq 0$	0%	0%
$r_3 \geq 0$	0%	0%
<i>Curvature</i>		
All the principal minors of $\widetilde{\mathbf{F}}$ are negative, and	0%	0%
$\widetilde{\mathbf{H}}$ is positive semidefinite	0%	0%

Figure 2. Estimated Distributions of the Shadow Shares from Constrained Model Evaluated at Mean Prices in 2005

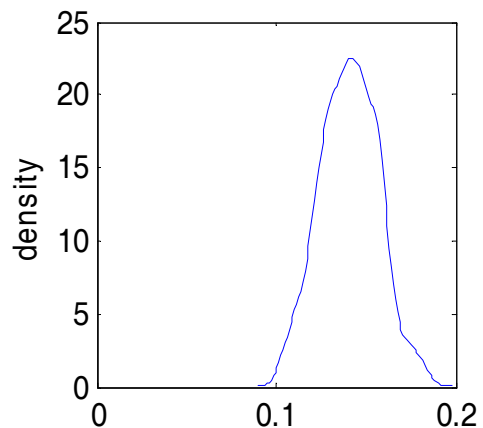


Figure 2.1: labor share

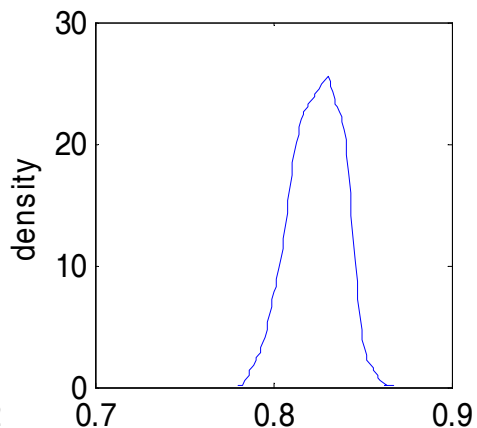


Figure 2.2: fund share

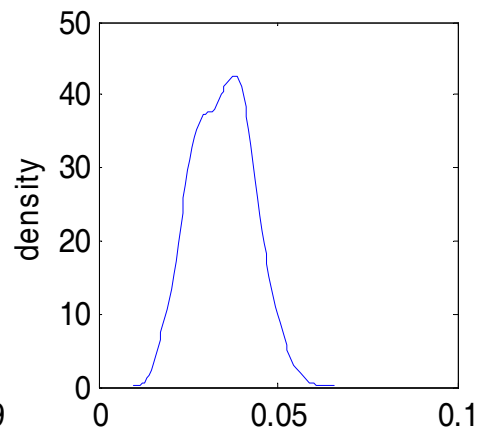


Figure 2.3: capital share

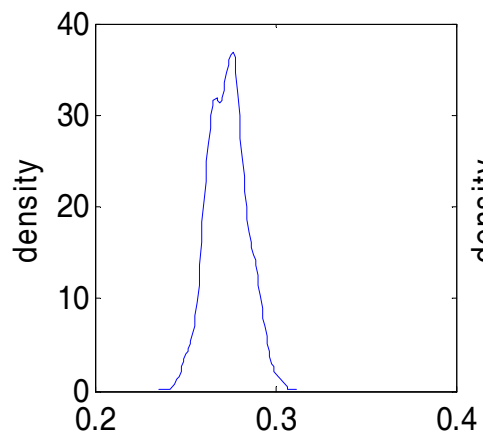


Figure 2.4: securities share

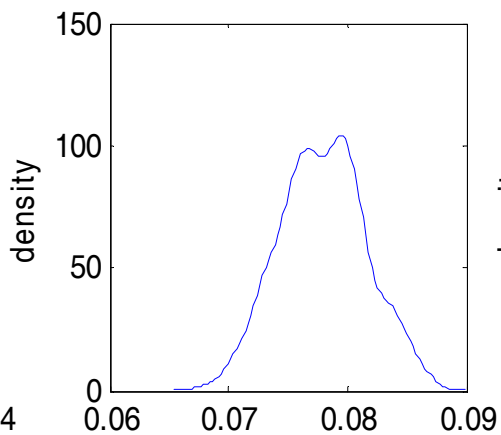


Figure 2.5: consumer loan share

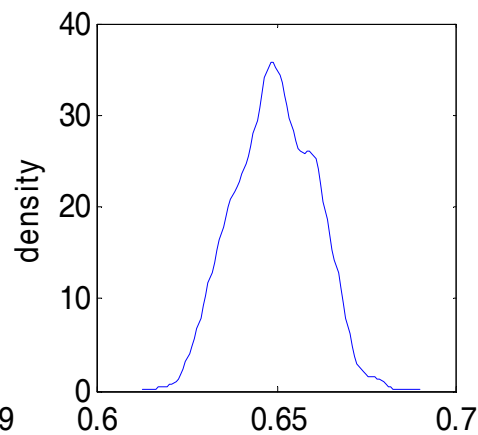


Figure 2.6: non-consumer loan share

TABLE 5.1 AVERAGE TECHNICAL EFFICIENCY

Year	Average technical efficiency	Standard deviation	90% posterior coverage regions
2000	0.9341	0.0048	(0.9259, 0.9418)
2001	0.9249	0.0057	(0.9151, 0.9339)
2002	0.9294	0.0052	(0.9203, 0.9376)
2003	0.9277	0.0054	(0.9183, 0.9361)
2004	0.9243	0.0057	(0.9144, 0.9331)
2005	0.9269	0.0055	(0.9174, 0.9357)

TABLE 5.2 DISTRIBUTION OF TECHNICAL EFFICIENCY ACROSS BANKS

Year	Minimum	Maximum	Standard deviation	5% percentile	95% percentile
2000	0.5242	0.9726	0.0335	0.9083	0.9585
2001	0.5245	0.9719	0.0365	0.8882	0.9531
2002	0.4589	0.9789	0.0406	0.8855	0.9616
2003	0.4593	0.9868	0.0441	0.8770	0.9673
2004	0.3717	0.9779	0.0476	0.8700	0.9638
2005	0.3508	0.9763	0.0474	0.8773	0.9603

TABLE 6. RETURNS TO SCALE

Year	Average returns to scale	Standard deviation	90% posterior coverage regions
2000	1.0365	0.0061	(1.0266, 1.0465)
2001	1.0394	0.0047	(1.0315, 1.0474)
2002	1.0413	0.0041	(1.0346, 1.0485)
2003	1.0446	0.0042	(1.0378, 1.0517)
2004	1.0509	0.0047	(1.0430, 1.0583)
2005	1.0560	0.0058	(1.0462, 1.0659)

TABLE 7.1. TECHNICAL CHANGE

Year	Average technical change	Standard deviation	90% posterior coverage regions
2000	0.0684	0.0085	(0.0540, 0.0829)
2001	0.0507	0.0055	(0.0415, 0.0598)
2002	0.0335	0.0030	(0.0282, 0.0383)
2003	0.0153	0.0030	(0.0098, 0.0199)
2004	-0.0051	0.0054	(-0.0143, 0.0040)
2005	-0.0247	0.0083	(-0.0380, -0.0102)

TABLE 7.2. TECHNICAL CHANGE ESTIMATES
FROM ALTERNATIVE MODELS

Year	Model 1	Model 2	Model 3
2000	0.0682 (0.0568, 0.0795)	0.0660 (0.0517, 0.0806)	0.0600 (0.0460, 0.0733)
2001	0.0504 (0.0434, 0.0576)	0.0509 (0.0415, 0.0605)	0.0454 (0.0366, 0.0535)
2002	0.0333 (0.0294, 0.0379)	0.0365 (0.0310, 0.0419)	0.0311 (0.0265, 0.0355)
2003	0.0151 (0.0108, 0.0198)	0.0206 (0.0154, 0.0260)	0.0159 (0.0098, 0.0213)
2004	-0.0053 (-0.0139, 0.0024)	0.0018 (-0.0075, 0.0111)	-0.0014 (-0.0115, 0.0092)
2005	-0.0248 (-0.0376, -0.0129)	-0.0157 (-0.0303, -0.0013)	-0.0179 (-0.0333, -0.0016)

Note: The 90% posterior coverage regions are shown in parentheses.

TABLE 8. PRODUCTIVITY CHANGE

Year	Average productivity change	Efficiency change		Scale effect		Technical change	
		Estimates	Contribution	Estimates	Contribution	Estimates	Contribution
2001	0.0662 (0.0530, 0.0794)	0.0092 (0.0013, 0.0172)	13.90%	0.0063 (0.0051, 0.0076)	9.52%	0.0507 (0.0415, 0.0598)	76.59%
2002	0.0311 (0.0211, 0.0409)	-0.0045 (-0.0124, 0.0034)	-14.47%	0.0020 (0.0017, 0.0024)	6.43%	0.0335 (0.0282, 0.0383)	107.72%
2003	0.0202 (0.0107, 0.0296)	0.0017 (-0.0059, 0.0094)	8.42%	0.0032 (0.0027, 0.0037)	15.84%	0.0153 (0.0098, 0.0199)	75.74%
2004	0.0041 (-0.0087, 0.0166)	0.0034 (-0.0046, 0.0113)	82.93%	0.0059 (0.0050, 0.0067)	143.90%	-0.0051 (-0.0143, 0.0040)	-124.39%
2005	-0.0225 (-0.0395, -0.0049)	-0.0026 (-0.0106, 0.0055)	11.56%	0.0047 (0.0039, 0.0056)	-20.89%	-0.0247 (-0.0380, -0.0102)	109.78%
Average	0.0198	0.0014	7.27%	0.0044	22.30%	0.0139	70.33%

Notes: The 90% posterior coverage regions are shown in parentheses.

TABLE 9. PRODUCTIVITY CHANGE WHEN INSTRUMENTAL VARIABLES ARE USED

Year	Average productivity change	Efficiency change		Scale effect		Technical change	
		Estimates	Contribution	Estimates	Contribution	Estimates	Contribution
2001	0.0491 (0.0179, 0.0784)	0.0021 (-0.0150, 0.0192)	4.29%	0.0040 (0.0004, 0.0081)	8.21%	0.0430 (0.0204, 0.0645)	87.50%
2002	0.0360 (0.0127, 0.0592)	0.0007 (-0.0162, 0.0178)	2.06%	0.0020 (0.0007, 0.0033)	5.57%	0.0333 (0.0192, 0.0472)	92.36%
2003	0.0263 (0.0062, 0.0463)	0.0008 (-0.0163, 0.0178)	2.97%	0.0022 (0.0010, 0.0036)	8.50%	0.0233 (0.0145, 0.0333)	88.53%
2004	0.0183 (-0.0039, 0.0407)	0.0032 (-0.0140, 0.0206)	17.48%	0.0029 (0.0012, 0.0052)	16.13%	0.0121 (-0.0004, 0.0247)	66.39%
2005	0.0033 (-0.0268, 0.0333)	0.0004 (-0.0173, 0.0179)	11.89%	0.0023 (0.0007, 0.0044)	71.30%	0.0005 (-0.0207, 0.0214)	16.81%
Average	0.0266	0.0014	7.74%	0.0027	21.94%	0.0224	70.32%

Notes: The 90% posterior coverage regions are shown in parentheses.