Accepted Manuscript

Returns to scale at large banks in the U.S.: A random coefficient stochastic frontier approach

Guohua Feng, Xiaohui Zhang

PII: S0378-4266(13)00413-5
DOI: http://dx.doi.org/10.1016/j.jbankfin.2013.10.012
Reference: JBF 4251

To appear in: Journal of Banking & Finance

Received Date: 22 June 2013
Revised Date: 15 October 2013
Accepted Date: 27 October 2013

Please cite this article as: Feng, G., Zhang, X., Returns to scale at large banks in the U.S.: A random coefficient stochastic frontier approach, Journal of Banking & Finance (2013), doi: http://dx.doi.org/10.1016/j.jbankfin.2013.10.012

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
Returns to scale at large banks in the U.S.: A random coefficient stochastic frontier approach

October 30, 2013
Abstract

This paper investigates the returns to scale of large banks in the U.S. over the period 1997–2010. This investigation is performed by estimating a random coefficient stochastic distance frontier model in the spirit of Tsionas (2002) and Greene (2005, 2008). The primary advantage of this model is that its coefficients can vary across banks, thereby allowing for unobserved technology heterogeneity among large banks in the U.S. We find that failure to consider unobserved technology heterogeneity results in a misleading ranking of banks and mismeasured returns to scale. Our results show that the majority of large banks in the U.S. exhibit constant returns to scale. In addition, our results suggest that banks of the same size can have different levels of returns to scale and there is no clear pattern among large banks in the U.S. concerning the relationship between asset size and returns to scale, due to the presence of technology heterogeneity.

*JEL classification:* C11; D24; G21.

*Keywords:* Returns to Scale; Random Coefficient Stochastic Distance Frontier Model; Bayesian Estimation.
1. Introduction

Over the past two decades, the increasing dominance of large banks in the U.S. banking industry, caused by fundamental regulatory changes and technological and financial innovations, has stimulated considerable research into returns to scale at large banks in the U.S. Specifically, major regulatory changes include the removal of restrictions on interstate banking and branching and the elimination of restrictions against combinations of banks, securities firms, and insurance companies, while 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. One of the most important consequences of these changes is the increasing concentration of industry assets among large banks. According to Jones and Critchfield (2005), the asset share of large banks (those with assets in excess of $1 billion) increased from 76 percent in 1984 to 86 percent in 2003. In the meantime, the average size of those banks increased from $4.97 billion to $15.50 billion. This has raised concern that some banks might be too large to operate efficiently, stimulating a substantial body of research into returns to scale at large banks in the U.S.. For excellent reviews, see Berger et al. (1993) and Berger et al. (1999).

However, few articles explicitly allow production technology to be heterogeneous, even though studies have found that unobserved technology heterogeneity is widely present in the U.S. banking industry. For example, a growing body of literature (Saloner and Shepard, 1995; Akhavein et al., 2005) suggests that diffusion of new technologies among banks takes time, because banks adopt new technologies at different times according to factors such as bank size, organizational structure, profitability, geographic location, and market structure. Specifically, Akhavein et al. (2005) finds that out of a sample of 96 large banks in the U.S., banks with more branches adopt new technologies earlier, as do those located in the New York Federal Reserve district. This slow diffusion process suggests that large banks in the U.S. do not always have access to the same technologies. To give another example, many studies (Coles et al., 2004; Berger et al., 2005) have found that banks with different organizational structures use different production technologies. Specifically,
centralized banks with their hierarchical structures tend to employ “hard” information-based production technologies (such as credit scoring technologies), whereas decentralized banks with their flatter organizational structures tend to employ “soft” information-based production technologies (such as traditional underwriting techniques). Further, these studies (for example, Canales and Nanda, 2012) finds that large banks differ in the degree of “centralizedness”, indicating that these banks may use different combinations of “hard” and “soft” information-based production technologies. To give a third example, a number of studies found that banks with different business models tend to employ different production technologies. For instance, Rossi (1998) find that mortgage banks (whether large or small) rely more heavily on automated lending technologies than do full-service commercial banks. In sum, all of these examples indicate that unobserved technology heterogeneity is widespread among large banks in the U.S., thus calling for a model that is suitable for modeling returns to scale in the presence of unobserved technology heterogeneity.

For the first time in this literature, we estimate a random coefficient translog stochastic distance frontier (SDF) model, which allows for unobserved technology heterogeneity. Essentially, this model is a variant of the random coefficient stochastic cost frontier model proposed by Tsionas (2002), with the former model based on the output distance function and the latter model based on the cost function. The main feature of this model is that its coefficients can vary across banks, thus allowing production technology to be heterogeneous across banks. Econometrically, this model can be obtained by permitting the coefficients of the standard fixed coefficient SDF model to vary across banks by drawing the coefficients from a multivariate normal distribution. More specifically, this drawing process can be achieved by first decomposing the bank-specific coefficient vector (denoted by $\beta_i$, where $i = 1, 2, ..., K$ indexes banks and $K$ is the number of banks) into two parts: a mean vector (denoted by $\overline{\beta}$) and a random vector (denoted by $\delta_i$) and then drawing the random vector from a multivariate normal distribution with mean zero. For excellent discussions on the specification of random coefficient stochastic frontier models, see Tsionas (2002) and Greene (2005, 2008).

---

1Hard information refers to information that is quantitative and can be credibly transmitted across physical or organizational distances, whereas soft information refers to the opposite.
A major advantage of the random coefficient translog SDF model over the commonly-used fixed coefficient SDF model is that the former model can provide a much better approximation to underlying arbitrary heterogeneous technologies than the latter model. This is because the fixed coefficient translog SDF model can only approximate the underlying true ‘average’ technology to the second order, whereas the random coefficient translog SDF model can not only approximate the underlying true average technology to the second order via its mean coefficients (i.e., $\beta$), but can also approximate each of the underlying true heterogeneous technologies to the second order via its firm-specific coefficients (i.e., $\beta_i$). Specifically, let $y$ denote the $1 \times M$ vector of outputs, $x$ denote the $1 \times N$ vector of inputs, and $w \equiv (y, x)$ denote the output and input vector. In addition, let $f^i (w)$ (for $i = 1, 2, ..., K$), a set of arbitrary output distance functions, represent the underlying ‘true’ heterogeneous technologies; $\ln D^i_o (w)$ (for $i = 1, 2, ..., K$) denote the random coefficient translog output distance function; and $\ln D_o (w)$ denote the fixed coefficient translog output distance function. For the random coefficient translog output distance function, it is straightforward to show by following the spirit of Diewert (1973) that the firm-specific translog output distance function, $\ln D^i_o (w)$, can approximate its corresponding arbitrary output distance function, $\ln f^i (w)$, to the second order at a point $w^*$, by solving the system of equations: $\ln D^i_o (w^*) = \ln f^i (w^*)$, $\partial \ln D^i_o (w^*) / \partial \ln w_m = \partial \ln f^i (w^*) / \partial \ln w_m$ (for $m = 1, 2, ..., M - 1, M + 1, ..., M - 1 + N$), and $\partial^2 \ln D^i_o (w^*) / \partial \ln w_m \partial \ln w_n = \partial^2 \ln f^i (w^*) / \partial \ln w_m \partial \ln w_n$ (for $1 < m < n < M + N$). With $\beta_i$ determined by solving the above system of equations, it is also straightforward to show that the translog output distance function with the mean coefficient vector (i.e., $\bar{\beta}$) can approximate the true ‘average’ technology (i.e., $\left[ \sum_{i=1}^{K} \ln f^i (w^*) \right] / K$) to the second order\textsuperscript{3}. For the fixed coefficient translog output distance function, note that it can be obtained from the random coefficient output distance function by setting the random component of $\beta_i$ to zero (i.e., $\delta_i = 0$). In other words, the fixed coefficient translog output distance function is essentially the translog output distance function with the mean coefficient vector (i.e., $\bar{\beta}$), suggesting that the fixed coefficient translog output

\textsuperscript{2}A detailed proof is available on request.

\textsuperscript{3}A detailed proof is available on request.
distance function can also approximate the true ‘average’ technology (i.e., $\sum_{i=1}^{K} \ln f^i(w^*) / K$) to the second order. However, unlike its random coefficient counterpart, the fixed coefficient translog output distance function is incapable of approximating $\ln f^i(w^*)$ to the second order.

The advantage of the random coefficient translog SDF model in approximating the underlying ‘true’ heterogeneous technologies means that we can estimate the returns to scale for each bank with more accuracy. As discussed above, with the random coefficient SDF model we can obtain a separate frontier for each bank (i.e., $\ln D^i(w)$), implying that we can measure returns to scale for each bank on the bank’s own frontier. In contrast, with the fixed coefficient SDF model we can only obtain one single frontier that provides a second order approximation to the true ‘average’ technology, implying that we are restricted to measuring returns to scale for all banks on this single frontier. This restriction implies that estimates of returns to scale for all banks that do not operate with the average technology would be biased. The consequences of this bias can be serious, especially when production technologies are very heterogeneous because in this case technologies for most banks would be different from the average technology. Taking our empirical results on the U.S. large banks for example, we compute the difference or bias in returns to scale (in absolute value) between the fixed coefficient SDF model and the random coefficient SDF model for each bank and find that the mean of the bias can be as large as 0.0735 and the maximum of the bias can be as large as 0.3752. We also calculate the Spearman rank correlation coefficient between the ranking based on the fixed coefficient SDF model and that based on the random coefficient SDF model, and find that due to the biases, there is little correlation between the two rankings. Thus to avoid the bias associated with the fixed coefficient SDF model, we choose to apply the random coefficient SDF model in this paper.

The random coefficient SDF model is estimated within a Bayesian framework. The primary reason for the choice of a Bayesian approach is that in contrast to the EM algorithm that is commonly used for finding maximum likelihood estimates of parameters in stochastic frontier models, the Bayesian procedure used in this study can produce, for each individual bank, a set of posterior distributions for all the model parameters (including latent variables) and any quantity of interest.
that can be computed as a function of the model parameters (Tsionas, 2002). In particular, the Bayesian procedure enables us to obtain, for each individual bank, a posterior distribution for our measure of returns to scale that can be computed as a nonlinear function of the parameters of the output distance function. In practice, this posterior distribution enables us to compute a credible interval for returns to scale for each bank in each period, which in turn can be used to determine if the bank faces increasing, constant, or decreasing returns to scale in the period (see Section 5.3).

Finally, we apply the above framework to the banks in the U.S. with assets in excess of $1 billion. Our results show that the majority of large banks in the U.S. exhibit constant returns to scale. In addition, our results suggest that banks of the same size can have different levels of returns to scale and there is no clear pattern among large banks in the U.S. concerning the relationship between asset size and returns to scale, due to the presence of technology heterogeneity.

The rest of the paper is organized as follows. In Section 2, we briefly discuss the output-distance-function-based measure of returns to scale. In Section 3, we present the random coefficient translog stochastic distance frontier model. In Section 4, we discuss the Bayesian procedure for the estimation of the true random translog stochastic distance frontier model. In Section 5, we describe data used in this work, apply our methodology to the U.S. banking industry, discuss the effects of incorporating unobserved technology heterogeneity, and report our estimates of returns to scale. Section 6 summarizes and concludes the paper.

2. The Output-Distance-Function-Based Measure of Returns to Scale

We begin by defining the output distance function, on which our measure of returns to scale is based. Consider a production technology where a bank uses \( N \) inputs \( x_t = (x_{t1}, x_{t2}, \ldots, x_{tN})' \) to produce \( M \) outputs \( y_t = (y_{t1}, y_{t2}, \ldots, y_{tM})' \) at time \( t = 1, 2, \ldots, T \). We assume that for time period \( t \), the production technology can be described by the output set

\[
P^t(x^t) = \{ y^t : y^t \text{ is producible from } x^t \}.
\]
Following Shephard (1970) or Färe et al. (1994), the output distance function is defined as

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

This function gives the minimum amount by which an output vector can be deflated and remains producible for a given input vector. It is non-decreasing, convex and linearly homogeneous in outputs, and non-increasing and quasi-convex in inputs. For notational simplicity, we follow the common practice in the literature and model the effect of time through an exogenous time variable, \( t \). Thus, the output distance function can be rewritten as \( D_o(y, x, t) \).

Let \( \varepsilon_n \) denote the elasticity of the output distance function with respect to \( x_n \)

\[ \varepsilon_n = \frac{\partial \ln D_o(y, x, t)}{\partial \ln x_n}. \] (2)

The output-distance-function-based measure of returns to scale (RTS) can then be defined as in Caves et al. (1982)

\[ RTS = -\sum_{n=1}^{N} \varepsilon_n. \] (3)

This measure is discussed in more details in Färe and Grosskopf (1994, p.103) and has been recently used in a number of studies such as Orea (2002).

3. The Random Coefficient Stochastic Distance Frontier Model

To estimate the output-distance-function-based measure of returns to scale given by (3), we need to parameterize and calculate the parameters of the output distance function, \( D_o(y, x, t) \). In this paper, we choose to parameterize \( D_o(y, x, t) \) as a translog function, mainly because it is easy to impose the linear homogeneity property with this functional form. Authors who have used a translog output distance function in empirical work include, but are not limited to, Färe et al. (1993) and O’Donnell and Coelli (2005). As discussed in the Introduction, the random coefficient translog stochastic distance frontier (SDF) model can be obtained by allowing the coefficients of the standard fixed coefficient translog SDF model to vary across banks. Thus in this section we first
specify the standard fixed coefficient translog SDF model and then specify the random coefficient translog SDF model.

3.1. Specification of the standard fixed coefficient stochastic distance frontier model

The standard fixed coefficient translog output distance function is written as

\[
\ln D_o(y, x, t) = a_0 + \sum_{m=1}^{M} a_m \ln y_m + \frac{1}{2} \sum_{m=1}^{M} \sum_{p=1}^{M} a_{mp} \ln y_m \ln y_p
\]

\[
+ \sum_{n=1}^{N} b_n \ln x_n + \frac{1}{2} \sum_{n=1}^{N} \sum_{j=1}^{N} b_{nj} \ln x_n \ln x_j + \delta \tau t + \frac{1}{2} \delta \tau \tau t^2
\]

\[
+ \sum_{n=1}^{N} \sum_{m=1}^{M} g_{nm} \ln x_n \ln y_m + \sum_{m=1}^{M} \delta_m t \ln y_m + \sum_{n=1}^{N} \rho_n t \ln x_n, \quad (4)
\]

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

\[
\sum_{m=1}^{M} a_m = 1; \quad \sum_{p=1}^{M} a_{mp} = 0, m = 1, \ldots, M; \quad \sum_{m=1}^{M} g_{nm} = 0, n = 1, \ldots, N; \quad \sum_{m=1}^{M} \delta_m = 0. \quad (5)
\]

(4) cannot be directly estimated since \( D_o(y, x, t) \) is not observable. To deal with this problem, we follow O’Donnell and Coelli (2005) and exploit the linear homogeneity property of \( D_o(y, x, t) \) to transform (4) into an estimable regression equation in the form of a standard stochastic frontier model. Specifically, normalizing (4) by output \( M \) gives

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

\[
= - \ln y_M + \ln \left[ D_o(y, x, t) \right]
\]

\[
= - \ln y_M - u, \quad (6)
\]

where the first equality is obtained by the linear homogeneity of \( D_o(y, x, t) \) in outputs and \( u \) is defined as follows \( u \equiv - \ln D_o(y, x, t) \geq 0 \). By definition, \( u \) is a measure of inefficiency, which
is unobservable and non-negative. Rearranging (6) yields

\[- \ln y_M = \ln D_o \left( \frac{y}{y_M}, x, t \right) + u. \] (7)

Assuming that \( u \) follows a distribution with non-negative support and adding an independently and identically normally distributed error term, \( v \), (7) can be further written as

\[- \ln y_M = \ln D_o \left( \frac{y}{y_M}, x, t \right) + u + v. \] (8)

The above procedure thus transforms (4) into (8), an estimable equation in the form of a standard stochastic frontier model with two error terms, with one (i.e., \( v \)) capturing statistical noise and the other (i.e., \( u \)) representing inefficiency. Note that, like other standard stochastic frontier models, (8) does not allow for technology heterogeneity (parameter variation). Thus, in this sense, we refer to (8) as the fixed coefficient SDF model.

By expanding the first term on the right hand side of (8), the fixed coefficient SDF model in (8) can be written more explicitly as

\[- \ln y_M = a_0 + \sum_{m=1}^{M-1} a_m \ln \left( \frac{y_m}{y_M} \right) + \frac{1}{2} \sum_{m=1}^{M-1} \sum_{p=1}^{M-1} a_{mp} \ln \left( \frac{y_m}{y_M} \right) \ln \left( \frac{y_p}{y_M} \right) + \sum_{n=1}^{N} b_n \ln x_n + \frac{1}{2} \sum_{n=1}^{N} \sum_{j=1}^{N} b_{nj} \ln x_n \ln x_j + \delta \tau t + \frac{1}{2} \delta \tau \tau t^2 + \sum_{n=1}^{N} \sum_{m=1}^{M-1} g_{nm} \ln x_n \ln \left( \frac{y_m}{y_M} \right) + \sum_{m=1}^{M-1} \delta_m t \ln \left( \frac{y_m}{y_M} \right) + \sum_{n=1}^{N} \rho_n t \ln x_n + u + v. \] (9)

Within a panel data framework, the fixed coefficient SDF model in (9) can be notationally simplified as

\[ q_{it} = z_{it}' \beta + u_{it} + v_{it}, \] (10)

where \( i = 1, \ldots, K \) indicate banks; \( t = 1, \ldots, T \) indicate time; \( q_{it} = - \ln y_{M,it} \); \( z_{it} \) is a \( s \times 1 \) vector
comprising all the variables which appear on the right hand side of (9); β refers to the corresponding vector of coefficients of the translog function (including the intercept); \( v_{it} \sim i.i.d. N(0, \sigma^2_v) \) as mentioned above; and \( u_{it} \sim i.i.d. \exp(\lambda^{-1}) \) with ‘exp’ denoting an exponential distribution with an unknown parameter \( \lambda \).

3.2. Specification of the random coefficient stochastic distance frontier model

Having defined the fixed coefficient SDF model, we now define the random coefficient SDF model. As noted in the Introduction, this can be done by permitting \( \beta \) in (10) to differ across banks

\[
q_{it} = z_{it}'\beta_i + u_{it} + v_{it},
\]

(11)

where \( \beta_i \) follows a multi-variate normal distribution

\[
\beta_i \sim i.i.d. N(\bar{\beta}, \Omega),
\]

(12)

where \( \bar{\beta} \) is the mean vector of \( \beta_i \) and \( \Omega = [\omega_{11}, \omega_{12}, \ldots; \omega_{1s}; \ldots; \omega_{s1}, \omega_{s2}, \ldots; \omega_{ss}] \) is a conformable positive definite covariance matrix. As noted in the Introduction, the model, given by (11) and (12), is a variant of the random coefficient stochastic cost frontier model proposed by Tsionas (2002), with the former model based on the output distance function and the latter model on the cost function.

Decomposing \( \beta_i \) into its mean vector (\( \bar{\beta} \)) and a random vector (\( \delta_i \)) yields

\[
\beta_i = \bar{\beta} + \delta_i,
\]

(13)

where \( \delta_i \sim i.i.d. N(0, \Omega) \). By substituting (13) into (11), the random coefficient SDF model in (11) can be rewritten as

\[
q_{it} = z_{it}'\bar{\beta} + z_{it}'\delta_i + u_{it} + v_{it}.
\]

(14)

where \( v_{it} \) and \( u_{it} \) are defined the same as in (10). Note that when \( \delta_i = 0_{s \times 1} \) (i.e., when there is no technology heterogeneity), the random coefficient translog SDF model in (14) will reduce to the
fixed coefficient SDF model in (10).

Following the common practice in this stochastic frontier literature, we assume that $u_{it}$ and $v_{it}$ are independent of each other and also of $z_{it}$. Also, it is convenient at this point to introduce some matrix notations. In what follows we use the following definitions: $q_i = (q_{i1}, \ldots, q_{iT})'$, $q = (q_{11}, \ldots, q_{1T}, \ldots, q_{K1}, \ldots, q_{KT})'$, $z_i = (z_{i1}, \ldots, z_{iT})$, $z = (z_1; \ldots; z_K)$, $u_i = (u_{i1}, \ldots, u_{iT})'$, $u = (u_{11}, \ldots, u_{1T}, \ldots, u_{K1}, \ldots, u_{KT})'$, $v = (v_{11}, \ldots, v_{1T}, \ldots, v_{K1}, \ldots, v_{KT})'$, and $\Delta = (\delta_1, \ldots, \delta_K)$.

4. Bayesian Estimation

In this section we briefly detail the Bayesian procedure for estimating the random coefficient SDF model (i.e., (14)). For the fixed coefficient SDF model (i.e., (10)), its Bayesian formulation has been widely documented in the stochastic frontier literature (Koop and Steel, 2003; O’Donnell and Coelli, 2005), and thus is not discussed in this paper.

We start by specifying priors for the parameters. To facilitate comparison of empirical results between the fixed coefficient SDF model and the random coefficient SDF model, we use the same priors for the parameters that are common to these two models. Following Koop and Steel (2003) and O’Donnell and Coelli (2005), we use a flat prior for $\bar{\beta}$

$$p(\bar{\beta}) \propto 1, \tag{15}$$

and the following distribution for $h_v$

$$p(h_v) \propto \frac{1}{h_v}, \quad \text{where } h_v = \frac{1}{\sigma_v^2} > 0. \tag{16}$$

(16) implies that $h_v$ is fully determined by the likelihood function — see the conditional posterior distribution for $h_v$ in equation (24).

As mentioned above, we choose an exponential distribution for $u_{it}$ (Tsionas, 2002; O’Donnell and Coelli, 2005). Since the exponential distribution is a special case of the gamma distribution,
the prior for $u_{it}$ can be written as

$$p(u_{it} | \lambda^{-1}) = f_{\text{Gamma}}(u_{it} | 1, \lambda^{-1}).$$ \hspace{1cm} (17)

According to Fernandez et al. (1997), in order to obtain a proper posterior we need a proper prior for the parameter, $\lambda$. Accordingly, we use the proper prior

$$p(\lambda^{-1}) = f_{\text{Gamma}}(\lambda^{-1} | 1, -\ln \tau^*),$$ \hspace{1cm} (18)

where $\tau^*$ is the prior median of the efficiency distribution. Our best prior knowledge of the efficiency of large banks in the U.S. is the mean efficiency value of 0.9279 reported by Feng and Serletis (2010) that examines the productivity and efficiency for large banks in the US during 2000–2005. To investigate the sensitivity of our results to extreme changes to $\tau^*$, we experiment with various values of $\tau^*$ ranging from 0.50 to 0.99. The results are always the same up to the number of digits presented in Section 5, implying that our results are very robust to enormous changes in $\tau^*$.

For the random coefficient SDF model (i.e., (14)), we need further to specify priors for $\delta_i$ and $\Omega$. Following Tsionas (2002), we use the following priors

$$\delta_i \sim \text{i.i.d.} N(0, \Omega),$$ \hspace{1cm} (19)

$$\Omega \sim IW(r, \Phi),$$ \hspace{1cm} (20)

where $IW(r, \Phi)$ is an Inverse Wishart density with $r$ degrees of freedom and scale matrix $\Phi$. Following Tsionas (2002), we set $r = 1$ and $\Phi = 10^{-6} I_s$, where $I_s$ is a $s \times s$ identity matrix.
The likelihood function can be shown to be

\[
L(q|\beta, h, u, \lambda^{-1}, \delta, \Omega) = \prod_{i=1}^{K} \prod_{t=1}^{T} \left[ \sqrt{h_v/2\pi} \exp \left( -\frac{h_v}{2} (q_{it} - z_{it}'\beta - z_{it}'\delta_i - u_{it})^2 \right) \right]^{K T/2} \exp \left[ -\frac{h_v}{2} v'v \right],
\]

where \( v \) is \( KT \times 1 \) vector with \( v_{it} = q_{it} - z_{it}'\beta - z_{it}'\delta_i - u_{it} \).

Using Bayes’s Theorem and combining the likelihood function in (21) and the prior distributions in (15) – (20), we obtain the posterior joint density function

\[
f(\beta, h, u, \lambda^{-1}, \delta, \Omega | q) \propto h_v^{KT/2-1} \prod_{i=1}^{K} \prod_{t=1}^{T} \left[ |\Omega|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} \delta_i'\Omega^{-1}\delta_i \right) \right] \prod_{i=1}^{K} \prod_{t=1}^{T} \left[ \lambda^{-1} \exp \left( -\lambda^{-1}u_{it} \right) \right] \exp (\lambda^{-1} \ln \tau^*) \times |\Omega|^{-\frac{r+K+1}{2}} \exp \left[ -\frac{1}{2} \text{tr} (\Phi\Omega^{-1}) \right].
\]

The full conditional posterior distributions for \( \beta, h, u, \lambda^{-1}, \delta, \Omega \) can be shown to be

\[
p(\beta | q, h, u, \lambda^{-1}, \Delta, \Omega) \propto f_{\text{Normal}} \left( \beta | \bar{b}, h_v^{-1} \left( zz' \right)^{-1} \right),
\]

\[
p(h | q, \beta, u, \lambda^{-1}, \Delta, \Omega) \propto f_{\text{Gamma}} \left( h_v | \frac{KT+1}{2}, \frac{1}{2} v'v \right),
\]

\[
p(u_{it} | q, \beta, h, \lambda^{-1}, \Delta, \Omega) \propto f_{\text{Normal}} \left( u_{it} | q_{it} - z_{it}'\beta - z_{it}'\delta_i - (h_v\lambda)^{-1}, h_v^{-1} \right) I(u_{it} \geq 0),
\]

\[
p(\lambda^{-1} | q, \beta, h, u, \Delta, \Omega) \propto f_{\text{Gamma}} \left( \lambda^{-1} | KT+1, u'\iota_{KT} - \ln \tau^* \right),
\]

\[
p(\delta_i | q, \beta, h, u, \lambda^{-1}, \Omega) \propto f_{\text{Normal}} \left( \delta_i | \bar{\delta}, (\Omega^{-1} + h_vz_{it}z_{it}')^{-1} \right),
\]

\[
p(\Omega | q, \beta, h, u, \lambda^{-1}, \Delta) \propto IW (r + K, \Phi + \Delta\Delta'),
\]

where \( \bar{b} = (zz')^{-1}z\zeta \), where \( \zeta \) is a \( KT \times 1 \) vector with \( \zeta_{it} = q_{it} - z_{it}'\delta_i - u_{it} \); \( \iota_{KT} \) is a \( KT \times 1 \)
vector of ones; and $\delta = \left( \Omega^{-1} + h_v z_i' z_i \right)^{-1} h_v z_i (q_i - z_i' \beta - u_i)$.

As noted above, the measure of returns to scale in (3) is a function of $\beta$, $h_v$, $u$, $\lambda^{-1}$, $\delta_i$ and $\Omega$. Let $g(\beta, h_v, u, \lambda^{-1}, \delta_i, \Omega)$ represent this function. In theory, we could obtain the moments of $g(\beta, h_v, u, \lambda^{-1}, \delta_i, \Omega)$ 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 (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.

5. Data and Empirical Results

5.1. Data

The data are obtained from the Reports of Income and Condition (Call Reports). The sample covers the period 1997.Q1 – 2010.Q1, a period that includes the most recent subprime crisis. We end our sample period in 2010.Q1 because after this period, many important data required for the construction of inputs or outputs are missing. We examine only continuously operating large 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, large banks are defined to be those with assets of at least $1 billion (in 1997 dollars) in the last three sample years. This gives a total of 354 banks over 53 quarters. In selecting the relevant variables, we follow the commonly-accepted intermediation approach (Sealey and Lindley, 1977). On the input side, three inputs are included: labor, $x_1$; purchased funds and deposits, $x_2$; and physical capital, $x_3$, which includes premises and other fixed assets. On the output side, three outputs are specified: securities, $y_1$, which includes all non-loan financial assets (i.e., all financial 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 GDP deflator to the base year 1997, except for the quantity of labor.

5.2. Model comparison

Before proceeding to model comparison, we first check if the monotonicity conditions of the output distance function are satisfied for the fixed coefficient SDF model and the random coefficient SDF model. When estimating each of the two models, we generate a total of 30,000 observations, and then discard the first 10,000 as a ‘burn-in’. For the fixed coefficient SDF model, monotonicity requires that

\[ \frac{\partial \ln D_o(y, x, t)}{\partial \ln x_n} = b_n + \sum_{j=1}^{N} b_{nj} \ln x_j + \sum_{m=1}^{M} g_{nm} \ln y_m + \rho_n t \leq 0 \]

for \( n = 1, 2, 3 \) and \( \frac{\partial \ln D_o(y, x, t)}{\partial \ln y_m} = a_m + \sum_{p=1}^{M} a_{mp} \ln y_p + \sum_{n=1}^{N} g_{nm} \ln x_n + \delta_m t \geq 0 \)

for \( m = 1, 2, 3 \). To simplify these nonlinear constraints, we follow O’Donnell and Coelli (2005) and deflate the sample data so that all output and input variables have a sample mean of one and the time trend has a sample mean of zero. When evaluated at these variable means, \( \frac{\partial \ln D_o(y, x, t)}{\partial \ln x_n} \) and \( \frac{\partial \ln D_o(y, x, t)}{\partial \ln y_m} \) collapse to \( b_n \) and \( a_m \) respectively, and the monotonicity conditions can therefore be expressed as \( b_n \leq 0 \) and \( a_m \geq 0 \). We first estimate the fixed coefficient SDF model without imposing these conditions. However, the monotonicity condition with respect to labor is violated. Thus, we reestimate this model with the six monotonicity conditions imposed, by following the Bayesian procedure discussed in O’Donnell and Coelli (2005). The estimated parameters and their associated 95% credible intervals are reported in Table 1. As this table indicates, all the six estimated parameters \( b_1, b_2, b_3, a_1, a_2, \) and \( a_3 \) now have the right signs (i.e., \( b_n \leq 0 \) for \( n = 1, 2, 3 \); and \( a_m \geq 0 \) for \( m = 1, 2, 3 \)), implying the monotonicity conditions are satisfied.

For the random coefficient SDF model, ideally, the same monotonicity conditions should be satisfied for each of the large banks. Practically, however, this is not possible in our particular case due to the large number of nonlinear constraints implied by these firm-specific conditions and the large number of parameters. Specifically, we first estimate the random coefficient SDF model without these monotonicity conditions imposed. However, these conditions are not satisfied for all the large banks. We then reestimate the model with these conditions imposed for each of
the banks. However, this imposition lead to some severe convergence problems. To avoid these problems, we require that these conditions be satisfied only by the mean coefficient vector (i.e., $\beta$), but not by the firm-specific coefficient vectors (i.e., $\beta_i's$). Specifically, we first estimate the random coefficient SDF model without imposing these conditions on the mean coefficients. However, the monotonicity condition with respect to labor is violated. We therefore reestimate this model with the six monotonicity conditions imposed on the mean coefficients. The estimated parameters and their associated 95% credible intervals are reported in Table 2. As can be seen, all the six estimated parameters $b_1$, $b_2$, $b_3$, $a_1$, $a_2$, and $a_3$ have the right signs, implying the monotonicity conditions are satisfied by the mean coefficient vector.

We employ the simulation inefficiency factor (SIF) to check the convergence performance of the sampling algorithms (see, for example, Kim et al. (1998) and Zhang et al. (2009)). For the fixed coefficient SDF model, the SIF values are displayed in the last column of Table 1. As can be seen, all SIF values are less than 41 and most of them are lower than 35, a good indication of the convergence of the sampler. For the random coefficient SDF model, we need to estimate a large number (436) of parameters, including the 28 elements of the mean vector ($\bar{\beta}$), the 406 ($= 28 \times (28 + 1) / 2$) elements of the symmetric parameter covariance matrix ($\Omega$), $h_v$, and $\lambda^{-1}$. To avoid potential convergence problems due to the large number of parameters, we assume that $\Omega$ is diagonal (Tsionas, 2002). With this assumption, the number of parameters to be estimated drops significantly from 436 to 58 (including the 28 elements of $\bar{\beta}$, the 28 diagonal elements of $\Omega$, $h_v$, and $\lambda^{-1}$). The SIF values for all the coefficients of the random coefficient SDF model are presented in the last column of Table 2. As can be seen from this table, all SIF values are less than 51, indicating that the simulated chains have converged.

We now turn to compare the estimation performance of the fixed coefficient SDF model (i.e., (10)) and that of the random coefficient SDF model (i.e., (14)), using a deviance information criterion (DIC), a measure of model fit. DIC was first proposed by Spiegelhalter et al. (2002) and
has since been widely used in Bayesian model selection problems. Formally, DIC is defined as:

\[
DIC = \bar{D} + p_D,
\]

where

\[
\bar{D} = E_{\theta|q} [D(\theta)],
\]

is the posterior expectation of the Bayesian deviance

\[
D(\theta) = -2 \log p(q|\theta),
\]

and \( p_D \), the “effective number of parameters”, is:

\[
p_D = \bar{D} - D(\bar{\theta}),
\]

where \( \theta \) is the vector of parameters intervening in the sampling model and \( \bar{\theta} \) is the posterior mean of \( \theta \). The idea is that models with smaller DIC should be preferred to models with larger DIC. Models are penalized both by the value of \( \bar{D} \), which favors a good fit, but also (in common with AIC and BIC) by the effective number of parameters \( p_D \). Since \( \bar{D} \) will decrease as the number of parameters in a model increases, the \( p_D \) term compensates for this effect by favoring models with a smaller number of parameters. Interested readers are referred to Spiegelhalter et al. (2002) for more details.

The DIC values for the fixed coefficient SDF model and the random coefficient SDF model are presented in Tables 1 and 2, respectively. We can see that the DIC value for the random coefficient SDF model is significantly smaller than that for the fixed coefficient SDF model, suggesting strongly that the random coefficient SDF model is preferred to the fixed coefficient SDF model for modeling technology heterogeneity among the large banks in the U.S.

Having established the superiority of the random coefficient SDF model, it is worth examining the consequences of failure to allow for technology heterogeneity. We first examine the conse-
quences of failure to allow for technology heterogeneity on the magnitudes of returns to scale. For this purpose, we plot in Figure 1 the probability density functions of the estimates of returns to scale from the two SDF models for the first quarter of year 2010. As the figure indicates, the estimates from the random coefficient SDF model vary within a much wider range than those from the fixed coefficient SDF model\(^4\), suggesting that the use of the latter model severely restricts the range of possible estimated values for returns to scale. To further examine the consequences on the magnitudes of returns to scale, we also compute the difference or bias in returns to scale (in absolute value) between the fixed coefficient SDF model and the random coefficient SDF model for each bank in the first quarter of each year. The results are summarized in Table 3. As the table indicates, the bias can be substantial. Taking 2010.Q1 for example, the mean of the bias is 0.0735 and the maximum of the bias is 0.3752.

We then examine the consequences of failure to allow for technology heterogeneity on the ranking of banks in terms of returns to scale (RTS). To this end, we calculate the Spearman rank correlation coefficient (\(\rho\)) between the ranking based on the fixed coefficient SDF model and that based on the random coefficient SDF model

\[
\rho = 1 - \frac{6 \sum_{j=1}^{n_k} (\text{Rank}_{j1} - \text{Rank}_{j2})^2}{n_k(n_k^2 - 1)},
\]

where \(n_k\) is the number of banks in the sample, \(\text{Rank}_{j1}\) is the rank of bank \(j\) based on the random coefficient SDF model, and \(\text{Rank}_{j2}\) is the rank of the same bank based on the fixed coefficient SDF model. If \(\rho = -1\), there is perfect negative correlation; if \(\rho = 1\), there is perfect positive correlation; and if \(\rho = 0\), there is no correlation.

Table 4 presents the estimates of \(\rho\) as well as their associated credible intervals (in parentheses) obtained using the posterior distributions of \(\rho\). As can be seen from the table, none of the 95% credible intervals contain one, suggesting that all the reported Spearman rank correlation coefficients are significantly different from one at the 5% level. This is confirmed by the point estimates,

\(^4\)In fact, we can theoretically prove that the range of RTS estimates obtained from the random coefficient SDF model is always wider than that obtained from its fixed coefficients counterpart. A detailed proof is available on request.
displayed in the first column of Table 4. Specifically, the Spearman rank correlation coefficients range from 0.0462 to 0.2010, suggesting that there is little correlation between the ranking based on the fixed coefficient SDF model and that based on the random coefficient SDF model.

Hence failure to consider technology heterogeneity results in mismeasured returns to scale and a misleading ranking of banks. Thus in what follows we will focus on the results from the random coefficient SDF model.

5.3. Results from the random coefficient SDF model

We first compute the percentage of large banks facing increasing, constant, or decreasing returns to scale for each sample year. This computation is performed by counting the number of cases where the 95% credible intervals are strictly greater than 1.0 (indicating increasing returns to scale), strictly less than 1.0 (indicating decreasing returns to scale), or contain 1.0 (indicating constant returns to scale). The results are presented in Table 5. Two findings emerge from this table. First, the majority of the large banks face constant returns to scale, a small percentage face decreasing returns to scale, and an even smaller percentage face increasing returns to scale. Specifically, on average 55.63% of the large banks in the U.S. face constant returns to scale, 33.82% face decreasing returns to scale, and 10.55% face increasing returns to scale. Second, the percentage of large banks facing increasing or decreasing returns to scale shows a "first increase and then decrease" pattern, while those facing constant returns to scale shows a "first decrease and then increase" pattern. Specifically, the percentage of large banks facing increasing returns to scale increases markedly from 5.37% in 1997.Q1 to 15.82% in 2007.Q1, then declines noticeably to 13.56% in 2010.Q1; the percentage of large banks facing constant returns to scale declines considerably from 62.43% in 1997.Q1 to 51.41% in 2005.Q1 and then increases gradually to 59.60% in 2010.Q1; and the percentage of large banks facing decreasing returns to scale increased considerably from 32.20% in 1997.Q1 to 39.27% in 2003.Q1 and then declines to 26.84% in 2010.Q1.

To check the robustness of these results, we also estimate a random coefficient SDF model with the inefficiency term \(u_{it}\) following a half normal distribution. The results are presented in Table 6. A comparison of the results in Table 5 and those in Table 6 reveals that the major
conclusions reached in the previous paragraph remain qualitatively valid, although we notice that there are some quantitative changes. First, the majority (67.62%) of the large banks in the U.S. face constant returns to scale, while a small percentage (22.74%) face decreasing returns to scale and an even smaller percentage (9.64%) face increasing returns to scale. Second, the percentage of large banks facing increasing / constant / decreasing returns to scale experiences almost the same temporal pattern as in the exponential model. Taking for example large banks facing constant returns to scale, the percentage of these banks declines considerably from 76.84% in 1997.Q1 to 62.15% in 2003.Q1 and then increases gradually to 66.95% in 2010.Q1.

We next turn to the estimate of returns to scale for each individual bank. Figure 2 presents the point estimate of returns to scale for each bank in the first quarter of 2010. Due to space limitations, we do not present the estimates for the other sample years, which are qualitatively similar to those shown in Figure 2. Two findings emerge from Figure 2. The first finding is that banks of the same or similar size can face different levels of returns to scale. To give an example, there are 6 banks whose total assets fall within the narrow range between $1.302 and $1.339 billion. Despite their similar sizes, these 6 banks face returns to scale that vary greatly from 0.79 to 1.21. This finding is not surprising because our estimates of returns to scale not only reflect differences in outputs and inputs across the banks but also differences in production technology across the banks.

The second finding emerging from Figure 2 is that in the presence of technology heterogeneity, there is no clear pattern among the large banks concerning the relationship between asset size and returns to scale. This lack of a relationship can be observed by visually examining the patterns of the scatter plots in Figure 2. To see this lack of a pattern more formally, we regress estimated returns to scale on a constant and total assets for the first quarter of each year, i.e., $RTS = \alpha_0 + \alpha_1 ASSET$, where $ASSET$ denotes total assets, $\alpha_0$ the constant, and $\alpha_1$ the coefficient for $ASSET$. For each of the sample period, we cannot reject that $\alpha_1 = 0$ at the 5% level of significance, indicating that there is no linear relationship between asset size and returns to scale for the large banks in each sample year. To further examine if there is a nonlinear relationship between asset size and returns to scale for the large banks, we regress estimated returns
to scale on a constant, total assets, and squared total assets for the first quarter of each year, i.e.,

\[ RTS = \gamma_0 + \gamma_1 \text{ASSET} + \gamma_2 \text{ASSET}^2, \]

where \( \gamma_0 \) is the constant, \( \gamma_1 \) the coefficient for \( \text{ASSET} \), and \( \gamma_2 \) the coefficient for \( \text{ASSET}^2 \). Again for each of the sample periods, we cannot reject that \( \gamma_1 = \gamma_2 = 0 \) at the 5% level of significance, confirming that there is no clear pattern among the large banks concerning the relationship between asset size and returns to scale.

This finding differs from the commonly found pattern concerning the relationship between asset size and returns to scale, i.e. banks face increasing returns to scale up to an optimum size, constant returns to scale at that point, and decreasing returns to scale above that point. This discrepancy might be explained by the presence of technology heterogeneity. When technology is heterogeneous, it is possible to have two banks of different sizes, where the smaller one shows decreasing returns to scale and the larger one shows increasing returns to scale, as long as the former bank is operating at a scale greater than its own optimum size, and the latter bank is operating at a scale smaller than its own optimum size. When this phenomenon is observed, the relationship between asset size and returns to scale is no longer independent of the technologies employed, and the commonly found pattern concerning relationship between asset size and returns to scale does not necessarily hold across banks employing different technologies.

The second finding is supported by the results of Daniel et al. (1973), which, to the best of our knowledge, is the only previous study that explicitly takes technology heterogeneity into account. Specifically, these researchers classify a sample of 967 banks in the U.S. into three technology groups — banks that did not use computers, banks that had used a computer less than one year, and banks that had used a computer more than one year — according to the technology the banks employ. By estimating a fixed coefficient Cobb-Douglas cost function separately for each of the three groups for the year 1968, Daniel et al. (1973) find that smaller banks without computers experience decreasing returns scale, while larger banks with computers experience increasing returns scale, suggesting that in the presence of technology heterogeneity, the commonly found pattern concerning relationship between asset size and returns to scale may not hold.

The second finding is also supported by results of studies on other industries. For example,
Murray and White (1983) examine returns to scale at Canadian credit unions. In this study, these authors classify a sample of 152 credit unions in Canada into three major technology groups — manual, tronics (credit unions adopting electronic bookkeeping machines), and computers (credit unions adopting more advanced computer systems) — according to the posting technology the credit unions employ. After estimating a fixed coefficient translog cost function separately for each of the three groups, Murray and White (1983) find that credit unions in the manual group exhibit decreasing returns to scale and those in the tronics group exhibit increasing returns to scale, although the credit unions in the manual group on average have much smaller assets than those in the tronics group. Similarly, they also find that credit unions in the tronics group exhibit smaller degrees of increasing returns to scale than those in the computer group, although the credit unions in the tronics group on average have much smaller assets than those in the computer group (see Table 4 of Murray and White, 1983). The results of Murray and White (1983) further confirm that in the presence of technological differences, the commonly found pattern concerning the relationship between asset size and returns to scale may disappear.

Having examined returns to scale for the large banks, we now turn to the estimates of two related measures — technical change and productivity growth. Within an output distance function framework, the Divisia productivity index \( \frac{d \ln TFP}{dt} \) can be calculated using the following equation (Diewert and Fox, 2010)

\[
\frac{d \ln TFP}{dt} = \frac{\partial \ln D_o(y, x, t)}{\partial \ln y_m} - \frac{\partial \ln D_o(y, x, t)}{\partial \ln x_n}.
\]

This output-distance-function-based Divisia productivity index can be further decomposed into two components: technical change and efficiency change (Feng and Zhang, 2012). Formally,

\[
\frac{d \ln TFP}{dt} = TC + \Delta TE,
\]
where

\[
TC = -\frac{\partial \ln D_o(y, x, t)}{\partial t},
\]

\[
\Delta TE = -d \ln \psi(t)/dt,
\]

where \(\psi(t)\) is the deviation of the output distance function from one due to technical inefficiency, i.e., \(D_o(y, x, t) \psi(t) = 1\). The first term \((TC)\) on the right side of (31) is an output-distance-function-based measure of technical change and the second term \((\Delta TE)\) is an output-distance-function-based measure of efficiency change. This decomposition resembles the famous Färe et al. (1994) decomposition, where the Caves et al. (1982) Malmquist productivity index is decomposed into the same two components. The difference is that the Färe et al. (1994) decomposition is performed within a discrete time Malmquist index framework, whereas the one in (31) is performed within a continuous Divisia index framework.

Table 7 reports the results on productivity growth, its two components (technical change and efficiency change), and the contribution of each of the two components to productivity growth. The results, displayed in the first column of Table 7, indicate that productivity grows in all years, at an average annual rate of 0.48%. Moreover, the estimates for productivity growth are rather stable over the sample period, ranging between 0.33% in 2000 to 0.73% in 1998. The decomposition of productivity growth in Table 7 identifies the forces behind the positive productivity growth. The results, displayed in the second column of Table 7, show that average annual technical change for the large banks is fairly stable and ranges between 0.43% and 0.53%, suggesting that on average, technical change has consistently contributed to the productivity growth at the large banks over the sample period. The results, displayed in the third column of Table 7, shows that average annual efficiency change is relatively small in magnitude compared with the average annual technical change. Further, it fluctuates around zero, suggesting that average annual efficiency change cannot be a major driver behind the productivity growth at the large banks. The dominance of technical change over efficiency change can also be seen from its percentage contribution: on average it
contributes 100% each year to productivity growth, while efficiency change contributes 0%.

6. Conclusion

Over the last two decades, the increasing concentration of industry assets among large banks prompted a substantial number of studies examining returns to scale at large banks in the U.S. However, few articles explicitly allow production technology to be heterogeneous, even though studies have found that unobserved technology heterogeneity is widely present among large banks in the U.S., due to the slow nature of diffusion process of new technologies, differences in organizational structure, and differences in business model. The widespread presence of unobserved technology heterogeneity thus calls for a model that is suitable for modeling returns to scale in such a situation. The standard fixed coefficient stochastic frontier models are unsuitable for modeling returns to scale in the presence of technology heterogeneity, because these models assume that all banks under study share a common vector of coefficients and thus a common production technology (Greene, 2008).

For the first time in this literature, we estimate a random coefficient translog stochastic distance frontier (SDF) model within a Bayesian framework. The main feature of this model is that its coefficients can vary across banks, thereby allowing production technology to be heterogeneous across banks. This coefficient variation feature enables one to measure returns to scale for each bank on the bank’s own frontier in the presence of technology heterogeneity. This implies that estimates of returns to scale based on the random coefficient SDF model not only reflect differences in input and output quantities across banks but also differences in technology across banks, thereby making this model a valuable tool for studying returns to scale in the presence of technology heterogeneity.

Our empirical results confirm that the random coefficient SDF model outperforms the fixed coefficient SDF model by a large margin, thus providing strong evidence for the presence of technology heterogeneity among large banks in the U.S. Our analysis further demonstrates that failure to allow for technology heterogeneity results in misreading results regarding returns to scale. Our re-
Results from the random coefficient SDF model indicate that the majority (55.63%) of the large banks in the U.S. face constant returns to scale, a small percentage (33.82%) face decreasing returns to scale, and an even smaller percentage (10.55%) face increasing returns to scale. In addition, our results suggest that banks of the same size may have different levels of returns to scale and that there is no clear pattern regarding the relationship between asset size and returns to scale for the large banks in the U.S.
References


