Journal of Banking & Finance 39 (2014) 135–145

Contents lists available at ScienceDirect

Journal of Banking & Finance

journal homepage: www.elsevier.com/locate/jbf

Returns to scale at large banks in the US: A random coefficient stochastic frontier approach ☆

Guohua Feng [a,*], Xiaohui Zhang [b]

[a] Department of Economics, The University of North Texas, Denton, TX 76201, USA
[b] School of Management and Governance, Murdoch University, Western Australia 6150, Australia

ARTICLE INFO

Article history:
Received 22 June 2013
Accepted 27 October 2013
Available online 21 November 2013

JEL classification:
C11
D24
G21

Keywords:
Returns to scale
Random coefficient stochastic distance frontier model
Bayesian estimation

ABSTRACT

This paper investigates the returns to scale of large banks in the US 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 US 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 US 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 US concerning the relationship between asset size and returns to scale, due to the presence of technology heterogeneity.

1. Introduction

Over the past two decades, the increasing dominance of large banks in the US 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 US. 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 US. For excellent reviews, see Berger et al. (1993, 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 US 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 US, 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 US 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"[1] information-based production technologies (such as credit scoring technologies), whereas decen-

☆ We would like to thank the referee for comments that greatly improved the paper.

* Corresponding author. Tel.: +1 (940) 565 2220; fax: +1 (940) 565 4426.
E-mail address: guohua.feng@unt.edu (G. Feng).

[1] Hard information refers to information that is quantitative and can be credibly transmitted across physical or organizational distances, whereas soft information refers to the opposite.

0378-4266/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jbankfin.2013.10.012

tralized 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 US, 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 $\boldsymbol{\beta}_i$, where $i = 1, 2, \ldots, K$ indexes banks and K is the number of banks) into two parts: a mean vector (denoted by $\bar{\boldsymbol{\beta}}$) and a random vector (denoted by $\boldsymbol{\delta}_i$) 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).

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. $\bar{\boldsymbol{\beta}}$), but can also approximate each of the underlying true heterogeneous technologies to the second order via its firm-specific coefficients (i.e. $\boldsymbol{\beta}_i$). Specifically, let $\boldsymbol{y}$ denote the $1 \times M$ vector of outputs, $\boldsymbol{x}$ denote the $1 \times N$ vector of inputs, and $\boldsymbol{w} \equiv (\boldsymbol{y}, \boldsymbol{x})$ denote the output and input vector. In addition, let $f^i(\boldsymbol{w})$ (for $i = 1, 2, \ldots, K$), a set of arbitrary output distance functions, represent the underlying 'true' heterogeneous technologies; $\ln D_o^i(\boldsymbol{w})$ (for $i = 1, 2, \ldots, K$) denote the random coefficient translog output distance function; and $\ln D_o(\boldsymbol{w})$ 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, $\ln D_o^i(\boldsymbol{w})$, can approximate its corresponding arbitrary output distance function, $\ln f^i(\boldsymbol{w})$, to the second order at a point $\boldsymbol{w}^*$, by solving the system of equations: $\ln D_o^i(\boldsymbol{w}^*) = \ln f^i(\boldsymbol{w}^*), \partial \ln D_o^i(\boldsymbol{w}^*)/\partial \ln w_m = \partial \ln f^i(\boldsymbol{w}^*)/\partial \ln w_m$ (for $m = 1, 2, \ldots, M-1, M+1, \ldots, M-1+N$), and $\partial^2 \ln D_o^i(\boldsymbol{w}^*)/\partial \ln w_m \partial \ln w_n = \partial^2 \ln f^i(\boldsymbol{w}^*)/\partial \ln w_m \partial \ln w_n$ (for $1 < m < n < M+N$).[2] With $\boldsymbol{\beta}_i$ 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. $\bar{\boldsymbol{\beta}}$) can approximate the true 'average' technology (i.e. $\left[\sum_{i=1}^{K} \ln f^i(\boldsymbol{w}^*)\right]/K$) to the second order.[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 $\boldsymbol{\beta}_i$ to zero (i.e. $\boldsymbol{\delta}_i = 0$). In other words, the fixed coefficient translog output distance function is essentially the translog output distance function with the mean coefficient vector (i.e. $\bar{\boldsymbol{\beta}}$), suggesting that the fixed coefficient translog output distance function can also approximate the true 'average' technology (i.e. $\left[\sum_{i=1}^{K} \ln f^i(\boldsymbol{w}^*)\right]/K$) to the second order. However, unlike its random coefficient counterpart, the fixed coefficient translog output distance function is incapable of approximating $\ln f^i(\boldsymbol{w}^*)$ 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. $\ln D_o^i(\boldsymbol{w})$), 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 US 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 US with assets in excess of $1 billion. Our results show that the

[2] A detailed proof is available on request.

[3] A detailed proof is available on request.