UNIVERSITY OF CALGARY

Three Essays in Productivity and Efficiency

by

Guohua Feng

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECONOMICS

CALGARY, ALBERTA

MAY, 2008

UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "Three Essays in Productivity and Efficiency" submitted by Guohua Feng in partial fulfillment of the requirements for the degree of Doctoral of Philosophy

Supervisor
Dr. Apostolos Serletis
Department of Economics

Dr. Francisco M. Gonzalez
Department of Economics

Dr. Daniel Gordon
Department of Economics

Dr. Lasheng Yuan
Department of Economics

Dr. Alexender David
School of Business

External Examiner
Dr. W. Erwin Diewert
Department of Economics
The University of British Columbia

May 1, 2008
Date

Abstract

This dissertation consists of three essays on productivity and efficiency of two U.S. major industries — manufacturing and banking industries. An abstract for each of the three essays follows.

Essay 1 (Chapter 2) takes the econometric approach to productivity measurement in United States manufacturing, using KLEM data over the period from 1953 to 2001. I am also interested in technical change bias, price elasticties, and elasticties of substitution in the U.S. manufacturing industry. I present the empirical comparison and evaluation of the effectiveness of three well-known locally flexible cost functions and the globally flexible Asymptotically Ideal Model, the latter modified to introduce technical change by means of factor-augmenting efficiency index approach. I show that the extended AIM model performs much better than the three locally flexible functional forms.

Essay 2 (Chapter 3) provides estimates of bank efficiency and productivity in the United States, over the period from 1998 to 2005, using (for the first time) the globally flexible Fourier cost frontier and estimated subject to full theoretical regularity conditions. I find that failure to incorporate monotonicity and curvature into the estimation results in mismeasured magnitudes of cost efficiency and misleading rankings of individual banks in terms of cost efficiency. I also find that the largest four bank subgroups (with assets greater than 400 million)experienced significant productivity gains and the smallest eight subgroups experienced insignificant productivity gains or even productivity losses.

Essay 3 (Chapte 4) 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, I 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 a clear downward trend in the growth rate of total factor productivity and my 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.

Acknowledgements

The deepest appreciation goes to my supervisor, Professor Apostolos Serletis. Throughout my Ph.D. program, he has constantly guided and encouraged me to challenge the toughest issues in the field of productivity and efficiency. Without his tireless and insightful supervision, the completion of this dissertation would not have been possible. In addition to his academic guidance, his thoughtful consideration and kindness in the way he lives his daily life served as an exceptional lesson to me. I truly believe that I am amazingly fortunate to have him as my supervisor.

My thanks also go to my dissertation committee members, Dr. Francisco M. Gonzalez, Dr. Daniel Gordon, and Dr. Lasheng Yuan, for their invaluable help and guidance. I would also like to thank Dr. Alexander David for having served on my committee and for his valuable advice. I am also very grateful to Dr. Zuzana Janko, Dr. John Boyce and Dr. Jean-Francois Wen for their kind help during my Ph.D. study. Special thanks go to Dr. W. Erwin Diewert, from whom I started to learn flexible functional forms and their applications to productivity issues. This, in fact, turned out to be a very important contribution to my dissertation.

Financial supports from Professor Serletis and the Department of Economics at the University of Calgary are greatly appreciated.

I would also like to express my heart-felt gratitude to my family. I deeply thank my father, sister and brother for their unconditional love and support on all the decisions I have made. Most importantly, I am extremely grateful to my wife, Juan, for her love, support and patience and for helping me keep my life in proper perspective and balance.

This dissertation is dedicated to the memory of my forever beloved mother.

Table of Contents

List of Tables

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

List of Figures

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

CHAPTER ONE

INTRODUCTION

1.1 Background

Productivity measurement and analysis has become widespread since Solow's (1957) decomposition of output growth into the contribution of input growth and a residual-based productivity term. The main thrust of the Solow article is to present a theoretical and empirical framework for distinguishing between shifts in the production structure and movements along the production structure. The resulting "residual" measure has been interpreted as a shift in the production structure and typically has been labeled technological change, or total productivity growth. Most of the literature follows the innovative works by Jorgenson and Griliches (1967), Diewert (1976), Berndt and Khaled (1979), Diewert and Wales (1987), and Fare *et al.* (1994) in investigating total factor productivity. This literature is interesting not only because of the critical importance of productivity growth to living standards in actual economies, but also because of the related interesting topics analyzed together with productivity growth, such as factor substitutability, convergence of macroeconomic structure and performance, and economic growth.

There are many different approaches to total factor productivity measurement and analysis. Among all the approaches, the econometric approach, which involves estimating the parameters of an aggregator function (cost, profit, production, or distance function), has arguably received the most attention. Compared with other approaches, it has the advantage of allowing for the careful testing of various structural and behavioral assumptions of a postulated model, i.e. non-competitive pricing behavior, non-constant returns, factor-augmenting technical change as well as embellishments like cost-of-adjustment parameters, rather than to simply impose those features a priori. Prominent works along this line include, but not limited to, Berndt and Khaled (1979), Denny, Fuss, and Waverman (1981) and Diewert and Wales (1987). These works typically treat producers as successful optimizers, and thus deviations from maximum outputs, from minimum cost

and from maximum profit are attributed exclusively to statistical noise.

The introduction of efficiency change as a source of productivity change into the traditional econometric approach has lead to the development of the more recent literature of stochastic frontier approach (SFA) to productivity analysis.[1] Recent studies suggest that not all producers are always so successful in solving their optimization problems. For example, not all of them succeed in utilizing the minimum inputs required to produce the outputs they choose to produce, given the existing technology (technical inefficiency), or succeed in allocating their inputs in a cost effective manner or allocate their outputs in a revenue maximizing manner, given the input and output prices they face (allocative inefficiency). The presence of technical and allocative inefficiency thus enables one to decompose the combined productivity changes into efficiency movements (efficiency change) and frontier shift components (technological change), among other components — see Kumbhakar and Lovell (2003) for an excellent review. Technically, the only difference between the traditional econometric approach and the stochastic frontier approach is that a two-component composite error term is usually assumed in the estimation of the stochastic frontier model with one capturing firm inefficiency and the other capturing statistical noise. In this sense, the stochastic frontier approach can be seen as a modified econometric approach.

As a modified econometric approach, the stochastic frontier approach has enjoyed great popularity in the analysis of productivity in the last fifteen years. It has been widely applied to the analysis of productivity in various industries using firm-level data. For example, Bauer (1990) applied it to the analysis of productivity in the U.S. airline industry; DeYoung and Hasan (1998) and Berger and Mester (2003), among many others, applied it to the analysis of efficiency and productivity in the U.S banking industry;

[1]The introduction of efficiency change as a source of productivity change was pioneered by Nishimizu and Page (1982), who used a deterministic translog production frontier. It is Bauer (1990) who was the first to employ the stochastic frontier approach to decompose productivity change.

and Atkinson *et al.* (2003) applied it the analysis of productivity in the U.S. electric utilities. It has also been applied to the analysis of the effect on firm productivity of exogenous variables characterizing the environment in which production occurs — i.e. public infrastructure, the degree of competitive pressure, network characteristics, ownership form, managerial ability — see for example, Simar *et al.* (1994), Battese and Coelli (1999), Huang and Liu (1994) and Mester (1997). In addition, the stochastic frontier approach has also been used in the analysis of growth convergence — see, for example, Kumbhakar and Wang (2005).

Despite its popularity, the econometric approach — including both the traditional approach and the more recent stochastic frontier approach — suffers from the following two problems. First, the functional form employed in the econometric approach suffers the problem of not having enough flexibility. Most of the previous studies in this literature employ a locally flexible functional forms, i.e. the generalized Leontief [see Diewert (1971)], translog [see Christensen *et al.* (1975)] and normalized quadratic [see Diewert and Wales (1987)] functional forms, which theoretically can attain flexibility only at a single point or in an infinitesimally small region. In the field of firm efficiency, researchers have found, however, that the translog function lacks enough flexibility in modelling industries which are composed of firms of widely varying sizes; see McAllister and McManus (1983) and Wheelock and Wilson (2001). Diewert and Lawrence (2002) also find that the NQ functional form, the only functional form that possesses the property that correct curvature conditions can be imposed globally without destroying the flexibility of the functional form, does not have enough flexibility in modelling the variations in price elasticities over time.

There are, however, two globally flexible functional forms which can provide greater flexibility than locally flexible functional forms: the Fourier flexible functional form [Gallant (1982)] and the Asymptotically Ideal Model (AIM) [Barnett *et al.* (1991)]. The

former is based on a Fourier series expansion and the latter is based on a linearly homogeneous multivariate Muntz-Szatz series expansion. Both of them are globally flexible in the sense that they are capable of approximating the underlying cost function at every point in the function's domain by increasing the order of the expansion, and thus have more flexibility than most of the locally flexible functional forms. Despite its greater flexibility, the AIM model has never been used to model productivity due to the computational complexity involved. Regarding the Fourier functional form, although some previous studies attempted to use it to model productivity and efficiency, all of them ignore the parametric relationship between the 'reparameterized' translog function and the trigonometric Fourier series of the Fourier function, and thus potentially reach perverse conclusions regarding productivity and efficiency.

The second problem with the econometric approach is that the estimated parameters of the dual functions (i.e. cost, profit functions) frequently violate the monotonicity and concavity constraints implied by economic theory. While permitting a parameterized function to depart from the neoclassical function space is usually fit-improving, it may also cause the failure of duality theory on which dual econometric models are based. As Barnett (2002, p. 199) put it, without satisfaction of all three theoretical regularity conditions (positivity, monotonicity and curvature) "the second-order conditions for optimizing behavior fail, and duality theory fails. The resulting first-order conditions, demand functions, and supply functions become invalid." With the econometric approach, estimates of productivity and efficiency are essentially functions of the estimates of parameters. Therefore, the inaccuracy or even wrongness of the estimates of the parameters from unconstrained models will eventually show up in the estimates of productivity and efficiency.

1.2 Objective and Significance of the Research

Motivated by the widespread practice of ignoring the theoretical regularity conditions and not using a globally flexible functional form, the purpose of this thesis is three-fold. First to show the superiority of the globally flexible Asymptotically Ideal model (AIM) over locally flexible functional forms in modelling productivity. This is done in the second chapter. In doing so, I extend the globally flexible AIM by incorporating technological change into it without destroying any of the neoclassical theoretical regularity conditions. By using the KLEM data for the U.S. manufacturing industry, I present an empirical comparison and evaluation of the effectiveness of four well-known flexible cost functions — the locally flexible generalized Leontief, translog, and normalized quadratic — and the globally flexible Asymptotically Ideal Model, in terms of their ability to capture the variation in total factor productivity and price elasticities. While the superiority of the globally flexible AIM over locally flexible functional forms is shown within the context of productivity measurement in this thesis, it can be easily generalized to other studies which involve technology or preference modeling.

The second purpose of this thesis is to show the importance of the incorporation of monotonicity and curvature in the analysis of firm efficiency using the stochastic frontier approach. This is done in the third chapter. More specifically, within the framework of cost frontier, I show, by comparing results from unconstrained and regularity-constrained models, that the violation of monotonicity and curvature constraints may lead to mis-measured magnitudes of cost efficiency and misleading rankings of individual banks in terms of firm efficiency. Intuitively, the violation of curvature at a data point (p_{jt}, y_{jt}) implies that the quantities of some outputs increase as their corresponding prices increase (holding other things constant); and the violation of monotonicity at that data point implies the quantities of some outputs decrease as total cost increases (holding other things

constant). Both of these two cases mean that the best practice firm is not minimizing its cost at (p_{jt}, y_{jt}). Therefore, cost efficiency, which is supposed to be measured relative to a cost-minimizing best practice bank, is not accurate when monotonicity and curvature are violated.

The third purpose is to show the importance of the incorporation of monotonicity and curvature in the construction of distance function based productivity indexes. The construction of distance function based productivity indexes by exploiting the duality between the output (input) distance function and the revenue (cost or profit) function has become an active research area — see, for example, Orea (2002). However, none of the previous studies has treated monotonicity and curvature conditions as maintained hypotheses. In the fourth chapter, I first derive a productivity index by exploiting the duality between the output distance function and the profit function. Within a Bayesian framework, I then demonstrate that violations of monotonicity and curvature constraints may lead to perverse conclusion regarding productivity. For example, when monotonicity is violated, we may reach a wrong conclusion, i.e. an increase in labor, holding other things fixed, may result in an increase in productivity growth.

1.3 Organization of This Thesis

The rest of the thesis is organized as follows. In Chapter 2, I take the econometric approach to productivity measurement in United States manufacturing, using KLEM data over the period from 1953 to 2001, and present an empirical comparison and evaluation of the effectiveness of four well-known flexible cost functions — the locally flexible generalized Leontief, translog, and normalized quadratic — and the globally flexible AIM model in modelling productivity. Chapter 3 provides estimates of bank efficiency and productivity in the United States, over the period from 1998 to 2005, using (for the

first time) the globally flexible Fourier cost functional form, as originally proposed by Gallant (1982), and estimated subject to global theoretical regularity conditions, using procedures suggested by Gallant and Golub (1984). The fourth chapter 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, I 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. Chapter 5 concludes the thesis.

CHAPTER TWO

PRODUCTIVITY TRENDS IN U.S. MANUFACTURING: EVIDENCE FROM THE NQ AND AIM COST FUNCTIONS

2.1 Introduction

The analysis and measurement of productivity performance has attracted a great deal of attention ever since Solow (1957) decomposed the growth in output into the growth of inputs and a residual-based productivity term. Most of the literature follows the innovative works by Jorgenson and Griliches (1967), Diewert (1976), Berndt and Khaled (1979), Diewert and Wales (1987), and Fare *et al.* (1994) in investigating total factor productivity. This literature is interesting not only because of the critical importance of productivity growth to living standards in actual economies, but also because of the related interesting topics analyzed together with productivity growth, such as factor substitutability, convergence of macroeconomic structure and performance, and economic growth.

There are four different approaches to total factor productivity (TFP) measurement — growth accounting, the index number approach, the distance function approach, and the econometric approach. In this paper, I briefly review each of these methods, and then take the econometric approach to productivity measurement in the United States. In doing so, I use manufacturing KLEM (capital, labor, energy, and intermediate materials) data, over the period from 1953 to 2001. I am also interested in technical change bias, price elasticities, and elasticities of substitution in the U.S. manufacturing industry. I also present an empirical comparison and evaluation of the effectiveness of four well-known flexible cost functions — namely, the locally flexible generalized Leontief [see Diewert (1971)], translog [see Christensen *et al.* (1975)], and normalized quadratic [see Diewert and Wales (1987)], and one globally flexible cost function, the Asymptotically Ideal Model [see Barnett *et al.* (1991)]. In this literature, there is no *a priori* view as to which flexible functional forms are appropriate, once they satisfy the regularity conditions of neoclassical microeconomic theory — positivity, monotonicity, and curvature.

I pay explicit attention to all three theoretical regularity conditions and argue that much of the older literature on total factor productivity measurement ignores economic regularity. I argue that unless economic regularity is attained by luck, flexible functional forms should always be estimated subject to regularity, as suggested by Barnett (2002) and Barnett and Pasupathy (2003). In fact, I follow Ryan and Wales (1998), Moschini (1999), Gallant and Golub (1984), and Serletis and Shahmoradi (2005, 2007) and treat the curvature property as a maintained hypothesis and build it into the models being estimated. I also address econometric regularity issues and highlight the challenge inherent with achieving both economic and econometric regularity.

I also extend the AIM model and introduce technical change in the AIM cost function by means of Thomsen's (2000) factor-augmenting efficiency index approach. The main advantage of this approach, unlike the generic time trend models of technical change, is that one can measure input specific productivity, changes in input productivity, as well as the contribution of each input to overall productivity. My empirical results show that the AIM cost function with technical change introduced through the factor-augmenting efficiency index approach performs better than traditional locally flexible function forms and gives more accurate estimates of total factor productivity.

The rest of the paper is organized as follows. Section 2 provides a brief review of the different approaches to total factor productivity measurement — growth accounting, the index number approach, the distance function approach, and the econometric approach. In Section 3 I follow the econometric approach and discuss in detail the four cost functions that I use as well as the relevant procedures for imposing concavity on each of these functions. In Section 4 I deal with data and econometric issues while in Section 5 I estimate the models, report on theoretical regularity violations, and report estimates of total factor productivity based on the best-performing model(s). The final section concludes the paper.

2.2 Productivity Measurement

As already noted, there are four different approaches to total factor productivity measurement — growth accounting, the index number approach, the distance function approach, and the econometric approach. In what follows, I briefly discuss each of these approaches.

2.2.1 Growth Accounting

Growth accounting was suggested by Solow (1957) as a method of estimating the growth of total factor productivity. Growth accounting calculation of total factor productivity requires the specification of a neoclassical production function. Consider, for example, the Cobb-Douglas production function,

$$Y = AK^{\alpha}L^{1-\alpha} \qquad \alpha \in (0,1),$$

where Y is (real) output, K is capital, L is labor, and α is the share of capital in output. A is a measure of the current level of technology, more commonly referred to as multi-factor growth productivity or total factor productivity (TFP) — if, for example, A increases by 1% and if the inputs (K and L) are unchanged, then output increases by 1%.

As noted by Carlaw and Lipsey (2003), total factor productivity can be calculated either as a geometric index in levels,

$$TFP = \frac{Y}{K^{\alpha}L^{1-\alpha}} = A, \tag{2.1}$$

or as an arithmetic index in rates of change,

$$\frac{\Delta A}{A} = \frac{\Delta Y}{Y} - \alpha\frac{\Delta K}{K} + (1-\alpha)\frac{\Delta L}{L} = \Delta TFP. \tag{2.2}$$

Equation (2.2) is the key equation in growth accounting. It defines the growth of total

factor productivity, $\Delta A/A$, as the growth in output that cannot be accounted for by growth in capital and labor. $\Delta A/A$ is called the Solow residual, after Robert Solow who suggested this method of estimating the growth of total factor productivity. It is also known as the rate of technical progress.

2.2.2 The Index Number Approach

The index number approach is an extension of (and complement to) growth accounting. It involves dividing a (real) output quantity index, Y, by an input quantity index, I, to obtain a measure of total factor productivity, A, as follows

$$A = \frac{Y}{I}.$$

The index number approach is widely used by the majority of statistical agencies that regularly produce productivity statistics. However, one critical issue regarding this approach is the selection of the appropriate indexes. In fact, statistical indexes are mainly characterized by their statistical properties. These properties were examined in great detail by Fisher (1922) and serve as tests in assessing the quality of a particular statistical index. They have been named, after Fisher, as 'Fisher's system of tests' — see Eichhorn (1976) for a detailed analysis as well as a comprehensive bibliography of Fisher's 'test' or 'axiomatic' approach to index numbers.

The index that Fisher (1922) found to be the best, in the sense of possessing the largest number of desirable statistical properties, has now become known as the 'Fisher ideal' index. Another index found to possess a very large number of such properties is the discrete time approximation to the continuous Divisia index, usually called the Törnqvist index or just the Divisia index (in discrete time). In fact, the primary advantage of the Fisher ideal index over the Divisia index is that the Fisher ideal index satisfies Fisher's

'factor reversal test' — which requires that the product of the price and quantity indexes for an aggregated good should equal actual expenditures on the component goods — while the discrete time approximation of the Divisia index fails that test. However, the magnitude of the error is very small — third order in the changes.

The index number approach does not require an aggregate production function, although the economic approach to statistical index numbers, pioneered by Diewert (1976), could be used for selecting the appropriate index — see also Diewert and Lawrence (1999) and Diewert and Nakamura (2003) for a detailed discussion. In particular, Diewert (1976) provided the link between aggregation theory and statistical index number theory by attaching economic properties to statistical indexes. These properties are defined in terms of the statistical indexes' ability to approximate a particular functional form for the unknown underlying aggregator function. For example, Diewert (1976) showed that the Divisia index is 'exact' for the linearly homogeneous translog and is, therefore, 'superlative' (since the translog is a flexible functional form).

2.2.3 The Distance Function Approach

The distance function approach to measuring total factor productivity seeks to separate total factor productivity in two components: changes resulting from a movement towards the production frontier (technical efficiency) and shifts in the frontier (technical change). The distance function was first introduced separately by Shephard (1953) in the context of production analysis and by Malmquist (1953) in the context of consumption analysis. But it was introduced as a theoretical productivity index by Caves *et al.* (1982), and then popularized as an empirical productivity index by Färe *et al.* (1994).[1]

[1]There is also a closely related literature on firm efficiency, using stochastic production (or cost/profit) frontiers. Like the Malmquist productivity index discussed below, the parametric stochastic frontier approach does not assume that firms are operating at their efficient level, and thus enables one to decompose the combined productivity changes into efficiency movements (efficiency change) and frontier shift components (technological change), among other components. A two-component composite error term is usually assumed in the estimation of the parametric model with one capturing firm inefficiency

Using technology in period t as the reference technology (which exhibits constant returns to scale), the output-based Malmquist productivity index is written as

$$m_o^t\left(\boldsymbol{y}_t, \boldsymbol{y}_{t+1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}\right) = \frac{d_o^t\left(\boldsymbol{y}_{t+1}, \boldsymbol{x}_{t+1}\right)}{d_o^t\left(\boldsymbol{y}_t, \boldsymbol{x}_t\right)},$$

where $d_o^t\left(\boldsymbol{y}_t, \boldsymbol{x}_t\right)$ is the output distance function; that is, the reciprocal of the maximum proportional expansion of the output vector $\boldsymbol{y}_t$, given inputs $\boldsymbol{x}_t$. Alternatively, the Malmquist index can be defined in terms of technology in time $t+1$. Färe *et al.* (1994) extend this approach by defining the Malmquist total factor productivity index as the geometric mean of these two indexes

$$m_o^t\left(\boldsymbol{y}_t, \boldsymbol{y}_{t+1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}\right) = \left[\frac{d_o^t\left(\boldsymbol{y}_{t+1}, \boldsymbol{x}_{t+1}\right)}{d_o^t\left(\boldsymbol{y}_t, \boldsymbol{x}_t\right)} \times \frac{d_o^{t+1}\left(\boldsymbol{y}_{t+1}, \boldsymbol{x}_{t+1}\right)}{d_o^{t+1}\left(\boldsymbol{y}_t, \boldsymbol{x}_t\right)}\right]^{1/2},$$

which can be equivalently written as

$$m_o^t\left(\boldsymbol{y}_t, \boldsymbol{y}_{t+1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}\right) = \frac{d_o^{t+1}\left(\boldsymbol{y}_{t+1}, \boldsymbol{x}_{t+1}\right)}{d_o^t\left(\boldsymbol{y}_t, \boldsymbol{x}_t\right)} \times \left[\frac{d_o^t\left(\boldsymbol{y}_{t+1}, \boldsymbol{x}_{t+1}\right)}{d_o^{t+1}\left(\boldsymbol{y}_{t+1}, \boldsymbol{x}_{t+1}\right)} \times \frac{d_o^t\left(\boldsymbol{y}_t, \boldsymbol{x}_t\right)}{d_o^{t+1}\left(\boldsymbol{y}_t, \boldsymbol{x}_t\right)}\right]^{1/2},$$

where the term outside the brackets on the right-hand side measures the change in relative efficiency between years t and $t+1$ and the geometric mean of the two ratios inside the brackets measures the shift in technology between the two periods evaluated at $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$. It is to be noted that the Malmquist total factor productivity index can also be measured on the best practice technologies when variable returns to scale are taken into account.

The Färe *et al.* (1994) distance function based productivity index has several advantages. It does not require a specific functional form, it does not require information on

and the other capturing statistical noise. For an excellent review of this literature, see Kumbhakar and Lovell (2003).

prices, and it can be implemented in a multiple-output setting with many inputs (with no separability assumptions being required). Most importantly, it does not assume that firms are operating at their efficient level, and thus enables one to decompose the combined productivity changes into efficiency movements and frontier shift components. However, as Carlaw and Lipsey (2003, pp. 464) put it, "in order to implement this technique, one must know everything about the state of technology at every point in time and at every level of aggregation that TFP is calculated. Unfortunately, this is not possible given the data available." Moreover, implicit in the distance function approach to measuring total factor productivity is the assumption that all units (firms, industries, or countries) being compared have the same production function, when in fact evidence suggests that even firms within the same industry do not have identical production functions.

2.2.4 The Econometric Approach

The econometric approach to productivity measurement involves estimating the parameters of an aggregator function — cost, profit, or production function. Productivity growth can then be expressed in terms of the estimated parameters.

Technical change (or productivity growth) is usually defined in the primal setup (production function), as any shift in the production frontier. In particular, assuming a production function

$$y = f(\boldsymbol{x}, t), \tag{2.3}$$

where y is output, f is a continuous twice differentiable nondecreasing and quasiconcave function of a vector of inputs $\boldsymbol{x} \geq \mathbf{0}$, and t denotes a technology index, then technical change is defined as

$$\partial f(\boldsymbol{x}, t) / \partial t.$$

Technical change can also be defined in the dual setup (cost function), under certain

conditions. In particular, if firms competitively minimize the cost of production subject to producing a given amount of output, then the technology (2.3) is completely described by the dual cost function

$$C = C(\boldsymbol{p}, y, t) = yc(\boldsymbol{p}, t), \tag{2.4}$$

with the second equality assuming constant returns to scale. In equation (2.4), C is a nondecreasing, linearly homogeneous and concave function of prices, $\boldsymbol{p} > \mathbf{0}$, and c is the corresponding unit cost function — for an excellent review of duality theory, see Diewert (1982).

To obtain equations that are amenable to estimation, I apply Shephard's lemma to equation (2.4) to get

$$x_i = \frac{\partial C(\boldsymbol{p}, y, t)}{\partial p_i}, \qquad i = 1, \cdots, n, \tag{2.5}$$

or a more convenient equation for estimation purposes, by dividing through by y,

$$\frac{x_i}{y} = \frac{1}{y}\frac{\partial C(\boldsymbol{p}, y, t)}{\partial p_i}, \qquad i = 1, \cdots, n.$$

Using the envelope theorem,

$$\frac{\partial C(\boldsymbol{p}, y, t)}{\partial t} = -\frac{\partial C(\boldsymbol{p}, y, t)/\partial y}{\partial f(\boldsymbol{x}, t)/\partial t},$$

the rate of technical change can be measured from the cost function as follows

$$TFP = \frac{\partial \ln f(\boldsymbol{x}, t)}{\partial t} = \frac{1}{y}\frac{\partial f(\boldsymbol{x}, t)}{\partial t}$$

$$= -\frac{\partial C(\boldsymbol{p}, y, t)/\partial t}{y \partial C(\boldsymbol{p}, y, t)/\partial y} = -\frac{\partial \ln C(\boldsymbol{p}, y, t)/\partial t}{\partial \ln C(\boldsymbol{p}, y, t)/\partial \ln y}$$

$$= -\epsilon_{ct}\epsilon_{cy}^{-1}, \tag{2.6}$$

where $\epsilon_{ct} = \partial \ln C(\boldsymbol{p}, y, t)/\partial t$ and $\epsilon_{cy} = \partial \ln C(\boldsymbol{p}, y, t)/\partial \ln y$. According to equation (2.6), total factor productivity is the product of the dual rate of cost diminution (ϵ_{ct}) and the dual rate of returns to scale (ϵ_{cy}^{-1}). Hence, under constant returns to scale (where ϵ_{cy} is equal to unity), total factor productivity is the negative of the dual rate of cost diminution, meaning that a 1% upward shift in the production function is equal to a 1% decrease in the cost of production.

By taking the derivative of each estimated factor demand equation with respect to time and dividing by the estimated demands, I can also obtain a measure of the effect of technical change on each input (denoted by τ_i below) — see, for example, Diewert and Wales (1992) and Kohli (1994) — as follows,

$$\tau_i = \frac{\partial \ln x_i(\boldsymbol{p}, y, t)}{\partial t}. \tag{2.7}$$

If $\tau_i > 0$ (< 0), then technical change is input i augmenting (reducing), meaning that more (less) of the input is required due to the passing of time. In fact, total factor productivity is a weighted average of τ_i's. Following Kohli (1994), I define these τ_i's to

be technical change biases for each input. In particular, if

$$TFP = -\tau_i, \tag{2.8}$$

for input i, then technical change is said to be completely unbiased (neutral), in the sense that all goods are affected to the same degree. This corresponds to a 'homothetic shift' of the isoquants leaving the marginal rate of substitution between any two inputs (measured along a ray through the origin) unaffected by technical change. If, however, (2.8) does not hold, then technical change is said to be biased. This corresponds to a 'non-homothetic shift' of the isoquants, meaning that the marginal rate of substitution between any two inputs is affected by technical change.

Factor substitution is calculated, using both Allen and Morishima elasticities of substitution. The Allen-Uzawa elasticity of substitution between inputs i and j is given by

$$\sigma_{ij}^{a}(\boldsymbol{p}, y, t) = \frac{C(\boldsymbol{p}, y, t)\, C_{ij}(\boldsymbol{p}, y, t)}{C_i(\boldsymbol{p}, y, t)\, C_j(\boldsymbol{p}, y, t)}, \tag{2.9}$$

where the i, j subscripts refer to the first and second partial derivatives of $C(\boldsymbol{p}, y, t)$ with respect to input prices p_i and p_j. The Morishima elasticity of substitution between inputs i and j is given by

$$\sigma_{ij}^{m}(\boldsymbol{p}, y, t) = \frac{p_j C_{ij}(\boldsymbol{p}, y, t)}{C_i(\boldsymbol{p}, y, t)} - \frac{p_j C_{jj}(\boldsymbol{p}, y, t)}{C_j(\boldsymbol{p}, y, t)}. \tag{2.10}$$

If $\sigma_{ij}^{a} > 0$ (that is, if increasing the j^{th} price increases the optimal quantity of input i), I say that inputs i and j are Allen-Uzawa (net) substitutes. If $\sigma_{ij}^{a} < 0$, they are Allen-Uzawa (net) complements. Similarly, if $\sigma_{ij}^{m} > 0$ (that is, if increasing the j^{th} price increases the optimal quantity of input i relative to the optimal quantity of input j), I say that input j is a Morishima (net) substitute for input i. If $\sigma_{ij}^{m} < 0$, input

j is a Morishima net complement to input i. The Allen elasticities provide immediate qualitative comparative-static information about the effect of price changes on absolute input shares, whereas the Morishima elasticities immediately yield both qualitative and quantitative information about the effect of price changes on relative input shares.

The familiar price elasticities,

$$\eta_{ij} = \frac{\partial x_i(\boldsymbol{p}, y, t)}{\partial p_j} \frac{p_j}{x_i(\boldsymbol{p}, y, t)}, \tag{2.11}$$

could also be calculated as

$$\eta_{ij} = s_j \sigma_{ij}^a,$$

where s_j is the cost share of input j in total production costs. Notice that the price elasticities must satisfy the following condition

$$\sum_{j=1}^{n} \eta_{ij} = 0, \qquad i = 1, \cdots, n.$$

By substituting different unit cost functions into (2.4), I can get different total cost functions. Clearly, the econometric approach overcomes the problems of the index number approach and the distance function approach and has the flexibility to incorporate pertinent features of the market and industry structures as well as technological features that affect the productivity of firms or industries.

In what follows, I take the econometric approach to productivity measurement (in the United States) and provide a comparison between three widely used locally flexible cost functional forms — the generalized Leontief, basic translog, and normalized quadratic — and the globally flexible AIM cost function.

2.3 Flexible Cost Functional Forms

2.3.1 The Generalized Leontief Cost Function

By substituting the GL unit cost function [see Diewert (1971)] into (2.4), I get the GL specification

$$C(\boldsymbol{p}, y, t) = y\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}p_i^{1/2}p_j^{1/2} + \sum_{i=1}^{n}\beta_{it}p_i t\right), \tag{2.12}$$

where $\beta_{ij} = \beta_{ji}$. Using Shephard's lemma (2.5), and dividing through by y, yields optimal input-output demand equations, as follows

$$\frac{x_i}{y} = \sum_{j=1}^{n}\beta_{ij}p_j^{1/2}p_i^{-1/2} + \beta_{it}t, \qquad i = 1, \cdots, n. \tag{2.13}$$

Notice that all the parameters of the GL cost function (2.12) can be obtained by estimating only (2.13). It is to be noted that when $i = j$ in (2.13), $p_j^{1/2}p_i^{-1/2} = 1$ and so β_{ii} is a constant term in the ith input-output equation. When $\beta_{ij} = 0$ for all i, j, $i \neq j$, then input-output demand equations are independent of relative prices and the cross-price elasticities are zero.

Caves and Christensen (1980) have shown that the GL has satisfactory local properties when technology is nearly homothetic and substitution is low. However, when technology is not homothetic and substitution increases, they show that the GL has a rather small regularity region.

Concavity of the cost function (2.12) requires that the Hessian matrix is negative semidefinite. I can therefore impose local concavity (that is, at the reference point) by evaluating the Hessian terms of (2.12) at the reference point, where all prices and output are unity, as follows

$$\boldsymbol{H}_{ij} = -\delta_{ij}\left(\sum_{j=1,j\neq i}^{n}\beta_{ij}/2\right) + (1-\delta_{ij})\,\beta_{ij}/2, \tag{2.14}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. By replacing $\boldsymbol{H}$ by $-1/2\ \boldsymbol{K}\boldsymbol{K}'$, where $\boldsymbol{K}$ is an $n \times n$ lower triangular matrix and $\boldsymbol{K}'$ its transpose, the above can be written as

$$-\frac{1}{2}\left(\boldsymbol{K}\boldsymbol{K}'\right)_{ij} = -\delta_{ij}\left(\sum_{j=1,j\neq i}^{n}\beta_{ij}/2\right) + \left(1-\delta_{ij}\right)\beta_{ij}/2. \tag{2.15}$$

There are two things that should be noted here. First, the β_{ii} $(i = 1, \cdots, n)$ do not appear in (2.15), thus leaving β_{ii} $(i = 1, \cdots, n)$ unrestricted. Second, the fact that the elements in the same row of $\boldsymbol{H}$ add to zero, that is

$$\sum_{j=1}^{n}\boldsymbol{H}_{ij} = -\left(\sum_{j=1,j\neq i}^{n}\beta_{ij}/2\right) + \sum_{j=1,j\neq i}^{n}\beta_{ij}/2 = 0, \qquad i = 1, \cdots, n,$$

implies the following restrictions on $\boldsymbol{K}$

$$\sum_{i=1}^{n}k_{ij} = 0, \qquad j = 1, \cdots, n, \tag{2.16}$$

i.e. the elements in the same column of $\boldsymbol{K}$ add to zero, where the k_{ij} terms are the elements of the replacement matrix $\boldsymbol{K}$. (2.16) can be easily shown by expanding out (2.15); a similar technique is also used by Fox and Diewert (1999) in imposing convexity of a Normalized Quadratic profit function in prices. Obtaining the main diagonal elements of $\boldsymbol{K}$, k_{ii}, expressed in terms of k_{ij} $(i \neq j)$ and then substituting them into (2.15), I will obtain β_{ij} $(1 \leq i < j \leq n)$ which are expressed only in terms of k_{ij} $(1 \leq j < i \leq n)$.

As an example, for the case of three inputs $(n = 3)$, I can use the restrictions (2.16) and the lower triangular structure of $\boldsymbol{K}$ in order to eliminate the diagonal elements of

$\boldsymbol{K}$, k_{ii} $(i = 1, 2, 3)$, as follows

$$k_{11} = -k_{21} - k_{31};$$
$$k_{22} = -k_{32};$$
$$k_{33} = 0.$$

Substituting the above restrictions in (2.15), I obtain

$$\beta_{12} = -k_{21}k_{11} = k_{21}(k_{21} + k_{31});$$
$$\beta_{13} = -k_{31}k_{11} = k_{31}(k_{21} + k_{31});$$
$$\beta_{23} = -\left(k_{21}k_{31} + k_{22}k_{32}\right) = -k_{21}k_{31} + k_{32}^2,$$

which guarantees concavity of the cost function at the reference point and may also induce concavity of the cost function at other data points. As already noted above, β_{11}, β_{22}, and β_{33} in this example are unrestricted and do not have to be expressed in terms of the elements of $\boldsymbol{K}$. Clearly, the flexibility of the GL is not destroyed because the $n(n-1)/2$ elements of $\boldsymbol{K}$ just replace the $n(n-1)/2$ elements of $\boldsymbol{H}$ in the estimation.

2.3.2 The Translog Cost Function

The translog specification, due to Christensen *et al.* (1975), is obtained by substituting the translog unit cost function into (2.4) to get

$$\ln C\left(\boldsymbol{p}, y, t\right) = \ln y + \beta_0 + \beta_t t + \sum_{i=1}^{n} \beta_i \ln p_i +$$

$$+ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \beta_{ij} \ln p_i \ln p_j + \sum_{i=1}^{n} \beta_{it} t \ln p_i + \frac{1}{2}\beta_{tt} t^2, \qquad (2.17)$$

where $\beta_{ij} = \beta_{ji}$. Homogeneity of degree one in prices (given y) implies the following restrictions

$$\sum_{i=1}^{n} \beta_i = 1, \ \sum_{i=1}^{n} \beta_{ij} = \sum_{j=1}^{n} \beta_{ji} = \sum_{i=1}^{n} \beta_{it} = 0. \tag{2.18}$$

Although I could estimate (2.17) directly, efficiency gains can be realized by estimating the optimal cost-minimizing input demand equations, transformed into cost-share equations, as follows

$$s_i = \frac{p_i x_i}{C} = \beta_i + \sum_{j=1}^{n} \beta_{ij} \ln p_j + \beta_{it} t, \tag{2.19}$$

where $\sum_{i=1}^{n} p_i x_i = C$.

Guilkey *et al.* (1983) show that the translog is globally regular if and only if technology is Cobb-Douglas. In other words, the translog performs well if substitution between all factors is close to unity. They also show that the regularity properties of the translog model deteriorate rapidly when substitution diverges from unity.

The Hessian matrix of the translog cost function at the reference point, where all prices and output are set to one, will be negative semidefinite if the following matrix is negative semidefinite

$$\boldsymbol{H}_{ij} = \beta_{ij} + \beta_i \beta_j - \delta_{ij} \beta_i, \qquad i, j = 1, \cdots, n, \tag{2.20}$$

with $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Local concavity can be imposed at the reference point as in Ryan and Wales (2000) by setting $\boldsymbol{H} = -\boldsymbol{K}\boldsymbol{K}'$, as follows

$$\beta_{ij} + \beta_i \beta_j - \delta_{ij} \beta_i = (-\boldsymbol{K}\boldsymbol{K}')_{ij}, \qquad i, j = 1, \cdots, n, \tag{2.21}$$

where (as before) $\boldsymbol{K}$ is a lower triangular matrix. Noting that $\sum_{j=1}^{n} \beta_{ij} = 0$ and

$\sum_{j=1}^{n} \beta_j = 0$ (see equation (2.18)), it can be easily shown that

$$\sum_{j=1}^{n} \boldsymbol{H}_{ij} = \sum_{j=1}^{n} \left(\beta_{ij} - \beta_i \delta_{ij} + \beta_i \beta_j\right) = 0, \tag{2.22}$$

i.e. the elements in the same row of $\boldsymbol{H}$ add to zero. Further, (2.22) implies the following restriction on the elements of $\boldsymbol{K}$

$$\sum_{i=1}^{n} k_{ij} = 0, \qquad j = 1, \cdots, n, \tag{2.23}$$

i.e., the elements in the same column of $\boldsymbol{K}$ add to zero. Again, (2.23) can be shown by expanding out $\boldsymbol{H} = -\boldsymbol{K}\boldsymbol{K}'$, where $\boldsymbol{H}$ satisfies (2.22). Combining (2.21) and (2.23), I can replace the elements of $\boldsymbol{B} = [\beta_{ij}]$ by those of $\boldsymbol{K}$. It should be noted that, unlike in the case of the generalized Leontief, β_{ii} $(i = 1, \cdots, n)$ are restricted in this case.

For the case with three inputs ($n = 3$), equations (2.21) and (2.23) imply the following restrictions on the elements of $\boldsymbol{K}$

$$\begin{aligned}
\beta_{11} &= -k_{11}^2 + \beta_1 - \beta_1^2 = -(k_{21} + k_{31})^2 + \beta_1 - \beta_1^2; \\
\beta_{12} &= -k_{11}k_{21} - \beta_1\beta_2 = (k_{21} + k_{31})k_{21} - \beta_1\beta_2; \\
\beta_{13} &= -k_{11}k_{31} - \beta_1\beta_3 = (k_{21} + k_{31})k_{31} - \beta_1\beta_3; \\
\beta_{22} &= -(k_{21}^2 + k_{22}^2) + \beta_2 - \beta_2^2 = -k_{21}^2 - {k_{32}}^2 + \beta_2 - \beta_2^2; \\
\beta_{23} &= -(k_{21}k_{31} + k_{22}k_{32}) - \beta_2\beta_3 = -k_{21}k_{31} + k_{32}^2 - \beta_2\beta_3; \\
\beta_{33} &= -(k_{31}^2 + k_{32}^2 + k_{33}^2) + \beta_3 - \beta_3^2 = -(k_{31}^2 + k_{32}^2) + \beta_3 - \beta_3^2,
\end{aligned}$$

which guarantee concavity of the cost function at the reference point and may also induce concavity of the cost function at other data points. Clearly, the flexibility of the translog specification is not destroyed because the $n(n-1)/2$ elements of $\boldsymbol{K}$ just replace the

$n(n-1)/2$ elements of $\boldsymbol{B}$ in the estimation.

2.3.3 The Normalized Quadratic Cost Function

The NQ model, due to Diewert and Wales (1987), can be obtained by substituting the NQ unit cost function into (2.4)

$$C(\boldsymbol{p}, y, t) = y\left[\sum_{i=1}^{n}\beta_i p_i + \frac{1}{2}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}p_i p_j}{\sum_{i=1}^{N}\alpha_i p_i} + \sum_{i=1}^{n}\beta_{it}p_i t\right], \tag{2.24}$$

where I impose two restrictions on the $\boldsymbol{B} \equiv [\beta_{ij}]$ matrix

$$\beta_{ij} = \beta_{ji}, \qquad \text{for all } i, j; \tag{2.25}$$

$$\boldsymbol{B}\boldsymbol{p}^* = \boldsymbol{0}, \quad \text{for some } \boldsymbol{p}^* > \boldsymbol{0}. \tag{2.26}$$

Further, the $\boldsymbol{\alpha}$ vector ($\boldsymbol{\alpha} > \boldsymbol{0}$) is usually predetermined.

For the NQ cost function, the unknown parameters in (2.24) can estimated by using the following system of factor demands

$$\frac{x_i}{y} = \beta_i + \sum_{j=1}^{n}\beta_{ij}\frac{p_i}{\sum_{i=1}^{n}\alpha_i p_i} - \frac{1}{2}\alpha_i\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}\frac{p_i}{\sum_{i=1}^{n}\alpha_i p_i}\frac{p_j}{\sum_{j=1}^{n}\alpha_j p_j}\right) + \beta_{it}t. \tag{2.27}$$

Before estimating the system in (2.27), I express the main diagonal elements of the $\boldsymbol{B}$ matrix, β_{ii}, in terms of its off-diagonal elements by using equation (2.26) and assuming that $\boldsymbol{p}^* = \boldsymbol{1}_n$. Thus, by estimating the input-output equations (2.27), I obtain estimates of β_i, the technical change parameters β_{it}, and the off-diagonal elements of the $\boldsymbol{B}$ matrix, β_{ij} ($i \neq j$). The main diagonal elements of the $\boldsymbol{B}$ matrix can be recovered from the restrictions imposed.

The Hessian matrix of the cost function (2.24) is obtained as follows

$$\nabla_{p_i p_j} C(\boldsymbol{p}, y, t) = \frac{\beta_{ij}}{\sum_{i=1}^{n} \alpha_i p_i} - \frac{\alpha_i \left(\sum_{j=1}^{n} \beta_{ij} p_j\right)}{\left(\sum_{i=1}^{n} \alpha_i p_i\right)^2}$$

$$- \frac{\alpha_i \left(\sum_{i=1}^{n} \beta_{ij} p_i\right)}{\left(\sum_{i=1}^{n} \alpha_i p_i\right)^2} + \frac{\alpha_i \alpha_j \left(\sum_{i=1}^{n} \sum_{j=1}^{n} p_i \beta_{ij} p_j\right)}{\left(\sum_{i=1}^{n} \alpha_i p_i\right)^3}. \tag{2.28}$$

Using the restrictions $\sum_{j=1}^{n} \beta_{ij} p_j^* = \mathbf{0}_n$ at the reference point, I have $\sum_{i=1}^{n} \sum_{j=1}^{n} p_i^* \beta_{ij} p_j^* = \sum_{i=1}^{n} \left(p_i^* \left(\sum_{j=1}^{n} \beta_{ij} p_j^*\right)\right) = 0$. Thus evaluating the above equation at $(\boldsymbol{p}^*, t^*)$ yields the following equation

$$\nabla_{p_i p_j} C(\boldsymbol{p}, y, t) = \frac{\beta_{ij}}{\left(\sum_{i=1}^{n} \alpha_i p_i^*\right)}. \tag{2.29}$$

Multiplying both sides of (2.29) by y and rearranging, I get $\nabla_{p_i p_j} C(\boldsymbol{p}, y, t) = \boldsymbol{\alpha}' \, \boldsymbol{p}^{-1} \boldsymbol{B}$. Thus the negative semidefiniteness of $\nabla_{p_i p_j} C(\boldsymbol{p}, y, t)$ at the reference point requires that $\boldsymbol{B}$ is negative semidefinite. More importantly, the negative semidefiniteness of $\boldsymbol{B}$ is not only the necessary condition for $\nabla_{p_i p_j} C(\boldsymbol{p}, y, t)$ to be concave locally at the reference point as I just showed, but it is also a sufficient condition for $\nabla_{p_i p_j} C(\boldsymbol{p}, y, t)$ to be concave globally (concave at every possible and imaginable point) — see Diewert and Wales (1987) for more details.

In practice, the concavity of $C(\boldsymbol{p}, y, t)$ may not be satisfied, in the sense that the estimated $\boldsymbol{B}$ matrix may not be negative semidefinite. In this case, to ensure global concavity (concavity at all possible prices) of the NQ cost function, I follow Diewert and Wales (1987) and impose

$$\boldsymbol{B} = -\boldsymbol{K}\boldsymbol{K}', \tag{2.30}$$

where $\boldsymbol{K}$ is a lower triangular matrix which satisfies

$$\boldsymbol{K}'\boldsymbol{p}^* = \mathbf{0}_n. \tag{2.31}$$

Note that (2.31) and the lower triangular structure of $\boldsymbol{K}$ imply

$$\sum_{i=1}^{n} k_{ij} = 0, \qquad j = 1, \cdots, n. \tag{2.32}$$

As an example, for the case of three inputs (2.30) and (2.32) imply

$$\begin{aligned}
\beta_{11} &= -k_{11}^2 = -\left(k_{21} + k_{31}\right)^2; \\
\beta_{12} &= -k_{11}k_{21} = \left(k_{21} + k_{31}\right)k_{21}; \\
\beta_{13} &= -k_{11}k_{31} = \left(k_{21} + k_{31}\right)k_{31}; \\
\beta_{22} &= -\left(k_{21}^2 + k_{22}^2\right) = -k_{21}^2 - k_{32}^2; \\
\beta_{23} &= -\left(k_{21}k_{31} + k_{22}k_{32}\right) = -k_{21}k_{31} + k_{32}^2; \\
\beta_{33} &= -\left(k_{31}^2 + k_{32}^2 + k_{33}^2\right) = -\left(k_{31}^2 + k_{32}^2\right).
\end{aligned}$$

That is, I replace the elements of $\boldsymbol{B}$ in the input-output equations (2.27) by the elements of $\boldsymbol{K}$, thus ensuring global curvature. It should be noted that in the case of the NQ cost model, concavity is imposed globally rather than locally at the reference point as I do in the case of the GL and translog specifications. The main advantage of the NQ specification comes from its property that correct curvature conditions can be imposed globally without destroying the flexibility of the functional form.

2.3.4 The AIM Cost Function

By specifying the unit cost function in (2.4) as a linearly homogeneous multivariate Müntz-Szatz series expansion, I get the AIM total cost function without technical change [see Barnett *et al.* (1991)]

$$C = g(\boldsymbol{p}, y) = y \left[\sum_{z \in A_\kappa} a_z \prod_{j=1}^{2^\kappa} p_{i_j}^{2^{-\kappa}} \right], \tag{2.33}$$

where κ is the order of expansion, a_z the unknown parameters, n the number of production factors, and $A_\kappa = \{(i_1, i_2, \cdots, i_{2^\kappa}) : i_1, i_2, \cdots, i_{2^\kappa} \in \{1, 2, \cdots, n); \; i_1 \leq i_2, \leq \cdots \leq i_{2^\kappa}\}$. For simplicity, I call a cost function without technical change the 'stripped-down cost function.'

Now I extend the Barnett *et al.* (1991) stripped-down AIM cost function to allow for technical change — a very valuable source of information about modeling technical change is Sato (1975). Instead of using the generic time trend to model technical change, as I did with the three locally flexible functional forms, I introduce technical change into the stripped-down AIM cost function using the efficiency index approach. In particular, I assume that the effects of technical change (t) on the production level y are purely factor-augmenting; that is, affecting each factor through a factor specific efficiency index, $e_i = e_i(t, y)$ — factor augmenting technical change was pioneered by Kohli (1981, 1982, 1991, 1993).

Thomsen (2000) shows generally that in order to obtain a total cost function with technical change and returns to scale, $C(\boldsymbol{p}, y, t)$, one can first figure out the stripped-down cost function denoted by $C^*(\boldsymbol{p}, y)$, and then divide $\boldsymbol{p}$ in $C^*(\boldsymbol{p}, y)$ by a factor specific efficiency index. Thomsen (2000) further shows that the efficiency index is capable of rendering any stripped-down cost function flexible in y and t. Under the assumption of

constant returns to scale I specify the efficiency index as[2]

$$\log e_i = \vartheta_i t \qquad i = 1, \cdots, n, \tag{2.34}$$

where ϑ_i indicates (when multiplied by 100) the percentage increase in the efficiency of factor i from period t to $t+1$. By dividing $\boldsymbol{p}$ in (2.33) by the above efficiency index, I obtain my AIM cost function with technical change

$$C = g(\boldsymbol{p}, y, t) = y\left[\sum_{z \in A_\kappa} a_z \prod_{j=1}^{2^\kappa} q^{2^{-\kappa}}\right]$$

$$= y\left[\sum_{z \in A_\kappa} a_z \prod_{j=1}^{2^\kappa} \left(p_{i_j} e_{i_j}^{-\theta_{i_j} t}\right)^{2^{-\kappa}}\right] = yQ(q), \tag{2.35}$$

where q_{i_j} is the efficiency-corrected price, defined as $q_{i_j} = p_{i_j}/e_{i_j}$, e is the efficiency index as defined by (2.34), and $Q(q)$ is the corresponding unit cost function. The main advantage of the efficiency index approach is that I can easily obtain a new AIM cost function with technical change which retains all of the theoretical properties of the stripped-down AIM cost function. Another advantage of this approach is that one can measure input-specific productivity, changes in input productivity, and the contribution of each input to overall productivity, unlike the generic time trend models of technical change. I shall discuss these advantages in more detail in what follows.

My AIM total cost function with technical change retains all the theoretical properties of the Barnett *et al.* (1991) stripped-down AIM cost function. First, my AIM total cost function with technical change is still globally flexible in the sense that it is capable of approximating the underlying cost function at every point in the function's domain by

[2]Returns to scale can be easily incorporated in the AIM cost function by modifying the efficiency index. Regarding the assumption of constant returns to scale, see the description of the data in Section 4.

increasing the order of expansion κ. Second, it can be clearly seen from (2.35) that the sum of the exponents of prices in each term is still $2^{\kappa}2^{-\kappa} = 1$, thus satisfying the property of global linear homogeneity. Third, as with the stripped-down AIM cost function, I can impose concavity and monotonicity on the coefficients of my AIM total cost function with technical change by requiring all the coefficients to be nonnegative. In particular, with nonnegative coefficients, the function $\left(p_{i_j} e_{i_j}^{-\vartheta_{i_j} t}\right)^{2^{-\kappa}}$ is increasing and concave in $\boldsymbol{p}$ for any fixed κ. Hence, according to Berge (1963, Theorem 1), $\prod_{j=1}^{2^{\kappa}} \left(p_{i_j} e_{i_j}^{-\vartheta_{i_j} t}\right)^{2^{-\kappa}}$ is increasing and quasiconcave jointly in all of its variables for any fixed κ. However, as shown by Diewert and Wales (1993) when global concavity is imposed on this functional form in this manner, it is not flexible and complements are ruled out.

Applying Shephard's lemma (2.5) to (2.35), and dividing through by y, yields optimal input-output demand equations, as follows

$$\frac{x_i}{y} = \frac{\partial}{\partial p_i} \left(\sum_{z \in A_{\kappa}} a_z \prod_{j=1}^{2^{\kappa}} \left(p_{i_j} e_{i_j}^{-\vartheta_{i_j} t}\right)^{2^{-\kappa}} \right). \tag{2.36}$$

The system of factor demand functions produced by applying Shephard's lemma to the κth partial sum of the cost function, $C_k(\boldsymbol{p}, y, t)$ will be called the AIM(κ) factor demand system, and the resulting input-output equations will be called the AIM(κ) input-output system.

In empirical applications, the approximation of the AIM cost function must be truncated at some finite value κ (i.e. finite partial sums). The order of approximation κ is usually determined empirically and stops when the elasticity estimates and the covariance matrix of the disturbances converge. By using formula (2.35) and (2.36), I now explicitly produce the first two partial sum of my expansion of the cost function in the four-factor case, AIM(1) and AIM(2) respectively.

For four goods ($n = 4$) and $\kappa = 1$, the AIM(1) cost function can be written as

$$C_{\kappa=1}(\boldsymbol{p}, y, t) = y\Big(\alpha_1 q_1 + \alpha_2 q_2 + \alpha_3 q_3 + \alpha_4 q_4 + \alpha_5 q_1^{1/2} q_2^{1/2} + \alpha_6 q_1^{1/2} q_3^{1/2} + \alpha_7 q_1^{1/2} q_4^{1/2} + \alpha_8 q_2^{1/2} q_3^{1/2} + \alpha_9 q_2^{1/2} q_4^{1/2} + \alpha_{10} q_3^{1/2} q_4^{1/2}\Big). \quad (2.37)$$

Applying Shephard's lemma (2.5) to (2.37) yields the factor demand equations of the AIM(1) model

$$\begin{aligned}
\frac{x_1}{y} &= \frac{1}{e^{\vartheta_1 t}}\left(\alpha_1 + \frac{1}{2}\alpha_5 q_1^{-1/2} q_2^{1/2} + \frac{1}{2}\alpha_6 q_1^{-1/2} q_3^{1/2} + \frac{1}{2}\alpha_7 q_1^{-1/2} q_4^{1/2}\right);\\
\frac{x_2}{y} &= \frac{1}{e^{\vartheta_2 t}}\left(\alpha_2 + \frac{1}{2}\alpha_5 q_1^{1/2} q_2^{-1/2} + \frac{1}{2}\alpha_8 q_2^{-1/2} q_3^{1/2} + \frac{1}{2}\alpha_9 q_2^{-1/2} q_4^{1/2}\right);\\
\frac{x_3}{y} &= \frac{1}{e^{\vartheta_3 t}}\left(\alpha_3 + \frac{1}{2}\alpha_6 q_1^{1/2} q_3^{-1/2} + \frac{1}{2}\alpha_8 q_2^{1/2} q_3^{-1/2} + \frac{1}{2}\alpha_{10} q_3^{-1/2} q_4^{1/2}\right);\\
\frac{x_4}{y} &= \frac{1}{e^{\vartheta_4 t}}\left(\alpha_4 + \frac{1}{2}\alpha_7 q_1^{1/2} q_4^{-1/2} + \frac{1}{2}\alpha_9 q_2^{1/2} q_4^{-1/2} + \frac{1}{2}\alpha_{10} q_3^{1/2} q_4^{-1/2}\right).
\end{aligned}$$

For $n = 4$ and $k = 2$, the AIM(2) cost function can written as

$$
\begin{aligned}
C_{\kappa=2}(p, y, t) = y \Big(& \alpha_1 q_1 + \alpha_2 q_2 + \alpha_3 q_3 + \alpha_4 q_4 \\
& + \alpha_5 q_1^{1/2} q_2^{1/2} + \alpha_6 q_1^{1/2} q_3^{1/2} + \alpha_7 q_1^{1/2} q_4^{1/2} \\
& + \alpha_8 q_2^{1/2} q_3^{1/2} + \alpha_9 q_2^{1/2} q_4^{1/2} + \alpha_{10} q_3^{1/2} q_4^{1/2} \\
& + \alpha_{11} q_1^{3/4} q_2^{1/4} + \alpha_{12} q_1^{1/4} q_2^{3/4} + \alpha_{13} q_1^{3/4} q_3^{1/4} + \alpha_{14} q_1^{1/4} q_3^{3/4} \\
& + \alpha_{15} q_1^{3/4} q_4^{1/4} + \alpha_{16} q_1^{1/4} q_4^{3/4} + \alpha_{17} q_2^{3/4} q_3^{1/4} + \alpha_{18} q_2^{1/4} q_3^{3/4} \\
& + \alpha_{19} q_2^{3/4} q_4^{1/4} + \alpha_{20} q_2^{1/4} q_4^{3/4} + \alpha_{21} q_3^{3/4} q_4^{1/4} + \alpha_{22} q_3^{1/4} q_4^{3/4} \\
& + \alpha_{23} q_1^{1/2} q_2^{1/4} q_3^{1/4} + \alpha_{24} q_1^{1/4} q_2^{1/2} q_3^{1/4} + \alpha_{25} q_1^{1/4} q_2^{1/4} q_3^{1/2} \\
& + \alpha_{26} q_1^{1/2} q_2^{1/4} q_4^{1/4} + \alpha_{27} q_1^{1/4} q_2^{1/2} q_4^{1/4} + \alpha_{28} q_1^{1/4} q_2^{1/4} q_4^{1/2} \\
& + \alpha_{29} q_1^{1/2} q_3^{1/4} q_4^{1/4} + \alpha_{30} q_3^{1/2} q_1^{1/4} q_4^{1/4} + \alpha_{31} q_4^{1/2} q_1^{1/4} q_3^{1/4} \\
& + \alpha_{32} q_2^{1/2} q_3^{1/4} q_4^{1/4} + \alpha_{33} q_2^{1/4} q_3^{1/2} q_4^{1/4} + \alpha_{34} q_2^{1/4} q_3^{1/4} q_4^{1/2} \\
& + \alpha_{35} q_1^{1/4} q_2^{1/4} q_3^{1/4} q_4^{1/4} \Big) . \qquad (2.38)
\end{aligned}
$$

Applying (2.5) to (2.38) yields the following system of factor demand equations for the

AIM(2) model

$$\begin{aligned}\frac{x_1}{y} = \frac{1}{e^{\vartheta_1 t}}\Bigg(&\alpha_1 + \frac{1}{2}\alpha_5 q_1^{-1/2} q_2^{1/2} + \frac{1}{2}\alpha_6 q_1^{-1/2} q_3^{1/2} + \frac{1}{2}\alpha_7 q_1^{-1/2} q_4^{1/2} \\ &+ \frac{3}{4}\alpha_{11} q_1^{-1/4} q_2^{1/4} + \frac{1}{4}\alpha_{12} q_1^{-3/4} q_2^{3/4} + \frac{3}{4}\alpha_{13} q_1^{-1/4} q_3^{1/4} \\ &+ \frac{1}{4}\alpha_{14} q_1^{-3/4} q_3^{3/4} + \frac{3}{4}\alpha_{15} q_1^{-1/4} q_4^{1/4} + \frac{1}{4}\alpha_{16} q_1^{-3/4} q_4^{3/4} \\ &+ \frac{1}{2}\alpha_{23} q_1^{-1/2} q_2^{1/4} q_3^{1/4} + \frac{1}{4}\alpha_{24} q_1^{-3/4} q_2^{1/2} q_3^{1/4} + \frac{1}{4}\alpha_{25} q_1^{-3/4} q_2^{1/4} q_3^{1/2} \\ &+ \frac{1}{2}\alpha_{26} q_1^{-1/2} q_2^{1/4} q_4^{1/4} + \frac{1}{4}\alpha_{27} q_1^{-3/4} q_2^{1/2} q_4^{1/4} + \frac{1}{4}\alpha_{28} q_1^{-3/4} q_2^{1/4} q_4^{1/2} \\ &+ \frac{1}{2}\alpha_{29} q_1^{-1/2} q_3^{1/4} q_4^{1/4} + \frac{1}{4}\alpha_{30} q_1^{-3/4} q_3^{1/2} q_4^{1/4} + \frac{1}{4}\alpha_{31} q_1^{-3/4} q_3^{1/4} q_4^{1/2} \\ &+ \frac{1}{4}\alpha_{35} q_1^{-3/4} q_2^{1/4} q_3^{1/4} q_4^{1/4}\Bigg); \end{aligned} \tag{2.39}$$

$$\begin{aligned}\frac{x_2}{y} = \frac{1}{e^{\vartheta_2 t}}\Bigg(&\alpha_2 + \frac{1}{2}\alpha_5 q_1^{1/2} q_2^{-1/2} + \frac{1}{2}\alpha_8 q_2^{-1/2} q_3^{1/2} + \frac{1}{2}\alpha_9 q_2^{-1/2} q_4^{1/2} \\ &+ \frac{1}{4}\alpha_{11} q_1^{3/4} q_2^{-3/4} + \frac{3}{4}\alpha_{12} q_1^{1/4} q_2^{-1/4} + \frac{3}{4}\alpha_{17} q_2^{-1/4} q_3^{1/4} \\ &+ \frac{1}{4}\alpha_{18} q_2^{-3/4} q_3^{3/4} + \frac{3}{4}\alpha_{19} q_2^{-1/4} q_4^{1/4} + \frac{1}{4}\alpha_{20} q_2^{-3/4} q_4^{3/4} \\ &+ \frac{1}{4}\alpha_{23} q_1^{1/2} q_2^{-3/4} q_3^{1/4} + \frac{1}{2}\alpha_{24} q_1^{1/4} q_2^{-1/2} q_3^{1/4} + \frac{1}{4}\alpha_{25} q_1^{1/4} q_2^{-3/4} q_3^{1/2} \\ &+ \frac{1}{4}\alpha_{26} q_1^{1/2} q_2^{-3/4} q_4^{1/4} + \frac{1}{2}\alpha_{27} q_1^{1/4} q_2^{-1/2} q_4^{1/4} + \frac{1}{4}\alpha_{28} q_1^{1/4} q_2^{-3/4} q_4^{1/2} \\ &+ \frac{1}{2}\alpha_{32} q_2^{-1/2} q_3^{1/4} q_4^{1/4} + \frac{1}{4}\alpha_{33} q_2^{-3/4} q_3^{1/2} q_4^{1/4} + \frac{1}{4}\alpha_{34} q_2^{-3/4} q_3^{1/4} q_4^{1/2} \\ &+ \frac{1}{4}\alpha_{35} q_1^{1/4} q_2^{-3/4} q_3^{1/4} q_4^{1/4}\Bigg); \end{aligned} \tag{2.40}$$

$$\frac{x_3}{y} = \frac{1}{e^{\vartheta_3 t}}\left(\alpha_3 + \frac{1}{2}\alpha_6 q_1^{1/2} q_3^{-1/2} + \frac{1}{2}\alpha_8 q_2^{1/2} q_3^{-1/2} + \frac{1}{2}\alpha_{10} q_3^{-1/2} q_4^{1/2} + \frac{1}{4}\alpha_{13} q_1^{3/4} q_3^{-3/4} + \frac{3}{4}\alpha_{14} q_1^{1/4} q_3^{-1/4} + \frac{1}{4}\alpha_{17} q_2^{3/4} q_3^{-3/4} + \frac{3}{4}\alpha_{18} q_2^{1/4} q_3^{-1/4} + \frac{3}{4}\alpha_{21} q_3^{-1/4} q_4^{1/4} + \frac{1}{4}\alpha_{22} q_3^{-3/4} q_4^{3/4} + \frac{1}{4}\alpha_{23} q_1^{1/2} q_2^{1/4} q_3^{-3/4} + \frac{1}{4}\alpha_{24} q_1^{1/4} q_2^{1/2} q_3^{-3/4} + \frac{1}{2}\alpha_{25} q_1^{1/4} q_2^{1/4} q_3^{-1/2} + \frac{1}{4}\alpha_{29} q_1^{1/2} q_3^{-3/4} q_4^{1/4} + \frac{1}{2}\alpha_{30} q_1^{1/4} q_3^{-1/2} q_4^{1/4} + \frac{1}{4}\alpha_{31} q_1^{1/4} q_3^{-3/4} q_4^{1/2} + \frac{1}{4}\alpha_{32} q_2^{1/2} q_3^{-3/4} q_4^{1/4} + \frac{1}{2}\alpha_{33} q_2^{1/4} q_3^{-1/2} q_4^{1/4} + \frac{1}{4}\alpha_{34} q_2^{1/4} q_3^{-3/4} q_4^{1/2} + \frac{1}{4}\alpha_{35} q_1^{1/4} q_2^{1/4} q_3^{-3/4} q_4^{1/4}\right); \tag{2.41}$$

$$\frac{x_4}{y} = \frac{1}{e^{\vartheta_4 t}}\left(\alpha_4 + \frac{1}{2}\alpha_7 q_1^{1/2} q_4^{-1/2} + \frac{1}{2}\alpha_9 q_2^{1/2} q_4^{-1/2} + \frac{1}{2}\alpha_{10} q_3^{1/2} q_4^{-1/2} + \frac{1}{4}\alpha_{15} q_1^{3/4} q_4^{-3/4} + \frac{3}{4}\alpha_{16} q_1^{1/4} q_4^{-1/4} + \frac{1}{4}\alpha_{19} q_2^{3/4} q_4^{-3/4} + \frac{3}{4}\alpha_{20} q_2^{1/4} q_4^{-1/4} + \frac{1}{4}\alpha_{21} q_3^{3/4} q_4^{-3/4} + \frac{3}{4}\alpha_{22} q_3^{1/4} q_4^{-1/4} + \frac{1}{4}\alpha_{26} q_1^{1/2} q_2^{1/4} q_4^{-3/4} + \frac{1}{4}\alpha_{27} q_1^{1/4} q_2^{1/2} q_4^{-3/4} + \frac{1}{2}\alpha_{28} q_1^{1/4} q_2^{1/4} q_4^{-1/2} + \frac{1}{4}\alpha_{29} q_1^{1/2} q_3^{1/4} q_4^{-3/4} + \frac{1}{4}\alpha_{30} q_1^{1/4} q_3^{1/2} q_4^{-3/4} + \frac{1}{2}\alpha_{31} q_1^{1/4} q_3^{1/4} q_4^{-1/2} + \frac{1}{4}\alpha_{32} q_2^{1/2} q_3^{1/4} q_4^{-3/4} + \frac{1}{4}\alpha_{33} q_2^{1/4} q_3^{1/2} q_4^{-3/4} + \frac{1}{2}\alpha_{34} q_2^{1/4} q_3^{1/4} q_4^{-1/2} + \frac{1}{4}\alpha_{35} q_1^{1/4} q_2^{1/4} q_3^{1/4} q_4^{-3/4}\right). \tag{2.42}$$

Concavity (in prices) requires that the Hessian matrix of the second derivatives of the cost function with respect to prices, $\nabla_{p_i p_j} C(\boldsymbol{p}, y, t)$, is negative semidefinite. In practice, concavity of the cost function may not be satisfied. In this case, I impose concavity fully (at every data point in the sample) on the AIM model using methods suggested by Gallant and Golub (1984) in the case of the Fourier cost function and recently used in the context of consumer demand systems estimation by Serletis and Shahmoradi (2005)

— I shall discuss this in detail in Section 4.2.

2.4 Data and Econometric Issues

I use annual KLEM (capital, labor, energy, and intermediate materials) data for total manufacturing in the United States over the period from 1953 to 2001. All series are from the website of the U.S. Bureau of Labor Statistics (BLS), at www.bls.gov/data/home.htm. The data consists of price and quantity indices for one output and four inputs (capital, labor, energy, and materials). All the price series have been normalized to one in 1953 and the quantity indices for output, capital, labor, energy, materials, and purchased business services have been obtained by dividing value of production or factor costs by the corresponding normalized price index. It is to be noted that I constructed the price and quantity indices for intermediate materials as subaggregates over the two components, materials and purchased business services, using the Fisher ideal index.

A major feature of the BLS data set is that constant returns to scale is built in by constructing input factor payments in such a way that they add up to the value of output. Thus, tests of returns to scale and scale bias are inappropriate, as are some tests of imperfect competition. Another feature of the BLS data set is that it provides the price and quantity series for purchased business services inputs. Directly collected data on purchased business services are relatively scant, and for that reason they have been ignored by similar studies in the past. However, there is ample evidence of an increased use of purchased business services by industries over the post-war period and there are two important issues to consider. The first is that a sizable and growing input should not be ignored in productivity measurement, if aggregate inputs are not to be underestimated and mismeasured. The other is the possibility of substitution between capital, labor, and services purchased from outside. Examples of the latter are the substitution of leased

equipment for owned capital and purchased accounting for services performed by payroll employees.

Maximum likelihood estimates of the three locally flexible cost functions with or without curvature imposed is straightforward, and can be approached in a variety of well-known ways. The estimation of the AIM model without curvature imposed can also be approached easily in the same way as with locally flexible cost functions.

The estimation, however, of the AIM model with curvature imposed cannot be approached in the usual way, and has to resort to some more advanced methods. For example, Gallant and Golub (1984) used the NPSOL subroutine of the Stanford Systems Optimization Laboratory to estimate the constrained Fourier cost function without technical change. Also Barnett *et al.* (1991) used numerical Bayesian estimation to solve a relatively simple constrained AIM cost function with only two factors (capital and labor) and no technical change. In this paper, I follow Gallant and Golub (1984) and Serletis and Shahmoradi (2005) and use the TOMLAB/NPSOL tool box with MATLAB — see http://tomlab.biz/products/npsol. NPSOL uses a sequential quadratic programming algorithm and is suitable for both unconstrained and constrained optimization of smooth (that is, at least twice-continuously differentiable) nonlinear functions.

2.4.1 Parametric Estimation of the Locally Flexible Forms

In order to estimate equation systems such as (2.13), (2.19), and (2.27), a stochastic component, ϵ_t, is added to the set of input-output equations or share equations as follows

$$\boldsymbol{w}_t = \boldsymbol{\psi}\left(\boldsymbol{p}_t, y, t, \boldsymbol{\theta}\right) + \boldsymbol{\epsilon}_t, \tag{2.43}$$

where $\boldsymbol{w} = (w_1, \cdots, w_n)'$ is the vector of input-output ratios in the case of the GL and NQ models and that of input shares in the case of the translog model. $\boldsymbol{\epsilon}_t$ is a

vector of stochastic errors and I assume that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Omega})$ where $\mathbf{0}$ is a null matrix and $\boldsymbol{\Omega}$ is the $n \times n$ symmetric positive definite error covariance matrix. $\boldsymbol{\psi}(\boldsymbol{p}_t, y, t, \boldsymbol{\theta}) = (\psi_1(\boldsymbol{p}_t, y, t, \boldsymbol{\theta}), \cdots, \psi_n(\boldsymbol{p}_t, y, t, \boldsymbol{\theta}))'$, and $\psi_i(\boldsymbol{p}_t, y, t, \boldsymbol{\theta})$ is given by the right-hand side of each of (2.13), (2.19), and (2.27).

In the case of the translog model, since the shares in (2.19) sum to unity, the random disturbances corresponding to the four share equations sum to zero and this yields a singular covariance matrix of errors. Barten (1969) has shown that full information maximum likelihood estimates of the parameters can be obtained by arbitrarily deleting any one equation. The resulting estimates are invariant with respect to the equation deleted and the parameter estimates of the deleted equation can be recovered from the restrictions imposed.

Another issue concerning my stochastic specification is that of endogeneity. At the individual firm level, it may be reasonably assumed that inputs prices on the right hand side of (2.43) are exogenous. At the more aggregated industry level (like U.S. manufacturing), however, input prices are less likely to be exogenous. In this literature, the possibility of endogeneity has been addressed by using iterative three-stage least squares (3SLS), but the results generally have been about the same as those with iterative Zellner estimation — see, for example, Barnett *et al.* (1991). Diewert and Fox (2004) also argue that instrumental variables estimation may be more biased, since the instruments may not be completely exogenous, and Burnside (1996) shows that results can vary markedly depending on the set of instruments used. In this paper, I choose to use the more commonly used iterative Zellner method of estimation.

The estimation is performed in TSP/GiveWin (version 4.5), using the LSQ procedure, and the regularity conditions are checked as follows:

- Positivity is checked by checking if the estimated cost is positive,

$$C(\boldsymbol{p}, y, t) > 0.$$

- Monotonicity is checked by direct computation of the values of the first gradient vector of the estimated cost function with respect to $\boldsymbol{p}$. It is satisfied if $\nabla_p C(\boldsymbol{p}, y, t) > 0$.

- Curvature requires the Hessian matrix of the cost function to be negative semidefinite and is checked by performing a Cholesky factorization of that matrix and checking whether the Cholesky values are nonpositive [since a matrix is negative semidefinite if its Cholesky factors are nonpositive — see Lau (1978, Theorem 3.2)]. Curvature can also be checked by examining the eigenvalues of the Hessian matrix provided that the monotonicity condition holds. It requires that these eigenvalues be negative or zero.

2.4.2 Semi-Nonparametric Estimation of the AIM(2) Cost Function

The AIM(2) factor demand system can be written as

$$z_t = \psi(\boldsymbol{p}, y, t, \boldsymbol{\theta}) + \epsilon_t, \tag{2.44}$$

where $\boldsymbol{z} = (z_1, \cdots, z_n)'$ is the vector of input-output ratios, $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \alpha_3, \cdots, a_{n^{2\kappa}}, \vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4)$, and $\psi(\boldsymbol{p}, y, t, \boldsymbol{\theta})$ is given by the the right hand side of (2.39)-(2.42).

As Gallant and Golub (1984, p. 298) put it,

> "all statistical estimation procedures that are commonly used in econometric research can be formulated as an optimization problem of the following type

[Burguete, Gallant and Souza (1982)]

$$\widehat{\boldsymbol{\theta}} \text{ minimizes } \varphi(\boldsymbol{\theta}) \text{ over } \Theta \tag{2.45}$$

with $\varphi(\boldsymbol{\theta})$ twice continuously differentiable in $\boldsymbol{\theta}$."

Notice that $\boldsymbol{\psi}(\boldsymbol{p}, y, t, \boldsymbol{\theta})$ is nonlinear in $\vartheta_1, \vartheta_2, \vartheta_3$, and ϑ_4, and therefore the AIM(2) factor demand system in (2.44) can be fitted using Gallant's (1975, p. 36) seemingly unrelated nonlinear regression method to estimate $\boldsymbol{\theta}$. Hence, $\varphi(\boldsymbol{\theta})$ has the form

$$\varphi(\boldsymbol{\theta}) = \frac{1}{T}\boldsymbol{\epsilon}_t'\boldsymbol{\epsilon}_t = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{z}_t - \boldsymbol{\psi}(\cdot))'\,\widehat{\Omega}^{-1}(\boldsymbol{z}_t - \boldsymbol{\psi}(\cdot)), \tag{2.46}$$

where $\widehat{\Omega}$ is an estimate of the error variance-covariance matrix of (2.44). In minimizing (2.44), I use the TOMLAB/NPSOL tool box with MATLAB. NPSOL uses a sequential quadratic programming algorithm and is suitable for both unconstrained and constrained optimization of smooth (that is, at least twice-continuously differentiable) nonlinear functions.

I first run an unconstrained optimization using (2.45). As results in nonlinear optimization are sensitive to the initial parameter values, to achieve global convergence, I randomly generated 500 sets of initial parameter values and chose the starting $\boldsymbol{\theta}$ that led to the lowest value of the objective function. I also check the regularity conditions, i.e. positivity, monotonicity, and curvature conditions, using the same methods as specified above for the three locally flexible functional forms.

In case where the curvature conditions are not satisfied at all observations, I then use the NPSOL nonlinear programming program to minimize $\varphi(\boldsymbol{\theta})$ subject to the constraint that the four eigenvalues of the Hessian matrix, $\boldsymbol{H}$, are non-positive. This is because a necessary and sufficient condition for the concavity of $\boldsymbol{H}$ is that all its eigenvalues are

nonpositive — see, for example, Morey (1986). The first derivatives of these eigenvalues are needed for the optimization algorithms and can be easily obtained using Matlab's Symbolic Math Toolbox. Thus, my constrained optimization problem can also be written as

$$\min_{\boldsymbol{\theta}} \varphi(\boldsymbol{\theta}) \text{ subject to } \varphi_i(\boldsymbol{p}, y, t, \boldsymbol{\theta}) < 0, \qquad i = 1, \cdots, n,$$

where $\varphi_i(\boldsymbol{p}, y, t, \boldsymbol{\theta})$, $i = 1, \cdots, n$, are the eigenvalues of the Hessian matrix of the AIM(2) cost function. With the constrained optimization method, I can impose curvature restrictions at any arbitrary set of points — at a single data point, over a region of data points, or fully (at every data point in the sample).

2.5 Empirical Evidence

2.5.1 Economic Regularity

Tables 2.1 - 2.4 contain a summary of results from the GL, translog, NQ, and AIM(2) models in terms of parameter estimates and theoretical regularity violations when the models are estimated without the curvature conditions imposed and with the curvature conditions imposed. Clearly, all models satisfy positivity and monotonicity at all sample observations when curvature is not imposed. However, all three locally flexible models — the GL, translog, and NQ — violate curvature at all 49 observations when curvature conditions are not imposed. Similarly, the AIM(2) model violates curvature at 33 data points when curvature is not imposed.

Because regularity hasn't been attained for any of the models, I follow the procedures discussed in Section 3 to impose curvature. In the case of the GL and translog models, I impose local curvature using the Ryan and Wales (2000) procedure. However, as noted by Ryan and Wales (2000), the ability of locally flexible models to satisfy curvature at sample observations other than the point of approximation, depends on the choice of the

approximation point. Thus, I estimate each model 49 times (a number of times equal to the number of observations) and report results for the best approximation point (best in the sense of satisfying the curvature conditions at the largest number of observations) — the best approximation point is 1982 for the GL and 1981 for the translog. In the case of the NQ model I impose global curvature following the procedure suggested by Diewert and Wales (1987). As for the AIM(2) model, I minimize $\varphi(\boldsymbol{\theta})$ subject to the constraint that the cost function is locally concave in 1981 and also subject to the constraint that it is fully concave (concave at every data point).

The estimation results of the three locally flexible functional forms with curvature imposed are reported in the second column of Tables 2.1-2.3. My findings in terms of regularity violations when the curvature conditions are imposed are disappointing in the case of the GL and translog models. In particular, the imposition of local curvature on the translog model reduces the number of curvature violations from 49 to 6. The performance of the GL is not satisfactory either, since the imposition of local curvature does not completely eliminate the curvature violations; it reduces the number of curvature violations from 49 to 2. As Barnett (2002, p.199) put it, without satisfaction of all three theoretical regularity conditions, "the second-order conditions for optimizing behavior fail, and duality theory fails. The resulting first-order conditions, demand functions, and supply functions become invalid." As expected, however, the imposition of global curvature (at all possible prices) on the NQ model reduces the number of curvature violations to zero, without any induced violations of monotonicity.

Using NPSOL I imposed the curvature condition on the AIM(2) model and report the results in the second and third columns of Table 2. 4 — the second column shows the results when the curvature constraint is imposed locally (in 1981) and the third column shows the results when the constraint is imposed at every data point in the sample. Clearly, the effect of imposing the curvature constraint locally is negligible, as

the number of curvature violations drops only from 33 to 32. However, the imposition of the curvature constraint at every data point in the sample has a significant impact on the AIM(2) model, as I obtain parameter estimates that are consistent with all three theoretical regularity conditions, at every data point in the sample; that is, fully.

2.5.2 Econometric Regularity

I have estimated input-output demand equations and share equations from aggregate time series data and highlighted the challenge inherent with achieving economic regularity and the need for economic theory to inform econometric research. Incorporating restrictions from economic theory seems to be gaining popularity as there are also numerous recent papers that estimate stochastic dynamic general equilibrium models using economic restrictions — see, for example, Aliprantis *et al.* (2007). With the focus on economic theory, however, I have ignored econometric regularity. In particular, I have ignored unit root and cointegration issues, because the combination of nonstationary data and nonlinear estimation in large models like the ones in this paper is an extremely difficult problem.

In this regard, it should be noted that I used two alternative unit root testing procedures — the augmented Dickey-Fuller (ADF) test [see Dickey and Fuller (1981)] and the non-parametric, $Z(t_{\hat{\alpha}})$ test of Phillips (1987) and Phillips and Perron (1987) — to deal with anomalies that arise when the data are not very informative about whether or not there is a unit root, and found that my input-output ratios, budget shares, and price variables are all integrated of order one [or I(1) in the terminology of Engle and Granger (1987)]. It follows then that for input-output demand equations and share equations to make any sense the variables must be cointegrated in levels; that is, the equation errors must be stationary. However, unit root test results on the residuals of the locally flexible systems — the generalized Leontief, translog, and normalized quadratic models — and

the globally flexible AIM model indicate that they are nonstationary.

If the errors are nonstationary, then there is no theory linking the left hand side to the right hand side variables in equation (2.43) or, equivalently, no evidence for the theoretical models in level form. In such cases, some important nonstationary variables might have been omitted. Allowing for first order serial correlation, as is usually done in the literature, is almost the same as taking first differences of the data if the autocorrelation coefficient is close to unity. In that case, the equation errors become stationary, but there is no theory for the models in first differences. Moreover, as argued by Serletis and Shahmoradi (2007), serial correlation correction increases the number of curvature violations and also leads to induced violations of monotonicity.

It is also to be noted that even if the errors are stationary and the estimates are super consistent, as argued by Attfield (1997) and Ng (1995), standard estimation procedures are inadequate for obtaining correctly estimated standard errors for coefficients in cointegrating equations. In that case, if the equations were all linear, the DOLS method of Stock and Watson (1993) or the FM-OLS method of Phillips (1995) could have been used to obtain correctly estimated standard errors. With my nonlinear models, however, some sort of modification of these procedures is called for, but this is a very difficult issue to deal with.

With the generalized Leontief and translog models failing both economic and econometric regularity and the NQ and AIM(2) models failing econometric regularity, in what follows I report total factor productivity estimates and elasticity estimates based only on the NQ and AIM(2) models.

2.5.3 Total Factor Productivity Trends

Figure 2.1 provides year-by-year total factor productivity estimates with the NQ and AIM(2) models, together with productivity measures formed from the Fisher ideal index

and the smoothed Fisher ideal index. Roughly speaking, the total factor productivity estimates from the NQ and AIM(2) models exhibit similar patterns. First, both of them show a general tendency to rise over the sample period. In particular, the estimates based on the NQ model rise markedly from 0.55% to 1.99% over the sample period whereas those based on the AIM(2) model rise moderately from 0.91% to 1.16%. Second, the two models have produced average total factor productivity measures which are very close to each other. In particular, the average total factor productivity from the NQ model is 1.08%, compared with 1.02% from the AIM(2) model.

To further evaluate the performance of the NQ and the AIM(2) models in capturing technical change, I calculate the productivity growth in U.S. manufacturing as a benchmark, using the Fisher ideal index. I first calculate the Fisher ideal quantity index for the four inputs as

$$g^t = \left[\frac{\sum_{j=1}^{n} p_j^t x_j^t}{\sum_{j=1}^{n} p_j^t x_j^{t-1}} \frac{\sum_{j=1}^{n} p_j^{t-1} x_j^t}{\sum_{j=1}^{n} p_j^{t-1} x_j^{t-1}} \right]^{1/2}$$

and then calculate the quantity index for the single output as $G^t = y^t / y^{t-1}$. The Fisher ideal total factor productivity index is then obtained as

$$\frac{G^t}{g^t} - 1$$

Following Fox (1996), I also obtain a smoothed Fisher ideal total factor productivity index by regressing the raw Fisher ideal series on a constant and a time trend and calculating the fitted values. Both of these indexes are plotted in Figure 2.1. Clearly, both the NQ and AIM(2) measures pass close by the mean of the raw Fisher ideal index series, which is volatile from year to year. Further, both of them evolve in a similar pattern as the smoothed Fisher ideal index which also shows a general tendency to rise from 0.98% to 1.31% over the sample period. In this sense, the productivity growth measures from both the NQ and AIM(2) models can be regarded as smoothed versions of that from the Fisher

ideal index. Generally speaking, both the NQ and AIM(2) models perform pretty well in modelling productivity growth in the U.S. manufacturing industry. However, a close look at Figure 2.1 reveals that the AIM(2) measure resembles the curve of the smoothed Fisher ideal index more closely. Moreover, the AIM(2) model captures (though not noticeably) the slowdown in productivity between 1974 and 1994, which is missed by the NQ model.

An advantage of the AIM(2) model over the NQ model is that total factor productivity can be easily decomposed into growth rates of input-specific efficiencies ($\vartheta's$). In particular, substituting the AIM(2) cost function into (2.6), I can obtain the specific total factor productivity formula for the AIM(2) model

$$TFP_{\text{AIM}} = \sum_{i=1}^{n} s_i \vartheta_i. \tag{2.47}$$

Equation (2.47) shows that total factor productivity estimates based on the AIM(2) model are an input cost-share weighted average of the growth rates of factor efficiencies. As shown in Table 2.4, the long-run growth rate of the efficiency of capital, ϑ_1, is 2.94% per year. For labor, energy, and materials, the long-run growth rates of efficiency are found to be 0.04% (ϑ_2), 5.47% (ϑ_3), and 0.96 % (ϑ_4), respectively. I further define

$$CT_i = \frac{s_i \vartheta_i}{TFP_{\text{AIM}}}$$

to be the contribution of factor i to total factor productivity, and plot in Figure 2.2 the contribution of each factor to total factor productivity. Clearly, capital and materials have been the dominant factors causing productivity growth, with energy having a moderate positive contribution to total factor productivity due to its small input cost share. Labor has a positive and small impact on total factor productivity.

2.5.4 Elasticity Estimates

I begin by presenting the own- and cross-price elasticities in Table 2.5, evaluated at the mean of the data. The signs of all own-price elasticities, η_{ii}, appear reasonable for both models since they are all negative (as predicted by the theory), with the absolute values being less than 1, indicating that the demands for all four inputs are inelastic. However, the AIM(2) model shows larger own-price elasticities in absolute value than the NQ model. In particular, η_{KK} from the AIM(2) is -0.522 which is about twice as large as that from the NQ model. Similarly, η_{LL} (-0.592), η_{EE} (-1.927), and η_{MM} (-0.297) from the AIM(2) model are about 3-10 times as large as their counterparts from the NQ model. This implies that capital, labor, energy, and materials are all more responsive to their own prices according to the AIM(2) than according to the NQ model.

I believe there are actually good reasons to graph the own-price elasticities that I have estimated. Figures 2.3-2.6 present the own-price elasticities for K, L, E, and M for each observation. Clearly, the own-price elasticities are negative at all data points for both models, as predicted by the theory. A prominent difference, as can seen from these figures, is that the own-price elasticities from the AIM(2) model show quite large variations, whereas those from the NQ trend over time. This problem of lacking variations in own-price elasticities over time with the NQ model was first noted by Diewert and Lawrence (2002) and referred to by them as 'the problem of trending elasticities.' As can seen below, this problem in the NQ model is also reflected in its cross elasticities and carried over to its Morishima elasticities. To cure this problem with the NQ model, Diewert and Lawrence (2002) suggested imposing flexibility at two sample points.

As with the own-price elasticities, the cross-price elasticities differ significantly between the two models (see Table 2.5). Moreover, results not presented here, but available upon request, indicate that the cross-price elasticities from the AIM(2) model vary considerably over time whereas those from the NQ are very stable. For example, while the

AIM(2) model shows the capital-labor substitution (η_{KL}) in the range between 0.43 to 0.93, the NQ model show a very stable η_{KL}, varying in a rather small range between 0.20 to 0.24. In addition, the two models show different relations between some of the four inputs. For example, both models classify capital and labor (see η_{KL} and η_{LK}) and energy and materials (see η_{EM} and η_{ME}) as substitutes, but are inconsistent in their classification of capital and materials (see η_{KM} and η_{MK}) and labor and materials (see η_{LM} and η_{ML}).

I now turn to the estimates of the Morishima elasticities of substitution, σ_{ij}^m ($i, j = K, L, E,$ and M), presented in Figures 2.7-2.18. Since the Morishima elasticities of substitution are just a simple function of related own- and cross-price elasticities (see equation (2.10)), the differences in own- and cross-price elasticities between the NQ and AIM(2) models also show up in the Morishima elasticities of substitution. In particular, the Morishima elasticities of substitution from the AIM(2) model vary considerably whereas those from the NQ model are very stable over the sample period. Moreover, the Morishima elasticities of substitution from the AIM(2) model are generally larger than the corresponding ones from the NQ model. Again, I am more interested in the Morishima elasticities of substitution obtained from the AIM(2) model and discuss them in more detail in what follows.

Let's consider first the Morishima elasticity of substitution between K and L, σ_{KL}^m, which represents the percentage change in the capital services to the labor quantity ratio, K/L, when the relative price P_L/P_K is changed by changing P_L and holding P_K constant. Figures 2.7 and 2.10 reveal that at each data point, $\sigma_{KL}^m > \sigma_{LK}^m > 0$, and the average estimated σ_{KL}^m is 1.303, compared with an average estimated σ_{LK}^m of 0.766. Thus, capital services and labor are Morishima substitutes, irrespective of whether the price of labor or the price of capital services changes.

Of particular interest are σ_{KE}^m and σ_{EK}^m in Figures 2.8 and 2.13. The estimates of σ_{KE}^m

are positive, but σ^m_{EK} is positive for most of the sample period (in particular, from 1960 to 1998) and negative from 1999 to 2001. Thus, capital services and energy are always Morishima substitutes when the price of energy changes, but can be either Morishima complements (as from 1960 to 1998) or Morishima substitutes (as for the rest of the sample period) when the price of capital services changes. In other words, an increase in the price of energy (holding the price of capital services constant) always leads to an increase in the K/E ratio, but an increase in the price of capital services (holding the price of energy constant) can lead to either an increase or a decrease in the E/K ratio. I also notice that at each data point between 1960 and 1998, when both σ^m_{KE} and σ^m_{EK} are positive, σ^m_{KE} is greater than σ^m_{EK}, and the average estimated σ^m_{KE} is 1.926 whereas the average estimated σ^m_{EK} is 0.247.

Next I consider σ^m_{LE} and σ^m_{EL} — see Figures 2.11 and 2.14. σ^m_{LE} is positive throughout, but σ^m_{EL} is negative prior to 1977 and positive afterwards. Thus, labor and energy are always Morishima substitutes when energy prices change, but they can be either Morishima complements or substitutes when the price of labor changes. The estimated σ^m_{KM} and σ^m_{MK} are always positive — see Figures 2.9 and 2.16. Thus, capital services and materials are Morishima substitutes irrespective of whether the price of materials changes or the price of capital services changes. The average estimated σ^m_{KM} is 0.221, compared with an average estimated σ^m_{MK} of 0.390. Similarly, the estimates of σ^m_{LM} and σ^m_{ML} are positive throughout (see Figures 2.12 and 2.17), indicating that labor and materials are Morishima substitutes irrespective of whether the price of materials changes or the price of labor changes. The average estimated σ^m_{LM} is 0.720, compared with an average estimated σ^m_{ML} of 0.967. Finally, the estimates of σ^m_{EM} are positive and those of σ^m_{ME} are positive for most of the sample period, but negative after 1994.

2.6 Conclusion

I have investigated productivity issues in the U.S. (total) manufacturing industry, in the context of three popular locally flexible functional forms — the generalized Leontief (GL), translog, and normalized quadratic (NQ) — and one globally flexible functional form — the Asymptotically Ideal Production Model (AIM). In doing so, I have extended the Barnett *et al.* (1991) AIM model, by incorporating (for the first time in the literature) technical change through the factor-augmenting efficiency index approach, proposed by Thomsen (2000).

I estimated the three locally flexible functional forms parametrically and the globally flexible functional form semi-nonparametrically and treated the curvature property as a maintained hypothesis. In particular, I imposed local curvature on the GL and translog models using procedures suggested by Ryan and Wales (2000), I imposed global curvature on the NQ using procedures suggested by Diewert and Wales (1987), and imposed local and global curvature on the AIM(2) model using procedures suggested by Gallant and Golub (1984) and more recently by Serletis and Shahmoradi (2005). I also showed that (with my data set) the imposition of local curvature does not always assure theoretical regularity, because of curvature violations at other points within the region of the data. I believe that this is a typical result in the literature that uses locally flexible functional forms and alert researchers to the kinds of problems that arise when all three theoretical regularity conditions are not satisfied — see also Barnett (2002) and Barnett and Pasupathy (2003).

I provided a comparison between the NQ and AIM cost functions, the only two models that satisfy all three theoretical regularity conditions. I found that the AIM(2) cost function with technical change introduced through the factor-augmenting efficiency index approach performs better than traditional locally flexible function forms and gives

more accurate estimates of total factor productivity. I also found that the elasticities from the AIM(2) model are generally larger and show more variation than those from the NQ model, which is consistent with Gallant and Golub (1984) who employed a different globally flexible functional form — the Fourier. Finally, I discussed the elasticities based on the AIM(2) model to shed some new light on the substitutability/complementarity relationship between capital, labor, energy, and materials.

Although I have achieved economic regularity (in terms of curvature, positivity, and monotonicity) with the NQ and AIM(2) models, I have not achieved econometric regularity (in terms of stationary equation errors), which makes interpreting my results difficult. Moreover, my econometric modeling assumes a serially uncorrelated Gaussian measurement error, or equivalently serially independent measurement error. To simultaneously achieve both economic and econometric regularity seems to be a challenging task and an area for potentially productive future research. It could also be that the econometric irregularity is caused by my treatment of technical change. That is, I treated technical change as being smooth over the sample period, but there are fairly large year to year fluctuations in technical change as well as secular trends in total factor productivity growth. Using the spline techniques pioneered by Diewert and Wales (1993) and Fox (1996) to model these trends in the context of the production models used in this paper is work that I am currently undertaking.

Finally, the Bayesian approach, pioneered by Terrell (1996) and Griffiths *et al.* (2000) in imposing regularity on linear factor demand systems, could also be directly used to estimate the first three locally flexible cost models presented in this paper, since all these three models are linear. For my AIM model with technical change, the application of Bayesian inference is more complicated due to its highly nonlinear nature and also the difficulty in finding reasonable informative priors. The Griffiths and Chotikapanich (1997) method can be used in this case after some appropriate modifications. The Bayesian

approach has two major advantages that traditional econometric methods commonly used for productivity estimation do not possess. First, the Bayesian approach provides exact (small-sample) inference on the productivity components (i.e. technical change, efficiency change, and returns to scale) which in many cases are nonlinear functions of estimated parameters, whereas the traditional methods provide only point estimates of these productivity components without statistical inference. Second, and even more importantly, the Bayesian approach allows us to incorporate the theoretical regularity restrictions of neoclassical microeconomic theory in the estimation. This can be done either by using the accept-reject algorithm — see Terrel (1996) — or the Metropolis-Hastings algorithm — see Griffiths *et al.* (2000).

TABLE 2.1

GENERALIZED LEONTIEF PARAMETER ESTIMATES

Parameter	Unrestricted	Local curvature imposed
β_{11}	.0966 (.000)	.0884 (.000)
β_{12}	.0438 (.017)	.0844 (.000)
β_{13}	−.0082 (.024)	−.0101 (.000)
β_{14}	.0633 (.000)	.0507 (.000)
β_{22}	.5108 (.000)	.3130 (.000)
β_{23}	.0006 (.925)	.0163 (.023)
β_{24}	−.1371 (.000)	−.0386 (.000)
β_{33}	.0650 (.054)	.0104 (.007)
β_{34}	.0264 (.001)	.0123 (.059)
β_{44}	.4033 (.000)	.3044 (.000)
β_{1t}	.0003 (.274)	.0001 (.045)
β_{2t}	−.0073 (.000)	−.0052 (.000)
β_{3t}	−.0001 (.430)	.0001 (.005)
β_{4t}	.0026 (.000)	.0011 (.000)
Positivity violations	0	0
Monotonicity violations	0	0
Curvature violations	49	2

Notes: Sample period, annual data 1953-2001 ($T = 49$).

TABLE 2.2

TRANSLOG PARAMETER ESTIMATES

Parameter	Unrestricted	Local curvature imposed
β_0	−.0512 (.000)	1.3182 (.000)
β_1	.1635 (.000)	.2546 (.000)
β_2	.4580 (.000)	.2568 (.000)
β_3	.0268 (.000)	.2265 (.000)
β_{11}	.1055 (.000)	.1176 (.000)
β_{12}	−.0553 (.000)	−.0747 (.000)
β_{13}	−.0097 (.000)	−.0113 (.216)
β_{22}	.2467 (.000)	.1810 (.000)
β_{23}	−.0175 (.000)	−.0575 (.000)
β_{33}	.0171 (.000)	.0886 (.000)
β_{1t}	.0024 (.000)	.0031 (.000)
β_{2t}	−.0071 (.000)	−.0050 (.000)
β_{3t}	.0003 (.015)	.0009 (.001)
β_t	−.0026 (.006)	−.0075 (.000)
β_{tt}	−.0001 (.002)	−.00002 (.550)
Positivity violations	0	0
Monotonicity violations	0	0
Curvature violations	49	6

Notes: Sample period, annual data 1953-2001 ($T = 49$).

TABLE 2.3

NQ PARAMETER ESTIMATES

Parameter	Unrestricted	Global curvature imposed
β_1	.1657 (.000)	.1677 (.000)
β_2	.4179 (.000)	.4251 (.000)
β_3	.0257 (.000)	.0254 (.000)
β_4	.3542 (.000)	.3556 (.000)
β_{12}	.0287 (.000)	.0331 (.003)
β_{13}	−.0057 (.005)	−.0059 (.015)
β_{14}	.0409 (.000)	.0324 (.000)
β_{23}	.0006 (.861)	.0059 (.126)
β_{24}	−.0868 (.000)	−.0197 (.199)
β_{34}	.0136 (.000)	.0100 (.002)
β_{1t}	.0006 (.074)	.0001 (.686)
β_{2t}	−.0078 (.000)	−.0062 (.000)
β_{3t}	−.0001 (.391)	−.0002 (.039)
β_{4t}	.0033 (.000)	.0010 (.082)
Positivity violations	0	0
Monotonicity violations	0	0
Curvature violations	49	0

Notes: Sample period, annual data 1953-2001 ($T = 49$).

TABLE 2.4. AIM(2) PARAMETER ESTIMATES

Parameter	Unconstrained estimates	Curvature constrained at 1981	Curvature constrained fully
ϑ_1	0.0072	0.0071	0.0294
ϑ_2	0.0154	0.0154	0.0004
ϑ_3	0.0079	0.0079	0.0547
ϑ_4	0.0048	0.0049	0.0096
α_1	36.5369	36.5660	−0.6979
α_2	48.0633	48.4436	−0.9681
α_3	−7.5918	−7.6265	−0.1435
α_4	57.7529	57.6138	−6.7273
α_5	132.6923	133.0500	11.1911
α_6	12.8664	12.7332	4.1975
α_7	226.2107	225.7530	19.8209
α_8	224.7218	224.7091	−9.3343
α_9	−98.4541	−98.3085	−27.8608
α_{10}	−99.6927	−99.4859	−25.8908
α_{11}	−73.6204	−73.8018	−1.6601
α_{12}	−148.9083	−149.3250	−1.6112
α_{13}	46.2347	46.1778	−1.5754
α_{14}	−3.6166	−3.6960	1.9261
α_{15}	−108.7704	−108.6506	6.1263
α_{16}	−134.5830	−134.1460	−47.3199
α_{17}	−158.5638	−159.1355	−3.1487
α_{18}	−51.6272	−51.4163	−5.8699
α_{19}	96.5016	95.9014	9.2258
α_{20}	−45.5792	−45.6127	28.8435
α_{21}	90.5297	90.5274	6.6752
α_{22}	−51.3034	−51.1016	30.2694
α_{23}	−46.7539	−46.9368	13.7951
α_{24}	182.7781	183.1389	−1.6796
α_{25}	−133.1851	−133.0411	4.7117
α_{26}	−21.2014	−21.2212	−33.1634
α_{27}	38.3530	38.5663	−10.8059
α_{28}	−15.7660	−15.6267	53.9644
α_{29}	−113.1907	−112.5742	−19.8109
α_{30}	127.3465	127.6929	−17.4536
α_{31}	−9.9053	−10.5092	49.3125
α_{32}	−132.7490	−131.3271	31.9598
α_{33}	−191.5419	−192.2739	36.1132
α_{34}	347.8034	347.3930	−63.0859
α_{35}	−20.7763	−21.4390	−28.3207
$S(\widehat{\theta})$	0.0070	0.0071	0.0103
Positivity violations	0	0	0
Monotonicity violations	0	0	0
Curvature violations	33	32	0

Note: Sample period, annual data 1953-2001 ($T = 49$).

TABLE 2.5

PRICE ELASTICITIES AT THE MEAN

Factor i	Model	Price elasticities			
		η_{iK}	η_{iL}	η_{iE}	η_{iM}
(K)	NQ	−.267	.216	−.040	.092
	AIM(2)	−.522	.804	.032	−.314
(L)	NQ	.080	−.071	.026	−.035
	AIM(2)	.314	−.592	−.011	.289
(E)	NQ	−.239	.409	−.547	.376
	AIM(2)	.264	−.249	−1.927	1.912
(M)	NQ	.042	−.044	.030	−.028
	AIM(2)	−.142	.335	.104	−.297

Note: Sample period, annual data 1953-2001 ($T = 49$).

Figure 2.1. Total Factor Productivity Estimates

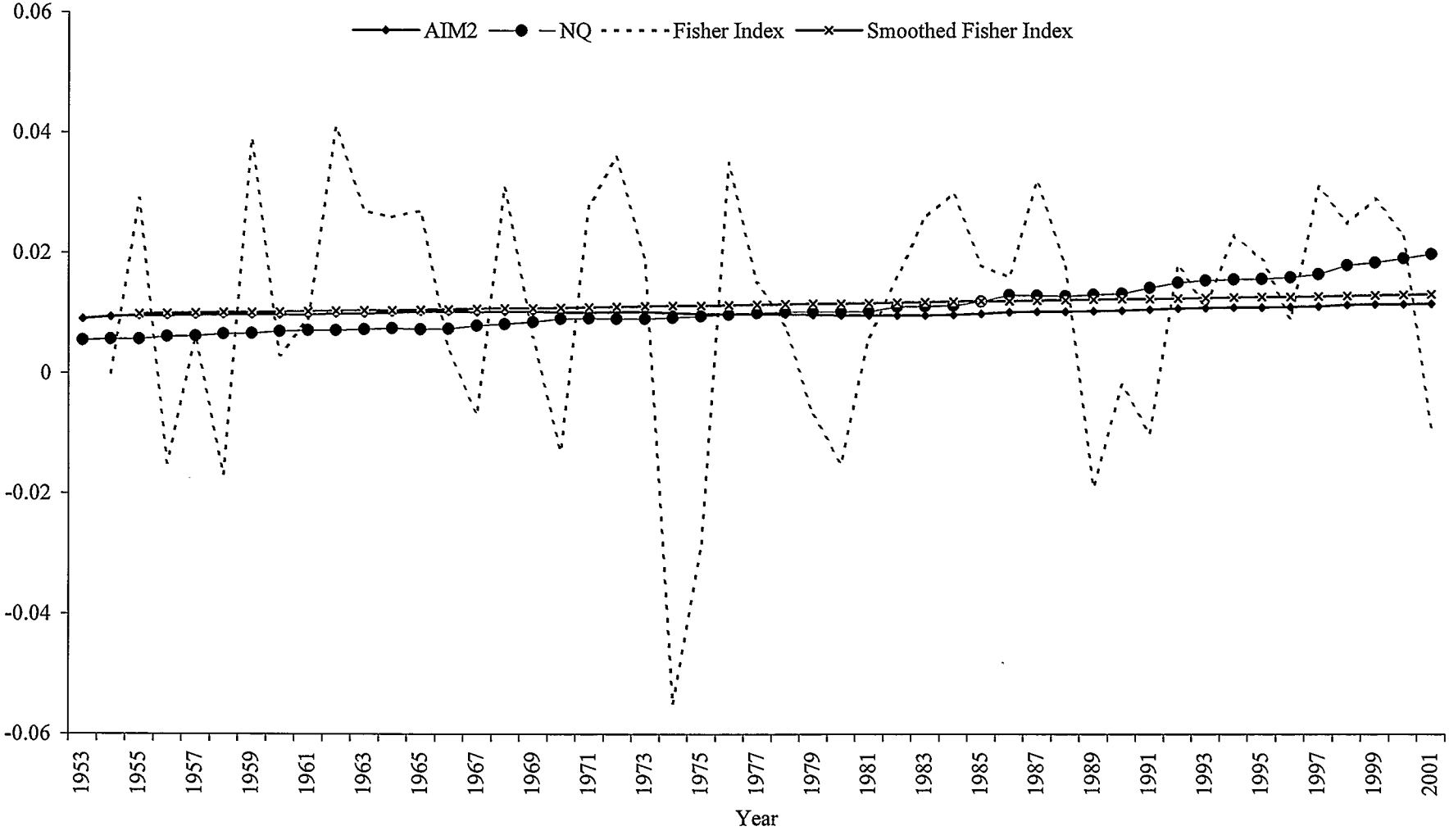


Figure 2.2. Factor Contributions to Total Factor Productivity

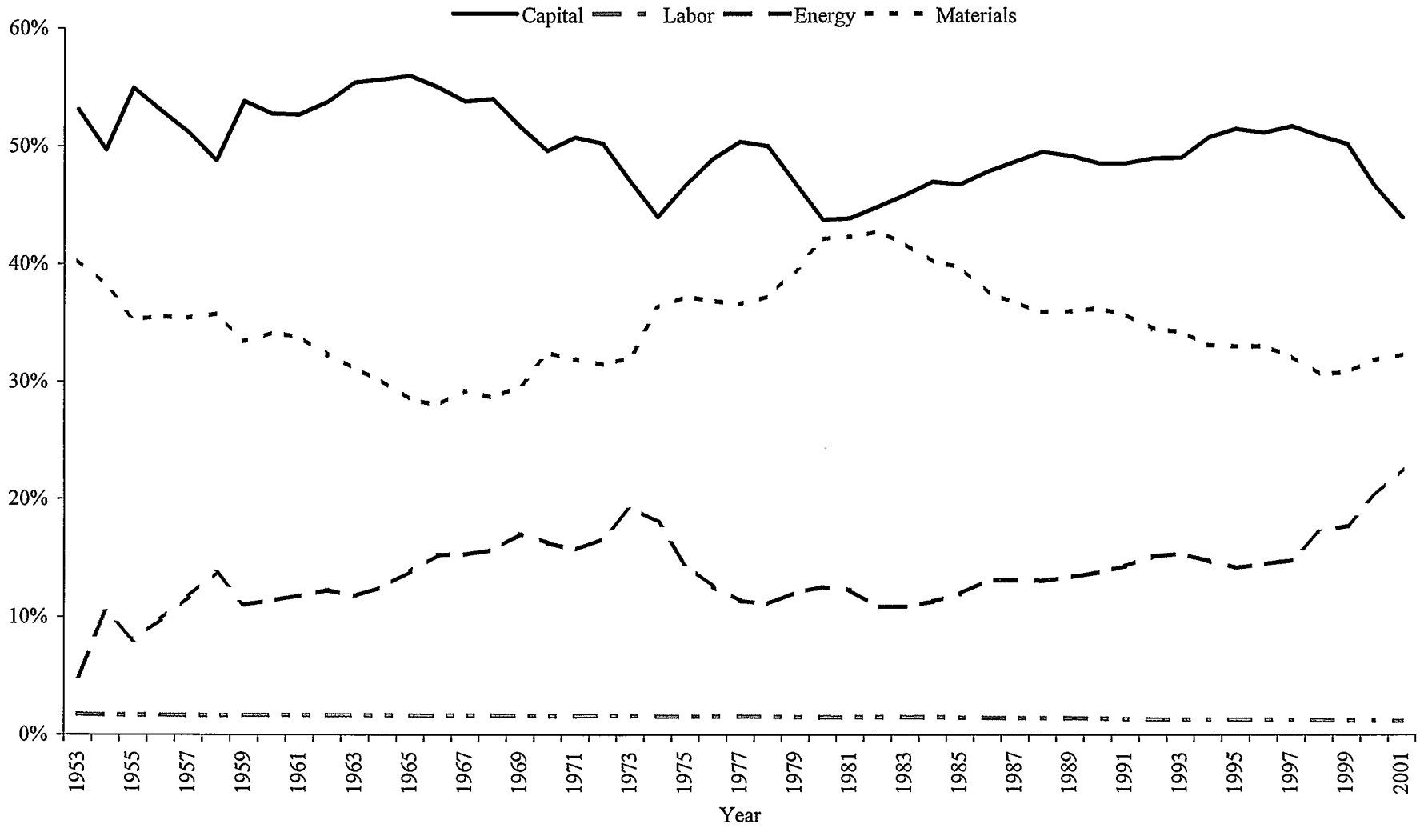


Figure 2.3. Own Price Elasticities for Capital

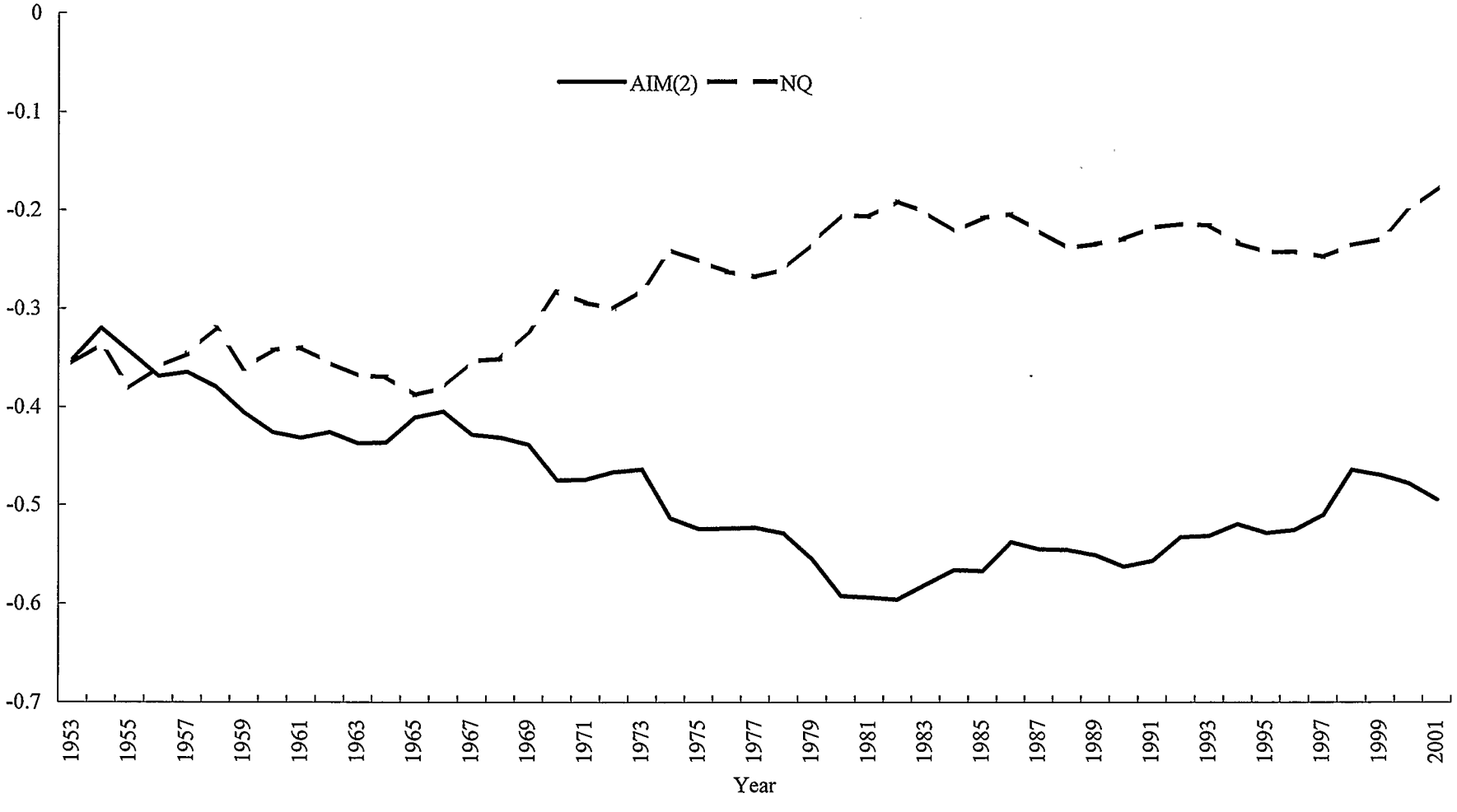


Figure 2.4. Own Price Elasticties for Labor

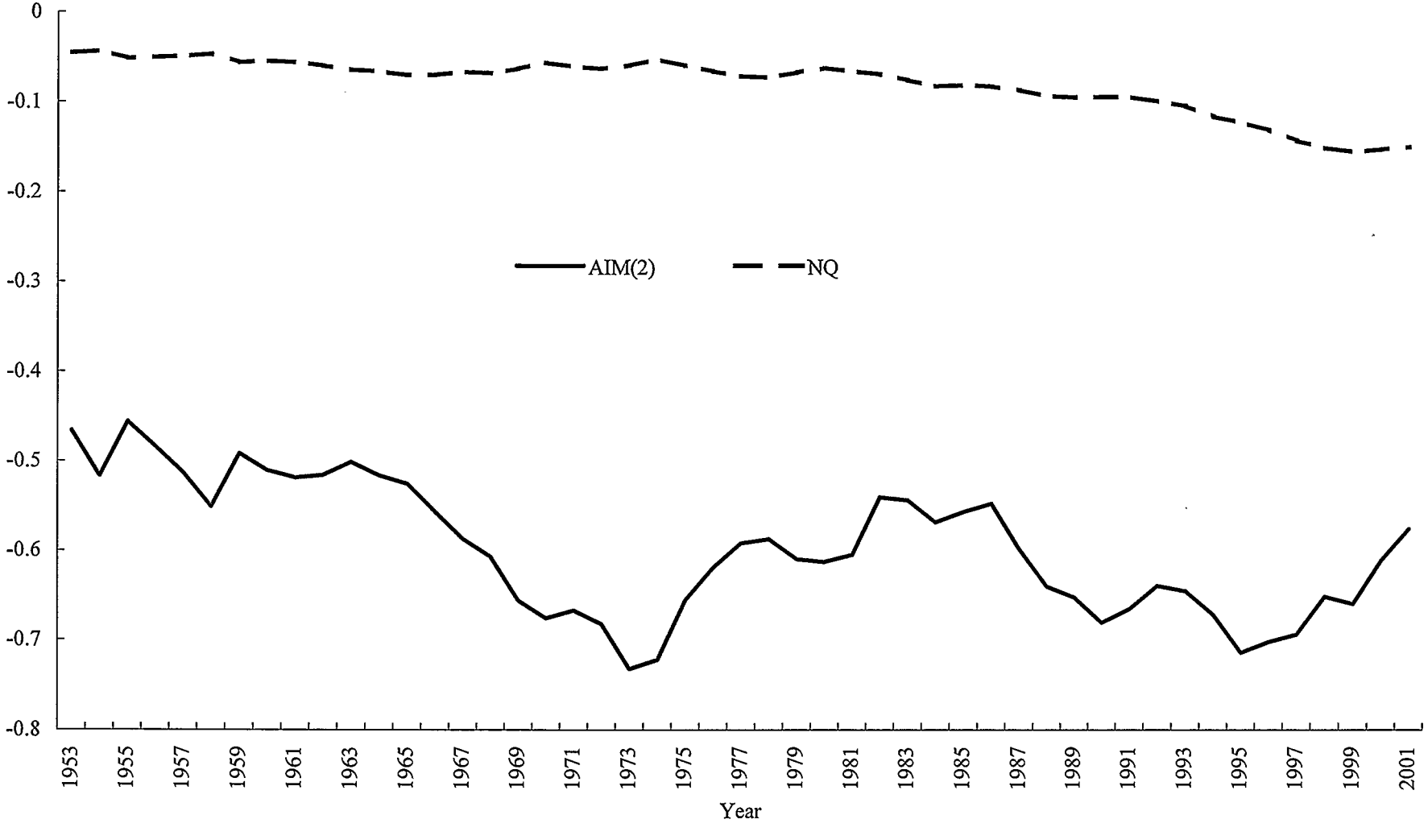


Figure 2.5. Own Price Elasticities for Energy

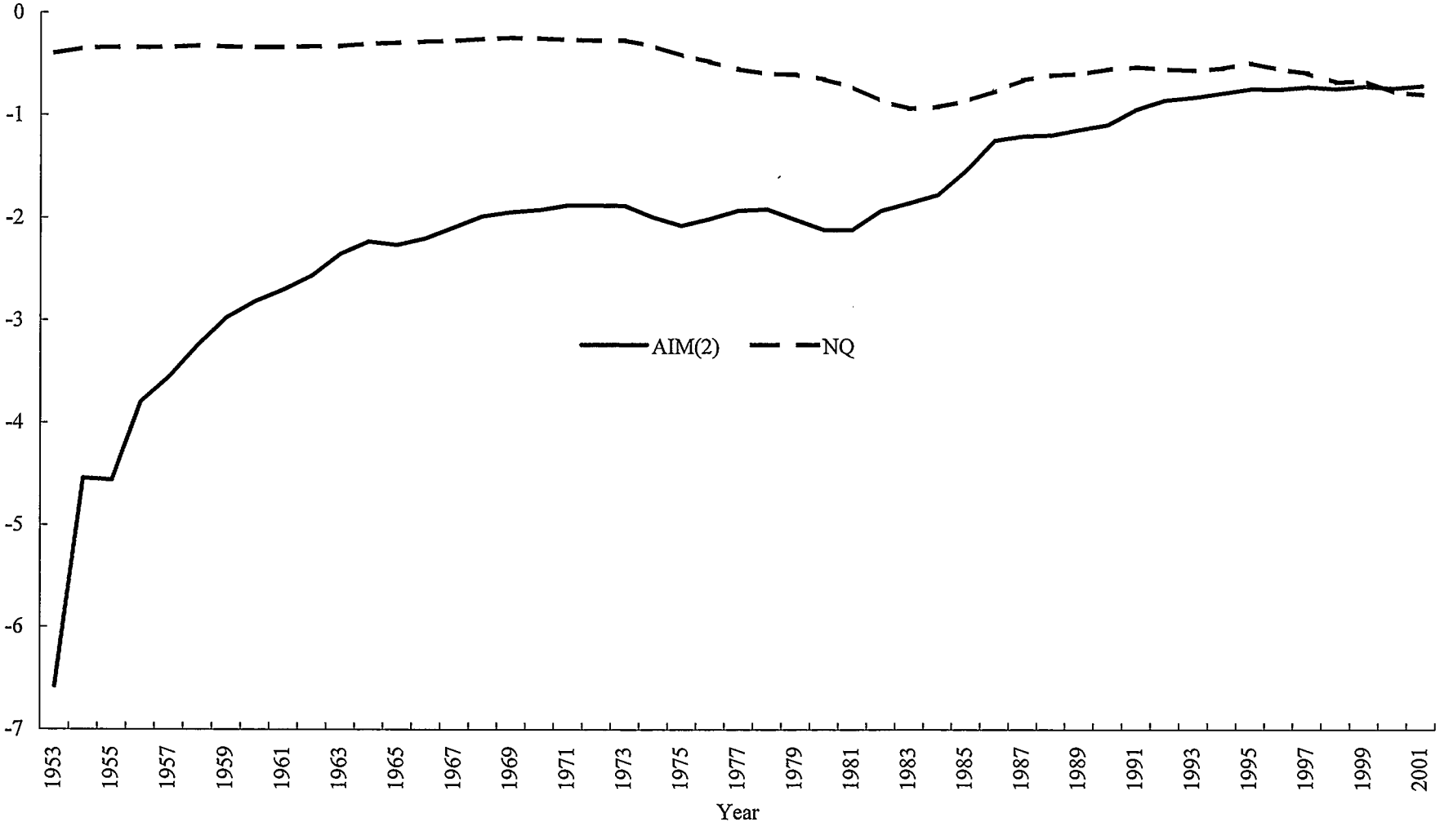


Figure 2.6. Own Price Elasticities for Materials

Figure 2.7. Morishima Elasticities of Substitution between K and L with the Price of L Changing

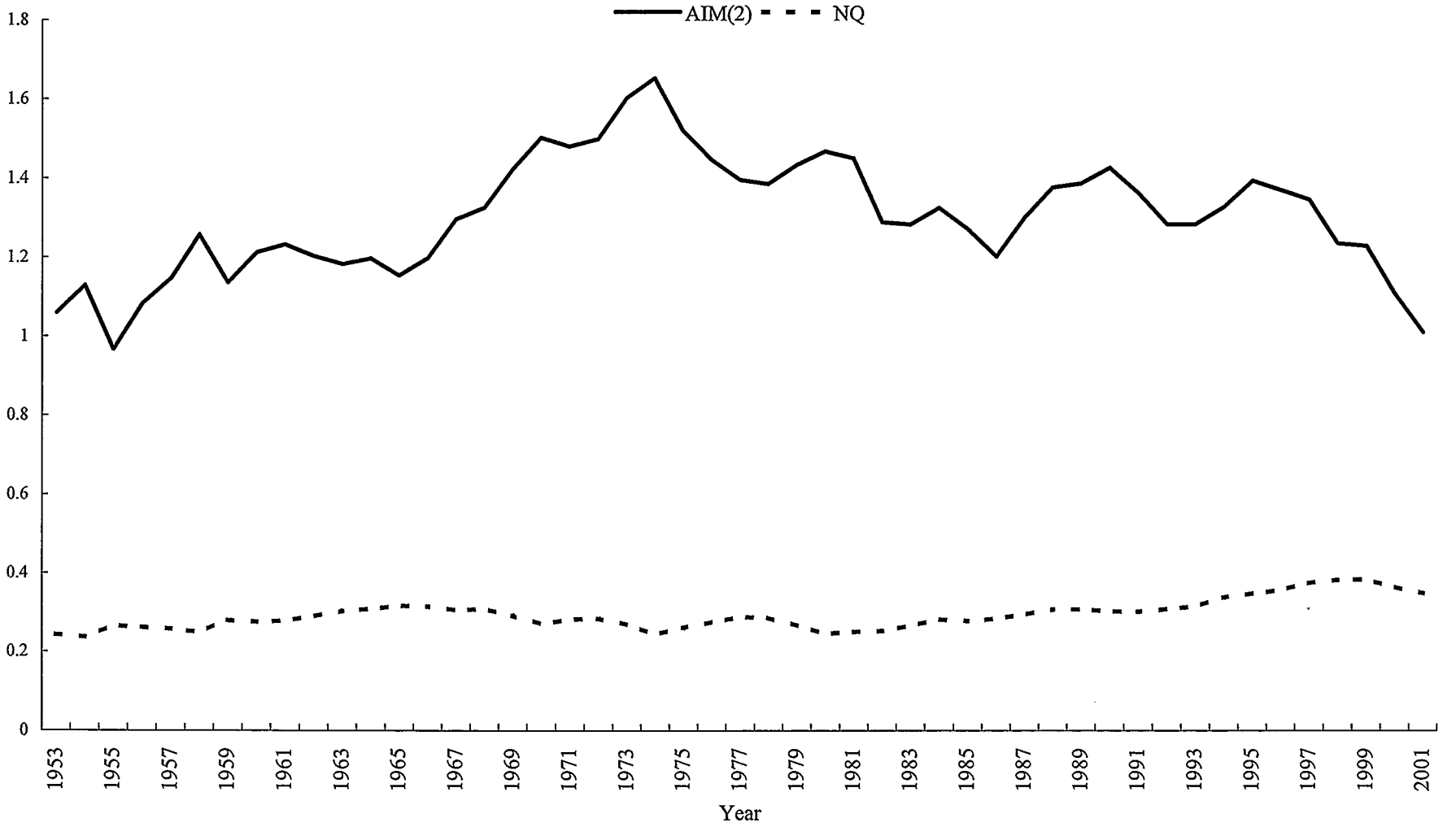


Figure 2.8. Morishima Elasticities of Substitution between K and E with the Price of E Changing

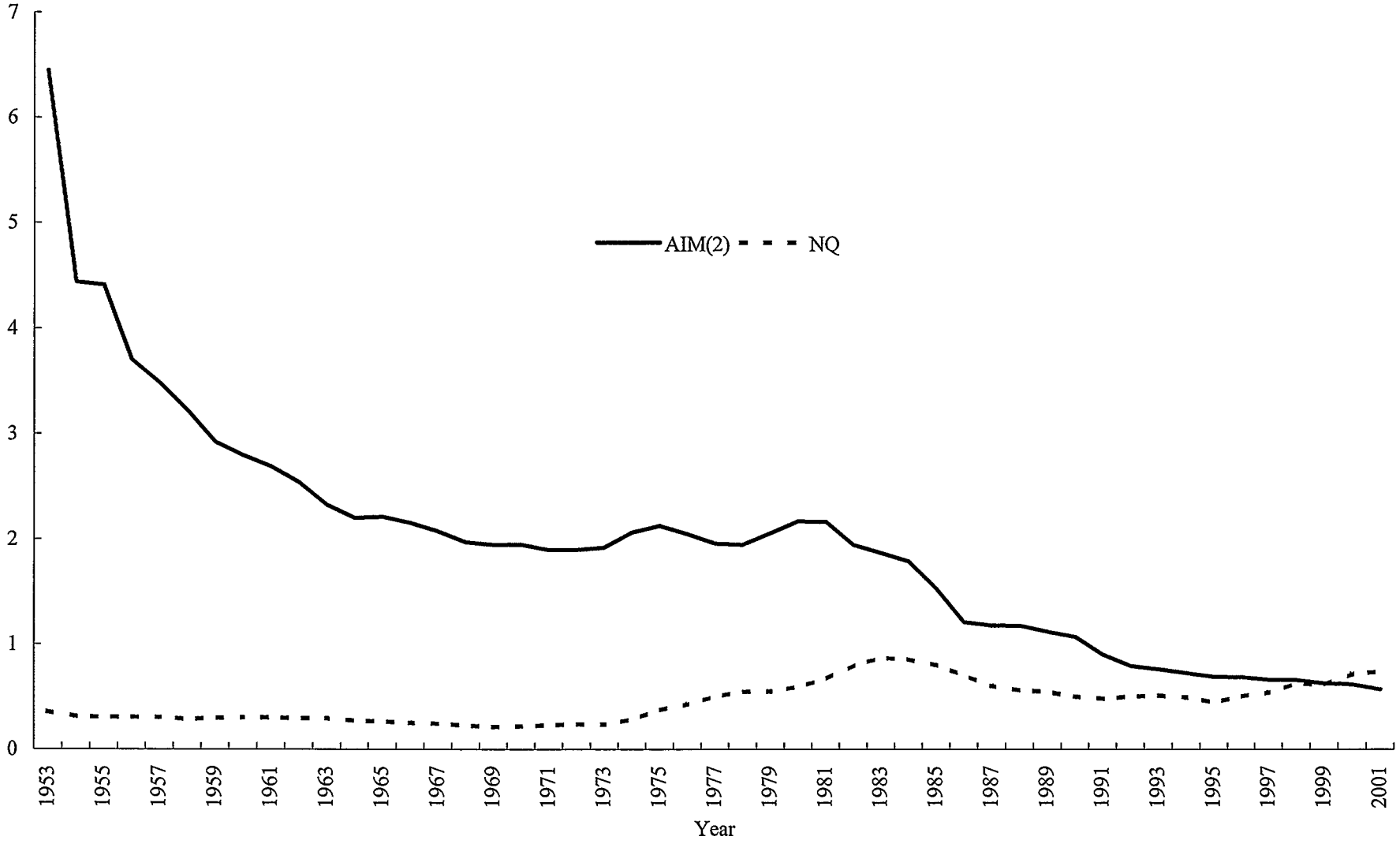


Figure 2.9. Morishima Elasticities of Substitution between K and M with the Price of M Changing

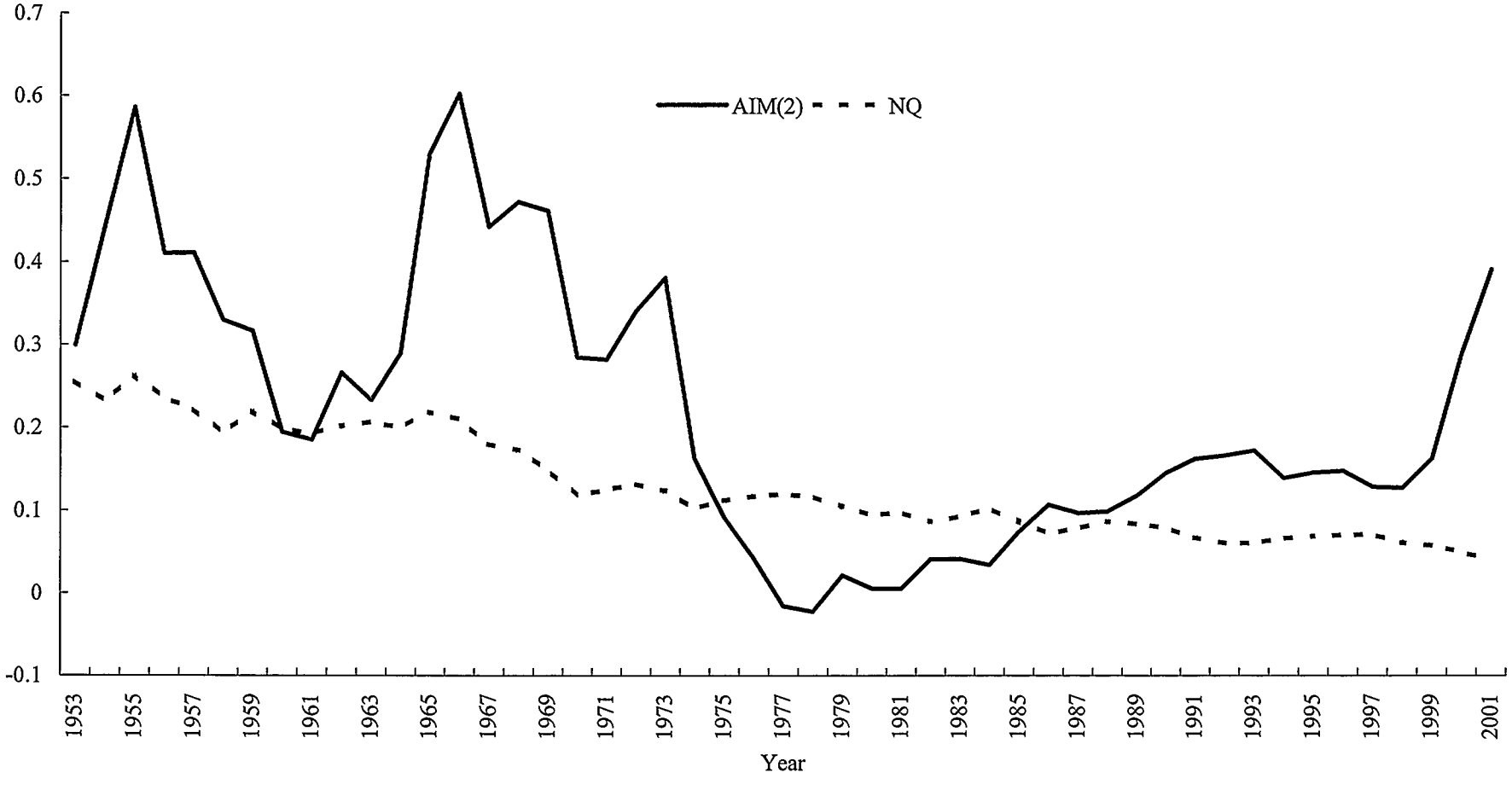


Figure 2.10. Morishima Elasticities of Substitution between L and K with the Price of K Changing

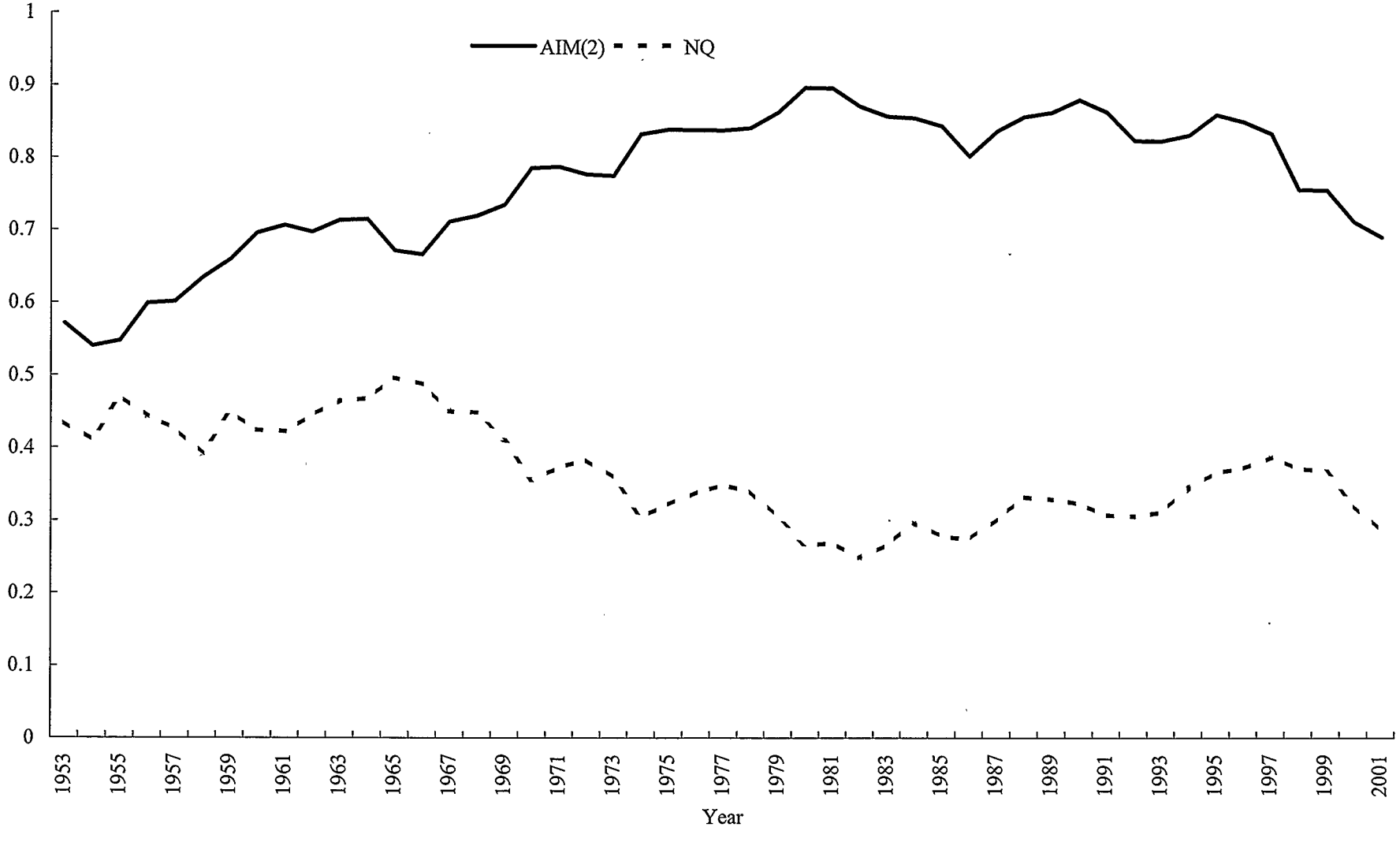


Figure 2.11. Morishima Elasticities of Substitution between L and E with the Price of E Changing

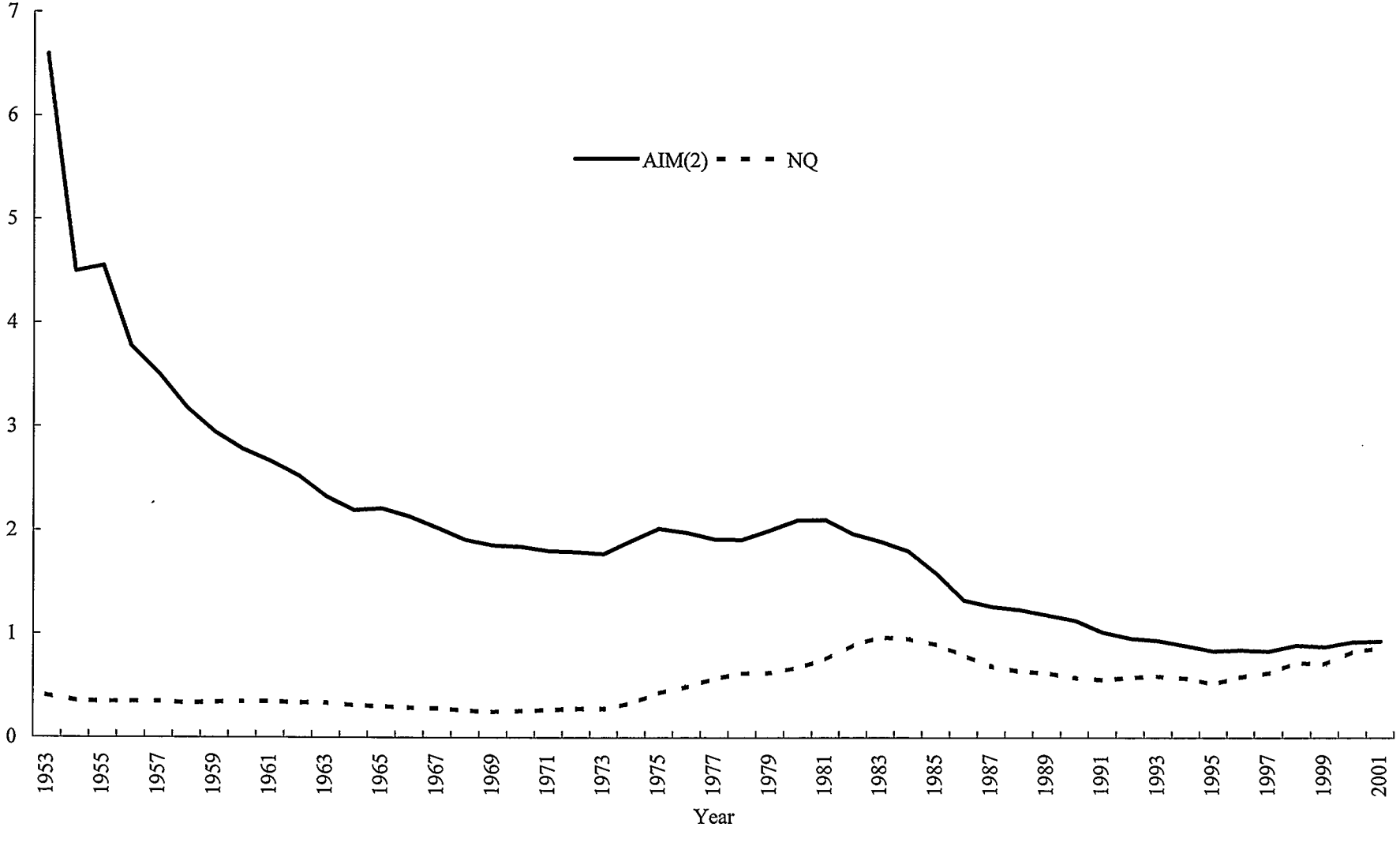


Figure 2.12. Morishima Elasticities of Substitution between L and M with the Price of M Changing

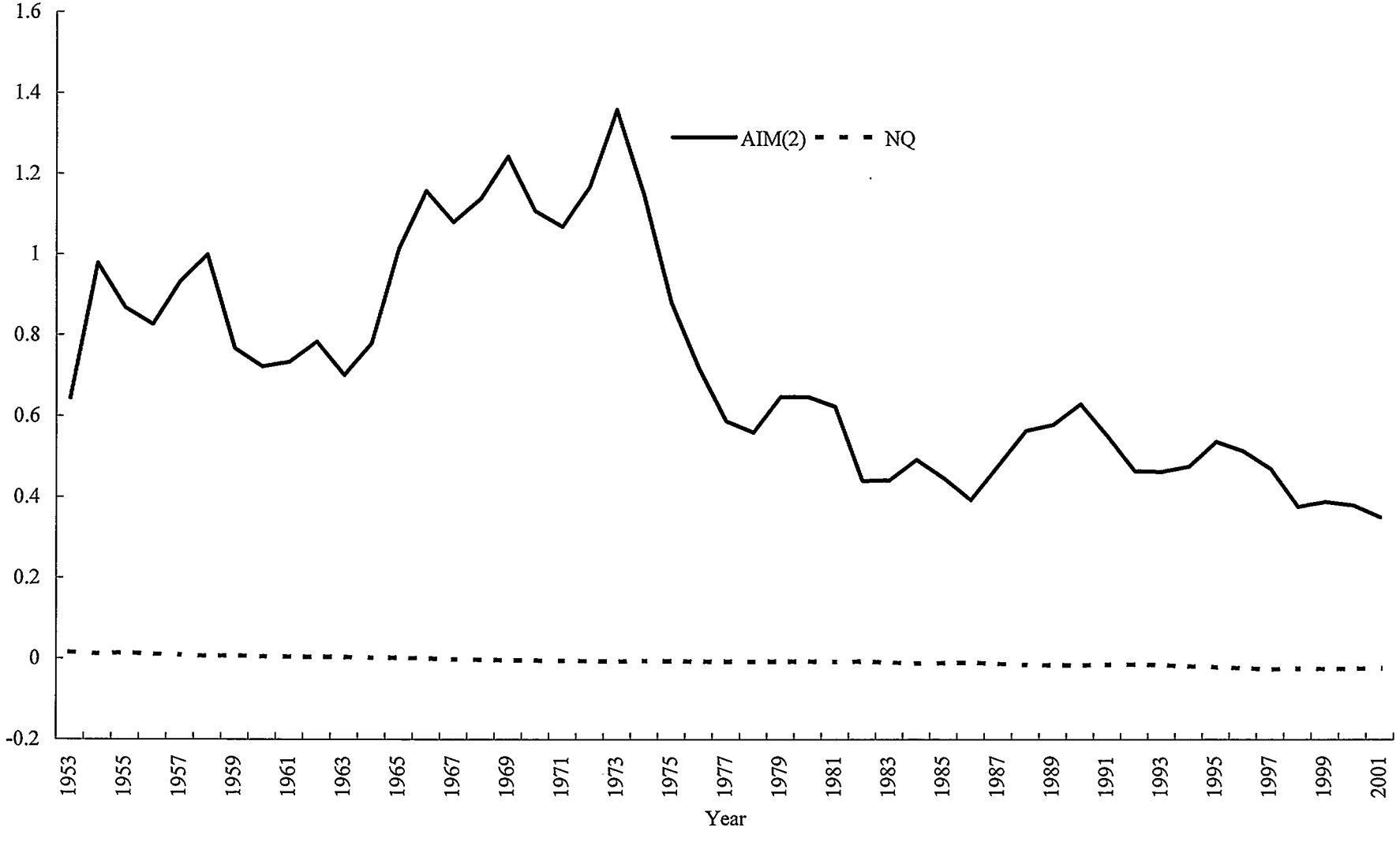


Figure 2.13. Morishima Elasticities of Substitution between E and K with the Price of K Changing

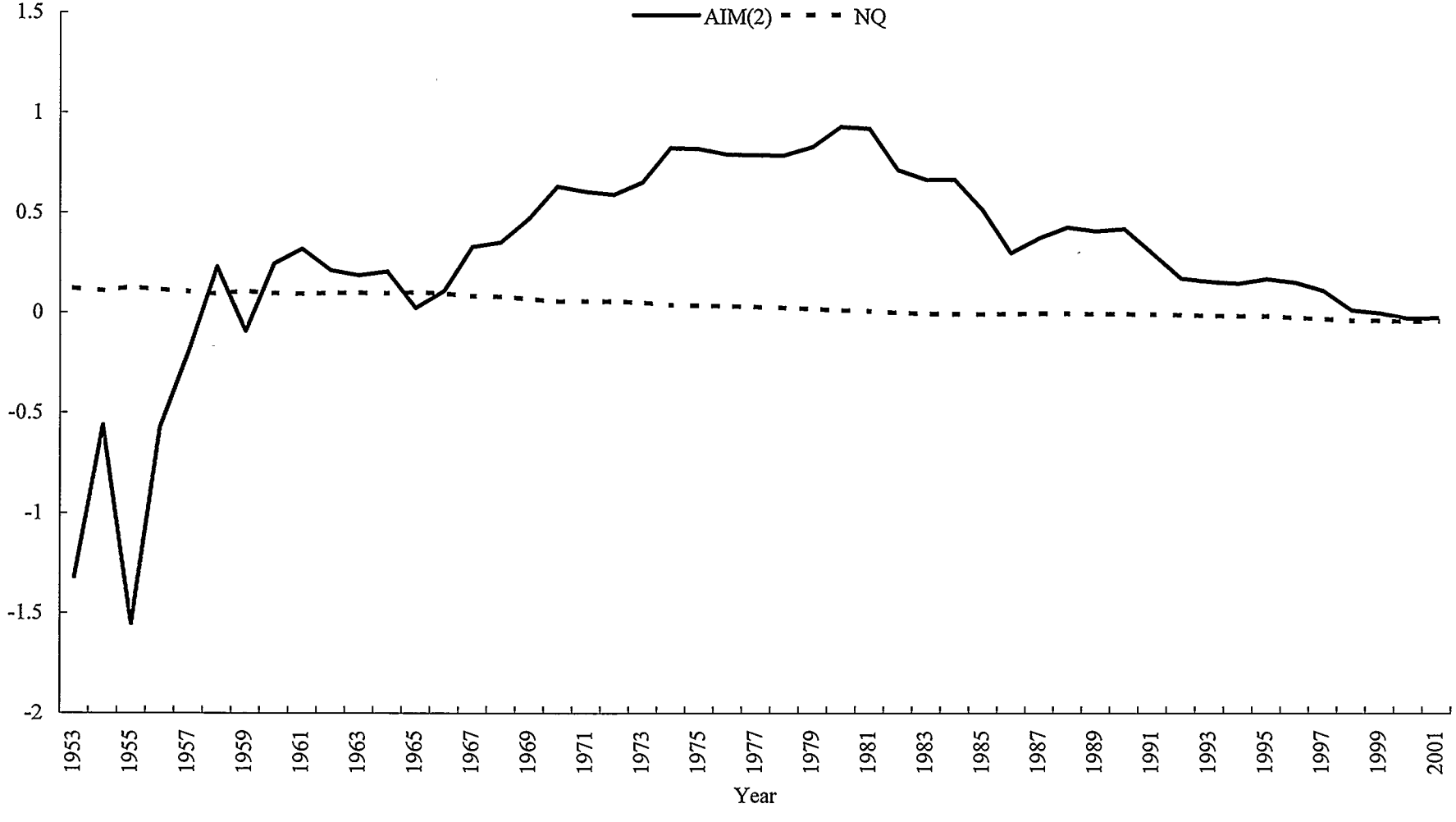


Figure 2.14. Morishima Elasticities of Substitution between E and L with the Price of L Changing

Figure 2.15. Morishima Elasticities of Substitution between E and M with the Price of M Changing

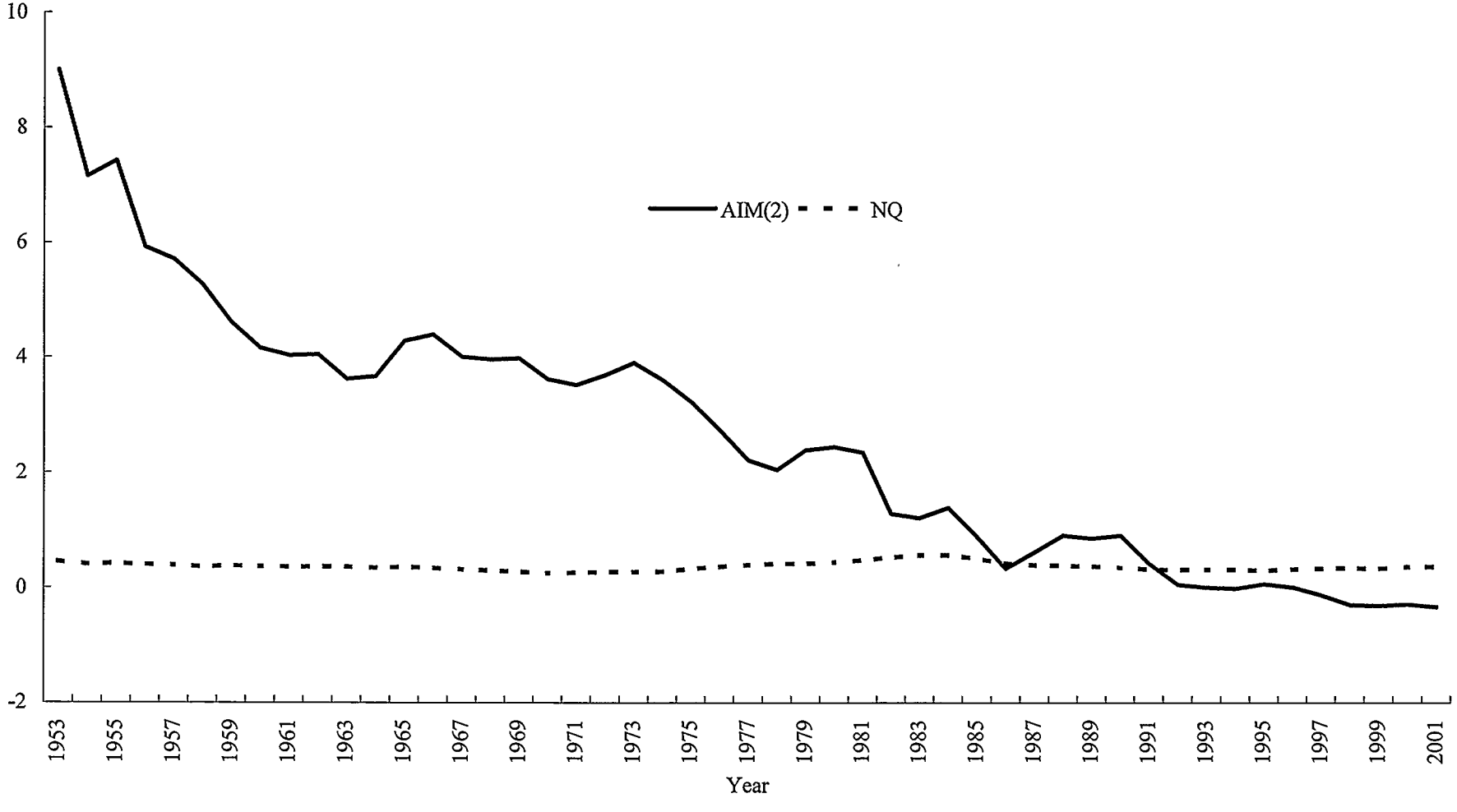


Figure 2.16. Morishima Elasticities of Substitution between M and K with the Price of K Changing

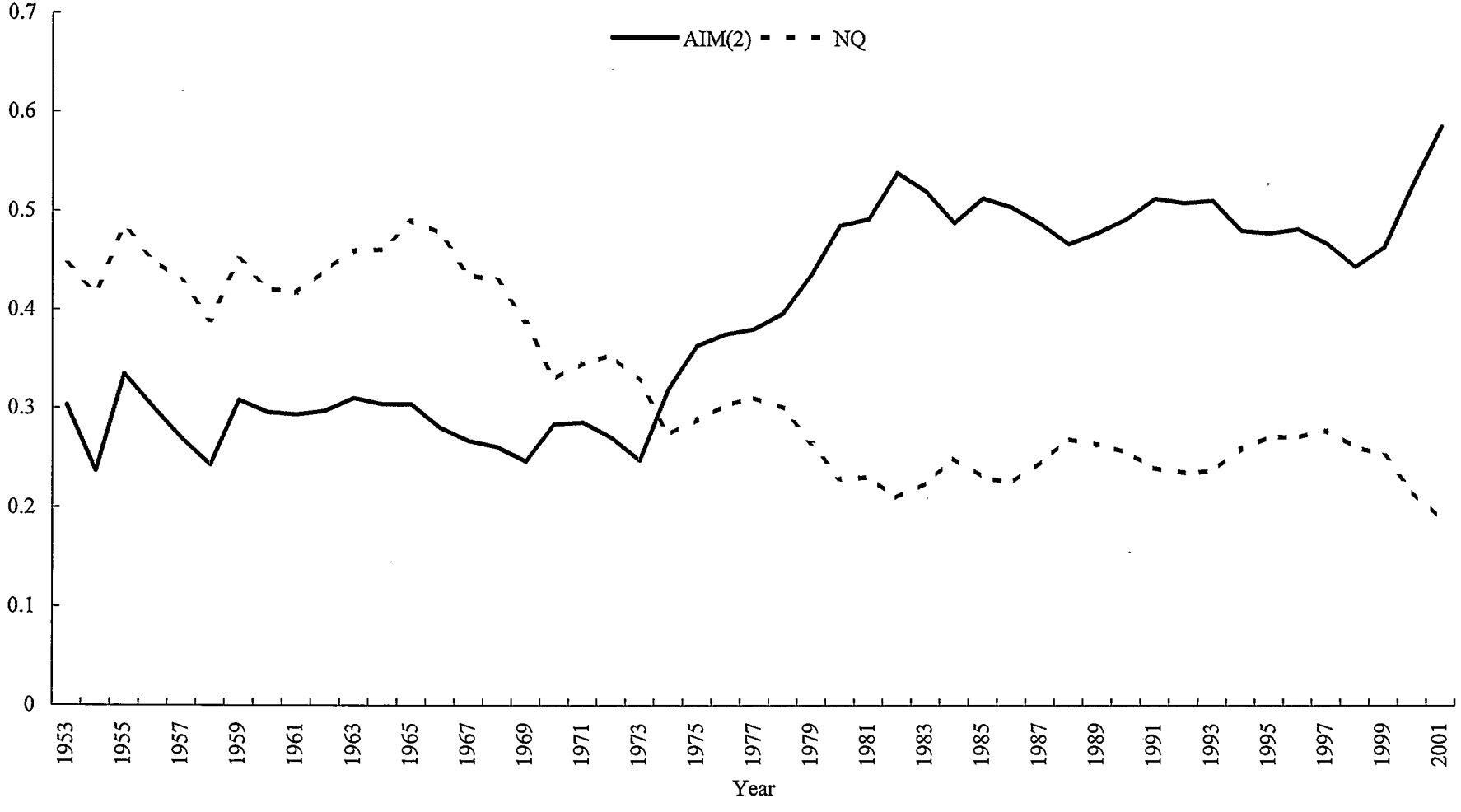


Figure 2.17. Morishima Elasticities of Substitution between M and L with the Price of L Changing

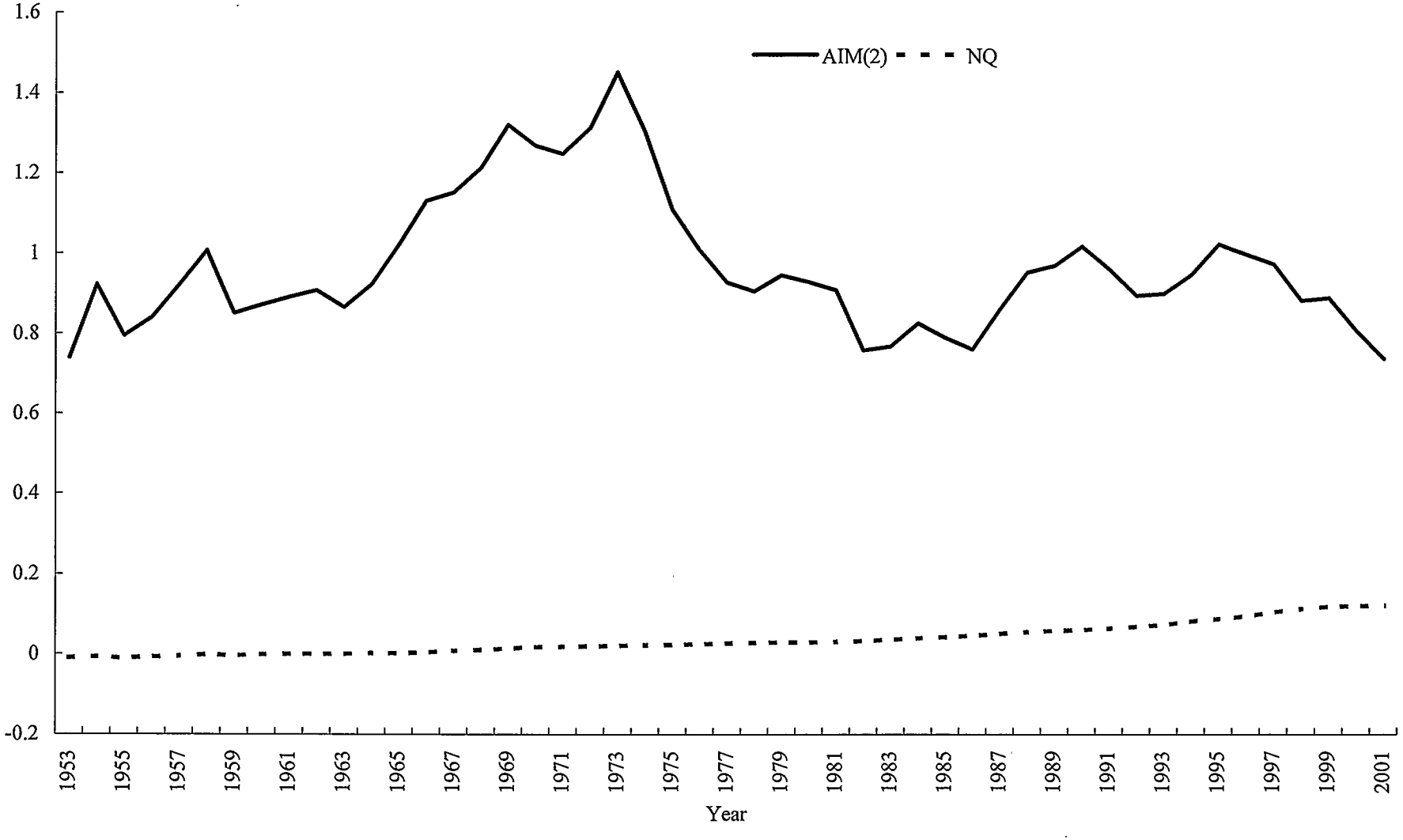
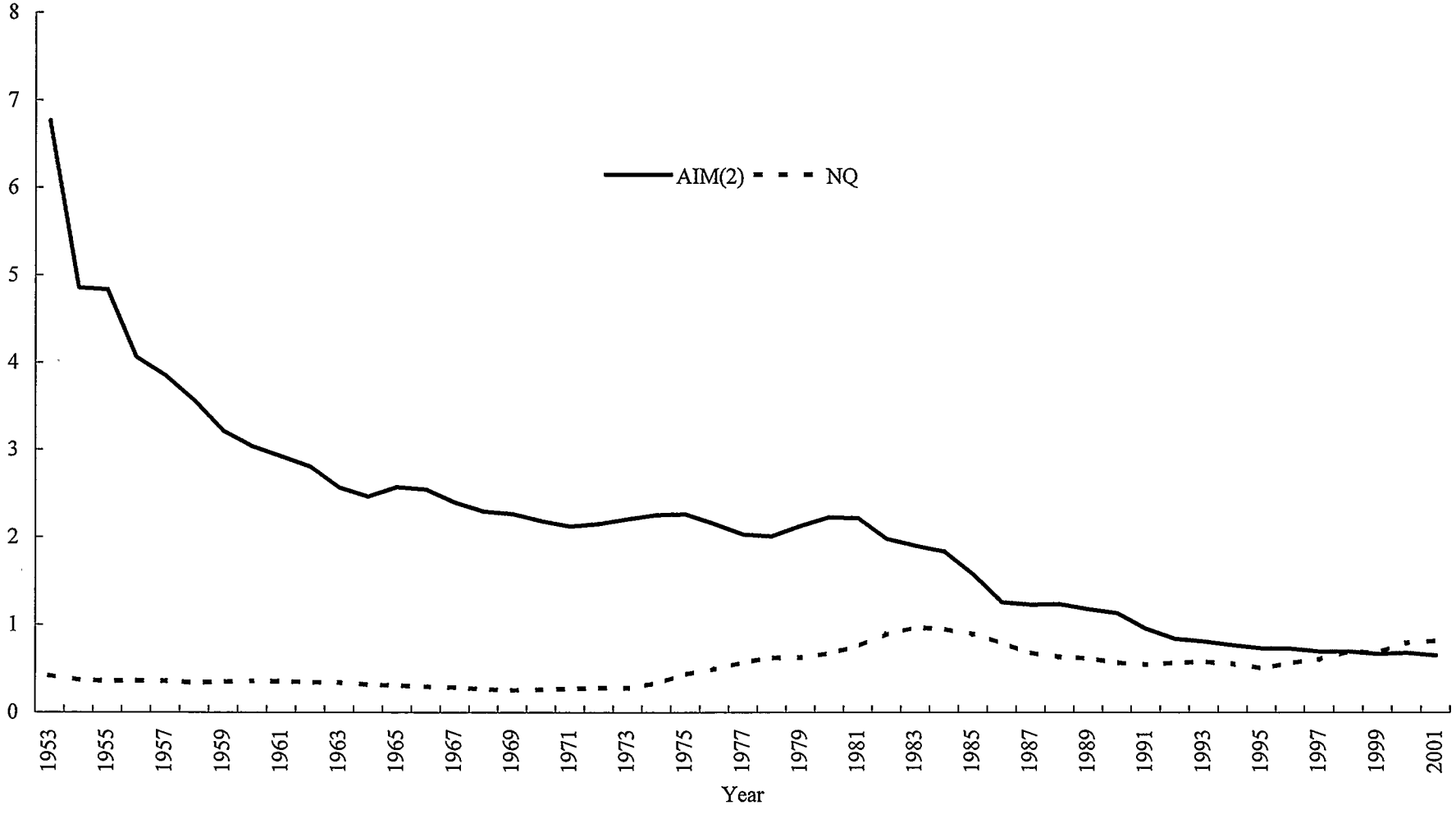


Figure 2.18. Morishima Elasticities of Substitution between M and E with the Price of E Changing



CHAPTER THREE

EFFICIENCY AND PRODUCTIVITY OF THE U.S. BANKING INDUSTRY, 1998-2005: EVIDENCE FROM THE FOURIER COST FUNCTION SATISFYING FULL REGULARITY CONDITIONS

3.1 Introduction

In the last twenty five years (from 1980 to 2005), the banking industry in the United States has been greatly transformed by numerous regulatory changes — see, for example, Lown *et al.* (2000), Kroszner and Strahan (2000), and Montgomery (2003) for a detailed list of regulatory changes. These changes, and particularly those related to the permission of interstate branching and combinations of banks, securities firms, and insurance companies, stimulated the decade-long consolidation in the industry characterized by the dramatic rise in merger and acquisition activities, the rapid decline in the number of commercial banks and the increasing concentration of industry assets among the very large banks (see Jones and Critchfield, 2005). On the other hand, various innovations in technology and applied finance were widespread and intensively adopted by the U.S. banking industry. These technological and financial innovations include, but not limited to, information processing and telecommunication technologies, the securitization and sale of bank loans, and the development of derivatives markets. The widespread and intensive use of information technologies and financial innovation has facilitated the rapid transfer of information at low cost, increased the scope and volume of non-traditional activities, and also helped facilitate consolidation of the industry (see Berger *et al.*, 1995; Berger, 2004).

The question of whether the unprecedented transformation has made the U.S. banking industry more efficient has stimulated a substantial body of efficiency studies — see, for example, surveys in Berger and Humphrey (1997) and Berger *et al.* (1999). One dimension of banking efficiency that attracted a lot of research interest (especially in studies prior to the 1990's) is scale efficiency and scope efficiency. The former is used to measure whether a banking firm is producing at optimal output levels; and the latter is used to measure whether it is producing at an optimal combination of outputs. The

other dimension of banking efficiency that has received increasing attention since the early 1990's is X-efficiency. X-efficiency is called 'frontier efficiency' in Bauer *et al.* (1998) and 'economic efficiency' in Kumbhakar and Lovell (2003). The interested reader is referred to Kumbhakar and Lovell (2003) for an excellent discussion of the relationship between different concepts of efficiency.

X-efficiency is a combination of technical efficiency and allocative efficiency, with the former referring to the ability of a firm to produce output from a given set of inputs and the latter referring to the extent to which a firm uses the inputs in the best proportions, given their prices. X-efficiency is most commonly measured by determining an industry's best-practice frontier and comparing how far each firm deviates from this frontier. However, previous studies revealed that X-inefficiency outweigh scale and scope inefficiencies by a considerable margin, and thus, as Bauer *et al.* (1998, p. 86) put it, "have a strong empirical association with higher probabilities of financial institution failures." According to Berger and Humphrey (1991), cost inefficiency consumes 25 percent or more of total costs, whereas scale inefficiency and allocative inefficiency consume only 5% or less. Therefore, in recent years, the research on the efficiency of the U.S. banking industry has increasingly focused on X-efficiency.

The literature investigating X-efficiency in the U.S. banking industry has been dominated by two methodologies: nonparametric Data Envelopment Analysis (DEA for short) and the parametric Stochastic Frontier Analysis (SFA for short). Two other less commonly used parametric approaches are the Thick Frontier Analysis (TFA for short, see Berger and Humphrey, 1991) and the Distribution Free Approach (DFA for short, see Berger, 1993). First put forward by Charnes *et al.* (1978), the DEA approach is a linear programming technique where the efficient frontier is formed as the piecewise linear combination that connects the set of best-practice observations in the data set under analysis, yielding a convex production possibility set (see Berger and Humphrey, 1997). However,

because DEA uses only the data on inputs and outputs and does not take direct account of input prices, it does not incorporate allocative inefficiency.

The SFA approach, based on the ideas of Aigner *et al.* (1977) and Meeusen and van den Broeck (1977), involves the estimation of a specific parameterized efficiency frontier with a composite error term consisting of nonnegative inefficiency and noise components. X-efficiency can thus be measured in terms of cost efficiency, revenue efficiency, or profit efficiency, depending on the type of frontier used. The DEA and SFA approaches generally give very different efficiency estimates. However, Bauer *et al.* (1998) and Rossi and Ruzzier (2000) argue that it is not necessary to have a consensus on which is the single best frontier approach for measuring efficiency. They also propose a series of criteria to evaluate if the inefficiency estimates obtained from different approaches are mutually consistent in terms of inefficiency scores and ranks.

Cost efficiency has received the most attention in the parametric analysis of efficiency of the U.S. banking industry. According to Berger and Humphrey (1997), 30 out of 38 studies that employed parametric techniques in the analysis of efficiency in the U.S. banking industry were reported to employ cost functions, and the rest employed profit functions — among these 38 parametric studies of the efficiency of the U.S. banking industry, several employed TFA and DFA. Despite its popularity, the cost frontier used in previous studies suffers from the following two problems. First, the estimated parameters of cost frontiers frequently violate the monotonicity and concavity constraints implied by economic theory, which eventually leads to wrong conclusions concerning efficiency levels. While permitting a parameterized function to depart from the neoclassical function space is usually fit-improving, it also causes the hypothetical best practice firm not to be fully efficient at those data points where theoretical regularity is violated.

Second, the cost frontier suffers the problem of not having enough flexibility. Most of the previous studies employ a translog functional form. Researchers have found, however,

that the translog function lacks enough flexibility in modelling the U.S. banking industry which is composed of banks of widely varying sizes (see McAllister and McManus, 1993; Wheelock and Wilson, 2001). In an attempt to increase flexibility, more recent studies employ a so called 'Fourier function' which is actually a translog function augmented with trigonometric Fourier terms. Although this so-called 'Fourier function' can improve the goodness of fit, it is not a true Fourier flexible functional form, in Gallant's (1982) original sense. In particular, the original Fourier flexible functional form consists of two components with the first component being a 'reparameterized' translog function and the second component a trigonometric Fourier series. It is important to note that these two components are not independent of each other. In fact, the scaled variables of outputs and input prices are not only used in the Fourier series, but also in the modified translog part. However, the so-called 'Fourier function' ignores the parametric relationship between the two components of the Fourier function, and just includes the scaled variables of outputs and input prices in the Fourier series. While this practice makes it a lot easier to use the Fourier function, it may be unable to reach close approximation in the Sobolev norm and may result in inconsistent parameter estimates.

Motivated by the widespread practice of ignoring the theoretical regularity conditions and not using a globally flexible functional form, as summarized in Table 3.1, the purpose of this paper is to reinvestigate the cost efficiency of the U.S. banking industry with more recent panel data over the sample period from 1998 to 2005, and by addressing the above two problems inherent in previous studies. In doing so, I take the SFA approach, and minimize the potential problem of using a misspecified functional form by employing a globally flexible functional form — Gallant's (1982) original Fourier flexible functional cost form. It should be noted that there are two globally flexible functional forms which can provide greater flexibility than locally flexible functional forms: the Fourier flexible functional form and the Asymptotically Ideal Model, introduced by Barnett *et al.* (1991).

The former is based on a Fourier series expansion and the latter is based on a linearly homogeneous multivariate Muntz-Szatz series expansion. Both of them are globally flexible in the sense that they are capable of approximating the underlying cost function at every point in the function's domain by increasing the order of the expansion, and thus have more flexibility than most of the locally flexible functional forms which theoretically can attain flexibility only at a single point or in an infinitesimally small region. In this study I employ the Fourier cost functional form which is both log-linear and globally flexible. In the implementation of it, I strictly follow Gallant's (1982) original specification of the functional form rather than just include the scaled variables of outputs and input prices in the Fourier series as previous studies did.

I also estimate the Fourier flexible cost function subject to full theoretical regularity. There are three approaches to incorporating curvature and/or monotonicity restrictions into flexible functional forms — the Cholesky factorization approach, the Bayesian approach, and the nonlinear constrained optimization approach. The Cholesky factorization approach can only guarantee the negative semidefiniteness of the Hessian matrix of a cost function in a region around the reference point (that is, a data point where curvature is imposed), and satisfaction of curvature at data points far away from the reference point can only be obtained by luck (see Ryan and Wales, 2000). This is not satisfactory especially when the sample size is large and violations of curvature are widespread. The Bayesian approach involves specifying prior distributions for parameters and inefficiency terms. However, the specification of prior distributions adds extra uncertainty to the outcome of the modelling exercise especially when researchers have no idea of how to parameterize a priori the unknown parameters (see Diewert, 2004; Greene, 2005). The nonlinear constrained optimization approach, originally proposed by Gallant and Golub (1984) and recently used by Serletis and Shahmoradi (2005) in the context of consumer demand systems, develops computational methods for imposing curvature restrictions at

any arbitrary set of points. Monotonicity can also be incorporated into the estimation of the cost function although the original Gallant and Golub (1984) paper does not do so. This method applies to any cost function as long as the Hessian matrix (or some transform of the Hessian matrix) and the first order conditions of the cost function can be explicitly specified. While the nonlinear constrained optimization method has many desirable properties, no attempt has been made in the stochastic frontier literature to use this method to incorporate monotonicity and curvature on parametric (cost or distance) functions.

The rest of the paper is organized as follows. Section 2 provides a brief review of stochastic cost frontiers. In Section 3 I present the Fourier cost function and detail the homogeneity, monotonicity, and curvature constraints implied by neoclassical microeconomic theory. In Section 4 I discuss the constrained nonlinear optimization methodology for imposing these constraints on the parameters of the Fourier cost function. Section 5 deals with the data description. In Section 6, I apply my model to panel data on U.S. banks, and discuss the effect of the incorporation of monotonicity and curvature on cost efficiency, and also report my estimates on cost efficiency for twelve different bank groups. Section 7 summarizes and concludes the paper.

3.2 Stochastic Cost Frontier

Within a panel data framework, the cost frontier model can be written as

$$C_{it} = f(\boldsymbol{X}_{it}, \boldsymbol{\rho})\, \tau_{it} \zeta_{it}, \qquad i = 1, \cdots, I, \quad t = 1, \cdots T. \tag{3.1}$$

This model decomposes the observed cost for firm i at time t, C_{it}, into three parts — (i) the actual frontier $f(\boldsymbol{X}_{it},\boldsymbol{\rho})$, which depends on $\boldsymbol{X}_{it}$, a vector of exogenous variables (i.e. input prices and output quantities), and $\boldsymbol{\rho}$, a vector of parameters, and which

represents the minimum possible cost of producing a given level of output with certain input prices; (ii) a non-negative term $\tau_{it} \geq 1$, measuring firm specific inefficiency; and (iii) a random error, ζ_{it}, which captures statistical noise. The deterministic kernel of the cost frontier is $f(\boldsymbol{X}_{it}, \boldsymbol{\rho})$, and the stochastic cost frontier is $f(\boldsymbol{X}_{it}, \boldsymbol{\rho})\zeta_{it}$. As required by microeconomic theory, $f(\boldsymbol{X}_{it}, \boldsymbol{\rho})$ is a linearly homogeneous and concave function in prices and also nondecreasing in both input prices and outputs.

I follow the common practice in this literature and assume that $f(\boldsymbol{X}_{it}, \boldsymbol{\rho})$ is a log-linear functional function. The stochastic cost function in (3.1) is rewritten as

$$c_{it} = \alpha + \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_{it} + v_{it}, \tag{3.2}$$

where $c_{it} = \ln C_{it}$; $\alpha + \boldsymbol{x}_{it}'\boldsymbol{\beta} = \ln f(\boldsymbol{X}_{it}, \boldsymbol{\rho})$; $u_{it} = \ln \tau_{it} \geq 0$; and $v_{it} = \ln \zeta_{it}$. $\boldsymbol{x}_{it}$ is the counterpart of $\boldsymbol{X}_{it}$ with the input prices and output quantities transformed to logarithms, $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameters, and α is the intercept. Thus the composite error term ε_{it} $(= u_{it} + v_{it})$ consists of two parts with u_{it} capturing the level of firm inefficiency and v_{it} capturing statistical noise.

In an empirical exercise, assumptions are commonly made about the two error components. Usually the v_{it}'s are assumed to be iid $N(0, \sigma^2)$ and independent of the u_{it}'s, an assumption I maintain throughout this paper. In the specification of the distribution for the u_{it}'s I assume

$$u_{it} = \eta_{it} u_i \tag{3.3}$$

where

$$\eta_{it} = \exp\left[-\eta_1(t-T) - \eta_2(t-T)^2\right], \qquad t = 1, \cdots, T, \tag{3.4}$$

where η_1 and η_2 are parameters to be estimated and the u_i's are assumed to be independently and identically distributed non-negative truncations of the $N(0, \sigma_u^2)$ distribution.

Note that the above exponential function of time, η_{it}, is a generalization of that proposed by Battese and Coelli (1992) in the sense that it relaxes the monotonicity of the temporal variation pattern of the efficiency term using a two parameter specification.

The cost efficiency of firm i at time t can then be defined as the ratio of minimum cost attainable in an environment characterized by $\exp(v_{it})$ to observed expenditure, as follows

$$CE_{it} = \frac{f(\boldsymbol{X}_{it}, \boldsymbol{\rho})\zeta_{it}}{C_{it}}$$

$$= \exp(\alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta} + v_{it} - c_{it})$$

$$= \exp(-u_{it}), \tag{3.5}$$

with $CE_{it} \leq 1$. Notice that $CE_{it} = 1$ if and only if $c_{it} = \alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta} + v_{it}$. For example, if a firm is 80% efficient, it could reduce costs by 20% simply by becoming fully efficient.

3.3 The Fourier Cost Function

I assume that $\alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta}$ in equation (3.2) is an M-output and N-input Fourier cost functional form, as follows

$$g(\boldsymbol{l}_{it}, \boldsymbol{q}_{it}, \vartheta) = u_0 + \boldsymbol{b}'\boldsymbol{z}_{it} + \frac{1}{2}\boldsymbol{z}'_{it}\boldsymbol{A}\boldsymbol{z}_{it}$$

$$+ \sum_{\alpha=1}^{E} \left\{ \widetilde{u}_{0\alpha} + 2\sum_{j=1}^{J} \left(\widetilde{u}_{j\alpha} \cos(j\lambda \boldsymbol{k}'_{\alpha}\boldsymbol{z}_{it}) - \widetilde{v}_{j\alpha} \sin(j\lambda \boldsymbol{k}'_{\alpha}\boldsymbol{z}_{it}) \right) \right\}, \tag{3.6}$$

where $u_0 = \alpha$, $\boldsymbol{\beta} = (\boldsymbol{b}, \widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}})$, and $\boldsymbol{\vartheta} = (u_0, \boldsymbol{b}, \widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}})$ is a vector of parameters to be estimated. $\boldsymbol{z}_{it} = (\boldsymbol{l}_{it}, \boldsymbol{q}_{it})'$ is a $(N + M)$ vector of rescaled log input prices, $\boldsymbol{l}_{it}$, and rescaled log outputs, $\boldsymbol{q}_{it}$. The procedure for this rescaling is the same as suggested by Gallant (1982)

$$\begin{aligned} l_n &= \ln p_n + \ln a_n > 0, \qquad n = 1, \cdots, N; \\ q_m &= \mu_m(\ln y_m + \ln \widetilde{a}_m) > 0, \qquad m = 1, \cdots, M, \end{aligned} \tag{3.7}$$

where p_n is the price for input n, y_m is the quantity for output m, and the location parameters $\ln a_n$ and $\ln \widetilde{a}_m$ are chosen as

$$\begin{aligned} \ln a_n &= -\min\{\ln p_n\} + 10^{-5}, \qquad n = 1, \cdots, N; \\ \ln \widetilde{a}_m &= -\min\{\ln y_m\} + 10^{-5}, \qquad m = 1, \cdots, M. \end{aligned} \tag{3.8}$$

In equation (4.5), $\boldsymbol{A} = -\sum_{\alpha=1}^{E} \widetilde{u}_{0\alpha} \lambda^2 \boldsymbol{k}_\alpha \boldsymbol{k}'_\alpha$; λ is a rescaling factor, and $\boldsymbol{k}_\alpha$ is a multi-index — an $(N + M)$ vector with integer components. As Gallant (1982) shows, the length of a multi-index, denoted as $|\boldsymbol{k}_\alpha|^* = \sum_{i=1}^{n} |k_{i\alpha}|$, reduces the complexity of the notation required to denote high-order partial differentiation and multivariate Fourier trigonometric terms (those sin and cos terms). Following Gallant (1982), these indexes are constructed using the following rules (the construction of these indexes is complex and is performed using MATLAB). First, the zero vector and any $\boldsymbol{k}_\alpha$ whose first non-zero element is negative are deleted. Second, every index with a common integer divisor is also deleted.

As a Fourier term is a periodic function in its arguments but the cost function is

not, the scaling of the data is also important. In empirical applications, to avoid the approximation from diverging from the true cost function, the data should be rescaled by a common scaling factor, λ, so that the input prices and output quantities lie in the interval $[0, 2\pi]$. The common scaling factor, λ, for input prices is defined analogously as in Gallant (1982). The parameters E (the number of terms) and J (the degree of the approximation) determine the degree of the Fourier polynomials. Thus, the Fourier cost function has $1 + (N + M) + E(1 + 2J)$ parameters to be estimated.

Substituting the cost frontier defined by (3.6) into (3.2), I obtain the basic panel data stochastic cost frontier model I am going to use in this paper

$$c_{it} = u_0 + \boldsymbol{b}'\boldsymbol{z}_{it} + \frac{1}{2}\boldsymbol{z}'_{it}\boldsymbol{A}\boldsymbol{z}_{it}$$

$$+ \sum_{\alpha=1}^{E} \left\{ \widetilde{u}_{0\alpha} + 2\sum_{j=1}^{J} \left(\widetilde{u}_{j\alpha} \cos\left(j\lambda \boldsymbol{k}'_{\alpha}\boldsymbol{z}_{it}\right) - \widetilde{v}_{j\alpha} \sin\left(j\lambda \boldsymbol{k}'_{\alpha}\boldsymbol{z}_{it}\right) \right) \right\}$$

$$+ u_{it} + v_{it}, \tag{3.9}$$

where all parameters and variables are defined the same as above.

3.3.1 Theoretical Regularity

As required by microeconomic theory, the Fourier cost function in (3.6) has to satisfy certain theoretical regularity conditions, i.e. homogeneity, monotonicity, and concavity. The restriction of linear homogeneity on the Fourier cost frontier can be imposed through reparameterization, as in Gallant(1982) and Gallant and Golub(1984),

$$\sum_{n=1}^{N} b_n = 1 \tag{3.10}$$

and

$$\widetilde{u}_{j\alpha} = \widetilde{v}_{j\alpha} = 0 \text{ if } \sum_{n=1}^{N} k_{n\alpha} \neq 0. \tag{3.11}$$

Restriction (3.10) guarantees the linear homogeneity of the first order terms, and (3.11) guarantees the linear homogeneity of both the second order terms and the Fourier trigonometric terms.

I now turn to the monotonicity and curvature constraints. For simplicity, the subscripts i and t for all variables are suppressed in this subsection to avoid notational cluster. Define $\nabla_z g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}) = \partial[g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta})]/\partial \boldsymbol{z}$, and $\nabla^2_{zz} g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}) = \partial[\nabla_z g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta})]/\partial \boldsymbol{z}$, where $\boldsymbol{z} = (\boldsymbol{l},\boldsymbol{q})$ as above. By the two equations defined in (3.7), it can be easily shown that

$$g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}) \equiv \ln f(p_1,\cdots,p_n,y_1,\cdots,y_m)$$

$$= \ln f\left(\frac{e^{l_1}}{a_1},\cdots,\frac{e^{l_n}}{a_n},\frac{e^{q_1}}{\widetilde{a}_1},\cdots,\frac{e^{q_m}}{\widetilde{a}_m}\right), \tag{3.12}$$

where $f(p_1,\cdots,p_n,y_1,\cdots,y_m) = f(\boldsymbol{X}_{it},\boldsymbol{\rho})$ is the cost frontier corresponding to the Fourier cost function. In what follows, I use $f(\boldsymbol{p},\boldsymbol{y})$ instead of $f(\boldsymbol{X}_{it},\boldsymbol{\rho})$. Taking the partial derivative of both sides of (3.12) with respect to $\boldsymbol{z}$, I can obtain the following equation

$$\frac{\partial f(\boldsymbol{p},\boldsymbol{y})}{\partial \widetilde{\boldsymbol{z}}} = f(\boldsymbol{p},\boldsymbol{y})\,\boldsymbol{Z}^{-1}\nabla_z g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}), \tag{3.13}$$

where $\widetilde{\boldsymbol{z}} = (\boldsymbol{p},\boldsymbol{y})$ and $\boldsymbol{Z}$ is a diagonal matrix with unscaled input prices $(p_1,\cdots,p_N)$ and outputs $(y_1,\cdots,y_M)$ on its main diagonal. With both $f(\boldsymbol{p},\boldsymbol{y})$ and $\boldsymbol{Z}^{-1}$ being positive,

monotonicity ($\partial [f(\boldsymbol{p}, \boldsymbol{y})] / \partial \boldsymbol{p} > 0$) requires

$$\begin{aligned} \nabla_{l_n} g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}) &= \partial [g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta})] / \partial l_n > 0, \qquad n = 1, \cdots, N; \\ \nabla_{q_m} g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}) &= \partial [g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta})] / \partial q_m > 0, \qquad m = 1, \cdots, M, \end{aligned} \tag{3.14}$$

where $\nabla_{l_n} g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta})$ has to satisfy $\sum_{n=1}^{N} \nabla_{l_n} g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}) = 1$, which can be derived from the fact that cost function is homogenous of degree one in prices, i.e.

$$\begin{aligned} \sum_{n=1}^{N} \nabla_{l_n} g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}) &= \sum_{n=1}^{N} \frac{\partial \ln f(\boldsymbol{p}, \boldsymbol{y})}{\partial \ln p_n} \\ &= \left[\sum_{n=1}^{N} \left(\frac{\partial f(\boldsymbol{p}, \boldsymbol{y})}{\partial p_n} p_n \right) \right] \frac{1}{f(\boldsymbol{p}, \boldsymbol{y})} = 1. \end{aligned} \tag{3.15}$$

In equation (3.15), the first equality can be obtained by using (3.13).

Concavity in input prices requires that the Hessian matrix, $\boldsymbol{H}$, of the cost frontier, $f(\boldsymbol{p}, \boldsymbol{y})$, is negative semidefinite. It can be easily shown that the element of the ith row and jth column of the Hessian matrix, $\boldsymbol{H}$, of the cost function $f(\boldsymbol{p}, \boldsymbol{y})$ is given by (see Appendix)

$$\boldsymbol{H}_{ij} = \left(s_i^{-1} \nabla_{\ln p_j} s_i + s_j - \delta_{ij} \right) \frac{\widetilde{X}_i(\boldsymbol{p},\boldsymbol{y})}{p_j}, \qquad i, j = 1, \cdots, N, \tag{3.16}$$

where s_i is the cost share for input i, $\nabla_{\ln p_j} s_i$ is the derivative of s_i with respect to the log price of input j, $\widetilde{X}_i$ is the demand for input i, obtained by Shephard's lemma as the first derivative of the cost function with respect to input price p_i, and $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

Since $s_i = \nabla_{l_i} g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta})$ and $\nabla_{\ln p_j} s_i = \nabla^2_{l_i l_j} g(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta})$, for $i = 1, \cdots, N$, equation (3.16)

can be rewritten as

$$\boldsymbol{H}_{ij} = \left(\left(\nabla_{l_i} g\left(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}\right) \right)^{-1} \nabla^2_{l_i l_j} g\left(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}\right) + \nabla_{l_j} g\left(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}\right) - \dot{\delta_{ij}} \right) \frac{\widetilde{X}_i\left(\boldsymbol{p},\boldsymbol{y}\right)}{p_j} \tag{3.17}$$

Since $\widetilde{X}_i\left(\boldsymbol{p},\boldsymbol{y}\right)$ is positive by the property of monotonicity, $\nabla_{l_j} g\left(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}\right)$ is positive by equation (3.6), and p_j is also positive, concavity of the Hessian matrix in my particular case is equivalent to requiring (in matrix notation) that

$$\boldsymbol{G} = \nabla^2_{ll} g\left(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}\right) + \nabla_l g\left(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}\right) \left(\nabla_l g\left(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}\right) \right)' - \operatorname{diag}\left(\nabla_l g\left(\boldsymbol{l},\boldsymbol{q},\boldsymbol{\vartheta}\right) \right) \tag{3.18}$$

be a negative semidefinite matrix. Thus, (3.14) and (3.18) are the constraints I need to incorporate into the estimation of the Fourier cost frontier defined in (3.6) — the monotonicity and curvature conditions are provided in Gallant (1982) without proof.

3.4 Constrained Optimization

In this section, I follow Gallant and Golub (1984) and show how the constrained nonlinear optimization approach can be used to impose the monotonicity and curvatures constraints given in (3.14) and (3.18) on the parameters of the Fourier cost function, defined in (3.6). Using the reparameterization method suggested by Battese and Corra (1977), the model above is parameterized in terms of σ_s^2 and γ, where $\sigma_s^2 = \sigma_v^2 + \sigma_u^2$ is the overall variance, and $\gamma = \sigma_u^2/\sigma_s^2$ is an indicator of the relative importance of noise and inefficiency variances. Under these assumptions, constrained optimization will give asymptotically efficient estimates for all the parameters.

With the distributional assumptions in Section 2, the log likelihood function for a

sample of I firms for T periods of time is given by

$$\begin{aligned}\ln L\left(\varphi\left(\boldsymbol{\theta}\right)\right) = & -\frac{1}{2}IT\left[\ln(2\pi)+\ln(\sigma_s^2)\right]-\frac{1}{2}I(T-1)\ln(1-\gamma)\\ & -\frac{1}{2}\sum_{i=1}^{I}\ln\left[(1+(\eta_i'\eta_i-1)\gamma\right]-I\ln\left(\frac{1}{2}\right)\\ & +\sum_{i=1}^{I}\ln\left(1-\Phi(-z_i^*)\right)+\frac{1}{2}\sum_{i=1}^{I}z_i^{*2}\\ & -\frac{1}{2}\sum_{i=1}^{I}\left[c_i-(\alpha+\boldsymbol{x}_i'\boldsymbol{\beta})\right]'\left[c_{it}-(\alpha+\boldsymbol{x}_i'\boldsymbol{\beta})\right]/\left[(1-\gamma)\sigma_s^2\right], \qquad (3.19)\end{aligned}$$

where $\boldsymbol{\theta}=(\vartheta,\sigma_s^2,\gamma,\eta_1,\eta_2)$, $z_i^*=\gamma\eta_i'\left[c_i-(\alpha+\boldsymbol{x}_i'\boldsymbol{\beta})\right]/\left\{\gamma(1-\gamma)\sigma_s^2\left[1+(\eta_i'\eta_i-1)\gamma\right]\right\}^{1/2}$, and $\Phi(\cdot)$ represents the distribution function for the standard normal random variables — see Battese and Coelli (1992) for details about the derivation of the log likelihood function and its derivatives in the production frontier context. Estimates of u_0, $\boldsymbol{\beta}$, σ_s^2, γ, η_1, and η_2 can be obtained by minimizing $-\ln L\left(\varphi\left(\boldsymbol{\theta}\right)\right)$, that is, maximizing the log likelihood function, $\ln L\left(\varphi\left(\boldsymbol{\theta}\right)\right)$ with respect to the parameters. In minimizing $-\ln L\left(\varphi\left(\boldsymbol{\theta}\right)\right)$, I use the TOMLAB/NPSOL tool box with MATLAB — see http://tomlab.biz/products/npsol. NPSOL uses a sequential quadratic programming algorithm and is suitable for both unconstrained and constrained optimization of smooth (that is, at least twice-continuously differentiable) nonlinear functions.

I first run an unconstrained optimization using (3.19) and check the theoretical regularity conditions of monotonicity and curvature. In case that the monotonicity and curvature conditions are not satisfied at all observations, I use the NPSOL nonlinear programming program to minimize $-\ln L\left(\varphi\left(\boldsymbol{\theta}\right)\right)$ with monotonicity and concavity imposed. Essentially, this becomes a constrained maximum likelihood problem.

While I follow Gallant and Golub (1984) and use nonlinear constrained optimization to impose curvature, I do not do it by constructing their submatrix $\boldsymbol{K}_{22}$ using a Householder

transformation and then deriving an indicator function for the smallest eigenvalue of $\boldsymbol{K}_{22}$ and it derivative. Instead, I work directly with the matrix $\boldsymbol{G}$ defined in (3.18), restricting its eigenvalues to be nonpostive. This is because a necessary and sufficient condition for negative semidefiniteness of $\boldsymbol{G}$ is that all its eigenvalues are nonpositive (see Morey, 1986). Compared with the Gallant and Golub (1984) approach where a reduced matrix $\boldsymbol{K}_{22}$ is sought, the direct restriction of the eigenvalues of $\boldsymbol{G}$ to be nonpositive seems more appealing.

It is well known that an $N \times N$ real symmetric matrix has N eigenvalues, with these eigenvalues being real numbers (see Magnus,1985). Let $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_N]$ then denote the N eigenvalues of $\boldsymbol{G}$, a real symmetric matrix defined in (3.18). The nonlinear curvature constraints for my constrained optimization problem can then be written as

$$\boldsymbol{\lambda}_n \leq 0, \qquad n = 1, \cdots, N.$$

The eigenvalues of $\boldsymbol{G}$ can be obtained by solving

$$|\boldsymbol{G} - \boldsymbol{\lambda} \boldsymbol{I}_N| = 0, \tag{3.20}$$

where $\boldsymbol{I}_N$ is an $N \times N$ identity matrix. Clearly, λ_n $(n = 1, \cdots, N)$ are functions of the elements of $\boldsymbol{G}$, denoted $\boldsymbol{G}_{ij}$, which are in turn linear functions of $\nabla^2_{l_i l_j} g(\boldsymbol{l}, \boldsymbol{q}, \boldsymbol{\vartheta})$ and $\nabla_{l_j} g(\boldsymbol{l}, \boldsymbol{q}, \boldsymbol{\vartheta})$ as can be seen from (3.18). In fact, in my case with $N = 3$, I have

$$\lambda_n(\boldsymbol{\vartheta}) = \lambda_n \left[\boldsymbol{G}_{11}(\boldsymbol{\vartheta}), \boldsymbol{G}_{12}(\boldsymbol{\vartheta}), \boldsymbol{G}_{13}(\boldsymbol{\vartheta}), \boldsymbol{G}_{22}(\boldsymbol{\vartheta}), \boldsymbol{G}_{23}(\boldsymbol{\vartheta}), \boldsymbol{G}_{33}(\boldsymbol{\vartheta}) \right], \tag{3.21}$$

for $n = 1, 2, 3$, where

$$\boldsymbol{G}_{ij}(\boldsymbol{\vartheta}) = \nabla^2_{l_i l_j} g(\boldsymbol{l}, \boldsymbol{q}, \boldsymbol{\vartheta}) + \nabla_{l_i} g(\boldsymbol{l}, \boldsymbol{q}, \boldsymbol{\vartheta}) \nabla_{l_j} g(\boldsymbol{l}, \boldsymbol{q}, \boldsymbol{\vartheta}) - \delta_{ij} \nabla_{l_i} g(\boldsymbol{l}, \boldsymbol{q}, \boldsymbol{\vartheta}), \tag{3.22}$$

for $i = 1, 2, 3$ and $j = i, \cdots, 3$. Explicit formulas for $\lambda_n(\vartheta)$ in terms of the $\boldsymbol{G}_{ij}$ elements can be easily obtained using the symbolic toolbox in MATLAB. After substituting (3.22) into $\lambda_n(\vartheta)$, the eigenvalues in terms of $\nabla^2_{l_i l_j} g(\boldsymbol{l},\boldsymbol{q},\vartheta)$ and $\nabla_{l_i} g(\boldsymbol{l},\boldsymbol{q},\vartheta)$ can be obtained.

As for the derivatives of $\lambda_n(\boldsymbol{\theta})$, they can be obtained using equation (3.21), as follows

$$\frac{\partial \lambda_n(\vartheta)}{\partial \vartheta} = \sum_{i=1}^{N}\sum_{j=i}^{N}\left\{\frac{\partial \lambda_n}{\partial \boldsymbol{G}_{ij}} \times \left[\frac{\partial\left[\nabla^2_{l_i l_j} g(\boldsymbol{l},\boldsymbol{q},\vartheta)\right]}{\partial \vartheta} + \right.\right.$$

$$\left.\left. + \nabla_{l_i} g(\boldsymbol{l},\boldsymbol{q},\vartheta)\frac{\partial\left[\nabla_{l_j} g(\boldsymbol{l},\boldsymbol{q},\vartheta)\right]}{\partial \vartheta} + \left[\nabla_{l_j} g(\boldsymbol{l},\boldsymbol{q},\vartheta) - \delta_{ij}\right]\frac{\partial\left[\nabla_{l_i} g(\boldsymbol{l},\boldsymbol{q},\vartheta)\right]}{\partial \vartheta}\right]\right\} \quad (3.23)$$

All of $\partial\left[\nabla^2_{l_i l_j} g(\boldsymbol{l},\boldsymbol{q},\vartheta)\right]/\partial\vartheta$, $\partial\left[\nabla_{l_i} g(\boldsymbol{l},\boldsymbol{q},\vartheta)\right]/\partial\vartheta$, and $\partial\left[\nabla_{l_j} g(\boldsymbol{l},\boldsymbol{q},\vartheta)\right]/\partial\vartheta$ can be easily computed. In my case with $N = 3$, each of (the eighteen) $\partial\lambda_n/\partial\boldsymbol{G}_{ij}$ (for $n = 1, 2, 3$, $i = 1, 2, 3$, and $j = i, \cdots, 3$) are calculated using the symbolic toolbox in MATLAB.

In addition to the imposition of concavity, the monotonicity constraints in (3.14) also need to be imposed, if monotonicity is violated. The derivatives for the monotonicity constraints, $\partial\left[\nabla_{l_n} g(\boldsymbol{l},\boldsymbol{q},\vartheta)\right]/\partial\vartheta$ and $\partial\left[\nabla_{q_m} g(\boldsymbol{l},\boldsymbol{q},\vartheta)\right]/\partial\vartheta$, also can be easily computed. Hence, my constrained maximum likelihood problem can be written as follows

$$\min_{\boldsymbol{\theta}} \varphi(\boldsymbol{\theta}) = -\ln L(\varphi(\boldsymbol{\theta})), \quad (3.24)$$

subject to

$$\boldsymbol{\lambda}_n(\vartheta) = 0, \qquad n = 1, \cdots, N; \quad (3.25)$$

$$\boldsymbol{W}_j(\vartheta) \geq 0, \qquad j = 1, \cdots, M + N; \quad (3.26)$$

where $\boldsymbol{\lambda}_n$ is the curvature constraint for each observation and $\boldsymbol{W}_j$ is the monotonicity constraint for each observation as shown in (3.14). As already noted, I can impose the regularity constraints locally (at single data point), regionally (over a region of data points), or fully (at every data point in the sample). After estimates of u_0, $\boldsymbol{\beta}$, σ_s^2, γ, η_1, and η_2 are obtained, σ_u^2 and σ_v^2 can then be calculated by using $\sigma_s^2 = \sigma_v^2 + \sigma_u^2$ and $\gamma = \sigma_u^2/\sigma_s^2$, both of which are discussed above.

Following Battese and Coelli (1992), the minimum-mean-squared-error predictor of the cost efficiency of the ith bank at time t, $CE_{it} = \exp(-u_{it})$ is

$$CE_{it} = E\left(\exp\{-u_{it}\}|\varepsilon_{it}\right)$$

$$= \left\{\frac{1-\Phi\left[\eta_{it}\sigma_i^* - (\mu_i^*/\sigma_i^*)\right]}{1-\Phi\left[-(\mu_i^*/\sigma_i^*)\right]}\right\}\exp\left(-\eta_{it}\mu_i^* + \frac{1}{2}\eta_{it}^2\sigma_i^{*2}\right), \tag{3.27}$$

where

$$\mu_i^* = \frac{\eta_i'\left(c_i - \alpha - x_i'\beta\right)\sigma_u^2}{\left(\sigma_v^2 + \eta_i'\eta_i\sigma_u^2\right)}; \tag{3.28}$$

$$\sigma_i^* = \frac{\sigma_v^2\sigma_u^2}{\left(\sigma_v^2 + \eta_i'\eta_i\sigma_u^2\right)}. \tag{3.29}$$

This framework allows us to calculate the efficiency level of each bank relative to the best-practice bank represented by the cost frontier.

While I follow Gallant and Golub (1984) in imposing the theoretical regularity conditions on the parameters of the Fourier flexible cost function, I extend Gallant's method in two ways. First, I extend Gallant's constrained non-linear optimization approach from a traditional factor demand system framework to a stochastic frontier framework. This extension involves the use of a much more complicated log likelihood function as the

objective function, rather than the simple least squares based objective function used in Gallant and Golub (1984). This is because a composed error term is assumed in the stochastic frontier framework, whereas a simple iid $N(0, \sigma^2)$ error term is assumed in the traditional factor demand system framework. Second, I extend Gallant's method from a time series framework to a panel data framework.

3.5 The Data

The data used in this study, obtained from the Reports of Income and Condition (Call Reports), cover the period from 1998 to 2005. I 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. There were 10,139 banks in the United States banking industry in 1998, and the number declined to 8,390 in 2005 due to industry consolidation. After deleting those observations whose input prices are negative or zero, I obtained a balanced panel of 6,010 observations for 8 years, from 1998 to 2005.

In choosing which financial accounts to specify as outputs versus inputs, I use the accounting balance-sheet approach of Sealey and Lindley (1977). All liabilities (core deposits and purchased funds) and financial equity capital provide funds and are treated as inputs. All assets (loans and securities) use bank funds and are treated as outputs. This approach is different from the intermediation approach which is consistent with the value added definition of output production by financial firms and with user-cost price evaluation of the services of outputs. An accurate representation of the intermediation approach can be found in Barnett (1987), Barnett and Hahm (1994), Barnett and Zhou (1994), Barnett *et al.* (1995), and Hancock (1991).

In this paper, three output quantities and three input prices are identified. The three

outputs are consumer loans, y_1; non-consumer loans, y_2, is composed of industrial and commercial loans and real estate loans; and securities, y_3, 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. All outputs are deflated by the Consumer Price Index (CPI) to the base year 1998. The three prices includes: the wage rate for labor, p_1; the interest rate for borrowed funds, p_2; and the price of physical capital, p_3. The wage rate equals total salaries and benefits divided by the number of full-time employees. The price of capital equals expenses on premises and equipment divided by premises and fixed assets. The price of deposits and purchased funds equals total interest expense divided by total deposits and purchased funds. Total cost is thus the sum of these three input costs. This specification of outputs and input prices is the same as or similar to most of the previous studies in this literature (see, for example, Akhigbea and McNulty, 2003; Stiroh, 2000; Berger and Mester, 2003). Thus, $M = N = 3$ in this paper. The three outputs and three input prices are then scaled, using the formulas specified in equation (3.7)–(3.8) of Section 3 for each of the twelve asset size classes, which I will discuss in more detail below.

The set of elementary multi-indexes that satisfy $\sum_{i=1}^{3} k_{i\alpha} = 0$ and have norm $k_{i\alpha} \leq 3$ are displayed in Table 3.2 — these three $k_{i\alpha}$ $(i = 1, 2, 3)$ are the three elements in the $\boldsymbol{k}_\alpha$ vector corresponding to the three input prices. For this set $E = 32$, and I take $J = 1$. While Chalfant and Gallant (1985) and Eastwood and Gallant (1991) have suggested that the number of parameters to be estimated should be equal to the number of effective sample observations raised to the power of 2/3, in this paper I set the number of parameters such that $k_{i\alpha} \leq 3$ in order to reduce the number of parameters to a manageable level, given that I also have to deal with hundreds of variables and thousands of highly non-linear constraints. Thus I have a total of $1 + (N + M) + E(1 + 2J) = 1 + (3 + 3) + 32 \times (1 + 2) = 103$ free parameters (that is, parameters estimated directly).

However, the effective number of parameters is 85 due to the following restrictions. The homogeneity restriction,

$$\sum_{i=1}^{3} b_i = 1, \tag{3.30}$$

reduces the number of free parameters by one. The remaining restrictions are due to the overparameterization of the $\boldsymbol{A}$ matrix. In particular, $\boldsymbol{A}$ is a 6×6 symmetric matrix which satisfies three linearly independent homogeneity restrictions,

$$\sum_{i=1}^{3} A_{ij} = 0, \qquad j = 1, \cdots, 6. \tag{3.31}$$

Moreover, the symmetry of the matrix $\boldsymbol{A}$ also implies

$$\sum_{j}^{3} A_{ij} = 0, \qquad i = 1, \cdots, 6. \tag{3.32}$$

Thus $\boldsymbol{A}$ can have at most 15 free parameters, and in the parameterization

$$\boldsymbol{A} = -\sum_{\alpha=1}^{32} u_{0\alpha} \lambda^2 k_\alpha k'_\alpha \tag{3.33}$$

15 of the $u_{0\alpha}$ parameters are free parameters and 17 parameters must be set equal to zero. These seventeen k_α parameters are listed in the last seventeen columns of Table 3.1.

Following Berger and Mester (2003), I add three more variables into the Fourier cost function: financial equity capital, $\widetilde{z}_1$, non-traditional banking activities, $\widetilde{z}_2$, and a time trend, t. Financial equity capital is treated as a fixed net input and off-balance-sheet items are treated as a fixed net output. The time trend t is intended to capture the effect of technological change on cost. In the treatment of non-traditional banking activities, I follow Boyd and Gertler (1994) and use an asset-equivalent measure (AEM) of these

non-traditional activities. I assume that all non-interest income is generated from off-balance-sheet assets, and that these non-traditional activities yield the same rate of return on assets (ROA) as traditional activities do. Thus, I transform the off-balance-sheet income into an equivalent asset. The two fixed net inputs are measured in 1998 constant dollars and used in logarithm form. When adding the $\widetilde{z}_1$, $\widetilde{z}_2$, and t variables in the Fourier cost function, these variables are used in linear and quadratic form (i.e. $\beta_{z_1}\widetilde{z}_1^2 + \beta_{z_{1s}}\widetilde{z}_1^2 + \beta_{z_2}\widetilde{z}_2 + \beta_{z_{2s}}\widetilde{z}_2^2 + \beta_t t + \beta_{tt}t^2$), and do not interact with the outputs and input prices in order to reduce the number of parameters to a manageable level and to lessen the effects of multicollinearity.

Separating banks into asset size classes is a common approach in assessing the performance of banks asset size. However, given the unique nature of the distribution of asset size for commercial banks in the United States, it is very difficult to categorize banks based upon asset size and also there is no industry standard on asset ranges. Over my sample period, from 1998 to 2005, around 85% of all commercial banks report less than $500 million in total assets. However, over that same time period, there exists a cluster of extremely large banks with over $3 billion in total assets that accounts for roughly 2.3% of all commercial banks. In this paper, I classify all banks into three groups: banks with over $500 million in total assets are classified as large banks, banks with assets between $100 million and $500 million are classified as medium banks, and banks with under $100 million in assets are classified as small banks.

This classification is mainly based on the standard asset size categories that are used by the Federal Financial Institutions Examination Council (FFIEC), as specified in forms 031, 032, 033, and 034. The only difference is that FFIEC sets the asset cap for medium banks to $300 million. The reason for this change is to keep consistency with the Financial Modernization Act and many previous studies which use $500 million as the lower limit for large banks. To reduce the computation time for each of the bank subgroups and in

order to avoid heterogeneity biases associated with asset size, I further classify each of the three bank groups into several subgroups. Specifically, I use cutoffs at $20 million, $40 million, $60 million, and $80 million within the small bank group; $200 million, $300 million, and $400 million within the medium bank group; and $1 billion and $3 billion within the large bank group. Table 3.3 presents the twelve bank subgroups, together with their corresponding asset ranges at 2000 dollars and at 2005 dollars, as well as the number of banks in each subgroup.

It is to be noted, however, that this classification keeps the asset ranges fixed for the asset classes from year to year. These fixed asset ranges raise a serious question regarding the usefulness of the results when a long sample period, such as this study's sample period, is under examination. To deal with this problem, an approach similar to that laid out in the Financial Modernization Act (FMA) is used. In particular, I define a community bank is defined to be an institution with average total deposits over the proceeding three years of no more than $500 million. Each subsequent year, the asset cap is adjusted upward by the growth in the CPI (for all urban consumers) unadjusted for seasonal variation for the previous year (see Federal Registry, 2000). The cap for each year is published in the Federal Registry, early in the year, along with the inflation rate used in the adjustment. For example, the official asset cap for community banks in 2005 is adjusted to $567 million (see Federal Registry, 2005). Consistent with the approach, all the asset size cutoffs are set at 2000 constant dollars, and are adjusted upward by the growth in the CPI.

3.6 Empirical Results

I use the TOMLAB/NPSOL tool box with MATLAB to estimate the model using panel data for each of the twelve bank subgroups. For each subgroup, the model is estimated

under four different levels of constraints: with no constraints imposed; with only the curvature constraint imposed; with only the monotonicity constraint imposed; and with both the monotonicity and curvature constraints imposed. For each of the latter three cases, I impose curvature and/or monotonicity in a stepwise manner — first locally and then globally in case that regularity is not satisfied when local imposition is employed. Tables 3.4-3.15 summarize the results for each of the twelve subgroups in terms of parameter estimates, together with the percentages of monotonicity and curvature violations. Due to space limitations, I report only the intercept, u_0, the coefficients on the first order terms, $\boldsymbol{b}$, the coefficients on the second order terms, $\widetilde{u}_{0\alpha}$, and the coefficients on the time trend and $\widetilde{z}_1$ and $\widetilde{z}_2$ variables.

A parametric bootstrapping method is usually used in constrained optimization to obtain statistical inference for the estimated parameters $(\widehat{\beta})$ or nonlinear transformations of these parameters $(\phi(\widehat{\beta})$, i.e. elasticities or efficiency) (see Gallant and Golub, 1984). This involves the use of Monte Carlo methods, generating a sample from the distribution of the inequality constrained estimator $\widehat{\beta}$, large enough to obtain a reliable estimate of the sampling distributions of $\widehat{\beta}$ and $\phi(\widehat{\beta})$. However, the possibility of the use of Monte Carlo methods depends on the complexity of the problem in question. For a simple problem where the objective function is simple and the number of observations and constraints is small, like the traditional factor demand problem with 24 observation in Gallant and Golub (1984), a few hundred simulations are easily affordable in terms of computing time. Unfortunately, this is not the case with my problem. The complicated objective function and the large number of observations and constraints render the Monte Carlo method almost unaffordable. In particular, it takes at least one hour of CPU time on a Pentium 4 PC to run the optimization problem once. A 500 simulation would take at least 500 hours. When coupled with the number of bank subgroups, 12 in my case, it would take over 6,000 hours of CPU time to obtain standard errors for all the twelve groups. This

is certainly unafforable at present. Therefore, only point estimates are provided for the estimated parameters $(\widehat{\beta})$ in the following tables.

When neither monotonicity nor curvature is imposed (see the second column of each table), both monotonicity and curvature are violated for each of the twelve subgroups, with the percentage of curvature violations ranging from 1.4% to 34.7% across subgroups and that of monotonicity violations ranging from 0.1% to 46.5%. Since regularity is not achieved for all of the twelve bank subgroups, I first impose curvature alone on the parameters of the cost function. Clearly, the imposition of curvature alone reduces the percentage of curvature violations to zero for each of the twelve bank subgroups, however, it does not guarantee the satisfaction of monotonicity at every data point for all the twelve subgroups (see the third column of each table). In particular, the percentage of monotonicity violations still ranges from 3.2% to 46.6% across bank subgroups when only curvature is imposed. I further notice that, while the imposition of curvature alone reduces the percentage of violation for all of the twelve bank subgroups, it may also induce more violations of monotonicity that otherwise would not have occurred. Taking bank subgroup one (see Table 3.4) for example, the percentage of monotonicity violations is 1.3% when no constraints are imposed, but increases to 5.7% when curvature alone is imposed. This confirms Barnett's (2002, p. 202) argument that "imposition of curvature may increase the frequency of monotonicity violations. Hence equating curvature alone with regularity, as has become disturbingly common in this literature, does not seem to ... be justified."

Similarly, the imposition of monotonicity alone reduces the percentage of monotonicity violations to zero for each of the twelve bank subgroups, but it does not guarantee the satisfaction of curvature at every data point (see the fourth column of each table). In particular, the percentage of curvature violations still ranges from 5% to 20% across subgroups when only monotonicity is imposed. I also notice that the imposition of

monotonicity alone may induce more violations of monotonicity that otherwise would not have occurred (see for example bank subgroup 1). This further confirms the argument of Barnett and Pasupathy (2003, p. 135) that "regularity requires satisfaction of both curvature and monotonicity conditions. Without both satisfied, the second order conditions for optimizing behavior fail and duality theory fails." I thus followed the procedures discussed in Sections 3 and 4 and imposed both curvature and monotonicity on the parameters of the Fourier cost function for each of the twelve bank subgroups. As expected, regularity is satisfied at every data point after curvature and monotonicity are globally imposed (see the fourth column in each of Tables 3.4-3.15).

A common practice in this literature is to derive cost efficiency measures from cost functions without theoretical regularity imposed. While permitting a parameterized function to depart from the neoclassical function space is usually fit-improving (this can be seen from the decrease in the log likelihood values as constraints are imposed), it also causes the hypothetical best practice firm not to be fully efficient at those data points where curvature and/or monotonicity are violated. In particular, the violation of curvature at a data point (p_{jt}, y_{jt}) implies that the quantities of some outputs increase as their corresponding prices increase (holding other things constant); and the violation of monotonicity at that data point implies the quantities of some outputs decrease as total cost increases (holding other things constant). Both of these two cases mean that the best practice firm is not minimizing its cost at (p_{jt}, y_{jt}). Therefore, cost efficiency, which is supposed to be measured relative to a cost-minimizing best practice bank, is not accurate for all the twelve bank subgroups when monotonicity and curvature are not imposed. In fact, I find that the difference in the eight-year mean efficiency between the unconstrained models and their corresponding curvature and monotonicity constrained versions range from -0.73% to 0.92% (see Table 3.16).[1] Hence, the failure to impose

[1]The eight-year mean efficiency for a subgroup is obtained by first averaging over eight years (from

monotonicity and curvature can produce misleading estimates of cost efficiency.

Another issue of particular interest is whether failure to impose theoretical regularity affects the ranking of individual banks in terms of cost efficiency. I calculate the Spearman rank correlation coefficient between unconstrained models and their corresponding (curvature and monotonicity) constrained versions, using the following formula

$$R = 1 - \frac{6\sum_{j=1}^{n_k}(\text{Rank}_{j1} - \text{Rank}_{j2})^2}{n_k(n_k^2 - 1)}, \tag{3.34}$$

where n_k is the number of banks in the subgroup, Rank_{j1} is the rank of bank i based on the constrained version of the model, and Rank_{j2} is the rank of the same bank based on the unconstrained version of the model.[2] If $R = -1$, there is perfect negative correlation; if $R = 1$, there is perfect positive correlation; and if $R = 0$, there is no correlation. As can be seen in Table 3.17, all of the twelve rank correlation coefficients are different than 1, indicating that the ranking of banks in term of cost efficiency changes due to the imposition of theoretical regularity.

Roughly speaking, the rank correlation coefficient between unconstrained and (theoretical regularity) constrained models is negatively related to the percentage of monotonicity and curvature violations. For example, bank subgroup 1, which has the lowest percentage of monotonicity violations (0.1%) and the lowest percentage of curvature violations (1.4%), has the highest rank correlation coefficient (0.9997); bank subgroup 12, which has

1998 to 2005), and then averaging over all banks in this subgroup, as follows

$$\text{8-year MEFF} = \frac{1}{n_k}\sum_{j=1}^{n_k}\left(\frac{1}{T}\sum_{t=1}^{T} CE_{jt}\right)$$

where $T = 8$ and n_k is the number of banks in the subgroup.

[2]Here, I use the time invariance cost efficiency for each bank, that is,

$$\text{EFF}_j = \frac{1}{T}\sum_{t=1}^{T} CE_{jt}.$$

the highest percentage of monotonicity violations (46.5%) and the highest percentage of curvature violations (34.1%), has the lowest rank correlation coefficient (0.8684). Hence, I alert researchers to the potential problems caused by failure to check for and impose (if necessary) theoretical regularity.

3.6.1 Cost Efficiency and Productivity of U.S. Banks

I now turn to the discussion of cost efficiencies by asset size class, reported in Table 3.18. Clearly, the mean efficiency for each of the twelve subgroups ranges from 82.19% to 91.78%, implying that about 8 % to 18 % of incurred costs over the sample period can be attributed to cost inefficiency relative to the best cost-practice banks. These results are similar to earlier estimates that examined commercial banks. Berger and Humphrey (1997), for example, report mean cost efficiency of 84% with a standard deviation 6% across 50 studies of US banks using parametric frontier techniques. Likewise, Berger and Mester (1997) report average cost efficiency of 87% using a large data set of almost 6,000 US commercial banks that were in continuous operation over the six-year period from 1990 to 1995.

There are several findings that emerge from Table 3.18. First, the largest two subgroups are less efficient than the other ten subgroups. In particular, the very largest subgroup (with assets greater than $3,000 million) is about 5.6% less efficient than the second largest subgroup and 7.8% less efficient than the third largest subgroup. The same subgroup is 6.3% to 9.6% less efficient than the medium sized and small banks. The second largest subgroup (with assets between $1,000 million and $3,000 million) is 1.2% less efficient than the third largest subgroup, and ranges from 0.9% to 3.9% less efficient than medium sized and small bank subgroups. Second, in general cost efficiency falls with bank size for banks with assets above $100 million except for subgroup 3 (with assets between $500 million and $1 billion) and subgroup 5 (with assets between $300

million and $400 million). However, cost efficiency increases with bank size for banks with assets below $200 million except for subgroup 9 (with assets between $60 million and $80 million) and subgroup 10 (with assets between $40 million and $60 million). These findings are partially consistent with Kaparakis *et al.* (1994) who applied a translog cost function to a large data set of almost 5,548 commercial banks in the United States. In particular, Kaparakis *et al.* (1994) also finds that banks with assets greater than $1,000 million are less efficient than smaller banks. However, he finds that average efficiency increases with bank size for banks with assets less than $500 million.

I am also interested in the time patterns of cost efficiency of the different bank subgroups, plotted in Figure 3.1. Several conclusions emerge. First, all the bank subgroups experienced a drastic decline in cost efficiency over the period from 1998 to 2004, and then showed an improvement in cost efficiency in 2005. For example, the cost efficiency of the largest bank subgroup (with assets greater than $3,000 million) declined from 94.79% in 1998 to 73.65% in 2004, and then resurged a little bit to 73.91% in 2005. The most efficient subgroup with assets between $100 million and $200 million shows a decline in cost efficiency from almost full efficiency in 1998 to 80.12% in 2004, and then shows a rebound to 84.47% in 2005. Second, the largest bank subgroup is consistently less efficient than the other bank subgroups. Further, the gap in cost efficiency between the largest bank subgroups and the other bank subgroups has increased. For example, the largest banks were 3.34% less efficient than the second largest bank subgroup in 1998, but 7.24% less efficient than the second largest bank subgroup in 2005.

The drastic decline in cost efficiency for all asset size classes during the first seven years of my sample period can be partially justified by the failure of banks to adjust to the rapid technological change of the best practice cost frontier. Figure 3.2 plots the technological change of the best practice cost frontier for all the size classes. Clearly, all twelve asset size classes have shown rapid technological change, with large banks being

more favored by the technological change. In particular, the largest size subgroup (with assets greater than $3,000 million) has seen the fastest technological change of around 6.71% per year; and even the second smallest size subgroup (with assets between $20 million and $40 million) — which has shown the lowest technological change — has also seen a technological change of around 1% per year. Rapid technological change, which makes feasible the production of given levels of outputs with fewer inputs (or, equivalently, the production of more outputs with given levels of inputs), could result in lower average bank efficiency, even if banks became increasingly productive over time. This can be clearly seen from equation (3.5).

The second reason may lie in unmeasured improvements in service quality and variety. Banks have provided an improved array of services (e.g., mutual funds, derivatives, on-line services, etc.) that increased bank costs, but at the same time were able to raise revenues to more than cover these costs. This is consistent with a strong improvement in profitability over the sample period. Another partial explanation for the decline in cost efficiency for the very large banks (those with assets greater than 1 billion) is that many of them have been engaged in geographical diversification and product diversification. The passage of the Riegle-Neal Interstate Banking and Branching Efficiency Act of 1994 undoubtedly helped spur large banks to spread across state lines and to grow. This development helped create large, geographically diversified branch networks that stretch across large regions and even coast-to-coast. The Gramm-Leach-Bliley Financial Services Modernization Act of 1999 allowed the largest banking organizations to engage in a wide variety of financial services, acquiring new sources of noninterest income and further diversifying their earnings. While these geographical diversification and product diversification have increased the large banks' profits, they also greatly increase their costs.

Finally, one thing that needs to be clarified here is that a lower cost efficiency does

not mean a lower productivity growth. In order to illustrate, I calculate the average productivity growth for each bank subgroup over the sample period. Within a cost frontier context, productivity growth can be decomposed into four components: a technological change term, a technical efficiency change term, an input allocative efficiency change term, and a scale effect term — see Kumbhakar and Lovell (2003) for more details. For simplicity, let's ignore the last term and call productivity growth, which is now composed of only the first three terms, 'net' productivity growth (NTFPG). Following Kumbhakar and Lovell (2003), I then express the net productivity change as

$$\text{NTFPG} = -\frac{\partial \ln f_{it}}{\partial t} - \frac{\partial u_{it}}{\partial t} \tag{3.35}$$

where the first term is the technological change of the best practice cost frontier and the second term is the change in cost efficiency, including both technical and allocative efficiency changes. The average annual net productivity growth for each of the twelve subgroups is plotted in Figure 3.3. Generally speaking, the net productivity growth rate increases with asset size, with the largest four bank subgroups (with assets greater than \$400 million) experiencing significant productivity gains (NTFPG $> 1\%$) and the smallest eight subgroups (with assets less than \$400 million) experiencing insignificant productivity gains (NTFPG $< 1\%$) or productivity losses (NTFPG < 0). In particular, the largest size subgroup, which has the lowest cost efficiency, shows the fastest average annual net productivity growth of 3.3% whereas, subgroup 7, which has the highest cost efficiency, shows a moderate average annual net productivity growth of 0.4%. This finding is also consistent with the view expressed by Berger (2003) and Bernanke (2006) and others that technological advances have favored larger banks at the expense of small lenders. However, these productivity gains by larger banks are mainly due to technological advances rather than cost efficiency gains.

3.7 Conclusion

The estimation of stochastic cost frontier is popular in the analysis of bank efficiency. However, the theoretical regularity conditions (especially those of monotonicity and curvature) required by neoclassical microeconomic theory have been widely ignored in the literature. In this paper, and for the first time in this literature, I use the globally flexible Fourier functional form, as originally proposed by Gallant (1982), and estimation procedures suggested by Gallant and Golub (1984) to impose the theoretical regularity conditions on the Fourier cost function. Hence, I provide estimates of bank efficiency in the United States using (for the first time) parameter estimates that are consistent with full regularity.

I find that failure to incorporate monotonicity and curvature into the estimation will result in mismeasured magnitudes of cost efficiency and also misleading bank rankings in terms of cost efficiency. Regarding cost efficiencies from my theoretical regularity constrained models, I find that the largest two subgroups are less efficient than the other subgroups. I also find that all twelve asset size classes show a decline in cost efficiency from 1998 to 2004, and then see a slight improvement in 2005. This decline in cost efficiency can be the result of adjustments to fast technical progress or unmeasured improvements in service quality and variety. For the very large banks, the decline in cost efficiency can be a result of their engagement in geographical diversification and product diversification after deregulation. Further, I find that the largest four bank subgroups (with assets greater than \$400 million) experienced significant productivity gains (NTFPG $>$ 1%) and the smallest eight subgroups (with assets less than \$400 million) experienced insignificant productivity gains (NTFPG $<$ 1%) or productivity losses.

In estimating bank efficiency and productivity in the United States, I have also high-

lighted the challenge inherent with achieving economic regularity and the need for economic theory to inform econometric research. Incorporating restrictions from economic theory seems to be gaining popularity as there are also numerous recent papers that estimate stochastic dynamic general equilibrium models using economic restrictions (see Aliprantis *et al.*, 2007). With the focus on economic theory, however, I have ignored econometric regularity. In particular, I have ignored unit root and cointegration issues, because the combination of nonstationary data and nonlinear estimation in large models like the ones in this paper is an extremely difficult problem. Dealing with these difficult issues is an area for potentially productive future research.

3.8 Appendix

Let $\widetilde{C}(\boldsymbol{p},\boldsymbol{y})$ denote the total cost function which is expressed in logarithmic form (i.e. translog or Fourier cost functions). By Shephard's lemma, $\widetilde{X}_i(\boldsymbol{p},\boldsymbol{y}) = \partial\widetilde{C}(\boldsymbol{p},\boldsymbol{y})/\partial p_i$, where $\widetilde{X}_i$ is the demand for input i. Also, let $s_i(\boldsymbol{p},\boldsymbol{y})$ denote the cost share for input i.

The element of the ith row and jth column of the Hessian matrix of $\widetilde{C}(\boldsymbol{p},\boldsymbol{y})$ can be derived as follows

$$\boldsymbol{H}_{ij} = \frac{\partial \ln \widetilde{X}_i(\boldsymbol{p},\boldsymbol{y})}{\partial \ln p_j}\frac{\widetilde{X}_i(\boldsymbol{p},\boldsymbol{y})}{p_j}$$

$$= \frac{\partial \ln\left[\widetilde{C}(\boldsymbol{p},\boldsymbol{y})\, p_i^{-1} s_i\right]}{\partial \ln p_j}\frac{\widetilde{X}_i(\boldsymbol{p},\boldsymbol{y})}{p_j}$$

$$= \left[\frac{\partial \ln s_i}{\partial \ln p_j} + \frac{\partial \ln \widetilde{C}(\boldsymbol{p},\boldsymbol{y})}{\partial \ln p_j} + \frac{\partial \ln p_i^{-1}}{\partial \ln p_j}\right]\frac{\widetilde{X}_i(\boldsymbol{p},\boldsymbol{y})}{p_j}$$

$$= \left(s_i^{-1}\frac{\partial s_i}{\partial \ln p_j} + s_j - \delta_{ij}\right)\frac{\widetilde{X}_i(\boldsymbol{p},\boldsymbol{y})}{p_j}$$

$$= \left(s_i^{-1}\nabla_{\ln p_j} s_i + s_j - \delta_{ij}\right)\frac{\widetilde{X}_i(\boldsymbol{p},\boldsymbol{y})}{p_j}, \qquad i,j = 1,\cdots,N,$$

where $\delta_{ij} = 1$ if $i = j$ and zero otherwise.

TABLE 3.1. A SUMMARY OF FLEXIBLE FUNCTIONAL FORMS
ESTIMATION OF COST EFFICIENCY OF US BANKS

Study	Model used	True Fourier	Curvature imposed
Ferrier and Lovell (1990)	Translog		No
Berger and Humphrey (1991)	Translog		No
Berger (1993)	Translog		No
Kaparakis *et al.* (1994)	Translog		No
Berger and Mester (1997)	Translog + Fourier trigonometric terms	No	No
Berger *et al.* (1997)	Translog + Fourier trigonometric terms	No	No
Peristiani (1997)	Translog		No
DeYoung (1997)	Translog		No
Mester (1997)	Translog		No
DeYoung *et al.* (1998)	Translog + Fourier trigonometric terms	No	No
Stiroh (2000)	Translog		No
Clark and Siems (2002)	Translog		No
Berger and Mester (2003)	Translog + Fourier trigonometric terms	No	No

Note: Some studies employed both cost and profit frontiers.

TABLE 3.2. ELEMENTARY MULTI-INDEXES

α	1	2	3	4	5	6	7	8	9	10	11
l_1	1	1	0	0	0	0	1	1	0	1	1
l_2	-1	0	1	0	0	0	-1	0	1	-1	0
l_3	0	-1	-1	0	0	0	0	-1	-1	0	-1
q_1	0	0	0	1	1	0	1	0	0	0	1
q_2	0	0	0	-1	0	1	0	1	0	0	0
q_3	0	0	0	0	-1	-1	0	0	1	1	0
$\lvert k_\alpha \rvert^*$	2	2	2	2	2	2	3	3	3	3	3

α	12	13	14	15	16	17	18	19	20	21	22
l_1	0	1	1	0	1	1	1	0	0	0	0
l_2	1	-1	0	1	-1	-1	-1	1	1	0	0
l_3	-1	0	-1	-1	0	0	0	-1	-1	0	0
q_1	1	0	0	0	-1	0	0	0	0	1	1
q_2	0	0	0	1	0	-1	0	-1	0	-2	0
q_3	0	1	1	0	0	0	-1	0	-1	0	2
$\lvert k_\alpha \rvert^*$	3	3	3	3	3	3	3	3	3	3	3

α	23	24	25	26	27	28	29	30	31	32
l_1	0	0	0	0	0	0	0	0	0	0
l_2	0	0	0	0	0	0	0	0	0	0
l_3	0	0	0	0	0	0	0	0	0	0
q_1	1	1	2	2	0	0	0	0	0	0
q_2	1	1	0	1	1	1	2	2	1	0
q_3	-1	0	1	0	-2	1	-1	1	0	1
$\lvert k_\alpha \rvert^*$	3	2	3	3	3	2	3	3	1	1

TABLE 3.3. BANK ASSET SIZE CLASSES

Bank groups	Asset size (in millions of 2000 dollars)	Asset size (in millions of 2005 dollars)	Number of banks	Share of banks
Large banks				
Group 1	assets $\geq 3,000$	assets $\geq 3,402$	141	2.3%
Group 2	$1,000 \leq$ assets $< 3,000$	$1134 \leq$ assets $< 3,000$	218	3.6%
Group 3	$500 \leq$ assets $< 1,000$	$567 \leq$ assets < 1134	381	6.3%
Medium banks				
Group 4	$400 \leq$ assets < 500	$453.6 \leq$ assets < 567	201	3.3%
Group 5	$300 \leq$ assets < 400	$340.2 \leq$ assets < 453.6	321	5.3%
Group 6	$200 \leq$ assets < 300	$226.8 \leq$assets ≤ 340.2	602	10.0%
Group 7	$100 \leq$ assets < 200	$113.4 \leq$assets ≤ 226.8	1262	21.0%
Small banks				
Group 8	$80 \leq$ assets < 100	$90.72 \leq$assets ≤ 113.4	477	7.9%
Group 9	$60 \leq$ assets < 80	$68.04 \leq$assets ≤ 90.72	597	9.9%
Group 10	$40 \leq$ assets < 60	$45.36 \leq$assets ≤ 68.04	669	11.1%
Group 11	$20 \leq$ assets < 40	$22.68 \leq$assets ≤ 45.36	813	13.5%
Group 12	assets ≤ 20	assets ≤ 22.68	328	5.5%
Total			6010	100%

TABLE 3.4. PARAMETER ESTIMATES FOR GROUP 1

Parameter	Unconstrained	Curvature only	Monotonicity only	Both Monotonicity and Curvature
u_0	7.3739	6.6017	7.7557	7.3661
b_1	0.6699	0.6858	0.9193	0.6671
b_2	0.0818	0.0949	−0.0971	0.0824
b_3	0.1557	0.1188	0.1955	0.1353
b_4	−0.4922	−0.5293	−0.0912	−0.2805
b_5	0.5639	0.6936	0.5790	0.3820
u_{01}	0.3094	0.2655	0.0954	0.3084
u_{02}	−0.2870	−0.2473	−0.2998	−0.2855
u_{03}	0.3850	0.3496	0.2252	0.3825
u_{04}	0.0021	−0.0460	−0.0041	0.0089
u_{05}	−0.0701	−0.0773	−0.0645	−0.0712
u_{06}	0.5085	0.2820	0.5583	0.3658
u_{07}	0.0112	0.0097	0.1887	0.0163
u_{08}	0.0837	0.0816	0.1284	0.0911
u_{09}	−0.0450	−0.0436	−0.0559	−0.0520
u_{010}	−0.0577	−0.0657	−0.0733	−0.0640
u_{011}	−0.1521	−0.1403	−0.4515	−0.1577
u_{012}	−0.3178	−0.2892	−0.2849	−0.3127
u_{013}	−0.3056	−0.2670	−0.2984	−0.3022
u_{014}	0.4326	0.3922	0.4842	0.4286
u_{015}	0.0062	0.0068	0.1567	0.0105
t	−0.1080	−0.1050	−0.1021	−0.1076
t^2	0.0045	0.0043	0.0033	0.0045
Nontrad	−0.0910	−0.0622	−0.1062	−0.0886
Nontrad^2	0.0136	0.0093	0.0171	0.0136
Equity	0.1726	0.2953	0.1996	0.1820
Equity^2	−0.0051	−0.0089	−0.0056	−0.0054
σ_s^2	0.1719	0.1690	0.1862	0.1716
γ	0.9473	0.9486	0.9543	0.9472
η_1	0.0574	0.0438	0.0309	0.0560
η_2	0.0449	0.0403	0.0436	0.0446
Log likelihood	611.5	586.9	593.4	572.1
Curvature violations	1.4%	0	37.5%	0
Monotonicity violations	0.1%	5.7%	0	0
Mean efficiency	0.8171			0.8219

TABLE 3.5. PARAMETER ESTIMATES FOR GROUP 2

Parameter	Unconstrained	Curvature only	Monotonicity only	Both Monotonicity and Curvature
u_0	12.8594	12.3481	13.2671	12.9243
b_1	0.8502	0.8710	0.8531	0.8048
b_2	0.0493	0.0484	0.0427	0.1095
b_3	0.1119	0.0972	0.1144	0.0718
b_4	−0.0299	−0.0214	0.0503	0.1488
b_5	0.3419	0.3793	0.3318	0.3613
u_{01}	−0.0274	−0.0434	−0.0449	−0.0734
u_{02}	−0.0518	−0.0229	−0.0408	0.0338
u_{03}	0.0174	0.0213	−0.0002	−0.0169
u_{04}	−0.0314	−0.0265	−0.0345	−0.0147
u_{05}	−0.0050	−0.0042	0.0042	−0.0024
u_{06}	−0.0540	−0.0515	−0.0603	−0.0633
u_{07}	0.0013	−0.0060	0.0207	0.0250
u_{08}	0.0261	0.0318	0.0242	0.0248
u_{09}	−0.0178	−0.0203	−0.0163	−0.0204
u_{010}	−0.0262	−0.0273	−0.0240	−0.0262
u_{011}	0.0345	0.0492	0.0122	0.0144
u_{012}	0.0316	0.0456	0.0263	0.0473
u_{013}	0.0133	0.0295	0.0082	0.0326
u_{014}	−0.0268	−0.0489	−0.0210	−0.0535
u_{015}	−0.0450	−0.0501	−0.0232	−0.0162
t	−0.0812	−0.0849	−0.0822	−0.0896
t^2	0.0043	0.0048	0.0044	0.0052
Nontrad	0.0371	0.0371	0.0389	0.0370
Nontrad^2	−0.0026	−0.0026	−0.0031	−0.0022
Equity	−0.6221	−0.6221	−0.8316	−0.8418
Equity^2	0.0299	0.0299	0.0389	0.0389
σ_s^2	0.0769	0.0769	0.0771	0.0745
γ	0.8851	0.8851	0.8866	0.8794
η_1	0.1523	0.1523	0.1819	0.1282
η_2	0.0794	0.0794	0.0923	0.0682
Log likelihood	934.6	926.9	933.2	919.4
Curvature violations	12.6%	0	13.2%	0
Monotonicity violations	6.0%	6.2%	0	0
Mean efficiency	0.8852	–	–	0.8820

TABLE 3.6. PARAMETER ESTIMATES FOR GROUP 3

Parameter	Unconstrained	Curvature only	Monotonicity only	Both Monotonicity and Curvature
u_0	11.6880	11.9781	11.7559	11.2504
b_1	0.6484	0.6503	0.6456	0.6142
b_2	0.0696	0.0798	0.0657	0.0936
b_3	0.0552	0.0948	0.0782	0.0956
b_4	0.2464	0.2113	0.1904	0.7056
b_5	0.7217	0.7357	0.5611	0.5993
u_{01}	−0.1411	−0.1574	−0.0803	−0.2350
u_{02}	0.1323	0.1550	0.0685	0.2363
u_{03}	−0.1258	−0.1401	−0.0657	−0.2247
u_{04}	−0.0275	−0.0332	−0.0127	−0.0135
u_{05}	−0.0181	−0.0175	−0.0104	−0.0120
u_{06}	−0.0316	−0.0207	−0.0222	−0.0234
u_{07}	−0.0395	−0.0430	−0.0539	0.0909
u_{08}	0.0511	0.0374	0.0379	0.0340
u_{09}	−0.0344	−0.0225	−0.0266	−0.0220
u_{010}	−0.0428	−0.0335	−0.0318	−0.0289
u_{011}	0.0318	0.0406	0.0486	−0.1009
u_{012}	0.1832	0.1929	0.1292	0.1390
u_{013}	0.1852	0.1964	0.1283	0.1394
u_{014}	−0.1932	−0.2029	−0.1406	−0.1495
u_{015}	−0.0228	−0.0303	−0.0366	0.1070
t	−0.0953	−0.0954	−0.0878	−0.0869
t^2	0.0045	0.0045	0.0039	0.0038
Nontrad	−0.0249	−0.0288	−0.0581	−0.0600
Nontrad^2	0.0063	0.0068	0.0104	0.0105
Equity	−1.0046	−1.0320	−0.9691	−1.0279
Equity^2	0.0450	0.0462	0.0429	0.0456
σ_s^2	0.0679	0.0673	0.0670	0.0664
γ	0.8095	0.8073	0.8022	0.7996
η_1	0.1468	0.1687	0.1791	0.1917
η_2	0.0834	0.0896	0.0998	0.1047
Log likelihood	1230.1	1227.5	1212.8	1211.4
Curvature violations	3.9%	0	1.9%	0
Monotonicity violations	14.2%	13.0%	0	0
Mean efficiency	0.8964	–	–	0.9038

TABLE 3.7. PARAMETER ESTIMATES FOR GROUP 4

Parameter	Unconstrained	Curvature only	Monotonicity only	Both Monotonicity and Curvature
u_0	16.3202	13.4177	13.7549	13.9571
b_1	0.8400	0.8932	0.8297	0.8821
b_2	0.0051	−0.0047	0.0699	0.0065
b_3	−0.1996	−0.1471	0.1099	−0.0925
b_4	0.6406	0.7147	0.8458	0.7259
b_5	−0.3814	−0.4059	0.0530	−0.3640
u_{01}	0.1010	0.0541	−0.1530	0.0817
u_{02}	−0.1280	−0.1116	0.1309	−0.1301
u_{03}	0.1540	0.1168	−0.0963	0.1366
u_{04}	0.0198	0.0152	−0.0123	0.0235
u_{05}	−0.0175	−0.0204	−0.0096	0.0105
u_{06}	−0.1723	−0.1521	−0.0834	−0.0881
u_{07}	0.1640	0.1770	0.2031	0.1396
u_{08}	0.1137	0.1006	0.0421	0.0993
u_{09}	−0.1032	−0.0920	−0.0285	−0.0784
u_{010}	−0.1216	−0.1030	−0.0380	−0.0845
u_{011}	−0.1620	−0.1587	−0.1988	−0.1310
u_{012}	−0.1940	−0.1791	−0.0568	−0.1841
u_{013}	−0.2006	−0.1820	−0.0645	−0.1827
u_{014}	0.2257	0.1921	0.0778	0.1969
u_{015}	0.1479	0.1590	0.1804	0.1266
t	−0.0676	−0.0649	−0.0678	−0.0600
t^2	0.0020	0.0020	0.0021	0.0026
Nontrad	0.0037	−0.0001	−0.0221	−0.0126
Nontrad^2	0.0044	0.0053	0.0071	0.0053
Equity	−1.6658	−1.1140	−1.4599	−1.3593
Equity^2	0.0828	0.0568	0.0727	0.0671
σ_s^2	0.0798	0.0777	0.0792	0.0908
γ	0.9076	0.9021	0.9048	0.8900
η_1	−0.0478	−0.0074	−0.0210	0.1715
η_2	0.0449	0.0604	0.0532	0.0678
Log likelihood	985.9	972.6	975.5	885.0
Curvature violations	27.4%	0	26%	0
Monotonicity violations	8.9%	5.3%	0	0
Mean efficiency	0.8948	–	–	0.8856

TABLE 3.8. PARAMETER ESTIMATES FOR GROUP 5

Parameter	Unconstrained	Curvature only	Monotonicity only	Both Monotonicity and Curvature
u_0	1.1076	1.7548	2.1774	1.1061
b_1	0.8642	0.8134	0.8777	0.8547
b_2	0.1917	0.1211	0.1995	0.1784
b_3	0.3425	0.1699	0.0957	0.3717
b_4	1.2590	1.2956	0.7711	1.2594
b_5	0.3305	0.2992	0.1345	0.3497
u_{01}	−0.5118	−0.4297	−0.2394	−0.5000
u_{02}	0.4670	0.3949	0.1955	0.4597
u_{03}	−0.4237	−0.3655	−0.1489	−0.4161
u_{04}	−0.0886	−0.0785	−0.0432	−0.0818
u_{05}	0.0153	0.0222	0.0154	0.0191
u_{06}	−0.0727	−0.0928	−0.0430	−0.0337
u_{07}	0.3283	0.3347	0.1879	0.3356
u_{08}	−0.0537	0.0054	0.0190	−0.0405
u_{09}	0.0478	−0.0037	−0.0208	0.0431
u_{010}	0.0563	0.0017	−0.0118	0.0481
u_{011}	−0.3064	0.0017	−0.1673	−0.3173
u_{012}	0.0579	0.0394	−0.0043	0.0578
u_{013}	0.0730	0.0486	0.0080	0.0578
u_{014}	−0.0795	−0.0509	−0.0106	−0.0649
u_{015}	0.3066	0.3250	0.1610	−0.0649
t	−0.0704	−0.0708	−0.0735	−0.0703
t^2	0.0041	−0.0884	0.0043	0.0038
Nontrad	−0.1196	−0.0884	−0.1454	−0.1237
Nontrad^2	0.0196	0.0158	0.0223	0.0203
Equity	0.7371	0.6224	0.7959	0.7308
Equity^2	−0.0402	−0.0343	−0.0433	−0.0401
σ_s^2	0.0586	0.0609	0.0579	0.0605
γ	0.8155	0.8179	0.8115	0.8132
η_1	0.2249	0.2020	0.2393	0.2228
η_2	0.0967	0.0916	0.1004	0.0996
Log likelihood	1246.2	1224.6	1238.3	1210.8
Curvature violations	23.4%	0	20.9%	0
Monotonicity violations	5.1%	3.2%	0	0
Mean efficiency	0.8975	–	–	0.8961

TABLE 3.9. PARAMETER ESTIMATES FOR GROUP 6

Parameter	Unconstrained	Curvature only	Monotonicity only	Both Monotonicity and Curvature
u_0	7.7279	6.7066	7.3981	6.3933
b_1	0.7662	0.6810	0.8205	0.7750
b_2	0.0372	0.1106	−0.0014	0.0679
b_3	0.1921	0.1692	0.1683	−0.0108
b_4	0.2053	0.3839	0.3832	−0.6297
b_5	0.1310	0.1047	0.2217	0.1336
u_{01}	−0.0341	−0.0749	−0.0900	0.2667
u_{02}	0.0112	0.0898	0.0497	−0.2775
u_{03}	−0.0282	−0.0642	−0.0832	0.2919
u_{04}	−0.0406	−0.0347	−0.0183	0.0069
u_{05}	−0.0100	−0.0098	−0.0073	0.0003
u_{06}	−0.0758	−0.0744	−0.0384	−0.0234
u_{07}	0.0147	0.0637	0.0455	−0.2440
u_{08}	−0.0200	−0.0091	0.0066	0.0493
u_{09}	0.0222	0.0145	0.0082	−0.0413
u_{010}	0.0077	0.0021	−0.0023	−0.0456
u_{011}	−0.0180	−0.0705	−0.0451	0.2499
u_{012}	−0.0311	−0.0299	0.0067	−0.0138
u_{013}	−0.0318	−0.0337	0.0049	−0.0179
u_{014}	0.0505	0.0419	0.0103	0.0264
u_{015}	0.0161	0.0650	0.0453	−0.2500
t	−0.0822	−0.0840	−0.0841	−0.0819
t^2	0.0047	0.0050	0.0049	0.0049
Nontrad	0.0484	0.0531	0.0427	0.0413
Nontrad^2	−0.0021	−0.0028	−0.0013	−0.0016
Equity	−0.2235	−0.0551	−0.2071	0.3698
Equity^2	0.0123	0.0037	0.0112	−0.0187
σ_s^2	0.0669	0.0665	0.0691	0.0719
γ	0.8736	0.8697	0.8766	0.8762
η_1	0.0585	0.0735	0.0647	0.0938
η_2	0.0552	0.0591	0.0561	0.0686
Log likelihood	1579.3	1559.5	1567.6	1528.5
Curvature violations	17.4%	0	17.4%	0
Monotonicity violations	8.1%	5.8%	0	0
Mean efficiency	0.8878	–	–	0.8890

TABLE 3.10. PARAMETER ESTIMATES FOR GROUP 7

Parameter	Unconstrained	Curvature only	Monotonicity only	Both Monotonicity and Curvature
u_0	6.0068	9.2679	5.6587	9.2110
b_1	0.6884	0.7039	0.7227	0.7683
b_2	0.1302	0.1014	0.1308	0.1054
b_3	0.2685	0.1194	0.2281	0.1085
b_4	0.8774	−0.7582	1.0423	−0.6621
b_5	0.1013	−0.1570	0.1091	−0.0876
u_{01}	−0.2710	0.2711	−0.2858	0.2395
u_{02}	0.2489	−0.2925	0.2555	−0.2766
u_{03}	−0.2164	0.3174	−0.2247	0.2989
u_{04}	−0.0634	−0.0529	−0.0388	−0.0269
u_{05}	−0.0215	−0.0148	−0.0091	−0.0010
u_{06}	−0.0034	−0.0145	0.0007	0.0046
u_{07}	0.1988	−0.2022	0.2111	−0.2064
u_{08}	−0.0233	0.0189	−0.0190	0.0130
u_{09}	0.0310	−0.0126	0.0231	−0.0097
u_{010}	0.0091	−0.0310	0.0098	−0.0194
u_{011}	−0.2084	0.1982	−0.2181	0.2070
u_{012}	−0.0124	−0.0966	−0.0087	−0.0749
u_{013}	−0.0075	−0.0910	−0.0026	−0.0662
u_{014}	0.0137	0.0986	0.0097	0.0772
u_{015}	0.1954	−0.2077	0.2077	−0.2154
t	−0.0577	−0.0572	−0.0593	−0.0584
t^2	0.0031	0.0030	0.0032	0.0031
Nontrad	0.0717	0.0756	0.0653	0.0711
Nontrad^2	−0.0021	−0.0028	−0.0017	−0.0025
Equity	−0.2653	−0.2342	−0.2424	−0.2603
Equity^2	0.0125	−0.2342	0.0108	0.0117
σ_s^2	0.0596	0.0605	0.0606	0.0613
γ	0.7285	0.7289	0.7309	0.7300
η_1	0.5445	0.5318	0.5369	0.5202
η_2	0.2418	0.2375	0.2391	0.2347
Log likelihood	2165.9	2152.7	2156.9	2143.1
Curvature violations	10.8%	0	8.0%	0
Monotonicity violations	22.1%	20.3%	0	0
Mean efficiency	0.9181	–	–	0.9178

TABLE 3.11. PARAMETER ESTIMATES FOR GROUP 8

Parameter	Unconstrained	Curvature only	Monotonicity only	Both Monotonicity and Curvature
u_0	5.0832	5.5241	5.4229	5.6218
b_1	0.7878	0.8074	0.7354	0.7780
b_2	0.1347	0.1106	0.1638	0.1307
b_3	0.1606	0.1013	0.1548	0.1355
b_4	0.5805	0.6484	0.5695	0.5083
b_5	0.1429	0.2829	0.2858	0.3314
u_{01}	−0.1841	−0.2200	−0.2056	−0.1941
u_{02}	0.1418	0.1735	0.1757	0.1544
u_{03}	−0.1301	−0.1600	−0.1569	−0.1377
u_{04}	−0.0495	−0.0415	−0.0420	−0.0389
u_{05}	0.0096	0.0153	0.0038	0.0060
u_{06}	−0.0187	−0.0157	−0.0128	−0.0108
u_{07}	0.1069	0.1312	0.1037	0.0922
u_{08}	−0.0034	0.0038	−0.0035	−0.0020
u_{09}	0.0166	0.0032	0.0098	0.0061
u_{010}	0.0062	−0.0065	0.0040	−0.0006
u_{011}	−0.1040	−0.1284	−0.1049	−0.0903
u_{012}	−0.0059	0.0235	0.0285	0.0356
u_{013}	0.0173	0.0443	0.0458	0.0530
u_{014}	−0.0106	−0.0295	−0.0446	−0.0436
u_{015}	0.1031	0.1244	0.1027	0.0864
t	−0.0502	−0.0500	−0.0487	−0.0486
t^2	0.0021	0.0022	0.0019	0.0020
Nontrad	0.0894	0.0879	0.0793	0.0775
Nontrad^2	−0.0088	−0.0085	−0.0076	−0.0073
Equity	0.2311	0.0942	0.0885	0.0754
Equity^2	−0.0109	−0.0034	−0.0035	−0.0026
σ_s^2	0.0690	0.0684	0.0692	0.0686
γ	0.8195	0.8165	0.8188	0.8161
η_1	0.3468	0.3527	0.3441	0.3507
η_2	0.1807	0.1811	0.1818	0.1829
Log likelihood	1307.2	1302.2	1303.3	1298.7
Curvature violations	19.2%	0	16.6%	0
Monotonicity violations	8.4%	6.9%	0	0
Mean efficiency	0.9145	–	–	0.9129

TABLE 3.12. PARAMETER ESTIMATES FOR GROUP 9

Parameter	Unconstrained	Curvature only	Monotonicity only	Both Monotonicity and Curvature
u_0	3.6871	4.9381	4.0945	5.3413
b_1	0.6942	0.6576	0.7199	0.7237
b_2	0.0930	0.1085	0.0792	0.0698
b_3	0.3937	0.2634	0.3275	0.1084
b_4	0.4650	−0.3527	0.3046	−0.6066
b_5	0.0291	−0.2014	0.0867	−0.0221
u_{01}	−0.1841	0.1708	−0.1242	0.2453
u_{02}	0.1683	−0.1748	0.1023	−0.2700
u_{03}	−0.1607	0.1874	−0.0976	0.2700
u_{04}	−0.0352	−0.0182	−0.0266	0.0001
u_{05}	−0.0178	−0.0177	−0.0123	−0.0130
u_{06}	−0.0149	0.0108	−0.0160	0.0026
u_{07}	0.0902	−0.1468	0.0360	−0.2222
u_{08}	−0.0672	−0.0269	−0.0442	0.0251
u_{09}	0.0682	0.0274	0.0484	−0.0194
u_{010}	0.0466	0.0055	0.0292	−0.0387
u_{011}	−0.0873	0.1434	−0.0344	0.2233
u_{012}	−0.0147	−0.0841	−0.0039	−0.0413
u_{013}	−0.0055	−0.0755	0.0057	−0.0312
u_{014}	0.0072	0.0810	−0.0016	0.0393
u_{015}	0.0956	−0.1413	0.0415	−0.2191
t	−0.0585	−0.0567	−0.0597	−0.0579
t^2	0.0036	0.0036	0.0036	0.0037
Nontrad	0.1336	0.1374	0.1303	0.1374
Nontrad^2	−0.0157	−0.0163	−0.0156	−0.0166
Equity	0.4182	0.4907	0.3689	0.4779
Equity^2	−0.0210	−0.0257	−0.0194	−0.0264
σ_s^2	0.0948	0.0934	0.0933	0.0909
γ	0.8429	0.8382	0.8392	0.8316
η_1	0.3910	0.3988	0.3900	0.4058
η_2	0.1875	0.1893	0.1864	0.1908
Log likelihood	1305.1	1292.3	1300.5	1280.9
Curvature violations	17.3%	0	15.6%	0
Monotonicity violations	5.6%	6.2%	0	0
Mean efficiency	0.8916	–	–	0.8936

TABLE 3.13. PARAMETER ESTIMATES FOR GROUP 10

Parameter	Unconstrained	Curvature only	Monotonicity only	Both Monotonicity and Curvature
u_0	7.7178	7.7385	7.0927	7.4785
b_1	0.4581	0.5502	0.5566	0.6193
b_2	0.3386	0.2337	0.2494	0.1889
b_3	0.0949	0.1325	0.1248	0.0854
b_4	0.1616	0.2084	0.2485	0.2047
b_5	−0.0204	0.0220	0.1263	0.0414
u_{01}	−0.0765	−0.0842	−0.1249	−0.0646
u_{02}	0.0955	0.0868	0.1189	0.0469
u_{03}	−0.0226	−0.0365	−0.0701	−0.0126
u_{04}	−0.0215	−0.0260	−0.0184	−0.0163
u_{05}	−0.0085	−0.0118	−0.0088	−0.0070
u_{06}	−0.0089	−0.0129	−0.0145	−0.0104
u_{07}	−0.0014	0.0173	0.0180	0.0156
u_{08}	−0.0366	−0.0386	−0.0238	−0.0098
u_{09}	0.0180	0.0240	0.0145	0.0018
u_{010}	0.0039	0.0118	0.0058	−0.0041
u_{011}	−0.0208	−0.0305	−0.0363	−0.0281
u_{012}	−0.0202	−0.0182	0.0087	−0.0204
u_{013}	−0.0125	−0.0104	0.0150	−0.0120
u_{014}	−0.0045	0.0073	−0.0270	0.0104
u_{015}	0.0107	0.0228	0.0315	0.0215
t	−0.0436	−0.0422	−0.0440	−0.0421
t^2	0.0039	0.0038	0.0039	0.0038
Nontrad	−0.0343	−0.0455	−0.0516	−0.0562
Nontrad^2	0.0066	0.0083	0.0085	0.0095
Equity	−0.3892	−0.3857	−0.3324	−0.3070
Equity^2	0.0211	0.0218	0.0170	0.0166
σ_s^2	0.0593	0.0600	0.0604	0.0610
γ	0.8038	0.8029	0.8065	0.8054
η_1	0.4957	0.4924	0.4878	0.4878
η_2	0.1885	0.1880	0.1865	0.1874
Log likelihood	1276.0	1261.9	1270.1	1257.2
Curvature violations	27.5%	0	22.4%	0
Monotonicity violations	8.4%	7.1%	0	0
Mean efficiency	0.9003	–	–	0.9001

TABLE 3.14. PARAMETER ESTIMATES FOR GROUP 11

Parameter	Unconstrained	Curvature only	Monotonicity only	Both Monotonicity and Curvature
u_0	2.4555	2.4245	2.2933	2.3250
b_1	0.7153	0.7153	0.7035	0.6845
b_2	0.0126	0.0123	0.0365	0.0321
b_3	−0.0280	−0.0280	0.0622	0.0609
b_4	−0.3451	−0.3451	0.2483	0.2173
b_5	0.7115	0.7115	0.3907	0.3571
u_{01}	−0.0803	−0.0803	−0.1619	−0.1266
u_{02}	0.0098	0.0097	0.0971	0.0768
u_{03}	−0.0078	−0.0080	−0.0875	−0.0696
u_{04}	−0.0338	−0.0345	0.0008	0.0003
u_{05}	−0.0045	−0.0105	0.0011	0.0016
u_{06}	−0.0474	−0.0458	−0.0358	−0.0323
u_{07}	−0.1212	−0.1211	0.0282	0.0233
u_{08}	0.0268	0.0266	0.0045	0.0060
u_{09}	−0.0328	−0.0327	−0.0082	−0.0094
u_{010}	−0.0447	−0.0446	−0.0199	−0.0201
u_{011}	0.1241	0.1241	−0.0247	−0.0194
u_{012}	0.1461	0.1461	0.0574	0.0486
u_{013}	0.1506	0.1506	0.0581	0.0495
u_{014}	−0.1382	−0.1376	−0.0487	−0.0389
u_{015}	−0.1085	−0.1084	0.0348	0.0279
t	−0.0315	−0.0315	−0.0342	−0.0339
t^2	0.0031	0.0029	0.0033	0.0033
Nontrad	−0.1056	−0.1056	−0.1309	−0.1292
Nontrad^2	0.0143	0.0142	0.0169	0.0167
Equity	0.8097	0.8097	0.7047	0.7264
Equity^2	−0.0571	−0.0571	−0.0512	−0.0521
σ_s^2	0.0706	0.0706	0.0728	0.0722
γ	0.8438	0.8438	0.8438	0.8421
η_1	0.4668	0.4668	0.4526	0.4605
η_2	0.2068	0.2068	0.2052	0.2071
Log likelihood	1283.9	1266.8	1258.9	1257.6
Curvature violations	3.8%	0	2.3%	0
Monotonicity violations	21.4%	25.1%	0	0
Mean efficiency	0.9034	–	–	0.9024

TABLE 3.15. PARAMETER ESTIMATES FOR GROUP 12

Parameter	Unconstrained	Curvature only	Monotonicity only	Both Monotonicity and Curvature
u_0	9.9718	8.9411	8.1803	11.0359
b_1	0.8387	0.7629	0.7857	0.6458
b_2	0.0200	0.0677	0.0479	0.1558
b_3	0.0051	−0.0150	0.0378	0.2757
b_4	−0.0320	0.0840	0.1272	1.4352
b_5	0.3304	0.1658	0.2147	0.4760
u_{01}	−0.0956	−0.0606	−0.0784	−0.5649
u_{02}	0.0018	0.0027	0.0084	0.5311
u_{03}	−0.0036	0.0116	−0.0010	−0.4995
u_{04}	−0.0754	−0.0664	−0.0203	−0.0495
u_{05}	−0.0191	−0.0112	−0.0017	−0.0163
u_{06}	−0.0472	−0.0437	−0.0116	−0.0158
u_{07}	0.0135	0.0438	0.0171	0.3617
u_{08}	0.0451	0.0469	0.0220	−0.0450
u_{09}	−0.0385	−0.0422	−0.0206	0.0439
u_{010}	−0.0536	−0.0538	−0.0294	0.0389
u_{011}	0.0005	−0.0307	−0.0021	−0.3706
u_{012}	0.0554	0.0095	0.0340	0.1043
u_{013}	0.0571	0.0113	0.0331	0.1016
u_{014}	−0.0513	−0.0087	−0.0363	−0.1090
u_{015}	−0.0055	0.0258	−0.0025	0.3537
t	−0.0517	−0.0512	−0.0563	−0.0637
t^2	0.0044	0.0043	0.0055	0.0062
Nontrad	0.0007	0.0111	0.0277	−0.0078
Nontrad^2	0.0047	0.0036	0.0002	0.0037
Equity	−1.4230	−1.1720	−1.0914	−2.6916
Equity^2	0.0871	0.0704	0.0590	0.1637
σ_s^2	0.0679	0.0706	0.0579	0.0632
γ	0.8513	0.8537	0.8083	0.8155
η_1	0.2559	0.2358	0.3856	0.3318
η_2	0.1018	0.0982	0.1357	0.1149
Log likelihood	660.6	651.8	627.5	603.1
Curvature violations	34.1%	0	24.9%	0
Monotonicity violations	46.5%	46.6%	0	0
Mean efficiency	0.8867	–	–	0.8856

TABLE 3.16. DIFFERENCES IN AVERAGE EFFICIENCY BETWEEN UNCONSTRAINED AND REGULARITY CONSTRAINED MODELS

Bank group	Difference in average efficiency
Large banks	
Group 1	−0.48%
Group 2	0.32%
Group 3	−0.73%
Medium banks	
Group 4	0.92%
Group 5	−0.05%
Group 6	−0.12%
Group 7	0.03%
Small banks	
Group 8	0.16%
Group 9	−0.20%
Group 10	0.01%
Group 11	0.10%
Group 12	0.11%

TABLE 3.17. SPEARMAN RANK CORRELATION COEFFICIENTS BETWEEN UNCONSTRAINED AND CONSTRAINED MODELS

Bank group	Rank correlation coefficient
Large banks	
Group 1	0.9997
Group 2	0.9861
Group 3	0.9792
Medium banks	
Group 4	0.9460
Group 5	0.9809
Group 6	0.9742
Group 7	0.9869
Small banks	
Group 8	0.9963
Group 9	0.9918
Group 10	0.9911
Group 11	0.9772
Group 12	0.8684

TABLE 3.18. COST EFFICIENCY (%) PER ASSET GROUP

Bank group	Mean	Min.	Max.	5% percentile	95% percentile
Large banks					
Group 1	82.19	42.50	98.95	66.68	97.42
Group 2	88.20	72.52	99.21	77.04	98.36
Group 3	90.38	72.22	99.45	81.71	98.08
Medium banks					
Group 4	88.56	71.86	98.58	77.89	97.86
Group 5	89.61	74.17	99.48	79.77	97.71
Group 6	88.90	72.90	99.40	78.97	98.20
Group 7	91.78	78.52	98.85	83.87	97.90
Small banks					
Group 8	91.29	77.16	98.79	83.11	98.12
Group 9	89.36	75.28	99.30	75.42	98.07
Group 10	90.01	75.56	98.96	80.48	97.93
Group 11	90.24	75.53	98.97	80.99	97.85
Group 12	88.56	70.32	98.97	78.18	97.58

Figure 3.1: Cost Efficiency per Asset Class

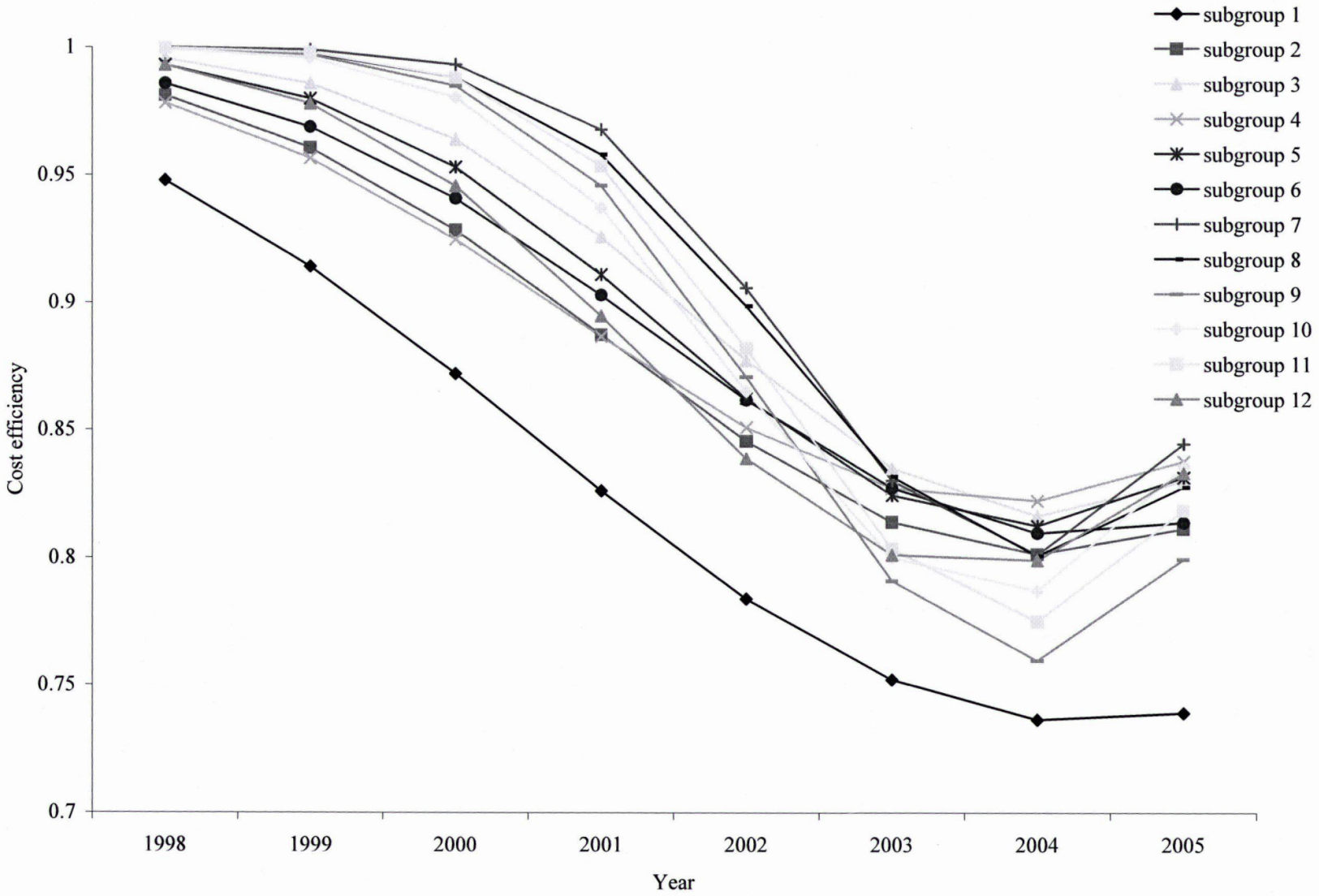


Figure 3.2: Technological Change per Asset Class 1998 - 2005

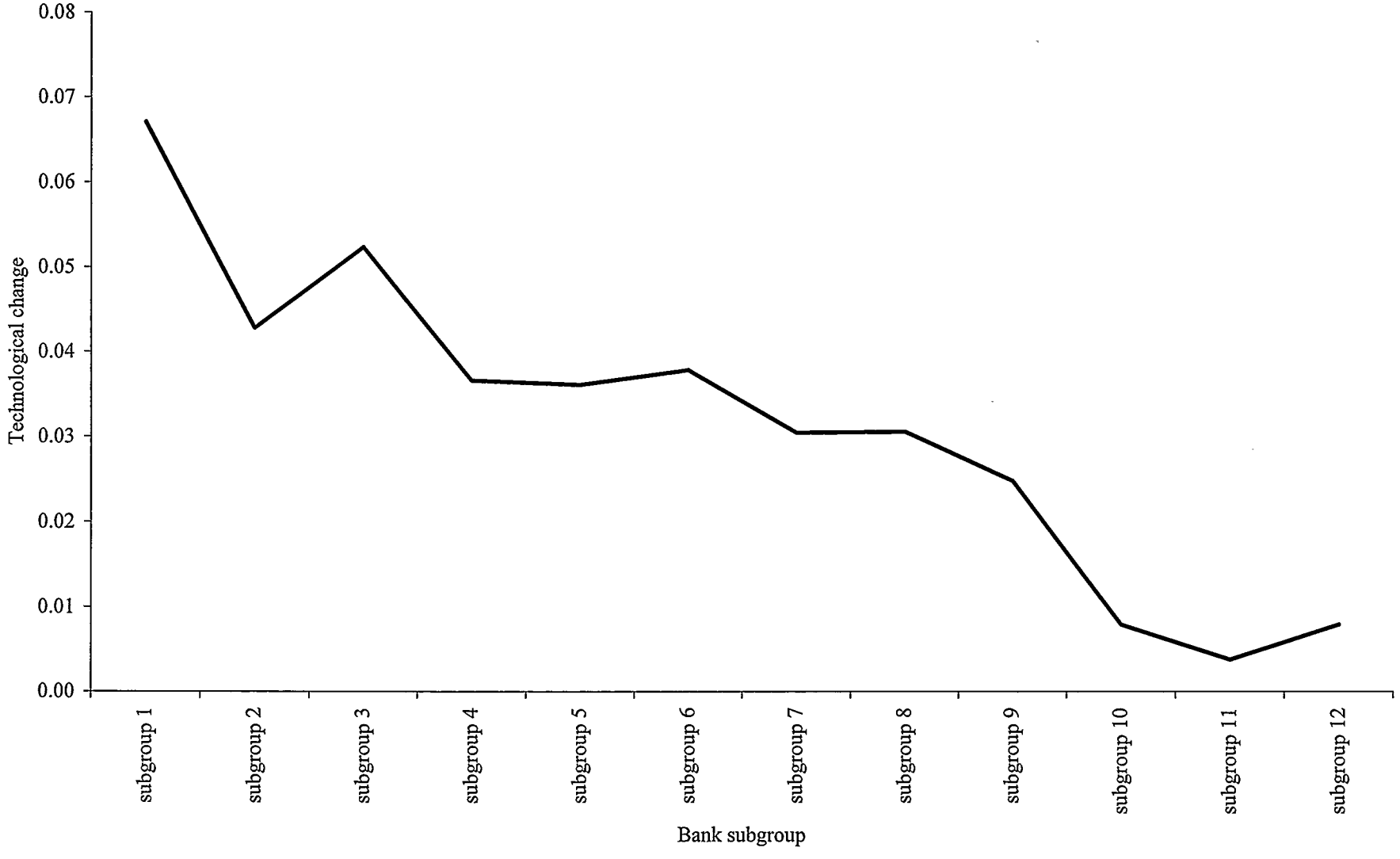
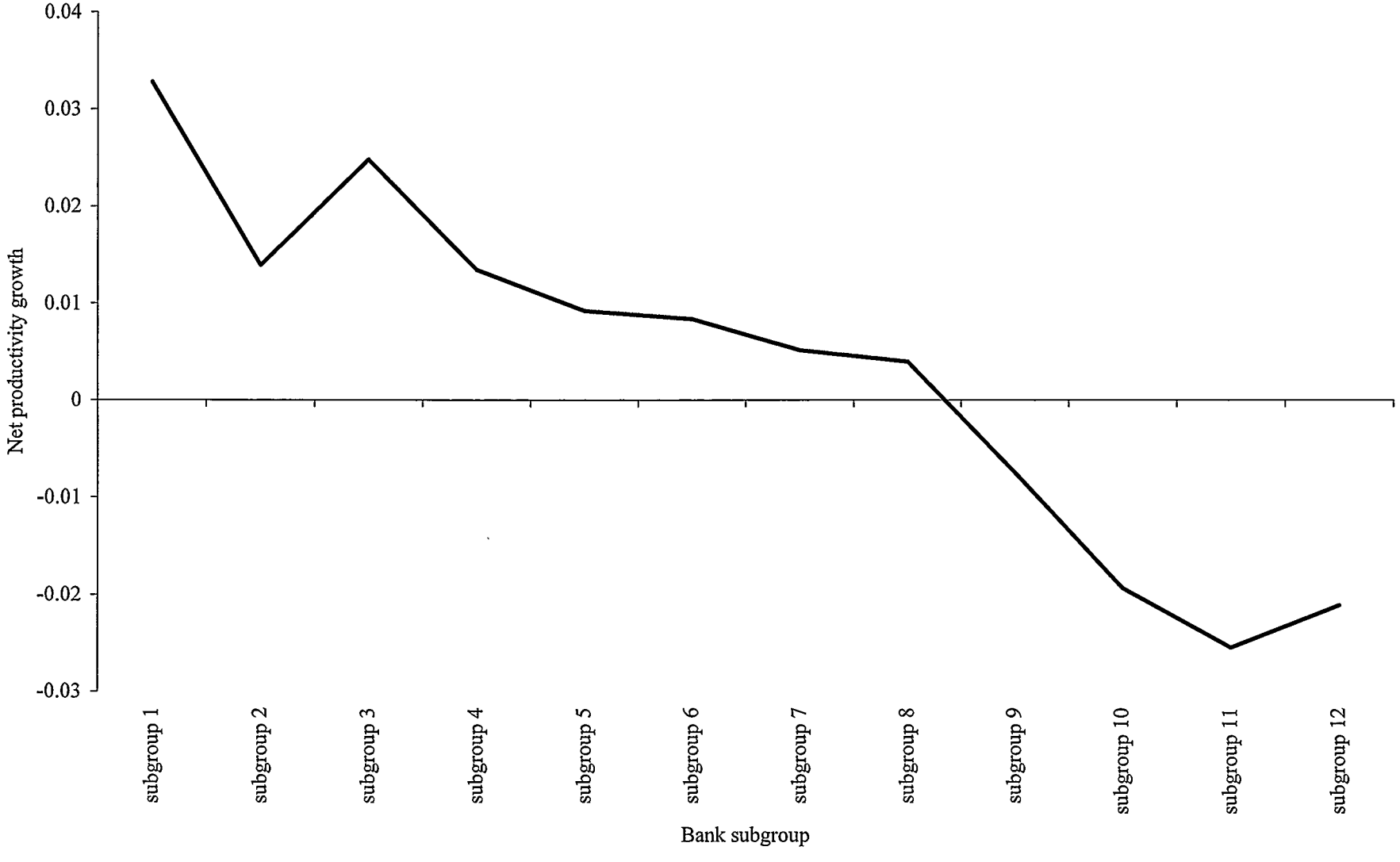


Figure 3.3: Net Productivity Growth per Asset Class: 1998-2005



CHPATER FOUR

EFFICIENCY, TECHNICAL CHANGE, AND RETURNS TO SCALE IN LARGE U.S. BANKS: PANEL DATA EVIDENCE ON BAYESIAN ESTIMATION OF THE OUTPUT DISTANCE FUNCTION

4.1 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, I 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 neccessitated 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, I 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, I 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. I 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, I 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.

I also pay explicit attention to theoretical regularity. I 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, I 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 my 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, I 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 I can reformulate those traditional methods within a constrained optimization method, as in Gallant and Golub (1984), in order to impose theoretical regularity, my 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, I 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. I 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 I 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 I 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 I apply my methodology to a panel data of 292 large banks in the United States, discuss the effects of incorporating monotonicity and curvature, and also report my estimates of total factor productivity growth and its components. The last section summarizes and concludes the paper.

4.2 Theoretical Framework

4.2.1 The Output Distance Function

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

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

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

$$D_o^t(\boldsymbol{y}^t, \boldsymbol{x}^t) = \inf_\theta \left\{\theta > 0 : \frac{\boldsymbol{y}^t}{\theta} \in P^t(\boldsymbol{x}^t)\right\}. \tag{4.1}$$

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

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

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

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

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

Output distance functions are non-decreasing, convex and linearly homogeneous in outputs, and non-increasing and quasi-convex in inputs — see Fare and Grosspokf (1994) for more details. From equation (4.1) it follows that

$$D_o^t(\boldsymbol{y}^t, \boldsymbol{x}^t) \leq 1. \tag{4.2}$$

In (4.2), the equality holds only if $\boldsymbol{y}^t$ is on the output isoquants, which are given by

$$\text{Isoq}P^t(\boldsymbol{x}^t) \equiv \left\{ \boldsymbol{y}^t \middle| D_o^t(\boldsymbol{y}^t, \boldsymbol{x}^t) = 1 \right\}, \tag{4.3}$$

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

To intuitively motivate the output distance function, I can consider $P^t(\boldsymbol{x}^t)$ 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 $\boldsymbol{y}$ 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 $\boldsymbol{y}$ 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 Farrel type output oriented measure of technical efficiency [see Kumbhakar and Lovell (2003)],

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

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, I 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 (4.1) can be rewritten as $D_o(\boldsymbol{x}, \boldsymbol{y}, t)$, which I will use throughout this paper. Deviation of the output distance function from one, due to technical inefficiency, can be accommodated as follows,

$$D_o(\boldsymbol{x}, \boldsymbol{y}, t)\psi(t) = 1, \tag{4.4}$$

where $\psi(t)$ is a function of a random variable, u, which will be discussed in more detail in Section 3. Equation (4.4) will be used below in the decomposition of productivity growth. Also, after specifying functional forms for $D_o(\boldsymbol{x},\boldsymbol{y},t)$ and $\psi(t)$, equation (4.4) will be econometrically estimated.

4.2.2 A Primal Divisia TFP Growth Index

Perfect Competition

I 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)]:

$$\left.\frac{d\ln TFP}{dt}\right|_{\text{Dual}} \equiv \sum_{m=1}^{M} \tilde{s}_m \dot{y}_m - \sum_{N=1}^{N} s_n \dot{x}_n, \tag{4.5}$$

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 (4.5), $\tilde{s}_m = p_m y_m / \sum_{m=1}^{M} p_m y_m$ 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. $\boldsymbol{w} = (w_1, \cdots, w_n)$ and $\boldsymbol{p} = (p_1, \cdots, p_n)$ are price vectors for inputs and outputs, respectively. In (4.5), 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 I 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, I replace the price information in (4.5) with quantity information by solving the following profit maximization problem in perfectly competitive markets.

$$\pi = \max_{\{y,x\}} \left\{ \sum_{m=1}^{M} p_m y_m - \sum_{n=1}^{N} w_m x_m : D_o(\boldsymbol{y}, \boldsymbol{x}, t)\, \psi(t) = 1 \right\}, \tag{4.6}$$

where the constraint is equivalent to $D_o(\boldsymbol{y}, \boldsymbol{x}, t) \leq 1$, 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 prefect 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

$$p_m = \mu \frac{\partial D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial y_m} \psi(t), \quad m = 1, \cdots, M, \tag{4.7}$$

where μ is the Lagrange multiplier. Multiplying both sides of (4.7) with $y_m / D_o(\boldsymbol{y}, \boldsymbol{x}, t)$ and rearranging yields

$$\frac{p_m \boldsymbol{y}_m}{D_o(\boldsymbol{y}, \boldsymbol{x}, t)} = \mu \psi(t) \frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial \ln y_m}, \quad m = 1, \cdots, M, \tag{4.8}$$

Summing up the M equations in (4.8) yields

$$\sum_{i=1}^{M} \frac{p_m \boldsymbol{y}_m}{D_o(\boldsymbol{y}, \boldsymbol{x}, t)} = \mu \psi(t) \sum_{i=1}^{M} \frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial \ln y_m}, \tag{4.9}$$

since the output distance function is linearly homogeneous in $\boldsymbol{y}$ and $D_o(y_m(\boldsymbol{p}, \boldsymbol{x}, t), \boldsymbol{x}, t)\psi(t) = 1$. Noting that $\sum_{i=1}^{M} \partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t) / \partial \ln y_m = 1$ by linear homogeneity of the output function in outputs [see equation (4.31) below], I divide (4.8) by (4.9) to obtain

$$\frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial \ln y_m} = \frac{p_m y_m}{R} = \tilde{s}_m, \quad m = 1, \cdots, M, \tag{4.10}$$

according to which the observed revenue share for the mth output, $\tilde{s}_m$, is equivalent to the elasticity of the distance function with respect to the mth output, $\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t) / \partial \ln y_m$ 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

$$w_n = \mu\psi(t)\frac{\partial D_o(\boldsymbol{y},\boldsymbol{x},t)}{\partial x_n}, \quad n = 1,\cdots,N, \tag{4.11}$$

where μ is the Lagrange multiplier. Multiplying both sides of (4.11) with $x_n/D_o(\boldsymbol{y},\boldsymbol{x},t)$ and rearranging yields

$$\frac{\partial \ln D_o(\boldsymbol{y},\boldsymbol{x},t)}{\partial \ln x_n} = \frac{1}{\mu\psi(t)}\frac{w_n x_n}{D_o(\boldsymbol{y},\boldsymbol{x},t)}, \quad n = 1,\cdots,N. \tag{4.12}$$

Summing up the N equations in (4.12) yields

$$\sum_{n=1}^{N}\frac{\partial \ln D_o(\boldsymbol{y},\boldsymbol{x},t)}{\partial \ln x_n} = \frac{1}{\mu\psi(t)}\sum_{n=1}^{N}\frac{w_n x_n}{D_o(\boldsymbol{y},\boldsymbol{x},t)}. \tag{4.13}$$

Dividing (4.12) by (4.13) yields (for $n = 1,\cdots,N$)

$$\frac{\partial \ln D_o(\boldsymbol{y},\boldsymbol{x},t)}{\partial \ln x_n}\frac{1}{\sum_{n=1}^{N}\partial \ln D_o(\boldsymbol{y},\boldsymbol{x},t)/\partial \ln x_n} = \frac{w_n x_n}{\sum_{n=1}^{N} w_n x_n} = s_n, \tag{4.14}$$

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 (4.13) is actually the shadow cost share.

Substituting (4.10) and (4.14) in (4.5) yields a primal measure of the Divisia total factor productivity growth index which needs only quantity information

$$\left.\frac{d\ln TFP}{dt}\right|_{\text{Primal}} = \sum_{m=1}^{M}\tilde{\omega}_m\dot{y}_m - \sum_{n=1}^{N}\omega_n\dot{x}_n, \tag{4.15}$$

where

$$\tilde{\omega}_m = \frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial \ln y_m}, \tag{4.16}$$

is the shadow revenue share for output m, and

$$\omega_n = \frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t) / \partial \ln x_n}{\sum_{n=1}^{N} \partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t) / \partial \ln x_n} \tag{4.17}$$

is the shadow cost share for input n. To further simplify the notation in (4.17), I define

$$\varepsilon_n = \frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial \ln x_n}, \tag{4.18}$$

and

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

so that ω_n in equation (4.18) 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.

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 I 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 I shall show below). I assume that each firm (bank) solves the following profit maximization problem

$$\max_{y} \pi = \left\{ \sum_{m=1}^{M} p_m (y_m) y_m - C(\boldsymbol{y}, \boldsymbol{w}, t) \right\}, \tag{4.19}$$

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

$$C(\boldsymbol{y}, \boldsymbol{w}, t) = \min_{x} \{ \boldsymbol{w}'\boldsymbol{x} : \; D_o(\boldsymbol{y}, \boldsymbol{x}, t) \psi(t) = 1 \}. \tag{4.20}$$

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 (4.19) are

$$p_m\left(1-m_m\right)=\lambda\frac{\partial C\left(\boldsymbol{y},\boldsymbol{w},t\right)}{\partial y_m},\quad m=1,\cdots,M, \tag{4.21}$$

where λ is the Lagrange multiplier for the profit maximization problem in (4.19), and

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

is the nonnegative ad valorem monopolistic markup for the mth output. Applying the envelope theorem to equation (4.20) with respect to the mth output, I obtain

$$\frac{\partial C\left(\boldsymbol{y},\boldsymbol{w},t\right)}{\partial y_m}=-\tilde{\lambda}\psi(t)\frac{\partial D_o\left(\boldsymbol{y},\boldsymbol{x},t\right)}{\partial y_m}, \tag{4.22}$$

where $\tilde{\lambda}$ is the Lagrangian multiplier for the cost minimization problem in (4.20). Substituting (4.22) into (4.21) yields

$$p_m\left(1-m_m\right)=-\lambda\tilde{\lambda}\psi(t)\frac{\partial D_o\left(\boldsymbol{y},\boldsymbol{x},t\right)}{\partial y_m},\quad m=1,\cdots,M. \tag{4.23}$$

Multiplying both sides of (4.23) by $y_m/D_o\left(\boldsymbol{y},\boldsymbol{x},t\right)$, yields

$$\frac{p_m\left(1-m_m\right)y_m}{D_o\left(\boldsymbol{y},\boldsymbol{x},t\right)}=-\lambda\tilde{\lambda}\psi(t)\frac{\partial \ln D_o\left(\boldsymbol{y},\boldsymbol{x},t\right)}{\partial \ln y_m},\quad m=1,\cdots,M. \tag{4.24}$$

Summing up the M equations in (4.24) yields

$$\sum_{i=1}^{M}\frac{p_m\left(1-m_m\right)y_m}{D_o\left(\boldsymbol{y},\boldsymbol{x},t\right)}=-\lambda\tilde{\lambda}\psi(t)\sum_{i=1}^{M}\frac{\partial \ln D_o\left(\boldsymbol{y},\boldsymbol{x},t\right)}{\partial \ln y_m},\quad m=1,\cdots,M. \tag{4.25}$$

Noting that $\sum_{i=1}^{M}\partial \ln D_o\left(\boldsymbol{y},\boldsymbol{x},t\right)/\partial \ln y_m=1$, by linear homogeneity of the output func-

tion in outputs, and dividing (4.24) by (4.25) yields

$$\frac{p_m\left(1-m_m\right)y_m}{\sum_{i=1}^{M}p_m\left(1-m_m\right)y_m}=\frac{\partial \ln D_o\left(\boldsymbol{y},\boldsymbol{x},t\right)}{\partial \ln y_m}, \tag{4.26}$$

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

Combining (4.14) and (4.26) gives

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

$$=\sum_{m=1}^{M}\frac{p_m\left(1-m_m\right)y_m}{\sum_{i=1}^{M}p_m\left(1-m_m\right)y_m}\dot{y}_m-\sum_{n=1}^{N}\frac{w_nx_n}{\sum_{n=1}^{N}w_nx_n}\dot{x}_n, \tag{4.27}$$

where $\tilde{\omega}_m$ and ω_n are defined separately in (4.16) and (4.17). According to (4.27), 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 (4.5).

It should be noted that (4.27) can be easily generalized to the case where market

power is present in both output and input markets. In that case

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

$$= \sum_{m=1}^{M} \frac{p_m\left(1-m_m\right) y_m}{\sum_{i=1}^{M} p_m\left(1-m_m\right) y_m} \dot{y}_m - \sum_{n=1}^{N} \frac{w_n\left(1-n_n\right) x_n}{\sum_{i=1}^{M} w_n\left(1-n_n\right) x_n} \dot{x}_n, \quad (4.28)$$

where

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

is the nonnegative ad valorem monopsony markdown for the nth input. According to (4.27), 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 (4.15) 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 I 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 — forexample, it allows an easy calculation of the contribution of the scale effect and a deep insight into important production structures.

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 Divisa total factor productivity index be non-decreasing in the output vector and non-increasing in the input vector. An examination of (4.15) reveals that the monotonicity property can be satisfied if

$$\begin{aligned} \frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial \ln y_m} &\geq 0; \\ \\ \frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial \ln x_n} &\leq 0, \end{aligned} \tag{4.29}$$

which is equivalent to the monotonicity conditions of the output distance function (i.e. $\partial D_o(\boldsymbol{y}, \boldsymbol{x}, t)/\partial y_m \geq 0$ and $\partial D_o(\boldsymbol{y}, \boldsymbol{x}, t)/\partial x_n \leq 0$), 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, I take the exponential of both sides of the primal Divisia total factor productivity growth index to obtain its corresponding total factor productivity index

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

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

$$\sum_{m=1}^{M} \tilde{\omega}_m = 1 \text{ and } \sum_{m=1}^{M} \omega_n = 1. \tag{4.30}$$

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

$$\sum_{m=1}^{M} \tilde{\omega}_m = \sum_{m=1}^{M} \frac{\partial \ln D_o\left(\boldsymbol{y}, \boldsymbol{x}, t\right)}{\partial \ln y_m} = \sum_{m=1}^{M} \left(\frac{\partial D_o\left(\boldsymbol{y}, \boldsymbol{x}, t\right)}{\partial y_m} y_m\right) \frac{1}{D_o\left(\boldsymbol{y}, \boldsymbol{x}, t\right)} = 1. \tag{4.31}$$

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 (4.29) and (4.30). This suggests that an estimation method that is capable of imposing the theoretical regularity conditions has to be employed.

4.2.3 Decomposition of the Primal Divisia TFP Growth Index

Equations (4.15), (4.29), and (4.30) provide a basic framework for further decomposing the total factor productivity growth index using the output distance function. In particular, totally differentiating equation (4.4) with respect to time (after taking logs of both sides) and rearranging yields

$$\sum_{m=1}^{M} \frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial \ln y_m} \dot{y}_m = -\frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial t} - \frac{d \ln \psi(t)}{dt} - \sum_{n=1}^{N} \frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial \ln x_n} \dot{x}_n. \quad (4.32)$$

Substituting (4.32) into (4.15) yields

$$\left.\frac{d \ln TFP}{dt}\right|_{\text{Primal}} = TC + \Delta TE + SC, \quad (4.33)$$

where

$$TC = -\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t) / \partial t \quad (4.34)$$

$$\Delta TE = -\partial \ln \psi(t) / \partial t \quad (4.35)$$

$$SC = (\varepsilon - 1) \sum_{n=1}^{N} \left(-\frac{\varepsilon_n}{\varepsilon}\right) \dot{x}_n \quad (4.36)$$

The first term in (4.33) 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.

4.3 The Translog Output Distance Function

In order to implement my total factor productivity growth index decomposition, I need to parameterize and calculate the parameters of an output distance function. Here I choose to parameterize $D_o(\boldsymbol{y}, \boldsymbol{x}, t)$ 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

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

$$+ \sum_{n=1}^{N} b_n \ln x_n + \frac{1}{2}\sum_{n=1}^{N}\sum_{j=1}^{N} b_{nj} \ln x_n \ln x_j + \delta_t t + \frac{1}{2}\delta_{tt} t^2$$

$$+ \sum_{n=1}^{N}\sum_{m=1}^{M} g_{nm} \ln x_n \ln y_m + \sum_{m=1}^{M} \delta_{ym} t \ln y_m + \sum_{n=1}^{N} \delta_{xn} t \ln x_n, \qquad (4.37)$$

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

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

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 (4.37), which is difficult to estimate directly, into an estimable regression model. I choose the Mth output for normalization, which leads to the following expression

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

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

a random error, v, yields the stochastic output distance function

$$-\ln y_M = a_0 + \sum_{m=1}^{M-1} a_m \ln\left(\frac{y_m}{y_M}\right) + \frac{1}{2}\sum_{m=1}^{M-1}\sum_{p=1}^{M-1} a_{mp} \ln\left(\frac{y_m}{y_M}\right)\ln\left(\frac{y_p}{y_M}\right)$$

$$+ \sum_{n=1}^{N} b_p \ln x_p + \frac{1}{2}\sum_{n=1}^{N}\sum_{j=1}^{N} b_{nj} \ln x_n \ln x_j + \delta_t t + \frac{1}{2}\delta_{tt} t^2$$

$$+ \sum_{n=1}^{N}\sum_{m=1}^{M-1} g_{nm} \ln x_n \ln\left(\frac{y_m}{y_M}\right) + \sum_{m=1}^{M-1} \delta_{ym} t \ln\left(\frac{y_m}{y_M}\right) + \sum_{n=1}^{N} \delta_{xn} t \ln x_n + u + v, \tag{4.38}$$

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. I assume that u follows an exponential distribution with scale parameter λ, which I will discuss in more detail in Section 4. Further, I assume that v and u are independent of each other, an assumption I maintain throughout this paper.

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

$$TE = \exp(-u) \tag{4.39}$$

$$TC = -\frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial t} = -\left(\delta_t + \delta_{tt} t + \sum_{m=1}^{M} \delta_{ym} \ln y_m + \sum_{n=1}^{N} \delta_{xn} \ln x_n\right) \tag{4.40}$$

$$RTS = -\sum_{n=1}^{N} \frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial \ln x_n}. \tag{4.41}$$

Equation (4.39) can then be used to obtain efficiency change, $\Delta TE = -du/dt$, and (4.41) can be used to obtain the scale effect,

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

4.3.1 Monotonicity Constraints

As required by microeconomic theory, the output distance function (4.37) has to satisfy the theoretical regularity conditions of monotonicity and curvature. Monotonicity requires that $D_o(\boldsymbol{y}, \boldsymbol{x}, t)$ is non-increasing in $\boldsymbol{x}$ and non-decreasing in $\boldsymbol{y}$. That is,

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

or, equivalently,

$$\frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial \ln x_n} \leq 0 \quad \text{and} \quad \frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial \ln y_m} \geq 0, \tag{4.42}$$

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

The monotonicity restrictions in (4.42) are critically important in ensuring that the

shadow revenue and cost shares are economically meaningful when decomposing the primal total factor productivity growth index (4.15), as I discussed above.

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

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

$$= b_n + \sum_{j=1}^{N} b_{nj} \ln x_j + \sum_{m=1}^{M} g_{nm} \ln y_m + \delta_{xn} t \leq 0, \text{ for } n = 1, \cdots, N; \tag{4.43}$$

$$r_m = \frac{\partial \ln D_o(\boldsymbol{y}, \boldsymbol{x}, t)}{\partial \ln y_m}$$

$$= a_m + \sum_{p=1}^{M} a_{mp} \ln y_p + \sum_{n=1}^{N} g_{nm} \ln x_n + \delta_{ym} t \geq 0, \text{ for } m = 1, \cdots, M. \tag{4.44}$$

Noting that [see equation (4.31)]

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

the monotonicity condition for the Mth output can be also rewritten as

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

4.3.2 Curvature Constraints

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

Hessian matrix

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

are negative, where

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

and

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

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

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

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

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

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

are non-negative, where

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

for $m, p = 1, \cdots, M$ and $\phi_{mp} = 1$ if $m = p$ and 0 otherwise. Note that factoring out $D_o(\boldsymbol{y}, \boldsymbol{x}, t)/y_m$ from the rows and $1/y_p$ from the columns of $\boldsymbol{H}$ does not change the signs of its principal minors. Hence, I can simplify the problem by considering the following matrix

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

where

$$\tilde{h}_{mp} = a_{mp} - r_m r_p - \phi_{mp} r_m. \tag{4.45}$$

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

4.4 Bayesian Estimation

With the translog function for $D_o(\boldsymbol{y},\boldsymbol{x},t)$, the stochastic output distance function in (4.38) can be rewritten in a panel data framework as

$$q_{it} = \boldsymbol{z}_{it}'\boldsymbol{\beta} + u_{it} + v_{it}, \tag{4.46}$$

where $i = 1, \cdots, K$ indicates firms, $t = 1, \cdots, T$ indicates time, $q_{it} = -\ln y_{3,it}$, $\boldsymbol{z}_{it}$ is a vector comprising all the variables which appear on the right hand side of (4.38), and $\boldsymbol{\beta}$ refers to the corresponding vector of coefficients of the translog function (including the

intercept).

The formulation of my empirical model as a random effects model (4.46) is convenient for Bayesian analysis. Although equation (4.38) can also be formulated as a fixed effects model and then estimated using Bayesian procedures, I prefer a random effects model for the following two reasons. First, with a fixed effects model I have to specify the same number of intercepts as that of observational units, which makes the implementation of Bayesian estimation methods cumbersome, since I 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 my 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 (4.39).

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

$$f\left(\boldsymbol{\beta}, h, \boldsymbol{u}, \lambda^{-1} \middle| \boldsymbol{q}\right) \propto L\left(\boldsymbol{q} \middle| \boldsymbol{\beta}, h, \boldsymbol{u}, \lambda^{-1}\right) p\left(\boldsymbol{\beta}, h, \boldsymbol{u}, \lambda^{-1}\right), \tag{4.47}$$

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

Under the assumption that the v_{it}'s are iid normal, the likelihood function in (4.47)

can be shown to be

$$L\left(\boldsymbol{q}\,|\boldsymbol{\beta},h,\boldsymbol{u},\lambda^{-1}\right)=\prod_{i=1}^{K}\prod_{t=1}^{T}\left\{\sqrt{\frac{h}{2\pi}}\exp\left[-\frac{h}{2}\left(q_{it}-\boldsymbol{z}_{it}'\boldsymbol{\beta}-u_{it}\right)\right]^{2}\right\}$$

$$\propto h^{K\times T/2}\exp\left[-\frac{h}{2}\boldsymbol{v}'\boldsymbol{v}\right], \tag{4.48}$$

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

The Bayesian model in (4.47) also requires choosing prior parameter values. I 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

$$p\left(\boldsymbol{\beta}\right)\propto I\left(\boldsymbol{\beta}\in R_{j}\right), \tag{4.49}$$

where $I\left(\cdot\right)$ 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, my 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.

I adopt the following prior for h

$$p(h)\propto h^{-1},\quad\text{where } h=\frac{1}{\sigma^{2}}>0. \tag{4.50}$$

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

As stated above, I 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

$$p\left(u_{it}\left|\lambda^{-1}\right.\right) = f_{\text{Gamma}}\left(u_{it}\left|1, \lambda^{-1}\right.\right), \tag{4.51}$$

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

$$p(\lambda^{-1}) = f_{\text{Gamma}}(\lambda^{-1}\,|1, -\ln\tau^*), \tag{4.52}$$

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

With the priors (4.49)-(4.52), my joint prior probability density function is therefore

$$f\left(\boldsymbol{\beta}, h, \boldsymbol{u}, \lambda^{-1}\right) = p\left(\boldsymbol{\beta}\right) p(h) p\left(\boldsymbol{u}\left|\lambda^{-1}\right.\right) p(\lambda^{-1})$$

$$\propto h^{-1} I\left(\boldsymbol{\beta} \in R_j\right) f_{\text{Gamma}}\left(\lambda^{-1}\,|1, -\ln\tau^*\right) \prod_{i=1}^{K}\prod_{t=1}^{T} f_{\text{Gamma}}(u_{it}\left|1, \lambda^{-1}\right.). \tag{4.53}$$

Finally, my 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 I am comfortable following Tsionas (2006), setting $\tau^* = 0.899$ in this study.

Combining the likelihood function in (4.48) and the joint prior distribution in (4.53) yields the posterior joint density function

$$f\left(\boldsymbol{\beta}, h, \boldsymbol{u}, \lambda^{-1} \,|\, \boldsymbol{q}\right) \propto h^{(KT/2-1)} \exp\left[-\frac{h}{2}\boldsymbol{v}'\boldsymbol{v}\right] I\left(\boldsymbol{\beta} \in R_j\right) \times$$

$$\times f_{\text{Gamma}}\left(\lambda^{-1}|1, -\ln\tau^*\right) \prod_{i=1}^{K}\prod_{t=1}^{T} f_{\text{Gamma}}(u_{it}\,|1, \lambda^{-1}). \qquad (4.54)$$

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

Let $g(\boldsymbol{\beta}, h, \boldsymbol{u}, \lambda^{-1})$ represent these functions of interest. In theory, I could obtain the moments of $g(\boldsymbol{\beta}, h, \boldsymbol{u}, \lambda^{-1})$ from the posterior density through integration. Unfortunately, these integrals cannot be computed analytically. Therefore, I 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, $\boldsymbol{u}$, and λ^{-1}can be shown to be

$$p\left(\lambda^{-1}\left|\boldsymbol{q},\boldsymbol{\beta},h,\boldsymbol{u}\right.\right)\propto f_{\text{Gamma}}\left(\lambda^{-1}\left|KT+1,\boldsymbol{u}'\boldsymbol{\iota}_{KT}-\ln\tau^{*}\right.\right), \tag{4.55}$$

$$p\left(h\left|\boldsymbol{q},\boldsymbol{\beta},\boldsymbol{u},\lambda^{-1}\right.\right)\propto f_{\text{Gamma}}\left(h\left|\frac{KT}{2},\frac{1}{2}\boldsymbol{v}'\boldsymbol{v}\right.\right); \tag{4.56}$$

$$p\left(\boldsymbol{\beta}\left|\boldsymbol{q},h,\boldsymbol{u},\lambda^{-1}\right.\right)\propto f_{\text{Normal}}\left[\boldsymbol{\beta}\left|\boldsymbol{b},h^{-1}\left(\boldsymbol{z}'\boldsymbol{z}\right)^{-1}\right.\right]I\left(\boldsymbol{\beta}\in R_{j}\right) \tag{4.57}$$

$$p\left(\boldsymbol{u}\left|\boldsymbol{q},\boldsymbol{\beta},h,\lambda^{-1}\right.\right)=f_{\text{Normal}}\left(\boldsymbol{u}\left|\boldsymbol{q}-\boldsymbol{z}'\boldsymbol{\beta}-(h\lambda)^{-1}\boldsymbol{\iota}_{KT},h^{-1}\mathbf{I}_{KT}\right.\right)\prod_{i=1}^{K}\prod_{t=1}^{T}I\left(\boldsymbol{u}_{it}\geq 0\right) \tag{4.58}$$

where $\boldsymbol{b}=\left(\boldsymbol{z}'\boldsymbol{z}\right)^{-1}\boldsymbol{z}'\left[\boldsymbol{q}-\mathbf{I}_{KT}\boldsymbol{u}\right]$, with $\boldsymbol{\iota}_{KT}$ 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 $I\left(\boldsymbol{\beta}\in R_{0}\right)$ in (4.57) equal to one and then drawing sequentially from the conditional posteriors in (4.55)–(4.58). Sampling from (4.55), (4.56), and (4.57) is straightforward. However, sampling from (4.58), a multivariate truncated normal distribution, is more complicated. Luckily, in my particular case, sampling from the multivariate truncated normal distribution (4.58) can be simplified as KT independent draws from the following univariate truncated normal distribution

$$p\left(\boldsymbol{u}_{it}\left|\boldsymbol{q},\boldsymbol{\beta},h,\lambda^{-1}\right.\right)=f_{\text{Normal}}\left(\boldsymbol{q}_{it}-\boldsymbol{z}_{it}'\boldsymbol{\beta}-(h\lambda)^{-1},h^{-1}\right)I\left(\boldsymbol{u}_{it}\geq 0\right), \tag{4.59}$$

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 (4.55), (4.56), and (4.58) is the same as in the case without monotonicity and curvature constraints. However, sampling from the multivariate normal distribution (4.57) is even more involved than sampling from the multivariate normal distribution (4.58) in that the region to which β 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 (4.57) — 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, I follow Griffiths *et al.* (2000) and sample the truncated multivariate normal distribution (4.57) using the Metropolis-Hastings algorithm, which in my 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 my case) as

$$\frac{\exp\left[\left(\beta^c - b\right)\left(hz'z\right)\left(\beta^c - b\right)\right]}{\exp\left[\left(\beta^j - b\right)\left(hz'z\right)\left(\beta^j - b\right)\right]}$$

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

4.5 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. I 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, I 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, I 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, I choose not to include it as an output. But I 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.

4.6 Empirical Results

4.6.1 Regularity Tests

I start with unconstrained parameter estimates and make 50,000 draws discarding the first 20,000 as a burn in. Table 4.1 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. I 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 $\widetilde{\boldsymbol{F}}$ and $\widetilde{\boldsymbol{H}}$. I first evaluate the posterior means of k_n and r_m and the principal minors of $\widetilde{\boldsymbol{F}}$ and $\widetilde{\boldsymbol{H}}$, 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 4.2, 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. I then evaluate the posterior coverage regions of k_n and r_m and of the principal minors of $\widetilde{\boldsymbol{F}}$ and $\widetilde{\boldsymbol{H}}$, 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 4.2, 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, I also generate the marginal density plots for the shadow input cost shares, ω_n for $n = 1, 2, 3$ in (4.15), and the shadow output revenue shares, $\tilde{\omega}_m$ for $m = 1, 2, 3$ in (4.15), 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 4.1.1-4.1.6 — 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 4.1.2 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, I follow the procedures specified in Section 4 to impose those constraints on the translog output distance function. Again, I 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 4.3, the monotonicity and curvature violations reported in Table 4.4, and the marginal densities for the shadow input and output shares are plotted in Figure 4.2.1–4.2.6. 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 4.2.1–4.2.6 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 I expected, monotonicity and curvature are satisfied by all measures after monotonicity and curvature are incorporated. In particular, k_m and r_m and the principal minors of $\widetilde{\boldsymbol{F}}$ and $\widetilde{\boldsymbol{H}}$ 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, I 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.

4.6.2 Results from the Constrained Model

Technical Efficiency

Table 4.5.1 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.

My 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, my 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 4.5.2 I 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%.

Returns to Scale

Table 4.6 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 4.6 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 as can be seen from (4.36). 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.

Technical Change

Table 4.7.1 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, I also estimated three alternative models to check the robustness of my results regarding the time pattern of technical change. In the first alternative model (Model 1), I 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), I 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 4.7.2.

The estimates of the rate of technical change from the first alternative model (Model 1), reported in the first column of Table 4.7.2, are almost the same as those in Table 4.7.1 (my 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 4.7.2, also follow the same pattern as in my 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 4.7.2, the time pattern of technical change is still almost the same. As I 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 my 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.

TFP Growth and Its Components

I now turn to a decomposition of the growth rate of total factor productivity, as shown in Table 4.8 — it should be noted that the first year in the sample period is dropped because I 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, I also calculate the percentage contribution of each of the three productivity components to total factor productivity growth, shown in brackets in Table 4.8.

Overall, the results presented in the first column of Table 4.8 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 I 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 4.8 identifies the forces that drive its decline. In particular, the estimates for efficiency changes, $-du/dt$, in the second column of Table 4.8 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 4.8, averaging 7.27% per year. The estimates reported in the fourth column of Table 4.8 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 my 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 4.8. 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.

4.6.3 Sensitivity Analysis

A possible problem with my estimation of the output distance function is endogeneity — that is, the regressors on the right hand side of equation (4.38) may not be exogenous. To investigate the robustness of my results to alternative estimation procedures, in this subsection I use instrumental variables.

The variables on the right hand of (4.38) can be classified into two types of variables — the output ratio variables (i.e. y_m/y_M, $m = 1, \cdots, M-1$) 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. Schmidit (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, I 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 I 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 4.9. A comparison of Tables 4.8 and 4.9 reveals that the major conclusions reached in the previous subsection are still valid, although I 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, I 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 4.9) 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 my earlier estimates. Finally, I also find that the estimates of the contributions of the three productivity components when instrumental variables are used are very similar to my 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, my major conclusions in the previous subsection are quite robust to the use of instrumental variables.

4.7 Conclusion

In this paper, I 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, I 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. I 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, I 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.

I 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). I 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, I need to adopt an estimation method which is capable of incorporating monotonicity and curvature conditions implied by neoclassical microeconomic theory. In this respect, I 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.

I applied my methodology to the panel data on 292 large banks in the United States over the period from 2000 to 2005. my 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. my 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 my 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. my 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, I 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. It should also be noted that the estimates of technical changes, technical efficiency, returns to scales, and productivity growth may be biased due to the selection problem associated with restricting the sample to surviving banks. However, since most of the banks that were driven out of the banking serives market were small banks with under 1 billion in assets – see Jones and Critchfield (2005), selection effects should be quite small.

TABLE 4.1

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 4.2 REGULARITY VIOLATIONS

Regularity conditions	Regularity violations (at the posterior mean)	pdf > 0 (in inadmissible region)
Monotonicity		
$k_1 \leq 0$	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%
Curvature		
All the principal minors of:		
$\widetilde{\boldsymbol{F}}$ are negative, and	100%	100%
$\widetilde{\boldsymbol{H}}$ is positive semidifinite	16.15%	100%

TABLE 4.3

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 4.4 REGULARITY VIOLATIONS (CONSTRAINED MODEL)

Regularity conditions	Regularity violations (at the posterior mean)	pdf > 0 (in inadmissible region)
Monotonicity		
$k_1 \leq 0$	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%
Curvature		
All the principal minors of		
$\widetilde{\boldsymbol{F}}$ are negative, and	0%	0%
$\widetilde{\boldsymbol{H}}$ is positive semidifinite	0%	0%

TABLE 4.5.1 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 4.5.2 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 4.6. 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 4.7.1. 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 4.7.2. 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 4.8. 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 4.9. 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.

Figure 4.1. Estimated Distributions of the Shadow Shares
from Unconstrained Model Evalauted at Mean Prices in 2005

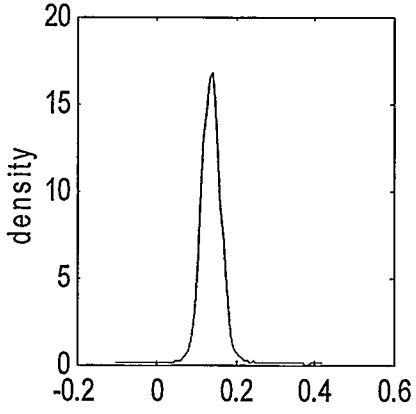


Figure 1.1: labor share

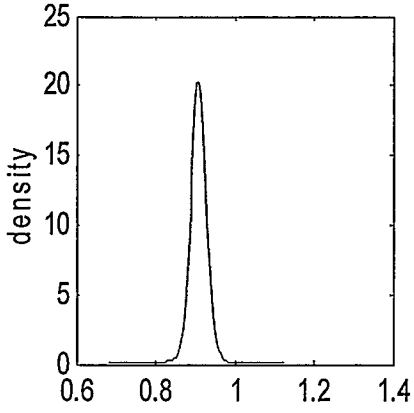


Figure 1.2: fund share

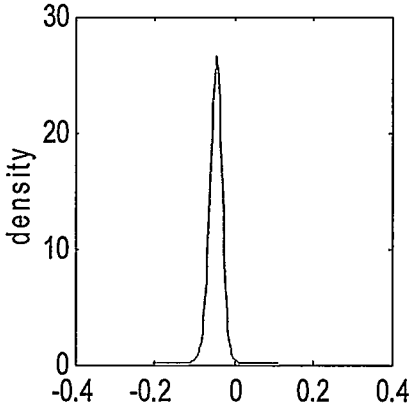


Figure 1.3: capital share

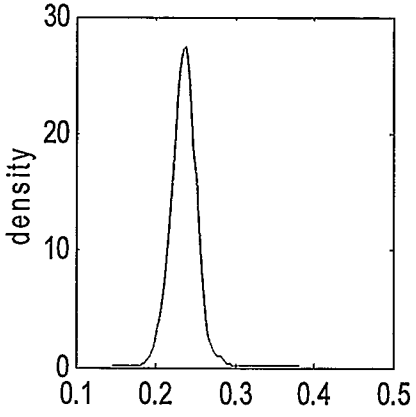


Figure 1.4: securities share

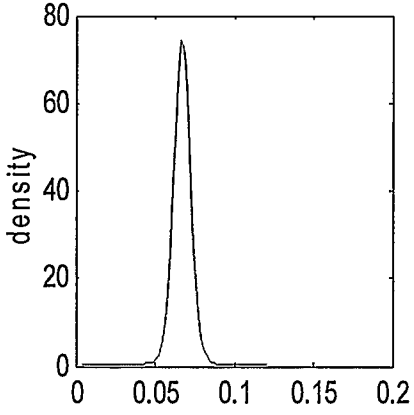


Figure 1.5: consumer loan share

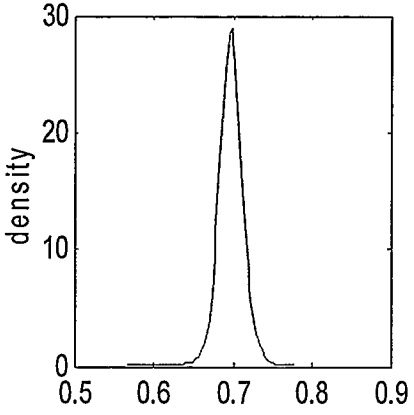


Figure 1.6: non-consumer loan share

Figure 4.2. Estimated Distributions of the Shadow Shares
from Constrained Model Evalauted at Mean Prices in 2005

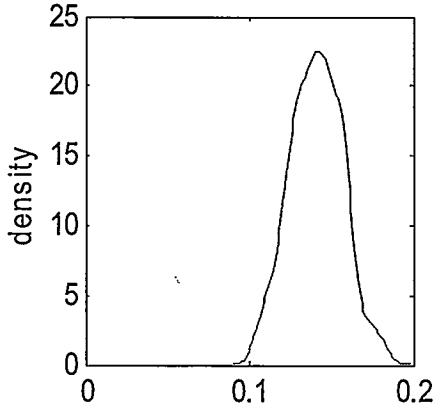


Figure 2.1: labor share

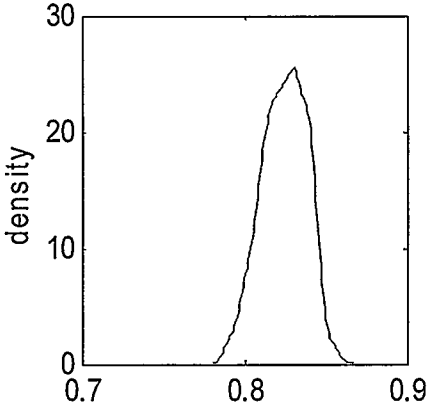


Figure 2.2: fund share

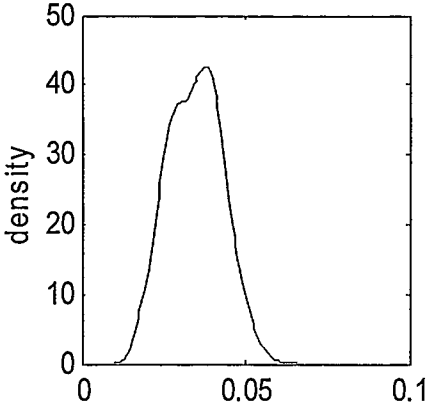


Figure 2.3: capital share

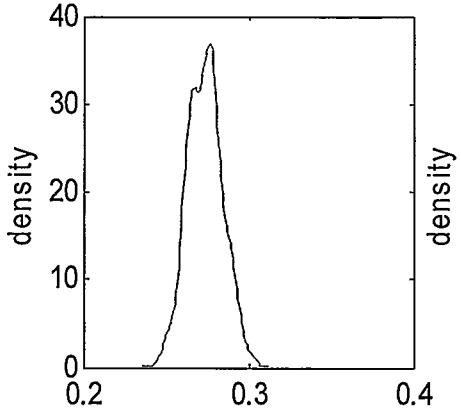


Figure 2.4: securities share

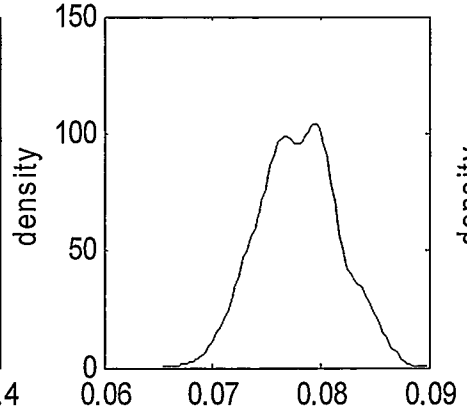


Figure 2.5: consumer loan share

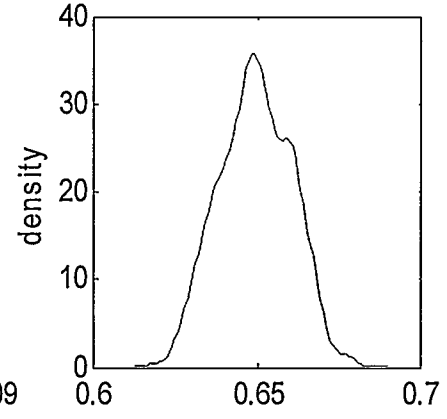


Figure 2.6: non-consumer loan share

CHAPTER FIVE

CONCLUSION

This thesis has focused on the productivity and efficiency issues of two U.S. major industries – manufacturing and banking industries in the context of traditional econometric and more recent stochastic frontier approaches.

In the second chapter of this thesis, I have investigated productivity issues in the U.S. (total) manufacturing industry, in the context of three popular locally flexible functional forms — the generalized Leontief (GL), translog, and normalized quadratic (NQ) — and one globally flexible functional form — the Asymptotically Ideal Production Model (AIM). In doing so, I have extended the Barnett *et al.* (1991) AIM model, by incorporating (for the first time in the literature) technical change through the factor-augmenting efficiency index approach, proposed by Thomsen (2000). I estimated the three locally flexible functional forms parametrically and the globally flexible functional form semi-nonparametrically and treated the curvature property as a maintained hypothesis. The results show that the imposition of local curvature on the GL and translog models does not always assure theoretical regularity. I then provided a comparison between the NQ and AIM cost functions, the only two models that satisfy all three theoretical regularity conditions. I found that the AIM(2) cost function with technical change introduced through the factor-augmenting efficiency index approach performs better than traditional locally flexible function forms and gives more accurate estimates of total factor productivity. I also found that the elasticities from the AIM(2) model are generally larger and show more variation than those from the NQ model, which is consistent with Gallant and Golub (1984) who employed a different globally flexible functional form — the Fourier. Finally, I discussed the elasticities based on the AIM(2) model to shed some new light on the substitutability/complementarity relationship between capital, labor, energy, and materials.

In the third chapter, I have investigated the cost efficiency of 6,010 commercial banks in the U.S. over the period from 1998 to 2005. Cost efficiency of an individual bank

is measured relative to a best practice cost frontier that is estimated using stochastic frontier techniques. In estimating the best practice cost frontier, I (for the first time in this literature) use the globally flexible Fourier functional form, as originally proposed by Gallant (1982), and estimation procedures suggested by Gallant and Golub (1984) to impose the theoretical regularity conditions on the Fourier cost frontier. I find that failure to incorporate monotonicity and curvature into the estimation will result in mismeasured magnitudes of cost efficiency and also misleading bank rankings in terms of cost efficiency. Regarding cost efficiencies from the theoretical regularity constrained models, I find that the largest two subgroups are less efficient than the other subgroups. We also find that all twelve asset size classes show a decline in cost efficiency from 1998 to 2004, and then see a slight improvement in 2005. Further, I find that the largest four bank subgroups (with assets greater than \$400 million) experienced significant productivity gains (NTFPG $> 1\%$) and the smallest eight subgroups (with assets less than \$400 million) experienced insignificant productivity gains (NTFPG $< 1\%$) or productivity losses.

In the fourth chapter, I propose a distance-function based primal Divisia total factor productivity growth index. In particular, I 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 perfect competition. I 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, I present a decomposition of productivity change, isolating the separate contributions of scale economies, technical change, and technical efficiency change. I 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. I then applied the methodology to the panel data on 292 large banks in the United States over

the period from 2000 to 2005. My results confirm that the monotonicity and concavity constrained model yields more accurate and favorable results than an unconstrained model. The 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 my 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.

Bibliography

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

[2] Akhigbea A., and J.E. McNulty. 2003. "The profit efficiency of small US commercial banks." *Journal of Banking and Finance* 27(2), 307-325.

[3] Alam, I.M.S. 2001. "A non-parametric approach for assessing productivity dynamics of large banks." *Journal of Money, Credit, Banking* 33, 121-139.

[4] Aliprantis, C.D., W.A. Barnett, B. Cornet, and S. Durlauf. 2007."Special issue editors' introduction: the interface between econometrics and economic theory." *Journal of Econometrics* 136, 325-329.

[5] Atkinson, S.E., C. Cornwell, and O. Honerkamp. 2003. "Measuring and decomposing productivity change: stochastic distance function estimation versus data envelopment analysis." *Journal of Business and Economic Statistics* 21, 284-294.

[6] Attfield, C.L.F. 1997."Estimating a cointegrating demand system." *European Economic Review* 41, 61-73.

[7] Barnett W.A. 1987. "The microeconomic theory of monetary aggregation". In *New Approaches to Monetary Economics*, Barnett WA, Singleton K (eds). Cambridge University Press: Cambridge, UK; 115-168.

[8] Barnett, W.A. 2002. "Tastes and technology: curvature is not sufficient for regularity." *Journal of Econometrics* 108, 199-202.

[9] Barnett W.A., J. Geweke, and M. Wolfe. 1991." Semi-nonparametric bayesian estimation of the asymptotically ideal production model." *Journal of Econometrics* 49(1/2), 5-50.

[10] Barnett W.A., and J.H. Hahm. 1994. "Financial-firm production of monetary services: a generalized symmetric barnett variable-profit-function approach." *Journal of Business and Economic Statistics* 12(1), 33-46.

[11] Barnett W.A., M. Kirova, and M. Pasupathy. 1995. "Estimating policy-invariant deep parameters in the financial sector when risk and growth matter." *Journal of Money, Credit, and Banking* 27(4), 1402-1430.

[12] Barnett WA, and M. Pasupathy. 2003. "Regularity of the generalized quadratic production model: a counterexample." *Econometric Reviews* 22(2), 135-154.

[13] Barnett, W.A. and P. Yue. 1988. "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 , 229-252.

[14] Barnett WA, and G. Zhou. 1994. "Financial-firms' production and supply-side monetary aggregation under dynamic uncertainty." Federal Reserve Bank of St. Louis *Review* 76(2), 133 -165.

[15] Barten, A.P. 1969. "Maximum likelihood estimation of a complete system of demand equations." *European Economic Review* 1, 7-73.

[16] Battese G.E., and T.J. Coelli. 1992. "Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India." *Journal of Productivity Analysis* 3(1/2), 153-169.

[17] Battese G.E., and T.J. Coelli. 1995. "A model for technical inefficiency effects in a stochastic frontier production function for panel data." Emprirical Economics 20, 325-332.

[18] Battese G.E., and G.S. Corra. 1977. "Estimation of a production frontier model: with application to the pastoral zone of eastern Australia." *Australian Journal of Agricultural Economics* 21(3), 169-179.

[19] Bauer, P. 1990. "Decomposing TFP growth in the presence of cost inefficiency, non-constant returns to scale and technological progress." *Journal of Productivity Analysis* 1, 287-301.

[20] Bauer P.W., and A.N. Berger, G.D. Ferrier, and D.B. Humphrey. 1998. "Consistency conditions for regulatory analysis of financial institutions: a comparison of frontier efficiency methods." *Journal of Economics and Business* 50(2), 85-114.

[21] Berge, C. 1963. *Topological Spaces.* New York: Macmillan..

[22] Berger A.N. 1993. "Distribution-free estimates of efficiency in the US banking industry and tests of the standard distributional assumptions." *Journal of Productivity Analysis* 4(3): 261-292.

[23] Berger A.N. 2004. "The economic effects of technological progress: evidence from the banking industry." *Journal of Money, Credit, and Banking* 35(2): 141-176.

[24] Berger A.N., R.S. Demsetz, and P.E. Strahan. 1999. "The consolidation of the financial services industry: causes, consequences, and the implications for the future." *Journal of Banking and Finance* 23(2-4): 135-94.

[25] Berger A.N., and D.B. Humphrey. 1991. "The dominance of inefficiencies over scale and product mix economies in banking." *Journal of Monetary Economics* 28(1):

117-148.

[26] Berger A.N., and D.B. Humphrey. 1997. "Efficiency of financial institutions: international survey and directions for future research." *European Journal of Operational Research* 98(2): 175-212.

[27] Berger A.N., A.K. Kashyap, and J.M. Scalise. 1995. "The transformation of the U.S. banking industry: what a long, strange trip its been." *Brookings Papers on Economic Activity* 1995(2): 55-218.

[28] Berger A.N., J.H. Leusner, and J.J. Mingo. 1997. "The efficiency of bank branches." *Journal of Monetary Economics* 40(1): 141-162.

[29] Berger, A.N. and L.J. Mester. 1997. "Inside the black box: what explains differences in the efficiencies of financial institutions?" *Journal of Banking and Finance* 21(7): 895-947.

[30] Berger A.N. and L.J. Mester. 2003. "Explaining the dramatic changes in the performance of U.S. banks: technological change, deregulation, and dynamic changes in competition." *Journal of Financial Intermediation* 12(1): 57-95.

[31] Bernanke B.S. 2006. "Community banking and community bank supervision in the twenty-first century." Remarks at the Independent Community Bankers of America National Convention and Techworld, Las Vegas, Nevada, March 8.

[32] Berndt, E.R. and M.S. Khaled. 1979. "Parametric productivity measurement and choice among flexible functional forms." *Journal of Political Economy* 87, 1220-1245.

[33] Bikker, J.A. and K. Haaf. 2002. "Competition, concentration and their relationship: an empirical analysis of the banking industry." *Journal of Banking and Finance* 26,

2191-2214.

[34] Boyd J, and M. Gertler. 1994. "Are banks dead? or are the reports greatly exaggerated." Federal Reserve Bank of Minneapolis *Quarterly Review* Sum: 2-23.

[35] Brummer, B., T. Glauben, and G. Thijssen. 2002. "Decomposition of productivity growth using distance functions: the case of dairy farms in three european countries." *American Journal of Agricultural Economics* 84, 628–644.

[36] Burnside, C. 1996. "Production function regressions, returns to scale, and externalities." *Journal of Monetary Economics* 37, 177-201.

[37] Burguete, J.F., A. R. Gallant, and G. Souza. 1982. "On unification of the asymptotic theory of nonlinear econometric models." *Econometric Reviews* 1, 151-190.

[38] Carlaw, K. I. and R. G. Lipsey. 2003. "Productivity, technology and economic growth: what is the relationship?" *Journal of Economic Surveys* 17, 457-495.

[39] Caves, D.W., and L.R. Christensen. 1980. "Global properties of flexible functional forms." *American Economic Review* 70, 422-432.

[40] Caves, D.W., L.R. Christensen, and W.E. Diewert. 1982. "The Economic theory of index numbers and the measurement of input, output, and productivity." *Econometrica* 50, 1393-1414.

[41] Chalfant J.A. and A.R. Gallant. 1985. "Estimating substitution elasticities with the Fourier cost function: some monte carlo results." *Journal of Econometrics* 28(2): 205-222.

[42] Charnes A., W.W. Cooper and E. Rhodes. 1978. "Measuring the efficiency of decision making units." *European Journal of Operations Research* 2, 429-444.

[43] Christensen, L., D.W. Jorgenson, and L.J. Lau.1975. "Transendendal logarithmic utility functions." *American Economic Review* 65, 367-364.

[44] Claessens S. and L. Laeven. 2003. "Financial development, property rights, and growth." Journal of Finance 58, 2401-2436.

[45] Clark JA, and T.F. Siems. 2002. "X-efficiency in banking: looking beyond the balance sheet." *Journal of Money, Credit, and Banking* 34(4), 987-1013.

[46] Coelli, T.J. and S. Perelman. 1999. "A comparison of parametric and non-parametric distance functions: with application to european railways." *European Journal of Operational Research* 117, 326–339.

[47] Denny, M., M. Fuss and L. Waverman. 1981. "The measurement and interpretation of total factor productivity in regulated industries, with an application to Canadian telecommunications". In: Cowing, T., Stevenson (Eds.), *Productivity Measurement in Regulated Industries.* New York: Academic Press, pp. 179–218.

[48] DeYoung R. 1997. "A diagnostic test for the distribution-free efficiency estimator: an example using U.S. commercial bank data." *European Journal of Operational Research* 98(2), 243-249.

[49] DeYoung R. and I. Hasan. 1998. "The performance of de novo commercial banks: A profit eciency approach." Journal of Banking & Finance 22, 565 - 587.

[50] DeYoung R., I. Hasan and B. Krichhoff. 1998. "The impact of out-of-state entry on the cost efficiency of local commercial banks." *Journal of Economics and Business* 50(2), 191-203.

[51] Dickey, D.A. and W.A. Fuller. 1981. "Likelihood ratio statistics for autoregressive time series with a unit root." *Econometrica* 49, 1057-72.

[52] Diewert, W. E. 1971. "An application of the shephard duality theorem: a generalized leontief production function." *Journal of Political Economy* 79, 481-507.

[53] Diewert, W.E. 1976. "Exact and superlative index numbers." *Journal of Econometrics* 4, 115-145.

[54] Diewert, W.E. 1982. "The duality approach to microeconomic theory." In Kenneth J. Arrow and Michael D. Intriligator (eds). *Handbook of Mathematical Economics* Vol. 2, Amsterdam: North Holland, pp. 535–599.

[55] Diewert, W.E. 2004. Preface . In *Functional Structure and Approximation in Econometrics*, Barnett WA, Binner J (eds). Elsevier: Amsterdam.

[56] Diewert, W.E. and A.O. Nakamura. 2003. "Index number concepts, measures of decompositions of productivity." *Journal of Productivity Analysis* 19, 127-159.

[57] Diewert, W.E. and D. Lawrence. 1999. "Measuring New Zealand's productivity." Wellington, New Zealand Treasury. http://www.treasury.govt.nz/workingpapers/1999/99-5.asp

[58] Diewert, W.E. and D. Lawrence. 2002. "The deadweight costs of capital taxation in Australia." In *Efficiency in the Public Sector*, Kevin J. Fox (ed.). Boston: Kluwer Academic Publishers, pp. 103-167.

[59] Diewert, W.E. and K.J. Fox. July 2004."On the estimation of returns to scale, technical progress and monopolistic markups." Working paper, Department of Economics, University of British Columbia.

[60] Diewert, W.E. and T.J. Wales.1987. "Flexible functional forms and global curvature conditions." *Econometrica* 55, 43-68.

[61] Diewert, W.E. and T.J. Wales.1992. "Quadratic spline models for producer's supply and demand functions." *International Economic Review* 33, 705-722.

[62] Diewert, W.E. and T.J. Wales. 1993. "Linear and quadratic spline models for consumer demand functions." *Canadian Journal of Economics* 26, 77-106.

[63] Dorfman, J.H. and C.S. McIntosh. 2001. "Imposing inequality restrictions: efficiency gains from economic theory." *Economics Letters* 71, 205-209.

[64] Eastwood B.J. and A.R. Gallant. 1991." Adaptive rules for semi-nonparametric estimators that achieve asymptotic normality." *Econometric Theory* 7(3), 307-340.

[65] Eichhorn, W.1976. "Fisher's tests revisited." *Econometrica* 44, 247-256.

[66] Engle, R.F. and C.W.J. Granger.1987. "Cointegration and error correction: representation, estimation and testing." *Econometrica* 55, 251-276.

[67] Färe, R., S. Grosskopf, M. Norris, and Z. Zhang. 1994. "Productivity growth, technical progress and efficiency change in industrialized countries." *American Economic Review* 84, 66-83.

[68] Färe, R. and S. Grosskopf. 1994. *Cost and Revenue Constrained Production.* Springer.

[69] Färe, R. and D. Primont. 1990. "A distance function approach to multi-output technologies." *Southern Economic Journal* 56, 879-891.

[70] Färe, R. and D. Primont. 2000. *Multi-Output Production and Duality: Theory and Applications.* Netherlands: Kluwer Academic Publishers.

[71] Federal Register. 2000. Volume 65, Number 51, 13867. March 15.

[72] Federal Register. 2005. Volume 70, Number 5, 1444. January 7.

[73] Feng, G. and A. Serletis. 2008. "Productivity trends in U.S. manufacturing: evidence from the NQ and AIM cost functions." *Journal of Econometrics* 142, 281-311

[74] Feng, G. and A. Serletis. 2008. "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* (forthcoming)

[75] Fernandez, C., J. Osiewalski, and M.F.J. Steel. 1997. "On the use of panel data in stochastic frontier models with improper priors." *Journal of Econometrics* 79, 169-193.

[76] Ferrier G.D. and C.A.K. Lovell. 1990. "Measuring cost efficiency in.banking: econometric and linear programming evidence." *Journal of Econometrics* 46(1/2), 229-245.

[77] Førsund, F.R. 1997. "The malmquist productivity index, TFP and scale." Taipei International Conference on Efficiency and Productivity Growth, June 20–21.

[78] Fisher, I. 1922. *The Making of Index Numbers: A Study of Their Varieties, Tests, and Reliability.* Boston: Houghton Mifflin .

[79] Fox, K.J. 1996. "Specification of functional form and the estimation of technical progress." *Applied Economics* 28 , 947-956.

[80] Fox, K.J. and Diewert, W. E. 1999. "Is the asia-pacific region different? technical progress bias and price elasticity estimates for 18 OECD countries 1960-1992." In *Economic Efficiency and Productivity Growth in the Asia-Pacific Region,* Fu, T.T., C.J. Huang, and C.A.K. Lovell (eds.), Northampton, MA : Edward Elgar Publishing , pp. 125-144.

[81] Gallant, A.R. 1975. "Seemingly unrelated nonlinear regressions." *Journal of Econometrics* 3, 35-50

[82] Gallant, A.R. 1982. "Unbiased determination of production technology." *Journal of Econometrics* 20(2), 285-323.

[83] Gallant, AR and G. Golub. 1984. "Imposing curvature restrictions on flexible functional forms." *Journal of Econometrics* 26(3), 295-321.

[84] Greene, W. 2005. "Reconsidering heterogeneity in panel data estimators of the stochastic frontier model." *Journal of Econometrics* 126(2), 269-303.

[85] Griffiths, W.E. 2004. "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, pp. 75-91.

[86] Griffiths, W.E. and D. Chotikapanich. 1997. "Bayesian methodology for imposing inequality constraints on a linear expenditure function with demographic factors." *Australian Economic Papers* 36, 321-341.

[87] Griffiths, W.E., C.J. O'Donnell. and A.T. Cruz. 2000. "Imposing regularity conditions on a system of cost and cost-share equations: a bayesian approach." *Australian Journal of Agricultural and Resource Economics* 44, 107-127.

[88] Griliches, Z. 2000. *R & D, Education, and Productivity: A Retrospective.* Cambridge, Harvard University Press.

[89] Guilkey, D., C. Lovell, and R. Sickles. 1983. " A comparison of the performance of three flexible functional forms." *International Economic Review* 24, 591-616.

[90] Hancock, D. 1991. *The Theory of Production for the Financial Firm.* Kluwer Academic: Boston.

[91] Huang, C.J. and J.T. Liu. 1994. "Estimation of a non-neutral stochastic frontier production function." *Journal of Productivity Analysis* 5(June), 171-180

[92] Hughes, J.P. and L.J. Mester. 1998. "Bank capitalization and cost: evidence of scale economies in risk management and signaling." *The Review of Economics and Statistics* 80, 314-325.

[93] Jones K.D. and T. Critchfield. 2005. "Consolidation in the U.S. banking industry: is the long, strange trip about to end?" *FDIC Banking Review* 17(4), 31-61.

[94] Jorgenson, D.W. and Griliches, Z. 1967. "The explanation of productivity change." *Review of Economic Studies* 34, 249-280

[95] Kaparakis E., S. Miller and A. Noulas. 1994. "Short-run cost inefficiency of commercial banks: a flexible stochastic frontier approach." *Journal of Money, Credit, and Banking* 26(4): 875-893.

[96] Kleit, A. and D. Terrell. 2001. "Measuring potential efficiency gains from deregulation of electricity generation: a bayesian approach." *Review of Economics and Statistics* 83, 523-530.

[97] Kohli, U.R. 1981. "Nonjointness and factor intensity in U.S. production." *International Economic Review* 22, 3-18.

[98] Kohli, U.R. 1982. "Production theory, technological change and the demand for imports." *European Economic Review* 18, 369-386.

[99] Kohli, U.R. 1991. *Technology, Duality and Foreign Trade: The GNP Function Approach to Modeling Imports and Exports.* Ann Arbor, MI: University of Michigan

Press .

[100] Kohli, U.R.1993. "U.S. technology and the specific factors model." *Journal of International Economics* 34, 115-136.

[101] Kohli, U.R. 1994. "Technological biases in U.S. aggregate production." *Journal of Productivity Analysis* 5, 5-22.

[102] Koop, G., J. Osiewalski, and M. Steel.1997. "Bayesian efficiency analysis through Individual effects: hospital cost frontiers." *Journal of Econometrics* 76, 77-105.

[103] Kroszner R. and P.E. Strahan. 2000. "Obstacles to optimal policy: the interplay of politics and economics in shaping bank supervision and regulation reforms." Working Paper 7582. National Bureau of Economic Research.

[104] Kumbhakar, S.C. and C.A.K. Lovell. 2003 *Stochastic Frontier Analysis.* Cambridge: Cambridge University Press .

[105] Kumbhakara, S.C. and H.J. Wang. 2005. "Estimation of growth convergence using a stochastic production frontier approach." *Economic Letters* 88, 300-305.

[106] Lau, L.J. 1978. "Testing and imposing monotonicity, convexity, and quasi-convexity constraints." In *Production Economics: A Dual Approach to Theory and Applications* Vol. 1, M. Fuss and D. McFadden (eds.), Amsterdam: North Holland , pp. 409-453.

[107] Lovell, C.A.K., S. Richardson, P. Travers, and L.L. Wood. 1994. "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, pp. 787-807.

[108] Lown C.S., C.L. Osler, P.E. Strahan and A. Sufi. 2000. "The changing landscape of the financial services industry: what lies ahead? Federal Reserve Bank of New York" *Economic Policy Review* 6(4), 39-54.

[109] Magnus J.R. 1985. "On differentiating eigenvalues and eigenvectors." *Econometric Theory* 1(2), 179-191.

[110] Malmquist, S. 1953. "Index numbers and indifference surfaces." *Trabajos de Estadistica* 4, 209-242.

[111] McAllister P.H. and D. McManus. 1993. "Resolving the scale efficiency puzzle in banking." *Journal of Banking and Finance* 17(2/3), 389-406.

[112] Meeusen W. and J. van den Broeck. 1977. "Efficiency estimation from Cobb-Douglas production functions with composed error." *International Economic Review* 18(2), 435-444.

[113] Mester L.J. 1997. "Measuring efficiency at U.S. banks: accounting for heterogeneity is important." *European Journal of Operational Research* 98(2), 230-242.

[114] Montgomery L. 2003. "Recent developments affecting depository institutions." *FDIC Banking Review* 15(2), 54-60.

[115] Morey E.R. 1986. "An introduction to checking, testing, and imposing curvature properties: the true function and the estimated function." *Canadian Journal of Economics* 19(2), 207-235.

[116] Moschini, G. 1999. "Imposing local curvature in flexible demand systems." *Journal of Business and Economic Statistics* 17, 487-490.

[117] Mundlak, Y. 1996. "Production function estimation: reviving the primal." *Econometrica* 64, 431-438.

[118] Ng, S. 1995. "Testing for homogeneity in demand systems when the regressors are nonstationary." *Journal of Applied Econometrics* 10, 147-163.

[119] Nishimizu, M. and J.M. Page Jr. 1982. "Total Factor Productivity Growth, Technological Progress and Technical Efficiency Change: Dimensions of Productivity Change in Yugoslavia, 1965-78." *The Economic Journal* 92(368), 920-936

[120] O'Donnell, C.J. and T.J. Coelli. 2005. "A bayesian approach to imposing curvature on distance functions." *Journal of Econometrics* 126, 493-523.

[121] Orea, L. 2002. "Parametric decomposition of a generalized malmquist productivity index." *Journal of Productivity Analysis* 18, 5-22.

[122] Panzar, J.C. and J.N. Rosse. 1987. "Testing for monopoly equilibrium." *Journal of Industrial Economics* 35, 443-456.

[123] Peristiani S. 1997. "Do mergers improve the X-efficiency and scale efficiency of US banks? evidence from the 1980s." *Journal of Money, Credit, and Banking* 29(3), 326-337.

[124] Phillips, P.C.B. 1987. "Time series regression with a unit root." *Econometrica* 55, 277-301.

[125] Phillips, P.C.B. 1995. "Fully modified least squares and vector autoregression." *Econometrica* 62, 1023-1078.

[126] Phillips, P.C.B. and P. Perron. 1987. "Testing for a unit root in time series regression." *Biometrica* 75, 335-346.

[127] Primont, D. and C. Sawyer. 1993. "Recovering the production technology from the cost function." *Journal of Productivity Analysis* 4, 347-352.

[128] Ray, S.C. and E. Desli. 1997. "Productivity growth, technical progress and efficiency change in industrialized countries: comment." *American Economic Review* 87, 1033-1039.

[129] Rossi M.A. and C.A. Ruzzier. 2000. "On the regulatory application of efficiency measures." *Utilities Policy* 9(2), 81-92.

[130] Ryan, D.L. and T.J. Wales.1998. "A simple method for imposing local curvature in some flexible consumer-demand systems." *Journal of Business and Economic Statistics* 16, 331-338.

[131] Ryan D.L. and T.J. Wales. 2000. "Imposing local concavity in the translog and generalized Leontief cost functions." *Economic Letters* 67(3), 253-260.

[132] Roberts, G.O., A. Gelman and W.R. Gilks. 1997. "Weak convergence and optimal scaling of random walk metropolis algorithms." *Annals of Applied Probability* 7, 110–120.

[133] Sato, K. 1975. *Production Functions and Aggregation.* Amsterdam: North-Holland.

[134] Schmit, P. 1988. "Estimation of fixed effect Cobb-Douglas system using panel data." *Journal of Econometrics* 37, 361-380.

[135] Sealey C. and J. Lindley. 1977. "Inputs, outputs, and a theory of production and cost at depository financial institutions." *Journal of Finance* 32(4), 1251-1266.

[136] Serletis A. and A. Shahmoradi. 2005. "Semi-nonparametric estimates of the demand for money in the united states.' *Macroeconomic Dynamics* 9(4), 542-559.

[137] Serletis, A. and A. Shahmoradi. 2007. "Flexible functional forms, curvature conditions, and the demand for assets." *Macroeconomic Dynamics* 11(4), 455 - 486.

[138] Shephard, R.W. 1953. *Cost and Production Functions.* Princeton: Princeton University Press .

[139] Shephard, R.W. 1970. *Theory of Cost and Production Functions.* Princeton: Princeton University Press.

[140] Simar, L., C.A.K. Lovell, and P.V. Eeckaut. 1994. "Stochastic frontiers incorporating exogenous influences on efficiency." Discussion Paper No. 9403, Institute de Statistique, Universite Catholique de Louvain, Belgium

[141] Solow, R. 1957. "Technical change and the aggregate production function." *Review of Economics and Statistics* 39 , 312-320.

[142] Stiroh, K.J. 2000. "How did bank holding companies prosper in the 1990s?" *Journal of Banking and Finance* 24(11), 1703-1745.

[143] Stock, J.H. and M.W. Watson. 1993. "A simple estimator of cointegrating vectors in higher order integrated systems." *Econometrica* 61, 783-820.

[144] Terrell, D. 1996. "Incorporating monotonicity and concavity conditions in flexible functional forms." *Journal of Applied Econometrics* 11, 179-194.

[145] Thomsen, T. 2000. "Short cuts to dynamic factor demand modelling." *Journal of Econometrics* 97, 1-23.

[146] Tierney, L. 1994. "Markov chains for exploring posterior distributions (with discussion)" *Annals of Statistics* 22 , 1701-1762.

[147] Tsionas, E.G. 2006. "Inference in dynamic stochastic frontier models." *Journal of Applied Econometrics* 21, 669-676.

[148] van den Broeck, J., G. Koop, J. Osiewalski, and M. Steel. 1994. "Stochastic frontier models: a bayesian perspective." *Journal of Econometrics* 46, 39–56.

[149] Wheelock D.C. and P.W. Wilson. 2001. "New evidence on returns to scale and product mix among U.S. commercial banks." *Journal of Monetary Economics* 47(3), 653-674.