ECONOMIC RESEARCH
FEDERAL RESERVE BANK OF ST. LOUIS

WORKING PAPER SERIES

Robust Non-parametric Quantile Estimation of Efficiency and Productivity Change in U.S. Commercial Banking, 1985-2004

Authors	David C. Wheelock, and Paul W. Wilson
Working Paper Number	2006-041B
Revision Date	April 2007
Citable Link	https://doi.org/10.20955/wp.2006.041
Suggested Citation	Wheelock, D.C., Wilson, P.W., 2007; Robust Non-parametric Quantile Estimation of Efficiency and Productivity Change in U.S. Commercial Banking, 1985-2004, Federal Reserve Bank of St. Louis Working Paper 2006-041. URL https://doi.org/10.20955/wp.2006.041

Published In	Journal of Business & Economic Statistics
Publisher Link	https://doi.org/10.1198/jbes.2009.06145

Robust Non-parametric Quantile Estimation of Efficiency and Productivity Change in U.S. Commercial Banking, 1985–2004

David C. Wheelock   Paul W. Wilson*

December 2007

Abstract

This paper uses a new non-parametric, unconditional, hyperbolic order-α quantile estimator to construct a hyperbolic version of the Malmquist index. Unlike traditional non-parametric efficiency estimators, the new estimator is both robust to data outliers and has a root-n convergence rate. We use this estimator to examine changes in the efficiency and productivity of U.S. banks between 1985 and 2004. We find that larger banks experienced larger efficiency and productivity gains than small banks, consistent with the presumption that recent changes in regulation and information technology have favored larger banks.

JEL classification nos.: C13, C14, G21

Keywords: efficiency, productivity, quantile estimation, commercial banks

*Wheelock: Research Department, Federal Reserve Bank of St. Louis, P.O. Box 442, St. Louis, MO 63166; email david.c.wheelock@stls.frb.org. Wilson: The John E. Walker Department of Economics, 222 Sirrine Hall, Clemson University, Clemson, SC 29634–1309; email pww@clemson.edu. Research described in this paper was conducted while Wilson was a visiting scholar in the Research Department of the Federal Reserve Bank of St. Louis. Xiaoshu Han, Justin Hauke, Hyun Kim, Aemit Lakdawala, J. Fred Thomas, and Joshua Ulrich provided capable assistance. Any remaining errors are solely our responsibility. The views expressed in this paper do not necessarily reflect official positions of the Federal Reserve Bank of St. Louis or the Federal Reserve System.

1 Introduction

The performance of firms, non-profit entities and other decision-making units (DMUs) is often assessed in terms of technical, cost, or other forms of efficiency. For managers, efficiency estimates can help identify opportunities for reducing costs or increasing revenues. Market analysts and researchers have used efficiency estimates to help predict failures, merger activity, and to examine the effects of innovations and regulatory changes. Often, dynamic effects are of interest, particularly in gauging the effects of regulatory reform, new methods of production, or other innovations. In addition to considering changes in technical efficiency over some time period, researchers often examine changes in productivity, changes in technology, changes in scale efficiency, etc.

Practitioners have used both parametric and non-parametric approaches to estimate efficiency. The typical approaches, however, either require potentially untenable specification assumptions or have other serious drawbacks. A common parametric approach, based on the ideas of Aigner et al. (1977) and Meeusen and van den Broeck (1977), involves estimation of a specific response function with a composite error term consisting of inefficiency and noise components. Often studies specify a translog response function; researchers have found, however, that the translog function is often a mis-specification, especially when producers are of widely varying sizes.[1] In an attempt to increase flexibility, researchers sometimes augment translog specifications with trigonometric terms. In order to maximize log-likelihoods when composite error terms are used, however, the number of additional terms is typically limited to a number that, in most cases, is probably far less than needed to minimize criteria such as asymptotic mean integrated square error.[2]

The inherent problems with parametric efficiency models have led many researchers to

[1]Empirical examples are provided by Cooper and McLaren (1996), Banks et al. (1997), Wheelock and Wilson (2001), and Wilson and Carey (2004). For Monte Carlo evidence, see Guilkey et al. (1983) and Chalfant and Gallant (1985).

[2]To our knowledge, no studies in the banking literature using this approach have optimized the number of terms with respect to asymptotic mean integrated square error or similar criteria. Gallant (1981, 1982) suggests as a rule of thumb $n^{2/3}$ terms, where n is the sample size; applied papers where models with composite errors are estimated have typically included far fewer terms. Researchers apparently limit the number of included terms due to the practical problems associated with maximizing highly non-linear log-likelihoods with respect to large numbers of parameters. Instead of trigonometric functions, one could use as basis functions members of a family of orthogonal polynomials (e.g., Laguerre or Legendre polynomials), but the problems of determining the optimal number of terms, and using these in a non-linear, maximum-likelihood framework, remain.

apply non-parametric methods. Non-parametric methods are popular because they avoid having to specify *a priori* a particular functional relationship to be estimated; the data are allowed to speak for themselves. Non-parametric approaches usually involve the estimation of a production or other set by either the free-disposal hull (FDH) of sample observations, or the convex hull of the FDH. Methods based on the convex hull of the FDH are collectively referred to as data envelopment analysis (DEA). Inefficiency is estimated by the distance from the location of a DMU in input/output space to an estimate of the boundary of support of the production set. In dynamic settings, Malmquist indices defined in terms of distance functions estimated by DEA methods are frequently used to measure changes in productivity; these indices are often decomposed into sub-indices giving measures of changes in efficiency, technology, etc.[3] The statistical properties of DEA estimators have been established, and bootstrap methods exist for making statistical inferences about the efficiency of individual firms based on DEA estimates, as well as productivity change, etc. measured by Malmquist indices and their component sub-indices.[4]

Despite their popularity, both DEA and FDH estimators have some obvious drawbacks. First, it has long been recognized that DEA and FDH estimates of inefficiency are sensitive to outliers in the data. Second, DEA as well as FDH estimators also suffer from the well-known *curse of dimensionality* that often plagues non-parametric estimators. The number of observations required to obtain meaningful estimates of inefficiency increases dramatically with the number of production inputs and outputs; for a given sample size, adding dimensions results in more observations falling on the *estimated* frontier. In many applications, including the one in this paper, there are simply too few observations available to obtain meaningful estimates of inefficiency using FDH or DEA.

Recently, some interesting and useful alternatives to FDH and DEA have been developed. Cazals et al. (2002) proposed a strategy based on estimation of expected minimum

[3]Productivity relates output quantities produced to input quantities used, whereas technical efficiency reflects how close a DMU lies to the boundary of the production set. A DMU could experience a change in productivity without a change in efficiency, or a change in efficiency without a change in productivity. For example, a DMU would experience no change in efficiency if an increase in its productivity was matched by an increase in production possibilities that left the DMU an unchanged distance from the boundary of the production set. Alam (2001) and Wheelock and Wilson (1999) are among recent applications of DEA that estimate changes in efficiency and productivity of U.S. commercial banks.

[4]See Simar and Wilson (2000b) for a survey, and Kneip et al. (2007) for more recent results, on the statistical properties of DEA estimators. See Simar and Wilson (1998, 2000a, 2000b) and Kneip et al. (2007) for details about the use of bootstrap methods to make inferences based on DEA estimates.

input functions or expected maximum output functions of order-m, with $m \in \{1,\ 2,\ \ldots\}$. Daouia (2003) and Aragon et al. (2005) offered an alternative involving estimation of input- and output-oriented conditional quantiles of order-α, with $\alpha \in (0, 1]$. Daouia and Simar (2007) extended the approach to a multivariate setting (i.e., where production involves both multiple inputs and multiple outputs). Both the order-m and α-quantile approaches involve estimating partial frontiers lying "close" to the full production frontier. Both estimators are robust with respect to outliers because they allow some observations to lie above the estimated partial frontier. Moreover, although fully non-parametric, both estimators achieve the classical, parametric root-n rate of convergence with no curse of dimensionality when used to estimate partial frontiers. Further, if the orders m or α are viewed as sequences of appropriate order in sample size n so that $m(n) \to \infty$ as $n \to \infty$ or $\alpha(n) \to 1$ as $n \to \infty$, the estimators can be interpreted as robust (with respect to outliers) estimators of the full frontier, although the root-n convergence rate is lost when the estimators are used to estimate the full frontier.

Although the order-m and order-α approaches overcome two of the problems associated with DEA and FDH estimators, a third issue remains—the decision whether to measure efficiency in the input- or the output-direction. The choice can be crucial—especially for estimating the efficiency of firms operating at the extremes of the size range. Irrespective of estimation method, or whether full or partial frontiers are estimated, small firms lying close to the production frontier in the input direction often lie much farther from the frontier in the output direction. Similarly, large firms lying close to the frontier in the output direction may lie far away from the frontier in the input direction. Thus, small (large) firms that appear relatively efficient when efficiency is estimated in the input (output) direction may appear highly inefficient when efficiency is measured in the output (input) direction. Consequently, true efficiency estimates (again, apart from estimation issues)—and the apparent amount of overall, or average, technical efficiency in a given sample—may depend crucially on the distribution of the data, the curvature of the frontier, and whether one uses an input- or output-orientation. Unfortunately, the choice between input- or output-orientation for efficiency measurement is often arbitrary.

In this paper, we describe an *unconditional*, hyperbolic order-α quantile estimator that shares the advantages of the Cazals et al. (2002) and Daouia and Simar (2007) estimators, but which avoids the third problem involving choice of orientation. Our estimator extends the ideas of Daouia and Simar (2007), and shares many of the properties of their conditional order-α estimators, but with the additional advantage of avoiding the choice between input- and output-orientations and the resulting sensitivity of results with respect to that choice. In addition, we define Malmquist indices and component sub-indices in terms of hyperbolic order-α quantiles, and demonstrate how these can be estimated.

Avoiding the choice between input- and output-orientation is potentially even more important in dynamic settings where Malmquist indices and their components are estimated. With cross-period comparisons needed to define such indices, the sensitivity of results to the choice of input- or output orientation is more likely to arise. Whereas a firm might lie near the middle of the range of the data for one period, it might lie near the steeply-sloped or nearly flat portions of the frontier prevailing in the other period, in which case estimates of productivity or efficiency change may be highly sensitive to the choice of input- or output-orientation.

Moreover, from a practical viewpoint, the cross-period comparisons used to estimate changes in technology often result in infeasible solutions when DEA or FDH estimators are used. In particular, it is sometimes the case that a firm's position in one period is either above or to the left of the frontier estimate in another period; in the former case, cross-period, input-oriented DEA or FDH efficiency estimates cannot be computed; in the latter case, cross-period output-oriented estimates cannot be computed. Similar problems exist for the input- and output-oriented estimators of Cazals et al. (2002) and Daouia and Simar (2007). Measuring efficiency along hyperbolic paths avoids these problems.

We use our estimator to produce new estimates of efficiency and productivity change for U.S. commercial banks between 1985 and 2004. The U.S. banking industry experienced rapid consolidation during these years, reflected in a reduction in the number of commercial banks from a post-war peak of 14,496 banks at the end of 1984 to 7,630 banks at the end of 2004. Consolidation coincided with dramatic changes in regulation, market structure, and in the use of information-processing technology by banks and their competitors. Bank failures accounted for a significant number of exits in the 1980s and early 1990s. Although failures have since been rare, analysts continue to question the long-run viability of commercial banks — especially smaller, "community" banks — as other intermediaries and financial markets increasingly encroach on the traditional deposit-taking and lending business of commercial banks.[5] The viability of banks would seem to hinge on how well they respond to changes in regulation, competition, and advances in information-processing technologies that shape the environment of banking by improving their efficiency and productivity. Although numerous studies have examined commercial bank efficiency, to date all have relied on the traditional approaches described above that either impose restrictive specification assumptions or have other undesirable properties.[6]

Our estimation results indicate that, in general, U.S. banks became more efficient between 1985 and 2004. However, only large banks—those with at least $1 billion of total assets—experienced significant productivity improvement. Our results are thus consistent with the presumption that branching deregulation and rapid advances in information technology have disproportionately benefited larger banks, and could help explain the relatively rapid decline in the number of small banks.

The rest of the paper unfolds as follows: Section 2 presents a statistical model and defines quantile distance functions. Section 3 discusses estimation methodology. Measures of productivity change and its components in terms of quantile distance functions are introduced in Section 4. We describe our data in Section 5, present empirical results in Section 6, and offer conclusions in Section 7.

[5]Berger (2003) describes the myriad advances in information and financial technology, and changes in regulation, that have affected the banking industry over these years, and discusses their implications for banks of different sizes and their competitors. Major changes in bank regulation since 1980 include the deregulation of deposit interest rates, the introduction of risk-based capital requirements, and the removal of legal restrictions on branching, first within states and later across state boundaries. Branching deregulation in particular promoted rapid consolidation of the industry.

[6]In particular, researchers have found that the translog functional form is a mis-specification of bank cost relationships (e.g., McAllister and McManus, 1993; Wheelock and Wilson, 2001), which calls into question the results of the numerous studies of commercial bank efficiency that impose this functional form. To date, studies that employ non-parametric estimation of commercial bank efficiency have used either DEA or FDH estimators. Kumbhakar et al. (2006) propose an interesting local maximum likelihood approach allowing for both noise and a stochastic inefficiency term while avoiding the need to specify the response function, but to our knowledge, their idea has (so far) not been used to examine commercial banks. See Berger and Humphrey (1997) for a survey of commercial bank efficiency studies.

2 A Statistical Model of Production

2.1 Some (Minimal) Assumptions

Given vectors $\boldsymbol{x} \in \mathbb{R}_+^p$ of p input quantities and $\boldsymbol{y} \in \mathbb{R}_+^q$ of q output quantities, standard microeconomic theory of the firm posits a production set at time t represented by

$$\mathcal{P}^t \equiv \{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{x} \text{ can produce } \boldsymbol{y} \text{ at time } t\}. \tag{2.1}$$

This set represents the set of feasible combinations of inputs and outputs at a given point in time, and may change with the passage of time.

The assumptions listed below are similar to those in Park et al. (2000), and serve to define a statistical model.

Assumption 2.1. *The production set $\mathcal{P}^t$ is compact and free disposal, i.e., if $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}^t$, $(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}) \in \mathcal{P}^t$, and $\widetilde{\boldsymbol{x}} \geq \boldsymbol{x}$, then $(\widetilde{\boldsymbol{x}}, \widetilde{\widetilde{\boldsymbol{y}}}) \in \mathcal{P}^t \; \forall \; \mathbf{0} \leq \widetilde{\widetilde{\boldsymbol{y}}} \leq \boldsymbol{y}$.*[7]

Assumption 2.2. *$(\boldsymbol{x}, \boldsymbol{y}) \notin \mathcal{P}^t$ if $\boldsymbol{x} = 0$, $\boldsymbol{y} \geq 0, \boldsymbol{y} \neq \mathbf{0}$, i.e., all production requires use of some inputs.*

Assumption 2.3. *The sample $\mathcal{S}_{n_t}^t = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n_t}$ of n_t observations on input and output quantities at time t are realizations of identically, independently distributed (iid) random variables with probability density function $f^t(\boldsymbol{x}, \boldsymbol{y})$ with support over $\mathcal{P}^t$.*

A point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}^t$ is said to be on the *frontier* of $\mathcal{P}^t$, denoted $\mathcal{P}^{t\partial}$, if $(\gamma^{-1}\boldsymbol{x}, \gamma\boldsymbol{y}) \notin \mathcal{P}^t$ for any $\gamma > 1$; let $(\boldsymbol{x}_0^{t\partial}, \boldsymbol{y}_0^{t\partial}) \in \mathcal{P}^{t\partial}$ denote such a point.

Assumption 2.4. *At the frontier, the density f^t is strictly positive, i.e., $f_0^t = f^t(\boldsymbol{x}_0^{t\partial}, \boldsymbol{y}_0^{t\partial}) > 0$, and sequentially Lipschitz continuous, i.e., for all sequences $(\boldsymbol{x}_n, \boldsymbol{y}_n) \in \mathcal{P}^t$ converging to $(\boldsymbol{x}_0^{t\partial}, \boldsymbol{y}_0^{t\partial})$, $|f^t(\boldsymbol{x}_n, \boldsymbol{y}_n) - f^t(\boldsymbol{x}_0^{t\partial}, \boldsymbol{y}_0^{t\partial})| \leq c_1 ||(\boldsymbol{x}_n, \boldsymbol{y}_n) - (\boldsymbol{x}_0^{t\partial}, \boldsymbol{y}_0^{t\partial})||$ for some positive constant c_1.*

Now let y^k denote the kth element of $\boldsymbol{y}$, $k = 1, \ldots, q$, and let $\boldsymbol{y}^{(k)} = \begin{bmatrix} y^1 & \ldots & y^{k-1} & y^{k+1} & \ldots & y^q \end{bmatrix}$ denote the vector $\boldsymbol{y}$ with the kth element deleted. In addition, let $\boldsymbol{y}^{(k)}(\eta) = \begin{bmatrix} y^1 & \ldots & y^{k-1} & \eta & y^{k+1} & \ldots & y^q \end{bmatrix}$ denote a vector similar to $\boldsymbol{y}$, but with

[7] Here and throughout, inequalities involving vectors are defined on an element-by-element basis; e.g., for $\widetilde{\boldsymbol{x}}$, $\boldsymbol{x} \in \mathbb{R}_+^p$, $\widetilde{\boldsymbol{x}} \geq \boldsymbol{x}$ means that some number $\ell \in \{0, 1, \ldots, p\}$ of the corresponding elements of $\widetilde{\boldsymbol{x}}$ and $\boldsymbol{x}$ are equal, while $(p - \ell)$ of the elements of $\widetilde{\boldsymbol{x}}$ are greater than the corresponding elements of $\boldsymbol{x}$.

η substituted for the kth element of $\boldsymbol{y}$. For each $k = 1, \ldots, 1$ define a function

$$g^k_{\mathcal{P}^t}\left(\boldsymbol{x}, \boldsymbol{y}^{(k)}\right) \equiv \max\left\{\eta \mid \left(\boldsymbol{x}, \boldsymbol{y}^{(k)}(\eta)\right) \in \mathcal{P}^t\right\}. \tag{2.2}$$

As discussed in Park et al. (2000), the production set $\mathcal{P}^t$ can be defined in terms of any of the functions $g^k_{\mathcal{P}^t}$. Along the lines of Park et al., the following analysis is presented in terms of $g^q_{\mathcal{P}^t}$, denoted simply as g^t.

Assumption 2.5. *At the frontier, $g^t(\cdot,\cdot)$ is (i) positive, i.e., $g^t(\boldsymbol{x}^{t\partial}_0, \boldsymbol{y}^{t\partial}_0) > 0$; (ii) continuously differentiable; and (iii) the first derivative is Lipschitz continuous, i.e., for all $(\boldsymbol{x}, \boldsymbol{y})$, $|g^t(\boldsymbol{x}, \boldsymbol{y}^{(q)}) - g^t(\boldsymbol{x}^{t\partial}_0, \boldsymbol{y}^{t\partial(q)}_0) - \nabla g^t(\boldsymbol{x}^{t\partial}_0, \boldsymbol{y}^{t\partial(q)}_0)'((\boldsymbol{x}, \boldsymbol{y}^{(q)}) - (\boldsymbol{x}^{t\partial}_0, \boldsymbol{y}^{t\partial(q)}_0))| \leq c_2||(\boldsymbol{x}, \boldsymbol{y}^{(q)}) - (\boldsymbol{x}_0, \boldsymbol{y}^{(q)}_0))||^2$ for some positive constant c_2, and for $k = 1, \ldots, p$ and $\ell = 1, \ldots, q-1$, $\frac{\partial}{\partial x^k} g^t(\boldsymbol{x}, \boldsymbol{y}^{(q)})|_{(\boldsymbol{x},\boldsymbol{y}^{(q)})=(\boldsymbol{x}^{t\partial}_0,\boldsymbol{y}^{t\partial(q)}_0)} > 0$ and $\frac{\partial}{\partial y^\ell} g^t(\boldsymbol{x}, \boldsymbol{y}^{(q)})|_{(\boldsymbol{x},\boldsymbol{y}^{(q)})=(\boldsymbol{x}^{t\partial}_0,\boldsymbol{y}^{t\partial(q)}_0)} < 0$.*

Assumptions 2.1 and 2.2 are standard in the economics literature (e.g., see Färe, 1988). Free disposability in Assumption 2.1 imposes monotonicity on the frontier $\mathcal{P}^{t\partial}$, while Assumption 2.2 merely says that there are no free lunches. Assumption 2.3 defines the sampling mechanism. While Assumptions 2.4 and 2.5 involve some complication, they are in the end mild assumptions, imposing weak conditions on the density f^t near the frontier, and some smoothness on the frontier itself.

2.2 Traditional Approaches

All studies of efficiency, productivity, etc., involve comparison of observed performance to some benchmark. In traditional, non-parametric studies, the frontier $\mathcal{P}^{t\partial}$ serves as the benchmark. One can use Shephard (1970) input or output distance functions given by

$$\theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t) \equiv \sup\left\{\theta > 0 \mid (\theta^{-1}\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}^t\right\} \tag{2.3}$$

and

$$\lambda(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t) \equiv \inf\left\{\lambda > 0 \mid (\boldsymbol{x}, \lambda^{-1}\boldsymbol{y}) \in \mathcal{P}^t\right\}, \tag{2.4}$$

(respectively) to measure distance from an arbitrary point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{p+q}_+$ to the boundary $\mathcal{P}^{t\partial}$ in the input direction or the output direction.

Under constant returns to scale (CRS), $\theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t) = \lambda(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t)^{-1}$. However, with variable returns to scale (VRS), the choice of orientation (either input or output) can have a large impact on measured efficiency. As discussed in Section 1, with VRS, a large firm could conceivably lie close to the frontier $\mathcal{P}^{t\partial}$ in the output direction, but far from $\mathcal{P}^{t\partial}$ in the input direction. Similarly, a small firm might lie close to $\mathcal{P}^{t\partial}$ in the input direction, but far from $\mathcal{P}^{t\partial}$ in the output direction. Such differences are related to the slope and curvature of $\mathcal{P}^{t\partial}$. Moreover, there seems to be no criteria telling the applied researcher whether to use the input- or output-orientation.[8]

The hyperbolic distance function

$$\gamma(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t) \equiv \sup\left\{\gamma > 0 \mid (\gamma^{-1}\boldsymbol{x}, \gamma\boldsymbol{y}) \in \mathcal{P}^t\right\} \tag{2.5}$$

reduces this ambiguity by measuring distance from the fixed point $(\boldsymbol{x}, \boldsymbol{y})$ to $\mathcal{P}^{t\partial}$ along the hyperbolic path $(\gamma^{-1}\boldsymbol{x}, \gamma\boldsymbol{y})$, $\gamma \in \mathbb{R}^1_{++}$. Note that for $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}^t$, $\theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t) \geq 1$, $\lambda(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t) \leq 1$, and $\gamma(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t) \geq 1$ by construction.[9] The measures $\theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t)$, $\lambda(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t)$, and $\gamma(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t)$ provide measures of the *technical efficiency* of a firm operating at input/output levels $(\boldsymbol{x}, \boldsymbol{y})$ at time t. Such a firm lying in the interior of $\mathcal{P}^t$ could become technically efficient by moving to either $(\boldsymbol{x}/\theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t), \boldsymbol{y}))$, $(\boldsymbol{x}, \boldsymbol{y}/\lambda(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t))$, or $(\boldsymbol{x}/\gamma(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t), \gamma(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t)\boldsymbol{y})$, or some other point along the frontier $\mathcal{P}^{t\partial}$.

2.3 The Quantile Approach

As discussed below in Section 3.1, estimation of the distance functions defined in (2.3), (2.4), and (2.5) incur the curse of dimensionality as well as (perhaps extreme) sensitivity to outliers. The order-m approach of Cazals et al. (2002) and the order-α approach of Daouia (2003), Aragon et al. (2005) Daouia and Simar (2007) avoid these problems by estimating features *close* to the boundary of the production set, rather than the boundary itself. As noted in Section 1, these partial frontier estimators can be interpreted as estimators of the full frontier $\mathcal{P}^{t\partial}$ when the orders m or α are viewed as sequences (of appropriate order) in

[8] In the case of parametric, stochastic frontier models along the lines of Aigner et al. (1977), one specifies a production, cost, or other relationship, which determines how efficiency is to be measured; e.g., when a production function is specified, efficiency is measured in the output direction. By contrast, the model specified by Assumptions 2.1–2.5 leaves open the question of the direction in which efficiency might be measured.

[9] The Shephard (1970) input and output distance functions defined in (2.3)–(2.4) are reciprocals of the corresponding Farrell (1957) measures. Färe et al. (1985) defined a Farrell-type hyperbolic measure that is the reciprocal of the measure defined here in (2.5).

the sample size. Here and in Section 3.2, we extend the input- and output-oriented approach of Daouia and Simar (2007) to a hyperbolic orientation.

The density $f^t(\boldsymbol{x}, \boldsymbol{y})$ introduced in Assumption 2.4 implies a probability function

$$H^t(\boldsymbol{x}_0, \boldsymbol{y}_0) = \Pr(\boldsymbol{x} \leq \boldsymbol{x}_0, \boldsymbol{y} \geq \boldsymbol{y}_0 \text{ at time } t). \tag{2.6}$$

Although this probability distribution function is non-standard, given the direction of the inequality for $\boldsymbol{y}$, it is well-defined. The function gives the probability of drawing, at time t, an observation from $f^t(\boldsymbol{x}, \boldsymbol{y})$ that weakly *dominates* the firm operating at $(\boldsymbol{x}_0, \boldsymbol{y}_0)$; an observation $(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}})$ weakly dominates $(\boldsymbol{x}_0, \boldsymbol{y}_0)$ if $\widetilde{\boldsymbol{x}} \leq \boldsymbol{x}_0$ and $\widetilde{\boldsymbol{y}} \geq \boldsymbol{y}_0$. Clearly, $H^t(\boldsymbol{x}_0, \boldsymbol{y}_0)$ is monotone, nondecreasing in $\boldsymbol{x}_0$ and monotone, non-increasing in $\boldsymbol{y}_0$.

The idea of *dominance* in the sense used here dates at least to the work of Deprins et al. (1984). As a practical matter, the idea is quite useful from the perspective of managers, policy makers, and others. While a set of firms may be ranked in terms of their estimated technical efficiencies or some other criteria, the manager of an inefficient firm may have little to learn from a more efficient firm unless the two firms use a similar mix of inputs to produce a similar mix of outputs. In other words, the more efficient firm may not be a relevant role model for the less efficient firm if they operate in very different regions of the input-output space. By contrast, a firm that *dominates* a less efficient firm is able to produce more with less, and consequently is likely to have management practices or other features that the less efficient firm should emulate.

Although efficiency is usually measured in either an input or an output direction in non-parametric studies, measurement along a *hyperbolic* path maintains a link with the idea of dominance. Using $H^t(\cdot, \cdot)$, a hyperbolic, α-quantile distance function can be defined by writing

$$\gamma^t_\alpha(\boldsymbol{x}, \boldsymbol{y}) \equiv \sup\left\{\gamma > 0 \mid H^t(\gamma^{-1}\boldsymbol{x}, \gamma\boldsymbol{y}) > (1 - \alpha)\right\} \tag{2.7}$$

for $\alpha \in (0, 1]$. For a fixed point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{p+q}_+$, $\gamma^t_\alpha(\boldsymbol{x}, \boldsymbol{y})$ gives the proportionate, simultaneous reduction in inputs and increase in outputs required to move from $(\boldsymbol{x}, \boldsymbol{y})$ along a path $(\gamma^{-1}\boldsymbol{x}, \gamma\boldsymbol{y})$, $\gamma > 0$, to a point that has probability $(1 - \alpha)$ of being weakly dominated at time t. By construction, for $\alpha \in (0, 1)$, $\gamma^t_\alpha(\boldsymbol{x}, \boldsymbol{y}) < \gamma(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t)$. The hyperbolic α-quantile at time t is defined by

$$\mathcal{P}^{t\partial}_\alpha \equiv \left\{(\gamma^t_\alpha(\boldsymbol{x}, \boldsymbol{y})^{-1}\boldsymbol{x}, \gamma^t_\alpha(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{y}) \mid (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}^t\right\}. \tag{2.8}$$

By definition, $H^t(\boldsymbol{x}_0, \boldsymbol{y}_0)$ is monotone, nondecreasing in $\boldsymbol{x}_0$ and monotone, non-increasing in $\boldsymbol{y}_0$; using this fact it is easy to show that $\mathcal{P}_\alpha^{t\partial}$ is monotone in the sense that if $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathcal{P}_\alpha^{t\partial}$, $(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}) \in \mathcal{P}_\alpha^{t\partial}$, and $\widetilde{\boldsymbol{x}} \geq \boldsymbol{x}_0$, then $\widetilde{\boldsymbol{y}} \geq \boldsymbol{y}_0$.[10]

Figure 1 illustrates the hyperbolic quantile for the simple case where $p = q = 1$ and $f^t(x, y)$ is uniform over the unit triangle with corners at (0,0), (1,0), and (1,1); hence the technology is characterized by constant returns to scale. The solid line shows $\mathcal{P}^{t\partial}$. Given $\alpha \in (0, 1)$, it is straightforward to solve analytically for $\gamma_\alpha^t(x, y)$, and hence the hyperbolic α-quantile can be traced out by solving for $\gamma_\alpha^t(x, y)$ for a variety of pairs (x, y). This has been done in Figure 1 for $\alpha = 0.99$, and $\mathcal{P}_\alpha^{t\partial}$ is illustrated by the dashed line.

Figure 2 provides another illustration of the hyperbolic quantile, again for $p = q = 1$, but with $f^t(x, y)$ uniform over a quarter-circle so that the technology displays variable returns to scale. The solid line shows the full frontier $\mathcal{P}^{t\partial} = \{(x, y) \mid x \in [0, 1],\ y = (2x - x^2)^{1/2}\}$. Some algebra reveals that the marginal distribution function for x, $F_x^t(x_0) = \Pr(x \leq x_0)$ is

$$F_x^t(x_0) = \begin{cases} 1 & \forall\ x_0 \geq 1; \\ \frac{4}{\pi}\left[\frac{x_0 - 1}{2}(2x_0 - x_0^2)^{1/2} + \frac{1}{2}\sin^{-1}(x_0 - 1) + \frac{\pi}{4}\right] & \forall\ x_0 \in (0, 1); \\ 0 & \forall\ x_0 \leq 0. \end{cases} \tag{2.9}$$

Then the joint density $H^t(x_0, y_0)$ is given by

$$H^t(x_0, y_0) = F_x^t(x_0) - F_x^t\left(1 - \left(1 - y_0^2\right)^{1/2}\right) - \frac{4}{\pi} y_0 \left[x_0 - 1 + \left(1 - y_0^2\right)^{1/2}\right] \tag{2.10}$$

for all $x_0 \in [0, 1]$, $y_0 \in [0, (2x_0 - x_0^2)^{1/2}]$. Using (2.10), the hyperbolic α-quantile $\mathcal{P}_\alpha^{t\partial}$ can be traced out; in Figure 2, this has been done for $\alpha = 0.99$, and $\mathcal{P}_\alpha^{t\partial}$ is illustrated by the dashed curve.

The probabilistic formulation used here is closely related to the work of Daouia and Simar (2007), which builds on earlier work by Daouia (2003) and Aragon et al. (2005). Daouia and Simar decompose the distribution function given in (2.6) to obtain

$$\begin{aligned} H^t(\boldsymbol{x}_0, \boldsymbol{y}_0) &= \underbrace{\Pr(\boldsymbol{x} \leq \boldsymbol{x}_0 \mid \boldsymbol{y} \geq \boldsymbol{y}_0)}_{F_{x|y}^t(\boldsymbol{x}_0|\boldsymbol{y}_0)} \underbrace{\Pr(\boldsymbol{y} \geq \boldsymbol{y}_0)}_{S_y^t(\boldsymbol{y}_0)} \\ &= \underbrace{\Pr(\boldsymbol{y} \geq \boldsymbol{y}_0 \mid \boldsymbol{x} \leq \boldsymbol{x}_0)}_{S_{y|x}^t(\boldsymbol{y}_0|\boldsymbol{x}_0)} \underbrace{\Pr(\boldsymbol{x} \leq \boldsymbol{x}_0)}_{F_x^t(\boldsymbol{x}_0)} \end{aligned} \tag{2.11}$$

[10]Suppose $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathcal{P}_\alpha^{t\partial}$. Then $\gamma^t(\boldsymbol{x}_0, \boldsymbol{y}_0) = 1$. Now consider $(\widetilde{\boldsymbol{x}}_0, \boldsymbol{y}_0)$, $\widetilde{\boldsymbol{x}}_0 \geq \boldsymbol{x}_0$. Necessarily, $H^t(\widetilde{\boldsymbol{x}}_0, \boldsymbol{y}_0) \geq H^t(\boldsymbol{x}_0, \boldsymbol{y}_0)$. Therefore $\gamma_\alpha^t(\widetilde{\boldsymbol{x}}_0, \boldsymbol{y}_0) \geq \gamma_\alpha^t(\boldsymbol{x}_0, \boldsymbol{y}_0) = 1$, and $(\gamma_\alpha^t(\widetilde{\boldsymbol{x}}_0, \boldsymbol{y}_0)^{-1}\widetilde{\boldsymbol{x}}_0,\ \gamma_\alpha^t(\widetilde{\boldsymbol{x}}_0, \boldsymbol{y}_0)\boldsymbol{y}_0) \in \mathcal{P}_\alpha^t$, with $\gamma_\alpha^t(\widetilde{\boldsymbol{x}}_0, \boldsymbol{y}_0)\boldsymbol{y}_0 \geq \boldsymbol{y}_0$. Moreover, $\gamma_\alpha^t(\widetilde{\boldsymbol{x}}_0, \boldsymbol{y}_0)^{-1}\widetilde{\boldsymbol{x}}_0 \geq \boldsymbol{x}_0$ since $H^t(\boldsymbol{x}_0, \gamma_\alpha^t(\widetilde{\boldsymbol{x}}_0, \boldsymbol{y}_0)\boldsymbol{y}_0) \leq H^t(\boldsymbol{x}_0, \boldsymbol{y}_0)$. Hence $\mathcal{P}_\alpha^t$ is monotonic in the sense we describe.

(the terms on the right-hand side of (2.11) also appear in Cazals et al., 2002, and Daraio and Simar, 2005). Working in a Farrell-type framework, Daouia and Simar define conditional quantile-based efficiency scores that are equivalent to the reciprocals of the Shephard-type input- and output-oriented conditional α-quantile distance functions given by

$$\theta_\alpha^t(\boldsymbol{x},\boldsymbol{y}) = \sup\left\{\theta \geq 0 \mid F_{x|y}^t(\theta^{-1}\boldsymbol{x} \mid \boldsymbol{y}) > (1-\alpha)\right\} \tag{2.12}$$

and

$$\lambda_\alpha^t(\boldsymbol{x},\boldsymbol{y}) = \inf\left\{\lambda \geq 0 \mid S_{y|x}^t(\lambda^{-1}\boldsymbol{y} \mid \boldsymbol{x}) > (1-\alpha)\right\}. \tag{2.13}$$

For $\alpha \in (0,1)$, $\theta_\alpha^t(\boldsymbol{x},\boldsymbol{y}) < \theta(\boldsymbol{x},\boldsymbol{y} \mid \mathcal{P}^t)$ and $\lambda_\alpha^t(\boldsymbol{x},\boldsymbol{y}) > \lambda(\boldsymbol{x},\boldsymbol{y} \mid \mathcal{P}^t)$ by construction. These implicitly define input and output conditional α-quantile frontiers given by

$$\mathcal{P}_{x,\alpha}^{t\partial} = \left\{(\theta_\alpha^t(\boldsymbol{x},\boldsymbol{y})^{-1}\boldsymbol{x},\boldsymbol{y}) \mid (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{P}^t\right\} \tag{2.14}$$

and

$$\mathcal{P}_{y,\alpha}^{t\partial} = \left\{(\boldsymbol{x},\lambda_\alpha^t(\boldsymbol{x},\boldsymbol{y})^{-1}\boldsymbol{y}) \mid (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{P}^t\right\}, \tag{2.15}$$

respectively.

Returning to the examples in Figures 1 and 2, input and output conditional α-quantiles are shown by the dotted curves for $\alpha = 0.99$. The steeper of the two shows the input-oriented conditional α-quantile; the other shows the output-oriented conditional α-quantile.[11] For any $\alpha \in (0,1)$, these frontiers differ from one another. The input-oriented conditional α-quantile $\mathcal{P}_{x,\alpha}^{t\partial}$ will necessarily have steeper slope than $\mathcal{P}^{t\partial}$, while the output-oriented conditional α-quantile $\mathcal{P}_{y,\alpha}^{t\partial}$ will have less steep slope than $\mathcal{P}^{t\partial}$.[12]

[11] In Figure 2, the input-oriented conditional α-quantile is obtained by deriving the survivor function $S_y^t(y_0) = 1 - F_x^t(1-(1-y_0^2)^{1/2}) - \frac{4}{\pi}y_0(1-y_0^2)\ \forall\ y_0 \in (0,1)$ and then using (2.10) and (2.11) to write $F_{x|y}^t(x_0 \mid y_0)$ as $H^t(x_0,y_0)/S_y^t(y_0)$. Similarly, the survivor function $S_{y|x}^t(y_0 \mid x_0)$ defined in (2.11) is equal to $H^t(x_0,y_0)/F_x^t(x_0)$, with an expression for $F_x^t(x_0)$ given in (2.9). Given $F_{x|y}^t(x_0 \mid y_0)$ and $S_{y|x}^t(y_0 \mid x_0)$, the input- and output-conditional α-quantiles can be traced as in Figure 2.

[12] This point is demonstrated by considering the decomposition in (2.11) and $y_0 = 0$ in either Figure 1 or 2. With $y_0 = 0$, $S_y^t(y_0) = 1$ and the first line of (2.11) yields $H^t(x_0,0) = F_{x|y}^t(x_0,0)$, and hence the input conditional α-quantile and the hyperbolic α-quantile intersect at $y_0 = 0$, as shown in the lower left corners of both Figures 1 and 2. For $y_0 > 0$, however, the first line of (2.11) suggests that $F_{x|y}^t(x_0 \mid y_0)$ must be strictly greater than $H^t(x_0,y_0)$, and hence with $y_0 > 0$, the input conditional α-quantile must lie to the left of and above the hyperbolic α-quantile. Hence the input conditional α-quantile has steeper slope than the hyperbolic α-quantile. Similar reasoning, starting with $x_0 = 1$ and considering the second line of (2.11), reveals that the output conditional α-quantile necessarily has less-steep slope than the hyperbolic α-quantile $\mathcal{P}_\alpha^{t\partial}$.

Before proceeding to a discussion of estimation strategy, note that if $\alpha = 1$, then the input and output conditional α-quantile distance functions $\theta_\alpha^t(\boldsymbol{x}, \boldsymbol{y})$ and $\lambda_\alpha^t(\boldsymbol{x}, \boldsymbol{y})$ defined in (2.12) and (2.13) are equivalent to Shephard (1970) input and output distance functions defined in (2.3) and (2.4). In this case, the distance functions measure distance either in the input direction or the output direction to $\mathcal{P}^{t\partial}$, rather than to a quantile lying within the interior of the set $\mathcal{P}^t$. Similarly, when $\alpha = 1$ the hyperbolic α-quantile distance function defined in (2.7) becomes equivalent to the Shephard-type hyperbolic distance function defined in (2.5). Choosing $\alpha < 1$, however, avoids some of the problems associated with estimation of boundaries of support (or distance to such boundaries) as discussed in the next section.

3 Estimation Methodology

3.1 Estimators for the Traditional Approach

Estimation of the Shephard input and output distance functions defined in (2.3) and (2.4), as well as of the hyperbolic distance function defined in (2.5), requires an estimator of the production set $\mathcal{P}^t$. Deprins et al. (1984) proposed the free-disposal hull (FDH) of the observations in $\mathcal{S}_{n_t}^t$, i.e.,

$$\widetilde{\mathcal{P}}(\mathcal{S}_{n_t}^t) = \bigcup_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{S}_{n_t}^t} \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}_+^{p+q} \mid \boldsymbol{y} \leq \boldsymbol{y}_i,\ \boldsymbol{x} \geq \boldsymbol{x}_i\}. \tag{3.1}$$

The variable-returns-to-scale (VRS) DEA estimator of $\mathcal{P}^t$ is the convex hull of $\widetilde{\mathcal{P}}(\mathcal{S}_{n_t}^t)$, given by

$$\widehat{\mathcal{P}}(\mathcal{S}_{n_t}^t) = \left\{(\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^{p+q} \mid \boldsymbol{y} \leq \boldsymbol{Y}\boldsymbol{\Gamma},\ \boldsymbol{x} \geq \boldsymbol{X}\boldsymbol{\Gamma},\ \boldsymbol{i}_{n_t}'\boldsymbol{\Gamma} = 1,\ \boldsymbol{\Gamma} \geq 0\right\} \tag{3.2}$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $(p \times n_t)$ and $(q \times n_t)$ matrices whose columns are the observed input and output vectors, $\boldsymbol{i}_{n_t}$ is an $(n_t \times 1)$ matrix of ones, and $\boldsymbol{\Gamma}$ is an $(n_t \times 1)$ matrix of weights.

DEA (VRS) estimators of $\theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t)$ and $\lambda(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t)$ are obtained by replacing $\mathcal{P}^t$ with either $\widetilde{\mathcal{P}}^t$ or $\widehat{\mathcal{P}}^t$ in (2.3) and (2.4). When the DEA (VRS) estimator of $\mathcal{P}^t$ is used, the resulting estimators of the input and output distance functions can be written as linear programs which can be solved to obtain estimates (see Simar and Wilson, 2000b for details). DEA (VRS) estimators based on $\widehat{\mathcal{P}}^t$ allow for varying (i.e., increasing, constant, and decreasing) returns to scale. Globally constant returns to scale can be imposed by replacing $\mathcal{P}^t$ in (2.3) and (2.4) with the convex cone of $\widehat{\mathcal{P}}^t$, denoted $\mathcal{V}\left(\widehat{\mathcal{P}}^t\right)$. The estimator $\mathcal{V}\left(\widehat{\mathcal{P}}^t\right)$ is obtained by dropping the constraint $\boldsymbol{i}_{n_t}\boldsymbol{\Gamma} = 1$ in (3.2).[13]

The asymptotic properties of the DEA (VRS) and FDH distance function estimators are discussed in Gijbels et al. (1999), Park et al. (2000), Simar and Wilson (2000b), and Kneip et al. (2007). In particular, consistency of DEA estimators requires assumptions in addition to A1–A4 listed above in Section 2, including convexity of $\mathcal{P}^t$ and sufficient smoothness of $\mathcal{P}^{t\partial}$ (see Simar and Wilson, 2000b for details). Under these assumptions, $\theta(\boldsymbol{x}, \boldsymbol{y} \mid \widehat{\mathcal{P}}^t) = \theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t) + O_p\left(n_t^{-2/(p+q+1)}\right)$ and $\theta(\boldsymbol{x}, \boldsymbol{y} \mid \widetilde{\mathcal{P}}^t) = \theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t) + O_p\left(n_t^{-1/(p+q)}\right)$ and similarly for output-oriented DEA and FDH estimators. The convergence rates are slow, reflecting the curse of dimensionality common to many non-parametric estimators. The FDH estimator has a slower convergence rate than the DEA estimator, but if $\mathcal{P}^t$ is non-convex, then only the FDH estimator is consistent. In addition to slow convergence rates and the curse of dimensionality, the DEA and FDH estimators also suffer from extreme sensitivity to outliers. For many applications, these problems are potentially acute.[14]

3.2 Quantile Estimation

Estimation of $\gamma_\alpha^t(\boldsymbol{x}, \boldsymbol{y})$, and hence $\mathcal{P}_\alpha^{t\partial}$, is straightforward. The empirical analog of the distribution function defined in (2.6) is given by

$$\widehat{H}(\boldsymbol{x}_0, \boldsymbol{y}_0 \mid \mathcal{S}_{n_t}^t) = n_t^{-1} \sum_{i=1}^{n} I(\boldsymbol{x}_i \le \boldsymbol{x}_0, \boldsymbol{y}_i \ge \boldsymbol{y}_0 \mid (\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{S}_{n_t}^t), \tag{3.3}$$

where $I(\cdot)$ denotes the indicator function. Then an estimator of $\gamma_\alpha^t(\boldsymbol{x}, \boldsymbol{y})$ is obtained by replacing $H^t(\cdot, \cdot)$ in (2.7) with $\widehat{H}(\cdot, \cdot \mid \mathcal{S}_{n_t}^t)$ to obtain

$$\widehat{\gamma}_{\alpha,n_t}^t(\boldsymbol{x}, \boldsymbol{y}) = \sup\left\{\gamma > 0 \mid \widehat{H}(\gamma^{-1}\boldsymbol{x}, \gamma\boldsymbol{y} \mid \mathcal{S}_{n_t}^t) > (1-\alpha)\right\}. \tag{3.4}$$

[13] A DEA (VRS) estimator of the hyperbolic distance function $\gamma(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{P}^t)$ can similarly be defined by replacing $\mathcal{P}^t$ with $\widehat{\mathcal{P}}^t$ in (2.5). Unfortunately, however, the estimator cannot be written as a linear program, though it would be easy in principle to adapt the numerical algorithm presented below in Section 3.2 to this problem. Doing so would perhaps result in considerable computational burden with large samples, since numerous linear programs would have to be solved to obtain a single estimate. In the CRS case, $\mathcal{P}^t = \mathcal{V}(\mathcal{P}^t)$ and it is easy to show that in such cases $\gamma(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{V}(\mathcal{P}^t)) = [\theta(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{V}(\mathcal{P}^t))]^{1/2}$. Hence, when returns to scale are globally constant, DEA estimates of the hyperbolic distance function can be computed by first computing input-oriented estimates under CRS and then taking their square roots.

[14] Several algorithms for detecting outliers in high dimensional spaces have been proposed (e.g., Wilson, 1993, 1995; Kuntz and Scholtes, 2000; Simar, 2003; and Porembski et al., 2005), but these involve substantial computational burden with large sample sizes.

Computing $\widehat{\gamma}^t_{\alpha,n_t}(\boldsymbol{x},\boldsymbol{y})$ is essentially a univariate problem. Given a point of interest $(\boldsymbol{x}_0,\boldsymbol{y}_0)$, it is easy to find initial values γ_a,γ_b such that $\gamma_a<\gamma_b$ that bracket the solution so that $\widehat{H}(\gamma_a^{-1}\boldsymbol{x}_0,\gamma_a\boldsymbol{y}_0 \mid \mathcal{S}^t_{n_t}) > (1-\alpha)$ and $\widehat{H}(\gamma_b^{-1}\boldsymbol{x}_0,\gamma_b\boldsymbol{y}_0 \mid \mathcal{S}^t_{n_t}) < (1-\alpha)$, and then solve for $\widehat{\gamma}_{\alpha,n_t}(\boldsymbol{x}_0,\boldsymbol{y}_0)$ using the bisection method. This method can be made accurate to an arbitrarily small degree. The following algorithm describes the procedure:

[1] Set $\gamma_a := 1$, $\gamma_b := 1$.

[2] If $\widehat{H}(\gamma_a^{-1}\boldsymbol{x},\gamma_a\boldsymbol{y} \mid \mathcal{S}^t_{n_t}) \leq (1-\alpha)$ then set $\gamma_a := 0.5\times\gamma_a$.

[3] Repeat step [2] until $\widehat{H}(\gamma_a^{-1}\boldsymbol{x},\gamma_a\boldsymbol{y} \mid \mathcal{S}^t_{n_t}) > (1-\alpha)$.

[4] If $\widehat{H}(\gamma_b^{-1}\boldsymbol{x},\gamma_b\boldsymbol{y} \mid \mathcal{S}^t_{n_t}) \geq (1-\alpha)$ then set $\gamma_b := 2\times\gamma_b$.

[5] Repeat step [4] until $\widehat{H}(\gamma_b^{-1}\boldsymbol{x},\gamma_b\boldsymbol{y} \mid \mathcal{S}^t_{n_t}) < (1-\alpha)$.

[6] Set $\gamma_c := (\gamma_a+\gamma_b)/2$ and compute $\widehat{H}(\gamma_c^{-1}\boldsymbol{x},\gamma_c\boldsymbol{y} \mid \mathcal{S}^t_{n_t})$.

[7] If $\widehat{H}(\gamma_c^{-1}\boldsymbol{x},\gamma_c\boldsymbol{y} \mid \mathcal{S}^t_{n_t}) \leq (1-\alpha)$ then set $\gamma_b := \gamma_c$; otherwise set $\gamma_a := \gamma_c$.

[8] If $(\gamma_b-\gamma_a) > \epsilon$, where ϵ is a suitably small tolerance value, repeat steps [6]–[7].

[9] If $\widehat{H}(\gamma_c^{-1}\boldsymbol{x},\gamma_c\boldsymbol{y} \mid \mathcal{S}^t_{n_t}) \leq (1-\alpha)$ set $\widehat{\gamma}_{\alpha,n_t}(\boldsymbol{x},\boldsymbol{y}) := \gamma_a$; otherwise set $\widehat{\gamma}_{\alpha,n_t}(\boldsymbol{x},\boldsymbol{y}) := \gamma_c$.

Note that finding γ_a first reduces the computational burden. Given γ_a, γ_b can be found using only the subset of sample observations that dominate the point $(\gamma_a^{-1}\boldsymbol{x},\gamma_a\boldsymbol{y})$. Moreover, only this same subset of observations need be used in the first pass through steps [6]–[7]. Upon reaching step [8], the relevant subset of observations can be further reduced each time γ_a is reset in step [7]. Setting the convergence tolerance ϵ in step [8] to 10^{-6} will yield solutions accurate to 5 decimal places, which is likely to be sufficient for most applications.[15] The algorithm presented here has been implemented in the freely-available *FEAR* library provided by Wilson (2007).

[15] Note that many rational decimal fractions become irrational numbers in the base-2 representation used by modern digital computers; e.g., the base-10 fraction 0.95 has no exact representation in base-2. To avoid problems with the logical comparisons in steps [7] and [9], comparisons should be made against $(1-\alpha-\nu)$ instead of $(1-\alpha)$, where ν is the smallest positive real number that can be represented on the computer architecture in use that yields the result $1-\nu\neq 1$. For machines using 64-bit IEEE arithmetic, this number is $2^{-53}\approx 1.110223\times 10^{-16}$.

Daouia and Simar (2007) describe a method for finding exact solutions for their input- and output-oriented conditional hyperbolic quantile estimators. It is similarly possible to obtain exact solutions for the unconditional hyperbolic quantile estimator; see Wheelock and Wilson (2007) for details. However, due to storage requirements, sorting, and the large number of logical comparisons required by the exact method, computing $\widehat{\gamma}_{\alpha,n_t}(\boldsymbol{x},\boldsymbol{y})$ using the bisection method is much faster than the exact method. Given that (i) in any dataset, variables typically have at most only a few significant digits, and (ii) our numerical solution can be made accurate to an arbitrary degree by choosing a suitably small value of the tolerance value ϵ, there seems to be no disadvantage in using the numerical procedure in applied research.

Some asymptotic results from Daouia and Simar (2007) have been extended to the hyperbolic α-quantile distance function estimator $\widehat{\gamma}^t_{\alpha,n_t}(\boldsymbol{x},\boldsymbol{y})$ of $\gamma^t_\alpha(\boldsymbol{x},\boldsymbol{y})$. First, Wheelock and Wilson (2007, Theorem 4.2) establish that $\widehat{\gamma}^t_{\alpha,n_t}(\boldsymbol{x},\boldsymbol{y})$ converges completely: $\widehat{\gamma}^t_{\alpha,n_t}(\boldsymbol{x},\boldsymbol{y}) \xrightarrow{c} \gamma^t_\alpha(\boldsymbol{x},\boldsymbol{y})$ as $n_t \rightarrow \infty$.[16] Hence, $\widehat{\gamma}^t_{\alpha,n_t}(\boldsymbol{x},\boldsymbol{y})$ is a strongly consistent estimator of $\gamma^t_\alpha(\boldsymbol{x},\boldsymbol{y})$.

In addition, assuming $\alpha \in (0,1)$ and $H^t(\gamma^{-1}\boldsymbol{x},\gamma\boldsymbol{y})$ is differentiable with respect to γ near $\gamma = \gamma^t_\alpha(\boldsymbol{x},\boldsymbol{y})$, Wheelock and Wilson (2007, Theorem 4.3) establish that $\widehat{\gamma}^t_{\alpha,n_t}(\boldsymbol{x},\boldsymbol{y})$ is asymptotically normally distributed, with convergence at the classical, parametric root-n rate; i.e.,

$$\sqrt{n_t}\left(\widehat{\gamma}^t_{\alpha,n_t}(\boldsymbol{x},\boldsymbol{y}) - \gamma^t_\alpha(\boldsymbol{x},\boldsymbol{y})\right) \xrightarrow{d} N\left(0, \sigma^2_\alpha(\boldsymbol{x},\boldsymbol{y} \mid t)\right) \tag{3.5}$$

where

$$\sigma_\alpha(\boldsymbol{x},\boldsymbol{y} \mid t)^2 = \alpha(1-\alpha)\left[\frac{\partial H^t\left(\gamma_\alpha(\boldsymbol{x},\boldsymbol{y})^{-1}\boldsymbol{x}, \gamma_\alpha(\boldsymbol{x},\boldsymbol{y})\boldsymbol{y}\right)}{\partial \gamma_\alpha(\boldsymbol{x},\boldsymbol{y})}\right]^{-2}. \tag{3.6}$$

Consequently, the hyperbolic quantile efficiency estimator $\widehat{\gamma}^t_{\alpha,n_t}(\boldsymbol{x},\boldsymbol{y})$ does not suffer from the curse of dimensionality that plagues most non-parametric estimators since its convergence rate depends solely on the sample size n_t and involves neither p nor q. These results are not surprising, given that similar results were obtained by Daouia and Simar (2007) for estimators of the input- and output-oriented conditional α-quantile distance functions defined in (2.12)–(2.13).

[16] A sequence of random variables $\{\zeta_n\}_{n=1}^\infty$ converges completely to a random variable ζ, denoted by $\zeta_{n_t} \xrightarrow{c} \zeta$, if $\lim_{n\rightarrow\infty}\sum_{j=1}^n \Pr\left(|\zeta_j - \zeta| \geq \epsilon\right) < \infty \;\forall\; \epsilon > 0$. This type of convergence was introduced by Hsu and Robbins (1947). Complete convergence implies, and is a stronger form of convergence than almost-sure convergence, which implies and is stronger than convergence in probability.

It is important to note that the results described above do not hold if $\alpha = 1$. In particular, if $\alpha = 1$, the hyperbolic α-quantile estimator defined in (3.4) measures distance along a hyperbolic path to the FDH of the sample data (see equation (3.1)). In this case, the estimator has an asymptotic Weibull distribution, with convergence rate $n_t^{-1/(p+q)}$. Similarly, if $\alpha = 1$, the conditional α-quantile distance function estimators described by Daouia and Simar (2007) become equivalent to FDH estimators of Shephard (1970) input and output distance functions, which converge at the rate $n^{-1/(p+q)}$, as discussed in Section 3.1.

On the other hand, if α is viewed a sequence in n_t tending to 1 (at an appropriate rate) as $n_t \to \infty$, then an interesting interpretation is possible. Daouia and Simar (2007) show that for their conditional α-quantile distance function estimators, allowing $\alpha \to 1$ as $n_t \to \infty$, permits their estimators to be interpreted as robust estimators of distance to the full frontier $\mathcal{P}^{t\partial}$, rather than of distance to the conditional α-quantiles $\mathcal{P}^{t\partial}_{x,\alpha}$ and $\mathcal{P}^{t\partial}_{y,\alpha}$. Similar results hold for the hyperbolic α-quantile distance function estimator. Wheelock and Wilson (2007, Theorem 4.4) show that provided Assumptions 2.3–2.5 hold and the order of $\alpha(n_t) > 0$ is such that $n_t^{(p+q+1)/(p+q)}(1-\alpha(n)) \to 0$ as $n_t \to \infty$, then for any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}^t$,

$$n_t^{1/(p+q)} \left(\gamma^t(\boldsymbol{x}, \boldsymbol{y}) - \widehat{\gamma}^t_{\alpha(n_t),n_t}(\boldsymbol{x}, \boldsymbol{y})\right) \xrightarrow{d} \text{Weibull}\left(\mu^{p+q}_{\mathcal{H},0}, p+q\right) \tag{3.7}$$

where $\mu_{\mathcal{H},0}$ is a constant. An expression for $\mu_{\mathcal{H},0}$ as well as proofs of the asymptotic results listed above are given in Wheelock and Wilson (2007). The result in (3.7) means that the hyperbolic α-quantile distance function estimator can be seen as a robust estimator of the hyperbolic distance (to the full frontier $\mathcal{P}^{t\partial}$) defined in (2.5) when α is regarded as a sequence in n tending to 1 at the appropriate rate. Consequently, estimators of the indices for efficiency change, productivity change, and technical change defined below Section 4 can be viewed in terms of either partial or full frontiers. Of course, when viewed as an estimator of distance to the full frontier, $\widehat{\gamma}^t_{\alpha(n),n}(\boldsymbol{x}, \boldsymbol{y})$ trades its root-n convergence rate when α is fixed for the slow convergence rate of FDH estimators, but retains the advantage of robustness.

4 Measuring and Estimating Changes in Performance

Given the interest in technical efficiency in cross-sectional contexts, it is natural to ask how efficiency evolves over time. In competitive industries, one would expect inefficient firms to be driven from the market, but this does not happen instantaneously, and firms that are inefficient today may become more efficient tomorrow and vice-versa.

Using the hyperbolic measure of efficiency defined in (2.7), an estimate of efficiency change is given by the ratio $\gamma_\alpha^{t_2}(\boldsymbol{x}_{it_2}, \boldsymbol{y}_{it_2})/\gamma_\alpha^{t_1}(\boldsymbol{x}_{it_1}, \boldsymbol{y}_{it_1})$, where $(\boldsymbol{x}_{it_1}, \boldsymbol{y}_{it_1})$ and $(\boldsymbol{x}_{it_2}, \boldsymbol{y}_{it_2})$ denote firm i's location in the input-output space at times $t_1 < t_2$. A value less than (equal to, greater than) unity indicates an increase (no change, a decrease) in technical efficiency measured relative to the α-quantiles at times t_1 and t_2.

To get an idea of industry-wide performance, one might consider *mean* changes in efficiency; given the multiplicative nature of the efficiency measures, researchers typically use geometric, rather than arithmetic, means. Let $\mathcal{I}(t_1, t_2)$ be the set of firms in existence at *both* times t_1 and t_2, and let $\#\mathcal{I}(t_1, t_2)$ denote the number of firms in this set. Then, for the quantile-based measure of efficiency, define

$$\mathcal{E}_\alpha(t_1, t_2) = \left[\prod_{i \in \mathcal{I}(t_1,t_2)} \frac{\gamma_\alpha^{t_2}(\boldsymbol{x}_{it_2}, \boldsymbol{y}_{it_2})}{\gamma_\alpha^{t_1}(\boldsymbol{x}_{it_1}, \boldsymbol{y}_{it_1})} \right]^{1/\#\mathcal{I}(t_1,t_2)}. \tag{4.1}$$

$\mathcal{E}_\alpha(t_1, t_2)$ measures the (geometric) mean change in efficiency between times t_1 and t_2, relative to the unconditional, hyperbolic α-quantiles at times t_1 and t_2.

In the case of one input and one output, one can judge productivity simply by the ratio of output to input quantities. If $\mathcal{P}^{t\partial}$ exhibits constant returns to scale everywhere, there is little difference between productivity and efficiency, although they might be measured differently. With variable returns to scale, however, technically efficient firms operating along $\mathcal{P}^{t\partial}$ in regions of either increasing or decreasing returns to scale will be less productive than technically efficient firms operating along the constant-returns region of $\mathcal{P}^{t\partial}$; they may also be less productive than some technically inefficient firms.

In cases of multiple inputs and multiple outputs, productivity cannot be measured reliably by simple ratios. Instead, in dynamic contexts, Malmquist indices are typically used to measure *changes* in productivity. These indices are usually defined in terms of the Shephard input and output distance functions defined in (2.3) and (2.4), which in turn are estimated by the DEA estimators discussed in Section 3; see Färe and Grosskopf (1996) for examples and discussion.

Malmquist indices can also be defined in terms of hyperbolic quantile measures. Define

$$\mathcal{P}_{\alpha}^{t} = \left\{(\boldsymbol{x},\boldsymbol{y}) \mid \boldsymbol{x} \geq \widetilde{\boldsymbol{x}},\ \boldsymbol{y} \in [0,\widetilde{\boldsymbol{y}}]\ \forall\ (\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}}) \in \mathcal{P}_{\alpha}^{t\partial}\right\}. \tag{4.2}$$

Then $\mathcal{P}_{\alpha}^{t\partial}$ is the closure of the compliment of the closed set $\mathcal{P}_{\alpha}^{t}$, just as $\mathcal{P}^{t\partial}$ is the closure of the compliment of $\mathcal{P}^{t}$. Recall that the operator $\mathcal{V}(\cdot)$ was introduced in Section 3.1 to denote the convex cone of a set in $\mathbb{R}_{+}^{p+q}$; hence $\mathcal{P}_{\alpha}^{t} \subseteq \mathcal{V}(\mathcal{P}_{\alpha}^{t})$. Also note that different distance functions can be defined by replacing $\mathcal{P}^{t}$ in (2.5) with some other set to measure distance from $(\boldsymbol{x},\boldsymbol{y})$ to the boundary of the other set; e.g., $\gamma(\boldsymbol{x},\boldsymbol{y} \mid \mathcal{V}(\mathcal{P}_{\alpha}^{t}))$ measures distance from $(\boldsymbol{x},\boldsymbol{y})$ along a hyperbolic path $(\gamma^{-1}\boldsymbol{x},\gamma\boldsymbol{y})$, $\gamma > 0$, the boundary of the set $\mathcal{V}(\mathcal{P}_{\alpha}^{t})$.

A hyperbolic-quantile-based Malmquist index is defined by the geometric mean

$$\mathcal{M}_{\alpha}(t_1,t_2) \equiv \left\{ \left[\prod_{i\in\mathcal{I}(t_1,t_2)} \frac{\gamma\left(\boldsymbol{x}_{it_2},\boldsymbol{y}_{it_2} \mid \mathcal{V}(\mathcal{P}_{\alpha}^{t_1})\right)}{\gamma\left(\boldsymbol{x}_{it_1},\boldsymbol{y}_{it_1} \mid \mathcal{V}(\mathcal{P}_{\alpha}^{t_1})\right)} \times \frac{\gamma\left(\boldsymbol{x}_{it_2},\boldsymbol{y}_{it_2} \mid \mathcal{V}(\mathcal{P}_{\alpha}^{t_2})\right)}{\gamma\left(\boldsymbol{x}_{it_1},\boldsymbol{y}_{it_1} \mid \mathcal{V}(\mathcal{P}_{\alpha}^{t_2})\right)} \right]^{\frac{1}{2}} \right\}^{\frac{1}{\#\mathcal{I}(t_1,t_2)}}. \tag{4.3}$$

This index provides a measure of the mean change in the productivity of firms from time t_1 to t_2. The index is analogous to those proposed by Färe et al. (1992, 1994), but with two important differences. First, productivity is benchmarked against the boundaries of $\mathcal{V}(\mathcal{P}_{\alpha}^{t_1})$ and $\mathcal{V}(\mathcal{P}_{\alpha}^{t_2})$, rather than the boundaries of $\mathcal{V}(\mathcal{P}^{t_1})$ and $\mathcal{V}(\mathcal{P}^{t_2})$. Second, the hyperbolic direction is used, rather than an input or output direction. As noted in the introduction, measurement along hyperbolic paths avoids the ambiguity discussed in Section 1.

The index $\mathcal{M}_{\alpha}(t_1,t_2)$ in (4.3) is a geometric mean of two ratios appearing inside the square brackets in (4.3). The first ratio inside the square brackets measures the change in productivity using as a benchmark the convex cone of the set bounded by the hyperbolic α-quantile $\mathcal{P}_{\alpha}^{t_1\partial}$ prevailing at time t_1, while the second ratio measures productivity change using as a benchmark the convex cone of the set bounded by the quantile $\mathcal{P}_{\alpha}^{t_2\partial}$ prevailing at time t_2. A particular firm either moves closer to each benchmark (becoming more productive), farther from each benchmark (becoming less productive), or closer to one and farther from the other. Values of the Malmquist index less than (equal to, greater than) unity indicate an increase (no change, a decrease) in productivity.

Malmquist indices can be decomposed to identify the sources of changes in productivity, and various decompositions of output- and input-oriented Malmquist indices have been proposed in the literature (see Wheelock and Wilson, 1999 for an example). Although many decompositions are possible, a measure of efficiency change such as the one defined in (4.1) and a measure of technical change are common to most decompositions that have appeared in the literature. In terms of hyperbolic α-quantiles, industry-wide technical change is measured by the geometric mean

$$\mathcal{T}_\alpha(t_1,t_2) \equiv \left\{ \left[\prod_{i\in\mathcal{I}(t_1,t_2)} \frac{\gamma_\alpha^{t_1}\left(\boldsymbol{x}_{it_1},\boldsymbol{y}_{it_1}\right)}{\gamma_\alpha^{t_2}\left(\boldsymbol{x}_{it_1},\boldsymbol{y}_{it_1}\right)} \times \frac{\gamma_\alpha^{t_1}\left(\boldsymbol{x}_{it_2},\boldsymbol{y}_{it_2}\right)}{\gamma_\alpha^{t_2}\left(\boldsymbol{x}_{it_2},\boldsymbol{y}_{it_2}\right)} \right]^{\frac{1}{2}} \right\}^{\frac{1}{\#\mathcal{I}(t_1,t_2)}}, \tag{4.4}$$

where $\gamma_\alpha^{t_k}\left(\boldsymbol{x}_{it_j},\boldsymbol{y}_{it_j}\right)$ measures distance from firm i's location at time t_j to the hyperbolic α quantile $\mathcal{P}_\alpha^{t_k}$ prevailing at time t_k, along a hyperbolic path.

The term inside the braces in (4.4) is a geometric mean of two ratios that measure the shift in the α-quantile relative to the ith firm's position at times t_1 and t_2. The first ratio will be less than (equal to, greater than) unity when distance from the point $(\boldsymbol{x}_{it_1},\boldsymbol{y}_{it_1})$ along the hyperbolic path $(\gamma^{-1}\boldsymbol{x}_{it_1},\gamma\boldsymbol{y}_{it_1})$, $\gamma>0$, to the hyperbolic α-quantile increases (remains the same, decreases) from time t_1 to t_2. Similarly, the second ratio will be less than (equal to, greater than) unity when distance from the point $(\boldsymbol{x}_{it_2},\boldsymbol{y}_{it_2})$ along the hyperbolic path $(\gamma^{-1}\boldsymbol{x}_{it_2},\gamma\boldsymbol{y}_{it_2})$, $\gamma>0$, to the hyperbolic α-quantile increases (remains the same, decreases) from time t_1 to t_2. Hence $\mathcal{T}_\alpha(t_1,t_2)(<,=,>)1$ indicates that on average, the hyperbolic α-quantile (shifts outward, remains unchanged, shifts inward).

Finally, note that one can define estimators of $\mathcal{E}_\alpha(t_1,t_2)$ and $\mathcal{T}_\alpha(t_1,t_2)$ by replacing the distance functions on the right-hand sides of (4.1) and (4.4) with the corresponding quantile-based estimators discussed previously in Section 3. To estimate $\mathcal{M}_\alpha(t_1,t_2)$ in (4.3), note that an observation $(\boldsymbol{x}_{it},\boldsymbol{y}_{it})$ can be projected onto the estimated hyperbolic α-quantile $\widehat{\mathcal{P}}_\alpha^{t\partial}$ by computing $(\widehat{\gamma}_\alpha^t(\boldsymbol{x}_{it},\boldsymbol{y}_{it})^{-1}\boldsymbol{x}_{it},\ \widehat{\gamma}_\alpha^t(\boldsymbol{x}_{it},\boldsymbol{y}_{it})\boldsymbol{y}_{it})$. Define the set of projected observations

$$\mathcal{S}_{n_tt}^* = \left\{\left(\widehat{\gamma}_\alpha^t(\boldsymbol{x}_{it},\boldsymbol{y}_{it})^{-1}\boldsymbol{x}_{it},\ \widehat{\gamma}_\alpha^t(\boldsymbol{x}_{it},\boldsymbol{y}_{it})\boldsymbol{y}_{it}\right)\right\}_{i=1}^{n_t}. \tag{4.5}$$

Then the hyperbolic distance functions that appear in (4.3) can be estimated by computing distance function estimates $\gamma(\boldsymbol{x},\boldsymbol{y}\mid\mathcal{V}(\widehat{\mathcal{P}}(\mathcal{S}_{n_tt}^*)))=\left[\theta(\boldsymbol{x},\boldsymbol{y}\mid\mathcal{V}(\widehat{\mathcal{P}}(\mathcal{S}_{n_tt}^*)))\right]^{1/2}$ by solving the appropriate linear programs and taking square roots (see footnote 13 for discussion).

5 Bank Production and Data

Distance function estimation using the estimators described in Section 2 requires the specification of production inputs and outputs. For our study of commercial banks, we define five inputs and five outputs which, with one exception (the measure of off-balance sheet output), are those used by Berger and Mester (2003). Our inputs are purchased funds, which consists of time deposits over $100,000, foreign deposits, federal funds purchased, and various other borrowed funds; core deposits, which consists of domestic transactions accounts, time deposits under $100,000 and savings deposits; labor; physical capital, which consists of premises and other fixed assets; and financial equity capital. Our outputs are consumer loans, business loans, real estate loans, securities, and off-balance sheet items, which consist of total non-interest income minus service charges on deposits.[17] With the exception of labor input (which is measured as full-time equivalent employees) and off-balance sheet items (which are measured in terms of net flow of income), inputs and outputs are stocks measured by dollar amounts reported on bank balance sheets, rather than number of loans or deposits, or loan income or deposit interest expenses. This approach is consistent with the widely used "intermediation" model of Sealey and Lindley (1977).

Our data come from Reports of Income and Condition (Call Reports) for all U.S. commercial banks at year-end 1985, 1994, and 2004. We omitted banks with missing or negative values for any input or output, and converted dollar values to constant year-2000 prices using the GDP deflator. Our sample consists of 11,993, 9,585, and 6,075 observations for 1985, 1994, and 2004, respectively, and comprises at least 95 percent of all commercial banks in operation in these years. Table 1 reports descriptive statistics for each input and output.

Although our annual sample sizes may seem large, at least by parametric standards, they are in fact small for the non-parametric DEA and FDH estimators given the high dimensionality of our application. With five inputs (p) and five outputs (q), we have $(p+q) = 10$ dimensions. The potential for the curse of dimensionality to affect DEA and FDH estimation can be gaged by a rough comparison of equivalent sample sizes. For the quantile, DEA, and FDH

[17]Of the various commonly used measures of off-balance sheet output, this definition is the most consistently measurable across banks and over time. We also used an "asset-equivalent" definition of off-balance sheet items in a second set of estimates and obtained qualitatively identical results except as noted below. See Clark and Siems (2002) for discussion of alternative measures of off-balance sheet items, and Berger and Mester (2003) for additional details about the computation of the other inputs and outputs.

estimators, we have convergence rates of $n^{-1/2}$, $n^{-2/(p+q+1)}$, and $n^{-1/(p+q)}$, respectively. Thus, to achieve the same order of magnitude in estimation error as obtained with the quantile estimator with $n = 100$ observations, the DEA estimator would require $\left(100^{-1/2}\right)^{-11/2} = 316,227$ observations, while the FDH estimator would require $\left(100^{-1/2}\right)^{-10} = 10^5$ observations.[18]

6 Empirical Results

We computed efficiency estimates for all U.S. commercial banks in 1985, 1994, and 2004, using the hyperbolic α-quantile estimator described previously. Table 2 reports the mean estimated change in efficiency $\mathcal{E}_\alpha(t_1, t_2)$ between 1985 and 1994 for banks in each asset-size quartile.[19] Quartile "Q1" consists of banks in the smallest-size quartile in a given year, "Q2" consists of those in the next smallest-size quartile, etc. Because the distribution of banks by size is extremely skewed, we divide "Q4," the largest-size quartile, into two groups. "Q4a" consists of banks in the largest-size quartile with total assets of less than \$1 billion, and "Q4b" consists of those with at least \$1 billion of total assets in a given year.[20] In empirical applications, a value for α must be chosen; apart from the theoretical properties mentioned above in Section 3.2 that follow from viewing α as a sequence converging to 1 as $n \to \infty$, our sample is finite, with fixed sample size. After examining results obtained with $\alpha = 0.95$, 0.99, and 0.999, we found that our qualitative results were robust with respect to the various choices of values for α. We report results in Table 2 (as well as in all of the tables that

[18] Illustrating the curse of dimensionality, FDH estimates of efficiency equal 1.0 for all observations in each year; in other words *all* observations lie on the FDH frontier. By contrast, only 7.9 to 8.8 percent of the sample observations in 1985, 1994, or 2004 yield DEA efficiency estimates equal to 1.0. Thus, all of the apparent inefficiency indicated by DEA estimates is due solely to the incorporation of an assumption of convexity on the production set by the DEA estimator. Recall that the DEA frontier estimator is merely a convexification of the FDH frontier estimator; the result here implies that many observations in a given year that lie on the FDH frontier estimate lie below facets of the DEA frontier estimate. Projecting a given number of observations in increasing numbers of orthogonal directions (i.e., increasing dimensionality of the problem) necessarily increases the chance that a given observation will become undominated by any of the other observations, and hence will lie on the FDH frontier estimate.

[19] Empirical results were obtained using the FEAR library (Wilson, 2007).

[20] The first, second, and third quartiles of the distribution consist of banks with total assets in the following ranges: For 1985: \$0–\$27.760, \$27.760–\$53.590, and \$53.590–\$286.700 million (year 2000 dollars); for 1994: \$0–\$32.900, \$32.900–\$62.960, and \$62.960–\$354.500 million (year 2000 dollars); and for 2004: \$0–\$48.120, \$48.120–\$92.210, and \$92.210–\$511.900 million (year 2000 dollars). In both 1985 and 1994, 349 banks had total assets exceeding \$1 billion (year 2000 dollars), and in 2004, 230 banks had total assets exceeding \$1 billion (year 2000 dollars).

follow) obtained with $\alpha = 0.99$.[21]

Row labels in Table 2 refer to the quartiles for 1985 and the column labels refer to the quartiles for 1994. Thus, for example, the upper-left most cell reports the geometric mean estimate of efficiency change for the 1,480 banks that were in quartile "Q1" in both 1985 and 1994. Mean values greater than 1 indicate a decrease in efficiency, whereas those less than 1 indicate an increase in efficiency. Changes that are statistically significant at 90, 95, and 99 percent are indicated by one, two, or three asterisks.[22]

Whereas the mean change in efficiency for the 1,480 banks in the smallest-size quartile in both 1985 and 1994 is not significantly different from zero, the 361 banks that moved from the smallest-size quartile ("Q1") in 1985 to the second quartile ("Q2") in 1994 experienced a statistically-significant mean efficiency decline of 7.5 percent. Similarly, the 88 banks in "Q1" in 1985 that moved to "Q3" in 1994 had a mean decline of 4.85 percent.

Not all groups of banks experienced a decline in efficiency. For example, the 22 banks that moved from "Q1" in 1985 to "Q4a" in 1994 had a mean efficiency increase of 17.2 percent. Our results indicate that larger banks tended to experience larger efficiency gains than small banks, with the largest gains obtained by banks in the largest-size quartile in 1994, regardless of their size in 1985.

Table 3 reports mean estimates of efficiency change for U.S. banks between 1994 and 2004. Qualitatively, the results are similar to those for 1985-1994. In general, banks with at least \$1 billion of assets experienced the largest efficiency gains. The only statistically

[21]Daouia and Simar (2007) employ a similar strategy in their example involving mutual funds data (Section 6.3 of their paper), and also note that for their results, "the choice of α is not so important." As discussed in Sections 1 and 3, the attractive statistical properties of the hyperbolic quantile estimator stem from the fact that quantiles lying *close* to the boundary of support of the input-output distribution, but not the boundary itself, are estimated. For a given sample size, α should be chosen "close" to 1, but not so close that the estimator collapses to the FDH estimator, losing the root-n convergence rate, etc. With $\alpha = 0.99$, one-percent of the sample will dominate any point projected onto the α-quantile. Daouia and Simar (Section 6.3) remark that in their example, $\alpha = 0.99$ yields results close to the FDH case, but their data include only 129 observations. In our application, the sample size is much larger in each of the three years we examine, and consequently even with $\alpha = 0.999$, our results are somewhat different from the FDH case.

[22]Under the null hypothesis of no change in efficiency, the measure defined in (4.1) equals 1. Replacing the unknown, true distance functions on the right-hand side of (4.1) with the corresponding estimators and then taking logs on both sides yields $\log\left(\widehat{\mathcal{E}}_\alpha(t_1,t_2)\right) = \frac{1}{\#\mathcal{I}(t_1,t_2)}\sum_{i\in\mathcal{I}(t_1,t_2)}\left[\log\left(\widehat{\gamma}_\alpha^{t_2}(\boldsymbol{x}_{it_2},\boldsymbol{y}_{it_2})\right) - \log\left(\widehat{\gamma}_\alpha^{t_1}(\boldsymbol{x}_{it_1},\boldsymbol{y}_{it_1})\right)\right]$. Under the null, $\widehat{\gamma}_\alpha^{t_2}(\boldsymbol{x}_{it_2},\boldsymbol{y}_{it_2})$ and $\widehat{\gamma}_\alpha^{t_1}(\boldsymbol{x}_{it_1},\boldsymbol{y}_{it_1})$ (and their logs) have the same mean; hence $\log\left(\widehat{\mathcal{E}}_\alpha(t_1,t_2)\right)$ has mean zero. The Lindeberg-Feller central limit theorem yields asymptotic normality, allowing significance testing.

significant decline in mean efficiency is for "Q1" banks in 1994 that moved to "Q2" in 2004.

Finally, Table 4 reports mean estimates of efficiency change for 1985 to 2004. Qualitatively, the estimates are similar to those for the 1985-94 and 1994-2004 sub-periods. In general, banks that had at least \$1 billion of assets in 2004 experienced the largest efficiency improvement. Again, however, the only statistically significant estimated decline in efficiency is for "Q1" banks in 1985 that moved to "Q2" in 2004.

Because efficiency is measured relative to the position of the quantile, observed changes in the efficiency of a bank over time can result either from a change in the bank's productivity, movement of the quantile, or from a combination of the two. Tables 5–7 report estimates of mean productivity change $\mathcal{M}(t_1, t_2)$ for 1985-1994, 1994-2004, and 1985-2004, respectively. Values less than 1 indicate productivity improvement, whereas values greater than 1 reflect a decline in productivity.

For 1985-94, the mean estimates reflect a general improvement in productivity across banks. For example, the banks that had at least \$1 billion of assets in both 1985 and 1994 ("Q4b") experienced a mean estimated productivity gain of 7.1 percent. Smaller banks tended to fair less well. The mean estimates of productivity change for banks in the two smallest-size quartiles in 1994 ("Q1" and "Q2") are not statistically different from zero, except for banks that moved from the smallest-size quartile ("Q1") in 1985 to the second quartile ("Q2") in 1994. For those banks we estimate a mean *decline* in productivity of 3.5 percent, which accounts for at least some of their estimated decline in efficiency.

The estimates reported in Table 6 indicate that productivity generally declined between 1994 and 2004 for all but the largest banks. For example, we estimate a mean productivity decline of 10.3 percent for banks located in quartile "Q1" in both years. Our estimates indicate declines of a similar order of magnitude for banks in the smallest-size quartiles, and statistically insignificant productivity gains for banks with at least \$1 billion of assets.[23]

Finally, for the entire period 1985-2004, our estimates indicate statistically significant productivity declines for smaller banks, and statistically significant improvements for larger banks. As shown in Table 7, we find a tendency for banks in the smallest three size quartiles in 2004 to have become less productive, regardless of their size in 1985. For banks with at

[23]We obtain statistically significant estimates of productivity improvement in the range of 5–15 percent for banks in the fourth quartile in 1994 that had at least \$1 billion of assets in 2004 ("Q4b"), when we replace the non-interest income measure of off-balance sheet items with an asset-equivalent measure.

least \$1 billion of assets in 2004, however, we find a tendency for productivity gains. For example, for banks with at least \$1 billion of assets in both 1985 and 2004, we estimate a mean improvement of 11.2 percent.

Tables 8–10 report estimates of mean technology change $\mathcal{T}(t_1, t_2)$, which reflect movements in the α-quantile. Values less than 1.0000 reflect outward movement of the quantile, whereas values greater than 1.0000 reflect inward movement, or reduced production possibilities. For 1985–94, the mean estimates by quartile, though statistically significant, are generally small. The estimated mean shift in the quantile for banks with at least \$1 billion of assets in both years, however, is a 14.5 percent outward movement. By contrast, for 1994-2004, our estimates indicate a general inward movement of the quantile, as do our estimates for 1985-2004 as a whole. For example, for banks in quartile "Q1" in both years, we estimate a mean 13.1 percent inward shift in the quantile between 1985 and 2004.

In addition to the estimates reported in Tables 2–10, which are based on the hyperbolic distance measure, we also produced estimates of efficiency change, productivity change, and technical change using estimates of the input- and output-oriented distance functions defined in (2.12) and (2.13).[24] Qualitatively, our estimates are similar across the three estimators for banks in the mid-size range, but in a number of cases are quite different for banks that are in the smallest or largest size groups in one or both periods. For banks at the extremes of the size distribution, the input- and output-oriented distance measures sometimes lead to opposite conclusions about the direction, as well as magnitude, of any change in performance. As discussed in Section 1, large banks that lie close to the quantile (or the full frontier) in the output-direction—and consequently, also in the hyperbolic-direction—may lie quite distant from the quantile (or the full frontier) in the input-direction. Similarly, very small banks that lie close to the quantile or full frontier in the input-direction—and consequently also in the hyperbolic-direction—may lie quite distant in the output-direction. Not surprisingly, we find that for large banks, estimates of efficiency, productivity, and technical change based on the hyperbolic distance measure are more similar to those produced by the output-oriented measure than they are to those produced by the input-measure, whereas for small banks, the hyperbolic-based estimates are more like those produced by the input-oriented measure.

[24] These results are available from the authors on request; we have not included these results here, in order to conserve space.

Overall, our estimation results based on hyperbolic α-quantiles indicate that U.S. banks generally became more efficient between 1985 and 2004. Our estimates suggest that much of this improvement was due to an inward shift of the α-quantile (and presumably also an inward shift of the production possibilities frontier), however, rather than significant productivity gain. Our results are thus consistent with the view of Berger (2003) that in the face of increased competition among banks and between banks and other financial intermediaries, advances in information processing and financial technologies have not increased bank productivity, but rather to a large extent have been "given away" to bank customers in the form of improved service quality and lower prices. Although banks have tended to become more efficient in the face of increased competition, with the exception of the largest institutions, banks have not generally experienced significant productivity improvement.

Our findings are also consistent with the view expressed by Berger (2003), Bernanke (2006) and others that technological advances have favored larger banks at the expense of small lenders. Traditionally, small, community banks have found their niche in lending to informationally-opaque borrowers, where the development of customer relationships is key to overcoming information gaps that inhibit profitable lending. Advances in technology have lowered information processing costs, brought improved credit-scoring methods, and enabled increasingly sophisticated management techniques, however, that have eroded some of the information benefits of relationship lending. Such advances appear to have enabled large banks and other intermediaries to capture market share from small, community banks. Our findings that larger banks have tended to experience larger gains in efficiency than smaller banks, and that only the largest banks have enjoyed significant productivity improvement, are consistent with these trends.

7 Conclusions

This paper describes an unconditional quantile estimator which we use to estimate the technical efficiency of U.S. commercial banks, and changes in efficiency and productivity between 1985 and 2004. Rapid advances in information and financial technology, and changes in regulation, are changing the environment in which banks operate. Although recently the banking industry has been profitable and failures have been rare, questions remain about the long-run viability of commercial banks. Advances in technology, some observers contend, have favored non-bank intermediaries and enabled them to capture market share from banks. Similarly, large banks have probably benefited more from advances in technology, and perhaps also from deregulation, than have small, community banks.

Most studies of bank efficiency or productivity change use data from the 1980s or early 1990s and, hence, provide no information about how these aspects of bank performance have changed more recently. Further, these studies all rely on empirical methods that impose questionable specification assumptions or that have other undesirable properties. By contrast, the present study examines efficiency and productivity change through 2004, and uses methods that avoid many of the problems inherent with past approaches.

Being fully non-parametric, the estimator described in this paper avoids the specification issues inherent with parametric estimators. The quantile estimator retains, however, the classical root-n convergence rate of parametric estimators, in contrast to the slow convergence rates of most non-parametric frontier estimators, such as the FDH and DEA estimators. Further, unlike FDH and DEA, which are acutely sensitive to outliers in the data, the quantile estimator is robust with respect to outliers.

Our research findings support the claim that recent changes in technology and regulation have tended to favor larger banks. Since 1985, banks with at least $1 billion of total assets have, on average, experienced larger gains in efficiency and productivity than have smaller banks. Indeed, though our results indicate a general improvement in efficiency, we find that only the largest banks have experienced significant productivity gains. Thus, an apparent inward shift in production possibilities, rather than significant productivity improvement, explains much of the gains in efficiency for all but those banks with at least $1 billion of assets. Our results are thus consistent with the view that deregulation and advances in information and financial technology have tended to favor the largest banks at the expense of small, community banks.

References

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

Alam, I. (2001), "A nonparametric approach for assessing productivity dynamics of large U. S. banks," *Journal of Money, Credit, and Banking,* 33, 121–139.

Aragon, Y., Daouia, A., and Thomas-Agnan, C. (2005), "Nonparametric frontier estimation: A conditional quantile-based approach," *Econometric Theory,* 21, 358–389.

Banks, J., Blundell, R., and Lewbel, A. (1997), "Quadratic engel curves and consumer demand," *Review of Economics and Statistics,* 79, 527–539.

Berger, A. N. (2003), "The economic effects of technological progress: Evidence from the banking industry," *Journal of Money, Credit, and Banking,* 35, 141–76.

Berger, A. N., and Humphrey, D. B. (1997), "Efficiency of financial institutions: International survey and directions for future research," *European Journal of Operational Research,* 98, 175–212.

Berger, A. N., and Mester, L. J. (2003), "Explaining the dramatic changes in performance of US banks: technological change, deregulation, and dynamic changes in competition," *Journal of Financial Intermediation,* 12, 57–95.

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, 2006.

Cazals, C., Florens, J. P., and Simar, L. (2002), "Nonparametric frontier estimation: A robust approach," *Journal of Econometrics,* 106, 1–25.

Chalfant, J. A., and Gallant, A. R. (1985), "Estimating substitution elasticities with the Fourier cost function," *Journal of Econometrics,* 28, 205–222.

Clark, J. A., and Siems, T. F. (2002), "X-efficiency in banking: Looking beyond the balance sheet," *Journal of Money, Credit and Banking,* 34, 987–1013.

Cooper, R. J., and McLaren, K. R. (1996), "A system of demand equations satisfying effectively global regularity conditions," *Review of Economics and Statistics,* 78, 359–364.

Daouia, A. (2003), "Nonparametric analysis of frontier production functions and efficiency measurement using nonstandard conditional quantiles," PhD dissertation, Groupe de Recherche en Economie Mathématique et Quantititative, Université des Sciences Sociales, Toulouse I, et Laboratoire de Statistique et Probabilités, Université Paul Sabatier, Toulouse III.

Daouia, A., and Simar, L. (2007), "Nonparametric efficiency analysis: A multivariate conditional quantile approach," *Journal of Econometrics,* 140, 375–400.

Daraio, C., and Simar, L. (2005), "Introducing environmental variables in nonparametric frontier models: A probabilistic approach," *Journal of Productivity Analysis,* 24, 93–121.

Deprins, D., Simar, L., and Tulkens, H. (1984), "Measuring labor inefficiency in post offices," In *The Performance of Public Enterprises: Concepts and Measurements,* ed. M. Marchand P. Pestieau and H. Tulkens, Amsterdam: North-Holland pp. 243–267.

Färe, R. (1988), *Fundamentals of Production Theory,* Berlin: Springer-Verlag.

Färe, R., and Grosskopf, S. (1996), *Intertemporal Production Frontiers: With Dynamic DEA,* Boston: Kluwer Academic Publishers.

Färe, R., Grosskopf, S., and Lovell, C. A. K. (1985), *The Measurement of Efficiency of Production,* Boston: Kluwer-Nijhoff Publishing.

Färe, R., Grosskopf, S., Lindgren, B., and Roos, P. (1992), "Productivity changes in Swedish pharmacies 1980–1989: A non-parametric Malmquist approach," *Journal of Productivity Analysis,* 3, 85–101.

—— (1994), "Productivity developments in Swedish hospitals: A Malmquist output approach," In *Data Envelopment Analysis: Theory, Methodology and Applications,* ed. Abraham Charnes, William W. Cooper, Arie Y. Lewin, and Lawrence M. Seiford, Dordrecht: Kluwer Academic Publishers.

Farrell, M. J. (1957), "The measurement of productive efficiency," *Journal of the Royal Statistical Society A,* 120, 253–281.

Gallant, A. R. (1981), "On the bias in flexible functional forms and an essentially unbiased form," *Journal of Econometrics,* 15, 211–245.

—— (1982), "Unbiased determination of production technologies," *Journal of Econometrics,* 20, 285–323.

Gijbels, I., Mammen, E., Park, B. U., and Simar, L. (1999), "On estimation of monotone and concave frontier functions," *Journal of the American Statistical Association,* 94, 220–228.

Guilkey, D. K., Lovell, C. A. K., and Sickles, R. C. (1983), "A comparison of the performance of three flexible functional forms," *International Economic Review,* 24, 591–616.

Hsu, P. L., and Robbins, H. (1947), "Complete convergence and the law of large numbers," *Proceedings of the National Academy of Sciences of the United States of America,* 33, 25–31.

Kneip, A., Simar, L., and Wilson, P. W. (2007), "Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models," *Econometric Theory.* Forthcoming.

Kumbhakar, S. C., Park, B. U., Simar, L., and Tsionas, E. G. (2006), "Nonparametric stochastic frontiers: A local likelihood approach," *Journal of Econometrics.* Forthcoming.

Kuntz, L., and Scholtes, S. (2000), "Measuring the robustness of empirical efficiency valuations," *Management Science,* 46, 807–823.

McAllister, P. H., and McManus, D. (1993), "Resolving the scale efficiency puzzle in banking," *Journal of Banking and Finance,* 17, 389–405.

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

Park, B., Simar, L., and Weiner, C. (2000), "FDH efficiency scores from a stochastic point of view," *Econometric Theory,* 16, 855–877.

Porembski, M., Breitenstein, K., and Alpar, P. (2005), "Visualizing efficiency and reference relations in data envelopment analysis with an application to the branches of a german bank," *Journal of Productivity Analysis,* 23(2), 203–221.

Sealey, C., and Lindley, J. (1977), "Inputs, outputs, and a theory of production and cost at depository financial institutions," *Journal of Finance,* 32, 1251–1266.

Shephard, R. W. (1970), *Theory of Cost and Production Functions,* Princeton: Princeton University Press.

Simar, L. (2003), "Detecting outliers in frontier models: A simple approach," *Journal of Productivity Analysis,* 20, 391–424.

Simar, L., and Wilson, P. W. (1998), "Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models," *Management Science,* 44, 49–61.

—— (2000a), "A general methodology for bootstrapping in non-parametric frontier models," *Journal of Applied Statistics,* 27, 779–802.

—— (2000b), "Statistical inference in nonparametric frontier models: The state of the art," *Journal of Productivity Analysis,* 13, 49–78.

Wheelock, D. C., and Wilson, P. W. (1999), "Technical progress, inefficiency, and productivity change in U. S. banking, 1984–1993," *Journal of Money, Credit, and Banking,* 31, 212–234.

—— (2001), "New evidence on returns to scale and product mix among US commercial banks," *Journal of Monetary Economics,* 47, 115–132.

—— (2007), "Non-parametric, unconditional quantile estimation for efficiency analysis with an application to Federal Reserve check processing operations," *Journal of Econometrics.* Forthcoming.

Wilson, P. W. (1993), "Detecting outliers in deterministic nonparametric frontier models with multiple outputs," *Journal of Business and Economic Statistics,* 11, 319–323.

—— (1995), "Detecting influential observations in data envelopment analysis," *Journal of Productivity Analysis,* 6, 27–45.

—— (2007), "FEAR: A software package for frontier efficiency analysis with R," *Socio-Economic Planning Sciences.* Forthcoming.

Wilson, P. W., and Carey, K. (2004), "Nonparametric analysis of returns to scale and product mix among US hospitals," *Journal of Applied Econometrics,* 19, 505–524.

Table 1: Summary Statistics for Inputs and Outputs

Variable	Mean	Median	Std. Dev.	Min	Max
1985 (11993 observations)					
purchased funds	116668	6230	2117428	28	138987254
core deposits	154791	41882	1036907	451	66596536
labor	117	26	1046	2	69286
physical capital	4323	909	40091	1	2504860
equity capital	17473	4544	154389	195	9548042
consumer loans	30201	6046	230908	16	13744537
business loans	100006	10082	1565607	3	90468741
real estate loans	43557	10110	374397	3	29309995
securities	112870	24235	1156479	1110	75603347
off-balance sheet items	2557	112	39772	0	2457114
1994 (9585 observations)					
purchased funds	101510	6287	1435418	60	85091523
core deposits	229579	50534	1526510	87	86193016
labor	130	28	968	2	60798
physical capital	5395	948	47303	1	2963253
equity capital	28050	5858	219618	167	12741688
consumer loans	39728	4365	310176	5	13361878
business loans	75780	8382	928236	15	55259794
real estate loans	96685	17930	731268	40	47816007
securities	142086	26927	1221160	747	69235206
off-balance sheet items	4416	139	58156	0	2568008
2004 (6075 observations)					
purchased funds	198812	16886	6117443	139	455604441
core deposits	270544	64640	3159656	2566	203900062
labor	136	33	2225	2	153158
physical capital	6012	1556	68568	0	3829974
equity capital	47537	9336	742035	355	46040144
consumer loans	37828	3860	912559	0	65069183
business loans	108916	12297	3347792	15	249880653
real estate loans	170730	38127	1598023	107	95453436
securities	193784	29860	4125693	715	264914267
off-balance sheet items	6892	222	231864	0	17313655

NOTE: Labor is measured in full-time equivalents; all other variables are measured in units of one thousand constant year-2000 dollars.

Table 2: Mean Estimates of Efficiency Change by Quartile, 1985–1994

	1994				
1985	Q1	Q2	Q3	Q4a	Q4b
Q1	1.0074 1480	1.0750*** 361	1.0485* 88	0.8228** 22	0.5293*** 2
Q2	0.9627*** 321	1.0038 1166	0.9942 460	0.9060*** 79	0.7421*** 2
Q3	0.7407 4	0.9812* 300	0.9885** 1232	0.9368*** 407	0.5645** 3
Q4a	— 0	0.9850 4	0.9735* 147	0.9274*** 1250	0.7792*** 104
Q4b	— 0	— 0	— 0	1.1015 2	0.8652*** 202

NOTE: Each cell contains two entries; the top entry gives the mean, and the bottom entry gives the number of observations. One, two, or three asterisks indicate significance at 90, 95, or 99-percent, respectively.

Table 3: Mean Estimates of Efficiency Change by Quartile, 1994–2004

	2004				
1994	Q1	Q2	Q3	Q4a	Q4b
Q1	0.9748***	1.0626***	0.9993	0.9669	0.3209***
	1079	263	72	31	2
Q2	0.9348***	0.9799***	0.9793**	0.9770	—
	233	725	318	84	0
Q3	0.9309	0.9769	0.9687***	0.9445***	0.7362***
	4	197	709	318	8
Q4a	—	—	0.9597*	0.9261***	0.6788***
	0	0	80	621	127
Q4b	—	—	—	—	0.3555***
	0	0	0	0	53

NOTE: Each cell contains two entries; the top entry gives the mean, and the bottom entry gives the number of observations. One, two, or three asterisks indicate significance at 90, 95, or 99-percent, respectively.

Table 4: Mean Estimates of Efficiency Change by Quartile, 1985–2004

	2004				
1985	Q1	Q2	Q3	Q4a	Q4b
Q1	0.9713**	1.1000***	1.0489	0.9866	0.4222***
	791	274	124	62	3
Q2	0.9133***	0.9930	0.9976	0.9017***	0.7395**
	297	497	295	129	5
Q3	0.9284**	0.9616***	0.9643***	0.9249***	0.4449***
	23	257	533	296	19
Q4a	—	1.0233	0.9602*	0.8898***	0.5725***
	0	6	80	397	106
Q4b	—	—	—	—	0.2459***
	0	0	0	0	35

NOTE: Each cell contains two entries; the top entry gives the mean, and the bottom entry gives the number of observations. One, two, or three asterisks indicate significance at 90, 95, or 99-percent, respectively.

Table 5: Mean Estimates of Productivity Change by Quartile, 1985–1994

	1994				
1985	Q1	Q2	Q3	Q4a	Q4b
Q1	0.9982	1.0352***	1.0039	0.8049**	0.8878***
	1476	361	88	22	2
Q2	0.9825	1.0084	0.9994	0.9459***	1.1588**
	320	1163	460	79	2
Q3	0.7623	0.9869	0.9982	0.9682***	0.8257
	4	299	1232	407	3
Q4a	—	1.0188	0.9727**	0.9534***	0.9273***
	0	4	147	1249	104
Q4b	—	—	—	1.1251	0.9292***
	0	0	0	2	202

NOTE: Each cell contains two entries; the top entry gives the mean, and the bottom entry gives the number of observations. One, two, or three asterisks indicate significance at 90, 95, or 99-percent, respectively.

Table 6: Mean Estimates of Productivity Change by Quartile, 1994–2004

	2004				
1994	Q1	Q2	Q3	Q4a	Q4b
Q1	1.1034***	1.1441***	1.0539	1.0133	0.6373
	1079	263	72	31	2
Q2	1.0679***	1.1044***	1.0609***	1.0672**	—
	233	725	318	84	0
Q3	1.0134	1.0897***	1.0863***	1.0394***	1.0375
	4	197	709	318	8
Q4a	—	—	1.072***	1.0331***	0.9906
	0	0	80	621	127
Q4b	—	—	—	—	0.9794
	0	0	0	0	53

NOTE: Each cell contains two entries; the top entry gives the mean, and the bottom entry gives the number of observations. One, two, or three asterisks indicate significance at 90, 95, or 99-percent, respectively.

Table 7: Mean Estimates of Productivity Change by Quartile, 1985–2004

	2004				
1985	Q1	Q2	Q3	Q4a	Q4b
Q1	1.0770***	1.1699***	1.0717**	1.0294	0.6371***
	786	274	124	62	3
Q2	1.0393**	1.1280***	1.1250***	1.0303	1.0592
	294	495	295	129	5
Q3	1.0352	1.0704***	1.0959***	1.0661***	0.8024*
	23	255	533	296	19
Q4a	—	1.1144	1.0675***	1.0160	0.9295***
	0	6	80	397	106
Q4b	—	—	—	—	0.8877**
	0	0	0	0	35

NOTE: Each cell contains two entries; the top entry gives the mean, and the bottom entry gives the number of observations. One, two, or three asterisks indicate significance at 90, 95, or 99-percent, respectively.

Table 8: Mean Estimates of Technology Change by Quartile, 1985–1994

	1994				
1985	Q1	Q2	Q3	Q4a	Q4b
Q1	0.9963***	1.0082***	0.9991	0.9782*	0.8488**
	1480	361	88	22	2
Q2	1.0065***	1.0115***	1.0047**	0.9919	1.0038
	321	1166	460	79	2
Q3	1.0129	1.0184***	1.0106***	0.9879***	0.9937
	4	300	1232	407	3
Q4a	—	1.065**	1.0229***	0.9937***	0.9609***
	0	4	147	1250	104
Q4b	—	—	—	1.1248***	0.8546***
	0	0	0	2	202

NOTE: Each cell contains two entries; the top entry gives the mean, and the bottom entry gives the number of observations. One, two, or three asterisks indicate significance at 90, 95, or 99-percent, respectively.

Table 9: Mean Estimates of Technology Change by Quartile, 1994–2004

	2004				
1994	Q1	Q2	Q3	Q4a	Q4b
Q1	1.1444***	1.0962***	1.0713***	1.0365***	1.2855**
	1079	263	72	31	2
Q2	1.1265***	1.1059***	1.0594***	1.0207***	—
	233	725	318	84	0
Q3	1.0989***	1.0961***	1.0842***	1.0327***	0.9877
	4	197	709	318	8
Q4a	—	—	1.0988***	1.0646***	1.1083***
	0	0	80	621	127
Q4b	—	—	—	—	1.6483***
	0	0	0	0	53

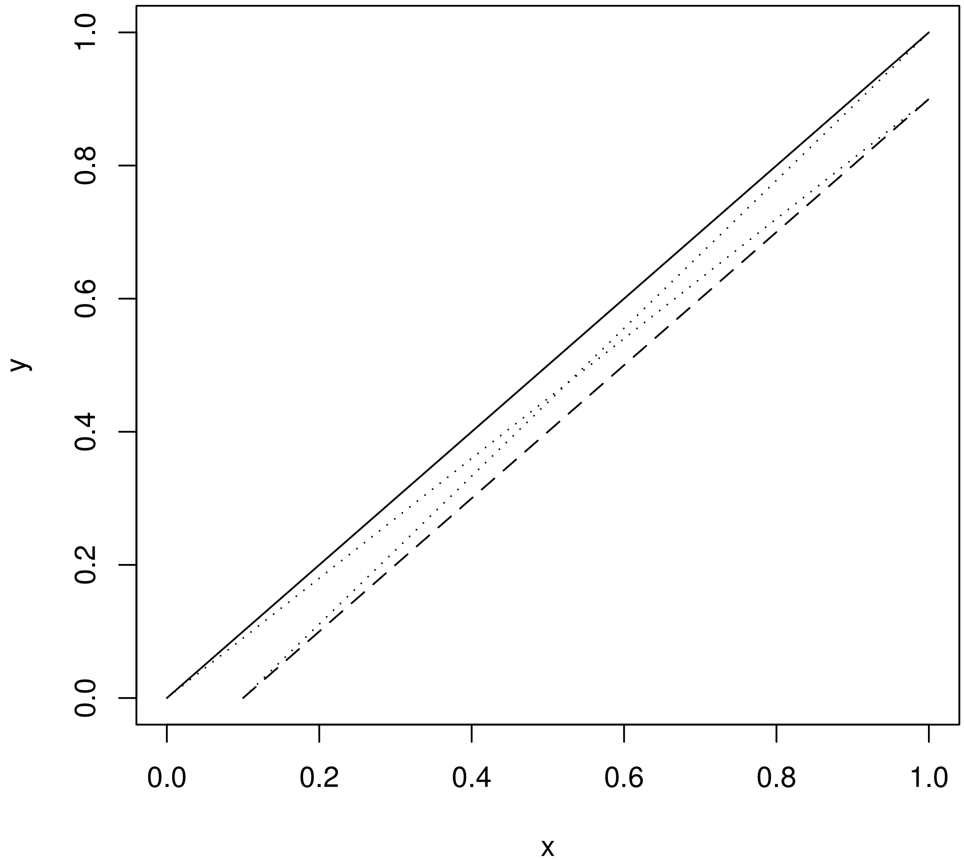
NOTE: Each cell contains two entries; the top entry gives the mean, and the bottom entry gives the number of observations. One, two, or three asterisks indicate significance at 90, 95, or 99-percent, respectively.

Table 10: Mean Estimates of Technology Change by Quartiles, 1985–2004

	2004				
1985	Q1	Q2	Q3	Q4a	Q4b
Q1	1.1308***	1.0918***	1.0375***	1.0027	0.9878
	791	274	124	62	3
Q2	1.1207***	1.1046***	1.0711***	1.0166*	0.9185***
	297	497	295	129	5
Q3	1.1224***	1.0937***	1.0746***	1.0113*	0.9518
	23	257	533	296	19
Q4a	—	1.1027***	1.0866***	1.0307***	1.0429**
	0	6	80	397	106
Q4b	—	—	—	—	1.4206***
	0	0	0	0	35

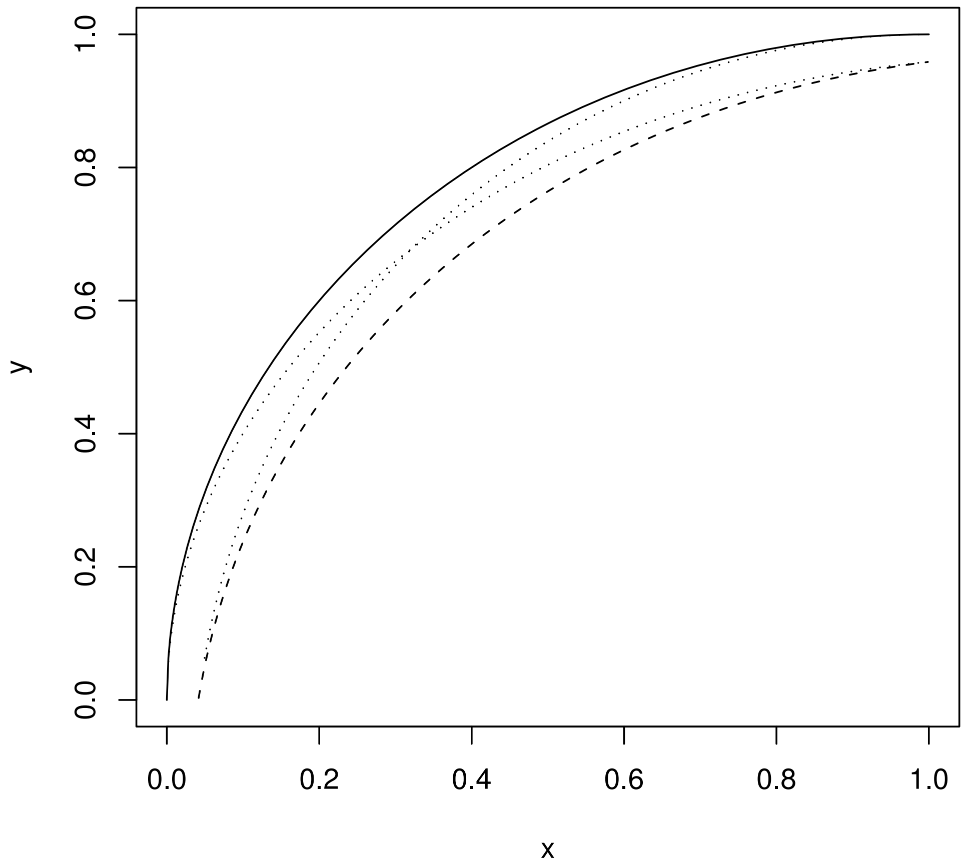
NOTE: Each cell contains two entries; the top entry gives the mean, and the bottom entry gives the number of observations. One, two, or three asterisks indicate significance at 90, 95, or 99-percent, respectively.

Figure 1: Conditional Input, Conditional Output, and Hyperbolic Quantiles ($\alpha = 0.99$; $f(x, y)$ uniform over a triangle)



NOTE: The solid curve shows the frontier $\mathcal{P}^{t\partial}$. The dashed curve illustrates the hyperbolic α quantile $\mathcal{P}^{t\partial}_{\alpha}$. The two dotted curves show conditional α-quantiles; the steeper of the two is the input conditional α-quantile, while the less-steeply sloped dotted curve corresponds to the output conditional α-quantile.

Figure 2: Conditional Input, Conditional Output, and Hyperbolic Quantiles ($\alpha = 0.99$; $f(x, y)$ uniform over a quarter-circle)



NOTE: The solid curve shows the frontier $\mathcal{P}^{t\partial}$. The dashed curve illustrates the hyperbolic α quantile $\mathcal{P}^{t\partial}_{\alpha}$. The two dotted curves show conditional α-quantiles; the steeper of the two is the input conditional α-quantile, while the less-steeply sloped dotted curve corresponds to the output conditional α-quantile.