2018-07-11

Essays on Distance Functions and Inefficiency Measurement

Esheba, Muna

UNIVERSITY OF CALGARY

Essays on Distance Functions and Inefficiency Measurement

by

Muna Esheba

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN ECONOMICS

CALGARY, ALBERTA

JULY, 2018

Abstract

This dissertation includes three essays on distance functions and inefficiency measurement. The main focus of the three essays is the measurement and determinants of technical inefficiency theoretically and empirically.

Essay 1 provides an up-to-date review that focuses on research methods, including different approaches to measuring technical inefficiency using distance functions, the development of modeling technical inefficiency, and the most common econometric estimation techniques. It also provides a useful guide on when these methods can be used and how to implement them. Regarding estimation issues, I address the important issues that should be managed in future applications while estimating technical inefficiency, including violation of theoretical and econometric regularity, the inaccurate choice of functional form, ignoring the possibility of heterogeneity and heteroskedasticity, and suffering from the endogeneity problem. I also discuss different approaches to deal with these issues, as well as potentially productive areas for future research.

Essay 2 derives the interactive effect between input and output technical inefficiencies theoretically using directional distance functions. This derivation solves the arbitrary decomposition of overall technical inefficiency into input and output components in previous studies. I argue that overall technical inefficiency equals the sum of input and output technical inefficiencies plus an interactive effect component which captures the interactions between them. I prove the results theoretically using exogenous and endogenous directional vectors.

Essay 3 investigates the relationships among input, output, and overall technical inefficiencies empirically using US banking data set. Using Bayesian estimation with the monotonicity conditions imposed at each observation, I estimate these inefficiencies separately using directional input, output, and technology distance functions. I model the overall technical inefficiency as a linear function of input and output technical inefficiencies, and a term

capturing the interactions between them. These determinants of overall technical inefficiency are estimated simultaneously with the variables that determine the frontier. I find significant evidence of the interactive effect between input and output technical inefficiencies which has a negative effect on the overall technical inefficiency. This result is robust to alternative directional vectors and model specifications, suggesting that the adjustability of both inputs and outputs is required for the improvement of efficiency.

Acknowledgments

First and foremost, I would like to thank God. I could never have written this dissertation without the faith I have in God, the Almighty.

With the support of many people, I was able to finish this dissertation. That being said, I would first like to express my sincere gratitude to my advisor Prof. Apostolos Serletis for his patience, motivation, immense knowledge, guidance, and the continuous support for overcoming numerous academic and general setbacks I have faced throughout the writing of this dissertation.

I would also like to thank the supervisory committee members; Prof. William David Walls, and Prof. Lasheng Yuan for their insightful comments, support, and encouragement.

The support of many faculty members at the Department of Economics was remarkable throughout the writing of this dissertation. My sincere thanks go to Prof. Francisco Gonzalez, Prof. Ana Ferrer, Prof. Robert Oxoby, Prof. Joanne Roberts, Prof. Jean-Francois Wen and Prof. Daniel Gordon.

I would like to thank Prof. Rolf Färe who read an earlier version of part of this dissertation and provided his feedback. His insightful comments and suggestions significantly improved this dissertation. My special thanks also go to Dr. Guohua Feng and Dr. Ali Jadidzadeh for their support and encouragement. My particular appreciation goes to Dr. Salwa Ahmed. Her generous support during this process was outstanding.

Financial support granted from the Libyan Ministry of Higher Education and Scientific Research, the Faculty of Graduate Studies, and the Department of Economics at the University of Calgary are also gratefully acknowledged.

Most importantly, I would like to thank my supportive family, my husband and my three wonderful children for supporting me spiritually throughout the writing of this dissertation and my life in general.

In memory of my father

who has contributed the most to make me the scholar I am today

Table of Contents

List of Tables

List of Figures and Illustrations

Overview

This dissertation includes three essays on distance functions and inefficiency measurement. The main focus of the three essays is the measurement and determinants of technical inefficiency theoretically and empirically. The importance of the measurement of technical inefficiency can be summarized by a notable quote by the physicist William Thomson; "If you cannot measure it, you cannot improve it". Measurement of technical inefficiency involves a comparison with the most efficient frontier. This comparison involves comparing observed input to minimum potential input required to produce the output or comparing observed output to maximum potential output obtainable from the input, or some combination of the two. The optimal is defined in terms of production frontiers, and inefficiency is technical.

Essays 1 and 2 deal with the measurement of technical inefficiency theoretically. More precisely, essay 1 is an up-to-date review of distance functions and the measurement of technical inefficiency that focuses on research methods. It also provides a useful guide on when these methods can be used and how to implement them. Essay 2 presents theoretical and illustrative methods to derive the interactive effect between input and output technical inefficiencies using directional distance functions. Essay 3 examines the relationships among input, output, and overall technical inefficiencies empirically.

In essay 1, I review and evaluate recent developments in several related areas, including different approaches to measuring technical inefficiency using distance functions, the development of modeling technical inefficiency in the stochastic frontier framework, and recent advances on the most common estimation techniques. In particular, I discuss and evaluate the radial measure of technical inefficiency given by the standard distance functions, the hyperbolic measure given by the hyperbolic distance function, and the directional measure given by the directional distance functions. I argue that the directional measure is preferable

to the radial and the hyperbolic measure. The development of modeling technical inefficiency regarding its temporal behavior, its classification, and its determinants are also discussed. With the aim of using the appropriate estimation techniques, I review recent advances on the most common estimation techniques, including maximum likelihood and Bayesian estimations. This essay also addresses the importance of attaining theoretical regularity applied by neoclassical microeconomic theory when violated, as well as econometric regularity when variables are non-stationary. Without regularity, technical inefficiency results are extremely misleading.

Regarding estimation issues, I address the important issues that should be managed in future applications while estimating technical inefficiency, including the inaccurate choice of functional form, ignoring the possibility of heterogeneity and heteroskedasticity, and suffering from the endogeneity problem. This essay also discusses different approaches to deal with these issues, as well as potentially productive areas for future research.

Most empirical studies that examine the technical inefficiency of production processes employ either an input or an output-oriented measurement technique. In terms of the former, researchers assume that outputs are exogenous and inputs endogenous and producers are fully capable of reallocating resources when improving efficiency. Similarly, by adopting an output-oriented measurement technique, it is assumed that inputs are exogenous and outputs endogenous and producers are fully capable of mixing production when improving efficiency. However, adopting an input (output) oriented measurement technique ignores the opposite output (input) orientation and this restriction may substantially bias the measures of producer inefficiency.

An efficiency survey by Berger, Hunter, and Timme (1993) suggests comparing these input and output approaches with a complete approach to investigate the relationships between input and output inefficiencies. However, few studies examine total technical inefficiency and decompose it into input and output components either by using a profit function or a direc-

tional technology distance function. Even though these studies disaggregate and quantify the impact of input and output on inefficiency, the arbitrary decomposition of total technical inefficiency into input and output inefficiency components results in concluding that total technical inefficiency equals the sum of input and output technical inefficiencies and shows no interactive effects between them.

Essay 2 differs from these studies by following Berger, Hunter, and Timme (1993) suggestion and comparing these input and output approaches with a complete approach using directional input, output, and technology distance functions. I derive the interactive effect between input and output technical inefficiencies theoretically using directional distance functions. This derivation solves the arbitrary decomposition of overall technical inefficiency into input and output components. I argue that overall technical inefficiency does not equal the sum of input and output technical inefficiencies as previous studies claim. It equals the sum of input and output technical inefficiencies plus an interactive effect component which captures the interactions between them. I prove the results theoretically using exogenous and endogenous directional vectors. I also use a numerical illustration to confirm my results.

The results indicate that ignoring the interactive effect between input and output technical inefficiencies results in a decomposition of overall technical inefficiency into input and output components that are significantly different from the ones that incorporate it. To the best of my knowledge, this essay is the first in the literature that derives the interactive effect between input and output technical inefficiencies theoretically.

Essay 3 builds on the theory presented in essay 2 and examines the relationships among input, output, and overall technical inefficiencies empirically. In doing so, I use annual data for US commercial banks over the period from 2001 to 2015, obtained from the Reports of Income and Condition (Call Reports). The market-average prices faced and determined exogenously rather than the actual prices paid or received by the bank are used, following Berger and Mester (2003). These market-average prices are more likely to be exogenous to

the bank than the bank-specific prices. The asset approach proposed by Sealey and Lindley (1977) is used to identify bank inputs and outputs.

I estimate input, output, and overall technical inefficiencies separately using the directional input distance function (DIDF), the directional output distance function (DODF) and the directional technology distance function (DTDF), respectively. I estimate these inefficiencies using Bayesian methods with the three commonly used directional vectors; the unit value, the observed input-output, and the optimal directional vectors. Given the observed estimation issues discussed in essay 1, the latter addresses the endogeneity of inputs and outputs by using systems of equations, consisting of DIDF (DODF, or DTDF) with the cost (revenue, or profit) minimizing (maximizing) first-order conditions, respectively. The latter also accounts for heterogeneity across banks by allowing the directional vectors to be endogenous and vary across banks. Regarding regularity violations, I find that the monotonicity conditions with respect to labor and all outputs are violated for all models at most observations. Therefore, all models are re-estimated with the monotonicity conditions imposed at each observation, by following the Bayesian procedure discussed in O'Donnell and Coelli (2005).

To investigate the relationships among input, output, and overall technical inefficiencies, I model the overall technical inefficiency as a linear function of input and output technical inefficiencies, and a term capturing the interactions between them, following Battese and Coelli (1995). These determinants of overall technical inefficiency are estimated simultaneously with the variables that determine the frontier.

Essay 3 contributes to the literature in many ways. First, to the best of my knowledge, it is the first in the literature that uses a complete approach to examine the relationships among input, output, and overall technical inefficiencies empirically using the same data set and the directional input, output, and technology distance functions with the three commonly used directional vectors; the unit value, the observed input-output, and the optimal directional

vectors. Second, the optimal directional vectors are allowed to be endogenous and vary across banks to account for heterogeneity across banks. Third, it pays explicit attention to the theoretical regularity conditions in order to produce inference that is consistent with neoclassical microeconomic theory.

In line with essay 2, the results show that overall technical inefficiency does not equal the sum of input and output technical inefficiencies, as previous studies claim. It equals the sum of input and output technical inefficiencies plus an interactive effect component which captures the interactions between them, where the increase in the output technical inefficiency reflects on a reduction on the input technical inefficiency and vice versa.

The results also show that both input and output technical inefficiencies have significant positive effects on the overall technical inefficiency. However, the interactive effect between input and output technical inefficiencies has a significant negative effect on the overall technical inefficiency. This result is robust to alternative directional vectors and model specifications. It is also consistent with the theoretical result obtained in essay 2. Banks with larger values of the interactive effect tend to have a lower level of overall technical inefficiency which indicates that they are more efficient.

The results also indicate that the value of the interactive effect between input and output technical inefficiencies depends on the choice of the directional vector in which the data are projected on the frontier and whether quantities and prices are taken into consideration. These results are quite significant, since these inefficiency components have different implications for bank performance, suggesting that the adjustability of both inputs and outputs is required for the improvement of bank efficiency.

Chapter 1

Distance Functions and the Measurement of Technical Inefficiency: A Review

1.1 Introduction

Measurement of technical inefficiency involves a comparison with the most efficient frontier. This comparison involves comparing observed input to minimum potential input required to produce the output or comparing observed output to maximum potential output obtainable from the input, or some combination of the two. The optimal is defined in terms of production frontiers, and inefficiency is technical.

Most of the literature reviews in this field focus on research outcomes and empirical applications. Battese (1992) provides a survey on production frontiers and technical efficiency that focuses on econometric models and empirical applications in agricultural economics. Greene (1993) provides a comprehensive review of the econometric approach to both technical and allocative inefficiency. More recently, Darku, Malla, and Tran (2013) provide a comprehensive and historical review that focuses on reviewing various agricultural efficiency studies and evaluating their methodologies and significant results. Parmeter and Kumbhakar (2014) review the econometric literature on the parametric and nonparametric estimation of technical efficiency.

This review paper contributes to the literature by providing an up-to-date review that focuses on research methods, including different approaches to measuring technical inefficiency using distance functions, the development of modeling technical inefficiency in the stochastic frontier framework, and the most common econometric estimation techniques. It also provides a useful guide on when these methods can be used and how to implement them.

In particular, I discuss and evaluate the radial measure of technical inefficiency given by the standard distance functions, the hyperbolic measure given by the hyperbolic distance function, and the directional measure given by the directional distance functions. Distance functions have the primary advantage of requiring only quantity information on inputs and outputs and serving as a direct measure of technical inefficiency. I argue that the directional measure is preferable to the radial and the hyperbolic measure. Measuring technical inefficiency with prices is also discussed. It can accommodate the joint estimation of both technical and allocative inefficiency, where allocative inefficiency is due to the failure of choosing the optimal input-output vector given relative input and output market prices. However, it depends on the availability of price information, and the satisfaction of the required behavioral assumptions; cost-minimizing (revenue or profit-maximizing) behavior.

The development of modeling technical inefficiency regarding its temporal behavior, its classification, and its determinants are also discussed. Regarding its temporal behavior, technical inefficiency in the stochastic frontier models is viewed first as time-invariant in cross-section and panel data models. This assumption is relaxed with the development of the time-variant technical inefficiency models. These models allow technical inefficiency to vary over time and across individual producers. Time-invariant and time-variant inefficiency models are developed to take both inefficiency components into account. More recently, four random components inefficiency models are proposed to account for both inefficiencies and heterogeneous technology since time-invariant and time-variant inefficiency models fail to explicitly account for unobserved heterogeneity or separate it from time-invariant inefficiency. The dynamic inefficiency models are proposed to capture the fact that the temporal behavior of inefficiency may be dynamic, where inefficiency evolves via an autoregressive process where past values of inefficiency determine the current value of inefficiency.

In contrast to the models that allow for the existence of extremely inefficient producers who cannot survive in highly competitive markets, the threshold inefficiency models truncate

the distribution of inefficiency by placing a threshold parameter of the minimum efficiency for survival on inefficiency. Thus, these models specify an upper bound to the distribution of inefficiency in addition to the zero lower bound.

While the threshold inefficiency models focus on the possibility of inefficient producers being out of the markets, recent studies develop the zero inefficiency models which focus on the possibility of producers being fully efficient. The zero inefficiency models can accommodate the presence of both fully efficient and inefficient producers in a probabilistic framework.

Heterogeneous inefficiencies models are proposed to capture heterogeneity in the inefficiency component by either including producer-specific characteristics in the inefficiency component or the mean, variance or both parameters of the inefficiency distribution.

With the aim of using the appropriate estimation techniques, I review the recent advances on the most common estimation techniques, including maximum likelihood and Bayesian estimations. This paper also addresses the importance of attaining the theoretical regularity applied by neoclassical microeconomic theory when violated, as well as the econometric regularity when variables are non-stationary. Without regularity, inefficiency results are extremely misleading. This paper also discusses techniques for imposing theoretical regularity and integration and cointegration techniques that can be used to manage the non-stationarity of the residuals.

Regarding estimation issues, I address the important issues that should be managed in future applications while estimating technical inefficiency, including the inaccurate choice of functional form, ignoring the possibility of heterogeneity and heteroskedasticity, and suffering from the endogeneity problem. This paper also discusses different approaches to deal with these issues, as well as potentially productive areas for future research.

The estimates of technical inefficiency can be distorted by the inaccurate choice of functional form for the production technology. This paper discusses several selection criteria

for choosing a particular functional form for the production technology based on theoretical properties such as its shape of the isoquants, its separability, its flexibility and its regular regions, and application properties such as homogeneity and translation properties. This paper also addresses empirical techniques that can be used to assess the ability of different functional forms to approximate the unknown underlying function.

The appropriate choice of functional form is not sufficient without accommodating heterogeneous technologies that may exist among producers or heterogeneity in the inefficiency term. Ignoring heterogeneity can lead to wrong conclusions concerning inefficiency measures since heterogeneity which is not captured by producer-specific characteristics is wrongly attributed to inefficiency. This paper addresses the importance of accommodating heterogeneity and discusses different approaches to account for both heterogeneous technologies and heterogeneity in the inefficiency term while estimating technical inefficiency.

Another potential issue when estimating technical inefficiency using distance functions is that inputs and outputs may be endogenous leading to biased and inconsistent estimates of the parameters of the production technology and the associated measures of inefficiency. This paper discusses different approaches to deal with this issue.

The rest of the paper is organized as follows. The next section presents theoretical backgrounds on the radial, the hyperbolic, and the directional measures of technical inefficiency using distance functions. Section 3 reviews the development of modeling technical inefficiency in the stochastic frontier framework. Section 4 gives a brief review of the recent advances on the most common estimation techniques. Section 5 discusses the estimation issues, and the last section summarizes and concludes the paper.

1.2 Distance Functions and Technical Inefficiency

There are several ways to measure technical inefficiency using distance functions. It can be measured radially using standard distance functions, hyperbolically using hyperbolic distance

function, or directionally using directional distance functions. It can also be measured by exploiting the duality between distance functions and cost, revenue and profit functions. The choice can be based on several selection criteria; the objective of the producers, exogeneity assumptions, data availability, and the complexity of the estimation procedures.

To briefly review some of the literature on the radial, hyperbolic, and directional measures of technical inefficiency using distance functions, consider a producer employing a vector of n inputs $x = (x_1, ..., x_n) \in \mathbb{R}^n_+$ available at fixed prices $w = (w_1, ..., w_n) \in \mathbb{R}^n_{++}$ to produce a vector of m outputs $y = (y_1, ..., y_m) \in \mathbb{R}^m_+$ that can be sold at fixed prices $p = (p_1, ..., p_m) \in \mathbb{R}^m_{++}$. Let $L(y)$ be the set of all input vectors x which can produce the output vector y

$$L(y) = \left\{x = (x_1, ..., x_n) \in R^N_+ : x \text{ can produce } y\right\}$$

and let $P(x)$ be the feasible set of outputs y that can be produced from the input vector x

$$P(x) = \left\{y = (y_1, ..., y_m) \in R^m_+ : y \text{ is producible from } x\right\}$$

The production technology T for a producer is defined as the set of all feasible input-output vectors

$$T = \left\{(x,\, y) : x \in \mathbb{R}^n_+,\, y \in \mathbb{R}^m_+,\, x \text{ can produce } y\right\}$$

Note that $(x, y) \in T \Leftrightarrow x \in L(y) \Leftrightarrow y \in P(x)$.

1.2.1 The Radial Measure of Technical Inefficiency

The radial measure of technical inefficiency is given by the standard distance functions. Distance functions are initially defined on the input or output production possibility sets by Debreu (1951) and Shephard (1953, 1970). While the input distance function considers the proportional contraction of inputs holding the outputs constant, the output distance function considers the proportional expansion of outputs holding the inputs constant. They are independent of the unit of measurement.

Input (output) distance functions ignore the opposite output (input) orientation, and this restriction may be unacceptable if the adjustability of both inputs and outputs is required. Thus, the choice of input versus output distance function as an alternative representation of production technology should be based on the objective of the producers and the exogeneity assumption of inputs and outputs. If the objective of the producers is to minimize cost which involves choosing the optimal quantities of inputs to produce a given output vector, the input distance function can be adopted to estimate technical inefficiency in the cost minimization problem. If the objective of the producers is to maximize revenues which involve producing the optimal quantities of outputs from a given input vector, the output distance function can be adopted to estimate technical inefficiency in the revenue maximization problem.

The Input Distance Function

Following Shephard (1953), the input distance function (IDF) can be defined relative to the input set $L(y)$ or the production technology T as follows

$$D_I(y,x) = \max_{\vartheta_I} \left\{ \vartheta_I : \frac{x}{\vartheta_I} \in L(y) \right\} = \max_{\vartheta_I} \left\{ \vartheta_I : \left(\frac{x}{\vartheta_I}, y \right) \in T \right\}$$

where $1/\vartheta_I$ represents the proportional contraction of inputs that is required to reach the inner boundary of the input set or the production frontier, holding the outputs constant. $D_I(y,x)$ is given by the ratio of the observed input to the minimum input required to produce the given output. Thus, for any x, $x/D_I(y,x)$ is the minimum input vector on the ray from the origin through x that can produce y, as can be seen in Figure 1.1. Efficient producers, who produce on the boundary of the input set or the production frontier, have $D_I(y,x) = 1$. Inefficiency is indicated by $D_I(y,x) > 1$.

The Debreu-Farrell input-oriented measure of technical efficiency is defined as

$$TE_I(y,x) = \min_{\vartheta_{FI}} \{ \vartheta_{FI} : \vartheta_{FI} x \in L(y) \} = \min_{\vartheta_{FI}} \{ \vartheta_{FI} : (\vartheta_{FI} x, y) \in T \}$$

Note that the Debreu-Farrell input-oriented measure of technical efficiency is the reciprocal of the IDF, $TE_I(y,x) = [D_I(y,x)]^{-1}$. $TE_I(y,x) \leq 1$ represents a radial reduction of inputs

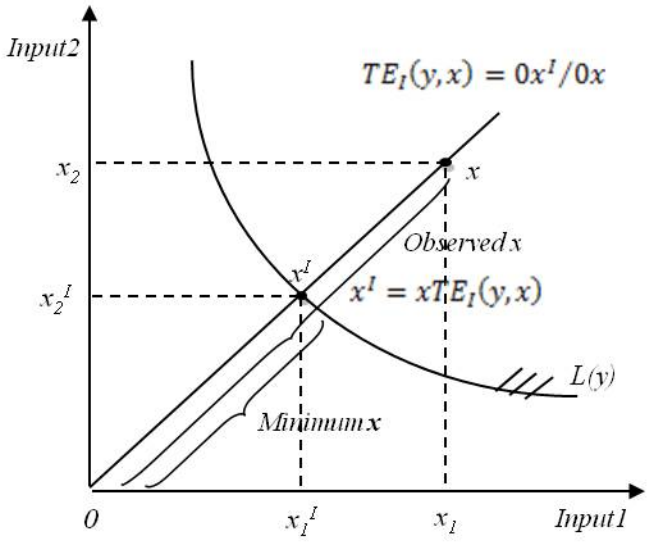

Figure 1.1: The Debreu-Farrell Input-Oriented Measure of Technical Efficiency

that is required to be considered as being efficient. Technical inefficiency is defined as

$$TI_I(y,x) = 1 - TE_I(y,x) = 1 - \frac{1}{D_I(y,x)}$$

where $0 \leq TI_I(y,x) \leq 1$. The IDF has the following properties [see Färe and Primont (1995), and Färe and Grosskopf (2004) for more details]

i) representation, $D_I(y,x) \geq 1$ iff $x \in L(y)$ or $(x,y) \in T$

ii) non-increasing and quasi-concave in outputs, and

iii) non-decreasing, concave, and linearly homogeneous in inputs,
$D_I(y,\lambda x) = \lambda D_I(y,x)$, $\lambda > 0$.

The IDF is also used to accommodate undesirable outputs. This is modeled by holding desirable outputs y constant and treating undesirable outputs b as inputs x; $D_I(y,x,b) = \max_{\vartheta_I}\left\{\vartheta_I : (\frac{x}{\vartheta_I}, \frac{b}{\vartheta_I}) \in L(y)\right\} = \max_{\vartheta_I}\left\{\vartheta_I : \left(\frac{x}{\vartheta_I}, \frac{b}{\vartheta_I}, y\right) \in T\right\}$. See, for example, Atkinson and Dorfman (2005). This credits the producer for reducing both inputs and undesirable outputs proportionally to reach the production frontier. However, if inputs are freely disposable so are undesirable outputs. Such a treatment of undesirable outputs is criticized due

to the implied strong disposability of undesirable outputs by Färe *et al.* (2005). Similarly, treating undesirable outputs as inputs allows substitutability or complementarity among them. Furthermore, these studies ignore the fact that the production of undesirable outputs is affected by the production of desirable outputs $b = f(y)$ not the opposite $y = f(b)$. Thus, treating undesirable outputs as inputs is inappropriate because it imposes incorrect theoretical restrictions on the production technology. Therefore, IDF can be used only to decrease inputs, holding both outputs constant. Assaf *et al.* (2013) use IDF and treat undesirable output as a technology shifter.

The Output Distance Function

Instead of looking at the proportional contraction of inputs holding the outputs constant, the output distance function (ODF) considers the proportional expansion of outputs holding the inputs constant. Following Shephard (1970), it is defined on the output set $P(x)$ or the production technology T as

$$D_O(x,y) = \min_{\vartheta_O} \left\{ \vartheta_O : \frac{y}{\vartheta_O} \in P(x) \right\} = \min_{\vartheta_O} \left\{ \vartheta_O : \left(x, \frac{y}{\vartheta_O} \right) \in T \right\}$$

where $1/\vartheta_O$ represents the proportional expansion of outputs that is required to reach the upper boundary of the output set or the production frontier, holding the inputs fixed. $D_O(x,y)$ is given by the ratio of the observed output to maximum potential output obtainable from the given input. Thus, for any y, $y/D_O(x,y)$ is the largest output vector on the ray from the origin through y that can be produced by x, as can be seen in Figure 1.2. If y is on the boundary of the output set or the production frontier, $D_O(x,y) = 1$, implying that the producer is operating at full technical efficiency. If y is within the boundary of the output set or the production frontier, $D_O(x,y) < 1$, indicating that the producer is operating with technical inefficiency.

The Debreu-Farrell output-oriented measure of technical efficiency is defined as

$$TE_O(x,y) = \max_{\vartheta_{FO}} \{ \vartheta_{FO} : \vartheta_{FO} y \in P(x) \} = \max_{\vartheta_{FO}} \{ \vartheta_{FO} : (x, \vartheta_{FO} y) \in T \}$$

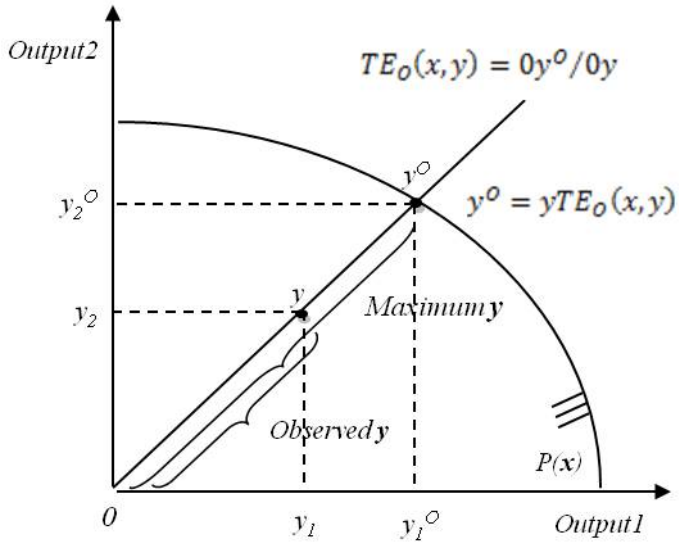

Figure 1.2: The Debreu-Farrell Output-Oriented Measure of Technical Efficiency

Note that the Debreu-Farrell output-oriented measure of technical efficiency is the reciprocal of the ODF, $TE_O(x,y) = [D_O(x,y)]^{-1}$. $TE_O(x,y) \geq 1$ represents a radial expansion of outputs that is required to achieve efficiency and the higher this measure, the lower the efficiency. Technical inefficiency is defined as

$$TI_O(x,y) = TE_O(x,y) - 1 = \frac{1}{D_O(x,y)} - 1$$

where $TI_O(x,y) \geq 0$. The ODF has the following properties [see Färe and Grosskopf (1994) for more details]

i) representation, $D_O(x,y) \leq 1$ iff $y \in P(x)$ or $(x,y) \in T$

ii) non-increasing and quasi-convex in inputs, and

iii) non-decreasing, convex and linearly homogeneous in outputs,
$D_O(x,\lambda y) = \lambda D_O(x,y)$, $\lambda > 0$.

The ODF is also used to accommodate undesirable outputs. This is modeled by holding inputs x constant and treating undesirable outputs b as desirable outputs y; $D_O(x,y) = \min_{\vartheta_O}\left\{\vartheta_O : \left(\frac{y}{\vartheta_O}, \frac{b}{\vartheta_O}\right) \in P(x)\right\} = \min_{\vartheta_O}\left\{\vartheta_O : \left(x, \frac{y}{\vartheta_O}, \frac{b}{\vartheta_O}\right) \in T\right\}$. See, for example, Färe *et al.*

(1993). Färe *et al.* (1989) show how to adjust inefficiency measures in the presence of undesirable outputs. This credits the producer for expanding both desirable and undesirable outputs proportionally to reach the production frontier. However, this can only be applied if the adjustability of both desirable and undesirable outputs is required. Producers have no control on reducing undesirable outputs without reducing desirable outputs and producing more desirable outputs require producing more undesirable outputs such as generating polluting as by-products of producing desirable outputs. That is, the undesirable outputs are inevitably produced unless the entire production process is terminated.

The standard input and output distance functions adjust the desirable and undesirable outputs proportionally at the same rate, which may not be aimed by producers who attempt to reduce undesirable outputs and increase desirable outputs simultaneously. Furthermore, these standard distance functions treat technical inefficiency as environmental inefficiency. Future studies comparing these inefficiencies and separating them would help to further understanding of how these inefficiencies differ.

1.2.2 The Hyperbolic Measure of Technical Inefficiency

The hyperbolic measure of technical inefficiency is given by the hyperbolic distance function. Following Färe *et al.* (1985), the hyperbolic distance function (HDF) can be defined relative to the production technology T as follows

$$D_H(x,y) = \min_{\vartheta_H} \left\{ \vartheta_H : \left(\vartheta_H x, \frac{y}{\vartheta_H} \right) \in T \right\}$$

where $1 \geq \vartheta_H > 0$ represents the proportional contraction of inputs and expansion of outputs that is required to reach the production frontier. Note that reducing ϑ_H implies expanding $1/\vartheta_H$. This is illustrated in figure 1.3 where the hyperbolic curve intersects with the production frontier at point $H = \left(\vartheta_H x, \frac{y}{\vartheta_H} \right)$. Efficient producers who produce on the boundary of the production frontier, have $D_H(x,y) = 1$. Inefficiency is indicated by $D_H(x,y) < 1$. The hyperbolic measure of technical efficiency proposed by Färe *et al.* (1985)

is defined as

$$TE_H(x,y) = \max_{\vartheta_{FH}} \left\{ \vartheta_{FH} : \left(\frac{x}{\vartheta_{FH}}, \vartheta_{FH} y \right) \in T \right\}$$

Note that the hyperbolic measure of technical efficiency is the reciprocal of the HDF, $TE_H(x,y) = [D_H(x,y)]^{-1}$. For technology frontiers with variable returns to scale, Nahm and Vu (2013) show that the hyperbolic measure of technical efficiency is the square of the HDF, $TE_H(x,y) = [D_H(x,y)]^2$. $TE_H(x,y) \geq 1$ under the assumption of weak disposability of inputs and outputs. Technical inefficiency is defined as

$$TI_H(x,y) = TE_H(x,y) - 1 = \frac{1}{D_H(x,y)} - 1$$

where $TI_H(x,y) \geq 0$. Färe *et al.* (2002) show that under constant returns to scale, the HDF is related to the standard input and output distance functions as $D_H(x,y) = [D_I(y,x)]^{-1/2} = [D_O(x,y)]^{1/2}$. Another type of relationship is developed by Simar and Vanhems (2012) and Daraio and Simar (2014) between the HDF and the directional technology distance function as $\ln D_H(x^*,y^*) = \vec{D}_T(x,y;g_x,g_y)$ where $x^* = \exp(x./g_x)$ and $y^* = \exp(y./g_y)$. The HDF has the following properties [see Färe *et al.* (1985, 1994) for more details]

i) representation, $D_H(x,y) \leq 1$ iff $(x,y) \in T$

ii) non-increasing in inputs and non-decreasing in outputs

iii) homogeneity, $D_H(\lambda^{-1}x, \lambda y) = \lambda D_H(x,y)$, $\lambda > 0$

iv) homogeneous of degree zero in inputs and outputs under constant returns to scale, and

v) almost homogeneous of degrees k_1, k_2 and k_3 if $D_H\left(\lambda^{k_1}x, \lambda^{k_2}y\right) = \lambda^{k_3} D_H(x,y)$.

In contrast to the standard distance functions, the HDF simultaneously contracts inputs and expands outputs proportionally without restricting either inputs or outputs to be constant. Moreover, it can be used for simultaneous contraction of inputs x and undesirable out-

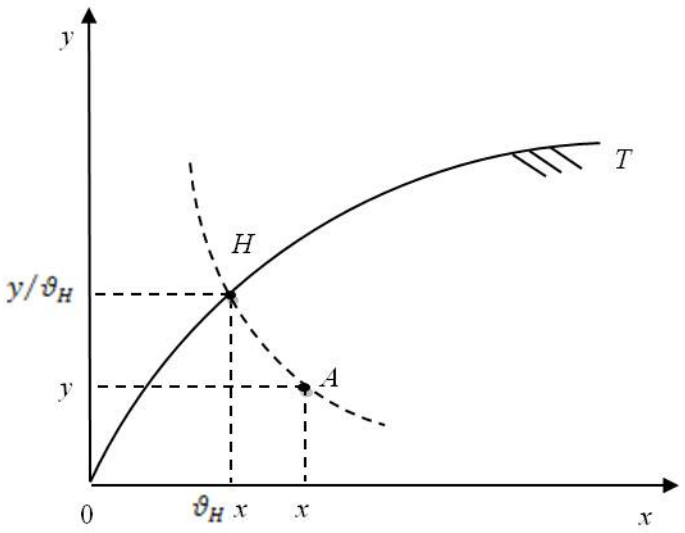

Figure 1.3: The Hyperbolic Measures of Technical Efficiency

puts b, and expansion of desirable outputs y; $D_H(x, y, b) = \min_{\vartheta_H} \left\{ \vartheta_H : \left(\vartheta_H x, \frac{y}{\vartheta_H}, \vartheta_H b \right) \in T \right\}$. See, for example, Cuesta *et al.* (2009) and Fang and Yang (2014).

The choice of the hyperbolic distance function can be adopted to estimate technical inefficiency assuming that producers can adjust both inputs and outputs when improving efficiency. Färe *et al.*(2002) show that the HDF is dual to the return-to-dollar function first proposed by Georgescu-Roegen (1951); $[D_H(x, y)]^{-2} \geq py/wx$. However, the HDF is occasionally used in the literature to measure technical inefficiency because of the non-linear optimization involved – see, for example, Cuesta and Zofio (2005) who introduce a method to estimate technical inefficiency for Spanish savings banks using the HDF.

1.2.3 The Directional Measure of Technical Inefficiency

Unlike the standard or hyperbolic distance functions, the directional distance functions (DDF) constitute an additive not proportional or multiplicative measure of technical inefficiency in a given direction g, therefore they are not scale-invariant. The additive nature of DDF allows for the treatment of non-positive inputs or outputs.

The Directional Technology Distance Function

The directional technology distance function (DTDF) generalizes the standard input and output distance functions, providing a tool to address efficiency issues in an integrated approach. It is introduced by Chambers *et al.* (1998) as a variant of the Luenberger (1995) shortage function. It allows for simultaneous contraction of inputs and expansion of outputs in terms of an explicit direction vector $g = (g_x, g_y)$, where $g_x \in R_+^N$ and $g_y \in R_+^M$ such that it contracts inputs in the direction g_x and expands outputs in the direction g_y. In particular, the DTDF is defined as

$$\vec{D}_T(x, y; g_x, g_y) = \max_{\theta_T} \{\theta_T : (x - \theta_T g_x, y + \theta_T g_y) \in T\} \tag{1.1}$$

Efficient producers who produce on the frontier of T have $\vec{D}_T(x, y; g_x, g_y) = 0$, implying that there is no further contraction of inputs and expansion of outputs that is feasible. Inefficiency is indicated by $\vec{D}_T(x, y; g_x, g_y) > 0$, with higher values indicating greater inefficiency when producers operate beneath the frontier of T. Eliminating technical inefficiency for producers who operate at point A would take the producers to point $B = \left(x^T, y^T\right) = (x - \theta_T g_x, y + \theta_T g_y)$ on the frontier of T, as can be seen in Figure 1.4. The DTDF serves as a technology-oriented measure of technical inefficiency

$$TI_T = \vec{D}_T(x, y; g_x, g_y)$$

As noted by Chambers *et al.* (1998), the DTDF has the following properties:

i) representation, $\vec{D}_T(x, y; g_x, g_y) \geq 0$ iff $(x, y) \in T$

ii) translation, $\vec{D}_T(x - \alpha g_x, y + \alpha g_y; g_x, g_y) = \vec{D}_T(x, y; g_x, g_y) - \alpha$, for $\alpha \in R$

iii) non-decreasing in x and non-increasing in y if inputs and outputs are freely disposable

iv) concave in (x, y)

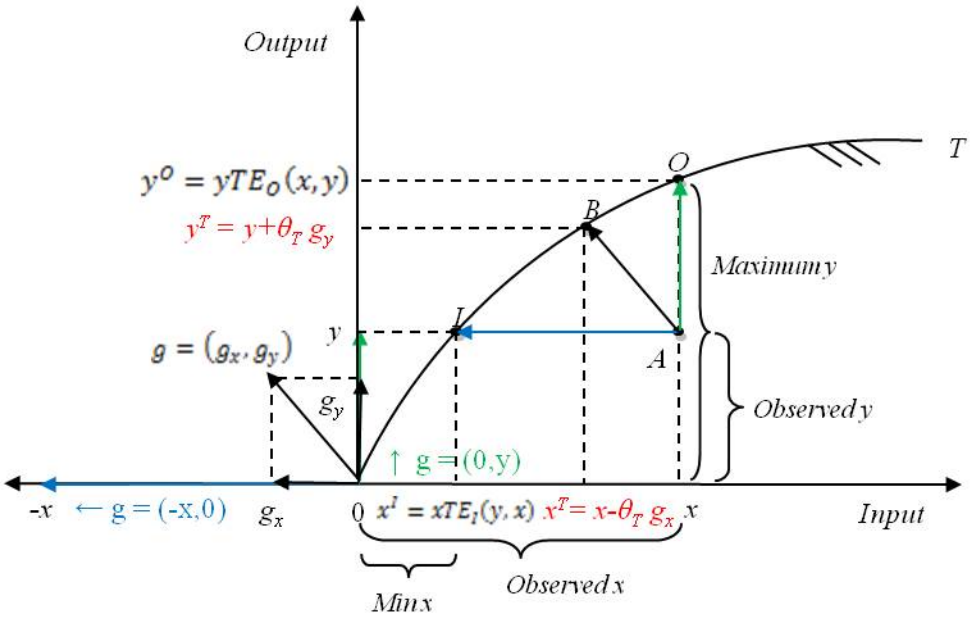

Figure 1.4: The Directional Measure of Technical Inefficiency

v) homogeneous of degree -1 in g, That is,
$\vec{D}_T(x, y; \lambda g_x, \lambda g_y) = \lambda^{-1}\vec{D}_T(x, y; g_x, g_y)$, for $\lambda > 0$, and

vi) homogeneous of degree $+1$ in x and y if the technology exhibits constant returns to scale, $\vec{D}_T(\lambda x, \lambda y; g_x, g_y) = \lambda\vec{D}_T(x, y; g_x, g_y)$, for $\lambda > 0$.

The advantage of DTDF over the IDF, ODF, and HDF is that it can accommodate desirable and undesirable outputs by allowing for non-radial or hyperbolic expansion of the desirable outputs and contraction of the undesirable outputs.

The Directional Input Distance Function

The inefficiency measures derived from the directional distance function depend on the choice of the directional vector, $g = (g_x, g_y)$. By setting $g_y = 0$, the directional vector becomes $g = (g_x, 0)$ and allows only for input contraction holding outputs fixed — see Figure 1.4. In this case, equation (1.1) becomes the directional input distance function (DIDF) that allows for only input contraction, $\vec{D}_T(x, y; g_x, 0) = \vec{D}_I(y, x; g_x)$

$$\vec{D}_I(y, x; g_x) = \max_{\theta_I}\{\theta_I : (x - \theta_I g_x) \in L(y)\} = \max_{\theta_I}\{\theta_I : (x - \theta_I g_x,\ y) \in T\}$$

Moreover, according to Chambers *et al.* (1996, 1998) and Fare and Grosskopf (2000), if the directional input vector, g_x, equals the observed input vector, x, (that is, $g_x = -x$), then

$$\vec{D}_I(y, x; g_x) = \vec{D}_I(y, x; -x) = 1 - \frac{1}{D_I(y, x)}$$

and in this case there is a relationship between the directional input distance function, $\vec{D}_I(y, x; -x)$, and the standard input distance function, $D_I(y, x)$. As can be seen in Figure 1.4, producers who operate at point A can hold output constant and contract input in the direction $g_x = -x$ to point I. The DIDF serves as an input-oriented measure of technical inefficiency

$$TI_I = \vec{D}_I(y, x; g_x)$$

The DIDF satisfies the following properties [see Chambers *et al.* (1996)]:

i) representation, $\vec{D}_I(y, x; g_x) \geq 0$ iff $x \in L(y)$ or $(x, y) \in T$

ii) translation, $\vec{D}_I(y, x - \alpha g_x; g_x) = \vec{D}_I(y, x; g_x) - \alpha$, for $\alpha \in R$

iii) concavity in inputs

iv) positive monotonicity in inputs. That is, $x' > x$ implies
$\vec{D}_I(y, x'; g_x) \geq \vec{D}_I(y, x; g_x)$

v) negative monotonicity in outputs. That is, $y' > y$ implies
$\vec{D}_I(y', x; g_x) \leq \vec{D}_I(g, x; g_x)$, and

vi) homogeneity of degree -1 in g_x. That is,
$\vec{D}_I(y, x; \lambda g_x) = \lambda^{-1}\vec{D}_I(y, x; g_x)$, for $\lambda > 0$.

The Directional Output Distance Function

By setting $g_x = 0$, the directional vector becomes $g = (0, g_y)$ and allows only for output expansion holding inputs fixed — see Figure 1.4. In this case, equation (1.1) reduces to

the directional output distance function (DODF) that allows for only output expansion, $\vec{D}_T(x, y; 0, g_y) = \vec{D}_O(x, y; g_y)$

$$\vec{D}_O(x, y; g_y) = \max_{\theta_o} \{\theta_O : (y + \theta_O g_y) \in P(x)\} = \max_{\theta_o} \{\theta_O : (x, y + \theta_O g_y) \in T\}$$

Moreover, as noted by Chambers *et al.* (1998) and Färe and Grosskopf (2000), if the directional output vector, g_y, equals the observed output vector, y (that is, $g_y = y$), then

$$\vec{D}_O(x, y; g_y) = \vec{D}_O(x, y; y) = \frac{1}{D_O(x, y)} - 1$$

and in this case there is a relationship between the directional output distance function, $\vec{D}_O(x, y; y)$, and the standard output distance function, $D_O(x, y)$. As can be seen in Figure 1.4, producers who operate at point A can hold input constant and expand output in the direction $g_y = y$ to point O. The DODF serves as an output-oriented measure of technical inefficiency

$$TI_O = \vec{D}_O(x, y; g_y)$$

The DODF satisfies the following properties [see Färe *et al.* (2005)]:

i) representation, $\vec{D}_O(x, y; g_y) \geq 0$ iff $y \in P(x)$ or $(x, y) \in T$

ii) translation, $\vec{D}_O(x, y + \alpha g_y; g_y) = \vec{D}_O(x, y; g_y) - \alpha$, for $\alpha \in R$

iii) concavity in outputs

iv) positive monotonicity in inputs. That is, $x' > x$ implies
$\vec{D}_O(x', y; g_y) \geq \vec{D}_O(x, y; g_y)$

v) negative monotonicity in outputs. That is, $y' > y$ implies
$\vec{D}_O(x, y'; g_y) \leq \vec{D}_O(x, y; g_y)$, and

vi) homogeneity of degree -1 in g_y. That is,
$\vec{D}_O(x, y; \lambda g_y) = \lambda^{-1} \vec{D}_O(x, y; g_y)$, for $\lambda > 0$.

The Directional Vector

The measures of technical inefficiency derived from the directional distance functions depend on the choice of the directional vector g in which the data are projected on the frontier of T. Technical inefficiency can be measured by choosing an exogenous or an endogenous directional vector. The former is a pre-specified directional vector, and the latter determines the direction through a specific endogenous behavior.

Exogenous Directional Vector

For an exogenous or a pre-specified directional vector, two widely used directions are the unit value direction $g = (-1, 1)$ and the observed input-output direction $g = (-x, y)$. The unit value direction $g = (-1, 1)$ implies that the amount by which a producer could decrease inputs and increase outputs will be $\vec{D}_T(x, y; -1, 1) \times 1$ units of x and y — see, for example, Färe *et al.* (2005). The advantage of choosing this directional vector is relied on its simplicity, its allowance for aggregation to the industry level, normalizing nature, and convenience in explaining the results of measurement. Specifically, an inefficiency measure based on the unit value directional vector indicates, regardless of the units of measurement, the number of units of each input (output) that should be contracted (expanded) to reach the production frontier. As noted by Färe and Grosskopf (2004), the inefficiency of the industry equals the sum of the directional distance functions for all producers when choosing a common directional vector for all producers.

Another widely used pre-specified direction is the observed input-output direction $g = (-x, y)$. This type of directional vector measures the simultaneous maximum proportional expansion of outputs and contraction of inputs that is feasible given the technology. It assumes that an inefficient producer can decrease inefficiency while decreasing inputs and increasing outputs in proportion to the initial combination of the actual inputs and outputs — see, for example, Färe, Grosskopf, and Weber (2004).

The pre-specified directional vector is extended in several directions. Koutsomanoli-

Filippaki *et al.* (2012) use the observed input-output averages direction $g = (\overline{x}, \overline{y})$. However, these producer-specific directional vectors cannot be aggregated to the industry level. Tzeremes (2015) uses a range directional vector $g = (g_x, g_y) = (R, 0)$ where the range of possible input reduction of a specific producer is defined as the input minus the minimum inputs observed; $R_{ik'} = x_{ik'} - \min_k \{x_{ik}\}$ given a set of producers $k = \{1, \dots K\}$. Färe *et al.* (2013) and Hampf and Kruger (2015) use a directional vector based on exogenous normalization constraints.

The main issue with the pre-specified directional vector is that the parameter estimates of the production technology and associated measures of technical inefficiency depend on the choice of the directional vector — see, for example, Atkinson and Tsionas (2016).

Endogenous Directional Vector

Alternatively, an endogenous directional vector such that it projects any inefficient producer to the cost (revenue or profit) minimizing (maximizing) benchmark can be chosen. See, for example, Malikov *et al.* (2016) for an endogenous direction vector projecting to the cost-minimizing benchmark, Feng *et al.* (2018) for an endogenous direction vector projecting to the profit-maximizing benchmark, and Atkinson and Tsionas (2016) for a set of directions that is consistent with cost minimization and profit maximization. However, further research is needed comparing the different choices of the directional vector and providing a framework to determine an optimal set of directions.

1.2.4 Technical Inefficiency Measures with Prices

Measuring technical inefficiency with prices can accommodate the joint estimation of both technical and allocative inefficiencies, where allocative inefficiency is due to the failure of choosing the optimal input-output vector given relative input and output market prices. However, it depends on the availability of price information, and the satisfaction of the required behavioral assumptions; cost-minimizing (revenue or profit-maximizing) behavior.

The Cost and Input Technical Inefficiency

The duality between cost and distance functions is introduced by Luenberger (1992), Färe and Primont (1995), and Chambers *et al.* (1996). They show that under weak input disposability, the cost function can be derived from the IDF or the DIDF by minimizing with respect to inputs and using either unconstrained or conditional optimization.

$$C(y,w) = \min_x \{wx / D_I(y,x)\} = \min_x \{wx : D_I(y,x) \geq 0\}$$

$$C(y,w) = \min_x \left\{wx - \vec{D}_I(y,x;g_x) \times wg_x\right\} = \min_x \left\{wx : \vec{D}_I(y,x;g_x) \geq 0\right\}$$

Following Luenberger (1992), Chambers *et al.* (1996) show that under weak input disposability, DIDF can be derived from the cost function by minimizing with respect to input prices and using either unconstrained or conditional optimization.

$$\vec{D}_I(y,x;g_x) = \min_w \{wx - C(y,w)/wg_x\} = \min_w \{wx - C(y,w) : wg_x = 1\}$$

The relationships between the IDF, the DIDF and the cost function can be represented as [see Chambers *et al.* (1998), and Färe and Grosskopf (2000)]

$$\frac{1}{D_I(y,x)} \geq \frac{C(y,w)}{wx} \text{ and } \vec{D}_I(y,x;g_x) \leq \frac{wx - C(y,w)}{wg_x}$$

The inequality can be turned into equality by including a residual multiplicative term to the IDF and an additive term to the DIDF to capture allocative inefficiency, where allocative inefficiency is due to the failure of choosing the cost-minimizing input vector given relative input market prices.

Ignoring the price information, technical inefficiency can be measured by the amounts by which a producer lies above its input isoquant. The Debreu-Farrell input-oriented measure of technical efficiency represents the radial reduction of inputs that is required to reach the inner boundary of the input set, holding the outputs constant. It is defined as the ratio of the minimum input required to produce the given output to the observed input. As can be seen in Figure 1.5, technical efficiency of producer D is $TE_I(y,x) = OE/OD$, where E is the minimum input vector on the ray from the origin through D that can produce y.

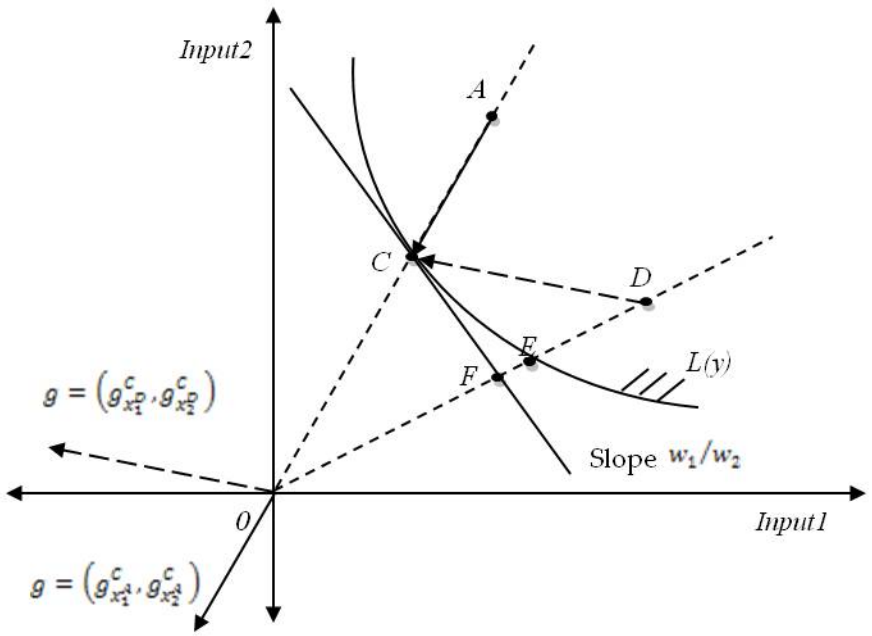

Figure 1.5: Cost and Input Technical Inefficiency

Cost efficiency is measured by the ratio of minimum or frontier cost over observed cost, $CE(x, y, w) = C(y, w)/wx = (wx^C)/wx \leq 1$, where x^C is the solution to the cost minimization problem, $C(y, w) = \min_x \{wx : f(x) \geq y, x \geq 0, (x, y) \in T\}$ where the isocost line is tangent to the input isoquant; point C as can be seen in Figure 1.5. Cost efficiency of producer D is $CE(x, y, w) = 0F/0D$. Allocative efficiency is measured by the ratio of minimum cost over frontier cost; $AE_I(x, y, w) = wx^C/wx^I$. Allocative efficiency of producer D is $AE_I(x, y, w) = 0F/0E$. Since $CE = AE_I \times TE_I$, then $TE_I = CE/AE_I = (0F/0D)/(0F/0E) = 0E/0D$.

Using the directional vector $g = \left(g_{x_1}^C, g_{x_2}^C\right)$, the directional measures of input technical inefficiency is measured by projecting any inefficient producer to the cost-minimizing bundle C where producers are both technically and allocatively efficient. All cost inefficiency for producers operate above the input isoquant can be regarded as measures of input technical inefficiency; points D and A in figure 1.5.

The Revenue and Output Technical Inefficiency

The duality between revenue and distance functions is introduced by Färe *et al.* (1993) and Färe and Grosskopf (2000). They show that the revenue function can be derived from the standard output distance function by maximizing with respect to outputs and using either unconstrained or conditional optimization.

$$R(x,p) = \max_{y} \{py / D_O(x,y)\} = \max_{y} \{py : D_O(x,y) \leq 1\}$$

$$R(x,p) = \max_{y} \left\{py + \vec{D}_O(x,y;g_y) \times pg_y\right\} = \max_{y} \left\{py : \vec{D}_O(x,y;g_y) \geq 0\right\}$$

The Luenberger (1992) and Chambers *et al.* (1998) duality theorem can also be expressed using DODF by maximizing with respect to output prices and using either unconstrained or conditional optimization[1].

$$\vec{D}_O(x,y;g_y) = \max_{p} \{(R(x,p) - py) / pg_y\} = \max_{p} \{R(x,p) - py : pg_y = 1\}$$

The relationships between the ODF, the DODF and the revenue function can be represented as [see Färe and Grosskopf (2000)]

$$\frac{1}{D_O(x,y)} \leq \frac{R(x,p)}{py} \text{ and } \vec{D}_O(x,y;g_y) \leq \frac{R(x,p) - py}{pg_y}$$

The inequality can be turned into equality by including a residual multiplicative term to the ODF and an additive term to the DODF to capture allocative inefficiency, where allocative inefficiency is due to the failure of choosing the revenue-maximizing output vector given relative output market prices.

Ignoring the price information, the Debreu-Farrell output-oriented measure of technical efficiency represents the radial expansion of outputs that is required to be considered as being technically efficient. It is defined as the ratio of the maximum potential output obtainable from the given input to the observed output. As can be seen in Figure 1.6, technical efficiency of producer D is $TE_O(x,y) = 0E/0D$, where E is the maximum output vector on the ray from the origin through D that can be produced by x.

[1]The proof of this duality can be deduced from Luenberger (1992).

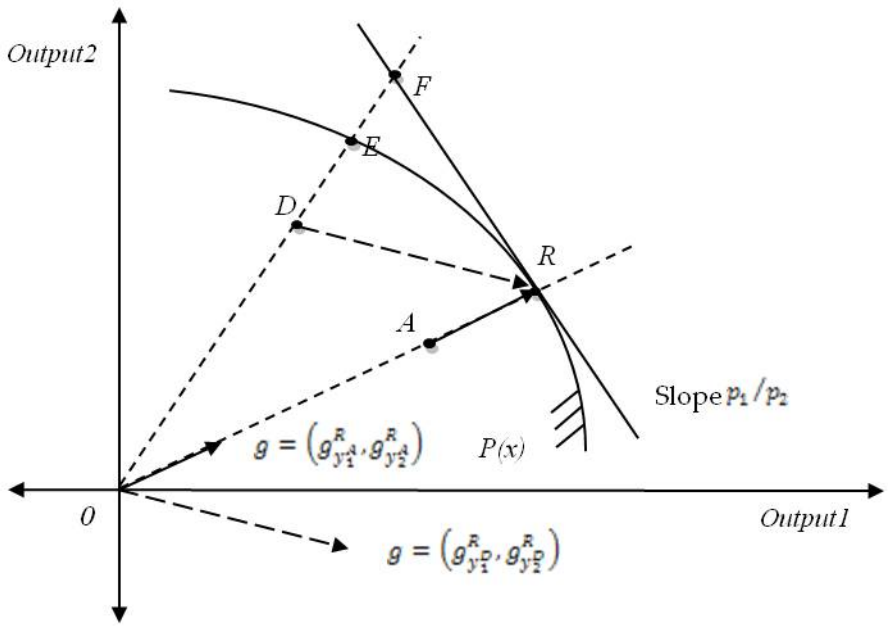

Figure 1.6: Revenue and Output Technical Inefficiency

Revenue efficiency is measured by the ratio of maximum or frontier revenue over observed revenue, $RE(y,x,p) = R(x,p)/py = py^R/py \geq 1$, where y^R is the solution to the revenue maximization problem, $R(x,p) = \max_y \{py : f(x) \geq y, y \geq 0, (x,y) \in T\}$ where the iso-revenue line is tangent to the output production curve; point R as can be seen in Figure 1.6. Revenue efficiency of producer D is $RE(y,x,p) = 0F/0D$. Allocative efficiency is measured by the ratio of maximum revenue over frontier revenue; $AE_O(y,x,p) = py^R/py^O$. Allocative efficiency of producer D is $AE_O(y,x,p) = 0F/0E$. Since $RE = AE_O \times TE_O$, then $TE_O = RE/AE_O = (0F/0D)/(0F/0E) = 0E/0D$.

Using the directional vector $g = \left(g^R_{y_1}, g^R_{y_2}\right)$, the directional measures of output technical inefficiency is measured by projecting any inefficient producer to the revenue-maximizing bundle R where producers are both technically and allocatively efficient. All revenue inefficiency for producers operate beneath the output production curve can be regarded as measures of output technical inefficiency; points D and A in figure 1.6.

The Profit and Overall Technical Inefficiency

The duality between profit and DTDF is introduced by Chambers *et al.* (1998). They show that the profit function can be derived from the DTDF by maximizing with respect to inputs and outputs and using either unconstrained or conditional optimization.

$$\begin{aligned}\pi(p,w) &= \max_{x,y}\left\{(py-wx)+\vec{D}_T(x,y;g_x,g_y)(pg_y+wg_x)\right\}\\ &= \max_{x,y}\left\{(py-wx):\vec{D}_T(x,y;g_x,g_y)\geq 0\right\}\end{aligned}$$

Chambers *et al.* (1998) show that the DTDF can be derived from the profit function by minimizing with respect to input and output prices using unconstrained optimization. This duality theorem can also be expressed using conditional optimization.

$$\begin{aligned}\vec{D}_T(x,y;g_x,g_y) &= \min_{p,w}\left\{\frac{\pi(p,w)-(py\ -\ wx)}{pg_y+wg_x}\right\}\\ &= \min_{p,w}\left\{\pi(p,w)-(py\ -\ wx):pg_y+wg_x=1\right\}\end{aligned}$$

The relationship between the DTDF and the profit function can be represented as [see Färe and Grosskopf (2000)]

$$\vec{D}_T(x,y;g_x,g_y)\leq\frac{\pi(p,w)-(py\ -\ wx)}{pg_y+wg_x}$$

The inequality can be turned into equality by adding a residual term that captures allocative inefficiency, where allocative inefficiency is due to the failure of choosing the profit-maximizing input-output vector given relative input and output market prices.

The hyperbolic and directional measures of technical inefficiency are appealing in the context of profit efficiency since they involve the adjustments of both inputs and outputs. Ignoring the price information, the directional measure of overall technical inefficiency is measured by projecting any inefficient producer to the frontier of T where producers are technically efficient using the directional vector $g=(g_x,g_y)$, such that inputs are contracted in the direction g_x and outputs are expanded in the direction g_y. Eliminating technical inefficiency for producers who operate at point D would take the producers to point $E=$

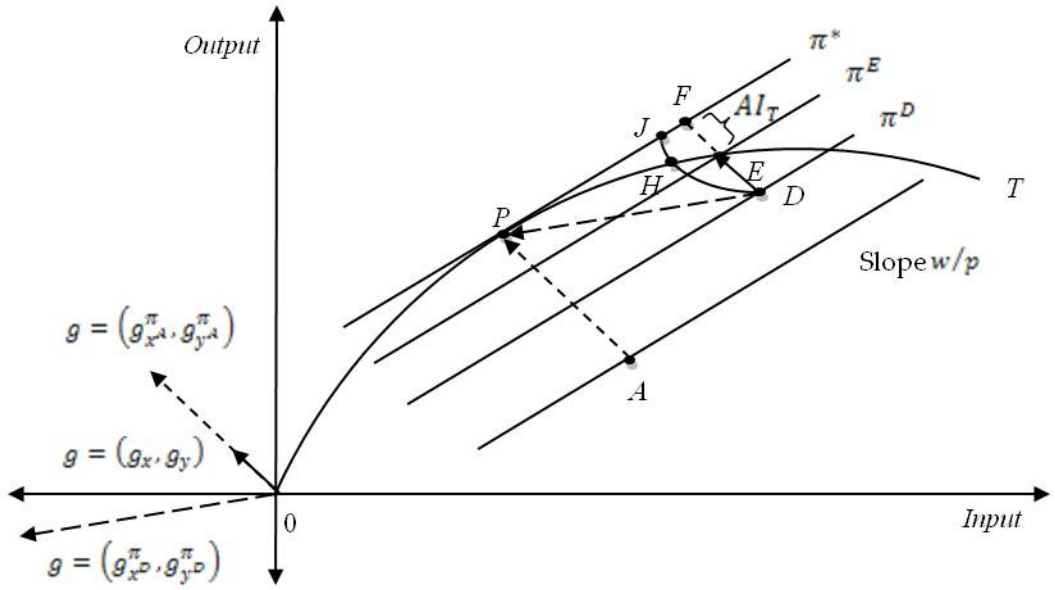

Figure 1.7: Profit and Overall Technical Inefficiency

$\left(x^E, y^E\right) = (x - \theta_T g_x, y + \theta_T g_y)$ on the frontier of T. In the context of the hyperbolic measure of technical inefficiency, eliminating technical inefficiency for producers who operate at point D would take the producers to point $H = \left(x^H, y^H\right) = \left(\vartheta_H x, \frac{y}{\vartheta_H}\right)$ on the frontier of T.

Profit of producers who operate at point D is less than maximum or frontier profit at point P, where the iso-profit line is tangent to the production frontier, as can be seen in Figure 1.7. In the context of the directional measure of inefficiency, profit inefficiency is measured by the ratio of the difference between maximum and observed profit normalized by the value of the directional vector; $PI(y, x, p, w) = (\pi(p, w) - (py - wx)) / (pg_y + wg_x)$. $\pi(p, w) = py^{\pi} - wx^{\pi}$, where (x^{π}, y^{π}) is the solution to the profit maximization problem, $\pi(p, w) = \max_{x,y} \{(py - wx) : f(x) \geq y, x \geq 0, y \geq 0, (x, y) \in T\}$. Profit inefficiency of producer D is $PI = \left(\pi^* - \pi^D\right) / (pg_y + wg_x)$. The technical inefficiency of producer D is $TI_T = \vec{D}_T(x, y; g_x, g_y) = \|DE\| / \|0g\|$. Since $PI = TI_T + AI_T$, then the residual following the path from point E to point F is the allocative inefficiency $AI_T = PI - TI_T$, where maximum profit π^* at point F equals that at point P. In the context of the hyperbolic measure of inefficiency, profit inefficiency PI represents the amount by which inputs and outputs

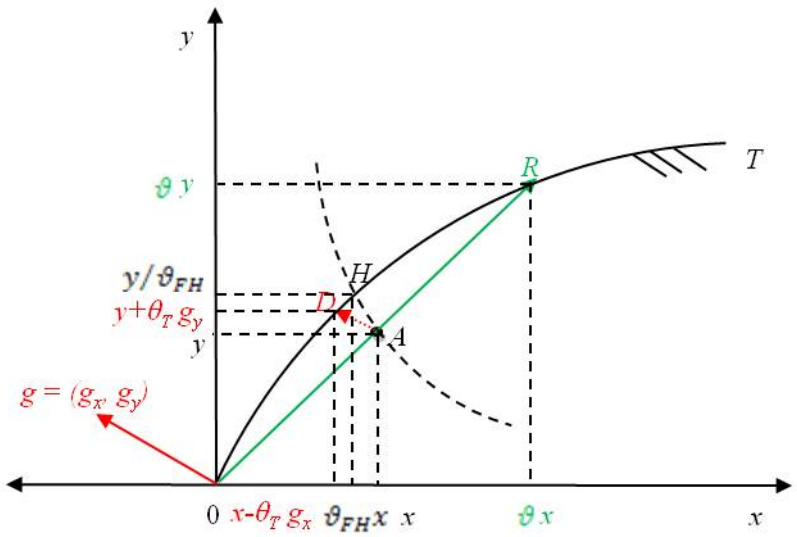

Figure 1.8: Radial, Hyperbolic, and Directional Measures of Technical Inefficiency

are projected hyperbolically to the maximum profit boundary through the hyperbolic path from point D to point J. Since $PI = TI_H \times AI_H$, then the residual $AI_H = PI/TI_H$ is the allocative efficiency.

Using the directional vector $g = \left(g_x^{\pi}, g_y^{\pi}\right)$, the directional measure of overall technical inefficiency is measured by projecting any inefficient producer to the profit-maximizing bundle P where producers are both technically and allocatively efficient. All profit inefficiency for producers operate beneath the production frontier can be regarded as measures of overall technical inefficiency — see Zofio *et al.* (2013).

To summarize, Figure 1.8 illustrates the projections of the observed input-output vector at point A by different types of distance functions. The standard distance function projects A proportionally onto R. The radial measure of technical efficiency is given by $TE_R\left(x, y\right) = \max_{\vartheta}\left\{\vartheta : (\vartheta x, \vartheta y) \in T\right\}$. However, the radial measure given by the standard distance function may produce high inefficiency measures even when the observed input-output vector is very close to the frontier — see Hudgins and Primont (2007). The HDF projects A hyperbolically onto H, where the intersection between the hyperbolic curve and the frontier of T is point $H = (\vartheta_{FH} x, y/\vartheta F_H)$. The hyperbolic measure of technical efficiency

is given by $TE_H(x,y) = \max_{\vartheta_{FH}} \{\vartheta_{FH} : (x/\vartheta_{FH}, \vartheta_{FH} y) \in T\}$. However, the hyperbolic measure given by the hyperbolic distance function is not always easy to implement due to the non-linear optimization involved. The DTDF provides greater flexibility by contracting inputs and expanding outputs simultaneously to project A onto D using the directional vector g. The directional measure of technical inefficiency is technology-oriented and given by $\vec{D}_T(x,y;g_x,g_y) = \max_{\theta_T} \{\theta_T : (x - \theta_T g_x, y + \theta_T g_y) \in T\}$. The important properties of the alternative distance functions that can be used for the measurement of technical inefficiency and the relationships among them can be summarized as in Table 1.1.

1.3 Modeling Technical Inefficiency

Unlike the deterministic frontier approach which assumes that all deviations from the efficient frontier are under the control of producers and considered as being technical inefficiency, the stochastic frontier approach introduces a random error term that captures exogenous stochastic factors beyond the control of producers into the specification of the frontier model in addition to the technical inefficiency term. The main advantage of the stochastic frontier approach is that it disentangles error term from inefficiency, thus providing more accurate measures of technical inefficiency.

This section discusses the development of modeling technical inefficiency regarding its temporal behavior, its classification, and its determinants in the stochastic frontier framework. Regarding its temporal behavior, technical inefficiency is viewed first as time-invariant in cross-section and panel data models. This assumption is relaxed with the development of the time-variant technical inefficacy models. These models allow technical inefficiency to vary over time and across individual producers. Time-invariant and time-variant inefficiency models are developed to take both inefficiency components into account. More recently, four random components inefficiency models are proposed to account for both inefficiencies and heterogeneous technology since time-invariant and time-variant inefficiency models fail

to explicitly account for unobserved heterogeneity or separate it from time-invariant inefficiency. The dynamic inefficiency models are proposed to capture the fact that the temporal behavior of inefficiency may be dynamic, where inefficiency evolves via an autoregressive process where past values of inefficiency determine the current value of inefficiency. While the threshold inefficiency models focus on the possibility of inefficient producers being out of the markets, recent studies develop the zero inefficiency models which focus on the possibility of producers being fully efficient. The zero inefficiency models can accommodate the presence of both fully efficient and inefficient producers in a probabilistic framework. Heterogeneous inefficiencies models are proposed to capture heterogeneity in the inefficiency component by either including producer-specific characteristics in the inefficiency component or the mean, variance or both parameters of the inefficiency distribution.

1.3.1 Time-Invariant Inefficiency Models

Time-invariant inefficiency models treat inefficiency as time-invariant. It is sometimes referred to as long-term or persistent inefficiency in the literature. It can be modeled using cross-section or panel data.

The Cross-Section Models

The early literature on the stochastic frontier framework uses cross-section models where specific distributions on inefficiency and error terms are assumed in order to estimate the production frontier. The distribution assumptions are also necessary to separate inefficiency from error term. The stochastic production frontier is first proposed independently by Aigner *et al.* (1977), and Meeusen and Van den Broeck (1977), while the first application is introduced by Battese and Cora (1977). It can be presented as

$$Y_i = \alpha + f(X_i; \beta) + \nu_i - u_i$$

where inputs, outputs, stochastic factors, and technical inefficiency vary only across producers. $(\nu_i - u_i)$ is a composed error term; $u_i \geq 0$ represents technical inefficiency and ν_i

represents random errors which is associated with random factors that can affect positively or negatively production. To separate the composed error term $\varepsilon_i = \nu_i - u_i$, several techniques are proposed in the literature — see section 1.4 for more details on estimation techniques.

The main issue with the cross-section model is that it relies on the strong assumption that inefficiency is independent of the regressors, the violation of which leads to inconsistent estimates of the model parameters as well as the measures of technical inefficiency.

The Panel Data Models

The use of panel data overcomes the limitations of cross-section models and has several advantages over them in time-invariant inefficiency models. It provides consistent estimates of inefficiency by adding more temporal observations for the same producer as long as the time series is sufficiently large. In addition, specific distribution assumption regarding technical inefficiency is no longer required, and all parameters of the model can be estimated using the traditional estimation procedures for panel data; fixed and random-effects. The stochastic production frontier can be presented as

$$Y_{it} = \alpha + f(X_{it}; \beta) + \nu_{it} - u_i$$

where inputs, outputs, and stochastic factors vary across time and producers but technical inefficiency varies only across producers. Even though this assumption is unrealistic, it may be the case if the time dimension of the panel is particularly short or if inefficiency is associated with management and there is no change in management during that period. However, if the time dimension is large, it seems unrealistic assuming constant inefficiency over time or for the inefficient producers to survive in the market.

The Fixed Effects Models

If technical inefficiency is considered as systematic and therefore u_i is treated as producer-specific constant or an unknown fixed parameter to be estimated, a fixed effect model can be implemented. No distribution assumption is required for u_i which is assumed to be correlated

with the regressors X_{it} or the random errors ν_{it}. Schmidt and Sickles (1984) define a fixed effects model of technical inefficiency as

$$Y_{it} = \alpha_i + f(X_{it}; \beta) + \nu_{it}$$

Since u_i is treated as fixed, it becomes producer-specific intercept $\alpha_i = \alpha - u_i$. They are regarded as fixed numbers that can be estimated as parameters, or eliminated by suitable transformation if the number of producers is too large. Schmidt and Sickles (1984) consider different procedures to estimate the fixed effects model; the within estimator, the generalized least square (GLS) estimator, and the maximum likelihood estimator. While the within estimator assumes no independence assumption between u_i and the regressors, the GLS estimator assumes that u_i are uncorrelated with the regressors. The maximum likelihood estimator assumes both distribution and independence assumptions. Sickles (2005) presents a wide variety of the identification of producer-specific technical inefficiency using panel estimators. Koop *et al.* (1997) describe procedures for Bayesian estimation of fixed effects models for technical inefficiency.

Using the within estimator, the fixed effects estimate $\widehat{\beta}$, also called the within estimate, can be estimated either by regressing $\widetilde{Y}_{it} = (y_{it} - \overline{y}_i)$ on $\widetilde{X}_{it} = (x_{it} - \overline{x}_i)$, where $\overline{y}_i = \sum_{i=1}^{T} y_{it}/T$ and $\overline{x}_i = \sum_{i=1}^{T} x_{it}/T$, thus eliminating α_i or equivalently by regressing y_{it} on x_{it} and a set of specific dummy variables for producers using ordinary least squares (OLS). As a result, $\widehat{\alpha}_i$ are obtained by averaging its residuals over time as $\widehat{\alpha}_i = \overline{y}_i - \overline{x}_i\widehat{\beta}$ or equivalently $\widehat{\alpha}_i$ are the estimated coefficients of the dummy variables. Technical inefficiency is estimated by comparing the estimated intercept of each producer to the maximum estimated value.

$$\widehat{u}_i = \max_j \{\widehat{\alpha}_j\} - \widehat{\alpha}_i$$

Producer-specific technical efficiency can be obtained from $TE_i = \exp(-\widehat{u}_i)$. However, this makes the producer with the highest intercept is regarded as fully efficient and thus inefficiency for other producers is relative to that producer. Feng and Horrace (2012) estimate

inefficiency relative to the least efficient producer instead of the highest efficient producer by comparing the estimated intercept of each producer to the minimum estimated value. They argue that these inefficiency estimates have smaller bias than those with the maximum estimated value when there are many producers operate close to the efficient frontier. However, in both cases, inefficiency is estimated as relative rather than absolute inefficiency. Furthermore, the intercept $\widehat{\alpha}_i$ captures all time-invariant unobserved heterogeneity not only those related to inefficiency. In addition, as pointed out by Kim and Schmidt (2000), Wang and Schmidt (2009) and Satchachai and Schmidt (2010), estimation of technical inefficiency based on the fixed effects estimator can be upward biased when the number of time series is small, and the number of cross-section observations is large. The max operator induces upward bias in $\widehat{\alpha} = \max_j \{\widehat{\alpha}_j\}$ which induces an upward bias in the inefficiency estimates $\widehat{u}_i$.

Wikstrom (2016) proposes a modified fixed effects estimator that does not suffer from bias in large cross-section observations using the second central moment of the inefficiency distribution to correct the intercept value obtained from the fixed effects estimator. He proposes a consistent estimator of α assuming half normal distribution and exponential distribution for u_i as $\widehat{\alpha} = \widehat{\mu}_\alpha + \widehat{\mu}_u$ where $\widehat{\mu}_\alpha = \sum_{i=1}^{N} \widehat{\alpha}_i / N$, $\widehat{\mu}_u = \widehat{\sigma}_\alpha^2 \left(2/\pi - 2\right)^{1/2}$ assuming half-normal distribution for u_i and $\widehat{\mu}_u = \left(\widehat{\sigma}_\alpha^2\right)^{1/2}$ assuming exponential distribution for u_i, $\widehat{\sigma}_\alpha^2 = \left(\sum_{i=1}^{N} \left(\widehat{\alpha}_i - \widehat{\mu}_\alpha\right)^2 / N\right) - \left(\widehat{\sigma}_v^2 / T\right)$. The modified fixed effects estimator of u_i is defined by

$$\widehat{u}_i = \widehat{\alpha} - \widehat{\alpha}_i$$

Fixed effects estimator has the advantage of not requiring a distribution assumption on inefficiency and allowing inefficiency to be correlated with any other variables. However, the time-invariance assumption of inefficiency is very restrictive and unreasonable for relatively long panels. Also, fixed effect time-invariant models are based on the assumption that all the time-invariant effects are parts of inefficiency and therefore inefficiency measures include any other source of time-invariant unobserved heterogeneity not only those related

to inefficiency, and it is not possible to identify unobserved heterogeneity from inefficiency—see, for example, Greene (2004). In addition, time-invariant regressors cannot be used in the specification of the model which lead to perfect multicollinearity between α_i and the time-invariant regressors.

The Random Effects Models

When the assumption of no correlation between the regressors and inefficiency is correct, then random effects models provide more efficient estimates than fixed effects models. Random effects time-invariant inefficiency models are introduced by Pitt and Lee (1981), Kumbhakar (1987), and Battese and Coelli (1988) in which inefficiency is treated as time-invariant. Inefficiency measures can be estimated by $E(u_i \mid \varepsilon_{it})$, where $\varepsilon_{it} = \nu_{it} - u_i$ using maximum likelihood estimation or the posterior mean $E(u_i \mid Y)$ using Bayesian estimation — see section 1.4 for more details on estimation techniques.

While fixed effect models allow for correlation between inefficiency and regressors, random effect models requires independence among them and do not allow for endogenous regressors in the model. This assumption is violated if inefficiency is related to the usage and quality of inputs and the production and quality of outputs. In addition, random effect time-invariant models are based on the assumption that all the time-invariant effects are parts of inefficiency. An advantage of the random effects models is that time-invariant regressors can be included in the model without leading to collinearity problem.

1.3.2 Time-Variant Inefficiency Models

Time-variant inefficiency is sometimes referred to as short-term or transient inefficiency in the literature. Estimates of technical inefficiency in the time-variant technical inefficiency models depend on model specifications, distribution assumptions, and the temporal behavior of inefficiency. The advantage of time-variant inefficiency model is that it allows the simultaneous specification of time-variant technical inefficiency, producer effects to account

for heterogeneous technologies, and technical change.

The Fixed Effects Models

The assumption of time-invariant inefficiency in Schmidt and Sickles (1984) model is relaxed by Cornwell *et al.* (1990) by replacing α_i with a quadratic function of time which allows technical inefficiency to vary over time and across individual producers. The model can be represented as

$$Y_{it} = \alpha_{it} + f(X_{it}; \beta) + v_{it}$$

where $\alpha_{it} = \theta_{0i} + \theta_{1i}t + \theta_{2i}t^2$. The model can be estimated by regressing the residuals for each producer $\left(\widehat{\varepsilon}_{it} = Y_{it} - X_{it}^{'}\widehat{\beta}\right)$ on a constant, time, and time-squared. The fitted values from this regression provide an estimate of α_{it}. Inefficiency measures are computed relative to the highest efficient producer over all time periods or the highest efficient producer in a given year. The latter modification allows the highest efficient producer to change from year to year.

$$\widehat{u}_{it} = \max_j \{\widehat{\alpha}_{jt}\} - \widehat{\alpha}_{it}$$

The advantages of this model are its independence of distribution assumptions on inefficiency and its allowance of inefficiency to vary across producers and time. However, it is quite restrictive in describing the temporal behavior of technical inefficiency which is assumed to be deterministic. Furthermore, this model cannot separate inefficiency from technical change since time appears in the inefficiency function.

Lee and Schmidt (1993) define technical inefficiency as the product of individual producer inefficiency and time effects; $\alpha_{it} = \theta_t\alpha_i$, where $\theta_t = \sum_t \delta_t$ with δ_t being a dummy variable for each period t and $\widehat{u}_{it} = \max_j \left\{\widehat{\theta}_t\widehat{\alpha}_i\right\} - \widehat{\theta}_t\widehat{\alpha}_i$. This specification differs from the time-invariant fixed effect model by allowing the inefficiency to vary over time. However, the temporal behavior of inefficiency is assumed to be the same for all producers.

The main issue with these fixed effects time-variant technical inefficiency models is that both models require many parameters to be estimated that can be limited to very short

panels. Moreover, inefficiency varies over time in both models by using time dummies or a time trend which prevents controlling for technical change. Greene (2005a, b) proposes what he calls true fixed effects model as

$$Y_{it} = \alpha_i + f(X_{it}; \beta) + v_{it} - u_{it}$$

Where α_i is unobserved time-invariant heterogeneity and is treated as a random variable that is correlated with X_{it} but does not capture inefficiency and can be estimated as parameters. True fixed effects model can be estimated by adding producer dummy variables to the model. However, the disadvantage of this model is that it introduces the incidental parameters (the number of fixed-effect parameters) problem which results in inconsistency for a finite number of producers and a fixed time due to the number of unknown parameters to be estimated increases with the number of producers — see, for example, Neyman and Scott (1948). Recent studies consider eliminating the problem of incidental parameters in the true fixed effects model by using within transformation to eliminate the producer effects for unobserved heterogenoty[2] — see, for example, Wang and Ho (2010) and Chen, Schmidt, and Wang (2014).

The Random Effects Models

In these models, time-variant inefficiency can be either identically and independently distributed (iid) across producers over time or modeled as a product of a deterministic function of time, $g(t; \gamma)$, and a non-negative time-invariant random variable u_i; $(u_{it} = g(t; \gamma)u_i)$ where γ is a parameter to be estimated. Thus $g(t; \gamma)$ allows the data to determine the temporal behavior of inefficiency instead of imposing it a priori. Technical inefficiency is then estimated from $\widehat{u}_{it} = \widehat{g}(t; \gamma)E(u_i \mid \varepsilon_i)$ or alternatively $\widehat{u}_i = E(u_i \mid \varepsilon_{it})$ — see Kumbhakar and Lovell (2000).

[2]Performing the within transformation on Greene (2005a) true fixed effects model yields $\widetilde{Y}_{it} = \widetilde{X}_{it}\beta + \widetilde{v}_{it} - \widetilde{u}_{it}$ where $\widetilde{Y}_{it} = y_{it} - \overline{y}_i$ are the deviations from the producer means, $\overline{y}_i = \sum_t y_{it}/T$. Similarly for $\widetilde{X}_{it}$, $\widetilde{v}_{it}$, and $\widetilde{u}_{it}$. The transformation from Y_{it} to $\widetilde{Y}_{it}$ is called the within transformation. Note that this transformation removes time-invariant heterogeneity α_i since $\widetilde{\alpha}_i = 0$. See Hsiao (2003) for a detailed discussion regarding the advantages of using within transformation.

Kumbhakar (1990) assumes that $u_{it} = u_i \left(1 + \exp\left(\gamma_1 t + \gamma_2 t^2\right)\right)^{-1}$. His specification allows technical inefficiency to increase or decrease over time monotonically, depending on the signs and magnitudes of γ_1 and γ_2. Battese and Coelli (1992) and Battese and Tessema (1993) assume that $u_{it} = \exp\left(-\gamma\left(t - T\right)\right) u_i$. Their specification implies that the temporal behavior of technical inefficiency is monotonic that allows inefficiency to increase or decrease over time exponentially, depending on the sign of γ. Technical inefficiency either increases at a decreasing rate when γ is positive or decreases at an increasing rate when γ is negative. Time-invariant model is obtained when γ is equal to zero. Kumbhakar and Wang (2005) assume that $u_{it} = \exp\left(-\gamma\left(t - \underline{t}\right)\right) u_i$. Technical inefficiency evolves over time according to $\exp\left(-\gamma\left(t - \underline{t}\right)\right)$ where $\underline{t}$ denotes the initial period and thus $u_{it} = u_i$ at time $\underline{t}$.

Lee and Schmidt (1993) assume that $u_{it} = \gamma_t u_i$ where γ_t are the parameters associated with the time dummy variables that need to be estimated. While Battese and Coelli (1992) and Battese and Tessema (1993) assume an unrealistic restriction that the temporal behavior of technical inefficiency is the same for all producers, Cuesta (2000) extends Battese and Coelli (1992) model to allow for greater flexibility for technical inefficiency to change over time by assuming $u_{it} = \exp\left(-\gamma_i\left(t - T\right)\right) u_i$, $u_{it} = \exp\left(g_i\left(t, T, z_{it}\right)\right) u_i$. His specification allows technical inefficiency to evolve over time at a different rate among producers; thus each producer has its own time path of technical inefficiency.

Cuesta and Orea (2002) and Feng and Serletis (2009) extend Battese and Coelli (1992) model by assuming $u_{it} = \exp\left(-\gamma_1\left(t - T\right) - \gamma_2\left(t - T\right)^2\right) u_i$. Their specification relaxes the monotonicity of the time path of inefficiency using two parameters specification. Therefore, their model allows inefficiency to be convex or concave and increasing some years and decreasing in others.

The main advantage of random effects over fixed effects time-variant technical inefficiency models is its allowance for the inclusion of time-invariant regressors in the model. However, random effect models require independence between inefficiency and regressors in the model

while this condition is not required in the fixed effects models.

Hausman and Taylor (1981) model is a mixture of fixed and random effects models that allows inefficiency to be uncorrelated with some but not all regressors and also allows for the inclusion of time-invariant regressors in the model. In that case, producer inefficiency can be consistently estimated and separated from the producer effects or the intercept as long as cross-section and temporal observations are large enough.

To separate producer heterogeneity or producer effects from inefficiency, where inefficiency is time-variant that can either be iid or a function of exogenous variables, Greene (2005a, b) adds a time-invariant random effect for unobserved heterogeneity and proposes what he calls true random effects model.

$$Y_{it} = (\alpha + w_i) + f(X_{it}; \beta) + v_{it} - u_{it}$$

where $\alpha_i = \alpha + w_i$ is unobserved time-invariant heterogeneity and is treated as a random variable that is uncorrelated with X_{it}. Note that Kumbhakar and Wang (2005) also introduce these producer-specific intercepts α_i to account for heterogeneous technologies. If α_i is treated as a random variable that is correlated with X_{it} but does not capture inefficiency, then the model turns to the true fixed effects model.

1.3.3 Time-Invariant and Time-Variant Inefficiency Models

Previous models for panel data focus either on time-invariant inefficiency or time-variant inefficiency. None of these models takes both inefficiencies into account. Mundlak (1961) notes that time-invariant inefficiency reflects the effects of inputs such as management; thus it is important to estimate it particularly in short panels. However, for large panels or if there are changes in management, time-variant inefficiency is also important to be estimated. Colombi *et al.* (2014) argue that time-variant inefficiency arises due to the failure of allocating resources appropriately in the short run. Tsionas and Kumbhakar (2014) note that estimating a model with only one inefficiency component with or without controlling

for producer effects for unobserved heterogeneity gives incorrect estimates of inefficiency. Kumbhakar and Heshmati (1995) propose a model in which technical inefficiency is assumed to have time-invariant and time-variant components.

$$Y_{it} = \alpha + f(X_{it}; \beta) + v_{it} - u_{i0} - u_{it}$$

where u_{i0} represents time-invariant inefficiency, u_{it} represent time-variant inefficiency, and $u_{i0} + u_{it}$ is total technical inefficiency. The error component is assumed to be independent of each other and also independent of X_{it}. The model can be estimated using several estimation techniques, including maximum likelihood estimation, Bayesian estimation, or three step procedure as follow: First, using a standard random effect model for panel data gives consistent estimates of the model parameters and predicted values of u_{i0} and u_{it}. Second, the time-invariant technical inefficiency can be estimated as $\widehat{u}_{i0} = \max_j \{\widehat{u}_{j0}\} - \widehat{u}_{i0}$. Finally, the time-variant technical inefficiency can be estimated by maximizing the log-likelihood function for pooled data $[r_{it} = \alpha + v_{it} - u_{it}]$ where $r_{it} = Y_{it} - f(X_{it}; \beta) + u_{i0}$. Estimates of u_{it} conditional on the estimated $(\varepsilon_{it} = \nu_{it} - u_{it})$ is obtained from $\widehat{u}_{it} = E(u_{it} \mid \varepsilon_{it})$ following Jondrow *et al.* (1982). Then, total technical efficiency is defined as the product of time-invariant and time-variant technical efficiencies.

$$\text{Total efficiency}_{it} = \exp\left[-\widehat{u}_{i0}\right] \times \exp\left[-\widehat{u}_{it}\right] \tag{1.2}$$

1.3.4 Four Random Components Inefficiency Models

Since time-invariant and time-variant inefficiency models fail to explicitly account for unobserved heterogeneity or separate it from time-invariant inefficiency, Kumbhakar *et al.* (2014), Colombi *et al.* (2014), Tsionas and Kumbhakar (2014) and Fillipini and Greene (2016) generalize the true random effects model proposed by Greene (2005a, b) by adding a time-invariant inefficiency and introduce four random components inefficiency models to account for both inefficiencies and heterogeneity. They decompose the time-invariant producer effect as a

producer effect and a time-invariant inefficiency effect.

$$Y_{it} = \alpha + f(X_{it};\beta) + \omega_i + v_{it} - u_{i0} - u_{it}$$

where the error term has four random components; ω_i are random producer effects for unobserved heterogeneity, u_{i0} are time-invariant inefficiency, u_{it} time-variant inefficiency, and v_{it} are random errors. Kumbhakar *et al.* (2014) estimate the model in three steps procedure. First, the model can be rearranged as $Y_{it} = \alpha + f(X_{it};\beta) + \xi_i + \varepsilon_{it}$ where $\xi_i = \omega_i - u_{i0}$ and $\varepsilon_{it} = \nu_{it} - u_{it}$ and ξ_i can be viewed as the producer specific component. Using a standard random effects model for panel data gives consistent estimates of the model parameters and predicted values of $\widehat{\xi_i}$ and $\widehat{\varepsilon}_{it}$. Second, estimates of u_{it} conditional on the estimated $(\varepsilon_{it} = \nu_{it} - u_{it})$ is obtained from $\widehat{u}_{it} = E(u_{it} \mid \varepsilon_{it})$ following Jondrow *et al.* (1982). Third, a similar procedure as in step two can be used to estimate time-invariant inefficiency component u_{i0}. Then total technical efficiency is defined as the product of time-invariant and time-variant technical efficiencies as in equation (1.2).

Note that Kumbhakar *et al.* (2014) use the procedure of Jondrow *et al.* (1982) which implicitly assumes that the marginal distribution of inefficiency given the observations is truncated normal. However, as shown by Cartinhour (1990) and Horrace (2005), the marginal distribution of a multivariate truncated normal distribution is not a truncated normal distribution.

Instead of the three steps procedure used in Kumbhakar *et al.* (2014), Colombi *et al.* (2014) consider a single step and use maximum likelihood estimation using results from the closed skew normal distribution. The maximum likelihood is asymptotically more efficient than the three steps procedure used in Kumbhakar *et al.* (2014) since it estimates all parameters simultaneously. However, Tsionas and Kumbhakar (2014) note that the maximum likelihood estimation used in Colombi *et al.* (2014) is computationally prohibitive when T is large because the likelihood function depends on a $(T + 1)$ dimensional integral of the normal distribution.

While Tsionas and Kumbhakar (2014) use Bayesian estimation to estimate the model using a large panel of US banks, Fillipini and Greene (2016) use simulated maximum likelihood estimation proposed by Greene (2005a, b) to estimate the parameters and the different random components of the model. Based on the moment generating function for the closed skew normal distribution proposed by Colombi *et al.* (2014), they estimate the values of technical efficiency. Total technical efficiency is defined as the product of time-invariant and time-variant technical efficiencies as in equation (1.2).

1.3.5 Dynamic Inefficiency Models

The temporal behavior of inefficiency in dynamic inefficiency models is dynamic where inefficiency evolves via an autoregressive process where past values of inefficiency determine the current value of inefficiency. However, few studies use these models to measure inefficiency — see, for example, Ahn and Sickles (2000), Tsionas (2006) and Emvalomatis (2012).

Ahn and Sickles (2000) assume that inefficiency follows the first order autoregressive process $AR(1)$, where the current inefficiency u_{it} depends on two components; the unadjusted portion of the past period inefficiency $(1-\rho_i)\,u_{i,t-1}$, where $0 < \rho_i \leq 1$ is the adjustment speed and the new unexpected inefficiency e_{it}.

$$u_{it} = (1-\rho_i)\,u_{i,t-1} + e_{it}$$

Tsionas (2006) applies Bayesian estimation to a panel of large US commercial banks and assumes that inefficiency evolves over time log-linearly and as a function of explanatory variables that reflect producer-specific characteristics to account for heterogeneity in inefficiency. Specific assumptions are also made for the initial value u_{i1}.

$$\ln u_{it} = z_{it}\delta + \rho \ln u_{i,t-1} + e_{it} \qquad \text{for } t = 2, ..., T$$

$$\ln u_{i1} = z_{i1}\delta/\left(1-\rho\right) + e_{i1} \qquad \text{for } t = 1$$

where $e_{it} \sim N(0, \sigma_e^2)$, $e_{i1} \sim N(0, \sigma_e^2/\left(1-\rho^2\right))$. Note that the specification of log-normality of inefficiency is used by Deprins and Simar (1989). However, the assumption of a log-normal

distribution for inefficiency cannot accommodate a situation where most producers are fully efficient. For $\ln u_{it}$ process to be stationary, $|\rho|$ should be less than one. Tsionas (2006) finds the posterior mean $\rho = 0.91$ which implies that the autoregressive process is almost static. Emvalomatis (2012) extends Tsionas (2006) model by separating unobserved heterogeneity from inefficiency where inefficiency evolves over time in a dynamic context.

1.3.6 Threshold Inefficiency Models

In contrast to the models that allow for the existence of extremely inefficient producers who cannot survive in highly competitive markets, the threshold inefficiency models truncate the distribution of inefficiency by placing a threshold parameter of the minimum efficiency for survival on inefficiency. Thus, these models specify an upper bound to the distribution of inefficiency in addition to the zero lower bound.

While Lee (1996) introduces a tail truncated half normal distribution with a threshold parameter θ; $u_i \sim N^+(0, \sigma_u^2)$, $0 \leq u_i \leq \theta$, Lee and Lee (2014) assume a uniform distribution, $u_i \sim U(0, \theta)$. Almanidis, Qian, and Sickles (2014) extend Lee (1996) model to panel data model and assume that u_{it} is drawn from a time-variant distribution with upper bound θ_t which is assumed to be the sum of weighted polynomials, $\theta_t = \sum_{i=0}^{N} b_i \left(t/T\right)^i$, where $t = 1, ..., T$ and b_i are constants. The threshold inefficiency models are useful for empirical studies focusing on estimating the inefficiency threshold.

1.3.7 Zero Inefficiency Models

While the threshold inefficiency models focus on the possibility of inefficient producers being out of the markets, recent studies develop the zero inefficiency models which focus on the possibility of producers being fully efficient.

Wheat *et al.* (2014) note that the probability of inefficiency being zero for any producer is zero in models that do not allow for the presence of fully efficient producers. However, Bos, Economidou, and Koetter (2010), and Bos *et al.* (2010) use latent class models and find

small groups of producers that are fully efficient. Kumbhakar, Parmeter, and Tsionas (2013) note that if the data represent a mixture of both fully efficient and inefficient producers, then models that impose inefficient behavior on all producers result in biased estimates of inefficiency. They introduce the zero inefficiency model which can accommodate the presence of both fully efficient and inefficient producers in a probabilistic framework.

Assuming that some producers are fully efficient where $u_i = 0$ for some i while others are inefficient where $u_i > 0$, the zero inefficiency model can be represented as

$$Y_i = \begin{array}{ll} f(X_i;\beta) + v_i & \text{with probability } p \\ f(X_i;\beta) + v_i - u_i & \text{with probability } (1-p) \end{array}$$

where p is the probability of a producer being fully efficient or the proportion of producers that are fully efficient and $(1-p)$ is the proportion of producers that are inefficient. Kumbhakar, Parmeter, and Tsionas (2013) define the estimates of inefficiency as $\widetilde{u}_i = (1-\widetilde{p}_i)\,\widehat{u}_i$, where $\widehat{u}_i$ is the zero inefficiency estimator of inefficiency with $p = 0$ and $\widetilde{p}_i$ is the estimate of the probability of being fully efficient.

Kumbhakar, Parmeter, and Tsionas (2013) and Rho and Schmidt (2015) propose the probability or the proportion of producers being fully efficient as a parametric function of a set of explanatory variables which determine full efficiency via a logit or a probit function. However, Tran and Tsionas (2016b) argue that misspecification of the parametric functional form of the probability of producers being fully efficient has implication for which producers are fully efficient as well as the estimates of technical inefficiency. They use a non-parametric formulation for the probability of producers being fully efficient via an unknown smooth function of explanatory variables which influence the likelihood that a producer is fully efficient.

Since zero inefficiency model deals a priori with two classes; fully efficient and inefficient producers, it does not face the problem of identifying the number of classes as in latent class models. However, Rho and Schmidt (2015) discuss the presence of the wrong skewness problem of Waldman (1982) as well as identification issues within the zero inefficiency models.

They argue that when all producers are fully efficient, it is not clear whether efficiency is due to p being close to 1, or σ_u^2 being close to zero which has important implications for conducting inference. Another concern with the zero inefficiency models is that the consistency of the estimates depends on the exogeneity of the regressors. Tran and Tsionas (2016a) investigate the endogeneity issues in the zero inefficiency models by using simultaneous equations setting allowing for one or more regressors to be endogenous.

1.3.8 Heterogeneous Inefficiency Models

Heterogeneous inefficiency models are proposed to capture heterogeneity in inefficiency by either including producer-specific characteristics Z in the inefficiency component or the mean, variance or both parameters of the inefficiency distribution. These models are also useful for understanding the relationships between inefficiency and its exogenous determinants.

Heterogeneous inefficiency models can be estimated either by using the two steps procedure by which inefficiency and explanatory variables Z are estimated sequentially or the one-step procedure by which the explanatory variables are estimated simultaneously with the other model parameters. However, the two steps procedure is criticized regarding its misspecification of the first step model, suffering from omitted variables bias if X and Z are correlated, and its bias from ignoring the impact of Z on inefficiency — see, for example, Caudill and Ford (1993), Battese and Coelli (1995), and Wang and Schmidt (2002).

Determinants of Inefficiency Models

In these models, inefficiency is modeled as a function of explanatory variables Z that reflect producer-specific characteristics and explain the differences in inefficiency across producers — see, for example, Deprins and Simar (1989), Kumbhakar *et al.* (1991), and Huang and Liu (1994) who add interaction terms between Z and the regressors, $z_i x_i$.

$$u_i = g\left(z_i, z_i x_i; \delta\right) + e_i$$

where δ are unknown parameters to be estimated and e_i is a random variable defined by the truncation of a normal distribution. If there are no interaction terms $z_i x_i$, the model reduces to the Deprins and Simar (1989) and Kumbhakar *et al.* (1991) models. Tsionas (2006) extends Kumbhakar *et al.* (1991) model to panel data model that allows for dynamic technical inefficiency; $\ln u_{it} = z_{it}\delta + \rho \ln u_{i,t-1} + e_{it}$. Srairi (2010) extends Kumbhakar *et al.* (1991) model to panel data where $u_{it} = g(z_{it}; \delta) + e_{it}$ to examine bank-specific variables that may explain the sources and differences of inefficiency across producers.

Determinants of inefficiency models encounter the issue of obtaining the non-negative inefficiency. The solution of Kumbhakar *et al.* (1991) to deal with this issue is $u_i = |N(z_i\delta, \sigma_u^2)|$. Reifschneider and Stevenson (1991) assume that $u_i = u_i^* + \exp(z_i\delta)$ where both $u_i^* \sim N^+(0, \sigma_u^2)$ and $\exp(z_i\delta)$ are positive. However, it is not necessary for both of them to be positive in order to obtain a positive u_i. Huang and Liu (1994) extend Reifschneider and Stevenson (1991) assumption by assuming only $u_i^* \geq -\exp(z_i\delta)$.

Determinants of Inefficiency Distribution Models

In these models, producer-specific characteristics can be included in the mean, variance or both parameters of the inefficiency distribution. Battese and Coelli (1995) and Wang and Ho (2010) assume that the mean of the inefficiency distribution can be modeled as a function of explanatory variables that reflect producer-specific characteristics.

$$u_{it} = g(z_{it}; \delta)\, u_i,\ u_{it} \sim N^+(\mu_{it}, \sigma_u^2),\ \mu_{it} = z_{it}\delta_m$$

Including producer-specific characteristics in the variance of the inefficiency distribution is first motivated by the possible presence of heteroscedasticity in inefficiency. Reifschneider and Stevenson (1991), Caudill and Ford (1993) and Caudill *et al.* (1995) assume that u are heteroskedastic and include the standard deviation in exponential form to ensure a positive estimate of the variance parameter for all Z and γ_u.

$$\sigma_{ui} = \exp(z_{ui}\gamma_u),\ \sigma_{vi} = \exp(\gamma_v)$$

It is also possible to assume that both u and v are heteroskedastic which referred to as the doubly heteroskedastic model in the literature. The variance parameters of u and v distributions are modeled as a function of explanatory variables z_{ui} and z_{vi} which may or may not be equivalent — see, for example, Hadri (1999) and Hadri *et al.* (2003).

$$\sigma_{ui} = \exp\left(z_{ui}\gamma_u\right),\ \sigma_{vi} = \exp\left(z_{vi}\gamma_v\right)$$

Including producer specific characteristics in both the mean and the variance of the inefficiency distribution allows for non-monotonic inefficiency across producers — see, for example, Wang (2002) and Wang and Schmidt (2002).

$$u_{it} \sim N^{+}(\mu_{it}, \sigma^2_{uit}),\ \mu_{it} = z_{it}\delta_m,\ \sigma^2_{uit} = \exp\left(z_{it}\gamma_u\right)$$

Kumbhakar and Wang (2005) assume that the variance parameter of v distribution can also be modeled as a function of explanatory variables z_{vi} besides the mean and the variance of the inefficiency distribution.

$$u_i \sim N^{+}(\mu_i, \sigma^2_{ui}),\ \mu_i = z_i\delta_m,\ \sigma^2_{ui} = \exp\left(z_{ui}\gamma_u\right),\ v_{it} \sim N(0, \sigma^2_{vi}),\ \sigma^2_{vi} = \exp\left(z_{vi}\gamma_v\right)$$

A critical question that needs to be considered regarding heterogeneous inefficiencies is whether heterogeneity in the inefficiency exists, or whether producer-specific inefficiency depends on a set of exogenous determinants. As suggested by Kim and Schmidt (2008), the presence of the determinants of inefficiency can be tested by regressing Y on X and Z and test the significance of the parameters of the determinants of inefficiency using F-test.

A summary of the main characteristics of technical inefficiency models that are widely used in the literature is presented in Table 1.2.

1.4 Estimation Techniques

Several econometric estimation techniques with recent developments are proposed in the literature to estimate technical inefficiency in the stochastic frontier framework. Fixed and

random effects estimators are briefly discussed in the previous section. Sickles (2005) summarizes different panel frontier estimators of technical inefficiency that are used in the literature. For the developments in the econometric estimation techniques, see, for example, Bauer (1990), Greene (1993) and Parmeter and Kumbhakar (2014). Since this literature is very extensive, this section gives a brief review of the most common estimation techniques, including maximum likelihood and Bayesian estimations.

1.4.1 Maximum Likelihood

Technical inefficiency in the stochastic frontier approach can be estimated using maximum likelihood estimation which requires a distribution assumption for the technical inefficiency as well as the random error in order to disentangle one from the other. Several distributions are assumed in the literature for technical inefficiency. The most frequently used being the half normal, exponential, gamma, truncated normal, and the skew normal distributions. Greene (1993) uses different distribution assumptions and shows that inefficiency measures are similar across different distributions. Berger and DeYoung (1997) find that assuming a truncated normal distribution for inefficiency gives similar but statistically significant estimates compared with the half normal assumption. However, Baccouche and Kouki (2003) find that estimates of technical inefficiency depend heavily on the distribution assumption. Since there is no consensus on whether technical inefficiency depends on the distribution assumption or not, further research is needed investigating this issue.

Maximum likelihood estimation (MLE) is based on the specification of the model through the joint probability density function (PDF), $f(Y, \theta)$. Assuming independence, the joint density of Y is the product of the densities of the individual observations $f_i(Y_i, \theta)$.

$$f(Y, \theta) = \prod_{i=1}^{N} f_i(Y_i, \theta)$$

Since the product is very large or very small number, it is more convenient to work with the

log-likelihood function.

$$L(Y,\theta) = \log f(Y,\theta) = \sum_{i=1}^{N} \log f_i(Y_i,\theta) = \sum_{i=1}^{N} \log l_i(Y_i,\theta)$$

where $L(Y,\theta)$ represents the likelihood of the parameters θ given the observed data Y. Note that $L(Y,\theta)$ gives the same parameter estimates since it is a monotonic transformation of $f(Y,\theta)$. The MLE of the parameters of the model can be obtained by maximizing the likelihood function with respect to the parameters. The estimated parameters are then used to obtain the estimate of technical inefficiency by using one of the inefficiency estimators.

Normal-Half Normal Models

The likelihood function for the normal–half normal cross-section models is derived by Aigner *et al.* (1977). Under the assumption that inefficiency has half normal distribution $u_i \sim N^{+}(0,\sigma_u^2)$; $f(u) = \left(\sigma_u\sqrt{2\pi}\right)^{-1} \exp\left(-u^2/2\sigma_u^2\right)$, the random errors have normal distribution $v_i \sim N(0,\sigma_v^2)$, and u_i and v_i are assumed to be identically and independently distributed, the likelihood function can be defined as the product of the densities of the composed error term $\prod_{i=1}^{N} f_\varepsilon(\varepsilon_i)$, where $f_\varepsilon(\varepsilon_i)$ is the density of the composed error term $\varepsilon_i = \nu_i - u_i$. Technical inefficiency can be estimated using Jondrow *et al.* (1982) estimator for the half normally distributed inefficiency for cross-section models

$$\widehat{u}_i = E(u_i \mid \varepsilon_i) = \sigma_S \left[\frac{\phi(\psi_i)}{1-\Phi(\psi_i)} - \psi_i\right] \tag{1.3}$$

where $\phi(.)$ is the density of the standard normal distribution, $\Phi(.)$ is the cumulative density function, $\sigma_S = \sigma\lambda/(1+\lambda^2) = (\sigma_u^2\sigma_v^2/\sigma^2)^{1/2} = \sigma_u\sigma_v/\sigma$, $\sigma = (\sigma_u^2+\sigma_v^2)^{1/2}$, $\psi_i = \lambda\varepsilon_i/\sigma$, and $\lambda = \sigma_u/\sigma_v$. The inefficiency estimator can be implemented by evaluating it at the estimated parameters $\left(\widehat{\alpha},\widehat{\beta},\widehat{\sigma}_u^2,\widehat{\sigma}_v^2\right)$ and the implied values of $\widehat{\lambda}$, $\widehat{\sigma}^2$, and $\widehat{\varepsilon}_i = Y_i - \widehat{\alpha} - X_i'\widehat{\beta}$. However, Wang and Schmidt (2009) show that the distribution of $\widehat{E}(u_i \mid \varepsilon_i)$ differs from the distribution of u_i unless $\sigma_v \to 0$ and converges to $E(u_i)$ as σ_v^2 increases which means that ε_i is no longer useful in predicting inefficiency through the conditional mean of Jondrow *et al.*

(1982) estimator as σ_v^2 increases. Battese and Coelli (1988) propose an alternative efficiency estimator given by

$$E(TE_i \mid \varepsilon_i) = E\left(\exp(-u_i) \mid \varepsilon_i\right) = \frac{1-\Phi\left(\sigma_S - \psi_i\right)}{1-\Phi\left(\psi_i\right)} \exp\left(\sigma_S \psi_i + \left(\left(\sigma_S\right)^2/2\right)\right)$$

Fried *et al.* (2008) note that the efficiency estimator of Battese and Coelli (1988) is preferable to $1 - E(u_i \mid \varepsilon_i)$ used in Jondrow *et al.* (1982) estimator because Jondrow *et al.* (1982) estimator is no more than the first order approximation to the more general infinite power series approximation, $\exp(-(u_i \mid \varepsilon_i)) = 1 - u_i + u_i^2/2 - u_i^3/3...$ However, Fried *et al.* (2008) and Kumbhakar *et al.* (2014) note that Jondrow *et al.* (1982) and Battesse and Coelli (1988) estimators are not consistent in cross-section models. Although they are unbiased, they are not consistent estimates of technical efficiency, since $p\lim E(u_i \mid \varepsilon_i) - u_i \neq 0$ or $E(u_i \mid \varepsilon_i)$ never approach u_i as the number of producers approaches infinity.

Greene (1990) notes that the half normal assumption for the distribution of inefficiency is relatively not flexible and implicitly assumes that most producers are nearly fully efficient. Furthermore, the distribution of the composed error term ε_i is no longer normal — see, for example, Horrace (2005). More precisely, it may be incorrectly skewed in the positive direction which leads to full efficiency measures for all producers[3]. Waldman (1982) finds that if ε_i are positively skewed in the wrong direction, the maximum likelihood estimates is equivalent to OLS estimates for $(\alpha, \beta, \sigma_u^2, \sigma_v^2)$ and zero for λ. The wrong skewness direction of ε_i, and consequently a zero maximum likelihood estimate of σ_u^2 results from the dependence of the maximum likelihood estimator for σ_u^2 on the skewness of ε_i in the normal-half normal model[4]. Feng, Horrace, and Wu (2013) suggest using constrained optimization methods to impose the restriction that $\sigma_u^2 > 0$ in the normal-half normal model. Hafner, Manner, and Simar (2018) generalize the inefficiency distribution that allows for the existence of the wrong

[3]Azzalini (1985) defines a continuous random variable ε to have a skew-normal distribution if it has density function $f(\varepsilon) = 2\phi(\varepsilon)\Phi(a\varepsilon)$, where a is a fixed arbitrary number. The distribution is right-skewed if $a > 0$ and is left-skewed if $a < 0$.

[4]Proper specification testing can be undertaken to check the sign of the skewness of the OLS residuals. See, for example, Kuosmanen and Fosgerau (2009).

skewness direction of ε_i while obtaining well-defined inefficiency measures.

Pitt and Lee (1981), Kumbhakar (1987), and Battese and Coelli (1988) extend the normal-half normal model proposed by Aigner *et al.* (1977) to the panel data time invariant inefficiency model. Technical inefficiency can be estimated using the extension of Jondrow *et al.* (1982) estimator to the panel data model

$$\widehat{u}_i = E(u_i \mid \varepsilon_{it}) = \varphi_i^N + \sigma_P \left[\frac{\phi\left(\varphi_i^N/\sigma_P\right)}{1 - \Phi\left(\varphi_i^N/\sigma_P\right)} \right] \tag{1.4}$$

where $\varphi_i^N = \left(-\sigma_u^2 \sum_{t=1}^{T} \varepsilon_{it}\right) / \left(\sigma_v^2 + T\sigma_u^2\right)$, and $\sigma_P = \left(\sigma_v^2 \sigma_u^2\right) / \left(\sigma_v^2 + T\sigma_u^2\right)$. Kumbhakar (1987) points out that these estimates are asymptotically consistent. Lee (1996) introduces a tail truncated half normal distribution to incorporate a bound for inefficiency since the least efficient producers cannot survive in highly competitive markets. He introduces the threshold parameter of the minimum efficiency for survival θ as $u_i \sim N^+(0, \sigma_u^2)$, $0 \leq u_i \leq \theta$. Thus, the variance of inefficiency depends on two parameters, σ_u^2 and θ.

Normal-Exponential Models

Aigner*et al.* (1977) and Meeusen and van den Broeck (1977) propose a likelihood function under the assumption that u_i have an exponential distribution; $f(u) = \theta \exp\left(-\theta u\right)$, $\theta > 0$ where $\theta = \sigma_u^{-1}$, and $v_i \sim N(0, \sigma_v^2)$. Technical inefficiency can be estimated using Jondrow *et al.* (1982) estimator for exponentially distributed inefficiency for cross-section models

$$\widehat{u}_i = E(u_i \mid \varepsilon_i) = \sigma_v \left[\frac{\phi\left(\varphi_i^E\right)}{\Phi\left(\varphi_i^E\right)} + \varphi_i^E \right] \tag{1.5}$$

where $\varphi_i^E = \left(\varepsilon_i \sigma_u - \sigma_v^2\right) / \left(\sigma_u \sigma_v\right)$. Kim and Schmidt (2000) extend the normal-exponential model proposed by Aigner *et al.* (1977) to the panel data time-invariant inefficiency model. Technical inefficiency can be estimated using the extension of Jondrow *et al.* (1982) estimator to the panel data model by replacing ε_i by $\overline{\varepsilon}_i$ and σ_v^2 by σ_v^2/T in equation (1.5). A recent simulation study by Horrace and Parmeter (2018) argue that assuming Laplace distribution for v_i and truncated Laplace distribution for u_i results in a Laplace model that performs relatively well compared to the normal exponential model when v_i is misspecified.

Normal-Gamma Models

Greene (1980a, b), Stevenson (1980, 1990), and Greene (1990) assume a gamma distribution for the inefficiency term where $f(u) = \left[\theta^P/\Gamma(P)\right] \exp(-\theta u)\, u^{P-1}$, $P > 0$, $\theta = \sigma_u^{-1}$, $\Gamma(P) = \int_0^\infty t^{P-1} e^{-t} dt$, and $v_i \sim N(0, \sigma_v^2)$. Stevenson (1980) only considers the Erlang form (integer values of P; 1.0 and 2.0) which produces a tractable formulation for $f_\varepsilon(\varepsilon_i)$ but greatly restricts the model. Beckers and Hammond (1987) derive the log-likelihood function for $f_\varepsilon(\varepsilon_i)$ without restricting P to be an integer, but the resulting functional form is intractable. When $P = 1$, the normal-gamma model returns to the normal-exponential model. The inefficiency estimator for the gamma model is

$$\widehat{u}_i = E(u_{it} \mid \varepsilon_{it}) = \frac{q(P, \varepsilon_{it})}{q(P-1, \varepsilon_{it})}$$

The normal-gamma distribution provides a more flexible parameterization of the distribution. However, the computational complexity of the maximum likelihood estimator restricts using this model in empirical studies. Several attempts are developed in the literature by Ritter and Simar (1997) and Greene (2003) among others to simplify the computation by using simulation methods.

Normal-Truncated Normal Models

Stevenson (1980) argues that the zero mean assumed in the Aigner *et al.* (1977) model is an unnecessary restriction and assumes that inefficiency follows the non-negative truncation distribution $u_i \sim N(\mu, \sigma_u^2)$; $f(u) = \left(\Phi(\mu/\sigma_u)\, \sigma_u \sqrt{2\pi}\right)^{-1} \exp\left(-(u-\mu)^2/2\sigma_u^2\right)$. Greene (1993) shows that the conditional expectation of technical inefficiency for the truncated normal distribution where μ is allowed to differ from zero in either direction is obtained by replacing ψ_i with $\psi_i^T = \lambda\varepsilon_i/\sigma + \mu/\lambda\sigma$ in equation (1.3).

Pitt and Lee (1981) extend the normal-truncated normal model to the panel data time-invariant inefficiency model. Battese and Coelli (1988) and Battese *et al.* (1989) provide the

extension of Jondrow *et al.* (1982) estimator to the panel data model

$$\widehat{u}_i = E(u_i \mid \varepsilon_{it}) = \varphi_i^T + \sigma_P \left[\frac{\phi\left(\varphi_i^T/\sigma_P\right)}{1 - \Phi\left(\varphi_i^T/\sigma_P\right)} \right]$$

where $\varphi_i^T = \left(\mu\sigma_v^2 - \sigma_u^2 \sum_{t=1}^{T} \varepsilon_{it}\right) / \left(\sigma_v^2 + T\sigma_u^2\right)$. By restricting μ to equal to zero $\mu = 0$, it returns to the estimator for the normal-half normal model in equation (1.4). Battese and Coelli (1988, 1992) derive the panel data extension to $E\left(\exp(-u_i) \mid \varepsilon_i\right)$ as

$$E\left(\exp(-u_i) \mid \varepsilon_{it}\right) = \left[\frac{\Phi\left[\left(\varphi_i^T/\sigma_P\right) - \sigma_P\right]}{\Phi\left(\varphi_i^T/\sigma_P\right)} \right] \exp\left(-\varphi_i^T + (\sigma_P/2)\right)$$

More recently, Almanidis, Qian, and Sickles (2014) specify inefficiency as a doubly truncated normal distribution. In addition to the zero lower bound, they specify an upper bound for inefficiency to exclude extremely inefficient producers. Their specification provides a closed form solution for $f_\varepsilon(\varepsilon_i)$ and the log-likelihood. Moreover, their specification results in non zero estimates of σ_u^2 in the presence of wrong skewness of the composed error term.

The truncated normal distribution can be used if producers are assumed to be inefficient since it has a mode at zero only when $\mu \leq 0$. Furthermore, it provides a way of introducing heterogeneity into the distribution of inefficiency by either including producer-specific characteristics in the mean, variance or both parameters of the inefficiency distribution.

Skew-Normal Models

Instead of focussing on the distribution of inefficiency, recent studies focus on the distribution of the composed error term, $f_\varepsilon(\varepsilon_i)$. For four random components models, the four random components $(\omega_i + \nu_{it} - u_{i0} - u_{it})$ can be treated as two terms since they can be written as the sum of the time-invariant terms $(\xi_i = \omega_i - u_{i0})$ and time-variant terms $(\varepsilon_{it} = \nu_{it} - u_{it})$. Time-invariant terms group together the producer-specific effects for unobserved heterogeneity ω_i and time-invariant inefficiency u_{i0} while time-variant terms group together the random errors ν_{it} and time-variant inefficiency u_{it}. These two terms are assumed to be given by the difference of a normal random variable and an independent left truncated at zero normal

random variable. Thus, each of the two terms has its own skew normal distribution rather than normal distribution[5].

The full unconditional log-likelihood function for this model based on the joint distribution of $(\varepsilon_{it}, \xi_i)$ is derived by Colombi *et al.* (2014). They estimate the four random components as $E\left(\exp(\omega_i) \mid y_i\right)$, and $E\left(\exp(t^{'}u_i) \mid y_i\right)$ where the first element of $E\left(\exp(t^{'}u_i) \mid y_i\right)$ is the conditional expected value of the time-invariant inefficiency u_{i0} for producer i. However, the computational complexity of the maximum likelihood estimator results from the $(T+1)$ dimensional multivariate normal integrals[6]. Tsionas and Kumbhakar (2014) note that the maximum likelihood estimator of Colombi *et al.* (2014) is computationally prohibitive when T is large. However, for time-invariant inefficiency models, the integral is one dimensional. For time-variant inefficiency models without a producer-specific component that accounts for heterogeneous technologies, it is a product of T one dimensional integral. Thus, the computational problem of the maximum likelihood estimator arises when estimating time-variant inefficiency with a producer-specific component that accounts for heterogeneous technologies or time-variant inefficiency along with time-invariant inefficiency. More recently, Fillipini and Greene (2016) exploit Butler and Moffitt (1982) formulation and propose a simplified density of y_i conditional on u_{i0} and ω_i that is the product over time of T univariate closed skew-normal densities.

For the importance of the distribution assumptions, if the interest is estimating producer-specific inefficiency, then distribution assumption on inefficiency is important. If the interest instead in comparing the ranking of producers, then using models with no distribution assumptions or following a recommendation by Ritter and Simar (1997) of using simple one-parameter distribution for inefficiency may be sufficient.

[5]See Gonzalez-Farias, Dominguez-Molina, and Gupta (2004), and Arellano-Valle and Azzalini (2006) for probability density function of the skew-normal distribution.

[6]See Genz and Bretz (2009) for a detailed review on computation methods of multi-normal integrals.

Confidence Intervals on Inefficiency

Since distributions imposed on v and u create distributions for $(u \mid \varepsilon)$ and $(\exp(-u) \mid \varepsilon)$ which can be used to construct confidence intervals on inefficiency, many studies show that it is possible to get confidence intervals for any of the technical inefficiency estimators. Hjalmarsson, Kumbhakar, and Heshmati (1996) develop confidence intervals for the Jondrow *et al.* (1982) estimator, and Bera and Sharma (1999) for the Battesse and Coelli (1988) estimator. Horrace and Schmidt (1996) derive upper and lower bounds on $(\exp(-u) \mid \varepsilon)$ based on lower and upper bounds of $(u \mid \varepsilon)$. However, Wheat *et al.* (2014) argue that the form of confidence intervals derived by Horrace and Schmidt (1996) is not minimum width since $f(u \mid \varepsilon)$ is truncated normal at zero and thus asymmetric. They propose a minimum width prediction interval for u given ε. Parmeter and Kumbhakar (2014) note that the narrower interval of Wheat *et al.* (2014) is preferable relative to Horrace and Schmidt (1996) intervals if the interest is to predict producer-specific inefficiency accurately.

1.4.2 Bayesian Estimation

Bayesian estimation of technical inefficiency is first introduced in the literature to the cross-section models by Van den Broeck *et al.* (1994) and Koop *et al.* (1994, 1995). Koop *et al.* (1997), Fernandez *et al.* (1997), and Osiewalski and Steel (1998) extend the use of Bayesian estimation to the panel data models. Koop *et al.* (1997) describe procedures for Bayesian estimation of both fixed and random effects models. Fernandez *et al.* (2000, 2002) extend the use of Bayesian estimation to the case where some of the outputs produced are undesirable to distinguish between technical and environmental inefficiency.

Bayesian approach treats the parameters of the model as random and conditional on the data instead of treating them as known or fixed and estimating them based on only information contained in the data. Instead of using the distribution of u conditional on ε; $E(u \mid \varepsilon)$, inference on technical inefficiency can be obtained using the conditional posterior distribution, $p(u \mid \theta_{-u}, Y)$ based on its marginal posterior, where θ_{-u} denotes all parameters

except the u. Bayesian inference, including point and interval estimation, evaluation of hypotheses, and prediction can be obtained from the posterior distribution.

The Prior Distribution

Initial beliefs and information that is not contained in the data through the likelihood function can be represented by the prior distributions of the parameters to be estimated including technical inefficiency. The prior distribution is presented in the form of a probability distribution, $p(\theta)$. It is classified as uninformative prior based on no prior knowledge that can be used for estimation or informative prior based on previous findings and theoretical predictions.

Specifying a uniform or flat prior distribution makes prior playing a small role in the estimation of the posterior distribution by relying on the data through the likelihood function. This is equivalent to specifying a prior distribution with a large variance that makes the prior distribution of the parameter values nearly flat. However, uninformative prior distribution is often improper. Fernandez *et al.* (1997) show that choosing an uninformative prior to the scale parameter leads to an improper prior.

Informative priors convey information and summarize existing knowledge about the parameters. Since normal distribution allows for negative numbers, it is not an appropriate prior distribution for technical inefficiency or scale parameters. Van den Broeck *et al.* (1994) find that the exponential distribution is more robust to prior assumptions than other distributions. Alvarez *et al.* (2014) compare an inverse Wishart, scaled inverse Wishart, and hierarchical inverse Wishart as possible priors for the scale parameter in multivariate models. They find that all priors work well except the inverse Wishart prior which is biased toward large values when the actual variance is small relative to the prior mean. In general, prior distribution for the scale parameter which plays an essential role in the estimation of technical inefficiency is crucial in any multivariate models and becomes more challenging as the dimension increases due to the quadratic growth in the number of parameters and the

need to force the matrix to remain non-negative definite. Informative priors can be used to impose restrictions from economic theory such as monotonicity and curvature restrictions — see, for example, Terrell (1996), or linear restrictions among the elements of the parameters — see, for example, Geweke (1993), or restrictions on inefficiency $u \geq 0$ — see, for example, Feng *et al.* (2018). Note that choosing a prior distribution that is conjugate to the likelihood leads to a posterior that has the same form as the prior.

The Posterior Distribution

Updating the prior information of the parameters can be done by combining the prior distribution $p(\theta)$ and the likelihood function $L(Y, \theta)$ in order to obtain the posterior distribution that is the basis of Bayesian estimation and defined by the Bayes Theorem as

$$p(\theta \mid Y) \propto L(Y, \theta) p(\theta)$$

where $p(\theta \mid Y)$ is the posterior distribution and it is proportion to the likelihood function times the prior. The posterior mean $E(\theta \mid Y)$ is the optimal Bayesian estimator of θ. However, when the model involves multidimensional parameters to be estimated, the posterior distribution is a joint posterior distribution. The marginal posterior distribution for a single given parameter θ_i is defined by integrating the joint posterior density of θ with respect to all elements of θ other than θ_i which may be too complicated for direct analytical integration or may not be analytically tractable. Implementing the Bayesian approach requires the use of an iterative Markov Chain Monte Carlo (MCMC) algorithm. Two common algorithms are the Gibbs sampling introduced by Geman and Geman (1984), and the Metropolis-Hastings algorithm introduced by Metropolis *et al.* (1953) and Hastings (1970).

When the joint posterior distribution is very complicated to work with, Gibbs sampling that uses draws from the conditional posterior distributions can be used to approximate joint and marginal distributions[7]. Thus, it is useful in cases in which the conditional posterior

[7]See, for example, Gelfand and Smith (1990), Casella and George (1992), Smith and Roberts (1993), Roberts and Smith (1994), Koop (1994), McCulloch and Rossi (1994), Dorfman (1997), and Geweke (1999) for more details on Gibbs sampling method.

distributions have relatively simple forms than the joint distribution, so it is possible to simulate from them. Gibbs sampling relies on the ability to partition θ as $\theta = (\theta_1, \theta_2, ..., \theta_p)$ where the θ_p may be multi-dimensional. The procedure proceeds as follows:

- Set initial values for the parameters θ^0 where the superscript 0 denotes the starting values and set $S = 0$.

- Generate random draws in sequence from the conditional distributions

  draw θ_1^{S+1} from $p(\theta_1 \mid Y, \theta_2^S, \theta_3^S, ..., \theta_p^S)$

  draw θ_2^{S+1} from $p(\theta_2 \mid Y, \theta_1^{S+1}, \theta_3^S, ..., \theta_p^S)$

  $\vdots$

  draw θ_p^{S+1} from $p(\theta_p \mid Y, \theta_1^{S+1}, \theta_2^{S+1}, ..., \theta_{p-1}^{S+1})$

  Set $S = S + 1$ and then iterate.

- Obtain series of draws, say $\left\{\theta^{(S)}, s = 1, ..., S\right\}$, where the superscript S denotes the S Gibbs iteration. For large enough number of iterations S, the samples or draws from the conditional distributions $\theta^{(S)}$ converge in distribution to the joint and marginal distributions of the parameters $p(\theta \mid Y)$ — see Casella and George (1992) for proof of convergence. In particular,

$$\left\{\theta^{(S)}, s = 1, ..., S\right\} \xrightarrow{D} p(\theta \mid Y)$$

- The first numbers of iterations that are required for the Gibbs sampler to converge are referred to as burn-in iterations. The sequence of last draws generated which converges in distribution to the full posterior can be considered as a sample from the joint posterior distribution of the parameters and used to provide inference about the parameters. For example, a histogram or kernel density of $\left\{\theta^{(S)}, s = 1, ..., S\right\}$ would provide the marginal posterior of θ. The

mean of these draws is an approximation to the posterior mean and provides a point estimate of these parameters.

To check whether the draws from the conditional posterior distributions are converged to the marginal posterior distribution, a test of convergence is proposed by Geweke (1992). If there is insufficient evidence for convergence, the number of draws should be increased.

While Gibbs sampling algorithm relies on conditional distributions, in some cases conditional posterior distributions are not from any known family distributions or are not available in closed form, so simulation from them is challenging. In such cases, the Metropolis-Hastings algorithm which is more general than the Gibbs sampling is an alternative MCMC algorithm that can be used to approximate the posterior distribution. The Metropolis-Hastings (MH) algorithm involves specifying a proposal density $q(\theta^* \mid \theta^S)$ where random draws generation is more comfortable than from the target density $p(\theta \mid Y)$. The procedure proceeds as follows:

- Set starting values for the parameters θ^0. These starting values can be OLS or MLE parameters estimates.

- Draw a number d from the standard uniform distribution $d \sim U(0,1)$ and draw a new or a candidate value of the parameters θ^* from the proposal density $q(\theta^* \mid \theta^S)$

- Compute the probability of accepting θ^*

$$r = \min\left(\frac{p(\theta^* \mid Y)/q(\theta^* \mid \theta^S)}{p(\theta^S \mid Y)/q(\theta^S \mid \theta^*)}, 1\right)$$

  where the numerator is the target density divided by the proposal density both evaluated at the new draw of the parameters, and the denominator is the same expression evaluated at the previous draw of the parameters. The most common choice of the proposal density is the multivariate normal $N(\mu, \sigma^2)$,

where the ratio simplifies to $p(\theta^* \mid Y)/p(\theta^S \mid Y)$ and σ^2 is to be chosen — see Chib and Greenberg (1995) for further details on choosing a proposal density for the Metropolis-Hastings algorithm. The covariance matrix σ^2 can be set as OLS or MLE estimate of the covariance matrix of the parameters.

- If $d < r$ then with probability r accept the proposal θ^* and set $\theta^{S+1} = \theta^*$; otherwise retain the old draw and set $\theta^{S+1} = \theta^S$. Thus, for this procedure to work, the proposal density $q(\theta^* \mid \theta^S)$ should be an excellent approximation to the target density $p(\theta^* \mid Y)$, otherwise, many candidates will be rejected.

  If the acceptance rate of this proposal is not satisfactory, it can be modified by multiplying its covariance matrix by a scaling factor h where a higher value for h means a lower acceptance rate. Roberts *et al.* (1997) show that if the proposal and target densities are normal densities, the optimal acceptance rate which minimizes the first order autocorrelations across the sample values of the algorithm approximately equal to 0.44 for one-dimensional models and 0.234 for higher dimensional models.

- Set $S = S + 1$ and repeat S times.

  The first MCMC iterations are discarded as a burn-in and estimates of the parameters can be obtained by simply averaging over the remaining iterations. Note that the Gibbs sampler can be seen as a special case of the MH algorithm where the candidate density $q(\theta^* \mid \theta^S)$ coincides with the target density and the acceptance probability assigned to every draw equals 1.

1.4.3 Theoretical Regularity

As required by microeconomic theory, the production technology has to satisfy the theoretical regularity conditions of monotonicity and curvature. Barnett (2002) notes that these theoretical regularity conditions can be violated unless imposed. However, Lau (1986) proves

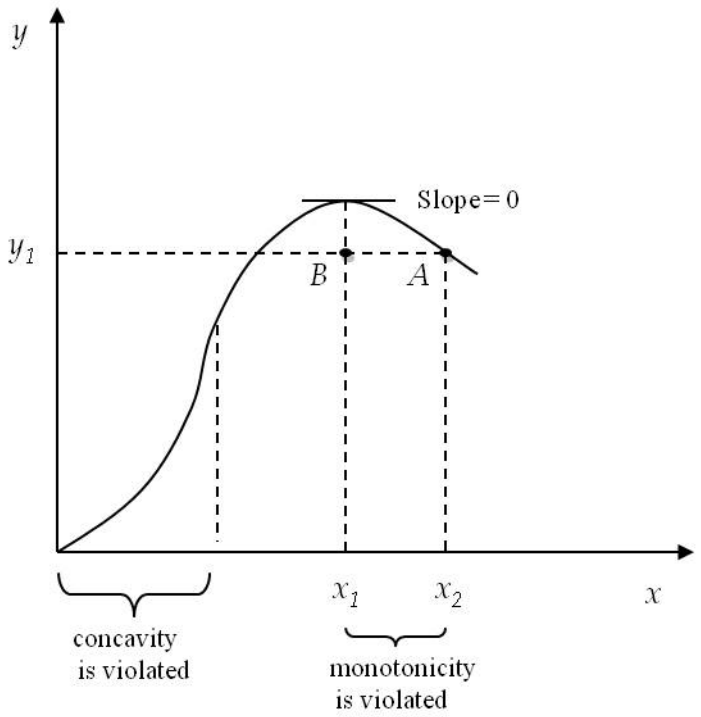

Figure 1.9: Violation of Monotonicity and Curvature Conditions

that imposing these regularity conditions globally can compromise the flexibility of flexible functional forms[8]. Ryan and Wales (2000) demonstrate that imposing curvature locally at a single point can be sufficient to achieve global regularity while preserving the flexibility of flexible functional forms. Terrell (1996) shows that imposing regularity conditions over small regions of data can preserve the flexibility of flexible functional forms. Wolff (2016) imposes regularity conditions locally, globally, and regionally on a flexible input demand system using the same data set. He finds that regional estimators outperform local and global estimators regarding the model fit to the sample data and preserving the flexibility of flexible functional forms.

Barnett (2002) and Barnett and Pasupathy (2003) note that monotonicity conditions are mostly disregarded in stochastic frontier estimation. However, monotonicity is particularly crucial regarding measuring technical inefficiency, where monotonicity implies that

[8]For example, imposing both monotonicity and curvature conditions globally on a translog functional form turns it into the Cobb-Douglas functional form.

additional units of inputs do not decrease outputs. Violation of monotonicity conditions can lead to an extreme misleading result such as the identification of a producer as being technically efficient when it is not. Consider an illustrative example of two producers A and B with non-monotone technology frontier — see Figure 1.9. Under this non-monotone technology frontier, producer A is efficient while producer B is inefficient since it operates beneath the production frontier. However, producer B produces the same output (y_1) as producer A using less input ($x_1 < x_2$). As a result, technical inefficiency measures of these two producers based on this non-monotone technology frontier reverses since in this case producer A is inefficient relative to producer B. Imposing monotonicity conditions prevents the production technology from exhibiting negative marginal productivities such as those implied by downward sloping production frontier; such as point A.

Curvature conditions are mainly required by microeconomic theory for the duality theorem to hold. Unless exploiting the duality theory and using the dual system specifications, measuring technical inefficiency does not require curvature conditions but monotonicity conditions to hold. In general, the regularity conditions can be checked as follows:

- Monotonicity can be checked by checking the first-order derivatives of the estimated production technology with respect to the inputs and outputs — see Table 1.1 for the properties of alternative distance functions.

- Concavity (convexity) can be checked using the unbordered Hessian matrix which is required to be negative (positive) semi-definite and can be checked by checking whether the Cholesky factors values are non-positive (negative) — see Lau (1978b).

If regularity conditions are not attained, the model can be estimated subject to imposed regularity conditions, thus treating these conditions as maintained hypotheses. This may require the use of Bayesian estimation to impose the inequality restrictions required to impose regularity conditions. Imposing regularity conditions can be either locally at one single

point of the domain of the regressor space or globally at the entire domain or regional on a connected subset of the domain. Regularity conditions can be imposed locally using techniques developed by Ryan and Wales (1998) or globally using techniques developed by Lau (1978b) and Diewert and Wales (1987) or regionally using techniques developed by Gallant and Golub (1984), Terrell (1996), Wolff *et al.* (2010), and Wolff (2016).

1.4.4 Econometric Regularity

Non-stationarity of the residuals of the production technology is an essential issue in estimating technical inefficiency because technical inefficiency measures are assessed from these estimated residuals. However, non-stationarity is mostly disregarded in inefficiency studies, mainly because standard methods for dealing with non-stationarity in linear models cannot be used with non-stationarity in nonlinear models. Nevertheless, ignoring the possibility of non-stationarity of the residuals can lead to misleading inefficiency results[9]. Barnett (1977) shows that consistency and asymptotic efficiency require stationarity assumptions as econometric regularity conditions.

Non-stationarity of the residuals of the production technology arises from non-stationarity of the dependent or explanatory variables or the omission of non-stationary variables. When all variables in the time series regression are integrated of order one; in short $I(1)$ in the context of Engle and Granger (1987), the production technology represents a cointegrating relationship, and the OLS estimator is super-consistent (is not only consistent, but it converges to the real value at a reasonable rate)[10]. If all variables are non-stationary, the production technology is a spurious relationship and the inefficiency measures are misleading. Additional complications arise if the production technology is not balanced where

[9]See Stock (1994) and Watson (1994) for a review of the econometric issues associated with non-stationary variables.

[10]While stationary time series is referred to as integrated of order zero and identified as $I(0)$, non-stationary time series that can be made stationary by taking the first difference, $[\triangle y_t = y_t - y_{t-1}]$, is referred to as integrated of order one, $I(1)$. In most cases, the order of integration of a time series to make it stationary is the lowest number of times it must be differenced.

different variables have different orders of integration or some variables are stationary, but other variables are non-stationary. Feng and Serletis (2008) find that input-output ratios, budget shares, and price variables are all integrated of order one, but the residuals are non-stationary.

Most common in the time series literature, serially correlated residuals are modeled by assuming a first order autoregressive $AR(1)$ process in the error terms as $\varepsilon_t = \rho\varepsilon_{t-1}+e_t$, where ρ is an unknown parameter and e_t is a non-autocorrelated random error term. The $AR(1)$ process is stationary when $|\rho| < 1$, and non-stationary random walk process when $\rho = 1$. Conventional unit root tests for stationarity test whether ρ is equal to one or significantly less than one. Alternatively, the augmented Dickey-Fuller (ADF) test proposed by Dickey and Fuller (1981), the non-parametric test of Phillips (1987), the numerical Bayesian test by Dorfman (1995), the test proposed by Harris and Tzavalis (1999) for dynamic panels, and the Fisher test by Maddala and Wu (1999) can be used to test for the unit root and the non-stationary on the residuals of the production technology.

If stationarity is not attained, cointegration techniques can be used to manage non-stationarity of the residuals[11]. If all variables are non-stationary, these variables must be cointegrated in levels given that inefficiency models are linear. Ng (1995) and Attfield (1997) argue that standard estimation techniques are inadequate for obtaining correctly estimated standard errors in cointegrated panels. Tsionas and Christopoulos (2001) apply panel cointegration techniques to estimate inefficiency using fully modified ordinary least squares (FM-OLS) proposed by Phillips and Hansen (1990), Phillips (1995), Phillips and Moon (1999), and Pedroni (2001) for cointegrated panels[12]. They find quantitatively important differences between their results and the results obtained by estimating inefficiency using standard estimation techniques. However, these cointegration techniques apply to linear models. Park

[11]When y_t and x_t are non-stationary $I(1)$ variables, then their difference, or any linear combination of them is $I(1)$ as well. In this case y_t and x_t are said to be cointegrated.

[12]Another estimation technique that can be used for cointegrated panels with higher order integrated systems is dynamic ordinary least square (DOLS) proposed by Stock and Watson (1993).

and Hahn (1999) consider the models linearized in the non-stationary variables. Lewbel and Ng (2005) propose a reformulation of the translog model that can be modified in a linear form to manage non-stationarity.

If finding cointegration between the $I(1)$ variables is failed, then a suitable solution is to convert the non-stationary series to stationary series by taking first differences if they are difference stationary or by de-trending or alternatively by including a trend variable in the model if they are trend stationary. However, Serletis and Shahmoradi (2007) argue that correction of serially correlated residuals increases the number of curvature violations and induces spurious violations of monotonicity.

Several attempts are proposed in the literature to develop estimation techniques for non-stationary models. Chang *et al.* (2001) extend earlier work by Phillips and Hansen (1990) and develop an estimator for nonlinear non-stationary models. Their estimator is consistent under reasonably general conditions, but the convergence rate critically depends on the type of the functional form. Han and Phillips (2010) propose a consistent GMM estimation method to the estimation of autoregressive roots near unity with both time series and panel data. However, their estimator has little bias even in very small samples. Therefore, with nonlinear non-stationary inefficiency models, further research is needed on the modification of the linear model cointegration techniques and developing the existing nonlinear cointegration techniques.

1.5 Estimation Issues

The estimates of technical inefficiency can be distorted by the inaccurate choice of functional form for the production technology, ignoring the possibility of heterogeneity and heteroskedasticity, and suffering from the endogeneity problem.

1.5.1 Functional Forms

The estimates of technical inefficiency can be distorted by the inaccurate choice of functional form for the production technology. Berger and Mester (1997) argue that a close fit of the actual data for the estimated production frontier is essential in estimating technical inefficiency because technical inefficiency is assessed as deviations from this production frontier. Giannakas, Tran, and Tzouvelekas (2003b) show that the inaccurate choice of functional form results in biased estimates of technical inefficiency, confidence intervals, and production elasticities, even asymptotically.

Regardless of not knowing the actual functional form, several properties of the production technology are known from economic theory, and several empirical techniques can be used to assess the ability of different functional forms to approximate the unknown underlying function. A functional form may be appropriate because of the correspondence theoretical properties, possibility and ease of application and empirical estimation, or a combination of these criteria. However, the reasons for choosing a particular functional form for the production technology are not explicitly stated in most studies.

Choosing a particular functional form for the production technology can be based on theoretical properties such as its shape of the isoquants, its separability, its flexibility, and its regular regions. Greene (1993) notes that the choice of functional form for the production technology has significant implications concerning the shape of the isoquants. Färe and Vardanyan (2016) compare the quadratic and translog functional forms regarding their ability to approximate convex frontiers of the input set and find that both functional forms provide a reliable approximation when a real frontier is assumed to be convex. Their findings validate the findings by Färe *et al.* (2010) and Chambers *et al.* (2013) that the translog functional form tends to yield convex frontier estimates even when the real frontier is concave. Therefore, the translog functional form that can approximate convex frontiers of the input set should behave relatively well when modeling input isoquants such as input distance

functions. On the other hand, assuming that the real production frontier is concave, Färe *et al.* (2010) and Chambers *et al.* (2013) simulation studies suggest that the concave frontier of the output set is better parameterized via a quadratic than with a translog functional form. Chambers *et al.* (2013) further find that the translog specification of a concave frontier can yield imprecise estimates of the technology. As a result, the quadratic functional form that can approximate concave frontiers of the output set should behave relatively well when modeling output isoquants such as standard or directional output distance functions. Separability properties are important for consistent aggregation. Thompson (1988) notes that both the translog and the quadratic are separable functional forms.

Choosing a particular functional form for specific studies can be based on a choice between functional forms that globally satisfy the theoretical regularity conditions for microeconomic theory and those that possess the flexibility. Flexible functional forms are functional forms that have a second order approximation property and are sufficiently flexible to ensure that the production elasticities and substitution elasticity are not restricted by the choice of the functional form[13]. However, the choice of functional forms for estimating technical inefficiency does not require flexible but regular functional forms that are consistent with economic theory. Greene (1980b) argues that flexible functional forms may suffer from multicollinearity due to a large number of parameters to be estimated and single equation estimates are likely imprecise.

Different distance functions have different application properties that influence the choice of functional forms such as homogeneity and translation properties — see table 1.1 for the properties of alternative distance functions. For instance, the choice of functional form to perform standard distance functions should be based on the satisfaction of the homogeneity property. Griffin *et al.* (1987) note that popular functional forms that are not

[13]For different definitions of the flexibility property, see, for example, Diewert (1971), Gallant (1981), and Barnett (1983). Diewert (1971) formalizes the notion of flexibility in functional forms by defining a second order approximation to an arbitrary function. Gallant (1981) proposes the Sobolev norm as a measure of global flexibility. See, for example, Griffin *et al.* (1987) for a comprehensive review of flexibility property.

linearly homogenous are logarithmic and augmented Fourier. Some functional forms can be linearly homogenous by incorporating the appropriate restrictions such as the quadratic, Cobb-Douglas, transcendental, constant elasticity of substitution, and the translog. On the other hand, the choice of functional form to perform directional distance functions should be based on the satisfaction of the translation property of directional distance functions. Chambers (1998) suggests two functional forms that satisfy the translation property; the logarithmic transcendental and the quadratic.

The Translog Functional Form

The translog functional form is introduced by Christensen *et al.* (1973). It is a generalization of the Cobb-Douglas and a locally flexible functional form providing a second order local approximation. Caves and Christensen (1980), Guilkey and Lovell (1980), Barnett and Lee (1985), and Barnett *et al.* (1985) argue that most locally flexible functional forms are not globally regular and have very small regions where theoretical regularity conditions are satisfied. The translog functional form can be defined over N inputs and M outputs as

$$\ln\left(D\left(x,y\right)\right)=\alpha_0+\sum_{n=1}^{N}\alpha_n\ln x_n+\sum_{m=1}^{M}\beta_m\ln y_m+\frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\alpha_{nn'}\ln x_n\ln x_{n'}$$
$$+\frac{1}{2}\sum_{m=1}^{M}\sum_{m'=1}^{M}\beta_{mm'}\ln y_m\ln y_{m'}+\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_{nm}\ln x_n\ln y_m$$

Symmetry requires $\alpha_{nn'}=\alpha_{n'n}$ $(n\neq n')$, and $\beta_{mm'}=\beta_{m'm}$ $(m\neq m')$. The translog functional form has many parameters to be estimated, for a total of $\left(k^2+3k+2\right)/2$ parameters including the intercept. It is linear in the parameters which can be restricted to satisfy the homogeneity property of standard and hyperbolic distance functions but cannot be restricted to satisfy the translation property of the directional distance functions.

The restrictions required for homogeneity of degree one in inputs are: $\sum_{n=1}^{N}\alpha_n=1$, $\sum_{n'=1}^{N}\alpha_{nn'}=0$, and $\sum_{n=1}^{N}\gamma_{nm}=0$. One way of imposing these restrictions on input distance function is to normalize the function by one of the inputs by setting the parameter of the homogeneity property $\lambda=1/x_N$ and obtaining $D_I\left(y,x/x_N\right)=D_I\left(y,x\right)/x_N$ — see, for

example, Sturm and Williams (2008). The restrictions required for homogeneity of degree one in outputs are: $\sum_{m=1}^{M}\beta_m = 1$, $\sum_{m'=1}^{M}\beta_{mm'} = 0$, and $\sum_{m=1}^{M}\gamma_{nm} = 0$. One way of imposing these restrictions on output distance function is to normalize the function by one of the outputs by setting $\lambda = 1/y_M$ and obtaining $D_O\left(x, y/y_M\right) = D_O\left(x, y\right)/y_M$ — see, for example, O'Donnell and Coelli (2005). The restrictions required for almost homogeneity of degrees -1, 1, and 1 are: $\sum_{m=1}^{M}\beta_m - \sum_{n=1}^{N}\alpha_n = 1$, $\sum_{m=1}^{M}\gamma_{nm} - \sum_{n'=1}^{N}\alpha_{nn'} = 0$, and $\sum_{m'=1}^{M}\beta_{mm'} - \sum_{n=1}^{N}\gamma_{nm} = 0$. One way of imposing these restrictions on hyperbolic distance function is to normalize the function by one of the inputs by setting $\lambda = 1/x_N$ and obtaining $D_H\left(x/x_N, yx_N\right) = D_H\left(x, y\right)/x_N$ or one of the outputs by setting $\lambda = 1/y_M$ and obtaining $D_H\left(xy_M, y/y_M\right) = D_H\left(x, y\right)/y_M$ — see, for example, Cuesta and Zofio (2005). The first and second order partial derivatives can be presented as

$$\frac{\partial \ln\left(D\left(x,y\right)\right)}{\partial \ln x_n} = \alpha_n + \sum_{n'=1}^{N}\alpha_{nn'}\ln x_{n'} + \sum_{m=1}^{M}\gamma_{nm}\ln y_m, \qquad \frac{\partial^2 \ln\left(D\left(x,y\right)\right)}{\partial \ln x_n \ln x_{n'}} = \alpha_{nn'}$$

$$\frac{\partial \ln\left(D\left(x,y\right)\right)}{\partial \ln y_m} = \beta_m + \sum_{m'=1}^{M}\beta_{mm'}\ln y_{m'} + \sum_{n=1}^{N}\gamma_{nm}\ln x_n, \qquad \frac{\partial^2 \ln\left(D\left(x,y\right)\right)}{\partial \ln y_m \ln y_{m'}} = \beta_{mm'}$$

However, the translog functional form is not monotonic or globally convex, as the Cobb-Douglas functional form. Caves and Christensen (1980) note that the translog functional form has satisfactory local properties when the technology is nearly homothetic, and the substitution between factors of production is high. Guilkey *et al.* (1983) show that the translog functional form is globally regular if and only if the technology is Cobb–Douglas. Färe and Vardanyan (2016) find that the translog functional form often violates theoretical regularity conditions and requires imposing the appropriate regularity conditions that significantly compromise its flexibility. Their results are consistent with the simulation results by Wales (1977) and Guilkey *et al.* (1983) that compare the performance of various functional forms including the translog.

The Quadratic Functional Form

Chambers (1998) suggests the quadratic functional form for the directional distance functions since its parameters can be restricted to satisfy the translation property. The quadratic functional form is introduced by Lau (1978a) and can be presented as

$$D(x,y) = \alpha_0 + \sum_{n=1}^{N} \alpha_n x_n + \sum_{m=1}^{M} \beta_m y_m + \frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N} \alpha_{nn'} x_n x_{n'} + \frac{1}{2}\sum_{m=1}^{M}\sum_{m'=1}^{M} \beta_{mm'} y_m y_{m'} + \sum_{n=1}^{N}\sum_{m=1}^{M} \gamma_{nm} x_n y_m \tag{1.6}$$

Symmetry requires $\alpha_{nn'} = \alpha_{n'n}$ $(n \neq n')$, and $\beta_{mm'} = \beta_{m'm}$ $(m \neq m')$. The quadratic functional form has many parameters to be estimated, for a total of $(k^2 + 3k + 2)/2$ parameters including the intercept. It is linear in the parameters which can be restricted to satisfy the translation property of the directional distance functions. The set of linear restrictions required for the translation property are: $\sum_{m=1}^{M} \beta_m g_{y_m} - \sum_{n=1}^{N} \alpha_n g_{x_n} = -1$; $\sum_{m'=1}^{M} \beta_{mm'} g_{y_m} - \sum_{n=1}^{N} \gamma_{nm} g_{x_n} = 0$; and $\sum_{m=1}^{M} \gamma_{nm} g_{y_m} - \sum_{n'=1}^{N} \alpha_{nn'} g_{x_n} = 0$. One way of imposing these restrictions is imposing them directly in equation (1.6) and obtaining a restricted version — see, for example, Atkinson and Tsionas (2016). Alternatively, these restrictions can be imposed by setting the parameter of the translation property α equal to an arbitrarily chosen input $\alpha = x_N$ or the negative of an arbitrarily chosen output $\alpha = -y_M$ and normalizing the corresponding direction vector $g_{x_N} = 1$ or $g_{y_M} = 1$. In the case of choosing $\alpha = x_N$, $\vec{D}_T(\widetilde{x} - x_N \widetilde{g}_x, y + x_N g_y; g_x, g_y) = \vec{D}_T(x, y; g_x, g_y) - x_N$, where $\widetilde{x} = (x_1, ..., x_{N-1})$ and $\widetilde{g}_x = (g_{x_1}, ..., g_{x_{N-1}})$ and the input x_N disapears from $\vec{D}_T(\widetilde{x} - x_N \widetilde{g}_x, y + x_N g_y; g_x, g_y)$ because $x_N - x_N(1) = 0$. In the case of choosing $\alpha = -y_M$, $\vec{D}_T(x + y_M g_x, \widetilde{y} - y_M \widetilde{g}_y; g_x, g_y) = \vec{D}_T(x, y; g_x, g_y) + y_M$, where $\widetilde{y} = (y_1, ..., y_{M-1})$ and $\widetilde{g}_y = (g_{y_1}, ..., g_{y_{M-1}})$ and the output y_M disapears from $\vec{D}_T(x + y_M g_x, \widetilde{y} - y_M \widetilde{g}_y; g_x, g_y)$ because $y_M - y_M(1) = 0$ — see, for example, Malikov *et al.* (2016). The first and second order partial derivatives can be presented as

$$\frac{\partial D(x,y)}{\partial x_n} = \alpha_n + \sum_{n'=1}^{N} \alpha_{nn'} x_{n'} + \sum_{m=1}^{M} \gamma_{nm} y_m, \qquad \frac{\partial^2 D(x,y)}{\partial x_n x_{n'}} = \alpha_{nn'}$$

$$\frac{\partial D\left(x,y\right)}{\partial y_m}=\beta_m+\sum_{m'=1}^{M}\beta_{mm'}y_{m'}+\sum_{n=1}^{N}\gamma_{nm}x_n,\qquad\frac{\partial^2 D\left(x,y\right)}{\partial y_m y_{m'}}=\beta_{mm'}$$

Thompson (1988) notes that the quadratic functional form is capable of satisfying global curvature restrictions without additional constraints in estimation. This is validated by Färe and Vardanyan (2016) results that the quadratic functional form satisfies global regularity without curvature restrictions while preserving its flexibility. A simulation study by Chambers *et al.* (2013) suggests that the quadratic functional form outperforms the translog in large samples with a relatively large amount of curvature. Diewert (2008) notes that curvature restrictions can be globally imposed on the quadratic functional form without losing flexibility. However, monotonicity cannot be imposed simultaneously with curvature without destroying the flexibility of the functional form. As noted by Barnett (2002), the imposition of global curvature on the quadratic functional form may induce spurious violations of monotonicity.

The Logarithmic-Transcendental Functional Form

Chambers (1998) suggests the logarithmic-transcendental functional form for the directional distance functions since it automatically satisfies the translation property. It is a flexible functional form providing a second order approximation. The logarithmic-transcendental or the transcendental-exponential functional form can be presented as

$$\begin{aligned}\exp\left(D\left(x,y\right)\right)=\alpha_0&+\frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\alpha_{nn'}\exp\left(\frac{x_n}{2}\right)\exp\left(\frac{x_{n'}}{2}\right)\\&+\frac{1}{2}\sum_{m=1}^{M}\sum_{m'=1}^{M}\beta_{mm'}\exp\left(-\frac{y_m}{2}\right)\exp\left(-\frac{y_{m'}}{2}\right)\\&+\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_{nm}\exp\left(\frac{x_n}{2}\right)\exp\left(-\frac{y_m}{2}\right)\end{aligned}$$

The logarithmic-transcendental functional form has fewer parameters to be estimated than the translog and the quadratic functional forms, for a total of $\left(k^2+k+2\right)/2$ parameters including the intercept. Symmetry requires $\alpha_{nn'}=\alpha_{n'n}$ $(n\neq n')$, and $\beta_{mm'}=\beta_{m'm}$ $(m\neq$

m'). The first order partial derivatives can be presented as

$$\frac{\partial \exp\left(D\left(x,y\right)\right)}{\partial \exp\left(\frac{x_n}{2}\right)} = \sum_{n'=1}^{N} \alpha_{nn'} \exp\left(\frac{x_{n'}}{2}\right) + \sum_{m=1}^{M} \gamma_{nm} \exp\left(-\frac{y_m}{2}\right)$$

$$\frac{\partial \exp\left(D\left(x,y\right)\right)}{\partial \exp\left(-\frac{y_m}{2}\right)} = \sum_{m'=1}^{M} \beta_{mm'} \exp\left(-\frac{y_{m'}}{2}\right) + \sum_{n=1}^{N} \gamma_{nm} \exp\left(\frac{x_n}{2}\right)$$

The second order partial derivatives can be presented as

$$\frac{\partial^2 \exp\left(D\left(x,y\right)\right)}{\partial \exp\left(\frac{x_n}{2}\right) \exp\left(\frac{x_{n'}}{2}\right)} = \alpha_{nn'}$$

$$\frac{\partial^2 \exp\left(D\left(x,y\right)\right)}{\partial \exp\left(-\frac{y_m}{2}\right) \exp\left(-\frac{y_{m'}}{2}\right)} = \beta_{mm'}$$

Empirical techniques can also be used to assess the ability of different functional forms to approximate the unknown underlying function. Several techniques are proposed in the literature; Monte Carlo simulations, parametric modeling, and constructive techniques. Monte Carlo simulations compare the approximation capabilities of different functional forms to the underlying technology. For applications of this technique, see, for example, Guilkey and Lovell (1980), Giannakas, Tran, and Tzouvelekas (2003b), Färe *et al.* (2010), Chambers *et al.* (2013), and Färe and Vardanyan (2016). Parametric modeling assesses the plausibility of different functional forms in fitting actual data. For applications of this technique, see, for example, Griffin *et al.* (1987), Giannakas, Tran, and Tzouvelekas (2003a), and Feng and Serletis (2008). The main issue with parametric modeling is that the actual functional form for the production technology is unknown. In this case, evaluating the performance of different functional forms in fitting actual data is beneficial if the interest is examining the data itself, not the functional forms. As noted by Giannakas, Tran, and Tzouvelekas (2003a), the appropriate functional form, in this case, is case-specific. The constructive technique provides a means of determining the preferable functional forms by deriving and graphically displaying their regular regions. For applications of this technique, see, for example, Caves and Christensen (1980), and Barnett *et al.* (1985, 1987).

1.5.2 Heterogeneity Issue

The appropriate choice of functional form for the production technology is not sufficient without accommodating heterogeneity in the production model. Heterogeneity can be in the technology by shifting the production frontier or in the inefficiency term by shifting the inefficiency distribution or both.

To account for heterogeneous technologies, producer-specific characteristics can be included directly in the functional form of the technology. Since inefficiency heterogeneity changes the location and scale parameters of the inefficiency distribution, heterogeneous inefficiency can be considered by including producer-specific characteristics either in the inefficiency term or the parameters of the inefficiency distribution. Greene (2002) argues that producer-specific characteristics are the primary sources of heterogeneity that are largely ignored in the inefficiency literature.

Heterogeneity in the Production Technology

Ignoring the possibility of heterogeneous technologies that may exist among producers can lead to wrong conclusions concerning inefficiency measures — see, for example, Casu and Molyneux (2003), and Bos and Schmiedel (2007). Brown and Glennon (2000) note that assuming a common production technology for all producers is a very restrictive assumption. According to Tsionas (2002), assuming a common technology for all producers may rank a producer as inefficient although it may employ different technology than other producers and fully utilize its own technology. Mester (1997), Greene (2005b, 2008), and Caiazza *et al.* (2016) note that heterogeneity causes biased estimates obtained from the stochastic frontier approach.

To illustrate the importance of accommodating heterogeneity in the production frontier when estimating technical inefficiency, consider an example of two producers A and B with production frontiers labeled with A and B, respectively — see Figure 1.10. Assuming a common frontier C for those two producers, the directional measure of overall technical inef-

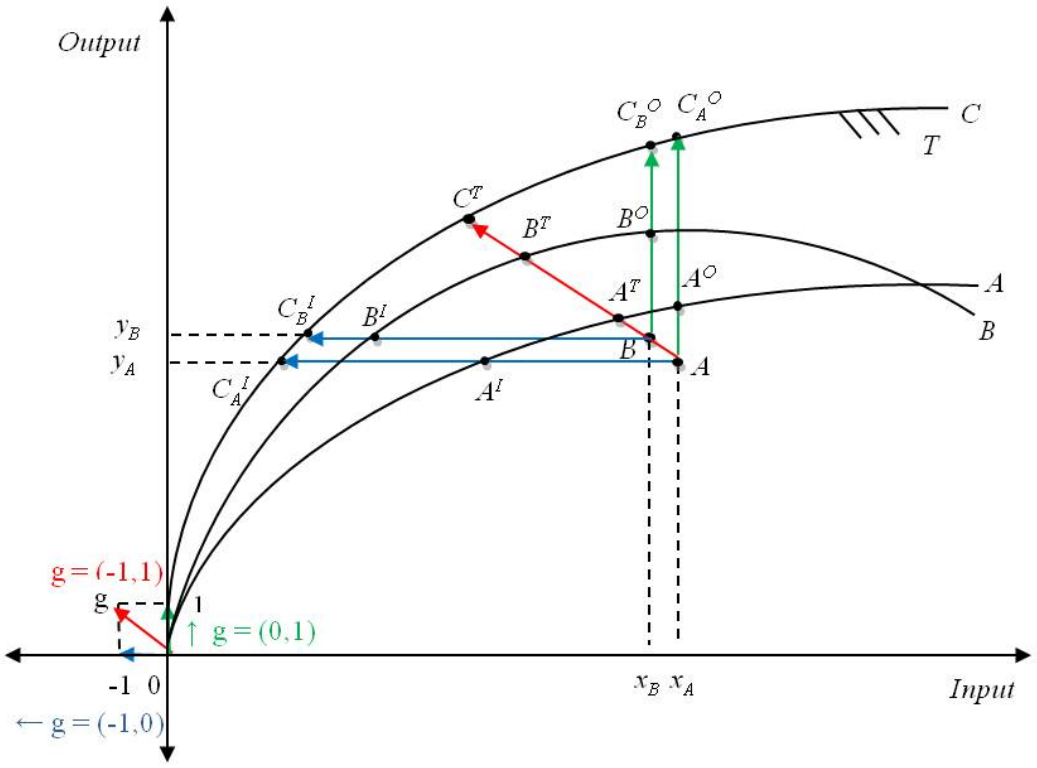

Figure 1.10: Heterogeneous Technologies and Technical Inefficiency

ficiency is given by $TI_T^{pooled}(x_A, y_A) = \left\|AC^T\right\| / \|0g\|$ for producer A and $TI_T^{pooled}(x_B, y_B) = \left\|BC^T\right\| / \|0g\|$ for producer B. Under this assumption, producer A is less efficient than producer B since $\left\|AC^T\right\| > \left\|BC^T\right\|$. However, considering each producer operating on its own frontier, then $TI_T^{own}(x_A, y_A) = \left\|AA^T\right\| / \|0g\|$ and $TI_T^{own}(x_B, y_B) = \left\|BB^T\right\| / \|0g\|$. Consequently, the efficiency ranking of these two producers reverses since in this case producer B is less efficient than producer A because of $\left\|BB^T\right\| > \left\|AA^T\right\|$. Similarly, the directional measure of output technical inefficiency is given by $TI_O^{pooled}(x_A, y_A) = \left\|AC_A^O\right\|$ for producer A and $TI_O^{pooled}(x_B, y_B) = \left\|BC_B^O\right\|$ for producer B assuming a common frontier C. Under this assumption, producer A is less efficient than producer B. However, considering each producer operating on its own frontier, then $TI_O^{own}(x_A, y_A) = \left\|AA^O\right\|$ and $TI_O^{own}(x_B, y_B) = \left\|BB^O\right\|$. Consequently, the efficiency ranking of these two producers reverses since in this case producer B is less efficient than producer A because of $\left\|BB^O\right\| > \left\|AA^O\right\|$. The directional measure of input technical inefficiency is given by $TI_I^{pooled}(x_A, y_A) = \left\|AC_A^I\right\|$ for producer A

and $TI_I^{pooled}(x_B, y_B) = \left\|BC_B^I\right\|$ for producer B assuming a common frontier C. Under this assumption, producer A is less efficient than producer B. However, considering each producer operating on its own frontier, then $TI_I^{own}(x_A, y_A) = \left\|AA^I\right\|$ and $TI_I^{own}(x_B, y_B) = \left\|BB^I\right\|$. Consequently, the efficiency ranking of these two producers reverses since in this case producer B is less efficient than producer A because of $\left\|BB^I\right\| > \left\|AA^I\right\|$.

There are different approaches to account for heterogeneous technologies. One approach is to introduce a producer-specific intercept to the model — see, for example, Greene (2005a, b). Cornwell *et al.* (1990) and Swamy and Tavlas (1995) assume that both the intercept and the slope parameters are random. Akhavein *et al.* (1997), Tsionas (2002), and Feng *et al.* (2018) assume a stochastic frontier model with random coefficients. Alternatively, varying coefficient models in which the coefficients are allowed to vary as functions of other variables can be used — see, for example, Hastie and Tibshirani (1993) and Tran (2014).

Another approach is to split producers exogenously into groups based on size, ownership, organizational structure or geographic regions, and estimate a model for each group — see, for example, Mester (1996), and Altunbas *et al.* (2001) among others. However, estimating a model for each group has a shortcoming of not using the information provided by the producers in the other groups — see, for example, Greene (1993, 2004b), Orea and Kumbhakar (2004), and Parmeter and Kumbhakar (2014).

Alternatively, the threshold models split producers into technology groups based on the threshold variable value, and the model of each group is estimated using the information provided by the producers in the other groups — see, for example, Hansen (1999, 2000) and Yélou, Larue, and Tran (2010) for a single threshold model and Almanidis (2013) for multiple thresholds model. However, Almanidis (2013) notes that the joint estimation of the threshold parameters requires a grid search over an enormous number of points which increases with the number of breakpoints. The solution is to use a sequential estimation of the threshold parameters. However, this method yields asymptotically efficient estimates only of the last

threshold parameter in the process. Bai (1997) suggests a refinement estimation of the threshold parameters, which amounts to re-estimating the threshold parameters backward, each time holding the estimates of the previous thresholds fixed. The refinement estimator is shown to be asymptotically efficient. More recently, Tsionas, Tran, and Michaelides (2017) propose a threshold stochastic frontier model that can accommodate multiple thresholds.

Heterogeneity in the Inefficiency

Ignoring the existence of heterogeneity in the inefficiency term can lead to inaccurate measures of the inefficiency since heterogeneity which is not captured by producer-specific characteristics is wrongly attributed to inefficiency. Heterogeneity in the inefficiency can be captured by including producer-specific characteristics in the mean, variance or both parameters of the inefficiency distribution — see section 1.3 for more details on these models.

To summarize, exogenous factors that are not under the control of the producer and affect the production are supposed to capture heterogeneous technologies and heterogeneity can be specified directly in the production frontier. If the exogenous factors are under the control of the producer, heterogeneity in the inefficiency can be captured by including producer-specific characteristics in the mean, variance or both parameters of the inefficiency distribution.

1.5.3 Heteroscedasticity Issue

Several inefficiency models are based on the assumptions that the random errors v and the inefficiency term u are homoscedastic or equivalently both σ_v^2 and σ_u^2 are constants. However, the random errors v and the inefficiency term u may be heteroscedastic. Heteroskedasticity refers to the models in which σ_v^2 and σ_u^2 are not constants but functions of explanatory variables that reflect producer-specific characteristics.

Kumbhakar and Lovell (2000) and Wang and Schmidt (2002) conclude that ignoring the heteroscedasticity of v results in consistent estimates of the parameters of the production technology but biased estimates of the intercept and inefficiency while ignoring the het-

eroscedasticity of u causes biased estimates of the parameters of the production technology as well as the estimates of inefficiency. To account for heteroscedasticity, the scale parameter of the distribution of the random error and inefficiency can be modeled as functions of explanatory variables that reflect producer-specific characteristics — see section 1.3 for more details on these models.

1.5.4 Endogeneity Issue

A potential issue when estimating technical inefficiency using distance functions is that inputs and outputs may be endogenous, meaning that they are correlated with the random errors or inefficiency or both and leading to biased and inconsistent estimates of the parameters of the production technology and the associated measures of inefficiency — see, for example, Atkinson and Primont (2002), Atkinson *et al.* (2003), and O'Donnell (2014).

The literature considers two approaches to deal with this issue; one approach relies on using instrumental variable estimation, and the other relies on employing a system of equations approach. The use of instruments involves the selection of instrumental variables that are uncorrelated with the composed errors term and the estimation of the stochastic frontier model with exogenous and endogenous variables and the reduced form equation for the endogenous variables which includes the exogenous variables and the instruments. Tran and Tsionas (2013) propose a simple generalized method of moments (GMM) procedure for estimating stochastic frontier models in the presence of endogenous variables. In comparing the use of instruments using GMM and Bayesian estimation, Assaf *et al.* (2013) find that Bayesian estimation provides more precise estimates compared to GMM. Tran and Tsionas (2015) consider an alternative procedure that does not involve the use of instruments and is based on Copula function to directly model and capture the dependency of the endogenous variables and the composed error term.

Alternatively, the endogeneity issue can be managed by using a system of equations approach. In order to meet the rank condition for the identification of the system, a total

number of potentially endogenous variables are required in the system as independent equations including the production technology. The choice of the system can be based on the behavioral assumptions of the producers, duality theory, and the endogeneity of inputs and outputs. As discussed earlier, the cost function is proved to be dual to the IDF and DIDF by Luenberger (1992), Färe and Primont (1995), and Chambers *et al.* (1996). While the revenue function is proved to be dual to the ODF and DODF by Färe *et al.* (1993) and Färe and Grosskopf (2000), the profit function is proved to be dual to the DTDF by Chambers *et al.* (1998) — see section 1.2 for more details.

If only inputs (outputs) are endogenous, choosing the first order conditions of cost minimization (revenue maximization) together with the IDF (ODF) or DIDF (DODF) is preferable. See Tsionas *et al.* (2015) as an example of a system based on the IDF and the first-order conditions for cost minimization. Coelli (2000) shows that OLS provides consistent estimates of an IDF (ODF) under the assumption of cost-minimizing (revenue-maximizing) behavior when estimating distance functions in a system of equations and indicates that the instrumental variables may not be required. If however both inputs and outputs are endogenous, choosing the first order conditions of profit maximization together with the HDF or DTDF is preferable. See Atkinson and Tsionas (2016) as an example of a system based on the DTDF and the assumption of profit-maximizing behavior. However, Malikov *et al.* (2016) consider the DTDF and the first order conditions for cost minimization leaving the endogeneity of outputs unaddressed. Feng *et al.* (2018) use the DODF and the first order conditions for profit maximization and thus treating inputs as fixed in the DODF and endogenous in the profit maximization.

The system approach is not only considered to be a procedure to manage the endogeneity issue but has several advantages. Berndt and Christensen (1973) argue that the use of the system approach overcome the multicollinearity issue that single equation may suffer due to a large number of parameters to be estimated. While evaluating different functional

forms using a single equation and a system of equations, Guilkey *et al.* (1983) find that the functional form considered in the system of equation outperforms the single equation concerning its bias. Furthermore, the system approach includes a considerable amount of information through the first order conditions and leads to more meaningful results.

1.6 Conclusion

This paper provides an up-to-date review that focusses on research methods, including different approaches to measuring technical inefficiency using distance functions, the development of modeling technical inefficiency in the stochastic frontier framework, and the most common econometric estimation techniques. It also provides a useful guide on when these methods can be used and how to implement them.

Regarding measuring technical inefficiency using distance functions, I discuss and evaluate the radial measure given by the standard distance functions, the hyperbolic measure given by the hyperbolic distance function, and the directional measure given by the directional distance functions. While the radial measure may produce high inefficiency measures even when the observed input-output vector is very close to the frontier, the hyperbolic measure is not always easy to implement due to the non-linear optimization involved. The directional measure is technology-oriented and provides greater flexibility by contracting inputs and expanding outputs simultaneously using an exogenous or an endogenous directional vector to reach the efficient frontier. A recent development is treating the directional vector as a parameter to be estimated with the other parameters of the model. However, further research is needed comparing the alternative choices of the endogenous directional vector.

Since theoretical and econometric regularity conditions are still disregarded in most efficiency studies, the paper addresses the importance of attaining the theoretical regularity applied by neoclassical microeconomic theory when violated, as well as the econometric regularity when variables are non-stationary. Without regularity, inefficiency results are ex-

tremely misleading. If regularity conditions are not attained, the model can be estimated subject to imposed regularity conditions, which may require the use of Bayesian estimation. If stationarity is not attained, cointegration techniques can be used to manage non-stationarity of the residuals. However, with nonlinear non-stationary inefficiency models, further research is needed on the modification of the linear model cointegration techniques and developing the existing nonlinear cointegration techniques.

Regarding estimation issues, the estimates of technical inefficiency can be distorted by the inaccurate choice of functional form for the production technology, ignoring the possibility of heterogeneity and heteroskedasticity, and suffering from the endogeneity problem. It is important in future applications to estimate technical inefficiency while managing these issues by one of the different procedures discussed in the paper.

The inaccurate choice of functional form results in biased estimates of technical inefficiency. This paper discusses several selection criteria for choosing a particular functional form for the production technology based on theoretical properties such as its shape of the isoquants, its separability, its flexibility and its regular regions, and application properties such as homogeneity and translation properties. It also addresses empirical techniques that can be used to assess the ability of different functional forms to approximate the unknown underlying function.

The appropriate choice of functional form is not sufficient without accommodating heterogeneous technologies that may exist among producers or heterogeneity in the inefficiency term. Ignoring heterogeneity can lead to wrong conclusions concerning inefficiency measures since heterogeneity which is not captured by producer-specific characteristics is wrongly attributed to inefficiency. This paper addresses the importance of accommodating heterogeneity and discusses different approaches to account for both heterogeneous technologies and heterogeneity in the inefficiency term while estimating technical inefficiency. In general, exogenous factors that are not under the control of the producer and affect the produc-

tion are supposed to capture heterogeneous technologies and heterogeneity can be specified directly in the production frontier. If the exogenous factors are under the control of the producer, heterogeneity in the inefficiency can be captured by including producer-specific characteristics in the mean, variance or both parameters of the inefficiency distribution. Including producer-specific characteristics in the scale parameter of the inefficiency distribution accounts for heteroscedasticity as well.

Another potential issue when estimating technical inefficiency using distance functions is that inputs and outputs may be endogenous leading to biased and inconsistent estimates of the parameters of the production technology and the associated measures of inefficiency. This paper discusses different approaches to deal with this issue, mainly the instrumental variable approach and the system of equations approach.

Table 1.1: A Summary of the Important Properties of Alternative Distance Functions

Property / Function	Feasibility	Monotonicity	Homogeneity Translation	Inputs	Outputs	Technical Inefficiency	Relationships
IDF $D_I(y,x)$ $=\max_{\vartheta_I}$: $\left(\frac{x}{\vartheta_I},y\right)$ $\in T$	$D_I(y,x)\geq 1$ iff $(x,y)\in T$ if $(x,y)$is on the frontier of T, then $D_I(y,x)=1$	$\nabla_x D_I(.)\geq 0$ $\nabla_y D_I(.)\leq 0$	homogeneity $D_I(y,\lambda x)$ $=\lambda D_I(y,x)$ $\lambda>0$	concave	quasi- concave	$TI_I(y,x)$ $=1-TE_I(y,x)$ $=1-\frac{1}{D_I(y,x)}$	under CRS; $D_I(y,x)=$ $1/D_o(x,y)$ duality; $1/D_I(y,x)\geq$ $C(y,w)/wx$
ODF $D_o(x,y)$ $=\min_{\vartheta_o}$: $\left(x,\frac{y}{\vartheta_O}\right)$ $\in T$	$D_o(x,y)\leq 1$ iff $(x,y)\in T$ if $(x,y)$is on the frontier of T, then $D_o(x,y)=1$	$\nabla_x D_o(.)\leq 0$ $\nabla_y D_o(.)\geq 0$	homogeneity $D_o(x,\lambda y)$ $=\lambda D_o(x,y)$ $\lambda>0$	quasi- convex	convex	$TI_o(x,y)$ $=TE_o(x,y)-1$ $=\frac{1}{D_o(x,y)}-1$	under CRS; $D_o(x,y)=$ $1/D_I(y,x)$ duality; $1/D_o(x,y)\leq$ $R(x,p)/py$
HDF $D_H(x,y)$ $=\min_{\vartheta_H}$: $\left(\vartheta_H x,\frac{y}{\vartheta_H}\right)$ $\in T$	$D_H(x,y)\leq 1$ iff $(x,y)\in T$ if $(x,y)$is on the frontier of T, then $D_H(x,y)=1$	$\nabla_x D_H(.)\leq 0$ $\nabla_y D_H(.)\geq 0$	almost homogeneous $D_H(\lambda^{-1}x,\lambda y)$ $=\lambda D_H(x,y)$ under CRS; $D_H(\lambda x,\lambda y)$ $=D_H(x,y)$	convex	convex	$TI_H(x,y)$ $=TE_H(x,y)-1$ $=\frac{1}{D_H(x,y)}-1$	under CRS; $D_H(x,y)$ $=\frac{1}{\sqrt{D_I(y,x)}}$ $=\sqrt{D_O(x,y)}$ duality; $[1/D_H(x,y)]^2$ $\geq py/wx$
DTDF $\vec{D}_T(.)$ $=\max_{\theta_T}$: $(x-\theta_T g_x,$ $y+\theta_T g_y)$ $\in T$	$\vec{D}_T(.)\geq 0$ iff $(x,y)\in T$ if $(x,y)$is on the frontier of T, then $\vec{D}_T(.)=0$	$\nabla_x \vec{D}_T(.)\geq 0$ $\nabla_y \vec{D}_T(.)\leq 0$	translation $\vec{D}_T(x-\alpha g_x,$ $y+\alpha g_y;g)=$ $\vec{D}_T(x,y;g)-\alpha$ homogeneity $\vec{D}_T(x,y;\lambda g)=$ $\lambda^{-1}\vec{D}_T(x,y;g)$	concave	concave	$TI_T(x,y)$ $=\vec{D}_T(.)$	$\vec{D}_T(x,y;0,g_y)$ $=\vec{D}_o(x,y;g_y)$ $\vec{D}_T(x,y;g_x,0)$ $=\vec{D}_I(y,x;g_x)$ duality; $\vec{D}_T(x,y;g)\leq$ $\frac{\pi(p,w)-(py-wx)}{pg_y+wg_x}$
DIDF $\vec{D}_I(y,x;g_x)$ $=\max_{\theta_I}$: $(y,x-\theta_I g_x)$ $\in T$	$\vec{D}_I(.)\geq 0$ iff $(x,y)\in T$ if $(x,y)$is on the frontier of T, then $\vec{D}_I(.)=0$	$\nabla_x \vec{D}_I(.)\geq 0$ $\nabla_y \vec{D}_I(.)\leq 0$	translation $\vec{D}_I(y,x-\alpha g_x)=$ $\vec{D}_I(y,x;g_x)-\alpha$ homogeneity $\vec{D}_I(y,x;\lambda g_x)=$ $\lambda^{-1}\vec{D}_I(y,x;g_x)$	concave	quasi- concave	$TI_I(y,x)$ $=\vec{D}_I(.)$	$\vec{D}_I(y,x;-x)=$ $\vec{D}_T(x,y;-x,0)$ $=1-1/D_I(y,x)$ duality; $\vec{D}_I(y,x;g_x)\leq$ $\frac{wx-C(y,w)}{wg_x}$
DODF $\vec{D}_o(x,y;g_y)$ $=\max_{\theta o}$: $(x,y+\alpha g_y)$ $\in T$	$\vec{D}_o(.)\geq 0$ iff $(x,y)\in T$ if $(x,y)$is on the frontier of T, then $\vec{D}_o(.)=0$	$\nabla_x \vec{D}_o(.)\geq 0$ $\nabla_y \vec{D}_o(.)\leq 0$	translation $\vec{D}_O(x,y+\alpha g_y)=$ $\vec{D}_O(x,y;g_y)-\alpha$ homogeneity $\vec{D}_O(x,y;\lambda g_y)=$ $\lambda^{-1}\vec{D}_O(x,y;g_y)$	quasi- concave	concave	$TI_o(x,y)$ $=\vec{D}_o(.)$	$\vec{D}_o(x,y;y)$ $=\vec{D}_T(x,y;0,y)$ $=[1/D_o(x,y)]-1$ duality; $\vec{D}_O(x,y;g_y)\leq$ $\frac{R(x,p)-py}{pg_y}$

Table 1.2: A Summary of the Main Characteristics of Technical Inefficiency Models

Inefficiency Models	Technical Inefficiency		ω_i	Heterogeneous Inefficiency			Models are Proposed by
	u_i	u_{it}		u	Mean μ	Variance σ^2	
Time-Invariant Cross-Section	u_i	NA	NA	NA	$u_i \sim N^+(0, \sigma_u^2)$	σ_u^2	Aigner *et al.* (1977), Meeusen & Van den Broeck (1977)
Time-Invariant Fixed Effects	$\alpha_i = \alpha - u_i$	NA	NA	NA	No distribution assumption	No distribution assumption	Schmidt & Sickles (1984)
Time-Invariant Random Effects	u_i	NA	NA	NA	$u_i \sim N^+(0, \sigma_u^2)$	σ_u^2	Pitt & Lee (1981), Battese & Coelli (1988)
Time-Variant Fixed Effects	NA	$\alpha_{it} = g(t)$	NA	NA	No distribution assumption	No distribution assumption	Cornwell *et al* (1990), Lee & Schmidt (1993)
Time-Variant Random Effects	NA	$u_{it} = g(t)u_i$	NA	NA	$u_i \sim N^+(0, \sigma_u^2)$	σ_u^2	Kumbhakar (1990), Battese & Coelli (1992)
True Fixed Effect	NA	u_{it}	α_i	NA	$u_{it} \sim N^+(0, \sigma_u^2)$	σ_u^2	Greene (2005a, b)
True Random Effect	NA	u_{it}	ω_i	NA	$u_{it} \sim N^+(0, \sigma_u^2)$	σ_u^2	Greene (2005a, b)
Time-Variant and Time-Invariant	u_{i0}	u_{it}	NA	NA	$u_{it} \sim N^+(0, \sigma_u^2)$ $u_{i0} \sim N^+(0, \sigma_{u_0}^2)$	σ_u^2 $\sigma_{u_0}^2$	Kumbhakar & Heshmati (1995)
Four Random Components	u_{i0}	u_{it}	ω_i	NA	$u_{it} \sim N^+(0, \sigma_u^2)$ $u_{i0} \sim N^+(0, \sigma_{u_0}^2)$	σ_u^2 $\sigma_{u_0}^2$	Colombi *et al.* (2014), Kumbhakar *et al.* (2014)
Dynamic	NA	$u_{it} = h(u_{i,t-1})$	NA	$u_{it} = g(z; \delta)$	$u_{i,t-1} \sim N^+(\mu_{i,t-1}, \sigma_u^2)$	σ_u^2	Ahn & Sickles (2000), Tsionas (2006)
Threshold	u_i	NA	NA	NA	$u_i \sim N^+(0, \sigma_u^2)$ $0 \leq u_i \leq \theta$	σ_u^2	Lee (1996)
Zero Inefficiency	$u_i = 0$ with p $u_i > 0$ with $(1-p)$	NA	NA	NA	$u_i \sim N^+(0, \sigma_u^2)$	σ_u^2	Kumbhakar *et al.* (2013), Rho & Schmidt (2015)

Note: ω_i denotes heterogeneous technologies and NA indicates that the component is not included in the inefficiency model.

Table 1.2 (Cont′d): A Summary of the Main Characteristics of Technical Inefficiency Models

Inefficiency Models	Technical Inefficiency		ω_i	Heterogeneous Inefficiency			Models are Proposed by
	u_i	u_{it}		u	Mean μ	Variance σ^2	
Heterogeneous Z on u	u_i	NA	NA	$u_i = g(z;\delta)$	$u_i \sim N^+(\mu, \sigma_u^2)$	σ_u^2	Deprins & Simar (1989), Huang & Liu (1994)
Heterogeneous Z on μ	NA	u_{it}	NA	$u_{it} = g(z;\delta)$	$u_{it} \sim N^+(\mu_{it}, \sigma_u^2)$ $\mu_{it} = z_{it}\delta_m$	σ_u^2	Battese & Coelli (1995)
Heterogeneous Z on σ_u	u_i	NA	NA	NA	$u_i \sim N^+(0, \sigma_{ui}^2)$	$\sigma_{ui} = \exp(z_{ui}\gamma_u)$	Reifschneider & Stevenson (1991)
Heterogeneous Z on σ_u and σ_v	u_i	NA	NA	NA	$u_i \sim N^+(0, \sigma_{ui}^2)$	$\sigma_{ui} = \exp(z_{ui}\gamma_u)$ $\sigma_{vi} = \exp(z_{vi}\gamma_v)$	Hadri (1999), Hadri *et al.* (2003)
Heterogeneous Z on μ and σ_u	NA	u_{it}	NA	NA	$u_{it} \sim N^+(\mu_{it}, \sigma_{uit}^2)$ $\mu_{it} = z_{it}\delta_m$	$\sigma_{uit}^2 = \exp(z_{it}\gamma_u)$	Wang (2002), Wang & Schmidt (2002)
Heterogeneous Z on μ, σ_u and σ_v	NA	$u_{it} = g(t)u_i$	α_i	NA	$u_i \sim N^+(\mu_i, \sigma_{ui}^2)$ $\mu_i = z_i\delta_m$	$\sigma_{ui}^2 = \exp(z_{ui}\gamma_u)$ $\sigma_{vi}^2 = \exp(z_{vi}\gamma_v)$	Kumbhakar & Wang (2005)

Note: Z denotes the explanatory variables that reflect producer specific characteristics and explain the differences in inefficiency across producers.

Chapter 2

Interactive Effects between Input and Output Technical Inefficiencies

2.1 Introduction

Inefficiency is an important performance indicator for producers, with inefficiency measures being more accurate than other indicators because they involve a comparison with the most efficient frontier. Inefficiency measurement involves a comparison between actual inputs with optimal inputs located on the relevant frontier, or a comparison between actual outputs with optimal outputs, or some combination of the two. The optimal is defined in terms of production frontiers and value duals, such as cost, revenue, and profit frontiers, and inefficiency is technical[1].

Most empirical studies that examine the inefficiency of production processes employ either an input or an output-oriented measurement technique. In terms of the former, researchers assume that outputs are exogenous and inputs endogenous and producers are fully capable of reallocating resources when improving efficiency. Inefficiency can be measured in terms of distance functions either by using an input distance function or a directional input distance function. Similarly, by adopting an output-oriented measurement technique, it is assumed that inputs are exogenous and outputs endogenous and producers are fully capable of mixing production when improving efficiency. In this case, inefficiency can be measured in terms of distance functions either by using an output distance function or a directional output distance function. However, adopting an input (output) oriented measurement technique ignores the opposite output (input) orientation and this restriction may substantially bias

[1]I would like to thank Prof. Rolf Färe for insightful comments and suggestions that significantly improved the paper.

the measures of producer inefficiency.

An efficiency survey by Berger, Hunter, and Timme (1993) suggests comparing these input and output approaches with a complete approach to investigate the relationships between input and output inefficiencies. However, few studies examine total technical inefficiency and decompose it into input and output components either by using a profit function or a directional technology distance function. Berger, Hancock, and Humphrey (1993) and Akhavein *et al.* (1997) apply a distribution-free approach and show no interactive effects between input and output technical inefficiencies by using a profit function. More recently, Barros *et al.* (2012) and Fujii *et al.* (2014) apply a nonparametric approach and decompose total technical inefficiency into input and output components by using a weighted directional distance function that takes into account the contribution of each input and output on total technical inefficiency.

Even though these studies disaggregate and quantify the impact of input and output on inefficiency, the arbitrary decomposition of total technical inefficiency into input and output inefficiency components results in concluding that total technical inefficiency equals the sum of input and output technical inefficiencies and shows no interactive effects between them.

In contrast to previous studies that decompose total technical inefficiency into input and output inefficiency components, this paper begins from the observation that technical inefficiency can arise from employing the wrong level of inputs as well as from producing at the wrong level of outputs, and that the adjustability of both inputs and outputs is required for the improvement of producer efficiency. I follow Berger, Hunter, and Timme (1993) suggestion and compare these input and output approaches with a complete approach using directional input, output, and technology distance functions. I derive the interactive effect between input and output technical inefficiencies using exogenous and endogenous directional vectors. I show that overall technical inefficiency does not equal the sum of input and output technical inefficiencies, as previous studies claim. It equals the sum of input

and output technical inefficiencies plus an interactive effect component which captures the interactions between them. This suggests that the overuse of inputs creates input technical inefficiency and affects output technical inefficiency. Similarly, the loss of output creates output technical inefficiency and affects input technical inefficiency. Ignoring the interactive effect between input and output technical inefficiency results in a decomposition of overall technical inefficiency into input and output components that are significantly different from the ones that incorporate it.

I prove the results using the relationships between the directional distance functions and both the standard input and output distance functions, and their dual representations (cost, revenue, and profit functions). I also show that the interactive effect between input and output technical inefficiencies derived from the directional technology distance function depends on the choice of the directional vector in which the data are projected on the frontier and whether quantities and prices are taken into consideration. These results are quite significant, since these inefficiency components have different implications for producer performance, suggesting that the adjustability of both inputs and outputs is required for the improvement of producer efficiency. To the best of my knowledge, this paper is the first in the literature that derives the interactive effect between input and output technical inefficiencies theoretically using the directional technology distance function.

The rest of the paper is organized as follows. The next section presents some theoretical background on radial and directional measures of technical inefficiency using distance functions. Sections 3 and 4 derive the interactive effect between input and output technical inefficiencies using the directional technology distance function assuming exogenous and endogenous directional vectors, respectively. Section 5 presents numerical illustration, and the last section summarizes and concludes the paper.

2.2 Theoretical Foundations

To briefly review some of the literature on radial and directional measures of technical inefficiency using distance functions, consider a producer employing a vector of n inputs $x = (x_1, ..., x_n) \in \mathbb{R}^n_+$ available at fixed prices $w = (w_1, ..., w_n) \in \mathbb{R}^n_{++}$ to produce a vector of m outputs $y = (y_1, ..., y_m) \in \mathbb{R}^m_+$ that can be sold at fixed prices $p = (p_1, ..., p_m) \in \mathbb{R}^m_{++}$. Let $L(y)$ be the set of all input vectors x which can produce the output vector y

$$L(y) = \left\{x = (x_1, ..., x_n) \in R^N_+ : x \text{ can produce } y\right\}$$

and let $P(x)$ be the feasible set of outputs y that can be produced from the input vector x

$$P(x) = \left\{y = (y_1, ..., y_m) \in R^m_+ : y \text{ is producible from } x\right\}$$

The production technology T for a producer is defined as the set of all feasible input-output vectors

$$T = \left\{(x,\, y) : x \in \mathbb{R}^n_+,\, y \in \mathbb{R}^m_+,\, x \text{ can produce } y\right\}$$

Note that $(x, y) \in T \Leftrightarrow x \in L(y) \Leftrightarrow y \in P(x)$.

2.2.1 The Input Distance Function

Following Shephard (1953), I can define the input distance function (IDF) relative to the input set $L(y)$ or the production technology T as follows

$$D_I(y, x) = \max_{\vartheta_I} \left\{\vartheta_I : \frac{x}{\vartheta_I} \in L(y)\right\} = \max_{\vartheta_I} \left\{\vartheta_I : \left(\frac{x}{\vartheta_I}, y\right) \in T\right\}$$

where $1/\vartheta_I$ represents the proportional contraction of inputs that is required to reach the inner boundary of the input set or the production frontier, holding the outputs constant. $D_I(y, x)$ is given by the ratio of the observed input to the minimum input required to produce the given output. Thus, for any x, $x/D_I(y, x)$ is the minimum input vector on the ray from the origin through x that can produce y, as can be seen in Figure 2.1. Efficient producers,

who produce on the boundary of the input set or the production frontier, have $D_I(y,x) = 1$. Inefficiency is indicated by $D_I(y,x) > 1$.

The Debreu-Farrell input-oriented measure of technical efficiency is defined as

$$TE_I(y,x) = \frac{1}{D_I(y,x)}$$

Note that the IDF is the reciprocal of the Debreu-Farrell input-oriented measure of technical efficiency. $TE_I(y,x) \leq 1$ represents a radial reduction of inputs that is required to be considered as being efficient. Technical inefficiency is defined as

$$TI_I(y,x) = 1 - TE_I(y,x) = 1 - \frac{1}{D_I(y,x)}$$

where $0 \leq TI_I(y,x) \leq 1$. The IDF has the following properties [see Färe and Primont (1995), and Färe and Grosskopf (2004) for more details]:

i) representation, $D_I(y,x) \geq 1$ iff $x \in L(y)$ or $(x,y) \in T$

ii) non-increasing and quasi-concave in outputs, and

iii) non-decreasing, concave, and linearly homogeneous in inputs,
$D_I(y,\lambda x) = \lambda D_I(y,x)$, $\lambda > 0$.

2.2.2 The Output Distance Function

Instead of looking at the proportional contraction of inputs holding the outputs constant, the output distance function (ODF) considers the proportional expantion of outputs holding the inputs constant. Following Shephard (1970), it is defined on the output set $P(x)$ or the production technology T as

$$D_O(x,y) = \min_{\vartheta_O} \left\{ \vartheta_O : \frac{y}{\vartheta_O} \in P(x) \right\} = \min_{\vartheta_O} \left\{ \vartheta_O : \left(x, \frac{y}{\vartheta_O} \right) \in T \right\}$$

where $1/\vartheta_O$ represents the proportional expansion of outputs that is required to reach the upper boundary of the output set or the production frontier, holding the inputs constant.

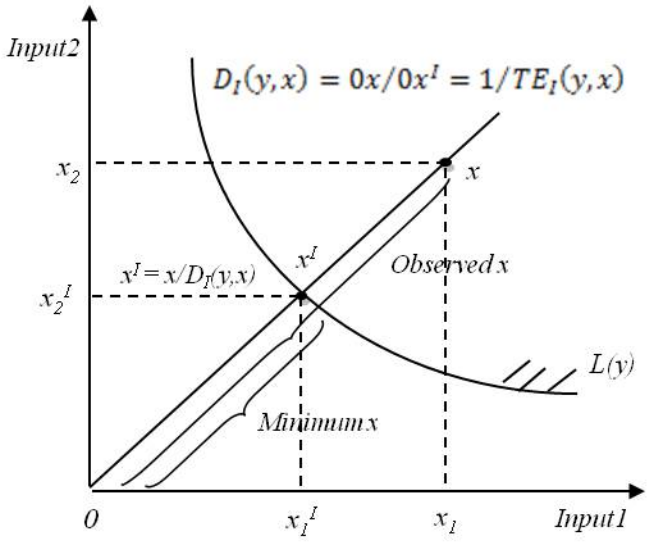

Figure 2.1: The Input Distance Function

$D_O(x, y)$ is given by the ratio of the observed output to maximum potential output obtainable from the given input. Thus, for any y, $y/D_O(x, y)$ is the largest output vector on the ray from the origin through y that can be produced by x, as can be seen in Figure 2.2. If y is on the boundary of the output set or the production frontier, $D_O(x, y) = 1$, implying that the producer is operating at full technical efficiency. If y is within the boundary of the output set or the production frontier, $D_O(x, y) < 1$, indicating that the producer is operating with technical inefficiency.

The Debreu-Farrell output-oriented measure of technical efficiency is defined as

$$TE_O(x, y) = \frac{1}{D_O(x, y)}$$

Note that the ODF is the reciprocal of the Debreu-Farrell output-oriented measure of technical efficiency. $TE_O(x, y) \geq 1$ represents a radial expansion of outputs that is required to achieve efficiency and the greater this measure, the smaller the efficiency. Technical inefficiency is defined as

$$TI_O(x, y) = TE_O(x, y) - 1 = \frac{1}{D_O(x, y)} - 1$$

where $TI_O(x, y) \geq 0$. The ODF has the following properties [see Färe and Grosskopf (1994) for more details]:

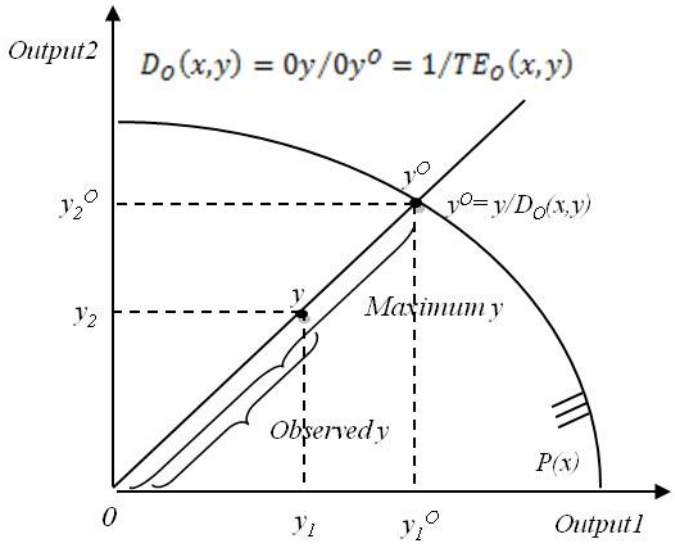

Figure 2.2: The Output Distance Function

i) representation, $D_O(x,y) \leq 1$ iff $y \in P(x)$ or $(x,y) \in T$

ii) non-increasing and quasi-convex in inputs, and

iii) non-decreasing, convex and linearly homogeneous in outputs,
$D_O(x,\lambda y) = \lambda D_O(x,y)$, $\lambda > 0$.

2.2.3 The Directional Technology Distance Function

The directional technology distance function (DTDF) generalizes Shephard's input and output distance functions, providing a tool to address efficiency issues in an integrated approach. It is introduced by Chambers *et al.* (1998) as a variant of the Luenberger (1995) shortage function. It allows for simultaneous contraction of inputs and expansion of outputs in terms of an explicit direction vector $g = (g_x, g_y)$, where $g_x \in R_+^N$ and $g_y \in R_+^M$ such that it contracts inputs in the direction g_x and expands outputs in the direction g_y. In particular, the DTDF is defined as

$$\begin{aligned}\vec{D}_T(x,y;g_x,g_y) &= \max_{\theta_T}\{\theta_T : (x-\theta_T g_x, y+\theta_T g_y) \in T\} \\ &= \max_{\theta_T}\{\theta_T : (x,y)+\theta_T(g_x,g_y) \in T\} \qquad (2.1)\end{aligned}$$

Efficient producers who produce on the frontier of T have $\vec{D}_T(x, y; g_x, g_y) = 0$, implying that there is no further contraction of inputs and expansion of outputs that is feasible. Inefficiency is indicated by $\vec{D}_T(x, y; g_x, g_y) > 0$, with higher values indicating greater inefficiency when producers operate beneath the frontier of T. A measure of technical inefficiency is defined as

$$TI_T = \vec{D}_T(x, y; g_x, g_y)$$

Eliminating technical inefficiency for producers who operate at point A would take the producers to point $B = \left(x^T, y^T\right) = (x - \theta_T g_x, y + \theta_T g_y)$ on the frontier of T, as can be seen in Figure 2.3. As noted by Chambers *et al.* (1998), the DTDF has the following properties:

i) the representation property; that is, $\vec{D}_T(x, y; g_x, g_y)$ completely characterizes the technology if inputs and outputs are strongly disposable,
$\vec{D}_T(x, y; g_x, g_y) \geq 0$ iff $(x, y) \in T$

ii) the translation property; that is, if (x, y) is translated into $(x - \alpha g_x, y + \alpha g_y)$, then the value of the directional distance function is reduced by α (for $\alpha \in R$)

$$\vec{D}_T(x - \alpha g_x, y + \alpha g_y; g_x, g_y) = \vec{D}_T(x, y; g_x, g_y) - \alpha$$

iii) non-decreasing in x if inputs are freely disposable; that is, $x' > x$ implies

$$\vec{D}_T(x', y; g_x, g_y) \geq \vec{D}_T(x, y; g_x, g_y)$$

iv) non-increasing in y if outputs are freely disposable; that is, $y' > y$ implies

$$\vec{D}_T(x, y'; g_x, g_y) \leq \vec{D}_T(x, y; g_x, g_y)$$

v) concave in (x, y)

vi) homogeneous of degree -1 in the directional vector, $g = (g_x, g_y)$

$\vec{D}_T(x, y; \lambda g_x, \lambda g_y) = \lambda^{-1} \vec{D}_T(x, y; g_x, g_y)$, for $\lambda > 0$

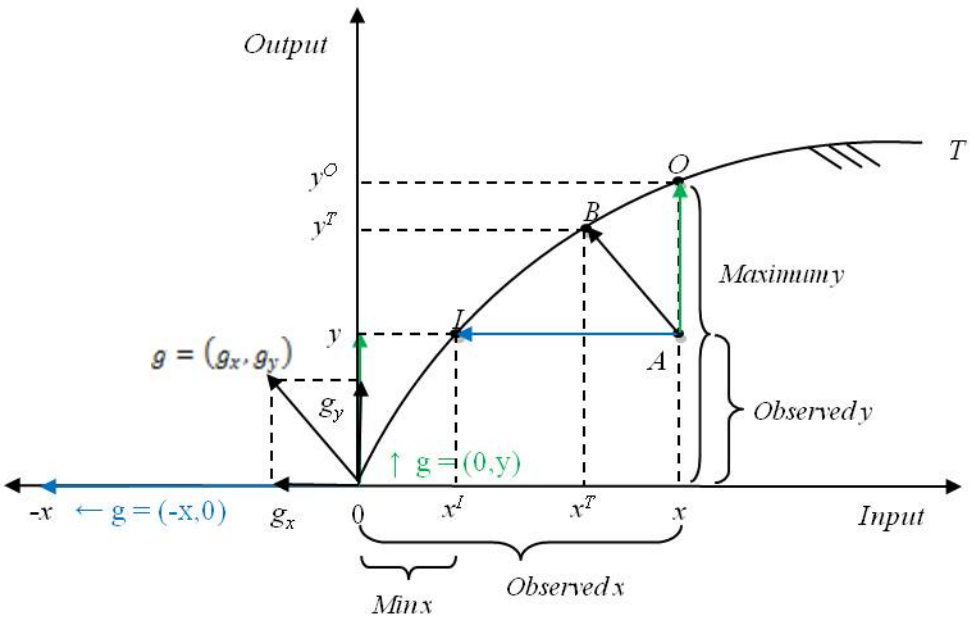

Figure 2.3: Directional Distance Functions with Different Directional Vectors

vii) homogeneous of degree $+1$ in x and y if the technology exhibits constant returns to scale, $\vec{D}_T(\lambda x, \lambda y; g_x, g_y) = \lambda \vec{D}_T(x, y; g_x, g_y)$, for $\lambda > 0$.

The Directional Input Distance Function

The inefficiency measures derived from the directional distance function depend on the choice of the directional vector, $g = (g_x, g_y)$. By setting $g_y = 0$, the directional vector becomes $g = (g_x, 0)$ and allows only for input contraction holding outputs fixed — see Figure 2.3. In this case, equation (2.1) becomes the directional input distance function (DIDF) that allows for only input contraction, $\vec{D}_T(x, y; g_x, 0) = \vec{D}_I(y, x; g_x)$

$$\vec{D}_I(y, x; g_x) = \max_{\theta_I} \{\theta_I : (x - \theta_I g_x) \in L(y)\} = \max_{\theta_I} \{\theta_I : (x - \theta_I g_x,\ y) \in T\}$$

Moreover, according to Chambers *et al.* (1996, 1998) and Fare and Grosskopf (2000), if the directional input vector, g_x, equals the observed input vector, x, (that is, $g_x = -x$), then

$$\vec{D}_I(y, x; g_x) = \vec{D}_I(y, x; -x) = 1 - \frac{1}{D_I(y, x)}$$

and in this case there is a relationship between the directional input distance function, $\vec{D}_I(y, x; -x)$, and the standard input distance function, $D_I(y, x)$. As can be seen in Figure 2.3, producers who operate at point A can hold output constant and contract input in the direction $g_x = -x$ to point I, which yields a value of $\vec{D}_I(y, x; -x) = \theta_I = 1 - 1/D_I(y, x) = 1 - 0x^I/0x = x^I x/0x$. The DIDF serves as an input-oriented measure of technical inefficiency

$$TI_I = \vec{D}_I(y, x; g_x)$$

The DIDF satisfies the following properties [see Chambers *et al.* (1996)]:

i) representation, $\vec{D}_I(y, x; g_x) \geq 0$ iff $x \in L(y)$ or $(x, y) \in T$

ii) translation, $\vec{D}_I(y, x - \alpha g_x; g_x) = \vec{D}_I(y, x; g_x) - \alpha$, for $\alpha \in R$

iii) concavity in inputs

iv) positive monotonicity in inputs. That is, $x' > x$ implies
$\vec{D}_I(y, x'; g_x) \geq \vec{D}_I(y, x; g_x)$

v) negative monotonicity in outputs. That is, $y' > y$ implies
$\vec{D}_I(y', x; g_x) \leq \vec{D}_I(g, x; g_x)$, and

vi) homogeneity of degree -1 in g_x. That is,
$\vec{D}_I(y, x; \lambda g_x) = \lambda^{-1}\vec{D}_I(y, x; g_x)$, for $\lambda > 0$.

The Directional Output Distance Function

By setting $g_x = 0$, the directional vector becomes $g = (0, g_y)$ and allows only for output expansion holding inputs fixed — see Figure 2.3. In this case, equation (2.1) becomes the directional output distance function (DODF) that allows for only output expansion, $\vec{D}_T(x, y; 0, g_y) = \vec{D}_O(x, y; g_y)$

$$\vec{D}_O(x, y; g_y) = \max_{\theta_o} \{\theta_O : (y + \theta_O g_y) \in P(x)\} = \max_{\theta_o} \{\theta_O : (x, y + \theta_O g_y) \in T\}$$

Moreover, as noted by Chambers *et al.* (1998) and Färe and Grosskopf (2000), if the directional output vector, g_y, equals the observed output vector, y (that is, $g_y = y$), then

$$\vec{D}_O(x, y; g_y) = \vec{D}_O(x, y; y) = \frac{1}{D_O(x, y)} - 1$$

and in this case there is a relationship between the directional output distance function, $\vec{D}_O(x, y; y)$, and the standard output distance function, $D_O(x, y)$. As can be seen in Figure 2.3, producers who operate at point A can hold input constant and expand output in the direction $g_y = y$ to point O, which yields a value of $\vec{D}_O(x, y; y) = \theta_O = [1/D_O(x, y)] - 1 = \left[0y^O/0y\right] - 1 = yy^O/0y$. The DODF serves as an output-oriented measure of technical inefficiency

$$TI_O = \vec{D}_O(x, y; g_y)$$

The DODF satisfies the following properties [see Färe *et al.* (2005)]:

i) representation, $\vec{D}_O(x, y; g_y) \geq 0$ iff $y \in P(x)$ or $(x, y) \in T$

ii) translation, $\vec{D}_O(x, y + \alpha g_y; g_y) = \vec{D}_O(x, y; g_y) - \alpha$, for $\alpha \in R$

iii) concavity in outputs

iv) positive monotonicity in inputs. That is, $x' > x$ implies
$\vec{D}_O(x', y; g_y) \geq \vec{D}_O(x, y; g_y)$

v) negative monotonicity in outputs. That is, $y' > y$ implies
$\vec{D}_O(x, y'; g_y) \leq \vec{D}_O(x, y; g_y)$, and

vi) homogeneity of degree -1 in g_y. That is,
$\vec{D}_O(x, y; \lambda g_y) = \lambda^{-1}\vec{D}_O(x, y; g_y)$, for $\lambda > 0$.

2.2.4 Duality Relationships

The standard and directional distance functions are primal representations of the technology. The dual representations of the technology are given by the profit, cost, and revenue

functions. Given input prices $w \in \mathbb{R}^n_{++}$ and output prices $p \in \mathbb{R}^m_{++}$, the following dual representations (profit, cost, and revenue functions) of the technology can be defined in terms of the directional distance functions as

$$\pi(p,w) = \max_{x,y} \left\{ (py - wx) : \vec{D}_T(x,y;g_x,g_y) \geq 0 \right\}$$

$$C(y,w) = \min_{x} \left\{ wx : \vec{D}_I(y,x;g_x) \geq 0 \right\}$$

$$R(x,p) = \max_{y} \left\{ py : \vec{D}_O(x,y;g_y) \geq 0 \right\}$$

The relationship between the directional technology (input or output) distance function and the profit (cost or revenue) function can be represented as [see Chambers *et al.* (1998), and Färe and Grosskopf (2000)]

$$\vec{D}_T(x,y;g_x,g_y) \leq \frac{\pi(p,w) - (py - wx)}{pg_y + wg_x}$$

$$\vec{D}_I(y,x;g_x) \leq \frac{wx - C(y,w)}{wg_x}$$

$$\vec{D}_O(x,y;g_y) \leq \frac{R(x,p) - py}{pg_y}$$

The right-hand side can be interpreted as a measure of profit (cost or revenue) inefficiency comparing observed profit (cost or revenue) to maximum profit (minimum cost or maximum revenue) normalized by the value of the directional vector. The left-hand side captures overall (input or output-oriented) technical inefficiency. The inequality can be turned into equality by adding a residual term that captures allocative inefficiency, where allocative inefficiency is due to the failure of choosing the profit-maximizing (cost-minimizing or revenue-maximizing) input-output (input or output) vector given relative input and output market prices. Thus, technical inefficiency is due to the overuse of inputs or the loss of production of outputs or both, and allocative inefficiency results from employing inputs and outputs in the wrong proportions.

Note that profit inefficiency is less than or equal to revenue inefficiency and greater than, less than, or equal to cost inefficiency since profits are greater than, less than, or equal to

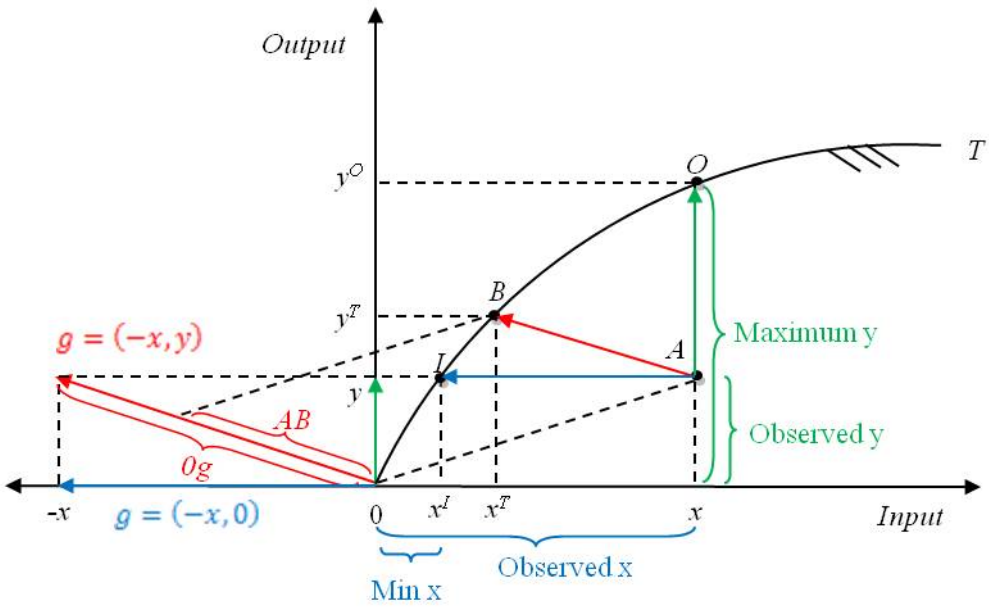

Figure 2.4: Inefficiency Measures with the Observed Input-Output Directional Vector

costs. Furthermore, revenue inefficiency is greater than, less than, or equal to cost inefficiency when profits are positive, negative, or zero, respectively.

2.3 Exogenous Directional Vectors

When only quantity information on input and output is available, and price information is unavailable, distorted or inaccurate, technical inefficiency can be measured by choosing a pre-specified directional vector such that it projects any inefficient producer to the frontier of T. Two widely used directions are the observed input-output direction $g = (-x, y)$ and the unit value direction $g = (-1, 1)$.

Suppose that the interest is to measure the simultaneous maximum proportional contraction of input and expansion of output that is feasible given the technology, the pre-specified directional vector $g = (-x, y)$ can be chosen. This type of directional vector assumes that an inefficient producer can decrease inefficiency while decreasing input and increasing output in proportion to the initial combination of the actual input and output.

Figure 2.4 illustrates how inefficiency can be measured using the pre-specified directional

vector $g = (-x, y)$. Producers who operate beneath the production frontier at point A are technically inefficient. The simultaneous maximum proportional contraction of input and expansion of output can be measured in terms of the lengths of x and y, with the use of the Pythagorean theorem, as

$$\vec{D}_T(x, y; g_x, g_y) = \vec{D}_T(x, y; -x, y) = \frac{\|AB\|}{\|Og\|} = \frac{\sqrt{(\|x\| - \|x^T\|)^2 + (\|y^T\| - \|y\|)^2}}{\sqrt{\|x\|^2 + \|y\|^2}} = \theta_T$$

The maximum proportional contraction of input holding output constant can be measured using the directional vector $g = (-x, 0)$. Technical inefficiency is considered to be input-oriented technical inefficiency and is defined by the difference of the lengths of x and x^I divided by the length of x

$$\vec{D}_T(x, y; -x, 0) = \vec{D}_I(y, x; -x) = \frac{\|x\| - \|x^I\|}{\|x\|} = 1 - \frac{\|x^I\|}{\|x\|} = 1 - \frac{1}{D_I(y, x)} = \theta_I$$

The maximum proportional expansion of output holding input constant can be measured using the directional vector $g = (0, y)$. In this case, technical inefficiency is considered to be output-oriented technical inefficiency and is defined as

$$\vec{D}_T(x, y; 0, y) = \vec{D}_O(x, y; y) = \frac{\|y^O\| - \|y\|}{\|y\|} = \frac{\|y^O\|}{\|y\|} - 1 = \frac{1}{D_O(x, y)} - 1 = \theta_O$$

A critical question that needs to be considered is whether $\theta_T = \theta_I + \theta_O$?

Proposition 1 *Let $\vec{D}_T(x, y; -x, 0) = \vec{D}_I(y, x; -x) = \theta_I$ be input-oriented technical inefficiency, $\vec{D}_T(x, y; 0, y) = \vec{D}_O(x, y; y) = \theta_O$ be output-oriented technical inefficiency, and $\vec{D}_T(x, y; -x, y) = \theta_T$ be overall technical inefficiency. Then $\theta_T = \theta_I + \theta_O - \theta_{IO}$, where θ_{IO} is the interactive effect between input and output-oriented technical inefficiencies.*

Proof *The sum of input and output-oriented technical inefficiencies is defined as $\theta_I + \theta_O$. Since $\theta_I = 1 - 1/D_I(y, x)$ and $\theta_O = [1/D_O(x, y)] - 1$, then $\theta_I + \theta_O = [1/D_O(x, y)] - [1/D_I(y, x)]$. Since $D_I(y, x) \geq 1$, then $1/D_I(y, x) \leq 1$ and since $1/D_I(y, x) = 1 - \theta_I \leq 1$,*

then $0 \leq \theta_I \leq 1$ *which implies that the maximum proportional contraction of input would not exceed the initial input and the resulting input could still produce the output (that is,* $x \geq x - \theta_I x$*,* $x \geq 0$ *and* $y\left(x - \theta_I x\right) = y$*). Similarly, since* $D_O\left(x, y\right) \leq 1$*, then* $1/D_O\left(x, y\right) \geq 1$ *and since* $1/D_O\left(x, y\right) = 1 + \theta_O \geq 1$*, then* $\theta_O \geq 0$*. Since* $0 \leq \theta_I \leq 1$ *and* $\theta_O \geq 0$*, then* $\theta_I + \theta_O \geq 0$*.*

The overall technical inefficiency θ_T *is* $0 \leq \theta_T \leq 1$ *to ensure that the inequality* $x \geq x - \theta_T x$ *holds. Furthermore, the directional technology distance function contracts input simultaneously with expanding output while the directional input distance function contracts input holding output fixed,* θ_T *is less than* θ_I *(that is,* $\theta_T < \theta_I$ *and* $y\left(x - \theta_T x\right) > y$*) which implies that more input is needed to produce the expanding output.*

Now, I have $0 \leq \theta_T \leq 1$*,* $0 \leq \theta_I \leq 1$*,* $\theta_T < \theta_I$*,* $\theta_O \geq 0$*, and* $\theta_I + \theta_O \geq 0$*. As a result,* $\theta_T \leq \theta_I + \theta_O$*. Thus, the inequality can be turned into equality by subtracting a residual term that captures the interactive effect between input and output-oriented technical inefficiencies,* θ_{IO}*, where* $\theta_T = \theta_I + \theta_O - \theta_{IO}$*, and the interactive effect is defined as the gap in the inequality, namely* $\theta_{IO} = \theta_I + \theta_O - \theta_T$*. Q.E.D.*

This derivation of the interactive effects based on the observed input-output directional vector is illustrated in Figure 2.5.

Another widely used pre-specified direction is the unit value direction $g = (-1, 1)$. Figure 2.6 illustrates how inefficiency can be measured using this directional vector. This type of directional vector implies that the amount by which a producer could decrease input and increase output will be $\vec{D}_T\left(x, y; -1, 1\right) \times 1$ units of x and y.

Producers who operate at points O, B, and I that belong to the boundary of the production frontier are overall, input and output technically efficient since their associated directional technology, input, and output distance functions in the feasible directions are zero; $\vec{D}_T\left(x, y; -1, 1\right) = \theta_T^1 = 0$, $\vec{D}_I\left(y, x; -1\right) = \theta_I^1 = 0$, and $\vec{D}_O\left(x, y; 1\right) = \theta_O^1 = 0$. However,

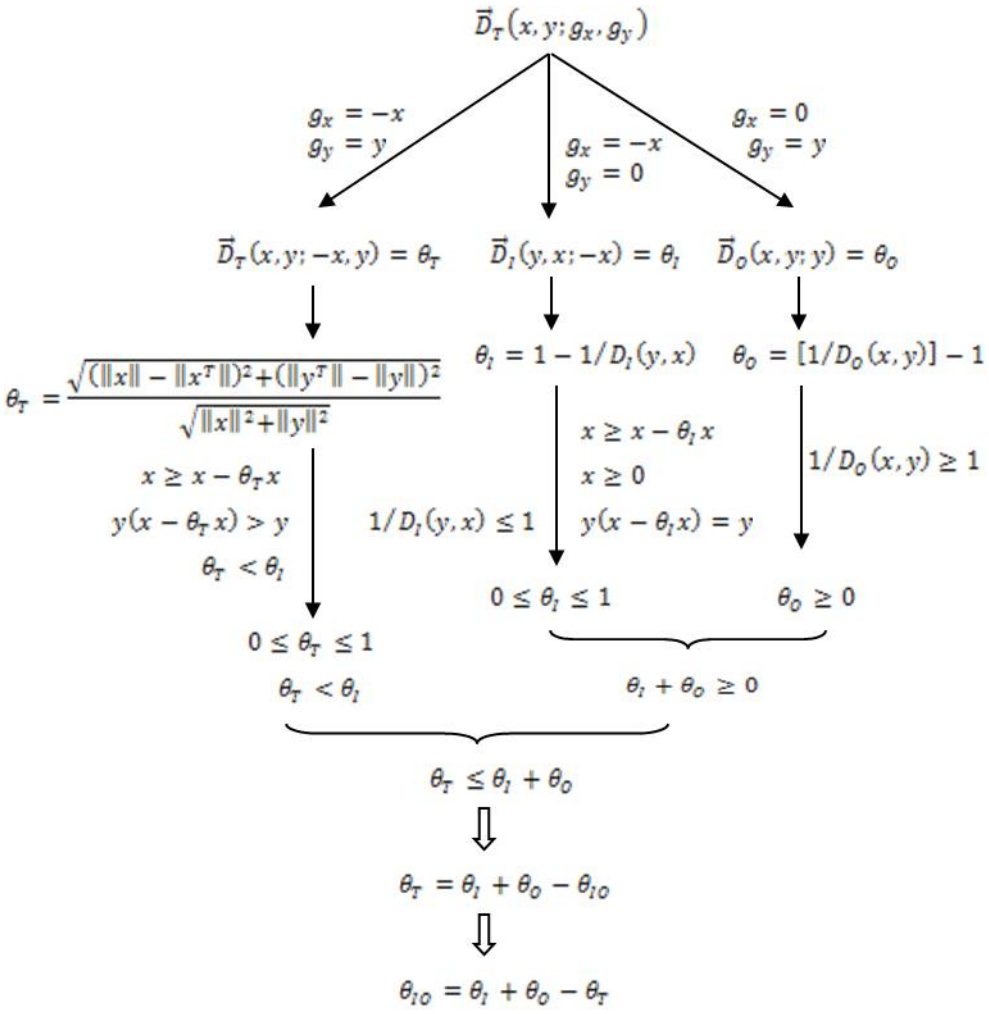

Figure 2.5: Interactive Effects Based on the Observed Input-Output Directional Vector

producers who operate at point A that is located beneath the production frontier are overall, input and output technically inefficient since their associated directional technology, input and output distance functions in the feasible directions are positive. The simultaneous maximum contraction of input and expansion of output can be measured in terms of the lengths of x and y, with the use of the Pythagorean theorem, as

$$\vec{D}_T\left(x,y;-1,1\right)=\frac{\|AB\|}{\|Og\|}=\frac{\sqrt{\left(\|x\|-\|x^T\|\right)^2+\left(\|y^T\|-\|y\|\right)^2}}{\sqrt{2}}=\theta_T^1$$

The maximum contraction of input holding output constant can be measured using the directional vector $g=(-1,0)$. Technical inefficiency is considered to be input-oriented technical inefficiency and is defined as

$$\vec{D}_I\left(y,x;-1\right)=\|AI\|=\|x\|-\left\|x^I\right\|=\theta_I^1$$

The maximum expansion of output holding input constant can be measured using the directional vector $g=(0,1)$. In this case, technical inefficiency is considered to be output-oriented technical inefficiency and is defined as

$$\vec{D}_O\left(x,y;1\right)=\|AO\|=\left\|y^O\right\|-\|y\|=\theta_O^1$$

Does $\theta_T^1=\theta_I^1+\theta_O^1$?

Proposition 2 *Let* $\vec{D}_T\left(x,y;-1,\,0\right)=\vec{D}_I\left(y,x;-1\right)=\theta_I^1$ *be input-oriented technical inefficiency,* $\vec{D}_T\left(x,y;0,1\right)=\vec{D}_O\left(x,y;1\right)=\theta_O^1$ *be output-oriented technical inefficiency, and* $\vec{D}_T\left(x,y;-1,1\right)=\theta_T^1$ *be overall technical inefficiency. Then* $\theta_T^1=\theta_I^1+\theta_O^1-\theta_{IO}^1$*, where* θ_{IO}^1 *is the interactive effect between input and output-oriented technical inefficiencies.*

Proof *The sum of input and output-oriented technical inefficiencies is* $\theta_I^1+\theta_O^1$*. Since* $\theta_I^1=\|x\|-\left\|x^I\right\|$*, then adding and subtracting* $\left\|x^T\right\|$ *yields* $\theta_I^1=\left(\|x\|-\left\|x^T\right\|\right)+\left(\left\|x^T\right\|-\left\|x^I\right\|\right)$*. Since* $\theta_O^1=\left\|y^O\right\|-\|y\|$*, then adding and subtracting* $\left\|y^T\right\|$ *yields* $\theta_O^1=\left(\left\|y^O\right\|-\left\|y^T\right\|\right)+$

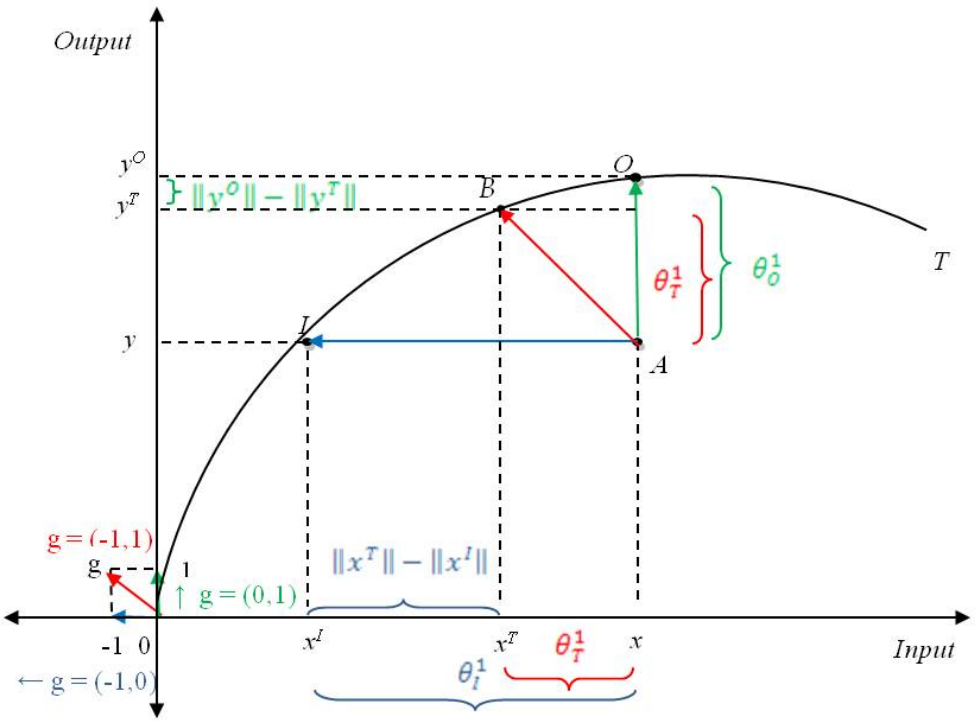

Figure 2.6: Inefficiency Measures with the Unit Value Directional Vector

$(\|y^T\| - \|y\|)$. *Thus,* $(\|x\| - \|x^T\|) + (\|x^T\| - \|x^I\|) + (\|y^O\| - \|y^T\|) + (\|y^T\| - \|y\|) = \theta_I^1 + \theta_O^1$, *and* $\theta_T^1 = \sqrt{(\|x\| - \|x^T\|)^2 + (\|y^T\| - \|y\|)^2}/\sqrt{2}$. *Since each point on the 45-degree line equates the variable measured on the vertical axis with the variable measured on the horizontal axis, then* $(\|x\| - \|x^T\|) = (\|y^T\| - \|y\|) = k$. *Substituting* $(\|x\| - \|x^T\|) = (\|y^T\| - \|y\|) = k$, *then* $\theta_T^1 = \sqrt{2k^2}/\sqrt{2} = k$, *and* $\theta_I^1 + \theta_O^1 = 2k + j$ *where* $j = (\|x^T\| - \|x^I\|) + (\|y^O\| - \|y^T\|)$. *As a result,* $\theta_T^1 \leq \theta_I^1 + \theta_O^1$. *Thus, the inequality can be turned into equality by subtracting a residual term that captures the interactive effect between input and output-oriented technical inefficiencies* θ_{IO}^1 *where* $\theta_T^1 = \theta_I^1 + \theta_O^1 - \theta_{IO}^1$ *and* $\theta_{IO}^1 = k + j$ *is the interactive effect between input and output-oriented technical inefficiencies. Q.E.D.*

This derivation of the interactive effects based on the unit value directional vector is illustrated in Figure 2.7.

Corollary 1 *Let θ_T^1 be overall technical inefficiency, θ_I^1 be input-oriented technical inefficiency, and θ_O^1 be output-oriented technical inefficiency derived using the unit value directional vectors. Then the interactive effect between input and output-oriented technical inefficiencies θ_{IO}^1 is related to θ_I^1 and θ_O^1 as*

$$\theta_{IO}^1 = \theta_I^1 + \left(\left\|y^O\right\| - \left\|y^T\right\|\right)$$

$$\theta_{IO}^1 = \theta_O^1 + \left(\left\|x^T\right\| - \left\|x^I\right\|\right)$$

While producers who operate on the production frontier have zero interactive effects since $\theta_T^1 = \theta_I^1 + \theta_O^1 = 0$, producers who operate beneath the production frontier have negative interactive effects since $\theta_T^1 < \theta_I^1 + \theta_O^1$. Furthermore, there is a relationship between the interactive effect θ_{IO}^1, and the input and output technical inefficiencies. The interactive effect between input and output technical inefficiencies equals the input technical inefficiency θ_I^1 plus the loss of production of output $\left\|y^O\right\| - \left\|y^T\right\|$ (that is part of output technical inefficiency) that is forgone to reduce input and eliminate part of the input technical inefficiency $\|x\| - \left\|x^T\right\|$ — see Figure 2.6. This suggests that the loss of output creates output technical inefficiency and has an effect on reducing input technical inefficiency. Intuitively, the loss of revenue from the production of output may encourage producers to reduce the input used in the production process. The interactive effect also equals the output technical inefficiency θ_O^1 plus the overuse of input $\left\|x^T\right\| - \left\|x^I\right\|$ (that is part of input technical inefficiency) that is used to produce more output and eliminate part of the output technical inefficiency $\left\|y^T\right\| - \|y\|$ — see Figure 2.6. This suggests that the overuse of input creates input technical inefficiency and has an effect on reducing output technical inefficiency. Intuitively, the excessive cost due to the overuse of input may encourage producers to produce more output to finance the cost.

Comparing these alternative ways to improve efficiency is crucial in eliminating overall technical inefficiency particularly when producers are not fully capable of reducing all the

overuse of input or producing all the loss of production while improving efficiency due to organizational, technological, or any other restrictions. The interactive effect between input and output technical inefficiencies tend to lower overall technical inefficiency since input inefficiency has an effect on reducing (improving) output inefficiency (efficiency) and output inefficiency has an effect on reducing (improving) input inefficiency (efficiency).

Consequently, including no price information on the directional vector such as $g = (-x, y)$ or $g = (-1, 1)$, the interactive effects between input and output technical inefficiencies have negative effect on the overall technical inefficiency. Producers with larger values of the interactive effect tend to have a lower level of overall technical inefficiency which indicates that they are more efficient.

2.4 Endogenous Directional Vectors

When information on input and output prices is available and producers are assumed to exhibit cost-minimizing (or revenue or profit-maximizing) behavior, technical inefficiency can be measured by choosing an endogenous direction vector such that it projects any inefficient producer to the cost-minimizing (or revenue or profit-maximizing) benchmark.

Following Zofio *et al.* (2013), the directional vector $g = \left(g_x^\pi, g_y^\pi\right)$ is assumed to satisfy the price normalization constraint $pg_y^\pi + wg_x^\pi = 1$ and projects any inefficient producer towards the profit-maximizing bundle (x^π, y^π) where producers are both technically and allocatively efficient[2]. Thus, the directional vector can be defined as

$$g = \left(g_x^\pi, g_y^\pi\right) = \left(\frac{x - x^\pi}{\pi\left(p, w\right) - \left(py - wx\right)}, \frac{y^\pi - y}{\pi\left(p, w\right) - \left(py - wx\right)}\right) \tag{2.2}$$

to ensure that $pg_y^\pi + wg_x^\pi = 1$. Then, the directional distance function, $\vec{D}_T\left(x, y; g_x^\pi, g_y^\pi\right)$, equals the loss of profit due to technical inefficiency and gives a measure of overall technical inefficiency in monetary values. Note that $pg_y^\pi + wg_x^\pi$ can be interpreted as the value of the

[2]The normalizing constraint of the value of the directional vector is used by Luenberger (1992) in the context of consumer theory.

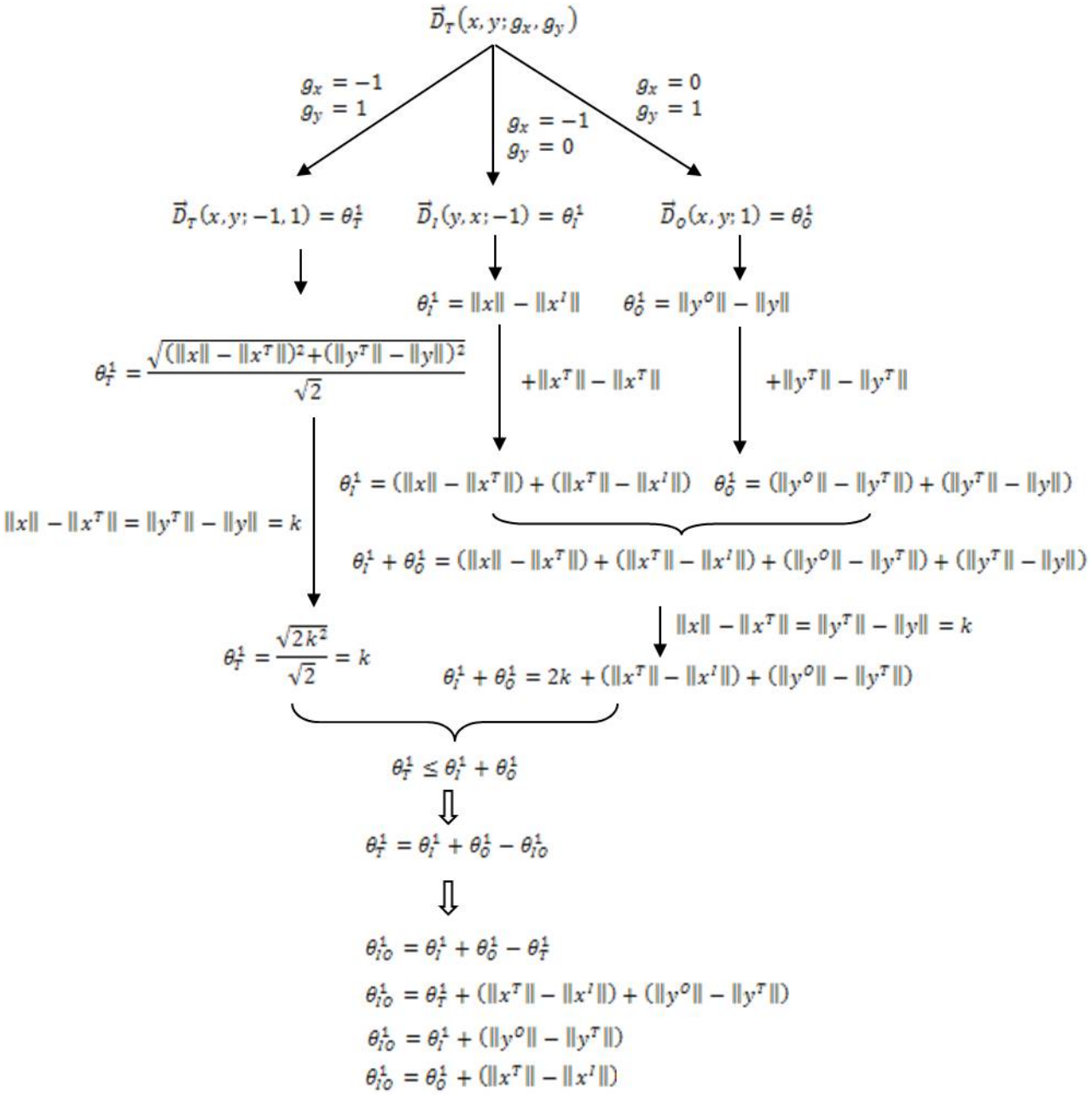

Figure 2.7: Interactive Effects Based on the Unit Value Directional Vector

directional vector and $(\pi(p,w)-(py-wx))$ is the difference between maximum profit and observed profit.

Proposition 3 *Let $(x,y)\in T$, (p,w) be the vector of output and input prices, and $g=\left(g_x^\pi,g_y^\pi\right)$ be a vector such that it satisfies $pg_y^\pi+wg_x^\pi=1$ and projects any inefficient producer to the profit-maximizing bundle (x^π,y^π), where producers are both technically and allocatively efficient. Then, $\vec{D}_T\left(x,y;g_x^\pi,g_y^\pi\right)=\theta_T^\pi=\pi(p,w)-(py-wx)$ and all profit inefficiency is technical, since allocative inefficiency equals zero.*

Proof *For every $(x,y)\in T$, the projected vector (x^π,y^π) based on the directional vector $g=\left(g_x^\pi,g_y^\pi\right)$ is $\left(x-\theta_T^\pi g_x^\pi,y+\theta_T^\pi g_y^\pi\right)\in T$ or, equivalently, $(x,y)+\theta_T^\pi\left(g_x^\pi,g_y^\pi\right)\in T$,* where $\theta_T^\pi=\vec{D}_T\left(x,y;g_x^\pi,g_y^\pi\right)$. Thus, $(py^\pi-wx^\pi)=(py-wx)+\theta_T^\pi\left(pg_y^\pi+wg_x^\pi\right)$. *Using the directional vector given in (2.2) yields, after some rearranging, $\theta_T^\pi=\pi(p,w)-(py-wx)$.* Q.E.D.

Suppose that the directional vector $g=\left(g_x^C,0\right)$ satisfies the price normalization constraint, $wg_x^C=1$ and projects any inefficient producer towards the cost-minimizing bundle $\left(x^C,y\right)$ where producers are both technically and allocatively efficient. Thus, the directional vector can be defined as

$$g=\left(g_x^C,0\right)=\left(\frac{x-x^C}{wx-C(y,w)},\,0\right) \tag{2.3}$$

to ensure that $wg_x^C=1$. Then, the directional input distance function $\vec{D}_I\left(y,x;g_x^C\right)$ equals the excessive cost due to technical inefficiency and gives a measure of input-oriented technical inefficiency in monetary values. Note that wg_x^C can be interpreted as the value of the directional vector and $(wx-C(y,w))$ is the difference between observed cost and minimum cost.

Proposition 4 *Let $x\in L(y)$, w be the vector of input prices, and $g=\left(g_x^C,0\right)$ a vector*

that satisfies $wg_x^C = 1$ *and projects any inefficient producer to the cost-minimizing bundle* $\left(x^C, y\right)$ *where producers are both technically and allocatively efficient. Then,* $\vec{D}_I\left(y, x; g_x^C\right) = \theta_I^C = wx - C\left(y, w\right)$ *and all cost inefficiency is technical, since allocative inefficiency equals zero.*

Proof *For every* $x \in L\left(y\right)$, *the projected vector* $\left(x^C, y\right)$ *based on the directional vector* $g = \left(g_x^C, 0\right)$ *is* $\left(x - \theta_I^C g_x^C, y\right) \in T$ *or, equivalently,* $\left(x - \theta_I^C g_x^C\right) \in L\left(y\right)$, *where* $\theta_I^C = \vec{D}_I\left(y, x; g_x^C\right)$. *Thus,* $wx^C = C\left(y, w\right) = wx - \theta_I^C wg_x^C$, *which after using the directional vector in (2.3), reduces to* $\theta_I^C = wx - C\left(y, w\right)$. Q.E.D.

Suppose that the directional vector $g = \left(0, g_y^R\right)$ satisfies the price normalization constraint $pg_y^R = 1$ and projects any inefficient producer towards the revenue-maximizing bundle $\left(x, y^R\right)$ where producers are both technically and allocatively efficient. Thus, the directional vector can be defined as

$$g \ = \left(0, g_y^R\right) = \left(0, \ \frac{y^R - y}{R\left(x, p\right) - py}\right) \tag{2.4}$$

to ensure that $pg_y^R = 1$. Then, the directional output distance function $\vec{D}_O\left(x, y; g_y^R\right)$ equals the loss of revenue due to technical inefficiency and gives a measure of output-oriented technical inefficiency in monetary values. Note that pg_y^R can be interpreted as the value of the directional vector and $\left(R\left(x, p\right) - py\right)$ is the difference between maximum revenue and observed revenue.

Proposition 5 *Let* $y \in P\left(x\right)$, p *be the vector of output prices, and* $g = \left(0, g_y^R\right)$ *a vector that satisfies* $pg_y^R = 1$ *and projects any inefficient producer to the revenue-maximizing bundle* $\left(x, y^R\right)$ *where producers are both technically and allocatively efficient. Then,* $\vec{D}_O\left(x, y; g_y^R\right) = \theta_O^R = R\left(x, p\right) - py$ *and all revenue inefficiency is technical, since allocative inefficiency equals zero.*

Proof *For every* $y \in P(x)$, *the projected vector* $\left(x, y^R\right)$ *based on the directional vector* $g = \left(0, g_y^R\right)$ *is* $\left(x, y + \theta_O^R g_y^R\right) \in T$ or, equivalently, $\left(y + \theta_O^R g_y^R\right) \in P(x)$, where $\theta_O^R = \vec{D}_O\left(x, y; g_y^R\right)$. *Thus,* $py^R = R(x,p) = py + \theta_O^R p g_y^R$, *which after using the directional vector in (2.4), reduces to* $\theta_O^R = R(x,p) - py$. Q.E.D.

Does $\theta_T^\pi = \theta_I^C + \theta_O^R$?

Proposition 6 *Let* $\vec{D}_T\left(x, y; g_x^\pi, g_y^\pi\right) = \theta_T^\pi$ *be overall technical inefficiency,* $\vec{D}_I\left(y, x; g_x^C\right) = \theta_I^C$ *be input-oriented technical inefficiency, and* $\vec{D}_O\left(x, y; g_y^R\right) = \theta_O^R$ *be output-oriented technical inefficiency. Then* $\theta_T^\pi = \theta_I^C + \theta_O^R \pm \theta_{IO}^{CR}$, *where* θ_{IO}^{CR} *is the interactive effect between input and output-oriented technical inefficiencies.*

Proof *The sum of input and output-oriented technical inefficiency can be defined as* $\theta_I^C + \theta_O^R$. *Since* $\theta_I^C = wx - C(y,w)$ *and* $\theta_O^R = R(x,p) - py$, *then* $\theta_I^C + \theta_O^R = R(x,p) - C(y,w) - (py - wx)$. *Since* $\theta_T^\pi = \pi(p,w) - (py - wx)$ *and* $\pi(p,w) \gtreqless C(y,w)$, *then* $\theta_T^\pi \gtreqless \theta_I^C + \theta_O^R$. *Thus, the inequality can be turned into equality by adding or subtracting a residual term that captures the interactive effect between input and output-oriented technical inefficiencies* θ_{IO}^{CR}, *where* $\theta_T^\pi = \theta_I^C + \theta_O^R \pm \theta_{IO}^{CR}$ *and the interactive effect is defined as the gap in the inequality, namely* $\theta_{IO}^{CR} = \theta_T^\pi \mp \left(\theta_I^C + \theta_O^R\right)$. Q.E.D.

This derivation of the interactive effects using the endogenous directional vectors $g = \left(g_x^\pi, g_y^\pi\right)$, $g = \left(g_x^C, 0\right)$, and $g = \left(0, g_y^R\right)$ is illustrated in Figure 2.8.

Consequently, including price information on the directional vector such as the directional vector that projects any inefficient producer to the profit-maximizing bundle, the interactive effects between input and output-oriented technical inefficiencies may have positive or

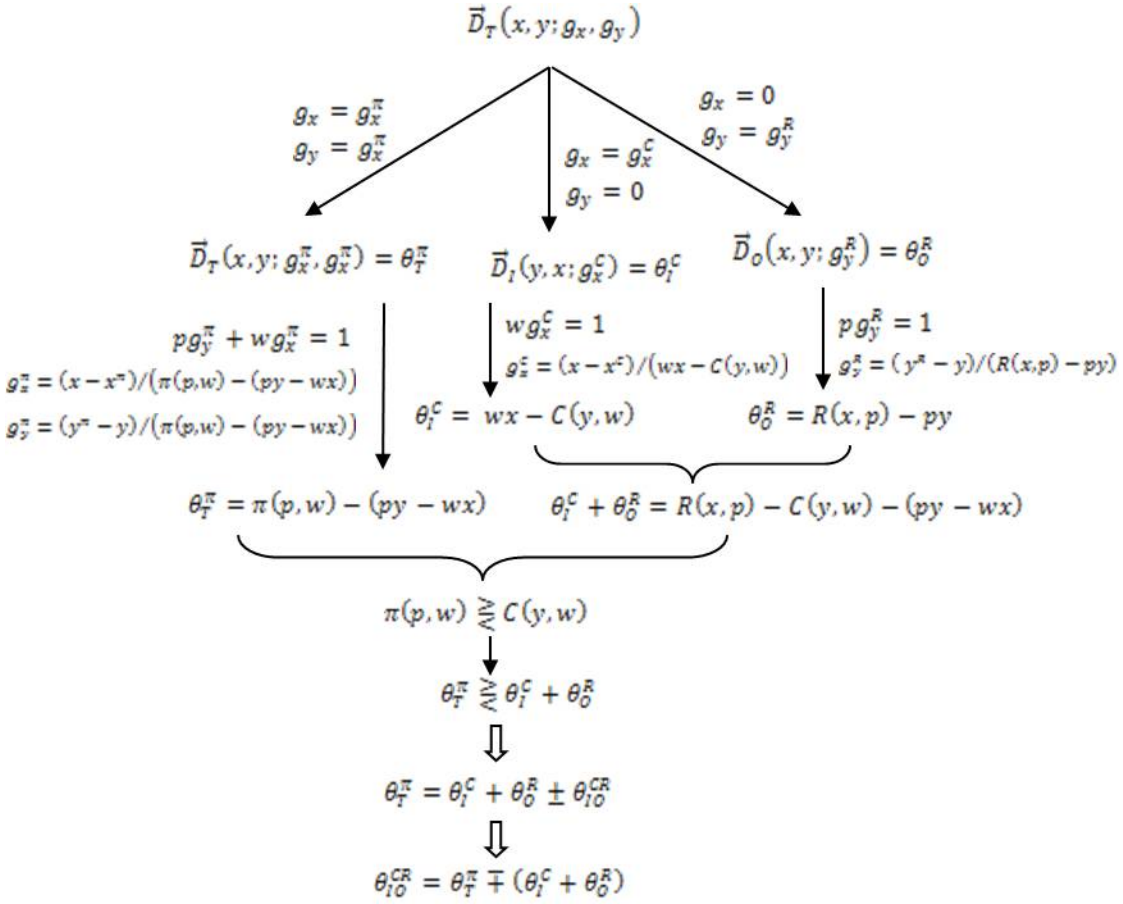

Figure 2.8: Interactive Effects with Endogenous Directional Vectors

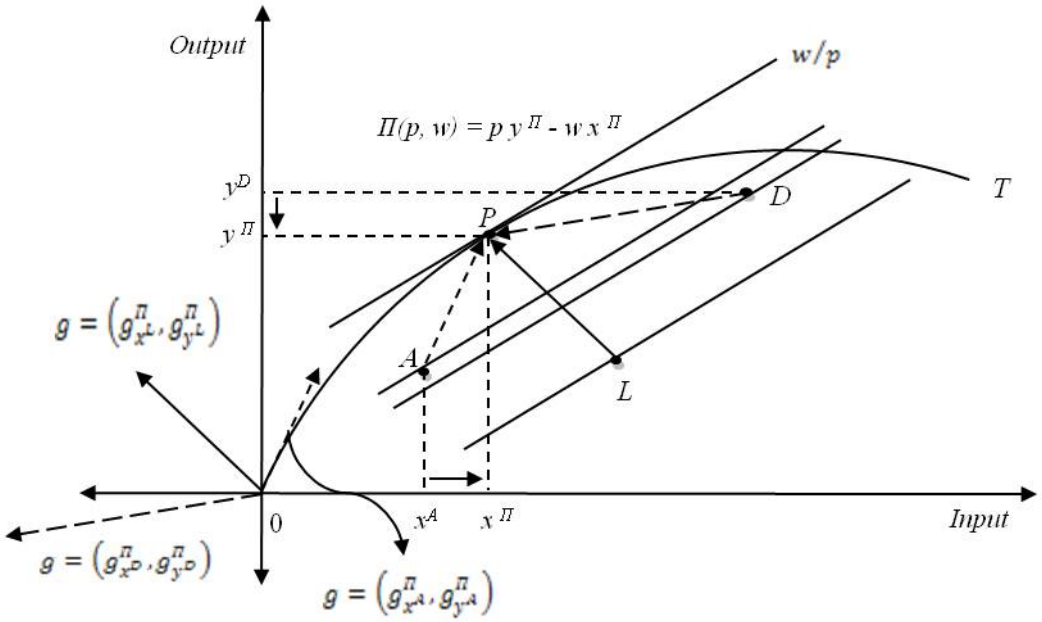

Figure 2.9: Endogenous Directional Vectors Projecting to the Profit-Maximizing Bundle

negative effects on the overall technical inefficiency, depending on the relationships between profits and costs. Given this observation, it would be interesting to examine the interactive effects between input and output technical inefficiencies empirically using this directional vector.

As noted by Zofio *et al.* (2013), the directional vector that projects any inefficient producer to the profit-maximizing bundle does not impose any sign restrictions on the adjustments of inputs and outputs. Figure 2.9 illustrates these directional vectors. As can be seen, these directional vectors may have negative components such that input is expanded (as, for example, from x^A to x^π) or output is contracted (as, for example, from y^D to y^π) to reach the frontier at the profit-maximizing benchmark P.

Similarly, the directional vector that projects any inefficient producer to the revenue-maximizing benchmark R may have negative components such that output is contracted (as, for example, from y_2^D to y_2^R) to reach the frontier, and the directional vector that projects any inefficient producer to the cost-minimizing benchmark C may have negative components such that input is expanded (as, for example, from x_2^D to x_2^C) to reach the frontier — see

Figure 2.10.

This paper shows that the derivation of the interactive effect between input and output technical inefficiencies from the directional distance function depends on the choice of the directional vector in which the data are projected on the frontier and whether quantities and prices are taken into consideration. When the directional vector includes no price information such as $g = (-x, y)$ or $g = (-1, 1)$, the interactive effects between input and output technical inefficiencies have negative effects on overall technical inefficiency and consequently lower overall technical inefficiency. When the directional vector includes price information, such as the directional vector that projects any inefficient producer to the profit-maximizing bundle, the interactive effects between input and output technical inefficiencies may have positive or negative effects on overall technical inefficiency, depending on the relationships between profits and costs.

From a theoretical perspective, this argument solves the arbitrary decomposition of overall technical inefficiency into input and output components. Overall technical inefficiency does not equal the sum of input and output technical inefficiencies as previous studies claim; it equals the sum of input and output technical inefficiencies plus an interactive effect component which captures the interactions between them. This suggests that the overuse of inputs creates input technical inefficiency and has an effect on output technical inefficiency. Also, the loss of output creates output technical inefficiency and has an effect on input technical inefficiency. From an applied perspective, producers either reduce inputs (costs), increase output (revenue) or reduce inputs and increase outputs (increase profit), depending on their objectives. However, the adjustability of both inputs and outputs is required for the improvement of producer overall efficiency, and no component of technical inefficiency should be ignored while improving efficiency since the interactive effects between input and output technical inefficiencies have an effect on the overall technical inefficiency.

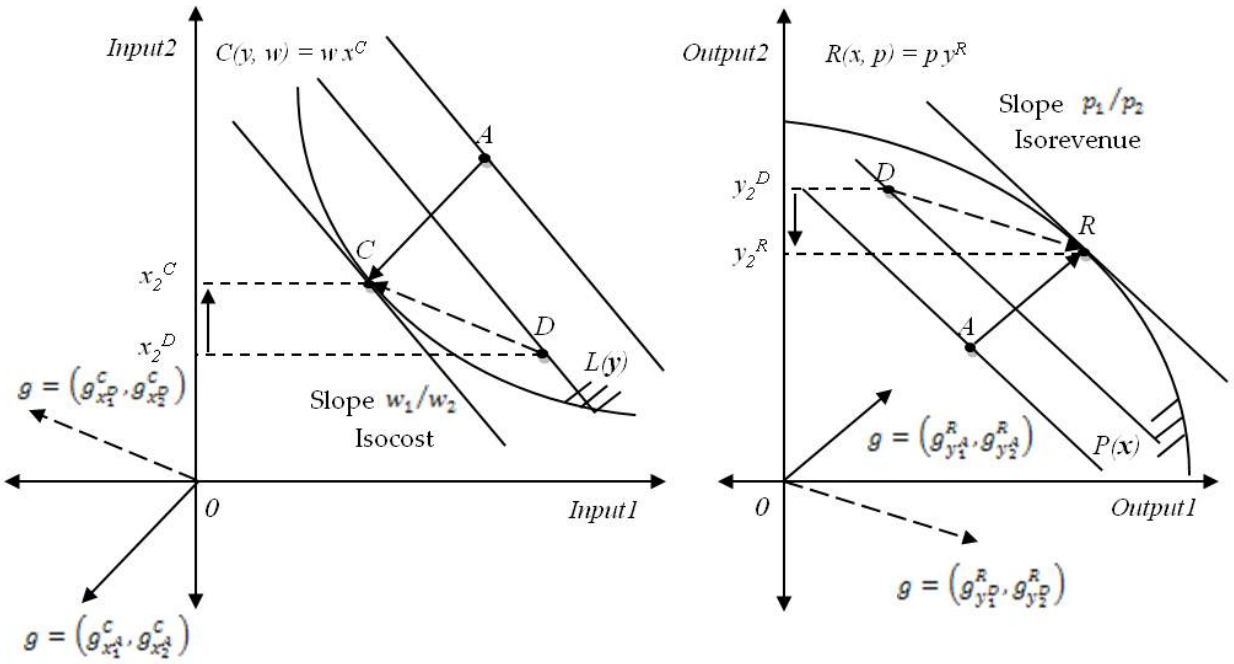

Figure 2.10: Endogenous Directional Vectors Projecting to the Cost-Minimizing or Revenue-Maximizing Bundle

2.5 Numerical Illustration

Consider seven hypothetical producers $P_1 - P_7$ who use one input to produce one output. Table 2.1 presents the data used in Figure 2.11 and the corresponding projection of the input-output vectors based on their associated directional vectors. For simplicity and convenience in measuring technical inefficiency, the pre-specified directional vector $g = (-1, 1)$ is used.

Producers 1, 2 and 3 who operate on the boundary of the production frontier are overall, input, and output technically efficient since their associated directional technology, input, and output distance functions in the feasible directions are zero; $\vec{D}_T(x, y; -1, 1) = \theta_T^1 = 0$, $\vec{D}_I(y, x; -1) = \theta_I^1 = 0$, and $\vec{D}_O(x, y; 1) = \theta_O^1 = 0$. However, producers 4, 5, 6, and 7 who operate beneath the production frontier are overall, input, and output technically inefficient since their associated directional technology, input, and output distance functions in the feasible directions are positive.

For example, producer 6 is overall technically inefficient since $\vec{D}_T(x, y; -1, 1) = \theta_T^1 =$

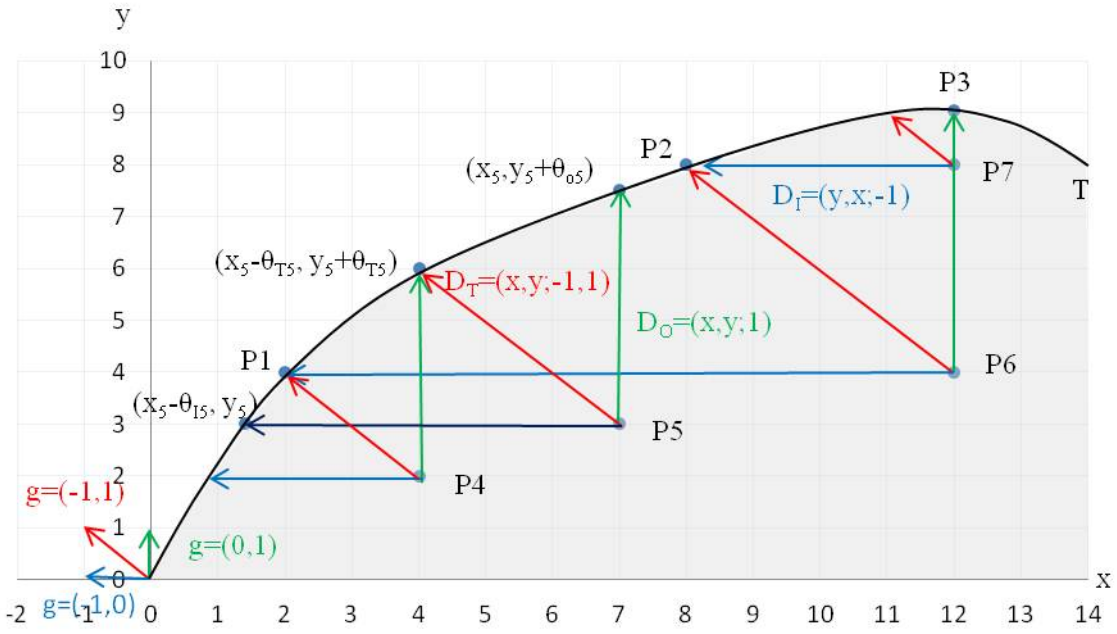

Figure 2.11: Graphical Representation of the Numerical Example Given in Table 2.1

$\|P6P2\| \,/\, \|0g\| = \sqrt{32}/\sqrt{2} = 4 > 0$ — see row 3 of Table 2.1. Substituting θ_T^1, the projected point $\left(x^T, y^T\right)$ is obtained as $\left(x^T, y^T\right) = (x - \theta_T^1 g_x, y + \theta_T^1 g_y) = (12 - 4 \times 1,\, 4 + 4 \times 1) = (8, 8)$ which is the same as the coordinates of producer 2 who operates on the boundary of the production frontier and is technically efficient.

If the direction vector $g = (-1, 0)$ is used, producer 6 is input technically inefficient since $\vec{D}_I\,(y, x; -1) = \theta_I^1 = x - x^I = 12 - 2 = 10 > 0$ — see row 5 of Table 2.1. Substituting θ_I^1, the projected point $\left(x^I, y\right)$ is obtained as $\left(x^I, y\right) = (x - \theta_I^1 g_x, y) = (12 - 10 \times 1,\, 4) = (2, 4)$ which is the same as the coordinates of producer 1 who operates on the boundary of the production frontier and is technically efficient.

Similarly, if the direction vector $g = (0, 1)$ is used, producer 6 is output technically inefficient since $\vec{D}_O\,(x, y; 1) = \theta_O^1 = y^O - y = 9 - 4 = 5 > 0$ — see row 7 of Table 2.1. Substituting θ_O^1, the projected point $\left(x, y^O\right)$ is obtained as $\left(x, y^O\right) = (x, y + \theta_O^1 g_y) = (12,\, 4 + 5 \times 1) = (12, 9)$ which is the same as the coordinates of producer 3 who operates on the boundary of the production frontier and is technically efficient. Following the same

procedure as above for producers 4, 5, and 7, the results are reported in Table 2.1.

Table 2.1 also shows the derivation of the interactive effects between input and output-oriented technical inefficiencies θ^1_{IO}. Setting $g_y = 0$, the directional vector $g = (-1, 0)$ allows for input contraction, holding output fixed. The directional input distance function $\vec{D}_I(y, x; -1)$ serves as an input-oriented measure of technical inefficiency $\theta^1_I = x - x^I$. Setting $g_x = 0$, the directional vector $g = (0, 1)$ allows for only output expansion, holding input fixed. The directional output distance function $\vec{D}_O(x, y; 1)$ serves as an output-oriented measure of technical inefficiency $\theta^1_O = y^O - y$. Row 11 of Table 2.1 shows the sum of input and output-oriented technical inefficiencies $\theta^1_I + \theta^1_O$. Row 12 shows the relationship between overall technical inefficiency θ^1_T and the sum of input and output-oriented technical inefficiencies $\theta^1_I + \theta^1_O$. The last four rows show the derivation of the interactive effects between input and output-oriented technical inefficiencies θ^1_{IO} with the use of proposition 2 and corollary 1. Producers 1, 2 and 3 who operate on the production frontier have zero interactive effects since $\theta^1_T = \theta^1_I + \theta^1_O = 0$. However, producers 4, 5, 6, and 7 who operate beneath the production frontier have negative interactive effects since $\theta^1_T < \theta^1_I + \theta^1_O$. The inequality can be turned into equality by subtracting a residual term that captures the interactive effect between input and output-oriented technical inefficiencies where the interactive effect θ^1_{IO} is defined as the gap in the inequality, namely $\theta^1_{IO} = \theta^1_I + \theta^1_O - \theta^1_T$. The interactive effect θ^1_{IO} can also be defined by using proposition 2 and corollary 1, namely $\theta^1_{IO} = \theta^1_T + \left(x^T - x^I\right) + \left(y^O - y^T\right)$, $\theta^1_{IO} = \theta^1_I + \left(y^O - y^T\right)$, or $\theta^1_{IO} = \theta^1_O + \left(x^T - x^I\right)$.

2.6 Conclusion

This paper presents theoretical and illustrative methods to derive the interactive effect between input and output technical inefficiencies using directional distance functions. This derivation solves the arbitrary decomposition of overall technical inefficiency into input and output components. The results show that overall technical inefficiency does not equal the

sum of input and output technical inefficiencies as previous studies claim. It equals the sum of input and output technical inefficiencies plus an interactive effect component which captures the interactions between them. I prove the results using the relationship between the directional distance functions and both the standard distance functions and their dual representations; cost, revenue, and profit functions. These results suggest that ignoring the interactive effect between input and output technical inefficiencies results in a decomposition of overall technical inefficiency into input and output components that are significantly different from the ones that incorporate it.

Table 2.1: Results for the Numerical Example

P	P_1	P_2	P_3	P_4	P_5	P_6	P_7
(x, y)	$(2, 4)$	$(8, 8)$	$(12, 9)$	$(4, 2)$	$(7, 3)$	$(12, 4)$	$(12, 8)$
θ_T^1	0	0	0	$\frac{\sqrt{8}}{\sqrt{2}} = 2$	$\frac{\sqrt{18}}{\sqrt{2}} = 3$	$\frac{\sqrt{32}}{\sqrt{2}} = 4$	$\frac{\sqrt{2}}{\sqrt{2}} = 1$
$\left(x^T, y^T\right)$	$(2, 4)$	$(8, 8)$	$(12, 9)$	$(2, 4)$	$(4, 6)$	$(8, 8)$	$(11, 9)$
$\theta_I^1 = x - x^I$	0	0	0	3.25	5.5	10	4
(x^I, y)	$(2, 4)$	$(8, 8)$	$(12, 9)$	$(0.75, 2)$	$(1.5, 3)$	$(2, 4)$	$(8, 8)$
$\theta_O^1 = y^O - y$	0	0	0	4	4.5	5	1
(x, y^O)	$(2, 4)$	$(8, 8)$	$(12, 9)$	$(4, 6)$	$(7, 7.5)$	$(12, 9)$	$(12, 9)$
$I = x^T - x^I$	0	0	0	1.25	2.5	6	3
$O = y^O - y^T$	0	0	0	2	1.5	1	0
$IO = \theta_I^1 + \theta_O^1$	0	0	0	7.25	10	15	5
$\theta_T^1 \leq \theta_I^1 + \theta_O^1$	$0 = 0$	$0 = 0$	$0 = 0$	$2 < 7.25$	$3 < 10$	$4 < 15$	$1 < 5$
$\theta_{IO}^1 = IO - \theta_T^1$	0	0	0	5.25	7	11	4
$\theta_{IO}^1 = \theta_T^1 + I + O$	0	0	0	5.25	7	11	4
$\theta_{IO}^1 = \theta_I^1 + O$	0	0	0	5.25	7	11	4
$\theta_{IO}^1 = \theta_O^1 + I$	0	0	0	5.25	7	11	4

Chapter 3

Interactive Effects between Input and Output Technical Inefficiencies in US Commercial Banking

3.1 Introduction

Most empirical studies that examine the inefficiency of production processes in banking employ either an input or an output-oriented measurement technique. In terms of the former, researchers assume that outputs are exogenous and inputs endogenous and banks are fully capable of reallocating resources when improving efficiency. A bank is an input technically efficient if it is capable of using minimal inputs to produce a given vector of outputs. Similarly, by adopting an output-oriented measurement technique, it is assumed that inputs are exogenous and outputs endogenous and banks are fully capable of mixing production when improving efficiency. An output technically efficient bank can produce maximal output from a given vector of inputs. However, adopting an input (output) oriented measurement technique ignores the opposite output (input) orientation, and this restriction may substantially bias the measures of technical inefficiency.

An efficiency survey by Berger, Hunter, and Timme (1993) suggests comparing these input and output approaches with a complete approach to investigate the relationships between input and output inefficiencies. However, few studies examine total technical inefficiency and decompose it into input and output components either by using a profit function or a directional technology distance function. Berger, Hancock, and Humphrey (1993) and Akhavein *et al.* (1997) apply a distribution-free approach and show no interactive effects between input and output technical inefficiencies by using a profit function. More recently, Barros *et al.* (2012) and Fujii *et al.* (2014) apply a nonparametric approach and decompose total

technical inefficiency into input and output components by using a weighted directional distance function that takes into account the contribution of each input and output on total technical inefficiency.

Even though these studies disaggregate and quantify the impact of input and output on inefficiency, the arbitrary decomposition of total technical inefficiency into input and output inefficiency components results in concluding that total technical inefficiency equals the sum of input and output technical inefficiencies and shows no interactive effects between them.

In contrast to these studies, this paper follows Berger, Hunter, and Timme (1993) suggestion and uses a complete approach, investigating the relationships among input, output, and overall technical inefficiencies in terms of the directional distance functions. Specifically, input, output and technology-oriented technical inefficiencies are estimated separately using directional input, output, and technology distance functions, respectively.

A potential issue when estimating technical inefficiency using directional distance functions is that inputs and outputs may be endogenous, meaning that they are correlated with the random errors or inefficiency or both and leading to biased and inconsistent estimates of the parameters of the production technology and the associated measures of inefficiency — see, for example, Atkinson and Primont (2002), Atkinson *et al.* (2003), and O'Donnell (2014). The literature considers two approaches to deal with this issue; one approach relies on using instrumental variable estimation, and the other relies on employing a system of equations approach. This paper follows the latter approach. Specifically, input, output, and technology-oriented technical inefficiencies are estimated separately using systems of equations, consisting of directional input (output, and technology) distance function with the cost (revenue, and profit) minimizing (maximizing) first-order conditions, respectively. Furthermore, the directional vectors of these models are allowed to be endogenous and vary across banks to account for heterogeneity across banks. The obtained estimates of the directional vectors can be interpreted as being optimal directional vectors — see Malikov *et al.* (2016).

The input, output, and overall technical inefficiencies are estimated with the three commonly used directional vectors; the unit value, the observed input-output, and the optimal directional vectors. The unit value and the observed input-output directional vector models are estimated without additional first-order condition equations. The optimal directional vector models are estimated using systems of equations, consisting of directional distance functions with the relevant first-order conditions.

To investigate the relationships among input, output, and overall technical inefficiencies, I model the overall technical inefficiency as a linear function of a vector of explanatory bank-specific variables that includes input technical inefficiency, output technical inefficiency, and a term capturing the interactions between them, following Battese and Coelli (1995). These bank-specific variables that determine overall technical inefficiency are estimated simultaneously with the variables that determine the frontier.

The above methodology is applied to a sample of 148 US commercial banks over the period from 2001 to 2015. The market-average prices faced and determined exogenously rather than the actual prices paid or received by the bank are used, following Berger and Mester (2003). These market-average prices are more likely to be exogenous to the bank than the bank-specific prices. The bank market-average price at a given year is the weighted average of the other banks prices at that year excluding the bank-specific price, where the weights are each bank respective market share at that year. To select the relevant variables, the commonly accepted asset approach proposed by Sealey and Lindley (1977) is used. It defines loans and other assets as outputs, while deposits and other liabilities are treated as inputs.

Regarding regularity violations, I find that the monotonicity conditions with respect to labor and all outputs are violated for all models at most observations. Therefore, all models are re-estimated with the monotonicity conditions imposed at each observation, by following the Bayesian procedure discussed in O'Donnell and Coelli (2005). Bayesian approach is

used, mainly because this approach can easily impose monotonicity conditions and estimate directional vectors that vary across banks.

This paper contributes to the literature in many ways. First, to the best of my knowledge, it is the first in the literature that uses a complete approach to examine the relationships among input, output, and overall technical inefficiencies empirically using the same data set and the directional input, output, and technology distance functions with the three commonly used directional vectors; the unit value, the observed input-output, and the optimal directional vectors. Second, the optimal directional vectors are allowed to be endogenous and vary across banks to account for heterogeneity across banks. Third, it pays explicit attention to the theoretical regularity conditions in order to produce inference that is consistent with neoclassical microeconomic theory.

In providing a comparison of the three estimates of technical inefficiencies in the case of the unit value, the observed input-output, and the optimal directional vector models, the empirical results show that overall technical inefficiency does not equal the sum of input and output technical inefficiencies, as previous studies claim. It equals the sum of input and output technical inefficiencies plus an interactive effect component which captures the interactions between them, where the increase in the output technical inefficiency reflects on a reduction on the input technical inefficiency and vice versa.

The results also show that both input and output technical inefficiencies have significant positive effects on the overall technical inefficiency. However, the interactive effect between input and output technical inefficiencies has a significant negative effect on the overall technical inefficiency. This result is robust to alternative directional vectors and model specifications. Banks with larger values of the interactive effect tend to have a lower level of overall technical inefficiency which indicates that they are more efficient. This suggests that the overuse of inputs creates input technical inefficiency and has an effect on reducing (improving) output technical inefficiency (efficiency) and therefore improving over-

all technical efficiency. Intuitively, the overuse of inputs whether physical inputs involving overuse of labor or overuse of financial inputs involving overpayment of interest creates input technical inefficiency and has an effect on reducing (improving) output technical inefficiency (efficiency). The overuse of labor and the overpayment of interest may encourage banks to produce more loans to pay salaries for its employees and interest rates on deposits.

Similarly, the loss of production of outputs creates output technical inefficiency and has an effect on reducing (improving) input technical inefficiency (efficiency) and therefore improving overall technical efficiency. Intuitively, the loss of production of loans creates output technical inefficiency and has an effect on reducing (improving) input technical inefficiency (efficiency). The loss of revenue from loans may encourage banks to reduce the number of labor used in the production process or lower the interest rates paid on deposits. Ignoring the interactive effect between input and output technical inefficiency results in a decomposition of overall technical inefficiency into input and output components that are significantly different from the ones that incorporate it.

The results also indicate that the value of the interactive effect between input and output technical inefficiencies depends on the choice of the directional vector in which the data are projected on the frontier and whether quantities and prices are taken into consideration. These results are quite significant, since these inefficiency components have different implications for bank performance, suggesting that the adjustability of both inputs and outputs is required for the improvement of bank efficiency.

The rest of the paper is organized as follows. The next section provides a brief review of directional distance functions that are used to measure input, output, and overall technical inefficiencies. Section 3 specifies the unit value, the observed input-output, and the optimal directional vector models and the interactive effects equation. Section 4 discusses the Bayesian procedure for estimating these models. Section 5 defines the data used in this paper. In Section 6, the methodology is applied to a sample of US commercial banks, and

the results are reported. Section 7 summarizes and concludes the paper.

3.2 The Directional Technology Distance Function

To model the banking production process and measure overall, input, and output technical inefficiencies, the directional technology, input, and output distance functions proposed by Chambers *et al.* (1998) are used. Consider a bank that uses $x \in R_+^N$ inputs to produce $y \in R_+^M$ outputs. This banking production process can be represented by the technology T; that is defined as the set of all feasible input-output vectors.[1]

$$T = \left\{(x, y) : x \in R_+^N, y \in R_+^M, \ x \text{ can produce } y\right\}$$

The directional technology distance function (DTDF) completely characterizes technology; that is, it is equivalent to T. It allows for simultaneous contraction of inputs and expansion of outputs in terms of an explicit direction vector $g = (g_x, g_y)$, where $g_x \in R_+^N$ and $g_y \in R_+^M$ such that it contracts inputs in the direction g_x and expands outputs in the direction g_y. In particular, the DTDF is defined as

$$\vec{D}_T\left(x, y; g_x, g_y\right) = \max_{\theta_T} \left\{\theta_T : (x - \theta_T g_x, y + \theta_T g_y) \in T\right\} \tag{3.1}$$

The measure of technical inefficiency derived from the DTDF is technology-oriented and depends on the choice of the direction vector g in which the data are projected on the frontier. Note that the DTDF constitutes an additive measure of technical inefficiency in a given direction g, where the zero value of $\vec{D}_T\left(x, y; g_x, g_y\right)$ implies full technological efficiency. Inefficiency is indicated by $\vec{D}_T\left(x, y; g_x, g_y\right) > 0$ with higher values indicating greater inefficiency when banks operate beneath the frontier of T. A measure of technical inefficiency is defined as

$$TI_T = \vec{D}_T\left(x, y; g_x, g_y\right)$$

[1]Standard properties are assumed on the technology T. See Chambers (1998) and Chambers *et al.* (1998). These properties include the axioms of the possibility of inaction; $(0, 0) \in T$, no free lunch; if $(x, y) \in T$ and $x = 0$ then $y = 0$, and free disposability of inputs and outputs; if $(x, y) \in T$ and $x' \geq x$, $y' \leq y$ then $(x', y) \in T$ and $(x, y') \in T$. It is also assumed that the technology T is closed and convex.

As noted by Chambers *et al.* (1998), the DTDF is non-negative, non-decreasing in x, non-increasing in y, and concave in (x, y). Moreover, it satisfies the following translation property

$$\vec{D}_T\left(x - \alpha g_x, y + \alpha g_y; g_x, g_y\right) = \vec{D}_T\left(x, y; g_x, g_y\right) - \alpha \tag{3.2}$$

where α is an arbitrary scaling factor. This property says that if the input-output vector (x, y) is translated into $(x - \alpha g_x, y + \alpha g_y)$, then the value of the DTDF is reduced by α (for $\alpha \in R$).

3.2.1 The Directional Input Distance Function

The directional input distance function (DIDF) can be derived from $\vec{D}_T\left(x, y; g_x, g_y\right)$ by setting $g_y = 0$. The directional vector $g = (g_x, 0)$ allows only for input contraction holding outputs fixed. In this case, equation (3.1) becomes the DIDF that allows for only input contraction, $\vec{D}_T\left(x, y; g_x, 0\right) = \vec{D}_I\left(y, x; g_x\right)$

$$\vec{D}_I\left(y, x; g_x\right) = \max_{\theta_I}\left\{\theta_I : (x - \theta_I g_x) \in L(y)\right\} = \max_{\theta_I}\left\{\theta_I : (x - \theta_I g_x,\ y) \in T\right\} \tag{3.3}$$

where $L(y)$ is the input set which represents the set of all input vectors x which can produce the output vector y, that is

$$L(y) = \left\{x \in R_+^N : x \text{ can produce } y\right\}$$

The DIDF serves as an input-oriented measure of technical inefficiency.

$$TI_I = \vec{D}_I\left(y, x; g_x\right)$$

As noted by Chambers *et al.* (1996), the DIDF is non-negative, non-decreasing in x, non-increasing in y, and concave in inputs x. Moreover, it satisfies the following translation property

$$\vec{D}_I\left(y, x - \alpha g_x; g_x\right) = \vec{D}_I\left(y, x; g_x\right) - \alpha \tag{3.4}$$

where α is an arbitrary scaling factor. This property says that if the input vector x is translated into $(x - \alpha g_x)$, then the value of the DIDF is reduced by α (for $\alpha \in R$).

3.2.2 The Directional Output Distance Function

The directional output distance function (DODF) can be derived from $\vec{D}_T(x, y; g_x, g_y)$ by setting $g_x = 0$. The directional vector $g = (0, g_y)$ allows only for output expansion holding inputs fixed. In this case, equation (3.1) reduces to the DODF that allows for only output expansion, $\vec{D}_T(x, y; 0, g_y) = \vec{D}_O(x, y; g_y)$.

$$\vec{D}_O(x, y; g_y) = \max_{\theta_o} \{\theta_O : (y + \theta_O g_y) \in P(x)\} = \max_{\theta_o} \{\theta_O : (x, y + \theta_O g_y) \in T\} \tag{3.5}$$

where $P(x)$ is the output set which represents the set of all output vectors y which can be produced using the input vector x, that is

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

The DODF serves as an output-oriented measure of technical inefficiency

$$TI_O = \vec{D}_O(x, y; g_y)$$

As noted by Färe *et al.* (2005), the DODF is non-negative, non-decreasing in x, non-increasing in y, and concave in outputs y. Moreover, it satisfies the following translation property

$$\vec{D}_O(x, y + \alpha g_y; g_y) = \vec{D}_O(x, y; g_y) - \alpha$$

where α is an arbitrary scaling factor. This property says that if the output vector y is translated into $(y + \alpha g_y)$, then the value of the DODF is reduced by α (for $\alpha \in R$)

3.3 Model Specification

To obtain the estimates of the directional distance functions and therefore the measure of technical inefficiencies, this section provides the parametric specification of these functions. This involves choosing a functional form, imposing the parameter restrictions for the translation property, modeling the interactive effects, and specifying the directional vector g.

3.3.1 The Quadratic Functional Form

To parameterize the functions in (3.1), (3.3), and (3.5), the quadratic functional form is used following Chambers (1998). The reason for choosing this functional form is that it is a second-order Taylor series approximation which is linear in the parameters and flexible enough to provide an excellent approximation to the actual production technology. Furthermore, the directional distance function satisfies a translation property, which can be easily imposed on a quadratic functional form.

To avoid any estimation biases that may arise due to potential changes in bank performance due to technological change, technical change is incorporated by a trend variable, t, while non-neutral technical change is modeled by including terms capturing the interaction between trend and inputs and trend and outputs, as is common in the literature. Thus, the directional distance functions defined in (3.1), (3.3), and (3.5) can be rewritten as $\vec{D}_T(x, y, t; g_x, g_y)$, $\vec{D}_I(y, x, t; g_x)$, and $\vec{D}_O(x, y, t; g_y)$, respectively.

3.3.2 Imposing the Restrictions

The translation property can be imposed by imposing a set of parameter restrictions that applied to the directional distance function directly during the estimation and estimating the restricted version of the directional distance function — see, for example, Atkinson and Tsionas (2016).

Alternatively, the translation property can be imposed by setting α equal to an arbitrarily chosen input or the negative of an arbitrarily chosen output which is specific to each bank, say $\alpha = -y_1$, and normalizing the corresponding directional vector $g_{y1} = 1$ — see, for example, Malikov *et al.* (2016). Using this transformation process and applying it to the empirical implementation that uses two inputs to produce two outputs, the translation property in equation (3.2) can be rewritten as

$$\vec{D}_T(x + y_1 g_x, y_2 - y_1 g_{y_2}, t; g_x, g_y) = \vec{D}_T(x, y, t; g_x, g_y) + y_1 \tag{3.6}$$

Note that the output y_1 disappears from the left-hand side of (3.6) because of $y_1 - y_1(1) = 0$. Rearranging equation (3.6) yields

$$\begin{aligned} y_1 &= \vec{D}_T\left(x + y_1 g_x, y_2 - y_1 g_{y_2}, t; g_x, g_y\right) - \vec{D}_T\left(x, y, t; g_x, g_y\right) \\ &= \vec{D}_T\left(x + y_1 g_x, y_2 - y_1 g_{y_2}, t; g_x, g_y\right) - u_T \end{aligned} \tag{3.7}$$

Adding a random error v_T to equation (3.7) yields the standard stochastic frontier model with two error terms, as follows

$$y_1 = \vec{D}_T\left(x + y_1 g_x, y_2 - y_1 g_{y_2}, t; g_x, g_y\right) + v_T - u_T \tag{3.8}$$

where v_T is a two-sided random error assumed to be identically and independently distributed (iid) with mean zero and variance $\sigma^2_{v_T} = \Sigma, v_T \sim N\left(0, \Sigma\right)$ and $\vec{D}_T\left(x, y, t; g_x, g_y\right) = u_T \geq 0$ is a one-sided error term which captures bank-specific overall technical inefficiency. Applying the quadratic functional form to the first term on the right-hand side of (3.8), (3.8) can be written more explicitly as

$$\begin{aligned} y_1 = \alpha_0 &+ \sum_{n=1}^{2} \alpha_n \widetilde{x}_n + \beta_2 \widetilde{y}_2 + \delta_t t + \frac{1}{2}\sum_{n=1}^{2}\sum_{n'=1}^{2} \alpha_{nn'} \widetilde{x}_n \widetilde{x}_{n'} + \frac{1}{2}\beta_{22}\left(\widetilde{y}_2\right)^2 \\ &+ \frac{1}{2}\delta_{tt} t^2 + \sum_{n=1}^{2} \gamma_{n2} \widetilde{x}_n \widetilde{y}_2 + \sum_{n=1}^{2} \delta_{tx_n} t \widetilde{x}_n + \delta_{ty_2} t \widetilde{y}_2 + v_T - u_T \end{aligned} \tag{3.9}$$

where $\widetilde{x}_n = x_n + y_1 g_{x_n}$ $(n = 1, 2)$, and $\widetilde{y}_2 = y_2 - y_1 g_{y_2}$, and y_1 corresponds to the dependent variable. The parameters of (3.9) must satisfy a set of parameter restrictions, including the usual restrictions for symmetry $\alpha_{nn'} = \alpha_{n'n}$ $(n \neq n')$ and $\beta_{mm'} = \beta_{m'm}$ $(m \neq m')$, and the following set of parameter restrictions that impose the translation property — see the Appendix for proof.

$$\beta_1 + \beta_2 g_{y_2} - \sum_{n=1}^{2} \alpha_n g_{x_n} = -1, \qquad \gamma_{n1} + \gamma_{n2} g_{y_2} - \sum_{n'=1}^{2} \alpha_{nn'} g_{x_{n'}} = 0,$$

$$\beta_{21} + \beta_{22} g_{y2} - \sum_{n=1}^{2} \gamma_{n2} g_{x_n} = 0, \quad \text{and} \quad \delta_{ty1} + \delta_{ty2} g_{y_2} - \sum_{n=1}^{2} \delta_{tx_n} g_{x_n} = 0$$

Within a panel data framework, the DTDF model in (3.9) can be notationally simplified as

$$y_{1,it} = \widetilde{R}_{T,it}(g)'\beta_T + v_{T,it} - u_{T,it} \tag{3.10}$$

where $i = 1, ..., K$ indicates banks; $t = 1, ..., T$ indicates time; $\widetilde{R}_T(g)$ is a vector of all the relevant variables on the right-hand side of (3.9) including a unity for the intercept term; and β_T is the corresponding vector of coefficients (including the intercept). Note that $\widetilde{R}_{T,it}(g)$ is a function of g where $g = (g_x, g_y)$. Formally, the dependent variable $y_{1,it} = [y_{1,11}, ..., y_{1,1T}, ..., y_{1,K1}, ...y_{1,KT}]'$, the vector of all the relevant variables on the right-hand side of (3.9), $\widetilde{R}_T(g) = \left[1\ \widetilde{x}_1\ \widetilde{x}_2\ \widetilde{y}_2\ t\ (\widetilde{x}_1)^2\ \widetilde{x}_1\widetilde{x}_2\ (\widetilde{x}_2)^2\ (\widetilde{y}_2)^2\ t^2\ \widetilde{x}_1\widetilde{y}_2\ \widetilde{x}_2\widetilde{y}_2\ t\widetilde{x}_1\ t\widetilde{x}_2\ t\widetilde{y}_2\right]'$, the vector of coefficients $\beta_T = [\alpha_0\ \alpha_1\ \alpha_2\ \beta_2\ \delta_t\ \alpha_{11}\ \alpha_{12}\ \alpha_{22}\ \beta_{22}\ \delta_{tt}\ \gamma_{12}\ \gamma_{22}\ \delta_{tx_1}\ \delta_{tx_2}\ \delta_{ty_2}]$, the vector of the random error $v_{T,it} = [v_{T,11}, ..., v_{T,1T}, ..., v_{T,K1}, ...v_{T,KT}]'$, and the vector of overall technical inefficiency $u_{T,it} = [u_{T,11}, ..., u_{T,1T}, ..., u_{T,K1}, ...u_{T,KT}]'$.

Similarly, the translation property of the DIDF in equation (3.4) can be imposed by setting α equal to an arbitrarily chosen input which is specific to each bank, say $\alpha = x_1$, and normalizing the corresponding directional vector $g_{x_1} = 1$. Using this transformation process, the translation property in equation (3.4) can be rewritten as

$$\vec{D}_I(y, x_2 - x_1 g_{x_2}; g_x) = \vec{D}_I(y, x; g_x) - x_1 \tag{3.11}$$

Note that the input x_1 disapears from the left-hand side of (3.11) because of $x_1 - x_1(1) = 0$. Rearranging equation (3.11) yields

$$\begin{aligned} -x_1 &= \vec{D}_I(y, x_2 - x_1 g_{x_2}; g_x) - \vec{D}_I(y, x; g_x) \\ &= \vec{D}_I(y, x_2 - x_1 g_{x_2}; g_x) - u_I \end{aligned} \tag{3.12}$$

where $\vec{D}_I(y, x; g_x) = u_I \geq 0$ represents bank-specific input technical inefficiency. Adding a random error v_I to equation (3.12) yields the standard stochastic frontier model with two error terms, as follows

$$-x_1 = \vec{D}_I(y, x_2 - x_1 g_{x_2}; g_x) + v_I - u_I \tag{3.13}$$

Applying the quadratic functional form to the first term on the right-hand side of (3.13), (3.13) can be written more explicitly as

$$-x_1 = \alpha_0 + \alpha_2\widetilde{x}_2 + \sum_{m=1}^{2}\beta_m y_m + \delta_t t + \frac{1}{2}\alpha_{22}\left(\widetilde{x}_2\right)^2 + \frac{1}{2}\sum_{m=1}^{2}\sum_{m'=1}^{2}\beta_{mm'}y_m y_{m'} + \frac{1}{2}\delta_{tt}t^2 + \sum_{m=1}^{2}\gamma_{2m}\widetilde{x}_2 y_m + \delta_{tx_2}t\widetilde{x}_2 + \sum_{m=1}^{2}\delta_{ty_m}ty_m + v_I - u_I \tag{3.14}$$

where $\widetilde{x}_2 = x_2 - x_1 g_{x_2}$, x_1 corresponds to the dependent variable, $u_I \geq 0$ is a one-sided error term which captures input technical inefficiency. The parameters of (3.14) must satisfy a set of parameter restrictions, including the usual restrictions for symmetry $\alpha_{nn'} = \alpha_{n'n}$ $(n \neq n')$ and $\beta_{mm'} = \beta_{m'm}$ $(m \neq m')$, and the following set of parameter restrictions that impose the translation property — see the Appendix for proof.

$$\alpha_1 + \alpha_2 g_{x_2} = 1, \qquad \gamma_{1m} + \gamma_{2m} g_{x_2} = 0, \quad \alpha_{11} - \alpha_{22} g_{x2}^2 = 0,$$

$$\alpha_{21} + \alpha_{22} g_{x_2} = 0, \quad \text{and} \quad \delta_{tx_1} + \delta_{tx_2} g_{x_2} = 0, \qquad (m = 1, 2)$$

Within a panel data framework, the DIDF model in (3.14) can be notationally simplified as

$$-x_{1,it} = \widetilde{R}_{I,it}(g)'\beta_I + v_{I,it} - u_{I,it} \tag{3.15}$$

where $\widetilde{R}_I(g)$ is a vector of all the relevant variables on the right-hand side of (3.14) including a unity for the intercept term; and β_I is the corresponding vector of coefficients (including the intercept).

When $g_x = 0$, the DTDF model in (3.9) reduces to the DODF that allows for only output expansion

$$y_1 = \alpha_0 + \sum_{n=1}^{2}\alpha_n x_n + \beta_2\widetilde{y}_2 + \delta_t t + \frac{1}{2}\sum_{n=1}^{2}\sum_{n'=1}^{2}\alpha_{nn'}x_n x_{n'} + \frac{1}{2}\beta_{22}\left(\widetilde{y}_2\right)^2 + \frac{1}{2}\delta_{tt}t^2 + \sum_{n=1}^{2}\gamma_{n2}x_n\widetilde{y}_2 + \sum_{n=1}^{2}\delta_{tx_n}tx_n + \delta_{ty_2}t\widetilde{y}_2 + v_O - u_O \tag{3.16}$$

where $\widetilde{y}_2 = y_2 - y_1 g_{y_2}$, y_1 corresponds to the dependent variable, $u_O \geq 0$ is a one-sided error term which captures output technical inefficiency. The parameters of (3.16) must satisfy

a set of parameter restrictions, including the usual restrictions for symmetry $\alpha_{nn'} = \alpha_{n'n}$ $(n \neq n')$ and $\beta_{mm'} = \beta_{m'm}$ $(m \neq m')$, and the following set of parameter restrictions that impose the translation property — see the Appendix for proof.

$$\beta_1 + \beta_2 g_{y_2} = -1, \qquad \gamma_{n1} + \gamma_{n2} g_{y_2} = 0, \quad \beta_{11} - \beta_{22} g_{y_2}^2 = 0,$$

$$\beta_{21} + \beta_{22} g_{y2} = 0, \quad \text{and} \quad \delta_{ty1} + \delta_{ty2} g_{y_2} = 0, \qquad (n = 1, 2)$$

Within a panel data framework, the DODF model in (3.16) can be notationally simplified as

$$y_{1,it} = \widetilde{R}_{O,it}(g)' \beta_O + v_{O,it} - u_{O,it} \tag{3.17}$$

where $\widetilde{R}_O(g)$ is a vector of all the relevant variables on the right-hand side of (3.16) including a unity for the intercept term; and β_O is the corresponding vector of coefficients (including the intercept).

3.3.3 Modeling the Interactive Effects

Following Battese and Coelli (1995), overall technical inefficiency, $u_{T,it}$, can be modeled as a linear function of a vector of explanatory bank-specific variables Z_{it} that are expected to influence $u_{T,it}$. Bank-specific variables Z_{it} include input technical inefficiency, $u_{I,it}$, output technical inefficiency, $u_{O,it}$, and a term capturing the interactions between input and output technical inefficiencies, $u_{I,it} \times u_{O,it}$

$$u_{T,it} = Z_{it}\delta + v_{u,it} \tag{3.18}$$

where δ is an unknown vector of coefficients (including the intercept) to be estimated, and $v_{u,it}$ is an error term that is defined by the truncation of a normal distribution.

3.3.4 Specifying the Directional Vector

In specifying the directional vector, there are two approaches in the literature. The first is to choose the directional vector a priori. The second approach is to let the data determine

the directional vector in which the bank's movement toward the efficient frontier is to be estimated. Specifically, the directional vector is treated as unknown parameters which are to be estimated. In this paper, both approaches are used and compared.

The Unit Value Directional Vector

When only quantity information on inputs and outputs is available, and price information is unavailable, distorted or inaccurate, technical inefficiency can be measured by choosing a pre-specified directional vector such that it projects any inefficient producer to the frontier of T. An example of a pre-specified directional vector is the unit value direction $g = (-1, 1)$ — see, for example, Park and Weber (2006), and Koutsomanoli-Filippaki *et al.* (2009). This type of directional vector implies that the amount by which a bank can decrease inputs and increase outputs will be $\vec{D}_T(x, y; -1, 1) \times 1$ units of x and y.

Input-oriented measures of technical inefficiency $u_{I,it}$ using the unit value direction can be obtained by setting $g_x = 1$ in the case of (3.15) and estimate the single equation DIDF subject to the usual symmetry restrictions and theoretical monotonicity restrictions. Similarly, output-oriented measures of technical inefficiency $u_{O,it}$ using the unit value direction can be obtained by setting $g_y = 1$ in the case of (3.17) and estimate the single equation DODF subject to the usual symmetry restrictions and theoretical monotonicity restrictions. Overall or technology-oriented measures of technical inefficiency $u_{T,it}$ using the unit value direction can be obtained by setting $g = (g_x, g_y) = (-1, 1)$ in the case of (3.10) and estimate the system of equations that includes the DTDF and the interactive effects equation in (3.18) subject to the usual symmetry restrictions and theoretical monotonicity restrictions, while the translation property is already imposed by construction. More specifically, the system can be written as

$$\begin{bmatrix} y_{1,it} \\ u_{T,it} \end{bmatrix} = \begin{bmatrix} \widetilde{R}_{T,it}(g)' & \\ & Z'_{it} \end{bmatrix} [\beta] - \begin{bmatrix} u_{T,it} \\ 0 \end{bmatrix} + \begin{bmatrix} v_{T,it} \\ v_{u,it} \end{bmatrix}$$

which can be written in a compressed form as

$$Y_{it} = R_{it}(g)\beta - u_{T,it}\iota + v_{it} \tag{3.19}$$

where $\iota = [1, 0]$ and $v_{it} = (v_{T,it}, v_{u,it})' \sim N(0, \Sigma)$. Bank-specific variables Z is a vector of the relevant variables on the right-hand side of (3.18) that are obtained using the unit value direction in equations (3.15) and (3.17). In this system of equations, bank-specific variables Z that determine overall technical inefficiency are estimated simultaneously with the variables that determine the frontier.

The Observed Input-Output Directional Vector

Another widely used pre-specified direction is the observed input-output direction $g = (-x, y)$ — see, for example, Färe, Grosskopf, and Weber (2004). This type of directional vector implies that a bank can decrease inefficiency while decreasing inputs and increasing outputs in proportion to the initial combination of the actual inputs and outputs.

Similar to the unit value directional vector, input-oriented measures of technical inefficiency $u_{I,it}$ using the observed input directional vector can be obtained by setting $g_{x_2} = x_2$ in the case of (3.15) and estimate the single equation DIDF subject to the usual symmetry restrictions and theoretical monotonicity restrictions. Output-oriented measures of technical inefficiency $u_{O,it}$ using the observed output directional vector can be obtained by setting $g_{y_2} = y_2$ in the case of (3.17) and estimate the single equation DODF subject to the usual symmetry restrictions and theoretical monotonicity restrictions. Overall or technology-oriented measures of technical inefficiency $u_{T,it}$ using the observed input-output directional vector can be obtained by setting $g_{x_1} = x_1$, $g_{x_2} = x_2$, and $g_{y_2} = y_2$ in the system of equations (3.19). Bank-specific variables Z is a vector of the relevant variables on the right-hand side of (3.18) that are obtained using the observed input and output directional vectors defined in the single equations in (3.15) and (3.17), respectively. The DTDF system is estimated subject to the symmetry restrictions and theoretical monotonicity restrictions, while the translation property is already imposed by construction.

The Optimal Directional Vector

When information on input and output prices is available, and the banking industry is assumed to exhibit cost-minimizing (or revenue or profit-maximizing) behavior, technical inefficiency can be measured by choosing an endogenous direction vector such that it projects any inefficient bank to the cost-minimizing (or revenue or profit-maximizing) benchmark. Therefore, the directional vector is treated as a parameter to be estimated — see, for example, Malikov *et al.* (2016) for a cost-optimal directional vector.

The Cost-Optimal Directional Vector

The bank cost-minimizing objective is defined as

$$C(y, w) = \min_{x} \{wx : (x, y) \in T\}$$

where $w \in R^N_{++}$ is the price vector for inputs. The bank cost-minimizing objective can be equivalently defined in terms of the DIDF in equation (3.3) as

$$C(y, w) = \min_{x} \left\{ wx : \vec{D}_I \left(y, x; g_x\right) \geq 0 \right\}.$$

Note that the cost function can be equivalently defined in terms of the DIDF to keep consistency with the endogeneity of inputs and exogeneity of outputs. Following Luenberger (1992) and Färe and Primont (1995), the constrained optimization problem can be represented by an unconstrained problem as

$$C(y, w) = \min_{x} \left\{ wx - \vec{D}_I \left(y, x; g_x\right) \times w' g_x \right\}$$

The corresponding first-order conditions are

$$w_n - \nabla_{x_n} \vec{D}_I(.) w' g_x = 0$$

Alternatively,

$$w_n = \nabla_{x_n} \vec{D}_I(.) \lambda_I \quad for\ n = 1, 2 \tag{3.20}$$

where $\lambda_I = w'g_x = \sum_{n=1}^{N} w_n g_{x_n}$ is the sum of direction-weighted cost and can be interpreted as the Lagrange multiplier. The first-order conditions in (3.20) are the inverse demand functions. In order to meet the rank condition for the identification of the model, a total of at least the total number of potentially endogenous variables is needed as independent equations in the system. As it is well-known, the system (3.20) is singular and only $(n-1)$ equations in (3.20) can be used for the estimation — see Barten (1969) for more details. The DIDF in (3.15) plus the $(n-1)$ first-order conditions in (3.20) provide a system of n equations. Precisely, the system consists of the DIDF in (3.15) and the first-order condition for w_2. Note that $\nabla_{x_2}\vec{D}_I(y, \widetilde{x}; g_x) = \nabla_{x_2}\vec{D}_I(y, x; g_x)$ by the translation property, then the first-order condition in (3.20) can be rewritten in terms of the parameters of (3.15) after adding an iid normal error term v_C as follows

$$w_2 = \lambda_I \left(\alpha_2 + \alpha_{22}\widetilde{x}_2 + \sum_{m=1}^{2} \gamma_{2m} y_m + \delta_{tx_2} t \right) + v_C \tag{3.21}$$

where $\widetilde{x}_2 = x_2 - x_1 g_{x_2}$. Note that, solving the first-order condition for $\widetilde{x}_2$ treats $\widetilde{x}_2$ as an endogenous variable (as opposed to x_2). The equivalence of working with $\widetilde{x}_2$ and working directly with x_2 holds because of $\partial\widetilde{x}_2 / \partial x_2 = 1$. By allowing g to differ across banks, (3.21) can be further written in a panel data framework as

$$w_{2,it} = \widetilde{R}_{C,it}(g_i)'\beta_C + v_{C,it} \tag{3.22}$$

where $\widetilde{R}_C$ is a vector of the relevant variables on the right-hand side of (3.21) and β_C is the corresponding vector of coefficients. β_C is a subset of β_I, which can be obtained using a selection matrix A_I that contains elements which are either 0 or 1, where $\beta_C = A_I\beta_I$.

The DIDF system (equations (3.15) and (3.22)) is a simultaneous equation model where the entire vector x is endogenous. The entire system can be written as

$$\begin{bmatrix} -x_{1,it} \\ w_{2,it} \end{bmatrix} = \begin{bmatrix} \widetilde{R}_{I,it}(g_i)' & \\ & \widetilde{R}_{C,it}(g_i)' \end{bmatrix} [\beta] - \begin{bmatrix} u_{I,it} \\ 0 \end{bmatrix} + \begin{bmatrix} v_{I,it} \\ v_{C,it} \end{bmatrix}$$

which can be written in a compressed form as

$$Y_{it} = R_{it}(g_i)\beta - u_{I,it}\iota + v_{it} \tag{3.23}$$

where $\iota = [1, 0]$, $v_{it} = (v_{I,it}, v_{C,it})' \sim N(0, \Sigma)$, and $g_i = (g_{x_{2i}})'$ for $i = 1, ..., K$. The DIDF system is estimated subject to the symmetry restrictions and theoretical monotonicity restrictions, while the translation property is already imposed by construction. The directional vector g_i is treated as unknown parameters which are estimated jointly with the remaining parameters in the system. The obtained estimates of the directional vector can be interpreted as being cost-optimal due to the inclusion of the cost-minimizing first-order conditions in the system — see Malikov *et al.* (2016). That is, the estimated DIDF direction captures the bank movement to the point on a technological frontier where costs are minimized.

The Revenue-Optimal Directional Vector

The bank revenue-maximizing objective is defined as

$$R(x, p) = \max_y \{py : (x, y) \in T\}$$

where $p \in R^M_{++}$ is the price vector for outputs. The bank revenue-maximizing objective can be equivalently defined in terms of the DODF in equation (3.5) as

$$R(x, p) = \max_y \left\{ py : \vec{D}_O(x, y; g_y) \geq 0 \right\}.$$

Note that the revenue function can be equivalently defined in terms of the DODF to keep consistency with the endogeneity of outputs and exogeneity of inputs. The constrained optimization problem can be represented by an unconstrained problem as

$$R(x, p) = \max_y \left\{ py + \vec{D}_O(x, y; g_y) \times p' g_y \right\}$$

The first-order conditions for the revenue maximization problem are

$$p_m + \nabla_{y_m} \vec{D}_O(.) p' g_y = 0$$

Alternatively,

$$p_m = -\nabla_{y_m} \vec{D}_O(.)\lambda_O \quad for\ m = 1, 2 \tag{3.24}$$

where $\lambda_O = p' g_y = \sum_{m=1}^{M} p_m g_{y_m}$ is the sum of direction-weighted revenues and can be interpreted as the Lagrange multiplier. The first-order conditions in (3.24) are the inverse supply functions. In order to meet the rank condition for the identification of the model, a total of at least the total number of potentially endogenous variables is needed as independent equations in the system. As it is well-known, the system (3.24) is singular and only $(m-1)$ equations in (3.24) can be used for the estimation. The DODF in equation (3.17) plus the $(m-1)$ equations given in equation (3.24) provide a system of m equations. Precisely, the system consists of the DODF in equation (3.17) and the first-order condition for p_2. Note that $\nabla_{y_2} \vec{D}_O(x, \widetilde{y}; g_y) = \nabla_{y_2} \vec{D}_O(x, y; g_y)$ by the translation property, then the first-order conditions in (3.24) can be rewritten in terms of the parameters of (3.17) after adding an iid normal error term v_R as follows

$$p_2 = -\lambda_O \left(\beta_2 + \beta_{22}\widetilde{y}_2 + \sum_{n=1}^{2} \gamma_{n2} x_n + \delta_{ty_2} t \right) + v_R \tag{3.25}$$

where $\widetilde{y}_2 = y_2 - y_1 g_{y_2}$. Note that, solving the first-order conditions for $\widetilde{y}_2$ treats $\widetilde{y}_2$ as an endogenous variable (as opposed to y_2). The equivalence of working with $\widetilde{y}_2$ and working directly with y_2 holds because of $\partial \widetilde{y}_2 / \partial y_2 = 1$. By allowing g to differ across banks, (3.25) can be further written in a panel data framework as

$$p_{2,it} = \widetilde{R}_{R,it}(g_i)' \beta_R + v_{R,it} \tag{3.26}$$

where $\widetilde{R}_R$ is a vector of the relevant variables on the right-hand side of (3.25) and β_R is the corresponding vector of coefficients. β_R is a subset of β_O, which can be obtained using a selection matrix A_O that contains elements which are either 0 or 1, where $\beta_R = A_O \beta_O$.

The DODF system (equations (3.17) and (3.26)) is a simultaneous equation model where

the entire vector y is endogenous. The entire system can be written as

$$\begin{bmatrix} y_{1,it} \\ p_{2,it} \end{bmatrix} = \begin{bmatrix} \widetilde{R}_{O,it}(g_i)' & \\ & \widetilde{R}_{R,it}(g_i)' \end{bmatrix} [\beta] - \begin{bmatrix} u_{O,it} \\ 0 \end{bmatrix} + \begin{bmatrix} v_{O,it} \\ v_{R,it} \end{bmatrix}$$

which can be written in a compressed form as

$$Y_{it} = R_{it}(g_i)\beta - u_{O,it}\iota + v_{it} \tag{3.27}$$

where $\iota = [1, 0]$, $v_{it} = (v_{O,it}, v_{R,it})' \sim N(0, \Sigma)$, and $g_i = (g_{y_{2i}})'$ for $i = 1, ..., K$. The DODF system is estimated subject to the symmetry restrictions and theoretical monotonicity restrictions, while the translation property is already imposed by construction. The directional vector g_i is treated as unknown parameters which are estimated jointly with the remaining parameters in the system. The obtained estimates of the directional vector can be interpreted as being revenue-optimal due to the inclusion of the revenue-maximizing first-order conditions in the system. That is, the estimated DODF direction captures the bank movement to the point on a technological frontier where revenues are maximized.

The Profit-Optimal Directional Vector

The bank profit-maximizing objective is defined as

$$\pi(p, w) = \max_{x,y} \{(p'y - w'x) : (x, y) \in T\}$$

where $w \in R^N_{++}$, and $p \in R^M_{++}$ are the price vectors for inputs and outputs, respectively. Chambers *et al.* (1998) show that the profit function can be equivalently defined in terms of the DTDF in equation (3.1) as

$$\pi(p, w) = \max_{x,y} \left\{ (p'y - w'x) : \vec{D}_T(x, y; g_x, g_y) \geq 0 \right\}.$$

Following Chambers *et al.* (1998), the constrained optimization problem can be represented by an unconstrained problem as

$$\pi(p, w) = \max_{x,y} \left\{ (p'y - w'x) + \vec{D}_T(x, y; g_x, g_y)(p'g_y + w'g_x) \right\}$$

The first-order conditions for the profit maximization problem are

$$-w_n + \nabla_{x_n} \vec{D}_T(.)(p'g_y + w'g_x) = 0$$

$$p_m + \nabla_{y_m} \vec{D}_T(.)(p'g_y + w'g_x) = 0$$

Alternatively,

$$w_n = \nabla_{x_n} \vec{D}_T(.)\lambda_T \qquad \text{for } n = 1, ..., N$$

$$p_m = -\nabla_{y_m} \vec{D}_T(.)\lambda_T \qquad \text{for } m = 1, ..., M \tag{3.28}$$

where $\lambda_T = p'g_y + w'g_x = \sum_{m=1}^{M} p_m g_{y_m} + \sum_{n=1}^{N} w_n g_{x_n}$ is the sum of direction-weighted profits and can be interpreted as the Lagrange multiplier — see Hudgins and Primont (2007). The first-order conditions in (3.28) are the inverse demand and supply functions. Note that $\nabla_{x_n} \vec{D}_T(\widetilde{x}, \widetilde{y}; g_x, g_y) = \nabla_{x_n} \vec{D}_T(x, y; g_x, g_y)$, and $\nabla_{y_m} \vec{D}_T(\widetilde{x}, \widetilde{y}; g_x, g_y) = \nabla_{y_m} \vec{D}_T(x, y; g_x, g_y)$ by the translation property, then the first-order conditions in (3.28) can be rewritten in terms of the parameters of (3.10) after adding iid normal error terms v_π as follows

$$w_n = \lambda_T \left(\alpha_n + \sum_{n'=1}^{2} \alpha_{nn'} \widetilde{x}_{n'} + \gamma_{n2} \widetilde{y}_2 + \delta_{tx_n} t \right) + v_{\pi w_n} \quad \text{for } (n = 1, 2)$$

$$p_2 = -\lambda_T \left(\beta_2 + \beta_{22} \widetilde{y}_2 + \sum_{n=1}^{2} \gamma_{n2} \widetilde{x}_n + \delta_{ty_2} t \right) + v_{\pi p}. \tag{3.29}$$

By allowing g to differ across banks, (3.29) can be further written in a panel data framework as

$$w_{n,it} = \widetilde{R}_{w_n,it}(g_i)'\beta_{w_n} + v_{\pi w_n,it} \qquad \text{for } (n = 1, 2)$$

$$p_{2,it} = \widetilde{R}_{p_2,it}(g_i)'\beta_{p_2} + v_{\pi p_2,it} \tag{3.30}$$

where $\widetilde{R}_{w_n}$ and $\widetilde{R}_{p_2}$ are the vectors of the relevant variables on the right-hand sides of (3.29) and β_{w_n} and β_{p_2} are the corresponding vectors of coefficients. β_{w_n} and β_{p_2} are subsets of β_T, which can be obtained using a selection matrix A_T that contains elements which are either 0 or 1, where $\beta_{w_n} = A_T\beta_T$ and $\beta_{p_2} = A_T\beta_T$.

In order to meet the rank condition for the identification of the model, a total of at least $N + M$ independent equations are needed in the system, where $N + M$ is the total number

of potentially endogenous variables (i.e., the total number of inputs and outputs) and is equal to 4 in this case. More precisely, the DTDF in equation (3.10) plus the $((n+m)-1)$ first-order conditions for profit maximization given in equation (3.30) provide a system of $(n+m)$ equations. This system of equations is a simultaneous equation model where the entire vector (x, y) is endogenous. Adding the interactive effects equation in (3.18), the entire system can be written as

$$\begin{bmatrix} y_{1,it} \\ w_{1,it} \\ w_{2,it} \\ p_{2,it} \\ u_{T,it} \end{bmatrix} = \begin{bmatrix} \widetilde{R}_{T,it}(g_i)' & & & & \\ & \widetilde{R}_{w_1,it}(g_i)' & & & \\ & & \widetilde{R}_{w_2,it}(g_i)' & & \\ & & & \widetilde{R}_{p_2,it}(g_i)' & \\ & & & & Z_{it}' \end{bmatrix} [\beta] - \begin{bmatrix} u_{T,it} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} v_{T,it} \\ v_{w_1,it} \\ v_{w_2,it} \\ v_{p_2,it} \\ v_{u,it} \end{bmatrix}$$

which can be written in a compressed form as

$$Y_{it} = R_{it}(g_i)\beta - u_{T,it}\iota + v_{it} \tag{3.31}$$

where $\iota = [1, 0, 0, 0, 0]$, $v_{it} = (v_{T,it}, v_{w_1,it}, v_{w_2,it}, v_{p_2,it}, v_{u,it})' \sim N(0, \Sigma)$, and $g_i = (g_{x_{1i}}, g_{x_2 i}, g_{y_{2i}})'$ for $i = 1, ..., K$. Bank-specific variables Z is a vector of the relevant variables on the right-hand side of (3.18) that are obtained using the cost-optimal and revenue-optimal directional vectors defined in the system of equations in (3.23) and (3.27), respectively. The DTDF system is estimated subject to the symmetry restrictions and theoretical monotonicity restrictions, while the translation property is already imposed by construction. The directional vector g_i is treated as unknown parameters which are estimated jointly with the remaining parameters in the system. The obtained estimates of the directional vector can then be

interpreted as being profit-optimal due to the inclusion of the profit-maximizing first-order conditions in the system. That is, the estimated DTDF direction captures the bank movement to the point on a technological frontier where profits are maximized.

Setting all the elements of g equal to ones; $g_{x_{n1}} = g_{x_{n2}} = ... = g_{x_{nk}} = 1$ for $(n = 1, 2)$, $g_{y_{21}} = g_{y_{22}} = ... = g_{y_{2k}} = 1$ without additional first-order condition equations, the system in (3.31) reduces to (3.19) in the case of the unit value directional vector. Setting $g_{x_{1i}} = x_{1_i}$, $g_{x_{2i}} = x_{2_i}$, and $g_{y_{2_i}} = y_{2_i}$ without additional first-order condition equations, the system in (3.31) reduces to (3.19) in the case of the observed input-output directional vector.

3.4 Bayesian Estimation

Bayesian approach is used to estimate the DIDF, DODF, and DTDF models using the unit value and the observed input-output directional vectors defined by (3.15), (3.17), and (3.19), and the cost-optimal, revenue-optimal, and profit-optimal directional vectors defined by (3.23), (3.27), and (3.31), respectively. Bayesian estimation involves using a Markov Chain Monte Carlo (MCMC) sampling algorithm to generate sequences of samples from the joint posterior distribution of inefficiency and the unknown parameters of the model. This paper uses Metropolis-Hastings algorithm introduced by Metropolis *et al.* (1953) and Hastings (1970). Bayesian estimation and MCMC sampling algorithm have been widely documented in the stochastic frontier literature and thus are not discussed in this paper — see, for example, Koop and Steel (2003), and O'Donnell and Coelli (2005). The reason for using this approach is that it combines prior information about the parameters with the information contained in the data through the likelihood function. Thus, prior information about the parameters such as monotonicity conditions implied by microeconomic theory can be easily imposed through the prior distribution of the parameters. Furthermore, directional vectors that vary across banks can be easily estimated with the Bayesian approach.

3.4.1 Prior Distributions

The use of Bayesian approach requires choosing prior distributions for the parameters β, Σ^{-1}, u_{it}, λ^{-1}, and g_i. For the ease of the comparison of the results among the three directional distance function models, the same prior distributions for the parameters are used. Following Gelfand *et al.* (1990), a normal prior distribution with zero mean and a large variance for β is used to ensure that the prior distribution for β is relatively uninformative.

$$p(\beta) \sim N(\beta_0, \Omega_\beta) I\left(\beta \in S_j\left(g_i\right)\right) \tag{3.32}$$

where β_0 is a vector of zeros and Ω_β is a diagonal matrix with 10^4 in diagonal elements. $I\left(\beta \in S_j\left(g_i\right)\right)$ is an indicator function which takes the value one if the constraints are satisfied and zero otherwise, and $S_j\left(g_i\right)$, which depends on g_i, is the set of permissible parameter values when no theoretical regularity constraints ($j = 0$) are imposed and when the theoretical regularity constraints ($j = 1$) must be satisfied. The indicator function restricts prior support to the region where the theoretical regularity constraints are satisfied.

For the covariance matrix Σ, the Wishart distribution is used first due to its conjugacy properties with the normal sampling model. However, it is found to be biased toward large values which result in large values for Σ and consequently large values for the inefficiency measures. The MCMC algorithm for a system of equations is also terminated after a small number of iterations due to the large values involved. Then, following O'Donnell and Coelli (2005), the following prior is used

$$p\left(\Sigma^{-1}\right) \propto \Sigma \tag{3.33}$$

which implies that Σ^{-1} is fully determined by the likelihood function — see the conditional posterior density for Σ^{-1} in equation (3.39).

As noted by van den Broeck *et al.* (1994), models based on exponential distribution are reasonably robust to changes in prior assumptions about the parameters. Therefore, an exponential distribution is used for the technical inefficiency u_{it} with an unknown parameter λ following Koop and Steel (2003); $u_{it} \sim i.i.d.\exp(\lambda^{-1})$. Since the exponential distribution

is a gamma distribution when its first parameter equals one, the prior for u_{it} can be written as

$$p\left(u_{it} \mid \lambda^{-1}\right) = f_{Gamma}\left(u_{it} \mid 1, \lambda^{-1}\right). \tag{3.34}$$

According to Fernandez *et al.* (1997), in order to obtain a proper posterior, a proper prior for the parameter λ should be used. Therefore, the parameter λ is assumed to have independent exponential prior with mean equals to $-1/\ln\tau^*$ following van den Broeck *et al.* (1994). The prior independence of λ leads to marginally prior independent of inefficiencies.

$$p(\lambda^{-1}) = f_{Gamma}\left(\lambda^{-1} \mid 1, -\ln\tau^*\right) \tag{3.35}$$

where τ^* is the prior estimate of the mean of the technical efficiency distribution — see, for example, Koop *et al.* (1997) and O'Donnell and Coelli (2005). My best prior knowledge of the efficiency of US banks is the mean efficiency value of 0.4583 for DTDF with fixed directional vector, and 0.9431 for DTDF with cost-optimal directional vector reported by Malikov *et al.* (2016) who apply a Bayesian DTDF-cost system approach to large US commercial banks for 2001–2010 period. The mean output technical efficiency value of 0.9279 is reported by Feng and Serletis (2010) who apply a Bayesian output distance function to US large banks for 2000–2005 period. To my knowledge, the input technical efficiency is not reported by any US banking study over a comparable period. However, the mean input technical efficiency value of 0.690 for all banks, 0.707 for small banks, and 0.735 for large banks are reported by Marsh *et al.* (2003) who apply a Bayesian input distance function to US commercial banks during 1990–2000. After reviewing the results of 50 US bank efficiency studies, Berger and Humphrey (1997) find that the average efficiency is 0.84. Since changing τ^* changes the prior moments, various values of τ^* within its possible range is experimented to assess the sensitivity of the results to changes to τ^*. The results are the same up to the number of digits presented in Section 6, implying that the results are very robust to large changes in τ^*.

To account for heterogeneity in the directional vectors for banks, prior distribution for g_i

is specified as a normal prior distribution with mean $G_0 = 1$ and a large variance to ensure that the prior distribution for g_i is relatively uninformative.

$$p(g_i) \sim N\left(G_0, \Omega_G\right) \tag{3.36}$$

where $g_i = \left(g_{x_{2_1}}, g_{x_{2_2}}, ..., g_{x_{2_K}}\right)$ for the DIDF-cost system, $g_i = \left(g_{y_{2_1}}, g_{y_{2_2}}, ..., g_{y_{2_K}}\right)$ for the DODF-revenue system, and $g_i = \left(g_{x_{1i}}, g_{x_{2i}}, g_{y_{2i}}\right)$ for the DTDF-profit system, with $i = 1, ..., K$ indexing banks. G_0 is a vector of ones, and Ω_G is a diagonal matrix with 10^4 in diagonal elements. Note that the directional vector that projects any inefficient bank to the cost (revenue or profit) minimizing (maximizing) benchmark does not impose any sign restrictions on the adjustments of inputs and outputs. Therefore, these directional vectors may have negative components such that inputs are expanded, or outputs are contracted to reach the frontier at the cost (revenue or profit) minimizing (maximizing) benchmark — see, for example, Zofio *et al.* (2013) for a profit-optimal directional vector and Atkinson *et al.* (2018) for a cost and profit-optimal directional vector.

Using the priors in (3.32)–(3.36), and assuming that the prior distributions of the parameters are independent, the joint prior probability density function is therefore

$$f\left(\beta, \Sigma^{-1}, u_{it}, \lambda^{-1}, g_i\right) = p(\beta)p\left(\Sigma^{-1}\right)p\left(u_{it} \mid \lambda^{-1}\right)p(\lambda^{-1})p(g_i). \tag{3.37}$$

3.4.2 Full Conditional Posterior Distributions

Let $\Gamma = (\beta, \Sigma^{-1}, u_{it}, \lambda^{-1}, g_i)$ denotes all the parameters of the model, and Γ_{-a} denotes all parameters other than a. To derive the likelihood function, the Jacobian transformation matrix from the vector of random errors to the endogenous variables (all the inputs and outputs) for the DTDF-profit system is defined as follows

$$J_{it}\left(g_i, \beta\right) = \frac{\partial\left(v_{T,it} - u_{T,it}, v_{w_1,it}, v_{w_2,it}, v_{p_2,it}\right)}{\partial\left(y_{1,it}, x_{1,it}, x_{2,it}, y_{2,it}\right)}.$$

The Jacobian transformation matrix from the vector of random errors $(v_{I,it} - u_{I,it}, v_{C,it})'$ to the endogenous variables (inputs) for the DIDF-cost system is

$$J_{it}(g_i, \beta) = \frac{\partial (v_{I,it} - u_{I,it}, v_{C,it})}{\partial (x_{1,it}, x_{2,it})}.$$

The Jacobian transformation matrix from the vector of random errors $(v_{O,it} - u_{O,it}, v_{R,it})'$ to the endogenous variables (outputs) for the DODF-revenue system is

$$J_{it}(g_i, \beta) = \frac{\partial (v_{O,it} - u_{O,it}, v_{R,it})}{\partial (y_{1,it}, y_{2,it})}.$$

Applying Jacobian transformation, the conditional density of the endogenous variables for bank i is

$$L_i(Y \mid \Gamma) \propto f_{Normal}(Y_i \mid R_i(g_i)\beta - u_i\iota, I_T \otimes \Sigma) \prod_{t=1}^{T} |\det(J_{it}(g_i, \beta))|$$

where $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{iT})'$, and I_T denotes an identity matrix of order T. The likelihood function of Y, given Γ is

$$L(Y \mid \Gamma) = \left[\prod_{i=1}^{K} f_{Normal}(Y_i \mid R_i(g_i)\beta - u_i\iota, I_T \otimes \Sigma)\right] \prod_{i=1}^{K}\prod_{t=1}^{T} |\det(J_{it}(g_i, \beta))|$$

Alternatively,

$$L(Y \mid \Gamma) = \left[\prod_{i=1}^{K}\prod_{t=1}^{T} f_{Normal}(Y_{it} \mid R_{it}(g_{it})\beta - u_{it}\iota, I \otimes \Sigma)\right] \prod_{i=1}^{K}\prod_{t=1}^{T} |\det(J_{it}(g_i, \beta))|. \quad (3.38)$$

Using Bayes' Theorem and combining the likelihood function in (3.38) and the joint prior distributions in (3.37), all terms that are constant with respect to the parameter can be ignored in order to obtain the full conditional posterior distribution for each parameter in the model. The full conditional posterior distributions for all the parameters are found to be

$$p\left(\Sigma^{-1} \mid Y, \Gamma_{-\Sigma^{-1}}\right) \propto f_{Gamma}\left(\Sigma^{-1} \mid \frac{KT}{2}, \frac{1}{2}(Q_{it} + u_{it}\iota)'(Q_{it} + u_{it}\iota)\right), \quad (3.39)$$

$$p(\beta \mid Y, \Gamma_{-\beta}) \propto f_{Normal}(\beta \mid Dd, D) \prod_{i=1}^{K}\prod_{t=1}^{T} |\det(J_{it}(g_i, \beta))| I(\beta \in S_j(g_i)), \quad (3.40)$$

$$p\left(\lambda^{-1} \mid Y, \Gamma_{-\lambda^{-1}}\right) \propto f_{Gamma}\left(\lambda^{-1} \mid KT + 1, u'\iota_{KT} - \ln\tau^*\right), \quad (3.41)$$

$$p\left(u_{it} \mid Y, \Gamma_{-u_{it}}\right) \propto f_{Normal}\left(u_{it} \mid -\frac{\left(Q_{it}'\left(g_i\right)\Sigma^{-1}\iota + \lambda^{-1}\right)}{\iota'\Sigma^{-1}\iota}, \frac{1}{\iota'\Sigma^{-1}\iota}\right) I\left(u_{it} \geq 0\right), \tag{3.42}$$

$$\begin{aligned} p\left(g_i \mid Y, \Gamma_{-g_i}\right) \propto & \left[\prod_{i=1}^{K} f_{Normal}\left(Y_i \mid R_i\left(g_i\right)\beta - u_i\iota, I \otimes \Sigma\right)\right] \prod_{i=1}^{K}\prod_{t=1}^{T} \left|\det\left(J_{it}\left(g_i, \beta\right)\right)\right| \\ & \times \prod_{i=1}^{K} f_{Normal}\left(g_i \mid G_0, \Omega_G\right) \prod_{i=1}^{K} I\left(\beta \in S_j\left(g_i\right)\right) \end{aligned} \tag{3.43}$$

where $D = \left(R_{it}\left(g_i\right)'\left(I \otimes \Sigma\right)^{-1} R_{it}\left(g_i\right) + \Omega_{\beta}^{-1}\right)^{-1}$, $d = R_{it}\left(g_i\right)'\left(I \otimes \Sigma\right)^{-1}\left(Y_{it} + u_{it}\iota\right) + \Omega_{\beta}^{-1}\beta_0$, and $Q_{it} = Y_{it} - R_{it}\left(g_i\right)\beta$. Note that the full conditional posterior is proportional to the function f where the missing normalizing constant can be computed by integrating the function f. The Metropolis-Hastings algorithm does not require knowledge of this normalizing constant — see Chen *et al.* (2000) for more details on this algorithm. $I\left(u_{it} \geq 0\right)$ is an indicator function that takes the value one if the constraint $u_{it} \geq 0$ is satisfied and zero otherwise.

Bayesian estimation for a single-equation stochastic directional distance function without additional first-order condition equations and with a pre-specified directional vector, $g = (-1, 1)$ or $g = (-x, y)$ can be implemented by setting $\prod_{i=1}^{K}\prod_{t=1}^{T} \left|\det\left(J_{it}\left(g_i, \beta\right)\right)\right| = 1$ and setting the relevant elements of g equal to ones or $g_{x_{1i}} = x_{1i}$, $g_{x_{2i}} = x_{2i}$, or $g_{y_{2i}} = y_{2i}$, and normalizing the relevant directional vector. Bayesian estimation without theoretical regularity constraints can be implemented by setting $I\left(\beta \in S_j\left(g_i\right)\right)$ in (3.40) and (3.43) equal to one and then drawing sequentially from the full conditional posteriors in (3.39)–(3.43).

3.4.3 Estimating the Interactive Effects

In the Bayesian framework, Koop *et al.* (1997) propose a model where a time-invariant inefficiency is assumed to be exponentially distributed with producer-specific mean inefficiencies λ_i and independent exponential priors; $u_i \sim i.i.d.\exp(\lambda_i^{-1})$ where $\lambda_i = \exp\left(Z_i'\delta\right)$. Following Koop *et al.* (1997), the inefficiency term $u_{T,it}$, can be specified to be time-variant inefficiency by including bank-specific time-varying covariates in the parameter of an exponential distribution as; $u_{T,it} \sim i.i.d.\exp(\lambda_{it}^{-1})$, and $\lambda_{it} = \exp\left(Z_{it}'\delta\right)$, where δ is an unknown vector of coefficients (including the intercept) to be estimated. Since the exponential distribution is a

gamma distribution when its first parameter equals one, the prior for $u_{T,it}$ can be written as

$$p\left(u_{T,it}, \delta \mid Z\right) = f_{Gamma}\left(u_{T,it} \mid 1, \exp\left(Z_{it}'\delta\right)\right) \tag{3.44}$$

Following Koop *et al.* (1997), the parameter vector δ is assumed to have a proper prior independent of the other parameters. A normal prior distribution with mean δ_0 and variance Ω_δ for δ is used.

$$p(\delta) \sim N(\delta_0, \Omega_\delta) \tag{3.45}$$

A large variance for δ is used to ensure that the prior distribution for δ is relatively uninformative where δ_0 is a vector of zeros and Ω_δ is a diagonal matrix with 10^4 in diagonal elements. Note that by conditioning on Y and Z, bank-specific variables Z are allowed to be correlated with the variables describing the frontier Y.

The full conditional posterior for δ and $u_{T,it}$ can be obtained by ignoring all terms that are constant with respect to δ and $u_{T,it}$, respectively. The full conditional posterior distributions for δ and $u_{T,it}$ are found to be

$$p\left(\delta \mid Y, \Gamma_{-\delta}\right) \propto f_{Normal}\left(\delta \mid \frac{\delta_0\Omega_\delta^{-1}\iota - Z_{it}'u_{T,it}}{\iota'\Omega_\delta^{-1}\iota}, \frac{1}{\iota'\Omega_\delta^{-1}\iota}\right), \tag{3.46}$$

$$p\left(u_{T,it} \mid Y, \Gamma_{-u_{T,it}}\right) \propto f_{Normal}\left(u_{T,it} \mid -\frac{\left(Q_{it}'\left(g_i\right)\Sigma^{-1}\iota + \mu_{it}\right)}{\iota'\Sigma^{-1}\iota}, \frac{1}{\iota'\Sigma^{-1}\iota}\right) I\left(u_{T,it} \geq \mu_{it}\right) \tag{3.47}$$

where $Q_{it}\left(g_i\right) = Y_{it} - R_{it}\left(g_i\right)\beta$, and $\mu_{it} = Z_{it}'\delta$.

3.5 Data

The annual data on US commercial banks used in this paper is obtained from the Reports of Income and Condition (Call Reports) over the period from 2001 to 2015. Only continuously operating banks are examined to avoid the impact of entry through new charters and exit through failure or merger, and to focus on the performance of a core of healthy, surviving banks during the sample period. The data sample consists of a balanced panel of a total of 148 banks (K=148) observed over 15 years, for a total of 2220 observations.

To select the relevant variables, the commonly accepted asset approach proposed by Sealey and Lindley (1977) is used. It defines loans and other assets as outputs, while deposits and other liabilities are treated as inputs. On the input side, two inputs are included; the quantity of labor, x_1, and the quantity of purchased funds and deposits, x_2. On the output side, two outputs are included; total loans y_1 which is composed of consumer loans, commercial and industrial loans, and real estate loans; and securities y_2 which includes all non-loan financial assets (i.e., all financial and physical assets minus the sum of total loans, and physical capital (premises and other fixed assets)), so that all financial assets are included.

While non-traditional banking activities are becoming increasingly important in identifying bank outputs, the imperfect data and the wide range of activities such as securitization, brokerage services, management of financial assets for depositors and borrowers, and others, make the measurement of non-traditional banking activities controversial. See Stiroh (2000) for a discussion of the different approaches to the measurement of non-traditional banking activities. To avoid the uncertainties associated with the measurement of non-traditional banking activities, it is not included as an additional output.

All the quantities of inputs and outputs are constructed by following the data construction method in Berger and Mester (2003). These quantities are deflated by the consumer price index CPI to the base year 2005, except for the quantity of labor. The data is normalized by dividing each input and output by its sample mean prior to the estimation following Färe *et al.* (2005). This normalization implies that $(x, y) = (1, 1)$ for a bank that uses mean inputs and produces mean outputs.

For the input and output prices, the actual price paid by the bank for each input or the bank-specific price of input is obtained by dividing total expenses on each input by the corresponding input quantity. Similarly, the actual price received by the bank for each output or the bank-specific price of output is obtained by dividing total revenues from each

output by the corresponding output quantity. Thus, for example, the bank-specific price of labor w_1 is obtained from expenses on salaries and benefits divided by the number of full-time employees x_1. The same approach is used to obtain w_2, p_1, and p_2.

Following Berger and Mester (2003), the market-average prices faced and determined exogenously rather than the actual prices paid or received by the bank are used. These market-average prices are more likely to be exogenous to the bank than the bank-specific prices. The bank market-average price at a given year is the weighted average of the other banks prices at that year excluding the bank-specific price, where the weights are each bank respective market share at that year. For example, the market-average price of labor that bank i faces in the labor market L at year t is obtained as

$$w_{1it} = \sum_{j=1, j\neq i}^{l} \left(\frac{x_{1jt}}{\sum_{h=1, h\neq i}^{l} x_{1ht}} \right) w_{1jt}$$

where l is the number of banks operating in labor market L, x_{1jt} is the number of full-time employees of bank j at year t and w_{1jt} is the bank j specific price of labor at year t, which is obtained from expenses on salaries and benefits divided by the number of full-time employees x_{1jt}. The same approach is used to obtain w_{2it}, p_{1it}, and p_{2it} for $i = 1, ..., K$ over the years $t = 1, ..., T$. Data summary statistics are presented in Table 3.1.

3.6 Empirical Results

To investigate the relationships among input, output, and overall technical inefficiencies, several models are estimated. Specifically, input, output, and technology-oriented technical inefficiencies are estimated separately using the Bayesian procedure outlined in Section 3.4 and directional input, output, and technology distance functions, respectively. All of these inefficiencies are estimated with the three commonly used directional vectors; the unit value, the observed input-output, and the optimal directional vectors. The unit value and the observed input-output directional vector models are estimated without additional first-order condition equations. The optimal directional vector models are estimated using systems

of equations, consisting of directional distance functions with the relevant first-order conditions, as discussed in Section 3.3. The unit value directional vector models are referred to as UDIDF, UDODF, and UDTDF, respectively. The observed input-output directional vector models are referred to as VDIDF, VDODF, and VDTDF, respectively. The optimal directional vector models are referred to as ODIDF, ODODF, and ODTDF, respectively. In total, nine models are estimated. For each of the nine models, a total of 450,000 observations are generated and then the first 150,000 observations are discarded as a burn-in. The simulation inefficiency factor (SIF) values for all the parameters of these models are estimated to check the mixing performance of the samplers following Kim *et al.* (1998). The SIF values for the unit value, the observed input-output, and the optimal directional vector models are reported in Tables $3.5 - 3.7$, suggesting that the samplers for these models have converged.

3.6.1 Imposing the Theoretical Regularity Conditions

As required by neoclassical microeconomic theory, the production technology has to satisfy the theoretical regularity conditions of monotonicity and curvature. Monotonicity requires that the directional distance function be non-decreasing in inputs and non-increasing in outputs. Therefore, monotonicity conditions of the DTDF imply the following restrictions

$$\frac{\partial \vec{D}_T(.)}{\partial x_n} = \alpha_n + \sum_{n'=1}^{2} \alpha_{nn'} x_{n'} + \left[\alpha_{nn} g_{x_n} + \alpha_{nn'} g_{x_{n'}} - \gamma_{n2} g_{y_2}\right] y_1 + \gamma_{n2} y_2 + \delta_{tx_n} t \geq 0 \quad (n = 1, 2)\,;$$

$$\begin{aligned}
\frac{\partial \vec{D}_T(.)}{\partial y_1} &= [\alpha_1 g_{x_1} + \alpha_2 g_{x_2} - \beta_2 g_{y_2} - 1] \\
&+ \left[\alpha_{11} g_{x_1}^2 + \alpha_{22} g_{x_2}^2 + \beta_{22} g_{y_2}^2 + \alpha_{12} g_{x_1} g_{x_2} - \gamma_{12} g_{x_1} g_{y_2} - \gamma_{22} g_{x_2} g_{y_2}\right] y_1 \\
&+ \left[\gamma_{12} g_{x_1} + \gamma_{22} g_{x_2} - \beta_{22} g_{y_2}\right] y_2 \\
&+ \left[\alpha_{11} g_{x_1} + \alpha_{12} g_{x_2} - \gamma_{12} g_{y_2}\right] x_1 \\
&+ \left[\alpha_{21} g_{x_1} + \alpha_{22} g_{x_2} - \gamma_{22} g_{y_2}\right] x_2 \\
&+ \left[\delta_{tx_1} g_{x_1} + \delta_{tx_2} g_{x_2} - \delta_{ty_2} g_{y_2}\right] t \leq 0;
\end{aligned}$$

$$\frac{\partial \vec{D}_T(.)}{\partial y_2} = \beta_2 + \left[\gamma_{12} g_{x_1} + \gamma_{22} g_{x_2} - \beta_{22} g_{y_2}\right] y_1 + \beta_{22} y_2 + \sum_{n=1}^{2} \gamma_{n2} x_n + \delta_{ty_2} t \leq 0. \tag{3.48}$$

When $g_{x_1} = g_{x_2} = 0$, monotonicity conditions of the DTDF in (3.48) reduces to the monotonicity conditions of the DODF. Monotonicity conditions of the DIDF imply the following restrictions

$$\frac{\partial \vec{D}_I(.)}{\partial x_1} = \left[1 - \alpha_2 g_{x_2}\right] + \alpha_{22} g_{x_2}^2 x_1 - \alpha_{22} g_{x_2} x_2 - \sum_{m=1}^{2} \gamma_{2m} g_{x_2} y_m - \delta_{tx_2} g_{x_2} t \geq 0;$$

$$\frac{\partial \vec{D}_I(.)}{\partial x_2} = \alpha_2 - \alpha_{22} g_{x_2} x_1 + \alpha_{22} x_2 + \sum_{m=1}^{2} \gamma_{2m} y_m + \delta_{tx_2} t \geq 0;$$

$$\frac{\partial \vec{D}_I(.)}{\partial y_m} = \beta_m + \sum_{m'=1}^{2} \beta_{mm'} y_{m'} - \gamma_{2m} g_{x_2} x_1 + \gamma_{2m} x_2 + \delta_{tym} t \leq 0 \quad (m = 1, 2).$$

Curvature restrictions can be imposed by ensuring that every principal minor of the Hessian matrix of odd order (even order) is non-positive (non-negative) — see, for example, Morey (1986). However, the US banking industry is highly regulated at both the federal and state levels, and different states change their regulatory restrictions at different times. This implies that the curvature condition would involve a bordered Hessian matrix that accounts for those regulatory restrictions. However, quantifying all those regulatory restrictions in the US banking industry is not an easy task. Furthermore, Barnett (2002) notes that the imposition of global curvature on the quadratic functional form may induce spurious violations of monotonicity. Therefore, directional distance functions are estimated subject to theoretical monotonicity only following Färe *et al.* (2005) and Feng *et al.* (2018).

The unit value, the observed input-output, and the optimal directional vector models are first estimated without imposing the monotonicity conditions. However, the monotonicity conditions with respect to labor and all outputs are violated for all models at most observations. Thus, to produce inference that is consistent with neoclassical microeconomic theory, all models are re-estimated with the monotonicity conditions imposed at each observation, by following the Bayesian procedure discussed in O'Donnell and Coelli (2005).

3.6.2 Technical Inefficiency Measures

Observation-specific posterior estimates of technical inefficiency are obtained from the posterior conditional mean of u. The average technical inefficiency measures for each sample year from the unit value, the observed input-output, and the optimal directional vector regularity-constrained models are summarized in Tables $3.2 - 3.4$, respectively.

As can be seen in Table 3.2, the estimated mean value of DTDF is equal to 181 in the unit value directional vector model. This average overall technical inefficiency measure indicates that each input should be contracted by 181 units of input, while each output should be expanded by 181 units of output on average for the bank to be technically efficient. While the estimated mean value of DIDF is equal to 80, the estimated mean value of DODF is equal to 149 in the unit value directional vector models. These average input and output technical inefficiency measures indicate that each input should be contracted by 80 units of input on average holding outputs fixed and each output should be expanded by 149 units of output on average holding inputs fixed for the bank to be technically efficient.

Comparing technical inefficiency measures across the unit value, the observed input-output, and the optimal directional vector models, the pre-specified unit value and observed input-output directional vector models which leave the endogeneity of inputs and outputs unaddressed produce lower estimates of technical inefficiency. This may result from the dependence of empirical estimates of directional distance functions on the choice of the directional vector.

It is apparent that the total average of input and output technical inefficiency measures are larger than the average overall technical inefficiency measures in the case of the unit value and the optimal directional vector models and smaller than the average overall technical inefficiency measures in the case of the observed input-output directional vector model. This implies that overall technical inefficiency does not equal the sum of input and output technical inefficiencies, as previous studies claim.

Note that the estimated value of DTDF is less than the estimated value of DIDF over the periods $2001-2003$, and less than the estimated value of DODF over the periods $2001-2008$ for the unit value directional vector models. The DTDF is also less than the estimated value of DODF in 2001 for the observed input-output directional vector model. Moreover, the DTDF is less than the estimated value of DIDF over the periods $2001-2005$ for the optimal directional vector model. This implies that models that measure technical inefficiency on one side of production tend to overestimate bank inefficiency measures. This suggests the importance of using models that incorporate both input and output inefficiencies.

Furthermore, output technical inefficiency is on average larger than input technical inefficiency in the unit value, the observed input-output, and the optimal directional vector models. This finding is consistent with Berger, Hancock, and Humphrey (1993) and English *et al.*(1993), who find that output inefficiency measures are as large or larger than input inefficiency measures. That is, most of the technical inefficiency in the US banking is in the form of loss of production, rather than overuse of inputs.

3.6.3 Results on the Interactive Effects

To focus on the relationships among input, output, and overall technical inefficiencies obtained from the systems of equations, consisting of DTDF and the interactive effect equations without and with the profit-maximizing first-order conditions, the estimated parameters of the DIDF and DODF for the three directional vector models are not discussed. The estimated parameters of the DTDF, their associated 95% Bayes intervals, and their SIF values from the unit value, the observed input-output, and the optimal directional vector regularity-constrained models are summarized in Tables $3.5-3.7$, respectively.

As can be seen in Tables $3.5-3.7$, UDTDF, VDTDF, and ODTDF models show that both input and output technical inefficiencies have significant positive effects on the overall technical inefficiency. However, the interactive effect between input and output technical inefficiencies δ_{IO} has a significant negative effect on the overall technical inefficiency. This

result is robust to alternative directional vectors and model specifications. Banks with larger values of the interactive effect tend to have a lower level of overall technical inefficiency which indicates that they are more efficient. This suggests that the overuse of inputs creates input technical inefficiency and has an effect on reducing (improving) output technical inefficiency (efficiency) and therefore improving overall technical efficiency. Intuitively, the overuse of inputs whether physical inputs involving overuse of labor or overuse of financial inputs involving overpayment of interest creates input technical inefficiency and has an effect on reducing (improving) output technical inefficiency (efficiency). The overuse of labor and the overpayment of interest may encourage banks to produce more loans to pay salaries for its employees and interest rates on deposits.

Similarly, the loss of production of outputs creates output technical inefficiency and has an effect on reducing (improving) input technical inefficiency (efficiency) and therefore improving overall technical efficiency. Intuitively, the loss of production of loans creates output technical inefficiency and has an effect on reducing (improving) input technical inefficiency (efficiency). The loss of revenue from loans may encourage banks to reduce the number of labor used in the production process or lower the interest rates paid on deposits.

The most clarifying insights come from comparing bank-specific input, output, and overall technical inefficiency measures over the years $2001 - 2015$. Figures $3.1 - 3.4$ show the interactions between input and output technical inefficiencies obtained based on the unit value directional vector models. Figures $3.5 - 3.8$ show the interactions between input and output technical inefficiencies obtained based on the observed input-output directional vector models. Figures $3.9 - 3.12$ show the interactions between input and output technical inefficiencies obtained based on the optimal directional vector models. It is apparent from all these figures that the increase in the output technical inefficiency reflects on a reduction on the input technical inefficiency and vice versa.

Differences in the magnitude of the interactive effects are observed when using a pre-

specified unit value and observed input-output directional vectors and endogenous directional vector. In particular, the interactive effects obtained based on the unit value and the observed input-output directional vector models appear to be of low magnitudes. On the other hand, considering banks heterogeneity in the directional vector, the interactive effect obtained based on the optimal directional vector model appears to be of higher magnitude. This may result from the dependence of empirical estimates of directional distance functions on the choice of the directional vector or suggest a relationship between technical and allocative inefficiency since eliminating technical inefficiency in this case requires choosing a directional vector that projects any inefficient bank to the optimal allocation of input-output vector given relative market prices. Further research is needed to investigate this issue. Table 3.8 presents the minimum, maximum, and mean of the estimates of optimal directional parameters. Letting the data select the directional vectors produces estimates of the directional parameters with a range of variation across banks. Precisely, the estimates of the directional parameters of $g_{x_{1i}}$ range from 0.85 to 1.12 in the DTDF-profit system model. While the estimates of the directional parameters of $g_{x_{2i}}$ range from 0.81 to 1.20 in the DIDF-cost system model, it ranges from 0.83 to 1.14 in the DTDF-profit system model. While the estimates of the directional parameters of $g_{y_{2i}}$ range from 0.56 to 1.42 in the DODF-revenue system model, it ranges from 0.80 to 1.17 in the DTDF-profit system model.

The parameter estimates of technical change δ_t, and δ_{tt} obtained based on the unit value, the observed input-output, and the optimal directional vector models appear to be of high magnitude. Specifically, it indicates significant technological advancement by the US banking industry over the period $2001 - 2015$, which seems realistic given the recent advances in information technologies and its effects on the US banking industry.

3.7 Conclusion

This paper investigates the relationships among input, output, and overall technical inefficiencies in US commercial banking over the period from 2001 to 2015. Specifically, these inefficiencies are estimated separately using DIDF, DODF, and DTDF, respectively. All of these inefficiencies are estimated with the three commonly used directional vectors; the unit value, the observed input-output, and the optimal directional vectors. The latter addresses the endogeneity of inputs and outputs by using systems of equations, consisting of DIDF (DODF, or DTDF) with the cost (revenue, or profit) minimizing (maximizing) first-order conditions, respectively. Furthermore, the directional vectors of these models are allowed to be endogenous and vary across banks to account for heterogeneity across banks. All of these models are estimated using Bayesian estimation with the monotonicity conditions imposed at each observation in order to produce inference that is consistent with neoclassical microeconomic theory.

To investigate the relationships among input, output, and overall technical inefficiencies, I model the overall technical inefficiency as a linear function of a vector of explanatory bank-specific variables that includes input technical inefficiency, output technical inefficiency, and a term capturing the interactions between them, following Battese and Coelli (1995). These bank-specific variables that determine overall technical inefficiency are estimated simultaneously with the variables that determine the frontier.

In providing a comparison of the three estimates of technical inefficiencies in the case of the unit value, the observed input-output, and the optimal directional vector models, the empirical results show that overall technical inefficiency does not equal the sum of input and output technical inefficiencies, as previous studies claim. It equals the sum of input and output technical inefficiencies plus an interactive effect component which captures the interactions between them, where the increase in the output technical inefficiency reflects on a reduction on the input technical inefficiency and vice versa.

The results also show that both input and output technical inefficiencies have significant positive effects on the overall technical inefficiency. However, the interactive effect between input and output technical inefficiencies has a significant negative effect on the overall technical inefficiency. This result is robust to alternative directional vectors and model specifications. This suggests that the overuse of inputs creates input technical inefficiency and has an effect on reducing (improving) output technical inefficiency (efficiency) and therefore improving overall technical efficiency. Similarly, the loss of production of outputs creates output technical inefficiency and has an effect on reducing (improving) input technical inefficiency (efficiency) and therefore improving overall technical efficiency. Ignoring the interactive effect between input and output technical inefficiencies results in a decomposition of overall technical inefficiency into input and output components that are significantly different from the ones that incorporate it.

The results also indicate that the value of the interactive effect between input and output technical inefficiencies depends on the choice of the directional vector in which the data are projected on the frontier and whether quantities and prices are taken into consideration. These results are quite significant, since these inefficiency components have different implications for bank performance, suggesting that the adjustability of both inputs and outputs is required for the improvement of bank efficiency.

Table 3.1: Data Summary Statistics

Variable	Mean	5th Percentile	Median	95th Percentile	Standard Deviation
			Financial Assets and Liabilities		
x_1	333.6545	70.0000	281.0000	741.5000	239.2208
x_2	1211.1000	312.4149	1030.9000	2701.0000	810.5971
y_1	911.6341	173.4561	750.6042	2179.6000	686.9423
y_2	402.4122	68.2664	328.4993	977.2095	308.4258
Total Assets	1350.7000	359.9879	1140.1000	3048.6000	918.6617
			Bank-Specific Price		
w_1	59.6961	37.4599	54.5960	91.3389	24.5955
w_2	0.0148	0.0014	0.0133	0.0341	0.0104
p_1	0.0611	0.0388	0.0591	0.0830	0.0227
p_2	0.0625	0.0217	0.0491	0.1435	0.0628
			Market Price		
w_1	58.0550	50.7531	56.8821	63.9205	3.9677
w_2	0.0148	0.0031	0.0139	0.0318	0.0093
p_1	0.0612	0.0457	0.0611	0.0807	0.0096
p_2	0.0578	0.0473	0.0585	0.0728	0.0079

Note: x_1, the quantity of labor; x_2, the quantity of purchased funds and deposits; y_1 total loans which includes consumer loans, commercial and industrial loans, and real estate loans; y_2, securities which includes all non-loan financial assets; w_1 and w_2 are the prices of x_1 and x_2, respectively; p_1 and p_2 are the prices of y_1 and y_2, respectively. All variables but labor and input and output prices are in thousands of real 2005 US dollars.

Table 3.2: Average Technical Inefficiency over Time Based on the Unit Value Directional Regularity-Constrained Models

Year	2001	2002	2003	2004	2005	2006	2007	2008
UDIDF	11.8294	21.1690	31.1416	41.7216	51.9758	62.5347	72.8167	81.9493
UDODF	38.0998	54.6814	70.2523	85.5096	101.4479	116.3714	130.8457	150.5572
UDTDF	5.9640	16.8900	31.0910	48.5342	69.1852	93.1282	120.3147	150.4222

Year	2009	2010	2011	2012	2013	2014	2015	Average
UDIDF	89.1105	98.1810	107.1871	116.5829	127.2603	134.9243	145.2475	79.5754
UDODF	166.3904	184.6105	195.5057	211.9362	224.2782	248.7085	255.3833	148.9719
UDTDF	184.0269	220.6380	260.8635	304.0676	350.6689	399.8923	453.1177	180.5870

Table 3.3: Average Technical Inefficiency over Time Based on the Observed Input-Output Directional Regularity-Constrained Models

Year	2001	2002	2003	2004	2005	2006	2007	2008
VDIDF	1.8966	3.3405	5.6046	8.6740	12.4314	16.9193	22.0298	27.4962
VDODF	13.0282	15.6542	19.2972	24.4736	31.0880	38.8666	48.0702	60.5054
VDTDF	4.3780	16.0843	30.5473	47.7446	67.7554	90.6375	116.4475	145.0767

Year	2009	2010	2011	2012	2013	2014	2015	Average
VDIDF	33.5365	40.4244	48.2354	56.7076	66.1682	75.5638	86.3371	33.6910
VDODF	72.8146	87.7846	101.5196	118.1862	135.8242	159.1207	178.3353	73.6379
VDTDF	176.8909	211.8751	250.1111	291.5732	336.4630	384.5791	436.7667	173.7954

Table 3.4: Average Technical Inefficiency over Time Based on the Optimal Directional Regularity-Constrained Models

Year	2001	2002	2003	2004	2005	2006	2007	2008
ODIDF	110.3720	132.2072	155.7102	173.8357	196.3768	221.2248	246.0232	265.5990
ODODF	44.5414	74.2772	107.1022	143.4714	184.6959	228.6927	275.8839	331.7615
ODTDF	72.2921	90.7688	116.7214	148.2288	189.0018	237.5864	293.8951	357.0892

Year	2009	2010	2011	2012	2013	2014	2015	Average
ODIDF	288.3704	317.0758	345.1988	382.4024	415.2908	443.7975	474.5047	277.8659
ODODF	387.7037	450.7542	510.5639	580.8563	650.9366	735.9464	806.1380	367.5550
ODTDF	428.9286	509.1689	597.3874	694.1378	798.7750	909.8987	1029.2937	431.5449

Table 3.5: Parameter Estimates for the Regularity-Constrained UDTDF Model

Parameter	Estimate	95% Bayes Interval	SIF
		The Frontier	
α_0	10.7097	(10.5691, 10.9404)	28.9949
α_1	0.3172	(0.3119, 0.3212)	26.2283
α_2	0.2938	(0.2892, 0.2986)	26.6859
β_2	-0.1710	(-0.1791, -0.1641)	23.4118
δ_t	11.3041	(11.1091, 11.5001)	30.0065
α_{11}	-0.0610	(-0.0643, -0.0580)	12.6863
α_{12}	0.0735	(0.0701, 0.0770)	14.2537
α_{22}	-0.0685	(-0.0711, -0.0645)	13.3931
β_{22}	-0.0141	(-0.0451, 0.0238)	5.3655
δ_{tt}	3.4889	(3.4595, 3.5139)	29.8515
γ_{12}	-0.0089	(-0.0346, 0.0245)	5.5158
γ_{22}	-0.0196	(-0.0437, 0.0221)	7.3338
δ_{tx_1}	-0.0026	(-0.0169, 0.0135)	12.0490
δ_{tx_2}	0.0026	(-0.0147, 0.0165)	12.4428
δ_{ty_2}	0.0014	(-0.0116, 0.0143)	11.5602
		The Interactive Effect Equation	
δ_0	2.0163	(2.0121, 2.0205)	29.2707
δ_I	5.7224	(5.7002, 5.7444)	29.8449
δ_O	5.7957	(5.7527, 5.8305)	30.0774
δ_{IO}	-1.2044	(-1.2258, -1.1833)	30.0231

Table 3.6: Parameter Estimates for the Regularity-Constrained VDTDF Model

Parameter	Estimate	95% Bayes Interval	SIF
		The Frontier	
α_0	8.3265	(8.3214, 8.3333)	53.1989
α_1	0.1966	(0.1921, 0.2004)	53.1405
α_2	0.3417	(0.3354, 0.3457)	53.1031
β_2	-0.1843	(-0.1881, -0.1823)	51.9249
δ_t	7.8266	(7.8234, 7.8304)	52.0048
α_{11}	-0.0348	(-0.0375, -0.0317)	52.3867
α_{12}	0.0693	(0.0633, 0.0770)	53.2463
α_{22}	-0.0640	(-0.0661, -0.0618)	51.5508
β_{22}	-0.0023	(-0.0064, 0.0016)	52.9642
δ_{tt}	3.1232	(3.1164, 3.1279)	51.2509
γ_{12}	-0.0390	(-0.0410, -0.0362)	51.3234
γ_{22}	0.0001	(-0.0019, 0.0019)	50.6460
δ_{tx_1}	0.0181	(0.0149, 0.0211)	51.8347
δ_{tx_2}	-0.0054	(-0.0092, -0.0019)	52.4963
δ_{ty_2}	-0.0090	(-0.0112, -0.0067)	51.2171
		The Interactive Effect Equation	
δ_0	2.4684	(2.4665, 2.4704)	49.9582
δ_I	1.5693	(1.5665, 1.5741)	53.0057
δ_O	3.5000	(3.4945, 3.5038)	53.1133
δ_{IO}	-0.8709	(-0.8740, -0.8676)	52.4857

Table 3.7: Parameter Estimates for the Regularity-Constrained ODTDF Model

Parameter	Estimate	95% Bayes Interval	SIF
		The Frontier	
α_0	10.4904	(10.4588, 10.5220)	19.3941
α_1	0.2325	(0.2281, 0.2370)	17.0840
α_2	0.2977	(0.2935, 0.3020)	17.4929
β_2	-0.2349	(-0.2384, -0.2315)	17.1590
δ_t	9.8543	(9.8116, 9.8996)	20.0275
α_{11}	-0.0525	(-0.0554, -0.0498)	12.7861
α_{12}	0.0671	(0.0634, 0.0709)	13.7315
α_{22}	-0.0662	(-0.0688,-0.0638)	12.6097
β_{22}	-0.0138	(-0.0448, 0.0257)	4.1099
δ_{tt}	9.0387	(8.9692, 9.1077)	19.1570
γ_{12}	-0.0055	(-0.0314, 0.0271)	4.5671
γ_{22}	-0.0215	(-0.0423, 0.0194)	6.7070
δ_{tx_1}	0.0017	(-0.0122, 0.0147)	7.2960
δ_{tx_2}	0.0029	(-0.0140, 0.0166)	9.0611
δ_{ty_2}	0.0049	(-0.0113, 0.0170)	8.8610
		The Interactive Effect Equation	
δ_0	1.2458	(1.2276, 1.2639)	19.6232
δ_I	0.9452	(0.9294, 0.9613)	19.9741
δ_O	1.3628	(1.3350, 1.3905)	20.0692
δ_{IO}	-2.5344	(-2.5660, -2.5026)	20.0838

Table 3.8: Estimates of Optimal Directional Parameters

Direction	ODIDF			ODODF			ODTDF		
Vector	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
$g_{x_{1i}}$	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.9928	0.8514	1.1156
$g_{x_{2i}}$	1.0063	0.8059	1.1996	0.0000	0.0000	0.0000	1.0049	0.8340	1.1441
$g_{y_{1i}}$	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$g_{y_{2i}}$	0.0000	0.0000	0.0000	1.0049	0.5568	1.4208	0.9950	0.8022	1.1661

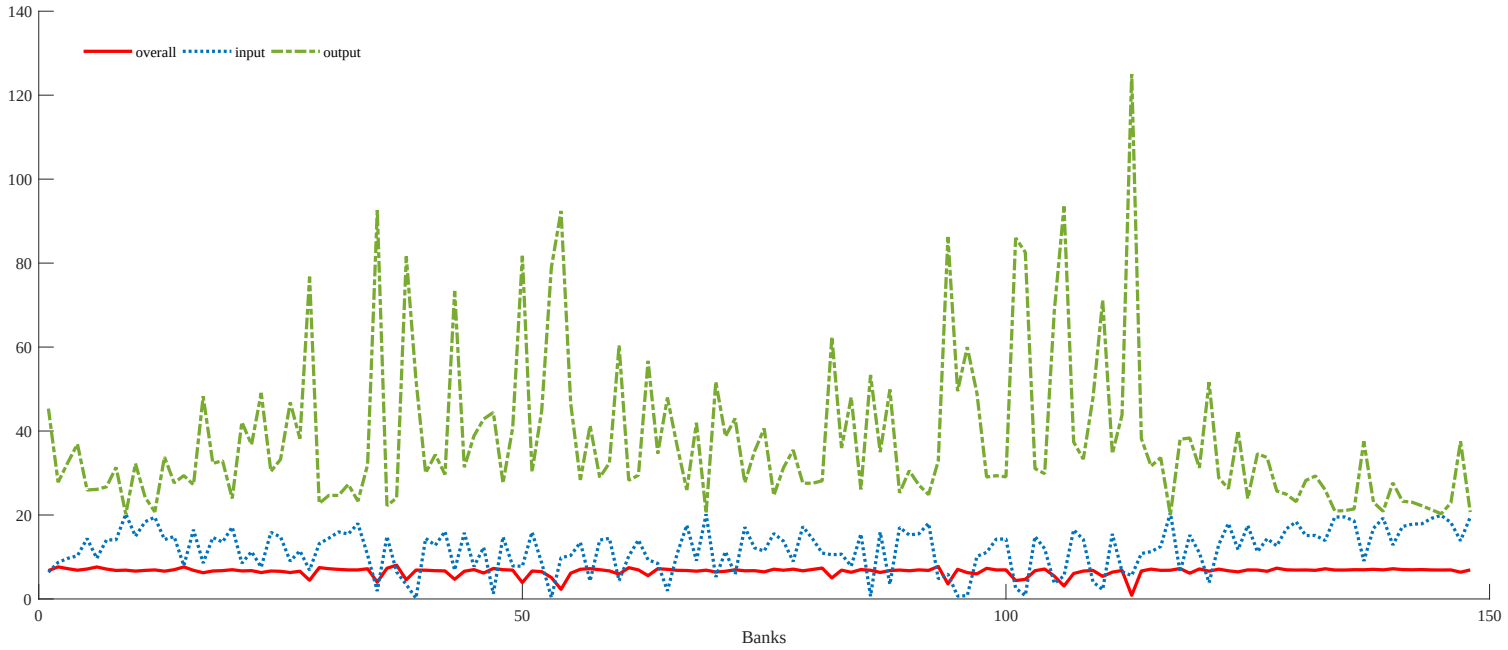

Figure 3.1: Technical Inefficiency Measures Based on the Unit Value Directional Vector in 2001

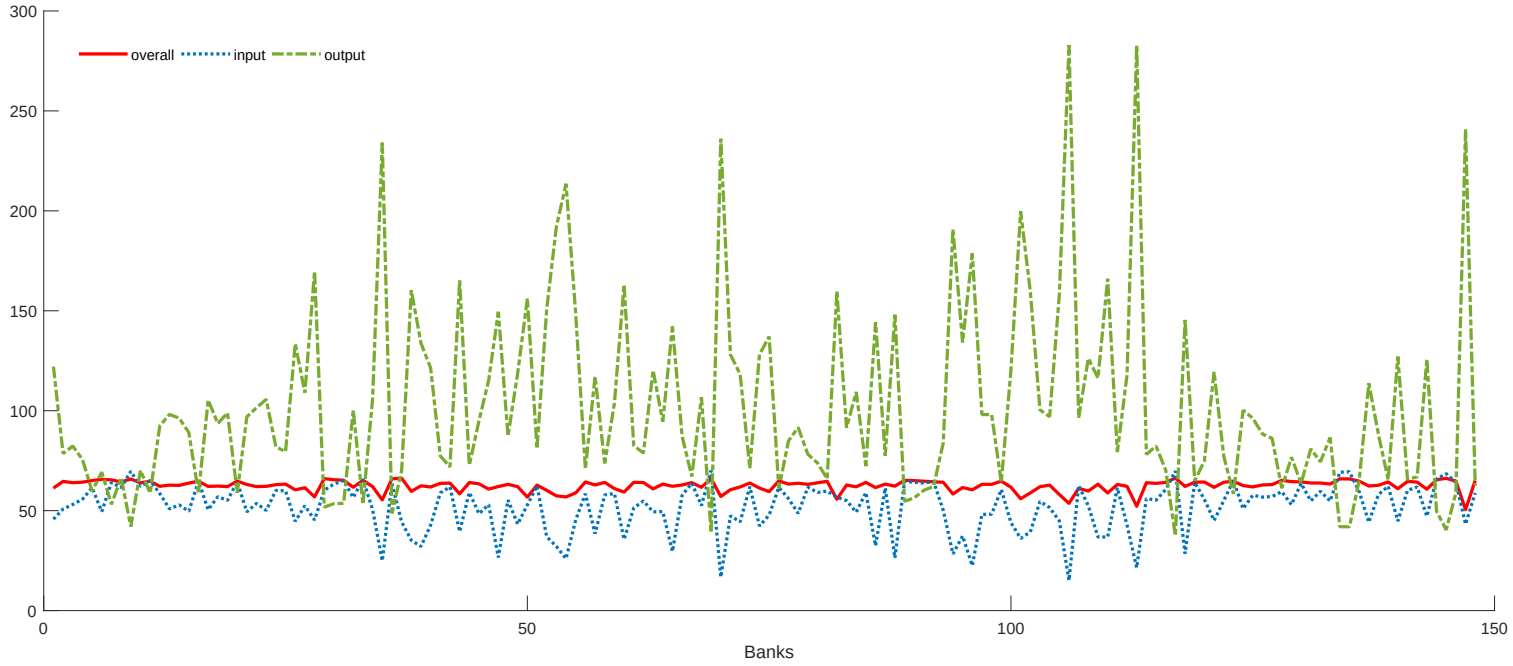

Figure 3.2: Technical Inefficiency Measures Based on the Unit Value Directional Vector in 2005

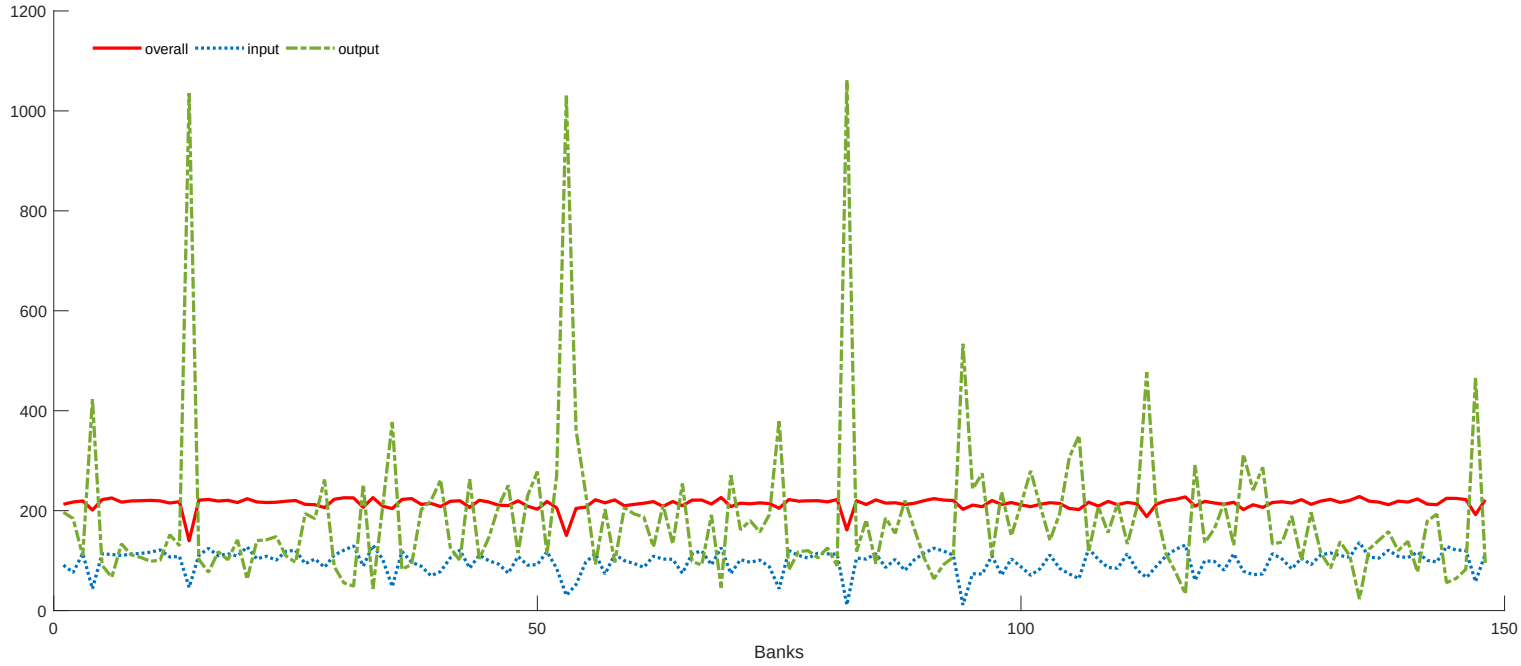

Figure 3.3: Technical Inefficiency Measures Based on the Unit Value Directional Vector in 2010

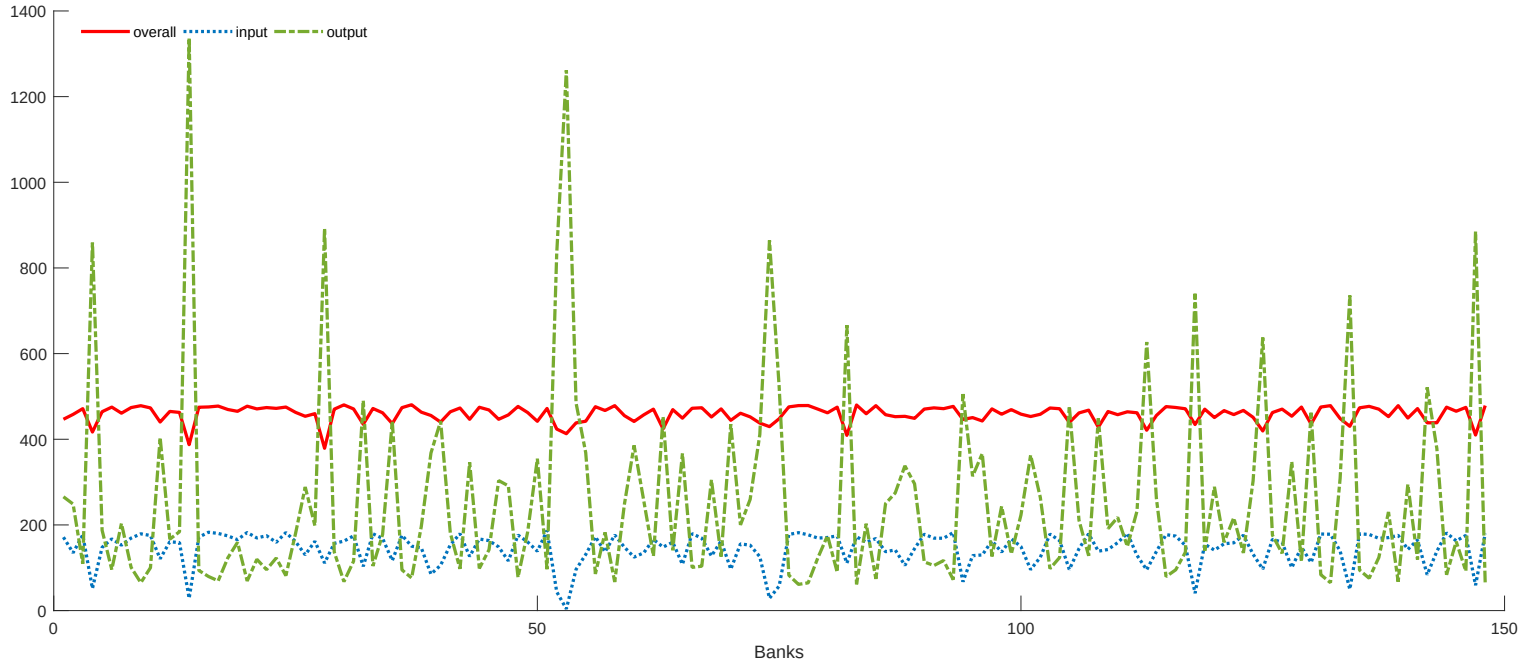

Figure 3.4: Technical Inefficiency Measures Based on the Unit Value Directional Vector in 2015

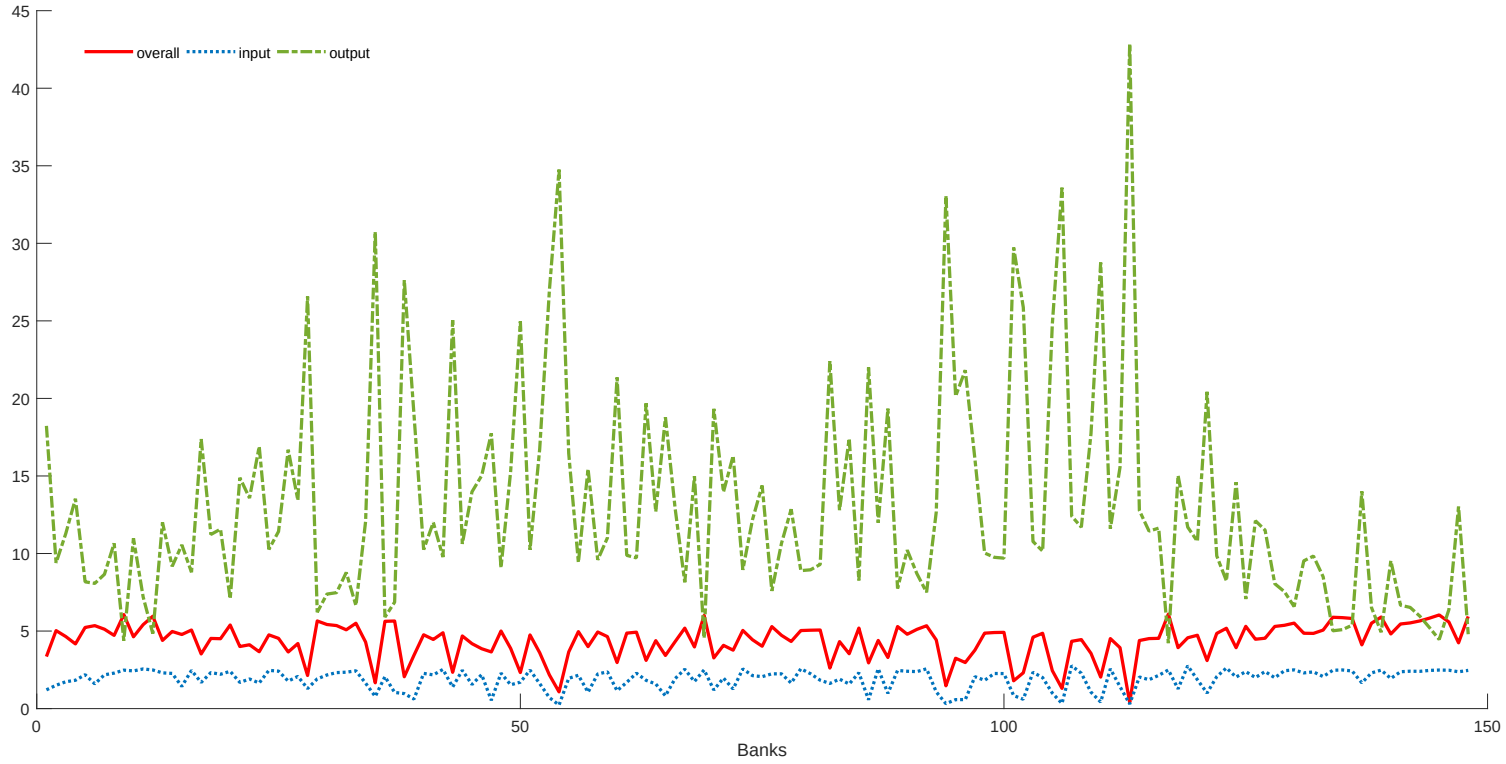

Figure 3.5: Technical Inefficiency Measures Based on the Observed Input-Output Directional Vector in 2001

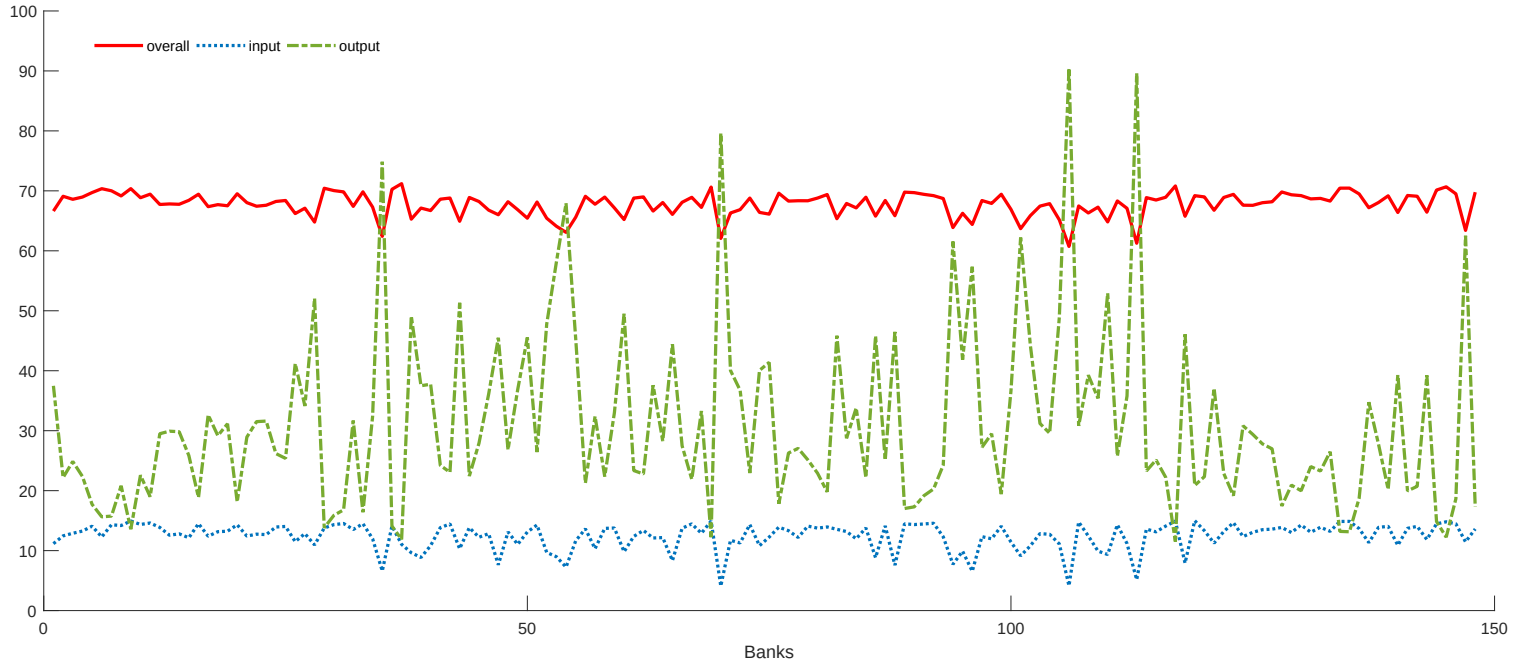

Figure 3.6: Technical Inefficiency Measures Based on the Observed Input-Output Directional Vector in 2005

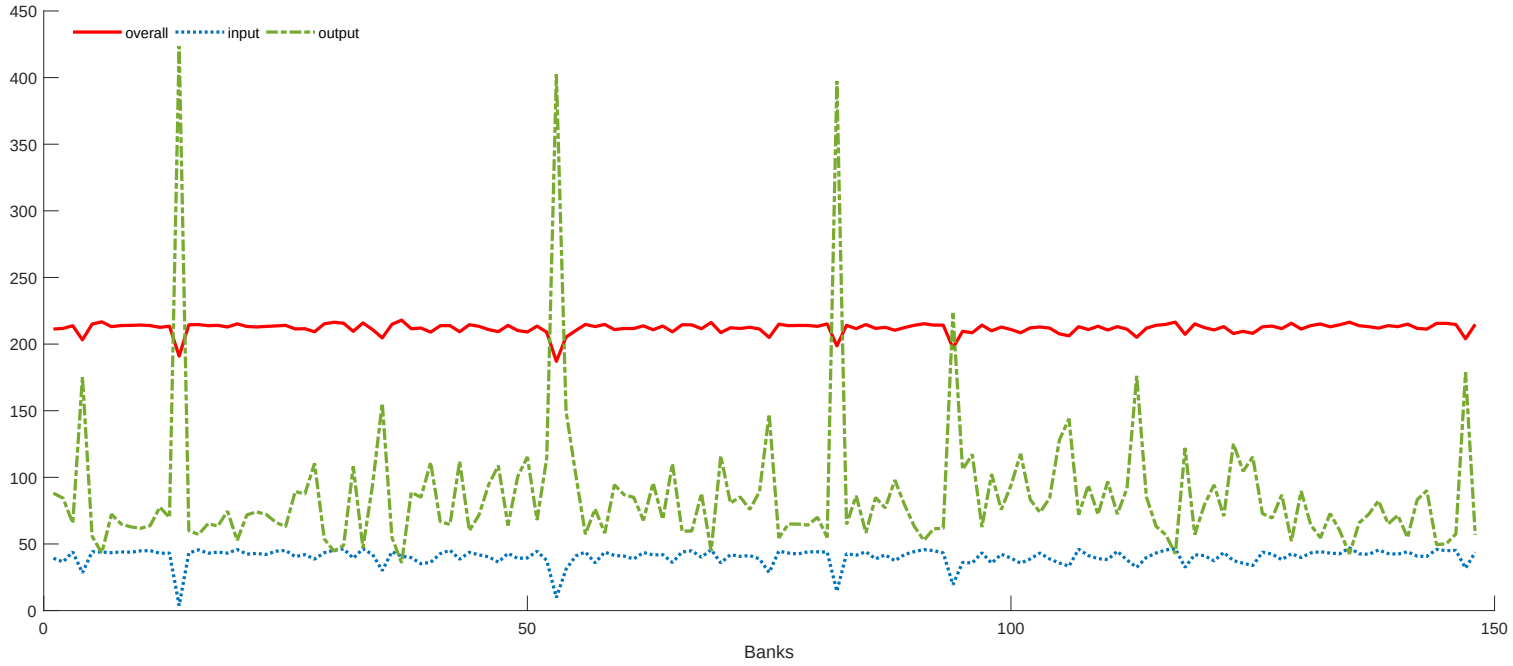

Figure 3.7: Technical Inefficiency Measures Based on the Observed Input-Output Directional Vector in 2010

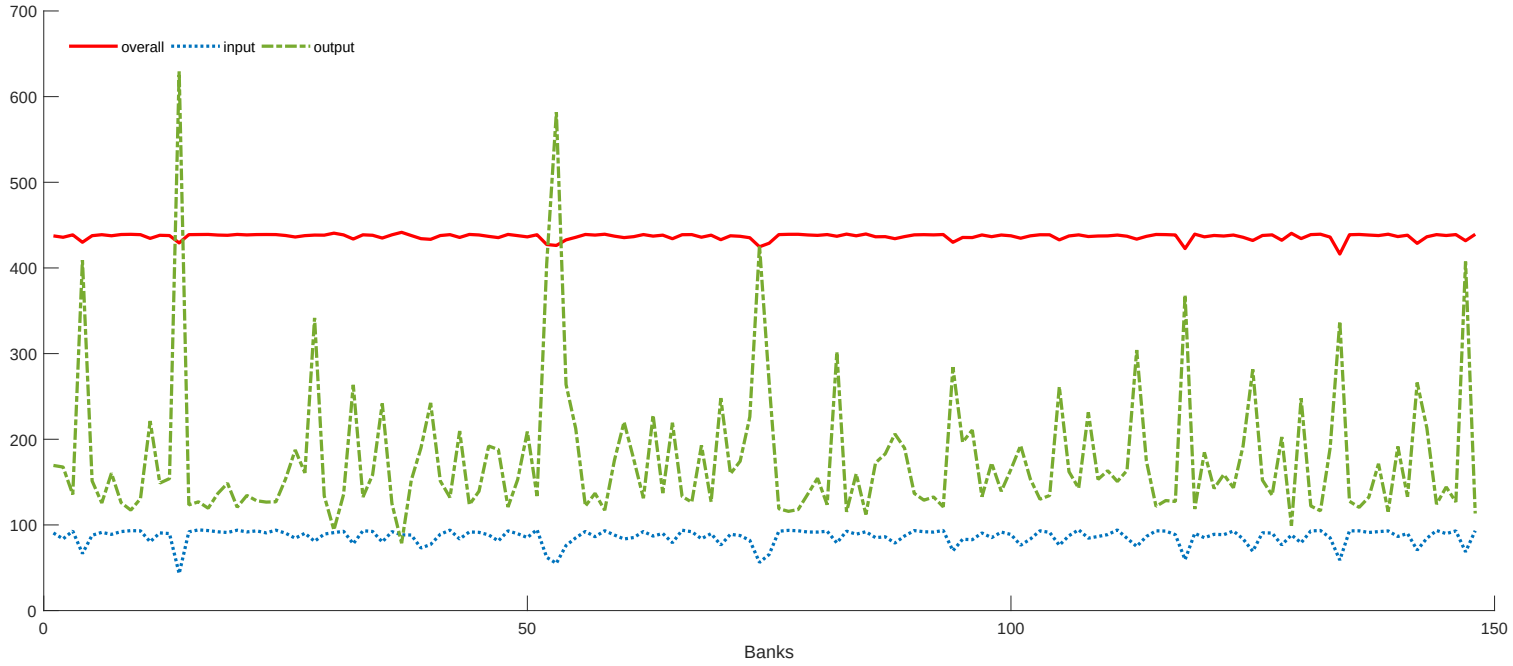

Figure 3.8: Technical Inefficiency Measures Based on the Observed Input-Output Directional Vector in 2015

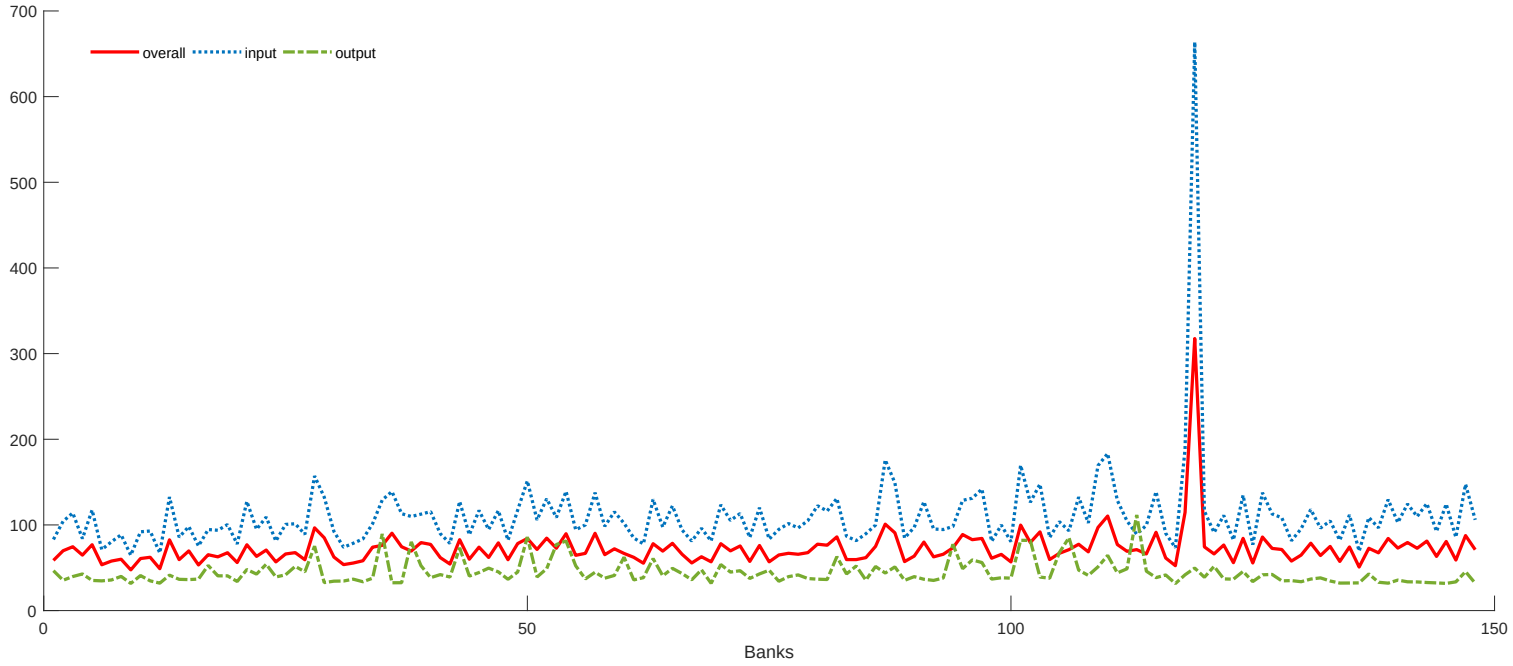

Figure 3.9: Technical Inefficiency Measures Based on the Optimal Directional Vector in 2001

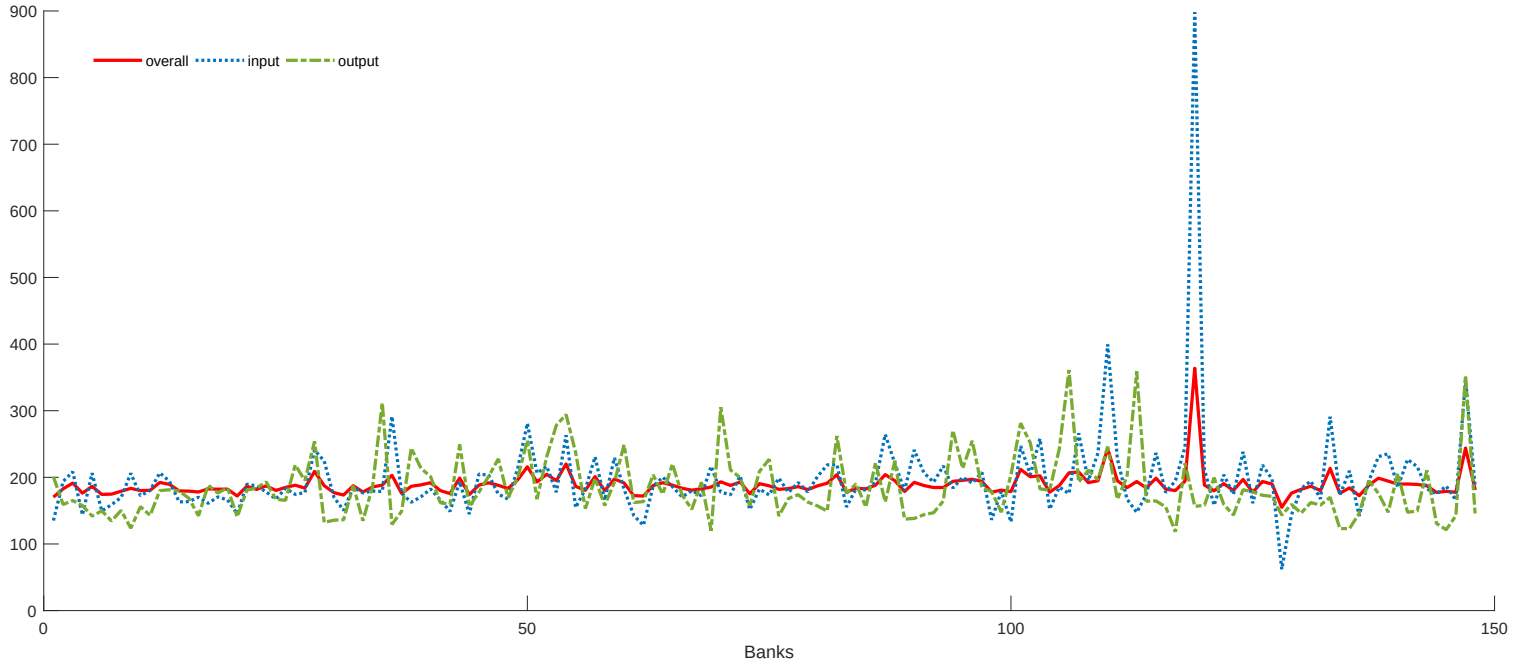

Figure 3.10: Technical Inefficiency Measures Based on the Optimal Directional Vector in 2005

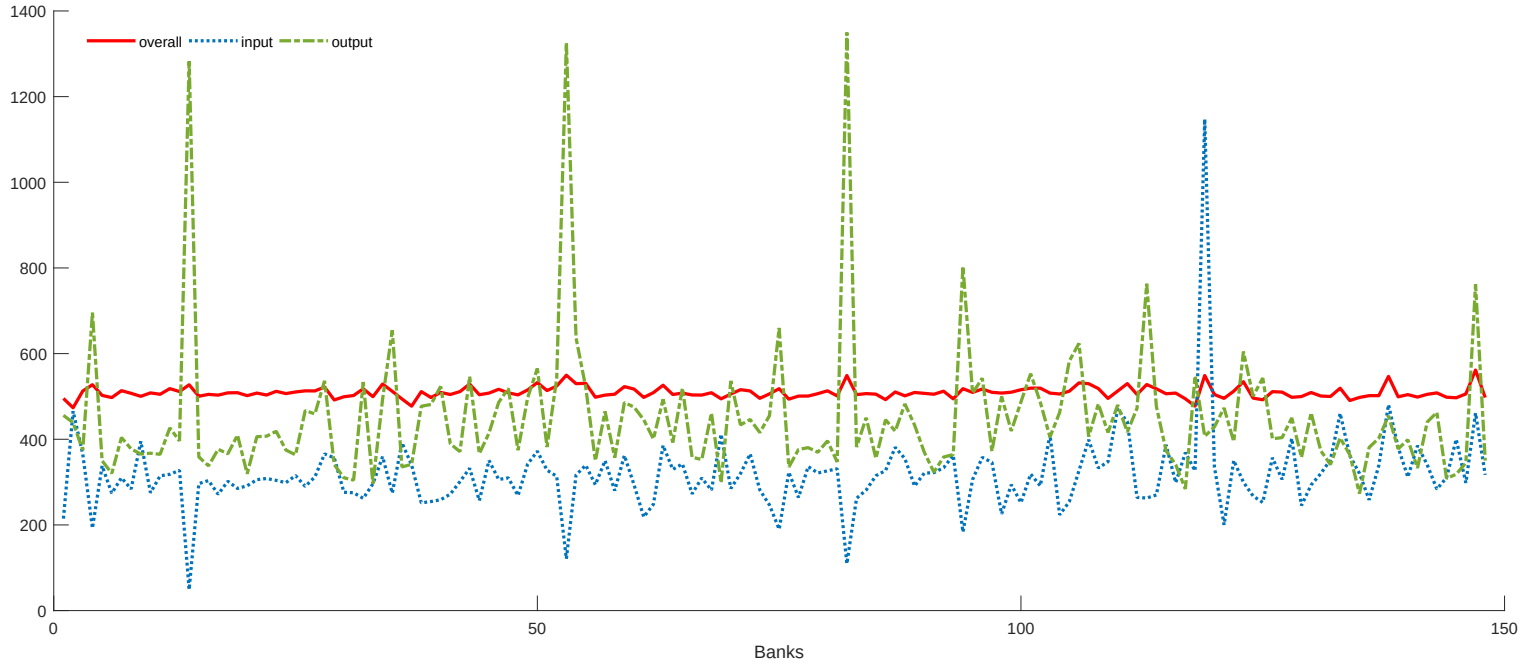

Figure 3.11: Technical Inefficiency Measures Based on the Optimal Directional Vector in 2010

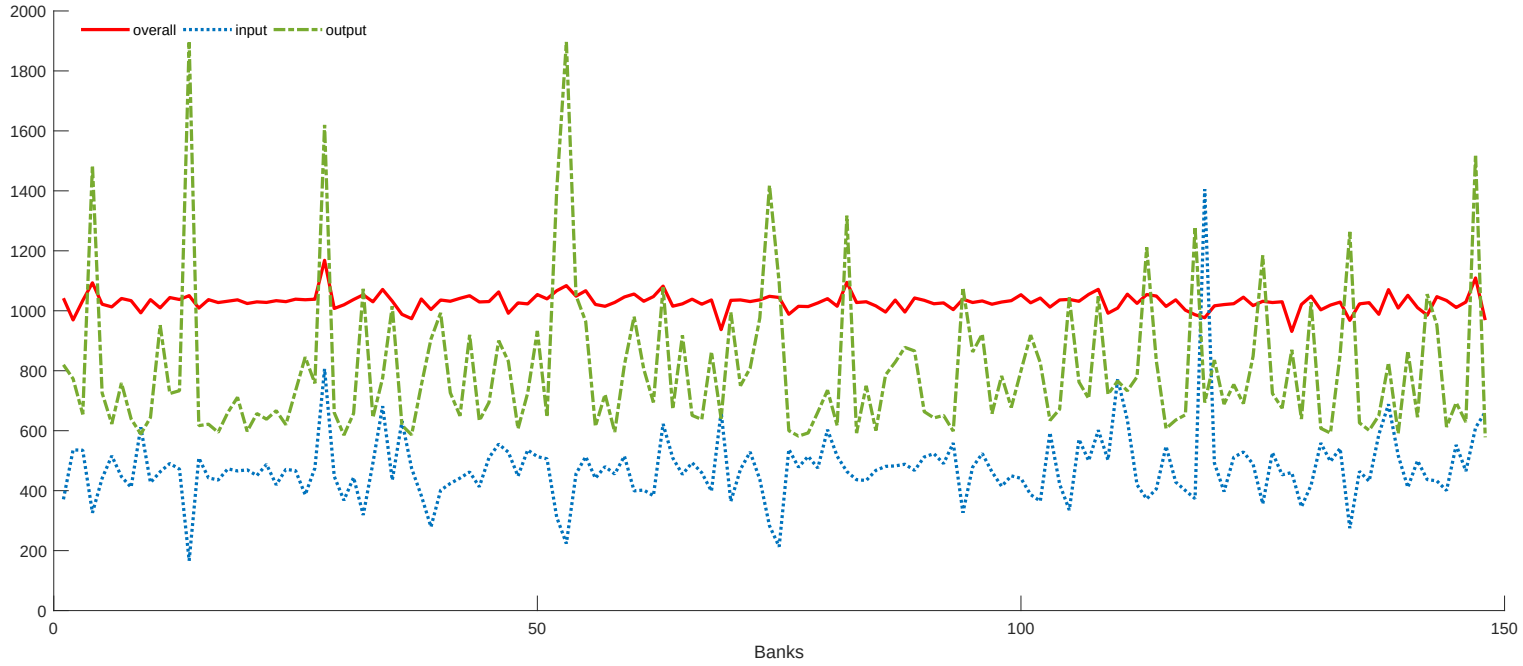

Figure 3.12: Technical Inefficiency Measures Based on the Optimal Directional Vector in 2015

Bibliography

[1] Ahn, S.C., Sickles, R.C., 2000. Estimation of long-run inefficiency levels: A dynamic frontier approach. *Econometric Reviews* 19, 461–492.

[2] Aigner, D.J., Lovell, C.A.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production functions. *Journal of Econometrics* 6, 21–37.

[3] Akhavein, J.D., Swamy, P.A., Taubman, S.B., 1997. A general Mmethod of deriving the inefficiencies of banks from a profit function. *Journal of Productivity Analysis* 8, 71–93.

[4] Almanidis, P., 2013. Accounting for heterogeneous technologies in the banking industry: A time-varying stochastic frontier model with threshold effects. *Journal of Productivity Analysis* 39, 191–205.

[5] Almanidis, P., Qian, J., Sickles, R., 2014. Stochastic frontier models with bounded inefficiency. Festschrift in honor of Peter Schmidt 47–81.

[6] Altunbas, Y., Evans, L., Molyneux, P., 2001. Bank ownership and efficiency. *Journal of Money, Credit and Banking* 33, 926–954.

[7] Alvarez, I., Niemi, J., Simpson, M., 2014. Bayesian inference for a covariance matrix. Conference on Applied Statistics in Agriculture 71–82. Available at http://works.bepress.com/jarad_niemi/2/

[8] Arellano-Valle, R.B., Azzalini, A., 2006. On the unification of families of skew-normal distribution. *Scandinavian Journal of Statistics* 33, 561–574.

[9] Assaf, A.G., Matousek, R., Tsionas, E.G., 2013. Turkish bank efficiency: Bayesian estimation with undesirable outputs. *Journal of Banking and Finance* 37, 506–517.

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

[11] Atkinson, S.E., Dorfman, J.H., 2005. Bayesian measurement of productivity and efficiency in the presence of undesirable outputs: Crediting electric utilities for reducing air pollution. *Journal of Econometric* 126, 445–468.

[12] Atkinson, S.E., Primont, D., 2002. Stochastic estimation of firm technology, inefficiency, and productivity growth using shadow cost and distance functions. *Journal of Econometric* 108, 203–225.

[13] Atkinson, S.E., Primont, D., Tsionas, M.G., 2018. Statistical inference in efficient production with bad inputs and outputs using latent prices and optimal directions. *Journal of Econometrics* 204, 131–146.

[14] Atkinson, S.E., Tsionas, M.G., 2016. Directional distance functions: Optimal endogenous directions. *Journal of Econometric* 190, 301–314.

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

[16] Azzalini A., 1985. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* 12, 171–178.

[17] Baccouche, R., Kouki, M., 2003. Stochastic production frontier and technical inefficiency: A sensitivity analysis. *Econometric Reviews* 22, 79–91.

[18] Bai, J., 1997. Estimating multiple breaks one at a time. *Economic Theory* 13, 315–352.

[19] Barnett, W.A., 1977. Pollak and Wachter on the household production function approach. *Journal of Political Economy* 85, 1073–1082.

[20] Barnett, W.A., 1983. Definitions of "second order approximation" and of "flexible functional form". *Economics Letters* 12, 31–35.

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

[22] Barnett, W.A., Lee, Y.W., 1985. The global properties of the minflex Laurent, generalized Leontief, and translog flexible functional forms. *Econometrica* 53, 1421–1437.

[23] Barnett, W.A., Lee, Y.W., Wolfe, M.D., 1985. The three-dimensional global properties of the minflex Laurent, generalized Leontief, and translog flexible functional forms. *Journal of Econometrics* 30, 3–31.

[24] Barnett, W.A., Lee, Y.W., Wolfe, M.D., 1987. The global properties of the two minflex Laurent flexible functional forms. *Journal of Econometrics* 36, 281–298.

[25] Barnett, W.A., Pasupathy, M., 2003. Regularity of the generalized quadratic production model: A counterexample. *Econometric Reviews* 22, 135–154.

[26] Barros, P.C., Managi, S., Matousek, R., 2012. The technical efficiency of the Japanese banks: Non-radial directional performance measurement with undesirable output. *Omega* 40, 1–8.

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

[28] Battese, G.E., 1992. Frontier production functions and technical efficiency: A survey of empirical applications in agricultural economics. *Agricultural Economics* 7, 185–208.

[29] Battese, G.E., Coelli, T.J., 1988. Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data. *Journal of Econometrics* 38, 387–399.

[30] Battese, G.E., Coelli T.J., 1992. Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India. *Journal of Productivity Analysis* 3, 153–169.

[31] Battese, G.E., Coelli T.J., 1995. A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics* 20, 325–332.

[32] Battese, G.E., Coelli T.J., Colby, T., 1989. Estimation of frontier production functions and the efficiencies of Indian farms using panel data from ICRISTAT's village level studies. *Journal of Quantitative Economics* 5, 327–348.

[33] Battese, G.E., Cora, G.S., 1977. Estimation of a production frontier model: with application to the pastoral zone of eastern Australia. *Australian Journal of Agricultural Economics* 21, 169–179.

[34] Battese, G.E., Tessema, G.A., 1993. Estimation of stochastic frontier production functions with time-varying parameters and technical efficiencies using panel data from Indian villages. *Agricultural Economics* 9, 313–333.

[35] Bauer, P., 1990. Recent developments in the econometric estimation of frontiers. *Journal of Econometrics* 46, 39–56.

[36] Beckers, D.E., Hammond, C.J., 1987. A tractable likelihood function for the normal-gamma stochastic frontier model. *Economics Letters* 24, 33–38.

[37] Bera, A.K., Sharma, S.C., 1999. Estimating production uncertainty in stochastic frontier production function models. *Journal of Productivity Analysis* 12, 187–210.

[38] Berger, A.N., DeYoung, R., 1997. Problem loans and cost efficiency in commercial banks. *Journal of Banking and Finance* 21, 849–870.

[39] Berger, A., Hancock, D., Humphrey, D., 1993. Bank efficiency derived from the profit function. *Journal of Banking and Finance* 17, 317–347.

[40] Berger, A.N., Humphrey, D.B., 1997. Efficiency of financial institutions: international survey and directions for future research. *European Journal of Operational Research* 98, 175–212.

[41] Berger, A., Hunter, W., Timme, S., 1993. The efficiency of financial institutions: A review and preview of research, past, present and future. *Journal of Banking and Finance* 17, 221–249.

[42] Berger, A., Mester, L., 1997. Inside the black box: What explains differences in the efficiencies of financial institutions? *Journal of Banking and Finance* 21, 895–947.

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

[44] Berndt, E.R., Christensen, L.R., 1973. The translog function and the substitution of equipment, structures and labor in U.S. manufacturing 1929-1968. *Journal of Econometrics* 1, 81–l14.

[45] Bos, J.W.B., Economidou, C., Koetter, M., 2010. Technology clubs, R&D and growth patterns: Evidence from EU manufacturing. *European Economic Review* 54, 60–79.

[46] Bos, J.W.B., Economidou, C., Koetter, M., Kolari, J.W., 2010. Do all countries grow alike? *Journal of Development Economics* 91, 113–127.

[47] Bos, J.W.B., Schmiedel, H., 2007. Is there a single frontier in a single European banking market? *Journal of Banking and Finance* 31, 2081–2102.

[48] Brown, J.A., Glennon, C., 2000. Cost structures of banks grouped by strategic conduct. *Applied Economics* 32, 1591–1605.

[49] Butler, J., Moffitt, R., 1982. A computationally efficient quadrature procedure for the one factor multinomial probit model. *Econometrica* 50, 761–764.

[50] Caiazza, S., Pozzolo, A.F., Trovato, G., 2016. Bank efficiency measures, M&A decision and heterogeneity. *Journal of Productivity Analysis* 46, 25–41.

[51] Cartinhour, J., 1990. One-dimensional marginal density functions of a truncated multivariate normal density function. *Journal of Communications in Statistics - Theory and Methods* 19, 197–203.

[52] Casella, G., George, E., 1992. Explaining the gibbs Ssampler. *The American Statistician* 46, 167–174.

[53] Casu, B., Molyneux, P., 2003. A comparative study of efficiency in European banking. *Applied Economics* 35, 1865–1876.

[54] Caudill, S.B., Ford, J.M., 1993. Biases in frontier estimation due to heteroskedasticity. *Economics Letters* 41, 17–20.

[55] Caudill, S.B., Ford, J.M., Gropper, D.M., 1995. Frontier estimation and firm-specific inefficiency measures in the presence of heteroskedasticity. *Journal of Business and Economic Statistics* 13, 105–111.

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

[57] Chambers, R.G., 1998. Input and output indicators. In: Färe, R., Grosskopf, S., Russell, R.R. (Eds), *Index Numbers: Essays in Honour of Sten Malmquist*. Kluwer Academic Publishers, Boston.

[58] Chambers, R.G., Chung, Y., Färe, R., 1996. Benefit and distance functions. *Journal of Economic Theory* 70, 407–419.

[59] Chambers, R.G., Chung, Y., Färe, R., 1998. Profit, directional distance functions, and Nerlovian efficiency. *Journal of Optimization Theory and Applications* 98, 351–364.

[60] Chambers, R., Färe, R., Grosskopf, S., Vardanyan, M., 2013. Generalized quadratic revenue functions. *Journal of Econometrics* 173, 11–21.

[61] Chang, Y., Park, J., Phillips, P., 2001. Nonlinear econometric models with cointegrated and deterministically trending regressors. *The Econometrics Journal* 4, 1–36.

[62] Chen, M.–H., Shao, Q.–M., Ibrahim, J.G., 2000. *Monte Carlo Methods in Bayesian Computation*. Springer, New York.

[63] Chen, Y.–Y., Schmidt, P., Wang, H.–J., 2014. Consistent estimation of the fixed effects stochastic frontier model. *Journal of Econometrics* 181, 65–76.

[64] Chib, S., Greenberg, E., 1995. Understanding the Metropolis Hastings algorithm. *The American Statistican* 49, 327–335.

[65] Christensen, L.R., Jorgenson, D.W., Lau, L.J., 1973. Transcendental logarithmic production frontiers. *The Review of Economics and Statistics* 55, 28–45.

[66] Coelli, T.J., 2000. On the econometric estimation of the distance function representation of a production technology. Discussion Paper 2000/42, Center for Operations Research and Econometrics, Universite Catholique de Louvain.

[67] Colombi, R., Kumbhakar, S., Martini, G.M., Vittadini, G., 2014. Closed-skew normality in stochastic frontiers with individual effects and long/short run efficiency. *Journal of Productivity Analysis* 42, 123–136.

[68] Cornwell, C., Schmidt, P., Sickles, R.C., 1990. Production frontiers with cross-sectional and time-series variation in efficiency levels. *Journal of Econometrics* 46, 185–200.

[69] Cuesta, R.A., 2000. A production model with firm-specific temporal variation in technical inefficiency: with application to Spanish dairy farms. *Journal of Productivity Analysis* 13, 139–158.

[70] Cuesta, R.A., Lovell, C.A.K., Zofio, J.L., 2009. Environmental efficiency measurement with translog distance functions: A parametric approach. *Ecological Economics* 68, 2232–2242.

[71] Cuesta, R., Orea, L., 2002. Mergers and technical efficiency in Spanish savings banks: A stochastic distance function approach. *Journal of Banking and Finance* 26, 2231–2247.

[72] Cuesta, R.A., Zofio, J.L., 2005. Hyperbolic efficiency and parametric distance functions: with application to Spanish savings banks. *Journal of Productivity Analysis* 24, 31–48.

[73] Daraio, C., Simar, L., 2014. Directional distances and their robust versions: Computational and testing issues. *European Journal of Operational Research* 237, 358–369.

[74] Darku, A.B., Malla, S., Tran, K.C., 2013. Historical review of agricultural efficiency studies. CAIRN Research Network.

[75] Debreu, G., 1951. The Coefficient of resource utilization. *Econometrica* 19, 273–92.

[76] Deprins, D., Simar, L., 1989. Estimating technical inefficiencies with correction for environmental conditions. *Annals of Public and Cooperative Economics* 60, 81–102.

[77] Dickey, D.A., Fuller, W.A., 1981. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49, 1057–1072.

[78] Diewert, W.E., 1971. An application of the shepherd duality theorem: A generalized Leontief production function. *Journal of Political Economy* 79, 481–507.

[79] Diewert, W.E., Fox, K.J., 2008. On the estimation of returns to scale, technical progress and monopolistic markups. *Journal of Econometrics* 145, 174–193.

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

[81] Dorfman, J.H., 1995. A numerical Bayesian test for cointegration of AR processes. *Journal of Econometrics* 66, 289–324.

[82] Emvalomatis, G., 2012. Adjustment and unobserved heterogeneity in dynamic stochastic frontier models. *Journal of Productivity Analysis* 37, 7–16.

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

[84] English, M., Grosskopf, S., Hayes, K., Yaisawarng, S., 1993. Output allocative and technical efficiency of the financial services sector. *Journal of Banking and Finance* 17, 349–366.

[85] Fang, X., Yang, F., 2014. Assessing Chinese commercial bank technical efficiency with a parametric hyperbolic distance function. *American Journal of Operations Research* 4, 124–131.

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

[87] Färe, R., Grosskopf, S., 2000. Notes on some inequalities in economics. *Economic Theory* 15, 227–233.

[88] Färe, R., Grosskopf, S., 2004. *New Directions: Efficiency and Productivity*. Kluwer Academic Publishers, Boston.

[89] Färe, R., Grosskopf, S., Lovell, C.A.K., 1985. *The Measurement of Efficiency of Production*. Kluwer Academic publishers, Boston.

[90] Färe, R., Grosskopf, S., Lovell, C.A.K., 1994. *Production Frontiers*. Cambridge University Press, New York.

[91] Färe, R., Grosskopf, S., Lovell, C.A.K., Pasurka, C., 1989. Multilateral productivity comparisons when some outputs are undesirable: A nonparametric approach. *The Review of Economics and Statistics* 71, 90–98.

[92] Färe, R., Grosskopf, S., Lovell, C.A.K., Yaisawarng, S., 1993. Derivation of shadow prices for undesirable outputs: A distance function approach. *The Review of Economics and Statistics* 75, 374–380.

[93] Färe, R., Grosskopf, S., Noh, D., Weber, W., 2005. Characteristics of a polluting technology: theory and practice. *Journal of Econometrics* 126, 469–492.

[94] Färe, R., Grosskopf, S., Weber, W., 2004. The effect of risk-based capital requirements on profit efficiency in banking. *Applied Economics* 36, 1731–1743.

[95] Färe, R., Grosskopf, S., Whittaker, G., 2013. Directional output distance functions: Endogenous directions based on exogenous normalization constraints. *Journal of Productivity Analysis* 40, 267–269.

[96] Färe, R., Grosskopf, S., zaim, O., 2002. Hyperbolic efficiency and return to the dollar. *European Journal of Operational Research* 136, 671–679.

[97] Färe, R., Lovell, C.A.K., 1978. Measuring the technical efficiency of production. *Journal of Economic Theory* 19, 150–162.

[98] Färe, R., Martins-Filho, C., Vardanyan, M., 2010. On functional form representation of multi-output production technologies. *Journal of Productivity Analysis* 33, 81–96.

[99] Färe, R., Primont, D., 1995. *Multi-Output Production and Duality: Theory and Applications*. Springer, Dordrecht.

[100] Färe, R., Vardanyan, M., 2016. A note on parameterizing input distance functions: Does the choice of a functional form matter? *Journal of Productivity Analysis* 45, 121–130.

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

[102] Feng, G., Serletis, A., 2008. Productivity trends in the US manufacturing: Evidence from the NQ and AIM cost functions. *Journal of Econometrics* 142, 281–311.

[103] Feng, G., Serletis, A., 2009. Efficiency and Productivity of the US Banking Industry, 1998–2005: Evidence from Fourier cost functions satisfying global regularity conditions. *Journal of Applied Econometrics* 24, 105–138.

[104] Feng, G., Serletis, A., 2010. Efficiency, technical change, and returns to scale in large US banks: Panel data evidence from an output distance function satisfying theoretical regularity. *Journal of Banking and Finance* 34, 127–138.

[105] Feng, G., Wang, C., Serletis, A. 2018. Shadow prices of CO2 emissions at US electric utilities: A random-coefficient, random-directional-vector directional output distance function approach. *Empirical Economics* 54, 231–258.

[106] Feng, Q., Horrace, W.C., 2012. Alternative technical efficiency measures: Skew, bias and scale. *Journal of Applied Econometrics* 27, 253–268.

[107] Feng, Q., Horrace, W.C., Wu, G.L. 2013. Wrong skewness and finite sample correction in parametric stochastic frontier models. Working Paper 154, Center for Policy Research, Syracuse University.

[108] Fernandez, C., Koop, G., Steel, M.F.J., 2000. A Bayesian analysis of multiple output stochastic frontiers. *Journal of Econometrics* 98, 47–79.

[109] Fernandez, C., Koop, G. , Steel, M.F.J., 2002. Multiple-output production with undesirable outputs: An application to nitrogen surplus in agriculture. *Journal of the American Statistical Association* 97, 432–442.

[110] Fernandez, C., Osiewalski, J., Steel, M., 1997. On the use of data in stochastic frontier models with improper priors. *Journal of Econometrics* 79, 169–193.

[111] Filippini, M., Greene, W.H., 2016. Persistent and transient productive inefficiency: A maximum simulated likelihood approach. *Journal of Productivity Analysis* 45, 187–196.

[112] Fried, H.O., Lovell, C.A.K., Schmidt, S.S., 2008. Efficiency and productivity. In: Fried, H.O., Lovell, C.A.K., Schmidt, S.S. (Eds.), *The Measurement of Productive Efficiency and Productivity Growth*. Oxford University Press, Oxford.

[113] Fujii, H., Managi, S., Matousek, R., 2014. Indian bank efficiency and productivity changes with undesirable outputs: A disaggregated approach. *Journal of Banking and Finance* 38, 41–50.

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

[115] Gallant, A.R., Golub, G.H., 1984. Imposing curvature restrictions on flexible functional forms. *Journal of Econometrics* 26, 295–322.

[116] Gelfand, A., Dey, D., 1994. Bayesian model choice: Asymptotic and exact calculations. *Journal of the Royal Statistical Society,* Series B 56, 501–514.

[117] Gelfand, A., Hills, S., Racine-Poon, A., Smith, A., 1990. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* 85, 972–985.

[118] Gelfand, A., Smith, A., 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85, 398–409.

[119] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. *Bayesian Data Analysis* (3th ed.). CRC press, New York.

[120] Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12, 609–628.

[121] Genz, A., Bretz, F., 2009. *Computation of Multivariate Normal and t Probabilities*. Springer, Dordrecht.

[122] Georgescu-Roegen, N., 1951. The aggregate linear production function and its application to von newman's economic model. In Koopmans, T. (Ed.), *Activity Analysis of Production and Allocation*. Wiley, New York.

[123] Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), *Bayesian Statistics*. Oxford University Press, Oxford.

[124] Geweke, J., 1993. Bayesian treatment of the independent student-t linear model. *Journal of Applied Econometrics* 8, S19-S40.

[125] Geweke, J., 1999. Using simulation methods for Bayesian econometric models: Inference, development and communication. *Econometric Reviews* 18, 1–73.

[126] Giannakas, K., Tran, K.C., Tzouvelekas, V., 2003a. On the choice of functional form in stochastic frontier modeling. *Empirical Economics* 28, 75–100.

[127] Giannakas, K., Tran, K.C., Tzouvelekas, V., 2003b. Predicting technical efficiency in stochastic production frontier models in the presence of misspecification: A Monte-Carlo analysis. *Applied Economics* 35, 153–161.

[128] Gonzalez-Farias, G., Dominguez-Molina, J.A., Gupta, A.K., 2004. The closed skew normal distribution. In: Genton, M. (Ed.), *Skew Elliptical Distributions and their Applications: A Journal beyond Normality*. Chapman and Hall/CRC, Florida.

[129] Greene, W.H., 1980a. Maximum likelihood estimation of econometric frontier functions. *Journal of Econometrics* 13, 27–56.

[130] Greene, W.H., 1980b. On the estimation of a flexible frontier production model. *Journal of Econometrics* 13, 101–115.

[131] Greene, W.H., 1990. A gamma-distributed stochastic frontier model. *Journal of Econometrics* 46, 141–163.

[132] Greene, W.H., 1993. The econometric approach to efficiency analysis. In: Fried, H.O., Lovell, C.A.K., Schmidt, S.S., (Eds), *The Measurement of Productive Efficiency: Techniques and Applications*. Oxford University Press, Oxford.

[133] Greene, W.H., 2002. *Fixed and Random Effects in Stochastic Frontier Models*. Stern School of Business, New York University.

[134] Greene, W.H., 2003. Simulated likelihood estimation of the normal-gamma stochastic frontier function. *Journal of Productivity Analysis* 19, 179–190.

[135] Greene, W.H., 2004. Distinguishing between heterogeneity and inefficiency: Stochastic frontier analysis of the world health organization's panel data on national health care systems. *Health Economics* 13, 959–980.

[136] Greene, W.H., 2005a. Fixed and random effects in stochastic frontier models. *Journal of Productivity Analysis* 23, 7–32.

[137] Greene, W.H., 2005b. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics* 126, 269–303.

[138] Griffin, R.C., Montgomery, J.M., Rister, M.E., 1987. Selecting functional form in production function analysis. *Western Journal of Agricultural Economics* 12, 216–227.

[139] Guilkey, D.K., Lovell, C.A.K., 1980. On the flexibility of the translog approximation. *International Economic Review* 21, 137–147.

[140] Guilkey, D.K., Lovell, C.A.K., Sickles, R., 1983. A comparison of the performance of three flexible functional forms. *International Economic Review* 24, 591–616.

[141] Hadri, K., 1999. Estimation of a doubly heteroskedastic stochastic frontier cost function. *Journal of Business and Economic Statistics* 17, 359–363.

[142] Hadri, K., Guermat, C., Whittaker, J., 2003. Estimation of technical inefficiency effects using panel data and doubly heteroskedastic stochastic production frontiers. *Empirical Economics* 28, 203–222.

[143] Hafner, C.M., Manner, H., Simar, L., 2018. The "wrong skewness" problem in stochastic frontier models: A new approach. *Econometric Reviews* 37, 380–400.

[144] Hampf, B., Kruger, J.J., 2015. Optimal directions for directional distance functions: An exploration of potential reductions of greenhouse gases. *American Journal of Agricultural Economics* 97, 920–938.

[145] Han, C., Phillips, P., 2010. GMM estimation for dynamic panels with fixed effects and strong instruments at unity. *Econometric Theory* 26, 119–151.

[146] Hansen, B.E., 1999. Threshold effects in non-dynamic panels: Estimation, testing, and inference. *Journal of Econometrics* 93, 345–368.

[147] Hansen, B.E., 2000. Sample splitting and threshold estimation. *Econometrica* 68, 575–603.

[148] Harris, R.D.F., Tzavalis, E., 1999. Inference for unit roots in dynamic panels where the time dimension is fixed. *Journal of Econometrics* 91, 201–226.

[149] Hastie, T., Tibshirani, R., 1993. Varying-coefficient models. *Journal of the Royal Statistical Society* 55, 757–796.

[150] Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrica* 57, 97–109.

[151] Hausman, J.A., Taylor, W.E., 1981. Panel data and unobservable individual effects. *Econometrica* 49, 1377–99.

[152] Hjalmarsson, L., Kumbhakar, S.C., Heshmati, A., 1996. DEA, DFA and SFA: A comparison. *Journal of Productivity Analysis* 7, 303–327.

[153] Horrace, W.C., 2005. Some results on the multivariate truncated normal distribution. *Journal of Multivariate Analysis* 94, 209–221.

[154] Horrace, W.C., Parmeter, C.F., 2018. A Laplace stochastic frontier model. *Econometric Reviews* 37, 260–280.

[155] Horrace, W.C., Schmidt, P., 1996. Confidence statements for efficiency estimates from stochastic frontier models. *Journal of Productivity Analysis* 7, 257–282.

[156] Hsiao, C., 2003. *Analysis of Panel Data*. Cambridge University Press, Cambridge.

[157] Huang, C.J., Liu, J.T., 1994. Estimation of a non-neutral stochastic frontier production function. *Journal of Productivity Analysis* 5, 171–180.

[158] Hudgins, L.B., Primont, D., 2007. Derivative properties of directional technology distance functions. In: Färe, R., Grosskopf, S., Primont, D. (Eds.), *Aggregation, Efficiency, and Measurement*. Springer, New York.

[159] Jondrow, J., Lovell, C.A.K., Materov, I.S., Schmidt, P., 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics* 19, 233–238.

[160] Kim, Y., Schmidt, P., 2000. A review and empirical comparison of Bayesian and classical approaches to inference on efficiency levels in stochastic frontier models with panel data. *Journal of Productivity Analysis* 14, 91–118.

[161] Kim, S., Shephard, N., Chib, S., 1998. Stochastic volatility: Likelihood inference and comparison with ARCH models. *The Review of Economic Studies* 65, 361–393.

[162] Koop, G., 1994. Recent progress in applied Bayesian econometrics. *Journal of Economic Surveys* 8, 1–34.

[163] Koop, G., Osiewalski, J., Steel, M.F., 1994. Bayesian efficiency analysis with a flexible form: The AIM cost function. *Journal of Business and Economic Statistics* 12, 339–346.

[164] Koop, G., Osiewalski, J., Steel, M.F., 1997. Bayesian efficiency analysis through individual effects: Hospital cost frontiers. *Journal of Econometrics* 76, 77–105.

[165] Koop, G., Steel, M.F., 2003. Bayesian analysis of stochastic frontier models. In: Baltagi, B. (Ed.), *A Companion to Theoretical Econometrics*. Blackwell, MA, USA.

[166] Koop, G., Steel, M.F., Osiewalsk, J., 1995. Posterior analysis of stochastic frontier models using Gibbs sampling. *Computational Statistics* 10, 353–373.

[167] Koutsomanoli-Filippaki, A., Margaritis, D., Staikouras, C., 2009. Efficiency and productivity growth in the banking industry of central and eastern Europe. *Journal of Banking and Finance* 33, 557–567.

[168] Koutsomanoli-Filippaki, A., Margaritis, D., Staikouras, C., 2012. Profit efficiency in the European union banking industry: A directional technology distance function approach. *Journal of Productivity Analysis* 37, 277–293.

[169] Kumbhakar, S.C., 1987. The specification of technical and allocative inefficiency in stochastic production and profit frontiers. *Journal of Econometrics* 34, 335–348.

[170] Kumbhakar, S.C., 1990. Production frontiers, panel data, and time-varying technical inefficiency. *Journal of Econometrics* 46, 201–211.

[171] Kumbhakar, S.C., Ghosh, S., McGuckin, J., 1991. A generalized production frontier approach for estimating determinants of inefficiency in US dairy farms. *Journal of Business and Economics Statistics* 9, 279–286.

[172] Kumbhakar, S.C., Heshmati, A., 1995. Efficiency measurement in Swedish dairy farms: An application of rotating panel data, 1976–88. *American Journal of Agricultural Economics* 77, 660–674.

[173] Kumbhakar, S.C., Lien, G., Hardaker, J.B., 2014. Technical efficiency in competing panel data models: A study of Norwegian grain farming. *Journal of Productivity Analysis* 41, 321–337.

[174] Kumbhakar, S.C., Lovell, C.A.K., 2000. *Stochastic Frontier Analysis*. Cambridge University Press, Cambridge, UK.

[175] Kumbhakar, S.C., Parmeter, C.F., Tsionas, E.G., 2013. A zero inefficiency stochastic frontier model. *Journal of Econometrics* 172, 66–76.

[176] Kumbhakar, S.C., Wang, H.–J., 2005. Estimation of growth convergence using a stochastic production function approach. *Economics Letters* 88, 300–305.

[177] Kuosmanen, T., Fosgerau, M., 2009. Neoclassical versus frontier production models? testing for the skewness of regression residuals. *The Scandinavian Journal of Economics* 111, 351–367.

[178] Lau, L.J., 1978a. Application of profit functions. In: Fuss, M., McFadden, D. (Eds.), *Production Economics: A Dual Approach to Theory and Applications*. North-Holland Publishing Co, Amsterdam.

[179] Lau, L.J., 1978b. Testing and imposing monotonicity, convexity, and quasi-convexity constraints. In: Fuss, M., McFadden, D. (Eds.), *Production Economics: A Dual Approach to Theory and Applications*. North-Holland Publishing Co., Amsterdam

[180] Lau, L.J., 1986. Functional forms in econometric model building. In: Griliches, Z., Intriligator, M.D. (Eds.), *Handbook of Econometrics*. Elsevier Science, Amsterdam.

[181] Lee, S., Lee, Y.H., 2014. Stochastic frontier models with threshold efficiency. *Journal of Productivity Analysis* 42, 45–54.

[182] Lee, Y.H., 1996. Tail truncated stochastic frontier models. *Journal of Economic Theory and Econometrics* 2, 137–152.

[183] Lee, Y.H., Schmidt, P., 1993. A production frontier model with flexible temporal variation in technical efficiency. In: Fried, H.O., Lovell, C.A.K., Schmidt, S.S. (Eds.), *The Measurement of Productive Efficiency: Techniques and Applications*. Oxford University Press, Oxford.

[184] Lewbel, A., Ng, S., 2005. Demand systems with nonstationary prices. *Review of Economics and Statistics* 87, 479–494.

[185] Luenberger, D., 1992. Benefit functions and duality. *Journal of Mathematical Economics* 21, 461–481.

[186] Luenberger, D., 1995. *Microeconomic Theory*. McGraw Hill, New York.

[187] Maddala, G.S., Wu, S., 1999. A comparative study of unit root tests with panel data and a new simple test. *Oxford Bulletin of Economics and Statistics* 61, 631–652.

[188] Malikov, E., Kumbhakar, S.C., Tsionas, E.G., 2016. A cost system approach to the stochastic directional technology distance function with undesirable outputs: The case of U.S. banks in 2001–2010. *Journal of Applied Econometrics* 31, 1407–1429.

[189] Marsh, T.L., Featherstone, A.M., Garrett, T.A., 2003. Input inefficiency in commercial banks: A normalized quadratic input distance approach. Working Paper 2003-036A Federal Reserve Bank of st. Louis. http://research.stlouisfed.org/wp/2003/2003-036.pdf

[190] McCulloch, R., Rossi, P.E., 1994. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics* 64, 207–240.

[191] Meeusen, W., Van den Broeck, J., 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review* 18, 435–444.

[192] Mester, L., 1996. A study of bank efficiency taking into account risk-preferences. *Journal of Banking and Finance* 20, 1025–1045.

[193] Mester, L., 1997. Measuring efficiency at U.S. banks: Accounting for heterogeneity is important. *European Journal of Operational Research* 98, 230–242.

[194] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21, 1087–1092.

[195] Morey, E.R., 1986. An introduction to checking, testing, and imposing curvature properties: The true function and the estimated function. *Canadian Journal of Economics* 19, 207–235.

[196] Mundlak, Y., 1961. Empirical production function free of management bias. *Journal of Farm Economics* 43, 44–56.

[197] Nahm, D., Vu, H.T., 2013. Measuring scale efficiency from a parametric hyperbolic distance function. *Journal of Productivity Analysis* 39, 83–88.

[198] Neyman, J., Scott, E.L., 1948. Consistent estimation from partially consistent observations. *Econometrica* 16, 1–32.

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

[200] O'Donnell, C.J., 2014. Econometric estimation of distance functions and associated measures of productivity and efficiency change. *Journal of Productivity Analysis* 41, 187–200.

[201] O'Donnell, C.J., Coelli, T.J., 2005. A Bayesian approach to imposing curvature on distance functions. *Journal of Econometrics* 126, 493–523.

[202] Orea, L., Kumbhakar, S.C., 2004. Efficiency measurement using a latent class stochastic frontier model. *Empirical Economics* 29, 169–183.

[203] Osiewalski, J., Steel, M.F., 1998. Numerical tools for the Bayesian analysis of stochastic frontier models. *Journal of Productivity Analysis* 10, 103–117.

[204] Park, J.Y., Hahn, S.B., 1999. Cointegrating regressions with time varying coefficients. *Econometric Theory* 15, 664–703.

[205] Park, K., Weber, W., 2006. A note on efficiency and productivity growth in the Korean banking industry, 1992–2002. *Journal of Banking and Finance* 30, 2371– 2386.

[206] Parmeter, C., Kumbhakar, S.C., 2014. Efficiency analysis: A primer on recent advances. *Foundations and Trends in Econometrics* 7, 191–385.

[207] Pedroni, P., 2001. Fully modified OLS for heterogeneous cointegrated panels. In: Baltagi, B.H., Fomby, T.B., Hill, R.C. (Eds.), *Non- Stationary Panels, Panel Cointegration, and Dynamic Panels (Advances in Econometrics)*. Emerald Group Publishing Limited.

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

[209] Phillips, P.C.B., 1995. Fully modified least squares and vector autoregression. *Econometrica* 63, 1023–1078.

[210] Phillips, P.C.B., Hansen, L.P., 1990. Statistical inference in instrumental variables regression with I(1) processes. *The Review of Economic Studies* 57, 99–125.

[211] Phillips, P.C.B., Moon, H., 1999. Linear regression limit theory for nonstationary panel data. Econometrica 67, 1057–1111.

[212] Pitt, M.M., Lee, L.–F., 1981. The measurement and sources of technical inefficiency in the Indonesian weaving industry. *Journal of Development Economics* 9, 43–64.

[213] Reifschneider, D., Stevenson, R., 1991. Systematic departures from the frontier: A framework for the analysis of firm inefficiency. *International Economic Review* 32, 715–723.

[214] Rho, S., Schmidt, P., 2015. Are all firms inefficient? *Journal of Productivity Analysis* 43, 327–349.

[215] Ritter, C., Simar, L., 1997. Pitfalls of normal-gamma stochastic frontier models. *Journal of Productivity Analysis* 8, 167–182.

[216] Roberts, G.O., Gelman, A., Gilks, W.R., 1997. Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability* 7, 110–120.

[217] Roberts, G.O., Smith, A.F.M., 1994. Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms. *Stochastic Processes and their Applications* 49, 207–216.

[218] Ryan, D.L., Wales, T.J., 1998. A simple method for imposing local curvature in some flexible consumer demand systems. *Journal of Business and Economic Statistics* 16, 331–338.

[219] Ryan, D.L., Wales, T.J., 2000. Imposing local concavity in the translog and generalized Leontief cost functions. *Economics Letters* 67, 253–260.

[220] Satchachai, P., Schmidt, P., 2010. Estimates of technical inefficiency in stochastic frontier models with panel data: Generalized panel jackknife estimation. *Journal of Productivity Analysis* 34, 83–97.

[221] Schmidt, P., Sickles, R.C., 1984. Production frontiers and panel data. *Journal of Business and Economic Statistics* 2, 367–374.

[222] Sealey, C.W., Lindley, J.T., 1977. Inputs, outputs, and a theory of production and cost at depository financial institutions. *Journal of Finance* 32, 1251–1266.

[223] Serletis, A., Shahmoradi, A., 2007. Flexible Functional Forms, Curvature Conditions, and the Demand for Assets. *Macroeconomic Dynamics* 11, 455–486.

[224] Shephard, R., 1953. *Cost and Production Functions*. Princeton University Press, Princeton.

[225] Shephard, R., 1970. *Theory of Cost and Production Functions*. Princeton University Press, Princeton.

[226] Sickles, R.C., 2005. Panel estimators and the identification of firm specific efficiency levels in parametric, semiparametric and nonparametric settings. *Journal of Econometrics* 126, 305–324.

[227] Simar, L., Vanhems, A., 2012. Probabilistic characterization of directional distances and their robust versions. *Journal of Econometrics* 166, 342–354.

[228] Smith, A.F.M., Roberts, C.O. , 1993. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society* B55, 3–23.

[229] Srairi, S.A., 2010. Cost and profit efficiency of conventional and Islamic banks in GCC countries. *Journal of Productivity analysis* 34, 45–62.

[230] Stevenson, R.E., 1980. Likelihood functions for generalized stochastic frontier estimation. *Journal of Econometrics* 13, 57–66.

[231] Stiroh, K.J., 2000. How did bank holding companies prosper in the 1990s? *Journal of Banking and Finance* 24, 1703–1745.

[232] Stock, J.H., 1994. Unit roots, structural breaks and trends. In: Engel, R., McFadden, D. (Eds.), *Handbook of Econometrics*. Elsevier, Amsterdam.

[233] Stock, J.H., Watson, M.W., 1993. A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica* 61, 783–820.

[234] Sturm, J.–E., Williams, B. , 2008. Characteristics determining the efficiency of foreign banks in Australia. *Journal of Banking and Finance* 32, 2346–2360.

[235] Swamy, P.A.V.B., Tavlas, G.S. 1995. Random coefficient models: Theory and applications. *Journal of Economic Surveys* 9, 165–96.

[236] Terrell, D., 1996. Incorporating monotonicity and concavity conditions in flexible functional forms. *Journal of Applied Econometrics* 11, 179–194.

[237] Thompson, G.D., 1988. Choice of flexible functional forms: Review and appraisal. *Western Journal of Agricultural Economics* 13, 169–183.

[238] Tran, K.C., 2014. Nonparametric estimation of functional-coefficient dynamic panel data model with fixed effects. *Economics Bulletin* 34, 1751–1761.

[239] Tran, K.C., Tsionas, E.G., 2013. GMM estimation of stochastic frontier model with endogenous regressors. *Economics Letters* 118, 233–236.

[240] Tran, K.C., Tsionas, E.G., 2015. Endogeneity in stochastic frontier models: Copula approach without external instruments. *Economics Letters* 133, 85–88.

[241] Tran, K.C., Tsionas, E.G., 2016a. On the estimation of zero-inefficiency stochastic frontier models with endogenous regressors. *Economics Letters* 147, 19–22.

[242] Tran, K.C., Tsionas, E.G., 2016b. Zero-inefficiency stochastic frontier model with varying mixing proportion: Semiparametric approach. *European Journal of Operational and Research* 249, 1113–1123.

[243] Tsionas, E.G., 2002. Stochastic frontier models with random coefficients. *Journal of Applied Econometrics* 17, 127–147.

[244] Tsionas, E.G., 2006. Inference in dynamic stochastic frontier models. *Journal of Applied Econometrics* 21, 669–676.

[245] Tsionas, E.G., Christopoulos D., 2001. Efficiency measurement with nonstationary variables: An application of panel cointegration techniques. *Economics Bulletin* 3,

1–7.

[246] Tsionas, E.G., Kumbhakar, S.C., 2014. Firm heterogeneity, persistent and transient technical inefficiency. *Journal of Applied Econometrics* 29, 110–132.

[247] Tsionas, E.G., Kumbhakar, S.C., Malikov, E., 2015. Estimation of input distance functions: A system approach. *American Journal of Agricultural Economics* 97, 1478–1493.

[248] Tsionas, E.G., Tran, K.C., Michaelides, P., 2017. Bayesian inference in threshold stochastic frontier models. *Empirical Economics* (Forthcoming).

[249] Tzeremes, N.G., 2015. Efficiency dynamics in Indian banking: A conditional directional distance approach. *European Journal of Operational Research* 240, 807–818.

[250] Van den Broeck, J., Koop, G., Osiewalski, J., Steel, M.F., 1994. Stochastic frontier models: A Bayesian perspective. *Journal of Econometrics* 61, 273–303.

[251] Waldman, D., 1982. A stationary point for the stochastic frontier likelihood. *Journal of Econometrics* 18, 275–279.

[252] Wales, T.J., 1977. On the flexibility of flexible functional forms. *Journal of Econometrics* 5, 183–193.

[253] Wang, H.–J., 2002. Heteroskedasticity and non-monotonic efficiency effects of a stochastic frontier model. *Journal of Productivity Analysis* 18, 241–253.

[254] Wang, H.–J., Ho, C.W., 2010. Estimating fixed-effect panel data stochastic frontier models by model transformation. *Journal of Econometrics* 157, 286–296.

[255] Wang, H.–J., Schmidt, P., 2002. One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. *Journal of Productivity Analysis* 18, 129–44.

[256] Wang, W.S., Schmidt, P., 2009. On the distribution of estimated technical efficiency in stochastic frontier models. *Journal of Econometrics* 148, 36–45.

[257] Watson, M.W., 1994. Vector autoregressions and cointegration. In: Engel, R., McFadden, D. (Eds.), *Handbook of Econometrics*, Elsevier: Amsterdam.

[258] Wheat, P., Greene, B., Smith, A. 2014. Understanding prediction intervals for firm specific inefficiency scores from parametric stochastic frontier models. *Journal of Productivity Analysis* 42, 55–65.

[259] Wikstrom, D., 2016. Modified Fixed Effects Estimation of Technical Inefficiency. *Journal of Productivity Analysis* 46, 83–86.

[260] Wolff, H., 2016. Imposing and testing for shape restrictions in flexible parametric models. *Econometric Reviews* 35, 1013–1039.

[261] Wolff, H., Heckelei, T., Mittelhammer, R.C., 2010. Imposing curvature and monotonicity on flexible functional forms: An efficient regional approach. *Computational Economics* 37, 309–339.

[262] Yélou, C., Larue, B., Tran, K.C., 2010. Threshold effects in panel data stochastic frontier models of dairy production in Canada. *Economic Modelling* 27, 641–647.

[263] Zofio, J.L., Pastor, J.T., Aparicio, J., 2013. The directional profit efficiency measure: On why profit inefficiency is either technical or allocative. *Journal of Productivity Analysis* 40, 257–266.

Appendix A

A.0.1 Proof of the Parameter Restrictions that Impose the Translation Property on the Directional Distance Functions

Since there are two inputs and two outputs in the empirical implementation, then applying the quadratic functional form to the left hand side of (3.2) yields

$$
\begin{aligned}
&\vec{D}_T\left(x-\alpha g_x, y+\alpha g_y, t; g_x, g_y\right) \\
&= \alpha_0 + \sum_{n=1}^{2}\alpha_n\left(x_n-\alpha g_{x_n}\right) + \sum_{m=1}^{2}\beta_m\left(y_m+\alpha g_{y_m}\right) \\
&+ \delta_t t + \frac{1}{2}\sum_{n=1}^{2}\sum_{n'=1}^{2}\alpha_{nn'}\left(x_n-\alpha g_{x_n}\right)\left(x_{n'}-\alpha g_{x_{n'}}\right) \\
&+ \frac{1}{2}\sum_{m=1}^{2}\sum_{m'=1}^{2}\beta_{mm'}\left(y_m+\alpha g_{y_m}\right)\left(y_{m'}+\alpha g_{y_{m'}}\right) \\
&+ \frac{1}{2}\delta_{tt}t^2 + \sum_{n=1}^{2}\sum_{m=1}^{2}\gamma_{nm}\left(x_n-\alpha g_{x_n}\right)\left(y_m+\alpha g_{y_m}\right) \\
&+ \sum_{n=1}^{2}\delta_{tx_n}t\left(x_n-\alpha g_{x_n}\right) + \sum_{m=1}^{2}\delta_{ty_m}t\left(y_m+\alpha g_{y_m}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \alpha_0 + \sum_{n=1}^{2}\alpha_n x_n + \sum_{m=1}^{2}\beta_m y_m + \delta_t t + \frac{1}{2}\sum_{n=1}^{2}\sum_{n'=1}^{2}\alpha_{nn'}x_n x_{n'} \\
&+ \frac{1}{2}\sum_{m=1}^{2}\sum_{m'=1}^{2}\beta_{mm'}y_m y_{m'} + \frac{1}{2}\delta_{tt}t^2 + \sum_{n=1}^{2}\sum_{m=1}^{2}\gamma_{nm}x_n y_m \\
&+ \sum_{n=1}^{2}\delta_{tx_n}t x_n + \sum_{m=1}^{2}\delta_{ty_m}t y_m \\
&+ \begin{bmatrix}
-\sum_{n=1}^{2}\alpha_n\alpha g_{x_n} + \sum_{m=1}^{2}\beta_m\alpha g_{y_m} \\
+\frac{1}{2}\sum_{n=1}^{2}\sum_{n'=1}^{2}\alpha_{nn'}\left(-x_n\alpha g_{x_{n'}} - x_{n'}\alpha g_{x_n} + \alpha^2 g_{x_n}g_{x_{n'}}\right) \\
+\frac{1}{2}\sum_{m=1}^{2}\sum_{m'=1}^{2}\beta_{mm'}\left(y_m\alpha g_{y_{m'}} + y_{m'}\alpha g_{y_m} + \alpha^2 g_{y_m}g_{y_{m'}}\right) \\
+\sum_{n=1}^{2}\sum_{m=1}^{2}\gamma_{nm}\left(x_n\alpha g_{y_m} - y_m\alpha g_{x_n} - \alpha^2 g_{x_n}g_{y_m}\right) \\
-\sum_{n=1}^{2}\delta_{tx_n}t\alpha g_{x_n} + \sum_{m=1}^{2}\delta_{ty_m}t\alpha g_{y_m}
\end{bmatrix}
\end{aligned}
\tag{A.1}
$$

Substitute the first three lines on the right-hand side of (A.1) with $\vec{D}_T\left(x,y,t;g_x,g_y\right)$ yields

$$\vec{D}_T\left(x-\alpha g_x,y+\alpha g_y,t;g\right)=\vec{D}_T\left(x,y,t;g_x,g_y\right)+\left[\begin{array}{c}-\sum_{n=1}^{2}\alpha_n\alpha g_{x_n}+\sum_{m=1}^{2}\beta_m\alpha g_{y_m}\\ +\frac{1}{2}\sum_{n=1}^{2}\sum_{n'=1}^{2}\alpha_{nn'}\left(-x_n\alpha g_{x_{n'}}-x_{n'}\alpha g_{x_n}+\alpha^2 g_{x_n}g_{x_{n'}}\right)\\ +\frac{1}{2}\sum_{m=1}^{2}\sum_{m'=1}^{2}\beta_{mm'}\left(y_m\alpha g_{y_{m'}}+y_{m'}\alpha g_{y_m}+\alpha^2 g_{y_m}g_{y_{m'}}\right)\\ +\sum_{n=1}^{2}\sum_{m=1}^{2}\gamma_{nm}\left(x_n\alpha g_{y_m}-y_m\alpha g_{x_n}-\alpha^2 g_{x_n}g_{y_m}\right)\\ -\sum_{n=1}^{2}\delta_{tx_n}t\alpha g_{x_n}+\sum_{m=1}^{2}\delta_{ty_m}t\alpha g_{y_m}\end{array}\right] \tag{A.2}$$

To satisfy the translation property in equation (3.2), the second term on the right-hand side of (A.2) should be $-\alpha$, that is

$$\begin{aligned}&-\sum_{n=1}^{2}\alpha_n\alpha g_{x_n}+\sum_{m=1}^{2}\beta_m\alpha g_{y_m}\\&+\frac{1}{2}\sum_{n=1}^{2}\sum_{n'=1}^{2}\alpha_{nn'}\left(-x_n\alpha g_{x_{n'}}-x_{n'}\alpha g_{x_n}+\alpha^2 g_{x_n}g_{x_{n'}}\right)\\&+\frac{1}{2}\sum_{m=1}^{2}\sum_{m'=1}^{2}\beta_{mm'}\left(y_m\alpha g_{y_{m'}}+y_{m'}\alpha g_{y_m}+\alpha^2 g_{y_m}g_{y_{m'}}\right)\\&+\sum_{n=1}^{2}\sum_{m=1}^{2}\gamma_{nm}\left(x_n\alpha g_{y_m}-y_m\alpha g_{x_n}-\alpha^2 g_{x_n}g_{y_m}\right)\\&-\sum_{n=1}^{2}\delta_{tx_n}t\alpha g_{x_n}+\sum_{m=1}^{2}\delta_{ty_m}t\alpha g_{y_m}\\&=-\alpha\end{aligned} \tag{A.3}$$

Following Feng *et al.* (2018) and dividing both sides of (A.3) by α and rearranging terms yields

$$\begin{aligned}
&\left(\sum_{m=1}^{2}\beta_m g_{y_m} - \sum_{n=1}^{2}\alpha_n g_{x_n} + 1\right) + \left(\sum_{m=1}^{2}\gamma_{nm} g_{y_m} - \sum_{n'=1}^{2}\alpha_{nn'} g_{x_{n'}}\right) x_n \\
&+ \left(\sum_{m'=1}^{2}\beta_{mm'} g_{y_{m'}} - \sum_{n=1}^{2}\gamma_{nm} g_{x_n}\right) y_m + \left(\sum_{m=1}^{2}\delta_{ty_m} g_{y_m} - \sum_{n=1}^{2}\delta_{tx_n} g_{x_n}\right) t \\
&+ \left(\frac{1}{2}\sum_{n=1}^{2}\sum_{n'=1}^{2}\alpha_{nn'} g_{x_n} g_{x_{n'}} + \frac{1}{2}\sum_{m=1}^{2}\sum_{m'=1}^{2}\beta_{mm'} g_{y_m} g_{y_{m'}} - \sum_{n=1}^{2}\sum_{m=1}^{2}\gamma_{nm} g_{x_n} g_{y_m}\right)\alpha \\
&= 0 \qquad \text{(A.4)}
\end{aligned}$$

When $g_{y1} = 1$ and $\alpha = -y_1$, (A.4) becomes

$$\begin{aligned}
&\left(\beta_1 + \beta_2 g_{y_2} - \sum_{n=1}^{2}\alpha_n g_{x_n} + 1\right) + \left(\gamma_{n1} + \gamma_{n2} g_{y_2} - \sum_{n'=1}^{2}\alpha_{nn'} g_{x_{n'}}\right) x_n \\
&+ \left(-\frac{1}{2}\sum_{n=1}^{2}\sum_{n'=1}^{2}\alpha_{nn'} g_{x_n} g_{x_{n'}} + \frac{1}{2}\left(\beta_{11} - \beta_{22} g_{y_2}^2\right) + \sum_{n=1}^{2}\gamma_{n2} g_{y_2} g_{x_n}\right) y_1 \\
&+ \left(\beta_{21} + \beta_{22} g_{y2} - \sum_{n=1}^{2}\gamma_{n2} g_{x_n}\right) y_2 + \left(\delta_{ty1} + \delta_{ty2} g_{y_2} - \sum_{n=1}^{2}\delta_{tx_n} g_{x_n}\right) t \\
&= 0 \qquad \text{(A.5)}
\end{aligned}$$

Then, for (A.5) to hold, the following restrictions must be satisfied:

$$\beta_1 + \beta_2 g_{y_2} - \sum_{n=1}^{2}\alpha_n g_{x_n} = -1, \qquad \gamma_{n1} + \gamma_{n2} g_{y_2} - \sum_{n'=1}^{2}\alpha_{nn'} g_{x_{n'}} = 0,$$

$$\beta_{21} + \beta_{22} g_{y2} - \sum_{n=1}^{2}\gamma_{n2} g_{x_n} = 0, \qquad \delta_{ty1} + \delta_{ty2} g_{y_2} - \sum_{n=1}^{2}\delta_{tx_n} g_{x_n} = 0, \text{ and}$$

$$\sum_{n=1}^{2}\gamma_{n2} g_{y_2} g_{x_n} - \frac{1}{2}\sum_{n=1}^{2}\sum_{n'=1}^{2}\alpha_{nn'} g_{x_n} g_{x_{n'}} + \frac{1}{2}\left(\beta_{11} - \beta_{22} g_{y_2}^2\right) = 0, \ (n = 1, 2).$$

When $g_{x1} = 1$, $\alpha = x_1$, and $g_y = 0$, (A.4) becomes

$$\begin{aligned}
&(1 - \alpha_1 - \alpha_2 g_{x_2}) - (\gamma_{1m} + \gamma_{2m} g_{x_2})\, y_m - \left(\frac{1}{2}\left(\alpha_{11} - \alpha_{22} g_{x2}^2\right)\right) x_1 \\
&- (\alpha_{21} + \alpha_{22} g_{x_2})\, x_2 - (\delta_{tx_1} + \delta_{tx_2} g_{x_2})\, t \\
&= 0
\end{aligned} \tag{A.6}$$

Then, for (A.6) to hold, the following restrictions must be satisfied:

$$\alpha_1 + \alpha_2 g_{x_2} = 1, \qquad \gamma_{1m} + \gamma_{2m} g_{x_2} = 0, \quad \alpha_{11} - \alpha_{22} g_{x2}^2 = 0,$$

$$\alpha_{21} + \alpha_{22} g_{x_2} = 0, \quad \text{and} \quad \delta_{tx_1} + \delta_{tx_2} g_{x_2} = 0, \qquad (m = 1, 2).$$

When $g_{y1} = 1$, $\alpha = -y_1$, and $g_x = 0$, the following restrictions must be satisfied for (A.5) to hold:

$$\beta_1 + \beta_2 g_{y_2} = -1, \qquad \gamma_{n1} + \gamma_{n2} g_{y_2} = 0, \quad \beta_{11} - \beta_{22} g_{y_2}^2 = 0,$$

$$\beta_{21} + \beta_{22} g_{y2} = 0, \quad \text{and} \quad \delta_{ty1} + \delta_{ty2} g_{y_2} = 0, \qquad (n = 1, 2).$$